BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN VĂN THỊNH

PHÁT TRIỂN PHƯƠNG PHÁP CHÚ THÍCH ẢNH DỰA TRÊN MẠNG HỌC SÂU

LUẬN ÁN TIẾN SĨ MÁY TÍNH

Hà Nội - 2026

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN VĂN THỊNH

PHÁT TRIỂN PHƯƠNG PHÁP CHÚ THÍCH ẢNH DỰA TRÊN MẠNG HỌC SÂU

LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Khoa học máy tính

Mã số: 9 48 01 01

Xác nhận của Học viện Khoa học và Công nghệ	**Người hướng dẫn 1** *(Ký và ghi rõ họ tên)*	**Người hướng dẫn 2** *(Ký và ghi rõ họ tên)*
	PGS. TS. Trần Văn Lăng	**TS. Văn Thế Thành**

Hà Nội - 2026

LỜI CAM ĐOAN

Tôi xin cam đoan luận án: "Phát triển phương pháp chú thích ảnh dựa trên mạng học sâu" là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của tập thể hướng dẫn. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ nguồn gốc. Các kết quả nghiên cứu của tôi được công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tôi làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hà Nội, ngày tháng năm 2026

Tác giả luận án

(Ký và ghi rõ họ tên)

NCS. Nguyễn Văn Thịnh

LỜI CẢM ƠN

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc và sự tri ân đặc biệt đến PGS.TS. Trần Văn Lăng, người Thầy hướng dẫn chính đã tận tình dìu dắt và đồng hành cùng tôi trong suốt thời gian thực hiện luận án. Với kiến thức chuyên môn sâu rộng và tư duy khoa học chặt chẽ, Thầy đã định hướng các vấn đề học thuật quan trọng và có ảnh hưởng lớn đến sự hình thành cũng như phát triển năng lực khoa học của tôi. Tôi luôn trân trọng và ghi nhớ rằng, nếu không có sự dẫn dắt và hỗ trợ tận tâm của Thầy, tôi khó có thể trưởng thành trên con đường học thuật và hoàn thành công trình này.

Tôi cũng xin chân thành cảm ơn TS. Văn Thế Thành, người Thầy đã đồng hành, định hướng và khích lệ tôi trong những bước đầu trên con đường học thuật. Thầy đã luôn dành cho tôi sự hỗ trợ và động viên quý báu. Sự giúp đỡ của Thầy đã tạo nền tảng quan trọng để tôi tiếp tục theo đuổi con đường nghiên cứu khoa học.

Tôi xin trân trọng cảm ơn Viện Hàn lâm Khoa học và Công nghệ Việt Nam, cùng các đơn vị trực thuộc như Viện Cơ học và Tin học Ứng dụng, Viện Công nghệ Thông tin, và đặc biệt là Học viện Khoa học và Công nghệ, đã tạo điều kiện thuận lợi về môi trường học thuật, cơ sở vật chất và các thủ tục cần thiết trong thời gian tôi thực hiện luận án.

Tôi cũng xin gửi lời cảm ơn chân thành đến Ban Giám hiệu Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Ban Lãnh đạo Khoa Công nghệ Thông tin, cùng các phòng ban chức năng và toàn thể đồng nghiệp trong Trường đã luôn quan tâm, hỗ trợ và tạo điều kiện thuận lợi để tôi vừa hoàn thành nhiệm vụ chuyên môn, vừa theo đuổi chương trình nghiên cứu sinh.

Tôi xin chân thành cảm ơn các thầy cô, đồng nghiệp và bạn bè đã chia sẻ kiến thức chuyên môn và luôn động viên tôi trong thời gian học tập.

Đặc biệt, tôi xin dành lời tri ân sâu sắc nhất đến gia đình thân yêu, những người luôn là điểm tựa tinh thần vững chắc và nguồn động viên lớn lao, giúp tôi kiên trì theo đuổi con đường học thuật và hoàn thành chặng đường nghiên cứu này.

Hà Nội, ngày tháng năm 2026

Tác giả luận án

(Ký và ghi rõ họ tên)

NCS. Nguyễn Văn Thịnh

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Từ viết tắt	Diễn giải tiếng Anh	Diễn giải tiếng Việt
AMR	Abstract Meaning Representation	Biểu diễn ngữ nghĩa trừu tượng
AMR-GT	Ground-truth Caption-based AMR	AMR chuyển đổi từ chú thích chuẩn (ground truth)
AMR-RG	Relationship Graph-based AMR	AMR chuyển đổi từ đồ thị quan hệ
AMR-GT&RG	AMR from Ground Truth Captions and Relationship Graph	Mô hình kết hợp hai nguồn AMR
BERT	Bidirectional Encoder Representations from Transformers	Mô hình biểu diễn mã hóa hai chiều từ Transformer
BLEU	Bilingual Evaluation Understudy	Độ đo BLEU
BLIP	Bootstrapping Language-Image Pretraining	Mô hình tiền huấn luyện ngôn ngữ-hình ảnh
CLIP	Contrastive Language-Image Pretraining	Mô hình huấn luyện tương phản ngôn ngữ và hình ảnh
CLIP-AMR-GPT	Contrastive Language-Image Pretraining, Abstract Meaning Representation, and Generative Pre-trained Transformer-based Image Captioning Model	Mô hình chú thích ảnh tích hợp đặc trưng từ CLIP, biểu diễn ngữ nghĩa trừu tượng AMR và tái xếp hạng ngôn ngữ bằng GPT
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
ConceptNet	ConceptNet Knowledge Graph	Đồ thị tri thức ConceptNet
Cross-Attention	Cross-Attention Mechanism	Cơ chế chú ý chéo giữa các nguồn dữ liệu
CV	Computer Vision	Thị giác máy tính

Từ viết tắt	Diễn giải tiếng Anh	Diễn giải tiếng Việt
Dual Attention	Dual Attention Mechanism	Cơ chế chú ý kép
GAT	Graph Attention Network	Mạng chú ý trên đồ thị
GCN	Graph Convolutional Network	Mạng nơ-ron tích chập đồ thị
GraphSAGE	Graph Sample and Aggregate	Phương pháp lấy mẫu và tổng hợp trên đồ thị
GT	Ground Truth	Dữ liệu chuẩn (gán nhãn thủ công)
LSTM	Long Short-Term Memory	Mạng bộ nhớ ngắn-dài hạn
AI	Artificial Intelligence	Trí tuệ nhân tạo
METEOR	Metric for Evaluation of Translation with Explicit ORdering	Độ đo METEOR
MS COCO	Microsoft Common Objects in Context	Tập dữ liệu MS COCO
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
ODwGCN	Object Detection with GCN	Mô hình phát hiện đối tượng kết hợp GCN
RGTranCNet	Relation Graph Transformer Network with ConceptNet	Mô hình chú thích ảnh sử dụng mạng Transformer với đồ thị quan hệ và ConceptNet
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation - LCS	Độ đo ROUGE-L
Scene Graph	Scene Graph	Đồ thị cảnh

Từ viết tắt	Diễn giải tiếng Anh	Diễn giải tiếng Việt
SPICE	Semantic Propositional Image Caption Evaluation	Độ đo SPICE
Transformer	Transformer Architecture	Kiến trúc Transformer
VRP	Visual Relationship Prediction	Dự đoán quan hệ thị giác

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

DANH MỤC THUẬT TOÁN

MỞ ĐẦU

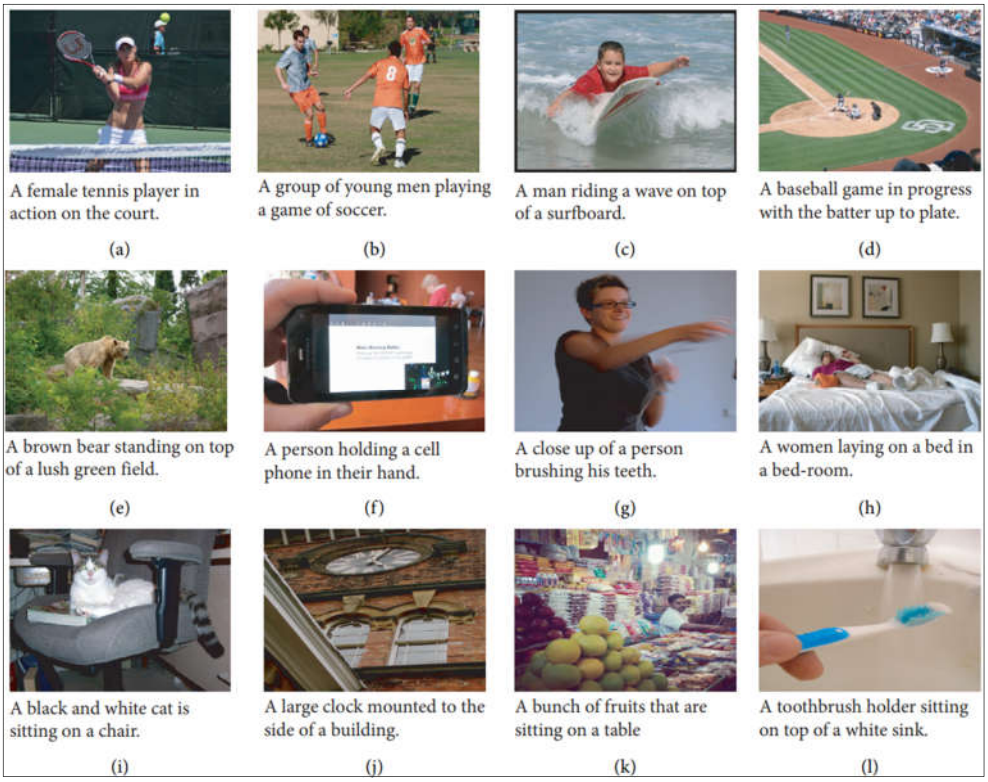
Chú thích ảnh tự động là bài toán tiêu biểu trong trí tuệ nhân tạo hiện đại, đòi hỏi phối hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên. Với sự phát triển nhanh của học sâu đa phương thức, bài toán này ngày càng thu hút sự quan tâm nhờ ứng dụng rộng rãi trong y tế, giáo dục, truyền thông và hỗ trợ người khiếm thị. Tuy nhiên, các mô hình hiện tại còn hạn chế trong việc khai thác đặc trưng hình ảnh, biểu diễn quan hệ ngữ nghĩa giữa các đối tượng, và tích hợp tri thức ngoài (tri thức bên ngoài tập dữ liệu huấn luyện). Do đó, phát triển phương pháp chú thích ảnh dựa trên mạng học sâu - kết hợp hiệu quả đặc trưng hình ảnh, tri thức ngữ nghĩa và biểu diễn trừu tượng - là hướng đi cần thiết. Phần Mở đầu trình bày lần lượt: (1) Tính cấp thiết của luận án; (2) Mục tiêu nghiên cứu; (3) Đối tượng và phạm vi nghiên cứu; (4) Phương pháp nghiên cứu; (5) Đóng góp chính của luận án; và (6) Bố cục luận án nhằm định hướng cho toàn bộ nội dung trình bày trong các chương tiếp theo.

1. Tính cấp thiết của luận án

Trí tuệ nhân tạo (Artificial Intelligence - AI), đặc biệt là học sâu (Deep Learning - DL), đã có những bước phát triển vượt bậc, góp phần đưa các hệ thống máy tính đến gần hơn với khả năng xử lý và nhận thức thông tin đa phương thức như con người [1]. Trong xu thế này, hai lĩnh vực vốn phát triển tương đối độc lập là thị giác máy tính (Computer Vision - CV) và xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) ngày càng được tích hợp nhằm giải quyết các bài toán phức tạp đòi hỏi sự phối hợp giữa khả năng "nhìn thấy" và "diễn đạt" của máy. Một trong những bài toán điển hình của sự kết hợp này là *chú thích ảnh* (Image Captioning - IC) - trong đó hệ thống cần phân tích nội dung hình ảnh đầu vào và sinh ra mô tả (chú thích) tương ứng dưới dạng ngôn ngữ tự nhiên [2-4].

Về mặt khái niệm, chú thích ảnh tự động được hiểu là quá trình tự động tạo ra mô tả dưới dạng ngôn ngữ tự nhiên sao cho phản ánh chính xác nội dung cốt lõi của ảnh, bao gồm các đối tượng và mối quan hệ giữa chúng [5, 6]. Đây là một nhiệm vụ học đa phương thức tiêu biểu, kết hợp hai lĩnh vực quan trọng của AI: thị giác máy tính để giải quyết việc hiểu ảnh và xử lý ngôn ngữ tự nhiên nhằm phát sinh mô tả cho hình ảnh đúng cú pháp và ngữ nghĩa [7]. Hình MĐ.1 minh họa đầu vào là một ảnh, và đầu ra là câu mô tả được sinh tự động bởi mô hình chú thích ảnh [8]. Nhờ khả năng gắn kết hai miền thông tin khác nhau (thị giác và ngôn ngữ), các mô hình chú thích ảnh đã được ứng dụng rộng rãi trong nhiều lĩnh vực, từ lập chỉ mục ảnh để tìm kiếm theo nội dung [9], hỗ trợ người khiếm thị [10], hỗ trợ chẩn đoán và kê đơn thuốc tự động trong ảnh y khoa [11], phân tích dữ liệu giao thông [12], chẩn đoán bệnh

cây trồng [13], điều hướng giao thông [14], đến việc sinh mô tả tự động cho ảnh trên các nền tảng mạng xã hội [15], trong kiểm duyệt nội dung [16] và an toàn lao động [17].



Hình MĐ.1. Một số ví dụ về ảnh đầu vào và câu mô tả được sinh ra bởi mô hình chú thích ảnh tự động. Mỗi cặp gồm ảnh và câu mô tả ngắn thể hiện khả năng của mô hình trong việc tạo chú thích ngôn ngữ tự nhiên phù hợp với nội dung thị giác (trích từ [8])

Về hướng tiếp cận, các phương pháp chú thích ảnh ban đầu chủ yếu sử dụng mạng nơ-ron tích chập (Convolutional Neural Network - CNN) để trích xuất đặc trưng và mạng nơ-ron hồi quy (Recurrent Neural Network - RNN hoặc Long Short-Term Memory - LSTM) để sinh chú thích, tiêu biểu là kiến trúc CNN-LSTM [3, 18]. Tuy nhiên, sự ra đời của kiến trúc Transformer cùng cơ chế tự chú ý (self-attention) đã mở ra hướng tiếp cận mới, giúp cải thiện khả năng học ngữ cảnh và tạo ra các chú thích phong phú hơn [18, 19]. Đồng thời, nhiều nghiên cứu đã tích hợp thêm các nguồn tri thức bên ngoài tập dữ liệu như ConceptNet, WordNet [20, 21], hoặc biểu diễn sâu hơn các mối quan hệ thông qua đồ thị quan hệ (relationship graph) hoặc đồ thị cảnh (scene graph) [22, 23]; đặc biệt là biểu diễn ngữ nghĩa trừu tượng (Abstract Meaning Representation - AMR) để tăng cường khả năng hiểu ngữ nghĩa và mối liên hệ giữa các đối tượng trong ảnh [24, 25]. Mặc dù đã đạt được nhiều kết quả tích cực, các mô hình chú thích ảnh vẫn còn tồn tại những thách thức lớn, bao gồm: (i) chú thích sinh ra thường thiếu chiều sâu ngữ nghĩa và không phản ánh đầy đủ quan hệ

giữa các đối tượng [26], (ii) mô hình khó khăn khi gặp các đối tượng hiếm hoặc chưa từng xuất hiện trong dữ liệu huấn luyện [27, 28], (iii) khả năng biểu diễn tri thức vẫn còn hạn chế do thiếu cơ chế tích hợp hiệu quả giữa thông tin hình ảnh và các dạng tri thức bên ngoài [29]. Bên cạnh đó, việc sử dụng tri thức ngữ nghĩa hiện nay còn rời rạc, chủ yếu được tích hợp vào một giai đoạn xử lý cụ thể, mà chưa có sự phối hợp nhịp nhàng giữa biểu diễn thị giác, quan hệ ngữ nghĩa và biểu diễn ngôn ngữ trong toàn bộ quy trình xử lý (pipeline) của mô hình.

Bên cạnh các thách thức về biểu diễn ngữ nghĩa và tích hợp tri thức, một vấn đề quan trọng khác trong nghiên cứu chú thích ảnh là khả năng mở rộng sang các ngôn ngữ ít tài nguyên như tiếng Việt. Phần lớn các công trình này tập trung vào tiếng Anh do sự sẵn có của các tập dữ liệu chuẩn quy mô lớn và các công cụ xử lý ngôn ngữ hoàn chỉnh, trong khi các ngôn ngữ khác còn thiếu dữ liệu gán nhãn và các bộ phân tích ngữ nghĩa tương ứng. Thách thức này đặt ra yêu cầu phát triển các phương pháp chú thích ảnh có tính tổng quát cao, có khả năng thích ứng với nhiều ngôn ngữ và miền dữ liệu khác nhau, thay vì phụ thuộc chặt chẽ vào đặc thù của một ngôn ngữ cụ thể.

Từ các phân tích trên có thể nhận thấy rằng, mặc dù các phương pháp chú thích ảnh dựa trên học sâu đã đạt được nhiều tiến bộ đáng kể, vẫn còn tồn tại những hạn chế mang tính hệ thống liên quan đến biểu diễn ngữ nghĩa, khai thác quan hệ giữa các đối tượng và tích hợp tri thức ngoài một cách hiệu quả. Những hạn chế này đặt ra các câu hỏi nghiên cứu (Research Question - RQ) cốt lõi mà luận án cần giải quyết, bao gồm:

- RQ1: Có cách gì để mô hình chú thích ảnh có thể biểu diễn và khai thác hiệu quả các mối quan hệ ngữ nghĩa giữa các đối tượng trong ảnh thay vì chỉ dựa trên đặc trưng thị giác cục bộ.
- RQ2: Việc tích hợp tri thức ngữ nghĩa bên ngoài (ví dụ: ConceptNet) vào quá trình sinh chú thích một cách có hệ thống nhằm cải thiện khả năng mô tả các đối tượng hiếm hoặc chưa xuất hiện trong dữ liệu huấn luyện được làm như thế nào.
- RQ3: Biểu diễn ngữ nghĩa trừu tượng, điển hình như AMR, có thể được khai thác như thế nào để nâng cao khả năng hiểu nội dung ảnh ở mức khái niệm và tăng tính nhất quán ngữ nghĩa của chú thích sinh ra.
- RQ4: Việc hợp nhất đồng thời đặc trưng thị giác, quan hệ cấu trúc và các dạng tri thức ngữ nghĩa đa nguồn trong một kiến trúc thống nhất có giúp cải thiện toàn diện chất lượng chú thích ảnh hay không.

Xuất phát từ các thách thức và câu hỏi nghiên cứu nêu trên, luận án lựa chọn đề tài **"Phát triển phương pháp chú thích ảnh dựa trên mạng học sâu"** với mục tiêu đề xuất các mô hình chú thích ảnh tích hợp đa nguồn thông tin, kết hợp chặt chẽ giữa biểu diễn thị giác, quan hệ ngữ nghĩa, tri thức ngoài và biểu diễn ngữ nghĩa trừu tượng. Về mặt học thuật, nghiên cứu này góp phần làm sâu sắc thêm hiểu biết về học sâu đa phương thức và biểu diễn ngữ nghĩa trong bài toán chú thích ảnh. Về mặt thực tiễn, các kết quả đạt được hướng tới việc xây dựng các hệ thống chú thích ảnh có độ chính xác cao, giàu ngữ nghĩa và phù hợp hơn với cách diễn đạt tự nhiên của con người, đáp ứng yêu cầu của nhiều ứng dụng thông minh trong bối cảnh AI hiện đại.

2. Mục tiêu nghiên cứu

Dựa trên các thách thức đã xác định, luận án đặt ra mục tiêu tổng quát là: ***Phát triển các phương pháp chú thích ảnh dựa trên mạng học sâu***, kết hợp hiệu quả đặc trưng thị giác, quan hệ ngữ nghĩa, tri thức ngoài tập dữ liệu và biểu diễn ngữ nghĩa trừu tượng nhằm cải thiện độ chính xác và tính tự nhiên của mô tả.

Trên cơ sở các câu hỏi nghiên cứu RQ1, RQ2, RQ3, RQ4 đã nêu, luận án cụ thể hóa mục tiêu nghiên cứu thành các nội dung cụ thể tương ứng như sau:

(i) Phân tích và đánh giá các hướng tiếp cận hiện có trong bài toán chú thích ảnh, từ đó xác định những hạn chế liên quan đến việc biểu diễn quan hệ giữa các đối tượng và tri thức ngữ nghĩa trong quá trình sinh chú thích.

(ii) Đề xuất phương pháp chú thích ảnh có khả năng biểu diễn và khai thác mối quan hệ giữa các đối tượng trong ảnh nhằm cải thiện khả năng hiểu cấu trúc cảnh vật và ngữ cảnh hình ảnh của mô hình.

(iii) Nghiên cứu cơ chế tích hợp và hợp nhất tri thức đa nguồn, bao gồm tri thức ngữ nghĩa từ các cơ sở tri thức bên ngoài (như ConceptNet, mô hình ngôn ngữ lớn) cùng đặc trưng thị giác và ngôn ngữ nhằm tăng chiều sâu ngữ nghĩa và khả năng diễn đạt tự nhiên của mô hình.

(iv) Khai thác và ứng dụng biểu diễn ngữ nghĩa trừu tượng, điển hình như AMR, để nâng cao khả năng biểu diễn và hiểu nội dung ảnh ở mức khái niệm trừu tượng.

(v) Tiến hành thực nghiệm và đánh giá các phương pháp được đề xuất tương ứng với các nội dung (ii)-(iv) trên các tập dữ liệu chuẩn (MS COCO, Flickr30K) bằng hệ thống độ đo BLEU, METEOR, ROUGE, CIDEr, SPICE và SCS nhằm kiểm chứng mức độ đạt được của các mục tiêu nghiên cứu và đánh giá khả năng tổng quát hóa của mô hình.

Các nội dung nghiên cứu được triển khai theo lộ trình kế thừa và mở rộng, trong đó thực nghiệm và đánh giá đóng vai trò kiểm chứng bắt buộc, đảm bảo tính thống nhất giữa cơ sở lý thuyết, thiết kế mô hình và kết quả thực nghiệm, đồng thời hướng đến giá trị ứng dụng thực tiễn trong các hệ thống thị giác-ngôn ngữ.

3. Đối tượng và phạm vi nghiên cứu

Để triển khai hiệu quả các mục tiêu nghiên cứu đã đề ra, việc xác định rõ đối tượng và phạm vi nghiên cứu là điều cần thiết nhằm đảm bảo tính tập trung, nhất quán và khả thi trong toàn bộ quá trình thực hiện luận án. Phần này trình bày cụ thể đối tượng nghiên cứu chính cùng với phạm vi nghiên cứu của luận án.

3.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của luận án là quá trình sinh chú thích bằng ngôn ngữ tự nhiên cho hình ảnh đầu vào, tập trung vào các mô hình học sâu dựa trên kiến trúc encoder-decoder. Cụ thể, luận án tập trung vào các yếu tố sau:

- Biểu diễn đặc trưng hình ảnh, kết hợp đặc trưng ngữ nghĩa thị giác-ngôn ngữ từ CLIP (Contrastive Language-Image Pretraining) với thông tin cấu trúc về các đối tượng và mối quan hệ giữa chúng được mô hình hóa thông qua đồ thị quan hệ nhằm cung cấp biểu diễn toàn diện và giàu ngữ cảnh hơn cho quá trình giải mã.

- Khai thác và tích hợp tri thức ngữ nghĩa, bao gồm tri thức bên ngoài tập dữ liệu (như ConceptNet và LLM), tri thức ngữ nghĩa trừu tượng (từ AMR), và tri thức ngữ cảnh trong ảnh nhằm tăng cường khả năng hiểu nội dung, suy luận ngữ nghĩa và tạo chú thích phù hợp hơn với nhận thức ngôn ngữ của con người;

- Mô hình hóa quá trình phát sinh chú thích, sử dụng các kiến trúc giải mã hiện đại như LSTM và Transformer có tích hợp cơ chế chú ý (dual attention, cross-attention, masked attention) để sinh chú thích mạch lạc, đúng cú pháp, ngữ nghĩa.

Như vậy, luận án không chỉ xem xét ảnh như một nguồn thông tin thị giác đơn thuần mà còn như một thực thể có cấu trúc ngữ nghĩa, cần được hiểu và mô tả bằng cách kết hợp nhiều dạng tri thức.

3.2. Phạm vi nghiên cứu

Để đảm bảo tính tập trung và khả thi, phạm vi nghiên cứu của luận án được giới hạn như sau:

- **Dữ liệu đầu vào và ngôn ngữ**: Luận án sử dụng hai tập dữ liệu chuẩn MS COCO và Flickr30K với chú thích bằng tiếng Anh nhằm đảm bảo khả năng so sánh với các phương pháp hiện có. Việc chưa thực nghiệm cho tiếng Việt chủ yếu do hạn chế về dữ liệu gán nhãn quy mô lớn và công cụ phân tích ngữ nghĩa trừu tượng tương ứng, không phải do giới hạn của mô hình đề xuất.

▪ **Kiến trúc mô hình**: Luận án tập trung vào kiến trúc encoder-decoder, trong đó encoder trích xuất đặc trưng từ nhiều nguồn gồm đối tượng, quan hệ, biểu diễn AMR, và đặc trưng ngôn ngữ - thị giác từ CLIP-ViT. Decoder sử dụng LSTM hoặc Transformer kết hợp với các cơ chế chú ý để hợp nhất thông tin và sinh chú thích tự nhiên, chính xác. Ở giai đoạn suy luận, áp dụng GPT re-ranking nhằm chọn chú thích tối ưu và đảm bảo tính mạch lạc.

▪ **Nguồn tri thức và biểu diễn ngữ nghĩa**: Luận án khai thác tri thức ngoài tập dữ liệu từ ba nguồn chính: ConceptNet để biểu diễn quan hệ ngữ nghĩa giữa các khái niệm, AMR để mô hình hóa cấu trúc ngữ nghĩa trừu tượng, và mô hình ngôn ngữ lớn (GPT) nhằm đánh giá, xếp hạng lại chú thích. Việc kết hợp các nguồn tri thức này giúp tăng cường khả năng hiểu ngữ cảnh và tính tự nhiên của mô tả ảnh.

▪ **Đầu ra**: Đầu ra là một câu mô tả bằng ngôn ngữ tự nhiên (tiếng Anh) phản ánh chính xác nội dung của ảnh đầu vào.

▪ **Thực nghiệm**: Các thực nghiệm được triển khai trên hai tập dữ liệu chuẩn MS COCO và Flickr30K, sử dụng các độ đo đánh giá phổ biến gồm BLEU, METEOR, ROUGE, CIDEr, SPICE, cùng độ đo ngữ nghĩa mới - Semantic Consistency Score (SCS) nhằm phản ánh mức độ tương đồng ngữ nghĩa giữa chú thích sinh ra và chú thích chuẩn. Việc mở rộng sang các ngôn ngữ khác hoặc tập dữ liệu không gán nhãn không nằm trong phạm vi của luận án.

4. Phương pháp nghiên cứu

Nhằm hiện thực hóa các mục tiêu nghiên cứu đã đề ra, trên cơ sở đối tượng và phạm vi nghiên cứu đã xác định, luận án vận dụng tổng hợp nhiều phương pháp nghiên cứu theo định hướng ứng dụng trong lĩnh vực trí tuệ nhân tạo, đặc biệt là học sâu đa phương thức (*multimodal deep learning*). Các phương pháp được tổ chức thành ba nhóm chính, phản ánh tuần tự tiến trình triển khai từ phân tích lý thuyết đến thiết kế mô hình và đánh giá thực nghiệm. Cụ thể như sau:

(i) **Về việc tổng quan tài liệu**: Luận án tiến hành khảo sát, tổng hợp và phân tích có hệ thống các công trình nghiên cứu liên quan đến bài toán chú thích ảnh, tập trung vào bốn nhóm phương pháp chủ đạo: (i) các mô hình dựa trên kiến trúc CNN-LSTM truyền thống; (ii) các phương pháp khai thác đồ thị quan hệ và đồ thị cảnh; (iii) các mô hình hiện đại sử dụng kiến trúc Transformer; (iv) các hướng tiếp cận tích hợp tri thức ngữ nghĩa từ các nguồn tri thức ngoài như ConceptNet, WordNet, LLMs; và (v) các phương pháp sử dụng AMR.

Quá trình khảo sát này nhằm xác lập nền tảng lý thuyết cho luận án, đồng thời nhận diện các hạn chế còn tồn tại trong nghiên cứu hiện tại, từ đó xác định khoảng

trống học thuật cần tập trung giải quyết.

(ii) **Trong việc thiết kế mô hình và phát triển thuật toán**: Trên cơ sở kết quả khảo sát, luận án đề xuất và phát triển chuỗi các mô hình chú thích ảnh theo kiến trúc encoder-decoder, có tính kế thừa và mở rộng qua từng giai đoạn. Các mô hình khai thác kết hợp đặc trưng thị giác, cấu trúc quan hệ giữa các đối tượng, tri thức ngữ nghĩa bên ngoài tập dữ liệu và biểu diễn ngữ nghĩa trừu tượng nhằm nâng cao khả năng hiểu nội dung ảnh và chất lượng sinh mô tả. Toàn bộ mô hình được cài đặt bằng ngôn ngữ lập trình Python, framework PyTorch và các thư viện liên quan khác.

(iii) **Về việc thực nghiệm và đánh giá mô hình**: Các mô hình được huấn luyện và đánh giá trên hai tập dữ liệu chuẩn MS COCO và Flickr30K. Hiệu quả của mô hình được đo lường thông qua các độ đo phổ biến như BLEU, METEOR, ROUGE- L, CIDEr và SPICE; đồng thời luận án đề xuất bổ sung độ đo SCS nhằm đánh giá mức độ tương đồng ngữ nghĩa ở không gian embedding. Bên cạnh so sánh định lượng, luận án còn thực hiện phân tích định tính để đánh giá khả năng hiểu nội dung ảnh và tính tự nhiên của ngôn ngữ sinh ra.

5. Những đóng góp chính của luận án

Luận án phát triển một chuỗi bốn mô hình chú thích ảnh phát triển theo hướng kế thừa và mở rộng, tập trung tăng cường khả năng biểu diễn ngữ nghĩa và hợp nhất tri thức đa nguồn nhằm nâng cao chất lượng mô tả ảnh. Các đóng góp chính bao gồm:

(i) Xây dựng mô hình OD-VR-Cap kết hợp đặc trưng đối tượng với quan hệ giữa các đối tượng thông qua đồ thị quan hệ và cơ chế chú ý kép, từ đó cải thiện khả năng nắm bắt ngữ nghĩa và cấu trúc cảnh khi sinh chú thích.

(ii) Phát triển mô hình RGTranCNet dựa trên Transformer decoder và tri thức ngữ nghĩa từ ConceptNet, tăng cường khả năng khai thác thông tin quan hệ và hỗ trợ xử lý hiệu quả các đối tượng ít xuất hiện trong dữ liệu huấn luyện.

(iii) Đưa ra mô hình AMR-GT&RG kết hợp hai dạng biểu diễn ngữ nghĩa trừu tượng - AMR trích xuất từ chú thích chuẩn và AMR-like chuyển đổi từ đồ thị quan hệ ảnh - nhằm nâng cao khả năng biểu diễn ngữ nghĩa sâu. Đồng thời, giới thiệu độ đo SCS để đánh giá mức độ tương đồng ngữ nghĩa giữa chú thích sinh ra và chú thích chuẩn.

(iv) Hoàn thiện khung hợp nhất tri thức đa nguồn thông qua mô hình CLIP-AMR-GPT, kết hợp đặc trưng thị giác - ngôn ngữ, biểu diễn đồ thị quan hệ và biểu diễn AMR/AMR-like bằng các cơ chế Cross-Fusion và Adaptive Attention. Giai đoạn suy luận áp dụng GPT re-ranking giúp tăng cường tính tự nhiên và độ chính xác ngữ nghĩa của chú thích.

6. Bố cục của luận án

Bên cạnh phần Mở đầu, Kết luận và Tài liệu tham khảo; luận án được tổ chức thành năm chương chính. Nội dung được sắp xếp theo tiến trình logic từ tổng quan lý thuyết đến đề xuất phương pháp và đánh giá thực nghiệm, cụ thể như sau:

- **Mở đầu**: Trình bày tính cấp thiết, mục tiêu, đối tượng, phạm vi và phương pháp nghiên cứu; đồng thời khái quát các đóng góp khoa học và bố cục tổng thể của luận án.

- **Chương 1**: Tổng quan các hướng nghiên cứu về chú thích ảnh dựa trên học sâu, bao gồm các phương pháp dựa trên đặc trưng thị giác, đồ thị quan hệ, tri thức ngữ nghĩa và biểu diễn AMR, qua đó xác định khoảng trống học thuật và định hướng nghiên cứu của luận án.

- **Chương 2 đến Chương 5**: Trình bày các mô hình chú thích ảnh được đề xuất, phát triển theo các hướng tiếp cận: từ khai thác đặc trưng thị giác và quan hệ đối tượng (OD-VR-Cap), tích hợp tri thức ngoài từ ConceptNet (RGTranCNet), kết hợp biểu diễn ngữ nghĩa trừu tượng AMR (AMR-GT&RG), đến hợp nhất đặc trưng đa nguồn và áp dụng GPT re-ranking (CLIP-AMR-GPT). Mỗi chương mô tả kiến trúc, quy trình huấn luyện, kết quả thực nghiệm và phân tích đánh giá tương ứng.

- **Kết luận và Kiến nghị**: Tổng kết kết quả nghiên cứu, khẳng định những đóng góp khoa học chủ yếu và đề xuất các hướng phát triển tiếp theo của luận án.

CHƯƠNG 1. CHÚ THÍCH ẢNH DỰA TRÊN MẠNG HỌC SÂU

Chương này cung cấp một cách tổng quan về nền tảng lý thuyết cho toàn bộ luận án thông qua việc khảo sát, phân tích và hệ thống hóa các hướng tiếp cận trong bài toán chú thích ảnh, với trọng tâm là các phương pháp dựa trên mạng học sâu. Mục 1.1 trình bày tổng quan về bài toán, đặc điểm và các thách thức trong việc sinh mô tả ngôn ngữ từ ảnh đầu vào. Mục 1.2 phân loại và phân tích các nhóm phương pháp nghiên cứu chính, từ các tiếp cận truyền thống đến các mô hình học sâu hiện đại. Trên cơ sở đó, Mục 1.3 xác định các khoảng trống nghiên cứu còn tồn tại và đề xuất định hướng phát triển nhằm nâng cao khả năng hiểu ngữ nghĩa và sinh ngôn ngữ của mô hình. Cuối cùng, Mục 1.4 giới thiệu khung thực nghiệm được sử dụng trong toàn bộ luận án, bao gồm tập dữ liệu, độ đo đánh giá và môi trường triển khai, làm cơ sở để đánh giá hiệu quả của các mô hình được trình bày trong các chương tiếp theo.

1.1. Giới thiệu

Chú thích ảnh tự động là một bài toán liên ngành tiêu biểu trong lĩnh vực trí tuệ nhân tạo, kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên [30]. Mục tiêu của bài toán là sinh ra một câu mô tả ngôn ngữ tự nhiên phản ánh nội dung ảnh đầu vào một cách chính xác, mạch lạc và phù hợp ngữ cảnh [31]. Khác với các tác vụ trong thị giác máy tính truyền thống như phân loại ảnh hoặc phát hiện đối tượng, bài toán chú thích ảnh đòi hỏi mô hình không chỉ nhận diện các thành phần trong ảnh mà còn hiểu được các mối quan hệ ngữ nghĩa giữa chúng để biểu đạt thành câu văn hoàn chỉnh, đáp ứng yêu cầu về cú pháp và ngữ nghĩa [32].

Với sự phát triển mạnh mẽ của học sâu đa phương thức, đặc biệt là các kiến trúc như CNN, RNN, và Transformer, hiệu quả của các hệ thống chú thích ảnh đã có những bước tiến đáng kể [31, 33]. Trong các kiến trúc này, CNN thường được sử dụng để trích xuất đặc trưng hình ảnh, trong khi RNN hoặc Transformer đảm nhiệm vai trò giải mã ngôn ngữ [34, 35]. Tuy nhiên, nhiều nghiên cứu chỉ ra rằng các mô hình này vẫn gặp khó khăn trong việc mô hình hóa sâu ngữ nghĩa của ảnh, đặc biệt là trong việc hiểu quan hệ giữa các đối tượng hoặc các khái niệm ngữ nghĩa không trực tiếp biểu hiện bằng đặc trưng thị giác [36, 37].

Để khắc phục hạn chế này, có hướng tiếp cận đã đề xuất như tích hợp thêm các nguồn tri thức ngoài như ConceptNet, WordNet hoặc sử dụng biểu diễn ngữ nghĩa trừu tượng như AMR nhằm mở rộng khả năng suy diễn và sinh ngôn ngữ của mô hình [20, 25, 38]. Dù vậy, cách thức tích hợp hiệu quả các nguồn tri thức này vào pipeline học sâu vẫn là một thách thức mở, đòi hỏi các giải pháp sáng tạo trong biểu diễn, huấn luyện và suy luận [31, 39].

Trên cơ sở đó, chương này được xây dựng nhằm cung cấp cái nhìn toàn diện về các hướng tiếp cận trong bài toán chú thích ảnh. Cụ thể, Mục 1.2 dùng để trình bày chi tiết các phương pháp này, qua đó làm rõ tiến trình phát triển và nền tảng lý thuyết cho các mô hình đề xuất trong các chương sau.

1.2. Hướng tiếp cận trong chú thích ảnh

Trên cơ sở tiến bộ của thị giác máy tính và xử lý ngôn ngữ tự nhiên, các phương pháp chú thích ảnh có thể được phân thành hai nhóm chính: (i) các phương pháp truyền thống và (ii) các phương pháp hiện đại dựa trên học sâu. Theo cách phân loại này, các mục tiếp theo lần lượt trình bày tổng quan về các phương pháp truyền thống (Mục 1.2.1), phân tích các phương pháp học sâu hiện đại - nền tảng cho các mô hình đề xuất trong luận án (Mục 1.2.2), và đưa ra đánh giá tổng hợp về ưu nhược điểm cũng như xu hướng phát triển của các hướng tiếp cận này (Mục 1.2.3).

1.2.1. Phương pháp chú thích ảnh truyền thống

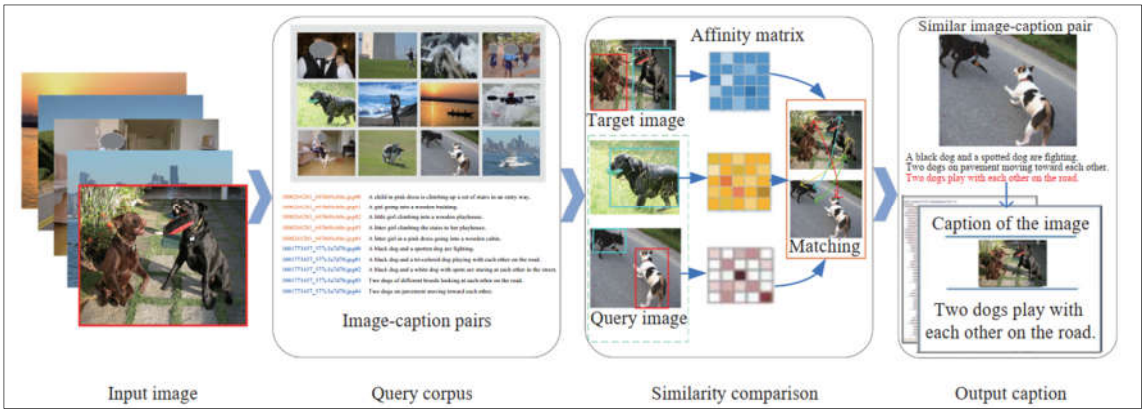
Trước khi các phương pháp học sâu trở nên phổ biến, bài toán chú thích ảnh đã được tiếp cận thông qua hai hướng truyền thống: (i) phương pháp dựa trên truy hồi thông tin (retrieval-based) và (ii) phương pháp dựa trên mẫu dữ liệu (template-based) [40]. Mặc dù không còn là xu hướng chủ đạo, hai hướng tiếp cận này vẫn đóng vai trò nền tảng, cung cấp những ý tưởng ban đầu cho việc phát sinh mô tả từ nội dung hình ảnh.

1.2.1.1. Phương pháp chú thích ảnh dựa trên truy hồi thông tin

Phương pháp chú thích ảnh dựa trên truy hồi thông tin hoạt động trên nguyên tắc tìm kiếm các ảnh tương tự từ một tập dữ liệu đã có chú thích, sau đó chọn ra chú thích phù hợp từ các ảnh tương tự đó làm chú thích đầu ra cho ảnh đầu vào. Quá trình này bao gồm ba bước chính: (i) trích xuất đặc trưng hình ảnh (thường bằng các mô tả SIFT, HOG hoặc BoVW); (ii) tính toán độ tương đồng giữa ảnh đầu vào và các ảnh trong tập dữ liệu; (iii) chọn chú thích của ảnh gần nhất (hoặc tổng hợp từ nhiều ảnh gần nhất) làm kết quả đầu ra [4, 40]. Hình 1.1 mô tả kiến trúc tổng quan của phương pháp chú thích ảnh dựa trên truy hồi thông tin, với đầu vào là một ảnh, đặc trưng ảnh được trích xuất và đối sánh với tập ảnh có sẵn, chú thích được chọn từ ảnh gần nhất hoặc tổ hợp từ nhiều ảnh tương tự để tạo ra mô tả đầu ra.

Ưu điểm chính của phương pháp này là đơn giản, nhanh và cho ra câu chú thích đúng cú pháp vì các câu được trích từ tập chú thích sẵn có. Tuy nhiên, nó phụ thuộc nhiều vào chất lượng và độ phủ của tập dữ liệu, dẫn đến ba hạn chế chính: (i) tính tổng quát hóa kém - không thể tạo mô tả cho những hình ảnh với nội dung chưa từng xuất hiện trong tập huấn luyện; (ii) khả năng biểu đạt ngữ nghĩa hạn chế - mô tả

thường bị trùng lặp, thiếu sáng tạo và không phù hợp với ngữ cảnh cụ thể; (iii) phụ thuộc vào chất lượng của tập dữ liệu và phương pháp tìm kiếm ảnh tương tự [40].



Hình 1.1. Khung của mô hình chú thích ảnh dựa vào truy hồi thông tin. Mô hình nhận đầu vào là một ảnh cần chú thích và truy vấn đến tập ảnh-chú thích đã được gán nhãn để tìm ảnh tương tự thông qua ma trận độ tương đồng. Sau đó, mô hình chọn cặp ảnh-chú thích gần nhất và sử dụng câu mô tả của ảnh đó làm chú thích cho ảnh đầu vào (trích từ [40]).

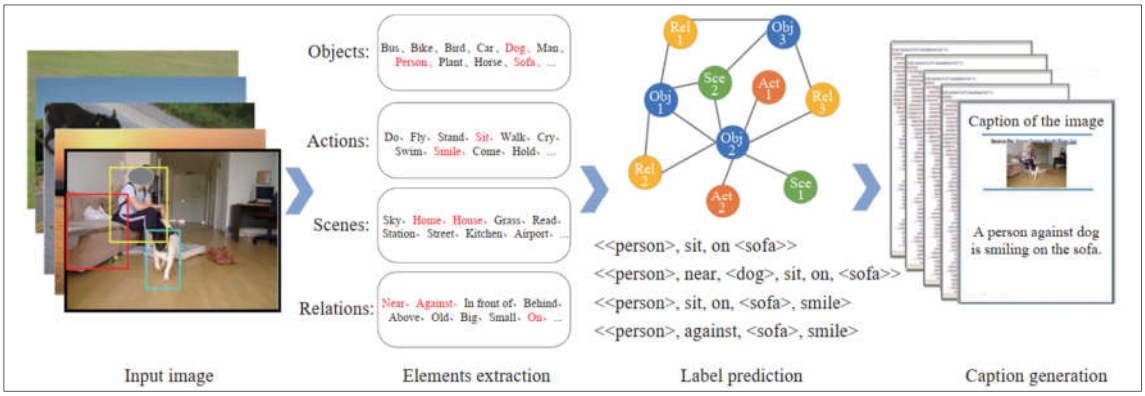
1.2.1.2. Phương pháp chú thích ảnh dựa trên mẫu dữ liệu

Khác với truy hồi thông tin, phương pháp dựa trên mẫu dữ liệu xây dựng chú thích mới cho mỗi ảnh dựa trên các mẫu câu định nghĩa trước (pre-defined templates). Về bản chất, các mẫu này là khung cú pháp có chỗ trống dành cho các thành phần nội dung như danh từ, động từ, trạng từ... được trích xuất từ ảnh thông qua các bước như: phát hiện đối tượng, xác định hành động và quan hệ giữa các đối tượng [40]. Ví dụ, một mẫu có thể là "There is a [*object*] on the [*location*]", trong đó hệ thống phát hiện ra các đối tượng và điền vào chỗ trống để tạo thành câu hoàn chỉnh như "There is a dog on the couch". Hình 1.2 minh họa kiến trúc cơ bản của phương pháp này: ảnh được đưa vào hệ thống nhận diện thị giác để trích xuất thông tin, sau đó các thông tin này được ánh xạ sang nhãn ngôn ngữ và điền vào mẫu câu có cấu trúc định sẵn.

Ưu điểm của phương pháp này là tạo ra câu chú thích đúng ngữ pháp, dễ kiểm soát và nhất quán về mặt ngôn ngữ. Tuy nhiên, hạn chế lại nằm ở chính cấu trúc cứng nhắc của mẫu câu: (i) khả năng biểu đạt bị giới hạn, không thể tạo ra các chú thích linh hoạt và đa dạng; (ii) khó mở rộng sang các ngữ cảnh phong phú; (iii) yêu cầu phát hiện đầy đủ và chính xác các thực thể và hành động - điều vốn khó đảm bảo với các kỹ thuật xử lý hình ảnh truyền thống [6, 40].

Mặc dù hai phương pháp truyền thống kể trên cung cấp những cách tiếp cận ban đầu có tính khả thi, song chúng chưa thể giải quyết được yêu cầu về tính linh hoạt, diễn đạt ngữ nghĩa phong phú và khả năng tổng quát hóa trên ảnh mới. Những hạn chế này chính là động lực thúc đẩy sự phát triển của các phương pháp dựa trên

mạng học sâu - vốn có khả năng học từ dữ liệu lớn, sinh mô tả mới và kết hợp được giữa biểu diễn hình ảnh và ngôn ngữ một cách mềm dẻo và mạnh mẽ hơn.



Hình 1.2. Khung của mô hình chú thích ảnh dựa vào mẫu dữ liệu. Mô hình thực hiện trích xuất các yếu tố ngữ nghĩa bao gồm đối tượng, hành động, cảnh nền và quan hệ từ ảnh đầu vào. Các yếu tố này được sử dụng để dự đoán cấu trúc nhãn dạng bộ ba, từ đó sinh ra câu mô tả ảnh tương ứng thông qua ánh xạ các mẫu đã học (trích từ [40]).

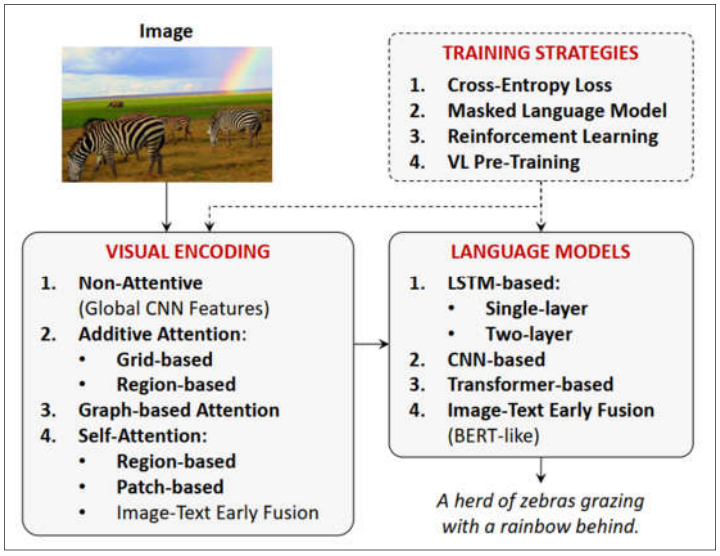
1.2.2. Phương pháp chú thích ảnh dựa trên học sâu

Các phương pháp chú thích ảnh dựa trên học sâu đã trở thành xu hướng chủ đạo trong những năm gần đây nhờ khả năng biểu diễn đặc trưng mạnh mẽ, học được các mẫu phức tạp giữa hình ảnh và ngôn ngữ. Khác với các phương pháp truyền thống dựa vào truy hồi hoặc mẫu dữ liệu có sẵn, các phương pháp chú thích ảnh hiện đại thường áp dụng mô hình học sâu để học ánh xạ trực tiếp giữa đặc trưng hình ảnh và mô tả ngôn ngữ. Khung làm việc phổ biến của các phương pháp này là kiến trúc mã hóa - giải mã (encoder - decoder). Trong đó, một mạng CNN hoặc mạng phát hiện đối tượng (Faster R-CNN, YOLO…) thường được sử dụng làm bộ mã hóa hình ảnh, giúp trích xuất đặc trưng thị giác; trong khi bộ giải mã ngôn ngữ thường là RNN, LSTM, GRU, hoặc Transformer decoder có nhiệm vụ sinh chú thích đầu ra dựa trên đặc trưng ảnh và ngữ cảnh trước đó.

Một thành phần quan trọng trong hầu hết các mô hình sinh chú thích là cơ chế chú ý (attention mechanism), giúp mô hình tập trung vào các vùng ảnh liên quan khi tạo ra mỗi từ trong câu mô tả. Ngoài ra, các mô hình hiện đại còn được tăng cường bằng tri thức ngữ nghĩa từ bên ngoài tập dữ liệu như ConceptNet, WordNet, LLMs hoặc sử dụng biểu diễn đồ thị ngữ nghĩa như scene graph hoặc AMR để cải thiện khả năng hiểu nội dung ảnh ở cấp độ ngữ nghĩa sâu hơn. Hình 1.3 minh họa cấu trúc tổng quát của mô hình chú thích ảnh encoder - decoder. Hệ thống sử dụng các phương pháp mã hóa thị giác như đặc trưng CNN toàn cục, attention theo lưới, theo vùng, attention dựa trên đồ thị. Các mô hình ngôn ngữ có thể dựa trên LSTM, CNN, Transformer hoặc mô hình hợp nhất sớm giữa văn bản và hình ảnh. Quá trình huấn

luyện có thể sử dụng các chiến lược như hàm mất mát cross-entropy, mô hình ngôn ngữ mặt nạ, học tăng cường và mô hình ngôn ngữ-thị giác huấn luyện trước.

Dựa trên kiến trúc tổng thể này, các phương pháp học sâu cho bài toán chú thích ảnh được phân thành các nhóm chính sau: (1.2.2.1) các phương pháp dựa trên CNN-LSTM, (1.2.2.2) các phương pháp sử dụng biểu diễn đồ thị, (1.2.2.3) các phương pháp dựa vào mạng Transformer, (1.2.2.4) phương pháp sử dụng tri thức bên ngoài tập dữ liệu, và (1.2.2.5) các phương pháp sử dụng AMR.



Hình 1.3. Tổng quan các thành phần chính trong hệ thống chú thích ảnh tự động bao gồm mã hóa hình ảnh, mô hình ngôn ngữ và chiến lược huấn luyện (trích từ [5]).

1.2.2.1. Phương pháp dựa trên CNN-LSTM

Những nghiên cứu tiên phong trong chú thích ảnh dựa trên học sâu thường áp dụng kiến trúc mã hóa - giải mã (encoder - decoder), trong đó CNN đảm nhiệm vai trò trích xuất đặc trưng hình ảnh, còn RNN/LSTM được dùng để sinh mô tả ngôn ngữ. Mô hình *Show and Tell* [3] là ví dụ tiêu biểu, đặt nền móng cho hướng tiếp cận này. Sau đó, *Show, Attend and Tell* [18] mở rộng với cơ chế chú ý, giúp mô hình tập trung vào các vùng ảnh quan trọng khi sinh từng từ trong chú thích.

Nhiều công trình đã cải tiến kiến trúc CNN-LSTM bằng cách sử dụng các mạng CNN huấn luyện trước như ResNet, DenseNet hoặc Inception để tạo véc-tơ đặc trưng hình ảnh, sau đó đưa vào LSTM tích hợp cơ chế chú ý để phát sinh mô tả [41-43]. Một số tác giả còn kết hợp mạng phát hiện đối tượng nhằm bổ sung thông tin từ các vùng cục bộ thay vì chỉ dựa trên đặc trưng toàn cục [44]. Ngoài ra, đã có các mô hình mở rộng cho những ngôn ngữ ngoài tiếng Anh, điển hình là hệ thống chú thích ảnh tiếng Hindi kết hợp CNN, cơ chế chú ý và LSTM [45]. Một số hướng nghiên cứu

khác tận dụng Bi-LSTM, học tăng cường sâu hay thuật toán tìm kiếm heuristic để nâng cao chất lượng sinh chú thích [46-48].

Các công trình kể trên đều được kiểm chứng trên những bộ dữ liệu chuẩn như MS COCO và Flickr30K, với độ đo BLEU, METEOR, CIDEr và ROUGE, cho thấy phương pháp CNN-LSTM đạt hiệu quả tốt trong giai đoạn đầu. Tuy nhiên, hạn chế chung của hướng tiếp cận này là việc chỉ khai thác đặc trưng toàn cục hoặc đặc trưng vùng đối tượng từ CNN, trong khi chưa xem xét đến quan hệ ngữ nghĩa giữa các đối tượng trong ảnh. Điều này dẫn đến việc mô hình còn thiếu khả năng biểu diễn cấu trúc ngữ nghĩa phức tạp, làm giảm tính chính xác và tính giàu ngữ nghĩa của chú thích sinh ra.

1.2.2.2. Phương pháp sử dụng biểu diễn đồ thị

Khác với các phương pháp CNN-LSTM truyền thống vốn chủ yếu khai thác đặc trưng toàn cục hoặc các vùng đối tượng riêng lẻ, các phương pháp chú thích ảnh sử dụng biểu diễn đồ thị tập trung mô hình hóa cấu trúc quan hệ giữa các đối tượng trong ảnh nhằm làm giàu thông tin ngữ nghĩa cho quá trình sinh mô tả. Trong nhóm phương pháp này, ảnh không còn được xem như tập hợp các vùng độc lập mà được biểu diễn dưới dạng đồ thị quan hệ (relationship graph) hoặc đồ thị cảnh (scene graph), trong đó các đỉnh tương ứng với đối tượng và các cạnh biểu diễn mối quan hệ giữa chúng. Ý tưởng cốt lõi là khai thác sự tương tác giữa các đối tượng để tổ chức lại nội dung ảnh theo cấu trúc logic, từ đó nâng cao khả năng hiểu ngữ cảnh và chất lượng chú thích sinh ra.

Một trong những hướng tiếp cận sớm là xây dựng đồ thị cảnh cho cả ảnh và câu mô tả. Yang và cộng sự [49] đề xuất mô hình học chung một từ điển ngôn ngữ cho đồ thị ảnh và đồ thị câu, qua đó thiết lập mối liên hệ chặt chẽ giữa biểu diễn thị giác và ngôn ngữ trong quá trình mã hóa-giải mã. Cách tiếp cận này giúp mô hình tạo ra các mô tả gắn liền với ngữ cảnh, song vẫn phụ thuộc mạnh vào dữ liệu ngôn ngữ. Yao và cộng sự [22] mở rộng hướng tiếp cận này bằng cách kết hợp R-CNN để phát hiện đối tượng và sử dụng GCN nhằm mô hình hóa các quan hệ không gian và ngữ nghĩa giữa các đối tượng. Nhờ đó, mô hình không chỉ nhận diện sự hiện diện của từng đối tượng mà còn nắm bắt được cách chúng liên kết trong không gian ảnh.

Một số nghiên cứu khác nhấn mạnh việc khai thác đồng thời các quan hệ tường minh và tiềm ẩn giữa các đối tượng. Song và cộng sự [50] cho thấy việc kết hợp nhiều loại quan hệ giúp cải thiện đáng kể chất lượng chú thích. Parseh và cộng sự [23] đề xuất phương pháp biểu diễn kép, trong đó đồ thị ngữ nghĩa và đồ thị không gian được

trích xuất song song, sau đó đưa vào LSTM với cơ chế chú ý đa phương thức để sinh mô tả giàu ý nghĩa hơn.

Bên cạnh các mô hình xây dựng đồ thị trực tiếp từ ảnh, một số công trình tập trung khai thác đồ thị cảnh trích xuất từ chú thích chuẩn. Gao và cộng sự [51] đề xuất khung hai giai đoạn, trong đó đồ thị cảnh được xây dựng thông qua kết hợp CNN-RNN-SVM và sau đó sử dụng RNN để sinh chú thích. Mặc dù đạt kết quả khả quan trên MS COCO, phương pháp này phụ thuộc nhiều vào dữ liệu ngôn ngữ và đồ thị không phản ánh trực tiếp nội dung thị giác của ảnh. Chen và cộng sự [52] đề xuất đồ thị cảnh trừu tượng (abstract scene graph) trích xuất từ chú thích chuẩn nhằm kiểm soát mức độ chi tiết và phong cách của chú thích đầu ra. Yan và cộng sự [53] cải tiến mô hình bằng cách kết hợp Transformer với kiến trúc hai tầng LSTM, trong đó Transformer học mức đóng góp tương đối của các nguồn đặc trưng trong quá trình dự đoán từ. Các công trình này cho thấy biểu diễn đồ thị trừu tượng giúp cải thiện độ mạch lạc và tính đa dạng của chú thích, tuy nhiên vẫn chưa trực tiếp khai thác đầy đủ quan hệ ngữ nghĩa xuất phát từ chính ảnh đầu vào.

Ngoài ra, Mozes và cộng sự [54] tiếp tục khai thác hướng này bằng cách tạo scene graph trực tiếp từ ảnh sử dụng MOTIFNET [55], sau đó ánh xạ sang bộ giải mã LSTM để sinh chú thích. Phương pháp này còn tích hợp GAT nhằm mã hóa tốt hơn các quan hệ giữa đối tượng. Thử nghiệm trên tập dữ liệu giao giữa Visual Genome và MS COCO cho thấy mô hình G-LSTM của họ vượt trội hơn rõ rệt so với các baseline CNN-LSTM, đặc biệt khi sử dụng scene graph từ chú thích chuẩn, qua đó khẳng định tầm quan trọng của việc mô hình hóa quan hệ ngữ nghĩa trong việc nâng cao chất lượng mô tả ảnh.

Từ các nghiên cứu trên có thể nhận thấy rằng, ưu điểm nổi bật của nhóm phương pháp sử dụng biểu diễn đồ thị là khả năng mô hình hóa trực tiếp sự tương tác giữa các đối tượng, vượt lên trên các véc-tơ đặc trưng thị giác tổng quát. Tuy nhiên, phần lớn các phương pháp hiện tại mới dừng ở việc khai thác quan hệ bề mặt (như quan hệ không gian hoặc sở hữu), chưa biểu diễn đầy đủ các khái niệm ngữ nghĩa trừu tượng như vai trò, mục đích hay sự kiện. Bên cạnh đó, việc liên kết và đồng bộ hóa hiệu quả giữa đặc trưng thị giác và đặc trưng đồ thị trong quá trình sinh chú thích vẫn còn là một thách thức mở. Những hạn chế này đặt ra nhu cầu phát triển các mô hình có khả năng kết hợp chặt chẽ giữa phát hiện đối tượng chính xác, mô hình hóa quan hệ ngữ nghĩa bằng đồ thị và cơ chế giải mã học sâu tích hợp chú ý, nhằm nâng cao hơn nữa độ chính xác và tính giàu ngữ nghĩa của chú thích ảnh.

1.2.2.3. Phương pháp dựa vào mạng Transformer

Sau thành công vượt bậc của kiến trúc Transformer trong xử lý ngôn ngữ tự nhiên [56], nhiều nghiên cứu gần đây đã mở rộng mô hình này cho bài toán chú thích ảnh trong bối cảnh thị giác-ngôn ngữ. So với các kiến trúc hồi quy như RNN hoặc LSTM, Transformer có ưu thế rõ rệt nhờ khả năng mô hình hóa hiệu quả các phụ thuộc dài hạn, hỗ trợ huấn luyện song song và đạt hiệu suất cao trên các tập dữ liệu quy mô lớn [53]. Ngoài ra, sự xuất hiện của các biến thể Transformer cho thị giác như Vision Transformer (ViT) [57], Swin Transformer [58] hay ConvNeXt [59] đã chứng minh khả năng mã hóa ảnh vượt trội thông qua việc mô hình hóa mối quan hệ toàn cục giữa các vùng trong ảnh, qua đó khắc phục hạn chế của kiến trúc CNN-LSTM vốn phụ thuộc nhiều vào đặc trưng cục bộ.

Trong hướng tiếp cận này, Wang và cộng sự [60] là một trong những công trình tiêu biểu khi đề xuất mô hình chú thích ảnh sử dụng Transformer nhằm thay thế hoàn toàn vai trò của LSTM trong bộ giải mã, qua đó cải thiện tính mạch lạc và ổn định của chú thích sinh ra. Tiếp nối nghiên cứu này, Yang và cộng sự [61] phát triển Transformer có khả năng nhận biết ngữ cảnh, giúp tăng cường liên kết giữa đặc trưng hình ảnh và mô tả ngôn ngữ. Cornia và cộng sự [19] đề xuất mô hình Meshed Transformer, trong đó sử dụng Memory-Augmented Encoder kết hợp với cơ chế Meshed Connectivity ở bộ giải mã, cho phép tối ưu hóa biểu diễn ngữ nghĩa-không gian giữa các đối tượng trong ảnh. Ngoài ra, Pan và cộng sự [58] giới thiệu X-Linear Attention Networks, áp dụng các khối chú ý phi tuyến nhằm khai thác tương tác hình ảnh-ngôn ngữ đồng thời theo cả chiều không gian và chiều kênh, từ đó nâng cao hiệu quả sinh mô tả.

Các mô hình chú thích ảnh dựa trên Transformer thường được đánh giá trên các tập dữ liệu chuẩn như MS COCO và đạt kết quả ấn tượng theo nhiều độ đo phổ biến như BLEU, METEOR, ROUGE, CIDEr và SPICE. Những kết quả này cho thấy Transformer đóng vai trò quan trọng trong việc nâng cao khả năng biểu diễn ngữ cảnh và tăng cường liên kết giữa thị giác và ngôn ngữ. Tuy nhiên, các mô hình Transformer thuần túy vẫn gặp khó khăn khi xử lý đối tượng hiếm, tình huống ngoài tập dữ liệu huấn luyện, cũng như thiếu cơ chế tích hợp tri thức ngữ nghĩa bên ngoài để bổ trợ cho quá trình sinh chú thích. Những hạn chế này đã thúc đẩy các hướng nghiên cứu kết hợp Transformer với tri thức ngoài nhằm nâng cao khả năng diễn giải và tính chính xác ngữ nghĩa của mô hình.

1.2.2.4. Phương pháp sử dụng tri thức bên ngoài tập dữ liệu

Nhằm khắc phục hạn chế về dữ liệu huấn luyện và tăng chiều sâu ngữ nghĩa

cho mô hình chú thích ảnh, nhiều nghiên cứu đã đề xuất tích hợp tri thức bên ngoài tập dữ liệu vào kiến trúc học sâu. Các nguồn tri thức này có thể được phân thành ba nhóm chính: (i) các cơ sở tri thức ngữ nghĩa như ConceptNet và WordNet; (ii) các mô hình ngôn ngữ lớn (Large Language Models - LLMs) như GPT hoặc LLaMA; và (iii) các mô hình ngôn ngữ-thị giác huấn luyện trước (Vision-Language Models - VLMs) cho phép học biểu diễn liên miền sâu giữa hình ảnh và văn bản. Ý tưởng cốt lõi của hướng tiếp cận này là cung cấp thêm thông tin về quan hệ ngữ nghĩa giữa các đối tượng, từ đó hỗ trợ mô hình diễn giải chính xác hơn nội dung ảnh, đặc biệt trong các trường hợp đối tượng hiếm hoặc ngữ cảnh phức tạp.

Một trong những công trình tiêu biểu là CNet-NIC do Zhou và cộng sự [62] đề xuất, trong đó tri thức từ ConceptNet được tích hợp vào bộ mã hóa của mô hình NIC [3], giúp mở rộng vốn từ và cải thiện sự đa dạng của chú thích sinh ra. Tuy nhiên, việc đưa tri thức vào một cách không kiểm soát có thể gây nhiễu thông tin và làm giảm hiệu quả huấn luyện. Để giải quyết vấn đề này, Hafeth và cộng sự [21] đề xuất mạng chú ý ngữ nghĩa cho phép tích hợp có chọn lọc tri thức từ ConceptNet vào các tầng chú ý của Transformer, giúp tăng cường chất lượng mô tả trong khi vẫn kiểm soát được mức độ ảnh hưởng của tri thức ngoài. Bên cạnh đó, Li và cộng sự [63] kết hợp Transformer với tri thức ngữ nghĩa nhằm mô hình hóa hiệu quả hơn các quan hệ giữa đối tượng trong ảnh, qua đó cải thiện khả năng biểu đạt nội dung hình ảnh.

Gần đây, sự phát triển của các mô hình thị giác-ngôn ngữ huấn luyện trước quy mô lớn đã mở ra một hướng tiếp cận mới cho bài toán chú thích ảnh. Các mô hình tiêu biểu như CLIP [64], BLIP [65], BLIP-2 [66], hay Flamingo [67] cho thấy khả năng học biểu diễn liên miền sâu giữa hình ảnh và văn bản thông qua huấn luyện trên các tập dữ liệu lớn, đa dạng, cho phép khai thác embedding đa phương thức mạnh mẽ nhằm cải thiện độ chính xác và khả năng khái quát mà không cần huấn luyện lại toàn bộ hệ thống. Nhiều nghiên cứu đã tận dụng hiệu quả các mô hình này trong chú thích ảnh, chẳng hạn như CLIP-Captioner [68], BCEFT [69], hay các kiến trúc kết hợp Vision Transformer với GPT trong mô hình ViT-GPT-2 [70], cũng như các phương pháp sử dụng CLIP kết hợp GPT cho bài toán chú thích ảnh viễn thám [71], cho thấy khả năng sinh mô tả hiệu quả trong bối cảnh zero-shot và giảm sự phụ thuộc vào tập dữ liệu gán nhãn lớn.

Nhìn chung, các phương pháp tích hợp tri thức ngữ nghĩa và mô hình đa phương thức huấn luyện trước đã chứng minh tiềm năng rõ rệt trong việc nâng cao độ chính xác ngữ nghĩa và khả năng tổng quát hóa của mô hình chú thích ảnh. Tuy nhiên, nhiều thách thức vẫn còn tồn tại, bao gồm: (i) phần lớn các phương pháp mới dừng ở mức liên kết từ vựng hoặc quan hệ bề mặt, chưa khai thác sâu các cấu trúc

ngữ nghĩa trừu tượng; (ii) việc tích hợp tri thức và embedding đa phương thức thiếu cơ chế điều phối theo ngữ cảnh ảnh có thể gây nhiễu thông tin; và (iii) chi phí tính toán cao khi sử dụng các mô hình ngôn ngữ lớn trong giai đoạn suy luận. Những vấn đề này cho thấy nhu cầu phát triển các kiến trúc hợp nhất tri thức có kiểm soát, kết hợp hiệu quả giữa đặc trưng thị giác-ngôn ngữ huấn luyện trước, tri thức cấu trúc và biểu diễn ngữ nghĩa trừu tượng nhằm nâng cao chất lượng chú thích ảnh trong các bối cảnh dữ liệu đa dạng.

1.2.2.5. Phương pháp sử dụng AMR

Abstract Meaning Representation (AMR) [72] là một dạng biểu diễn ngữ nghĩa trừu tượng cho câu ngôn ngữ tự nhiên, trong đó ngữ nghĩa được mã hóa dưới dạng đồ thị với cấu trúc "who does what to whom" (ai làm gì với ai). Khác với các biểu diễn cú pháp bề mặt, AMR tập trung biểu diễn ngữ nghĩa cốt lõi của câu, cho phép các câu có cùng nội dung ngữ nghĩa nhưng khác nhau về cấu trúc cú pháp được ánh xạ về cùng một đồ thị AMR. Nhờ khả năng trừu tượng hóa này, AMR đã chứng minh tính hữu ích trong nhiều bài toán xử lý ngôn ngữ tự nhiên, đặc biệt là các nhiệm vụ đòi hỏi biểu diễn ngữ nghĩa sâu.

Trong bối cảnh bài toán chú thích ảnh, việc sử dụng AMR được xem là một hướng tiếp cận tiềm năng nhằm khắc phục hạn chế của các mô hình dựa thuần vào đặc trưng thị giác hoặc các quan hệ tường minh giữa đối tượng, vốn chưa thể hiện đầy đủ cấu trúc ngữ nghĩa sâu của nội dung ảnh. Tuy nhiên, số lượng các công trình khai thác AMR trong chú thích ảnh hiện vẫn còn hạn chế và chủ yếu tập trung vào việc xử lý ngôn ngữ hơn là thiết lập mối liên kết trực tiếp với thông tin thị giác.

Một số nghiên cứu tiên phong đã bước đầu khai thác AMR thông qua việc trích xuất biểu diễn ngữ nghĩa từ chính các chú thích chuẩn. Neto và cộng sự [73] đề xuất sử dụng AMR như một tầng trung gian trong bài toán chú thích dày đặc cho hình ảnh (dense image captioning), trong đó các chú thích chuẩn được chuyển sang dạng AMR tuyến tính nhằm giảm biến thiên cú pháp và giúp mô hình học biểu diễn ngữ nghĩa tổng quát hơn. Bhattacharyya và cộng sự [24] giới thiệu mô hình ReCAP, tích hợp thông tin vai trò ngữ nghĩa (Semantic Role Labeling - SRL) vào quá trình sinh chú thích, cho phép kiểm soát tốt hơn cấu trúc predicate-argument và đảm bảo tính mạch lạc ngữ nghĩa của câu mô tả. Bên cạnh đó, Kim và cộng sự [26] tập trung khai thác quan hệ chủ ngữ - tân ngữ nhằm giảm thiên lệch dữ liệu và cải thiện sự nhất quán trong nội dung chú thích sinh ra.

Gần đây hơn, một số nghiên cứu đã mở rộng hướng tiếp cận này theo hướng kiểm soát nội dung chú thích dựa trên biểu diễn ngữ nghĩa trừu tượng. Basioti và cộng

sự [25] đề xuất kỹ thuật Semantic Structure Alignment (SSA) trong bối cảnh controllable captioning, kết hợp nhiều đồ thị AMR trích xuất từ chú thích chuẩn thành một meta-graph nhằm điều hướng nội dung mô tả theo các vùng ảnh. Mặc dù phương pháp này bước đầu cho thấy khả năng kết nối giữa ngữ nghĩa trừu tượng và không gian thị giác, biểu diễn AMR vẫn chủ yếu được sinh từ ngôn ngữ, chưa phản ánh trực tiếp cấu trúc quan hệ thực tế trong ảnh đầu vào.

Tổng hợp các công trình liên quan cho thấy một hạn chế chung của hầu hết các phương pháp sử dụng AMR hiện nay là AMR chỉ được khai thác từ phía ngôn ngữ, khiến biểu diễn ngữ nghĩa trừu tượng không gắn kết chặt chẽ với nội dung thị giác của ảnh. Việc ánh xạ trực tiếp thông tin thị giác và các quan hệ giữa đối tượng trong ảnh sang cấu trúc AMR tương thích vẫn còn là một bài toán mở. Bên cạnh đó, chi phí xử lý đồ thị, sự thiếu hụt các công cụ parser AMR phù hợp cho dữ liệu hình ảnh, cũng như việc thiếu một cơ chế thống nhất để tích hợp đồng thời AMR và đặc trưng thị giác vào bộ giải mã ngôn ngữ, tiếp tục là những thách thức lớn trong hướng nghiên cứu này.

Những hạn chế nêu trên cho thấy nhu cầu phát triển các phương pháp chú thích ảnh có khả năng kết nối chặt chẽ giữa biểu diễn ngữ nghĩa trừu tượng và cấu trúc quan hệ trong ảnh, đồng thời tích hợp hiệu quả các nguồn thông tin này vào kiến trúc giải mã hiện đại. Đây cũng chính là khoảng trống nghiên cứu quan trọng, tạo tiền đề cho các hướng tiếp cận mới trong việc khai thác AMR nhằm nâng cao chiều sâu và độ chính xác ngữ nghĩa của chú thích ảnh tự động.

1.2.3. Đánh giá tổng quan các hướng tiếp cận

Từ tổng quan đã trình bày trong các mục trước, có thể nhận thấy rằng lĩnh vực chú thích ảnh tự động đã trải qua nhiều giai đoạn phát triển, với sự chuyển dịch rõ nét từ các hướng tiếp cận truyền thống dựa trên truy hồi thông tin và mẫu dữ liệu sang các mô hình học sâu hiện đại tích hợp thị giác và ngôn ngữ. Các mô hình CNN-LSTM đóng vai trò nền tảng, thiết lập khung mã hóa - giải mã ban đầu; trong khi đó, các kiến trúc Transformer đã nâng cao khả năng mô hình hóa ngữ cảnh và biểu diễn ngôn ngữ tự nhiên một cách linh hoạt hơn. Đồng thời, việc kết hợp đồ thị quan hệ (scene graph, relationship graph) và tri thức ngữ nghĩa từ bên ngoài (ConceptNet, WordNet, LLMs) đã góp phần cải thiện khả năng biểu đạt ngữ nghĩa và tăng cường tính nhất quán trong mô tả.

Tuy nhiên, đa số các hướng tiếp cận hiện hành vẫn còn tồn tại nhiều hạn chế chưa được giải quyết triệt để. Cụ thể: (i) các chú thích thường thiếu chiều sâu ngữ nghĩa và chưa thể hiện rõ được vai trò, hành động giữa các đối tượng; (ii) khả năng

biểu diễn và suy luận các quan hệ phức tạp còn hạn chế, nhất là khi chỉ dựa trên đặc trưng thị giác bề mặt; (iii) việc mô tả chính xác các đối tượng mới hoặc ít gặp trong tập huấn luyện vẫn là một thách thức; (iv) các biểu diễn ngữ nghĩa trừu tượng như AMR - dù tiềm năng - vẫn chưa được khai thác đầy đủ trong khung mô hình chú thích ảnh.

Việc nhận diện các điểm mạnh, điểm yếu và khoảng trống còn tồn tại ở mỗi nhóm tiếp cận là cơ sở quan trọng để định hướng nghiên cứu các phương pháp mới có khả năng tăng cường biểu diễn ngữ nghĩa, cải thiện tính khái quát và nâng cao hiệu quả mô tả trong những tình huống đa dạng và phức tạp hơn.

1.3. Khoảng trống và định hướng nghiên cứu

Dựa trên tổng quan và đánh giá các hướng tiếp cận được trình bày ở Mục 1.2, phần này phân tích những khoảng trống nghiên cứu còn tồn tại trong lĩnh vực chú thích ảnh tự động và đề xuất các định hướng giải quyết tiềm năng. Việc nhận diện rõ các khoảng trống giúp xác lập luận cứ khoa học cho việc phát triển các mô hình chú thích ảnh tích hợp tri thức ngữ nghĩa và biểu diễn ngữ nghĩa trừu tượng - định hướng xuyên suốt trong các chương tiếp theo của luận án.

1.3.1. Khoảng trống nghiên cứu

Mặc dù lĩnh vực chú thích ảnh tự động đã có những bước tiến đáng kể trong thập kỷ qua nhờ vào sự phát triển mạnh mẽ của các mô hình học sâu, nhưng vẫn còn tồn tại nhiều khoảng trống nghiên cứu mà các công trình hiện tại chưa giải quyết triệt để. Dưới đây là năm vấn đề đổi bật:

(i) **Biểu diễn và khai thác quan hệ giữa các đối tượng còn hạn chế**: Các mô hình đang có chủ yếu sử dụng đặc trưng toàn cục hoặc vùng cục bộ, trong khi quan hệ giữa các đối tượng (về không gian, ngữ nghĩa hoặc hành động) chưa được khai thác sâu. Việc biểu diễn cấu trúc ảnh dưới dạng đồ thị mới dừng ở mức đặc trưng bổ trợ, chưa được tích hợp chặt chẽ trong pipeline sinh chú thích.

(ii) **Khó mô tả đối tượng ngoài tập huấn luyện và việc tích hợp tri thức ngữ nghĩa chưa hiệu quả**: Do mô hình hiện tại phụ thuộc chặt vào dữ liệu huấn luyện và không gian từ vựng giới hạn, rất khó để nhận diện hoặc sinh mô tả phù hợp cho những đối tượng chưa từng xuất hiện trong quá trình học. Đồng thời, việc tận dụng các nguồn tri thức ngữ nghĩa bên ngoài - như ConceptNet, WordNet hay các mô hình ngôn ngữ lớn (GPTs, LLaMA…) - vẫn đang ở giai đoạn sơ khởi, gặp nhiều thách thức cả về cách biểu diễn tri thức lẫn cơ chế tích hợp để không gây nhiễu cho quá trình học, dẫn đến khả năng cân bằng giữa đặc trưng thị giác và tri thức ngữ nghĩa bên ngoài kém hiệu quả, làm giảm tính tổng quát hóa và chất lượng mô tả của mô hình.

(iii) **Ứng dụng AMR trong chú thích ảnh còn sơ khai**: Dù AMR đã chứng minh hiệu quả vượt trội trong nhiều tác vụ xử lý ngôn ngữ tự nhiên, nhưng việc kết hợp AMR với thông tin thị giác trong mô hình sinh chú thích cho hình ảnh vẫn chưa được nghiên cứu đầy đủ. Hiện tại, còn thiếu các giải pháp nhất quán để ánh xạ cấu trúc đồ thị từ ảnh sang biểu diễn AMR, cũng như cơ chế học biểu diễn AMR một cách hiệu quả.

(iv) **Hạn chế trong biểu diễn ngữ nghĩa sâu từ đặc trưng thị giác**: Phần lớn các mô hình hiện tại vẫn dựa trên đặc trưng trích xuất từ các mạng CNN hoặc các backbone như Faster R-CNN, ResNet, … Tuy nhiên, các đặc trưng này thường phản ánh sự hiện diện hoặc ngữ cảnh cục bộ của đối tượng, trong khi thiếu khả năng biểu diễn các liên kết ngữ nghĩa giữa các đối tượng cũng như liên kết giữa đặc trưng thị giác và ngôn ngữ. Điều này dẫn đến các mô tả đơn giản, thiếu chiều sâu và không phản ánh đầy đủ nội dung ảnh.

(v) **Hạn chế trong đánh giá chất lượng ngữ nghĩa:** Mặc dù các độ đo BLEU, METEOR, ROUGE, CIDEr hay SPICE phần nào đánh giá độ khớp cú pháp và chất lượng bề mặt, vẫn tồn tại khoảng cách đáng kể với nhận định của con người - nhất là khi đòi hỏi phản ánh chiều sâu ngữ nghĩa, đa dạng biểu đạt hay kiến thức mở rộng. Việc kết hợp đồng thời nhiều độ đo giúp đánh giá toàn diện hơn, nhưng vẫn chưa thể đo lường chính xác chiều sâu ngữ nghĩa và sự đa dạng biểu đạt của các chú thích.

Những khoảng trống nêu trên không chỉ cho thấy giới hạn hiện tại của các mô hình chú thích ảnh, mà còn mở ra nhiều hướng nghiên cứu giàu tiềm năng nhằm cải thiện chiều sâu ngữ nghĩa, khả năng tổng quát hóa và suy diễn của hệ thống thị giác - ngôn ngữ.

1.3.2. Định hướng nghiên cứu

Để khắc phục các khoảng trống đã được chỉ ra, luận án đề xuất một số định hướng nghiên cứu cụ thể, gắn chặt với từng vấn đề còn tồn tại, cụ thể như sau:

(i) **Tăng cường khai thác quan hệ giữa các đối tượng**: Xây dựng pipeline gồm phát hiện đối tượng, dự đoán quan hệ và xây dựng đồ thị quan hệ từ ảnh, sau đó khai thác toàn diện cấu trúc đồ thị bằng các kỹ thuật như mạng tích chập đồ thị nhằm làm giàu biểu diễn hình ảnh trước khi sinh chú thích.

(ii) **Tích hợp tri thức ngoài vào mô hình một cách có kiểm soát**: Sử dụng các nguồn tri thức như ConceptNet, WordNet, Visual Genome và các mô hình thị giác - ngôn ngữ huấn luyện trước (CLIP, BLIP-2…) để tạo embedding ngữ nghĩa phong phú. Các embedding này được đưa vào mô hình thông qua mô-đun hợp nhất tri thức (Knowledge Fusion), sử dụng cơ chế hợp nhất (fusion) để điều hòa tỷ lệ đóng

góp giữa tri thức ngoài và đặc trưng thị giác, đồng thời áp dụng cơ chế chú ý cross-modal để tinh chỉnh tương tác giữa hai miền và loại bỏ nhiễu không mong muốn. Cách tiếp cận này cho phép mô hình vừa mở rộng khả năng nhận diện và mô tả các đối tượng mới, vừa đảm bảo tính nhất quán và chiều sâu ngữ nghĩa cho các chú thích ảnh.

(iii) **Ứng dụng AMR vào quá trình học và sinh chú thích ảnh**: Kết hợp AMR theo hai hướng: (1) trích xuất AMR từ chú thích chuẩn để học biểu diễn ngữ nghĩa và điều chỉnh mô hình sinh chú thích; (2) ánh xạ từ đồ thị quan hệ sang đồ thị AMR-like từ ảnh để đưa vào mô hình. Embedding AMR có thể được học qua GCN hoặc các kỹ thuật học đồ thị, sau đó tích hợp vào bộ giải mã Transformer thông qua các cơ chế chú ý như masked multi-head hoặc cross-modal attention.

(iv) **Tăng cường biểu diễn ngữ nghĩa sâu từ đặc trưng thị giác**: Sử dụng các backbone hiện đại như Vision Transformer, Swin Transformer hoặc các mô hình thị giác-ngôn ngữ huấn luyện trước quy mô lớn (CLIP, BLIP-2, PaLI, Flamingo, Kosmos-2) làm bộ mã hóa; kết hợp huấn luyện đa nhiệm (nhận dạng đối tượng, vai trò, hành động) và cơ chế cross-modal attention để thu được đặc trưng ảnh mô tả toàn diện nội dung ảnh, tạo nền tảng cho việc tạo chú thích giàu ngữ nghĩa.

(v) **Đề xuất độ đo đánh giá ngữ nghĩa**: Đưa ra độ đo mới nhằm đánh giá tính tương đồng ngữ nghĩa giữa chú thích phát sinh và chú thích chuẩn, khắc phục hạn chế của các độ đo truyền thống vốn chủ yếu dựa trên so khớp n-gram và phản ánh chính xác hơn chất lượng ngữ nghĩa của mô tả [74].

Các định hướng nghiên cứu nêu trên không chỉ mở rộng khả năng mô hình hóa trong không gian thị giác - ngữ nghĩa mà còn đặt nền tảng cho các hướng tiếp cận học sâu đa phương thức có tích hợp tri thức. Đặc biệt, việc khai thác biểu diễn ngữ nghĩa trừu tượng như AMR, kết hợp với các mô hình ngôn ngữ - thị giác và độ đo ngữ nghĩa chuyên biệt, kỳ vọng nâng cao chất lượng mô tả cả về chiều sâu nội dung lẫn khả năng tổng quát hóa. Các định hướng này được cụ thể hóa thông qua bốn mô hình được trình bày tương ứng trong Chương 2, 3, 4 và 5 của luận án.

1.4. Phương pháp thực nghiệm và đánh giá

Nhằm đảm bảo tính nhất quán và khách quan trong quá trình kiểm thử các mô hình đề xuất, luận án xây dựng một quy trình thực nghiệm có cấu trúc rõ ràng, bao gồm ba thành phần chính: (i) dữ liệu thực nghiệm, (ii) độ đo đánh giá, và (iii) môi trường triển khai thực nghiệm.

1.4.1. Dữ liệu thực nghiệm

Trong bài toán chú thích ảnh, các bộ dữ liệu chuẩn thường bao gồm các cặp ảnh - chú thích được gán nhãn thủ công. Trong luận án này, ba nguồn dữ liệu chính được sử dụng gồm MS COCO và Flickr30K cho nhiệm vụ sinh chú thích ảnh, cùng với Visual Genome cho bài toán dự đoán quan hệ giữa các đối tượng trong ảnh. Việc lựa chọn kết hợp các tập dữ liệu này nhằm đánh giá một cách toàn diện khả năng tổng quát hóa và mở rộng của mô hình trên những tập dữ liệu có quy mô, mức độ chi tiết và độ phức tạp khác nhau.

- **MS COCO** (Microsoft Common Objects in Context) [75]: Là một trong những tập dữ liệu lớn và đa dạng nhất (hơn 330.000 ảnh) cho bài toán chú thích ảnh, mỗi ảnh đi kèm 5 chú thích ngôn ngữ tự nhiên do con người tạo ra. Tập dữ liệu có mức đa dạng cao về bối cảnh, đối tượng và tương tác giữa các đối tượng, và là tiêu chuẩn phổ biến để đánh giá các mô hình chú thích ảnh. *Thiết lập dùng trong luận án:* luận án tuân theo Karpathy split [76] với 82.783 ảnh huấn luyện, 5.000 ảnh kiểm định (validation) và 5.000 ảnh kiểm thử (testing); mỗi ảnh có 5 chú thích thu thập qua Amazon Mechanical Turk (AMT). MS COCO được dùng làm chuẩn chính để huấn luyện và đánh giá tất cả các mô hình.

- **Flickr30K** [77]: Gồm 31.783 ảnh, mỗi ảnh có 5 chú thích do người gán nhãn. Đây là tập dữ liệu cỡ vừa, thường được dùng để đánh giá hiệu quả mô hình trên ảnh từ bối cảnh thực tế hơn. *Thiết lập dùng trong luận án:* Karpathy split [78] với 29.783 ảnh huấn luyện, 1.000 ảnh kiểm định và 1.000 ảnh kiểm thử; được dùng làm đánh giá bổ sung nhằm kiểm tra khả năng khái quát hoá ngoài miền MS COCO.

- **Visual Genome** [79]: Tập dữ liệu phong phú về ngữ nghĩa, chứa hơn 100.000 ảnh kèm mô tả chi tiết, chú thích vùng đối tượng, thuộc tính, quan hệ giữa các đối tượng và câu hỏi - trả lời. Trong luận án này, Visual Genome chỉ được sử dụng để huấn luyện/tinh chỉnh phần dự đoán quan hệ (phục vụ xây dựng đồ thị quan hệ); việc sử dụng này không ảnh hưởng tới quy trình đánh giá chú thích trên MS COCO/Flickr30K.

Để đảm bảo thống nhất dữ liệu đầu vào và khả năng so sánh công bằng giữa các mô hình, quy trình chuẩn hoá dữ liệu thống nhất được áp dụng như sau:

Tiền xử lý văn bản: Từ vựng được xây dựng từ tập huấn luyện; sau khi loại bỏ các từ xuất hiện dưới 5 lần, kích thước tập từ vựng đạt 10.010 đối với MS COCO và 7.414 đối với Flickr30K. Độ dài tối đa của câu chú thích được đặt là 16 từ.

Tiền xử lý ảnh: Ảnh được chuẩn hoá kích thước đầu vào phù hợp với backbone thị giác của từng mô hình (ví dụ 224×224), và việc chuẩn hoá theo thống

kê của bộ pretrain tương ứng được thực hiện. Các đặc trưng/vùng bổ sung được trích xuất theo mô tả của từng chương.

Hình 1.4. Ví dụ về ảnh và các câu mô tả chú thích trong tập dữ liệu MS COCO (phải) và Flickr30K (trái). Mỗi ảnh đi kèm 5 chú thích ngôn ngữ tự nhiên do người gán nhãn.

Để minh họa dữ liệu huấn luyện, Hình 1.4 trình bày một ví dụ từ mỗi tập dữ liệu MS COCO (ảnh 000000000632.jpg) và Flickr30K (ảnh 1001633352.jpg), trong đó một ảnh được gán nhãn với 5 câu mô tả (*ground truth captions*) do con người tạo ra. Các mô tả này phản ánh sự đa dạng về ngôn ngữ trong việc biểu đạt nội dung thị giác.

1.4.2. Các độ đo đánh giá

Hiệu quả của các mô hình chú thích ảnh trong luận án được đánh giá dựa trên mức độ tương đồng giữa chú thích sinh ra và chú thích chuẩn (ground truth), sử dụng năm độ đo phổ biến trong lĩnh vực chú thích ảnh gồm: BLEU, METEOR, ROUGE, CIDEr và SPICE.

1.4.2.1. Độ đo BLEU

BLEU (Bilingual Evaluation Understudy) là một trong những độ đo phổ biến và có ảnh hưởng lớn trong việc đánh giá tự động chất lượng bản dịch máy và các mô hình sinh văn bản, bao gồm cả bài toán chú thích ảnh. BLEU được đề xuất bởi Papineni và cộng sự (2002) [80] nhằm mục tiêu tạo ra một phương pháp đánh giá tự động, nhanh chóng, chi phí thấp, và có tương quan cao với đánh giá của con người. Nguyên lý cốt lõi của BLEU là: một bản dịch được xem là tốt nếu nó có mức độ tương đồng cao với một hoặc nhiều bản dịch tham chiếu do con người tạo ra.

Khác với các tiêu chí đánh giá ngữ nghĩa hoặc cú pháp phức tạp, BLEU sử dụng cách tiếp cận dựa trên thống kê n-gram để so sánh mức độ trùng khớp giữa các cụm từ trong câu sinh ra với các câu chuẩn.

BLEU được tính toán dựa trên hai thành phần chính: (1) trung bình của các độ chính xác n-gram đã được điều chỉnh (modified n-gram precision), và (2) yếu tố phạt độ dài (brevity penalty) để tránh ưu tiên các bản dịch quá ngắn hoặc quá dài.

Đầu tiên, tính độ chính xác n-gram đã được điều chỉnh p_n:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \quad (1.1)$$

Trong đó:

- $Count_{clip}(n-gram) = min\big(Count(n-gram), max_{ref} Count_{ref}(n-gram)\big)$: số lượng *n-gram* trong chú thích sinh ra được giới hạn bởi số lượng tối đa của $\mathrm{n-gram}$ đó trong bất kỳ chú thích chuẩn nào.
- Tổng được tính trên toàn bộ tập chú thích sinh ra (candidates), thường ở cấp độ corpus để tăng độ ổn định.

Tiếp theo, tính yếu tố phạt độ dài (brevity penalty - BP):

$$BP = \begin{cases} 1 & nếu\ c > r \\ e^{(1-r/c)} & nếu\ c \le r \end{cases} \quad (1.2)$$

Trong đó:

- c: Độ dài tổng của tất cả chú thích sinh ra trong tập kiểm tra.
- r: Độ dài tham chiếu hiệu quả, được tính bằng cách chọn độ dài chú thích chuẩn gần nhất với chú thích sinh ra cho từng ảnh và tổng hợp lại (trong chú thích ảnh, thường có nhiều tham chiếu cho mỗi hình ảnh, nên lấy độ dài trung bình hoặc gần nhất).

Cuối cùng, giá trị BLEU được tính như sau:

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n log p_n\right) \quad (1.3)$$

Hoặc dưới dạng log để dễ phân tích hơn:

$$log\ BLEU = min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n log p_n \quad (1.4)$$

Thông thường, sử dụng $N = 4$ (tức là tính đến 4-gram) và trọng số đồng đều $w_n = {}^1/_4 = 0.25$. Giá trị BLEU nằm trong khoảng $[0, 1]$, với 1 biểu thị chú thích

hoàn toàn trùng khớp với tham chiếu. Trong chú thích ảnh, thường báo cáo BLEU-N riêng lẻ (ví dụ: BLEU-1 cho unigram, BLEU-4 cho 4-gram) để đánh giá chi tiết. BLEU đánh giá tốt trên tập dữ liệu lớn, có khả năng phân biệt rõ ràng giữa các bản dịch tốt và kém khi xét trung bình trên toàn bộ tập kiểm thử. Tuy nhiên, độ đo này có thể không phản ánh chính xác chất lượng trên từng câu riêng lẻ và không xét đến ngữ nghĩa sâu sắc nếu không có sự trùng khớp bề mặt.

1.4.2.2. Độ đo METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) là một độ đo được phát triển để đánh giá chất lượng đầu ra của các hệ thống sinh văn bản, bao gồm cả dịch máy và chú thích ảnh tự động. Được đề xuất bởi Banerjee và Lavie (2005) [81], METEOR khắc phục một số hạn chế của BLEU bằng cách kết hợp cả precision và recall, đồng thời đưa vào yếu tố đánh giá thứ tự từ ngữ nhằm phản ánh tốt hơn tính mạch lạc và đúng đắn về cú pháp-ngữ nghĩa của câu mô tả được sinh ra.

Trong bài toán chú thích ảnh, METEOR đánh giá mức độ tương đồng giữa câu mô tả sinh ra từ mô hình và các chú thích chuẩn do con người gán nhãn, bằng cách thực hiện khớp từ linh hoạt giữa hai câu.

Đầu tiên, thực hiện cơ chế khớp từ và tính điểm:

METEOR thực hiện so sánh ở mức unigram (từng từ riêng lẻ) giữa câu mô tả của mô hình và các câu tham chiếu. Quá trình khớp từ bao gồm ba giai đoạn chính:

- Khớp chính xác (Exact match): hai từ giống nhau hoàn toàn.
- Khớp theo gốc từ (Stem match): hai từ giống nhau sau khi rút gọn về gốc (ví dụ: “running” và “run”).
- Khớp theo từ đồng nghĩa (Synonym match): hai từ thuộc cùng nhóm đồng nghĩa trong WordNet.

Mỗi từ trong câu mô tả chỉ được khớp với một từ duy nhất trong câu tham chiếu, và ưu tiên khớp theo thứ tự trên.

Tiếp đó, tính toán độ chính xác (precision), độ bao phủ (recall) và F-mean:

- Precision (P): Tỷ lệ từ trong câu mô tả khớp với câu tham chiếu.

$$P = \frac{m}{t}$$

- Recall (R): Tỷ lệ từ trong câu tham chiếu được khớp với câu mô tả.

$$R = \frac{m}{r}$$

▪ F_mean: Trung bình điều hòa có trọng số giữa Precision và Recall, với trọng số nghiêng về Recall nhằm phản ánh mức độ bao phủ ngữ nghĩa:

$$F\_mean = \frac{10.P.R}{R + 9P} \quad (1.5)$$

Trong đó:

▪ m là số từ được khớp,

▪ t là tổng số từ trong câu mô tả,

▪ r là tổng số từ trong câu tham chiếu.

Sau đó, tính mức phạt phân mảnh (Fragmentation Penalty): Để đánh giá mức độ liên tục và đúng thứ tự của các từ khớp, METEOR xác định số cụm liên tục (chunks) gồm các từ khớp có thứ tự giống nhau trong cả hai câu. Số lượng cụm càng lớn thì câu mô tả càng rời rạc, và bị phạt nhiều hơn.

Công thức tính hình phạt:

$$Penalty = 0.5.\left(\frac{ch}{m}\right)^3 \quad (1.6)$$

Trong đó:

▪ ch là số cụm từ khớp liên tiếp,

▪ m là số từ khớp,

Cuối cùng, điểm số METEOR được tính bằng:

$$METEOR = F\_mean.(1 - Penalty) \quad (1.7)$$

Giá trị METEOR nằm trong đoạn [0,1], giá trị càng cao thể hiện câu mô tả gần với câu tham chiếu về bề mặt nội dung và cấu trúc ngôn ngữ.

Trong các nghiên cứu chú thích ảnh, METEOR được sử dụng phổ biến bên cạnh các độ đo khác như BLEU, ROUGE và CIDEr để đánh giá chất lượng mô tả tự động. Ưu điểm nổi bật của METEOR là khả năng đánh giá linh hoạt hơn so với BLEU nhờ hỗ trợ các phép ánh xạ từ mở rộng, đồng thời có tương quan cao hơn với đánh giá của con người, đặc biệt trong các tập dữ liệu có nhiều phong cách viết câu chú thích khác nhau như MS COCO hoặc Flickr30K.

1.4.2.3. Độ đo CIDEr

CIDEr (Consensus-based Image Description Evaluation) là một độ đo tự động được đề xuất bởi Vedantam và cộng sự (2015) [82] nhằm đánh giá chất lượng mô tả ảnh sinh ra bởi mô hình học sâu dựa trên mức độ phù hợp với sự đồng thuận của con người. Khác với các độ đo như BLEU hay METEOR vốn thiên về n-gram bề mặt,

CIDEr được thiết kế đặc biệt cho bài toán chú thích ảnh với khả năng phản ánh sự đa dạng và độ tự nhiên của các câu mô tả do con người tạo ra.

a) Nguyên lý hoạt động

Trong bài toán chú thích ảnh, với mỗi ảnh đầu vào I, giả sử có một tập các câu mô tả tham chiếu của con người $S_I = \{s_{i1}, s_{i2}, \dots, s_{iM}\}$, và một câu mô tả do mô hình sinh ra c_i. CIDEr đo lường độ tương đồng giữa câu sinh ra và tập các câu tham chiếu dựa trên phân tích thống kê n-gram (với n từ 1 đến 4), kết hợp với kỹ thuật TF-IDF để điều chỉnh trọng số theo độ phổ biến và đặc trưng của mỗi n-gram.

Điểm mạnh của CIDEr là nó không chỉ đo lượng độ khớp bề mặt, mà còn khuyến khích việc sử dụng các từ vựng quan trọng (salient words), hiếm và giàu thông tin trong ngữ cảnh hình ảnh.

b) Biểu diễn câu và trọng số TF-IDF

Mỗi câu được biểu diễn bằng một tập các n-gram. Với mỗi n-gram w_k, trọng số TF-IDF được tính như sau:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \Omega} h_l(s_{ij})} . log\left(\frac{|I|}{\sum_{I_p \in I} min(1, \sum_q h_k(s_{pq}))}\right) \quad (1.8)$$

Trong đó:

- $h_k(s_{ij})$: số lần xuất hiện của *n-gram* w_k trong câu tham chiếu s_{ij},
- Ω: tập hợp tất cả các *n-gram* trong tập dữ liệu,
- $|I|$: tổng số ảnh trong tập dữ liệu,

c) **Tính điểm CIDEr theo từng bậc n-gram**

Sau khi tính trọng số TF-IDF, điểm CIDEr theo bậc n-gram n được tính dựa trên độ đo tương tự cosine giữa véc-tơ đặc trưng của câu sinh ra c_i và các câu tham chiếu s_{ij}:

$$CIDEr_n(c_i, S_i) = \frac{1}{m}\sum_{j=1}^{m} \frac{g^n(c_i) . g^n(s_{ij})}{\|g^n(c_i)\| . \|g^n(s_{ij})\|} \quad (1.9)$$

Trong đó:

- $g^n(c_i)$: véc-tơ TF-IDF của các n-gram độ dài n trong câu sinh ra,
- m: số câu tham chiếu.

d) Tính điểm tổng hợp CIDEr

Điểm CIDEr tổng quát được tính bằng trung bình cộng các điểm theo từng bậc n-gram từ 1 đến 4:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{4} w_n . CIDEr_n(c_i, S_i) \quad (1.10)$$

Với trọng số $w_n = \frac{1}{4}$ cho mọi n. Theo thông lệ đánh giá trong cộng đồng, biến thể CIDEr-D được sử dụng bổ sung clipping n-gram (giới hạn số lần lặp theo mức tối đa trong tham chiếu) và phạt độ dài dạng Gaussian (khuyến khích độ dài câu hợp lý). Công thức CIDEr-D còn bao gồm một hệ số mở rộng thang điểm; khi hiển thị, các công cụ đánh giá thường nhân thêm hệ số hiển thị (ký hiệu "×100").

CIDEr là độ đo đặc thù cho bài toán chú thích ảnh, được thiết kế để phản ánh mức độ phù hợp với nhận thức chung của con người về nội dung ảnh. So với BLEU và METEOR, CIDEr cho thấy tương quan cao hơn với đánh giá thủ công của con người, đặc biệt khi số câu tham chiếu tăng lên (từ 5 đến 50 câu).

1.4.2.4. Độ đo ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) là một tập hợp các độ đo được đề xuất bởi Lin (2004) [83] nhằm đánh giá chất lượng của bản tóm tắt tự động bằng cách so sánh với một hoặc nhiều bản tóm tắt tham chiếu do con người tạo ra. Mặc dù ban đầu được phát triển cho bài toán tóm tắt văn bản, ROUGE cũng được áp dụng hiệu quả trong các bài toán sinh ngôn ngữ tự nhiên khác, bao gồm chú thích ảnh tự động, nơi các câu mô tả sinh ra được so sánh với các câu chuẩn về mức độ trùng khớp nội dung và trật tự từ vựng.

Trong bài toán chú thích ảnh, ROUGE đo lường mức độ tương đồng giữa câu mô tả của mô hình và tập các chú thích chuẩn dựa trên tiêu chí n-gram trùng khớp và chuỗi con chung dài nhất. Các biến thể phổ biến được sử dụng bao gồm ROUGE-N và ROUGE-L.

a) ROUGE-N: Thống kê trùng khớp n-gram

ROUGE-N đo tỷ lệ *n-gram* trong câu mô tả sinh ra mà cũng xuất hiện trong các câu tham chiếu. Với $n = 1$, ta có ROUGE-1 (trùng khớp từ đơn); với $n = 2$, ta có ROUGE-2 (trùng khớp cặp từ liên tiếp),…

Công thức tính ROUGE-N (dạng recall):

$$ROUGE - N = \frac{\sum_{S \in RefSummaries} \sum_{gram \in S} Count_{match}(gram_n)}{\sum_{S \in RefSummaries} \sum_{gram \in S} Count(gram_n)} \quad (1.11)$$

Trong đó:

- $Count_{match}$: số lần xuất hiện của *n-gram* trong cả câu mô tả và câu tham chiếu.
- Mẫu số là tổng số *n-gram* trong tất cả các câu tham chiếu.
- ROUGE-N là độ đo thiên về *recall*, nên đánh giá cao các câu mô tả bao phủ được nhiều n-gram quan trọng trong câu tham chiếu.

b) ROUGE-L: Chuỗi con chung dài nhất (Longest Common Subsequence)

ROUGE-L đánh giá dựa trên độ dài của chuỗi con chung dài nhất (LCS) giữa hai câu, phản ánh mức độ bảo toàn thứ tự từ ngữ giữa câu sinh và câu tham chiếu. So với n-gram, LCS không yêu cầu các từ trùng phải nằm liên tiếp nhưng phải giữ thứ tự xuất hiện.

Công thức ROUGE-L sử dụng F-measure:

$$R_{lcs} = \frac{LCS(X,Y)}{len(X)}, P_{lcs} = \frac{LCS(X,Y)}{len(Y)}$$

$$F_{lcs} = \frac{(1+\beta^2).P_{lcs}.R_{lcs}}{R_{lcs}+\beta^2.P_{lcs}} \quad (1.12)$$

Trong đó:

- X: câu tham chiếu, Y: câu mô tả sinh ra,
- $LCS(X,Y)$: độ dài chuỗi con chung dài nhất,
- β thường đặt rất lớn (≈ 8) để ưu tiên recall hơn precision.

Trong các nghiên cứu chú thích ảnh tự động, các biến thể như ROUGE-1, ROUGE-2, và ROUGE-L thường được sử dụng để đánh giá mức độ bao phủ nội dung và tính mạch lạc ngôn ngữ của mô tả sinh ra. Chúng đặc biệt hữu ích trong việc đánh giá khả năng tạo câu gần nghĩa với mô tả của con người, và có thể bổ sung cho các độ đo như BLEU hay CIDEr để đưa ra cái nhìn toàn diện hơn về hiệu suất mô hình.

1.4.2.5. Độ đo SPICE

SPICE (Semantic Propositional Image Caption Evaluation) là một độ đo đánh giá tự động cho bài toán sinh chú thích ảnh, được đề xuất bởi Anderson và cộng sự (2016) [84] với mục tiêu phản ánh sát hơn đánh giá của con người so với các độ đo truyền thống như BLEU, METEOR, ROUGE hay CIDEr. Khác với các phương pháp dựa trên so khớp *n-gram*, SPICE đo lường mức độ tương đồng về ngữ nghĩa giữa chú thích sinh ra và các câu tham chiếu, thông qua biểu diễn đồ thị cảnh (scene graph).

Phương pháp này tập trung vào việc phân tích nội dung ngữ nghĩa (semantic propositional content) bằng cách chuyển đổi chú thích sinh ra và các chú thích chuẩn

thành đồ thị cảnh, sau đó tính toán sự tương đồng dựa trên các đối tượng, thuộc tính, và mối quan hệ. SPICE khắc phục hạn chế của các độ đo dựa trên *n-gram* như BLEU hoặc ROUGE bằng cách ưu tiên so khớp ngữ nghĩa, giúp đánh giá tính chính xác, chi tiết, và nổi bật của chú thích mà không phụ thuộc vào từ vựng chính xác.

Việc tính điểm số SPICE được thực hiện như sau:

Đầu tiên, biểu diễn câu bằng scene graph. Với mỗi câu mô tả c, SPICE chuyển đổi câu này thành một đồ thị ngữ nghĩa $G(c) =< O(c), E(c), K(c) >$. Trong đó:

- $O(c)$: tập các đối tượng (*objects*),
- $E(c) \subseteq O(c) \times R \times O(c)$: tập quan hệ giữa các đối tượng,
- K$(c) \subseteq O(c) \times A$: tập thuộc tính gắn với đối tượng.

Tiếp theo, định nghĩa $T(G(c)) = O(c) \cup E(c) \cup K(c)$ là tập hợp tất cả các bộ ba ngữ nghĩa từ câu sinh ra. Tương tự, $T(G(S))$ là hợp nhất của các bộ ba từ câu tham chiếu.

Gọi toán tử so khớp là $\otimes$, với nghĩa là đếm số bộ ba giống nhau giữa 2 tập. Khi đó:

- Precision:

$$P(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$$

- Recall:

$$R(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

- F-score (SPICE Score):

$$SPICE(s,S) = F_1(c,S) = \frac{2.P(c,S).R(c,S)}{P(c,S) + R(c,S)} \quad (1.13)$$

Các bộ ba được xem là khớp nếu giống nhau về từ nguyên (lemma) hoặc thuộc cùng nhóm đồng nghĩa theo WordNet. SPICE không chấm điểm một phần: nếu chỉ một thành tố trong bộ ba sai thì không được tính điểm.

Giá trị SPICE nằm trong khoảng [0, 1], cao hơn biểu thị khớp ngữ nghĩa tốt hơn. Có thể phân tích theo từng loại (ví dụ: SPICE-objects, SPICE-relations) để đánh giá chi tiết. Phương pháp này đã được chứng minh qua các thực nghiệm trên MS COCO và Flickr8K, cho thấy tương quan cao với đánh giá con người trên tập dữ liệu lớn, và hỗ trợ tốt cho việc so sánh mô hình chú thích ảnh.

Mỗi độ đo nêu trên đều có ưu điểm và hạn chế riêng. Trong thực tế, nhiều nghiên cứu sử dụng kết hợp các độ đo (BLEU, METEOR, CIDEr, ROUGE, SPICE) để có cái nhìn toàn diện hơn. Tuy nhiên, vẫn còn tồn tại khoảng cách đáng kể giữa các đánh giá tự động và nhận định thực tế của con người, đặc biệt khi chú thích cần phản ánh thông tin ngữ nghĩa sâu, sự đa dạng biểu đạt hoặc các kiến thức mở rộng. Điều này đặt ra nhu cầu phát triển các độ đo mới tập trung vào đánh giá ngữ nghĩa và khả năng tổng quát hóa của mô hình, là một hướng nghiên cứu đầy tiềm năng trong tương lai.

1.4.3. Môi trường triển khai thực nghiệm

Toàn bộ mô hình được triển khai bằng ngôn ngữ Python và một số framework, thư viện chính như sau:

- PyTorch: Xây dựng và huấn luyện mô hình học sâu.
- Torch-Geometric: Xử lý và học biểu diễn trên đồ thị (GCN).
- NeuralAMR (AMR Parser): Xử lý biểu diễn chú thích chuẩn thành đồ thị AMR.

Việc huấn luyện mô hình được thực hiện trên Google Colab Pro, GPU NVIDIA Tesla, sử dụng batch size, learning rate, và số epoch được tối ưu hóa cho từng mô hình riêng biệt. Các thành phần thực nghiệm nêu trên đóng vai trò cốt lõi trong việc kiểm thử và đánh giá hiệu quả của các mô hình được đề xuất trong luận án.

1.5. Kết chương

Chương này cung cấp một cái nhìn tổng quan có hệ thống về lĩnh vực nghiên cứu chú thích ảnh, đặc biệt tập trung vào các phương pháp dựa trên mạng học sâu. Thông qua việc phân tích và phân loại các hướng tiếp cận như CNN-LSTM, mô hình dựa trên đồ thị, Transformer, tích hợp tri thức ngữ nghĩa từ nguồn ngoài, và biểu diễn trừu tượng AMR, chương đã làm rõ tiến trình phát triển, ưu điểm và những hạn chế còn tồn tại của từng nhóm phương pháp. Từ đó, các khoảng trống nghiên cứu quan trọng đã được xác định, bao gồm: hạn chế trong biểu diễn quan hệ và ngữ nghĩa sâu, khó khăn khi mô tả các đối tượng ngoài tập huấn luyện, và sự thiếu vắng các độ đo phản ánh đầy đủ chất lượng ngữ nghĩa trong sinh mô tả. Ngoài ra, chương cũng đã giới thiệu khung thực nghiệm thống nhất bao gồm tập dữ liệu, độ đo đánh giá và môi trường cài đặt - những thành phần được sử dụng xuyên suốt trong các chương tiếp theo. Những nội dung trình bày trong chương này đóng vai trò làm nền tảng lý thuyết và định hướng kỹ thuật quan trọng cho việc phát triển và triển khai bốn mô hình chú thích ảnh tự động được đề xuất trong luận án.

CHƯƠNG 2. MÔ HÌNH CHÚ THÍCH ẢNH SỬ DỤNG BIỂU DIỄN ĐỒ THỊ QUAN HỆ GIỮA CÁC ĐỐI TƯỢNG

Trong các mô hình chú thích ảnh gần đây, phần lớn phương pháp vẫn dựa vào đặc trưng toàn cục của ảnh hoặc đặc trưng vùng độc lập, dẫn đến hạn chế trong việc nắm bắt ngữ cảnh và quan hệ giữa các đối tượng. Để cải thiện khả năng mô hình hóa cấu trúc ngữ nghĩa này, chương 2 giới thiệu mô hình **OD-VR-Cap**, đại diện cho bước khởi đầu trong chuỗi phát triển của luận án. Mô hình kết hợp đặc trưng đối tượng với thông tin quan hệ thông qua biểu diễn đồ thị quan hệ và cơ chế chú ý kép, giúp tăng cường khả năng hiểu cấu trúc cảnh và ngữ nghĩa khi sinh chú thích.

Chương này đưa ra một cơ chế tích hợp thông tin quan hệ vào quá trình sinh chú thích, qua đó nâng cao khả năng mô hình nắm bắt cấu trúc ngữ nghĩa của ảnh và tạo nền tảng cho các hướng mở rộng ở những chương tiếp theo. Các đóng góp của chương này được công bố trong công trình [CT1], đồng thời cũng đóng vai trò nền tảng cho các phát triển trình bày trong các công trình [CT4], [CT5] và [CT6].

Nội dung của Chương được tổ chức như sau: Mục 2.1 trình bày hướng tiếp cận; Mục 2.2 mô tả chi tiết mô hình OD-VR-Cap; Mục 2.3 trình bày thiết lập thực nghiệm và kết quả; cuối cùng, Mục 2.4 kết luận chương và thảo luận vai trò của mô hình trong chuỗi nghiên cứu của luận án.

2.1. Giới thiệu

Các mô hình tạo chú thích ảnh đạt kết quả tốt chủ yếu dựa trên mô hình học sâu kết hợp cơ chế chú ý và được tổ chức theo khung mã hóa - giải mã [85]. Ở khung này, bộ mã hóa (image encoder) học biểu diễn nội dung trực quan và ánh xạ ảnh thành véc-tơ đặc trưng, còn bộ giải mã (language decoder) - thường là mô hình ngôn ngữ - nhận véc-tơ này để phát sinh câu mô tả. Nhìn chung, các công trình này có thể quy về hai nhóm chính: (i) CNN-RNN/LSTM [3, 76], và (ii) các biến thể tăng cường cơ chế chú ý [18, 86].

Tuy nhiên, khi ảnh được nén về một véc-tơ duy nhất bởi các mạng CNN huấn luyện trước, mô hình khó nắm bắt ngữ nghĩa ở cấp đối tượng cũng như sự tương tác giữa các các đối tượng này; đồng thời việc căn chỉnh hoặc liên kết (alignment) giữa vùng ảnh và các từ trong câu thường kém hiệu quả [87]. Do đó, một hướng cải thiện tự nhiên là trích xuất đối tượng trong ảnh và mô hình hóa các quan hệ giữa chúng để bộ mã hóa giữ lại thông tin giàu ngữ cảnh hơn, tạo tiền đề cho bộ giải mã sinh chú thích chính xác hơn. Mặt khác, các phương pháp có cơ chế chú ý đã chứng minh hiệu quả khi giúp bộ giải mã tập trung động vào những phần liên quan của ảnh thay vì dựa toàn bộ ảnh [85]. Dẫu vậy, nhiều nghiên cứu chỉ khai thác đặc trưng thị giác vùng

ảnh, trong khi ngữ nghĩa nhãn lớp và thông tin quan hệ giữa các đối tượng chưa được khai thác đầy đủ. Vì thế, cần một cơ chế chú ý kép kết hợp chú ý thị giác (trên đặc trưng vùng) với chú ý ngữ nghĩa/đồ thị (trên nhãn lớp và quan hệ), để gán trọng số thích nghi cho từng thành phần đặc trưng khi giải mã, qua đó cải thiện chất lượng chú thích.

Xuất phát từ các hạn chế nêu trên, chương này đề xuất tiếp cận mã hóa đồ thị quan hệ kết hợp chú ý kép nhằm nâng cao độ chính xác. Cụ thể, đồ thị quan hệ được xây dựng bằng cách phát hiện đối tượng và dự đoán quan hệ giữa các cặp đối tượng; cấu trúc này cho phép mô hình tổ chức lại ngữ cảnh ảnh theo dạng liên kết, giúp hiểu nội dung một cách hệ thống hơn trước khi đưa vào bộ giải mã. Song song, chú ý kép đồng thời khai thác đặc trưng thị giác và đặc trưng ngữ nghĩa từ đồ thị, nhờ đó tăng khả năng chọn lọc thông tin phù hợp ở mỗi bước sinh từ. Đóng góp trọng tâm của chương được tóm lược như sau:

- Đưa ra mô hình nâng cao độ chính xác phát hiện đối tượng kết hợp mạng tích chập đồ thị, mô hình này có thể dễ dàng tích hợp vào các kỹ thuật phát hiện đối tượng khác.

- Đưa ra mô hình dự đoán mối quan hệ giữa các đối tượng trong ảnh dựa vào đặc trưng vùng đối tượng, thông tin ngữ cảnh và tri thức quan hệ giữa các đối tượng trong tập dữ liệu. Từ đó, tạo ra đồ thị quan hệ để biểu diễn nội dung hình ảnh đầy đủ ngữ nghĩa.

- Giới thiệu cách biểu diễn đồ thị quan hệ của hình ảnh thành véc-tơ đặc trưng, cách biểu diễn này khai thác được ngữ nghĩa của mối quan hệ dựa trên mạng tích chập đồ thị.

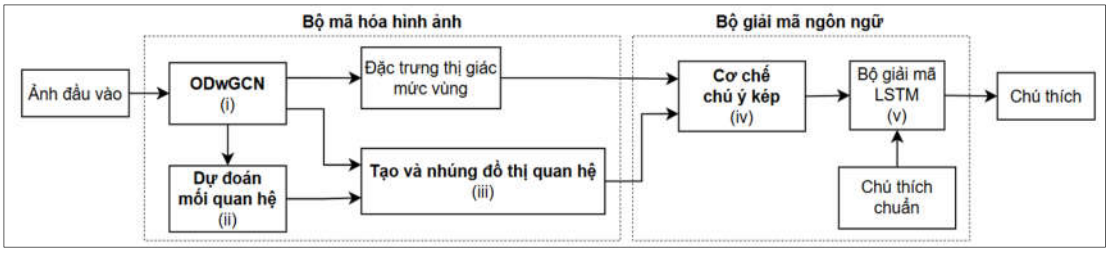
- Đưa ra cơ chế chú ý kép tập trung vào cả đặc trưng vùng đối tượng và đặc trưng các đỉnh trong đồ thị quan hệ của hình ảnh. Từ đó, xây dựng mô hình chú thích ảnh dựa vào đồ thị quan hệ của hình ảnh và mạng LSTM với cơ chế chú ý kép.

2.2. Phương pháp chú thích ảnh đề xuất

Trong chương này, mô hình chú thích ảnh được xây dựng theo kiến trúc mã hóa - giải mã (encoder-decoder), như minh họa trong Hình 2.1, gồm hai thành phần:

- Bộ mã hóa hình ảnh (Image Encoder): học và biểu diễn nội dung ảnh dưới dạng các đặc trưng giàu thông tin ngữ nghĩa, bao gồm (i) phát hiện vùng đối tượng cải tiến bằng ODwGCN; (ii) dự đoán quan hệ giữa các đối tượng bằng mô hình VRP+RK; và (iii) xây dựng - biểu diễn đồ thị quan hệ (R-Graph/R-Graph*) dưới dạng embedding.

▪ Bộ giải mã ngôn ngữ (Language Decoder): tạo câu mô tả bằng ngôn ngữ tự nhiên dựa trên đặc trưng đầu ra của bộ mã hóa; trong nghiên cứu này, bộ giải mã gồm (iv) cơ chế chú ý kép (dual attention) và (v) mạng LSTM nhằm kết hợp đồng thời đặc trưng vùng đối tượng và đặc trưng cấu trúc từ đồ thị quan hệ.



Hình 2.1. Kiến trúc tổng thể của mô hình chú thích ảnh OD-VR-Cap. (i) ODwGCN phát hiện các vùng đối tượng; (ii) Mô-đun dự đoán quan hệ giữa các đối tượng; (iii) Xây dựng và biểu diễn đồ thị quan hệ dưới dạng embedding; (iv) cơ chế chú ý kép kết hợp đặc trưng vùng và đặc trưng đồ thị; (v) Bộ giải mã LSTM sinh câu chú thích dựa trên các đặc trưng kết hợp và chú thích chuẩn.

2.2.1. Kiến trúc tổng thể của mô hình đề xuất

Hình 2.1 mô tả toàn bộ mô hình OD-VR-Cap như một chuỗi xử lý liên tiếp gồm bộ mã hóa hình ảnh và bộ giải mã ngôn ngữ. Trên cơ sở đó, các thuật toán trong mục này được sắp xếp theo đúng thứ tự dòng dữ liệu từ ảnh đầu vào đến câu chú thích đầu ra, đồng thời mỗi thuật toán được gắn trực tiếp với một khối chức năng trong kiến trúc tổng thể.

Cụ thể, trong bộ mã hóa hình ảnh, khối phát hiện đối tượng cải tiến **ODwGCN** được trình bày trước, bao gồm quy trình học trọng số/embedding để hiệu chỉnh phát kết quả phát hiện đối tượng của các mô hình huấn luyện trước; nội dung này tương ứng với **Thuật toán 2.1-2.2**. Tiếp theo, khối dự đoán quan hệ **VRP+RK** được mô tả như một tác vụ phân lớp quan hệ giữa các cặp vùng đối tượng, trong đó tri thức quan hệ được học từ cơ sở tri thức và được đưa vào bộ phân lớp; quy trình dự đoán quan hệ giữa hai vùng đối tượng được cụ thể hóa bằng **Thuật toán 2.3**. Sau khi có tập đối tượng và tập quan hệ, đồ thị quan hệ G và đồ thị quan hệ mở rộng G^* được xây dựng; việc học embedding cho các nút của G^* nhằm tạo đặc trưng cấu trúc - ngữ nghĩa phục vụ giải mã được mô tả trong **Thuật toán 2.4**.

Trong bộ giải mã ngôn ngữ, hai nguồn đặc trưng đầu vào gồm đặc trưng vùng đối tượng F và embedding đồ thị Z^* được tích hợp thông qua cơ chế chú ý kép. Quy trình tính điểm liên kết, chuẩn hóa trọng số chú ý và tạo hai véc-tơ ngữ cảnh $c_t^{(v)}$ và $c_t^{(g)}$ tại mỗi bước thời gian được mô tả trong **Thuật toán 2.5**. Các véc-tơ ngữ cảnh này sau đó được đưa vào LSTM để cập nhật trạng thái và sinh từ tiếp theo theo các công thức (2.7)-(2.8), từ đó tạo thành chuỗi chú thích hoàn chỉnh.

Để thuận tiện cho việc mô tả chi tiết mô hình đề xuất, các ký hiệu và định nghĩa liên quan sau được sử dụng:

▪ Tập dữ liệu được sử dụng cho bài toán chú thích ảnh gồm có N_T mẫu, mỗi mẫu được biểu diễn dưới dạng một cặp (I, S). Trong đó, I là ảnh đầu vào và $S = \{s_1, s_2, \dots, s_{N_S}\}$ là chú thích chuẩn tương ứng. Mỗi s_i đại diện cho từ thứ i trong câu, $\forall i \in \{1, \dots, N_S\}$.

▪ B, N_B lần lượt là tập vùng đối tượng và số vùng đối tượng phát hiện được (detected boxes) trong ảnh.

▪ C, N_C lần lượt là tập nhãn lớp và số lượng nhãn lớp đối tượng trong tập dữ liệu.

▪ Y là ma trận độ tin cậy thu được từ kết quả của mô hình phát hiện đối tượng huấn luyện trước, $\hat{Y}$ là ma trận độ tin cậy sau khi thực hiện điều chỉnh bằng mạng GCN.

▪ K là cơ sở tri thức chứa N_E thực thể (entity) bao gồm các $subjects$, $objects$ và $predicates$; E là tập thực thể.

▪ R, N_R lần lượt là tập mối quan hệ và số lượng mối quan hệ có trong K.

▪ $G^{(1)}, G^{(2)}, G, G^*$ lần lượt là đồ thị tương quan nhãn đối tượng, đồ thị thực thể, đồ thị quan hệ và đồ thị quan hệ mở rộng; $X^{(1)}, X^{(2)}$ là ma trận đặc trưng của $G^{(1)}$ và $G^{(2)}$; X là ma trận đặc trưng của G^*.

▪ $h_v^{(l)}$ là trạng thái ẩn của đỉnh v tại tầng l của GCN, h_t là trạng thái ẩn của LSTM cell tại thời điểm t.

▪ Có 3 hàm tổng mất mát hay chi phí (cost function), ký hiệu $L_1(\varphi)$ cho bài toán chú thích ảnh, $L_2(\varphi)$ cho bài toán huấn luyện mạng GraphSAGE theo phương pháp có giám sát trên đồ thị $G^{(1)}$ dựa vào nhãn của các vùng đối tượng trong ảnh và $L_3(\varphi)$ cho bài toán huấn luyện mạng GraphSAGE theo phương pháp không giám sát dựa vào độ tương tự của các đỉnh lân cận trong đồ thị $G^{(2)}$ và $G^{(4)}$.

Trong đó:

$$L_1(\varphi) = -\frac{1}{N_T}\sum_{i=1}^{N_T}\sum_{t=1}^{N_S^{(i)}} log\, P(s_t^{(i)} | s_1^{(i)}, s_2^{(i)}, \dots, s_{t-1}^{(i)}, f_I^{(i)}; \varphi) \quad (2.1)$$

Trong (2.1), $N_S^{(i)}, s_t^{(i)}, f_I^{(i)}$ lần lượt là số từ trong chú thích, từ đúng tại thời điểm t, đặc trưng ảnh của mẫu dữ liệu thứ i; $P(s_t^{(t)} | s_1^{(i)}, s_2^{(i)}, \dots, s_{t-1}^{(i)}, f^{(i)})$ là xác suất

dự đoán từ $y_t^{(i)}$ tại thời điểm t của mẫu thứ i dựa trên các từ đã dự đoán trước đó và đặc trưng của ảnh đầu vào.

$$L_2(\varphi) = -\frac{1}{N_B}\sum_{i=1}^{N_B}\sum_{j}^{N_C}\left(y_{ij}\, log\left(\hat{y}_{ij}\right);\varphi\right) \tag{2.2}$$

Trong (2.2), $y_{ij}, \hat{y}_{ij} \in \{0,1\}$ lần lượt là nhãn chuẩn và nhãn dự đoán của mô hình phát hiện đối tượng kết hợp GCN cho bounding box i thuộc về lớp đối tượng j.

$$L_3(\varphi) = \sum_{v\in V^{(2)}}\left(\sum_{u\in N(v)} -log\left(\sigma\left(h_v^{(N_L)}\cdot h_u^{(N_L)}\right)\right) + \sum_{n\in N'} log\left(1-\sigma\left(h_v^{(N_L)}\cdot h_n^{(N_L)}\right)\right);\varphi\right) \tag{2.3}$$

Trong (2.3), σ là hàm sigmoid, $h_v^{(N_L)}$ là véc-tơ embedding của đỉnh v tại tầng cuối (N_L), $N(v)$ là tập các đỉnh lân cận của v, $N'(v)$ là tập các đỉnh được chọn ngẫu nhiên không phải lân cận của v.

Ba hàm mất mát L_1, L_2, L_3 được xây dựng tương ứng với ba nhiệm vụ riêng biệt trong mô hình đề xuất. Mỗi hàm đảm nhiệm một mục tiêu tối ưu cụ thể và được sử dụng trong các giai đoạn huấn luyện khác nhau, phù hợp với bản chất của từng bài toán. Cách tổ chức này giúp đảm bảo sự tách biệt rõ ràng giữa các mục tiêu học và duy trì tính nhất quán trong quá trình xây dựng mô hình.

Dựa trên các ký hiệu và định nghĩa trên, các mục tiếp theo trình bày chi tiết từng thành phần trong mô hình đề xuất theo đúng luồng xử lý của kiến trúc tổng thể.

2.2.2. Bộ mã hóa hình ảnh

Bộ mã hóa hình ảnh bao gồm: (i) mô hình phát hiện đối tượng kết hợp GCN, gọi là ODwGCN; (ii) mô hình dự đoán mối quan hệ giữa các đối tượng trong ảnh; (iii) đồ thị quan hệ, bao gồm cách tạo đồ thị quan hệ và cách biểu diễn các nút của đồ thị quan hệ.

2.2.2.1. Mô hình phát hiện đối tượng kết hợp mạng tích chập đồ thị

Các mô hình phát hiện đối tượng phổ biến như SSD, Faster R-CNN hay YOLO, vốn được huấn luyện sẵn trên các tập dữ liệu lớn, đã chứng minh được hiệu quả trong nhiều bài toán thị giác máy tính, bao gồm cả chú thích ảnh. Tuy nhiên, khi xử lý các hình ảnh có nhiều đối tượng hoặc cấu trúc phức tạp, các mô hình này thường gặp khó khăn trong việc phân biệt và định danh chính xác các đối tượng. Hạn chế chính nằm ở chỗ chúng chỉ tập trung vào đặc trưng cục bộ của từng đối tượng riêng

lẻ, trong khi bỏ qua mối quan hệ ngữ cảnh giữa các đối tượng trong cùng một khung hình, dẫn đến khả năng nhận diện sai hoặc thiếu nhất quán trong một số trường hợp.

Để khắc phục vấn đề này, nghiên cứu đề xuất một mô hình phát hiện đối tượng cải tiến, ký hiệu là ODwGCN, được minh họa trong Hình 2.2. Mô hình gồm hai giai đoạn chính:

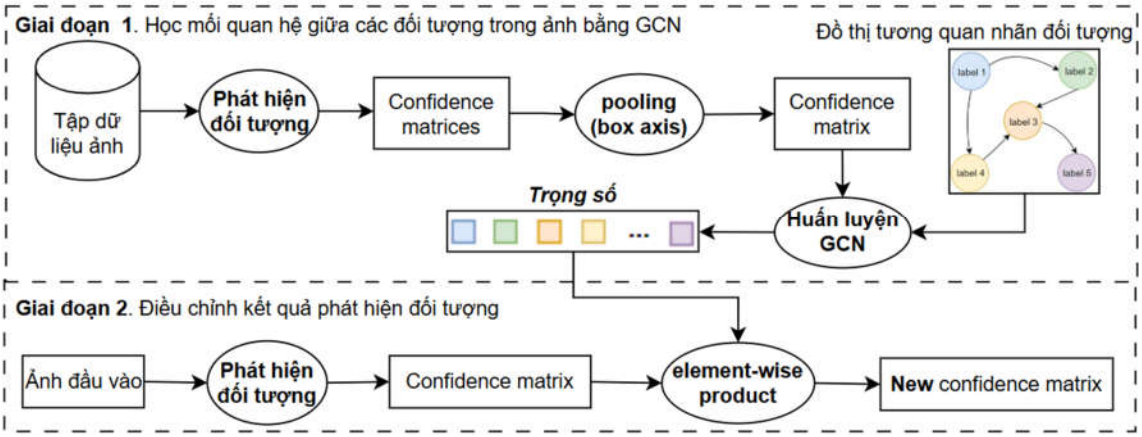
(a) Ở giai đoạn đầu, mô hình học các mối quan hệ đồng xuất hiện giữa các đối tượng trong ảnh bằng cách sử dụng mạng tích chập đồ thị (Graph Convolutional Network - GCN) để mô hình hóa sự phụ thuộc ngữ cảnh.

(b) Ở giai đoạn tiếp theo, kết quả phát hiện từ các mô hình nền tảng (SSD, Faster R-CNN, YOLO) được hiệu chỉnh lại dựa trên các mối quan hệ học được ở giai đoạn trước, thông qua hệ số điều chỉnh tương quan giữa các đối tượng.

Cách tiếp cận này giúp tăng độ chính xác trong nhận diện đối tượng, đặc biệt đối với các ảnh chứa nhiều đối tượng tương tác hoặc có bố cục phức tạp, đồng thời cung cấp nguồn đặc trưng đầu vào giàu ngữ nghĩa hơn cho giai đoạn xây dựng đồ thị quan hệ ở phần sau của mô hình.

a) Giai đoạn 1: Học mối quan hệ đồng xuất hiện giữa các đối tượng trong ảnh bằng GCN

Ở giai đoạn đầu này, mô hình xây dựng một đồ thị biểu diễn sự tương quan giữa các nhãn đối tượng nhằm mô tả mức độ đồng xuất hiện của chúng trong cùng một ảnh. Dựa trên đồ thị này, mạng tích chập đồ thị được thiết kế và huấn luyện để học các mẫu quan hệ đồng xuất hiện giữa các đối tượng, với dữ liệu đầu vào là kết quả phát hiện từ các mô hình phát hiện đối tượng được huấn luyện trước.



Hình 2.2. Mô hình phát hiện đối tượng cải tiến kết hợp mạng tích chập đồ thị (ODwGCN) gồm hai giai đoạn: học quan hệ nhãn và hiệu chỉnh ma trận độ tin cậy

a.1) Đồ thị tương quan nhãn đối tượng

Đồ thị này biểu diễn mối quan hệ đồng xuất hiện giữa các nhãn lớp đối tượng trong tập dữ liệu hình ảnh và được xây dựng dựa trên nguyên lý của mô hình ML-GCN [88]. Mục tiêu là khai thác sự phụ thuộc ngữ cảnh giữa các đối tượng để hỗ trợ việc hiệu chỉnh các phát hiện sai hoặc thiếu nhất quán trong giai đoạn nhận dạng. Trong kiến trúc ML-GCN, mỗi đỉnh trong đồ thị tương ứng với một nhãn lớp đối tượng của tập dữ liệu, còn đặc trưng của đỉnh được biểu diễn thông qua véc-tơ nhúng từ (word embedding) của nhãn lớp đó. Ma trận kề được xác định bằng cách thống kê tần suất đồng xuất hiện giữa các nhãn trong cùng một ảnh, qua đó phản ánh mức độ liên kết giữa các lớp đối tượng.

Trong nghiên cứu này, đặc trưng của các đỉnh được trích xuất trực tiếp từ kết quả phát hiện đối tượng trong ảnh thay vì sử dụng véc-tơ nhúng từ. Cách tiếp cận này giúp đồ thị biểu diễn chính xác hơn mối quan hệ thực tế giữa các đối tượng trong không gian hình ảnh, đồng thời tăng cường khả năng khai thác thông tin ngữ cảnh, phục vụ cho việc hiệu chỉnh kết quả phát hiện trong các giai đoạn tiếp theo.

Định nghĩa 2.1. Đồ thị tương quan nhãn LC-Graph $G^{(1)} = \left(V^{(1)}, A^{(1)}\right)$ là một đồ thị có hướng bao gồm:

– Tập đỉnh $V^{(1)} = \{v_i^{(1)} \in C, \forall i = \overline{1, N_C}\}$, mỗi đỉnh $v_i^{(1)}$ đại diện cho một nhãn lớp trong tập dữ liệu ảnh.

– Ma trận kề $A^{(1)} = \{a_{ij}^{(1)} \in \{0,1\}, \forall i, j = \overline{1, N_C}\}$, mỗi phần tử $a_{ij}^{(1)}$ trong ma trận kề biểu thị sự tồn tại của một cạnh có hướng giữa 2 đỉnh $v_i^{(1)}$ và $v_j^{(1)}$.

Quá trình tạo ma trận kề $A^{(1)}$ được tiến hành theo các bước sau:

(1) Trước hết, gọi l_i, l_j lần lượt là nhãn của lớp đối tượng o_i và o_j; đồng thời, m_{ij} biểu thị số lần cặp nhãn l_i và l_j cùng xuất hiện trong các ảnh của tập dữ liệu huấn luyện. Khi đó, ta xác định ma trận tần suất đồng xuất hiện $M \in N^{N_C \times N_C}$ như sau:

$$M = \{m_{ij} \geq 0, \forall i, j = \overline{1, N_C}, i \neq j$$

(2) Tiếp theo, gọi n_i là tổng số lần xuất hiện của nhãn l_i trong toàn bộ tập dữ liệu huấn luyện. Dựa trên đó, ma trận hệ số tương quan giữa các nhãn lớp được tính bằng công thức:

$$D = \left\{d_{ij} = \frac{m_{ij}}{n_i}, \forall i, j = \overline{1, N_C}\right\}$$

(3) Sau khi thu được ma trận D, tiến hành nhị phân hóa để loại bỏ các quan hệ nhiễu, thu được ma trận tương quan nhị phân - đồng thời cũng là ma trận kề của đồ thị.

$$A^{(1)} = \left(a_{ij}^{(1)} = \begin{bmatrix} 0, & if\ d_{ij} < \tau \\ 1, & if\ d_{ij} \geq \tau \end{bmatrix}, \forall i,j = \overline{1,N_C} \right)$$

Trong đó, $\tau \in (0,1)$ là ngưỡng lọc nhằm loại bỏ các cạnh có tần suất đồng xuất hiện thấp, giúp giảm nhiễm trong đồ thị. Giá trị $\tau = 0.5$ được sử dụng trong các thực nghiệm của nghiên cứu này.

a.2) Mạng tích chập đồ thị cho mô hình phát hiện đối tượng

Mạng tích chập đồ thị được giới thiệu lần đầu trong công trình *"Semi-Supervised Classification with Graph Convolutional Networks"* [89], với mục tiêu xử lý bài toán phân lớp bán giám sát trên dữ liệu có cấu trúc đồ thị. Cốt lõi của phương pháp này nằm ở cơ chế lan truyền thông tin giữa các đỉnh lân cận, cho phép cập nhật biểu diễn của mỗi đỉnh dựa trên đặc trưng của các đỉnh liền kề. Cụ thể, tại mỗi lớp GCN, véc-tơ đặc trưng của một đỉnh mới được tính bằng cách tổng hợp (thường là trung bình có trọng số) các đặc trưng của các đỉnh kề cùng với chính đỉnh đó ở lớp trước. Cách làm này giúp mô hình khai thác được mối quan hệ cục bộ trong cấu trúc đồ thị, từ đó cải thiện hiệu quả biểu diễn và phân biệt giữa các nhóm đỉnh có liên kết mạnh. Tuy nhiên, khi số tầng GCN tăng lên, hiện tượng đồng nhất hóa đặc trưng (over-smoothing) có thể xảy ra: các đỉnh nằm trong các vùng đồ thị có cấu trúc *k-hop* tương tự nhau dần trở nên khó phân biệt, vì véc-tơ đặc trưng của chúng trở nên gần như giống nhau ở các lớp sâu hơn. Điều này làm giảm khả năng biểu diễn ngữ nghĩa của mô hình, đặc biệt trong các bài toán cần phân biệt tinh tế giữa các quan hệ phức tạp.

Để khắc phục hạn chế này, nghiên cứu sử dụng GraphSAGE [90] thay cho GCN truyền thống nhằm học biểu diễn cho các đỉnh dựa trên mối quan hệ giữa các đối tượng trong ảnh. Khác với cơ chế gộp trung bình toàn cục của GCN, GraphSAGE tổng hợp thông tin từ tập các đỉnh lân cận thông qua một hàm lấy mẫu và hàm gộp (aggregator) để tạo ra véc-tơ đại diện cho vùng lân cận. Véc-tơ này sau đó được kết hợp với đặc trưng của chính đỉnh đang xét để sinh ra biểu diễn mới. Cách tiếp cận này giúp giữ được sự khác biệt giữa các đỉnh, đồng thời mở rộng khả năng tổng quát hóa cho các cấu trúc đồ thị phức tạp trong bài toán phát hiện đối tượng.

Với mạng GraphSAGE có đầu vào là đồ thị LC-Graph$\left(G^{(1)}\right)$ theo Định nghĩa 2.1, ký hiệu là GraphSAGENet1, quá trình thực hiện giai đoạn 1 được thể hiện qua **Thuật toán 2.1**, gồm 3 bước như sau:

Bước 1 (detect objects): Các mô hình phát hiện đối tượng huấn luyện trước được sử dụng để xác định và trích xuất các vùng chứa đối tượng trong tập ảnh huấn luyện. Sau khi xử lý một ảnh đầu vào I, hệ thống thu được ma trận độ tin cậy $Y \in R^{N_B \times N_C}$, trong đó, mỗi phần tử y_{ij} biểu diễn xác suất mà vùng đối tượng thứ i được dự đoán thuộc về lớp đối tượng j, $\forall i = 1, \dots, N_B; \forall j = 1, \dots, N_C$.

Bước 2 (pooling): Chọn giá trị đại diện cho các lớp đối tượng trong ma trận confidence (*có thể mà max, min hoặc average*). Giá trị max được chọn trong thực nghiệm của nghiên cứu này, từ đó, mỗi ma trận độ tin cậy của ảnh thu được một véc-tơ N_C chiều. Khi đó, $X^{(1)} = \{x_i^{(1)} \in R^{N_C} | i = 1,2, \dots, N_T\}$ là tập véc-tơ thu được từ việc chọn giá trị đại diện cho các đối tượng trong ma trận độ tin cậy.

$$x_{ij}^{(1)} = \underset{k=\overline{1,N_B}}{max}\{Y_{kj}^{(i)}\}, \forall j = \overline{1, N_C}, \forall i = \overline{1, N_T} \quad (2.4)$$

Tập véc-tơ $X^{(1)}$ làm đầu vào cho mạng tích chập đồ thị, đóng vai trò là tập véc-tơ đặc trưng của các đỉnh trong đồ thị LG-Graph.

Bước 3: Huấn luyện **GraphSAGENet1** theo phương pháp học có giám sát (supervised) với hàm mất mát là L_2 để tạo ra embedding cho các nhãn lớp đối tượng, đây cũng là véc-tơ trọng số (w) giúp điều chỉnh kết quả của các mô hình phát hiện đối tượng huấn luyện trước.

Véc-tơ trọng số $w \in R^{N_C}$ phản ánh mức độ ảnh hưởng ngữ cảnh giữa các nhãn đối tượng. Kích thước của w tương ứng với số lớp đối tượng N_C, và các giá trị của w được học từ dữ liệu thông qua GraphSAGE nhằm đảm bảo phù hợp với cấu trúc đồng xuất hiện của các nhãn trong tập huấn luyện.

Thuật toán 2.1 mô tả quá trình học trọng số cho các nhãn lớp đối tượng dựa trên đồ thị LC-Graph. Trước hết, các mô hình phát hiện đối tượng huấn luyện trước được sử dụng để tạo ma trận độ tin cậy Y. Từ đó, mỗi ảnh được biểu diễn bằng một véc-tơ đặc trưng $x^{(1)}$ thông qua phép pooling theo từng lớp đối tượng. Tập các véc-tơ này được sử dụng làm đặc trưng đầu vào cho GraphSAGENet1 nhằm học các trọng số ngữ cảnh giữa các nhãn đối tượng.

Sau khi **GraphSAGENet1** đã được huấn luyện, mô hình đã học được tập ma trận trọng số $W^{(1)}$. Khi đó, thuật toán phát sinh embedding cho các đỉnh của đồ thị được mô tả như ở **Thuật toán 2.2**.

Thuật toán 2.1. LearningWeightStage1($G^{(1)}$, $\widehat{M}$)

Đầu vào: Tập dữ liệu ảnh I, đồ thị $G^{(1)}\left(V^{(1)}, A^{(1)}\right)$, mô hình phát hiện đối tượng huấn luyện trước $\widehat{M}$.

Đầu ra: Véc-tơ trọng số w.

Begin

$Y^{(i)} = \widehat{M}(I_i), \forall i = \overline{1..N_T}$

$X^{(1)} = []$

for $i = 1$ to N_T do

$x_{ij}^{(1)} = \underset{k=1,N_B}{max}\left\{Y_{kj}^{(i)}\right\}, \forall j = \overline{1,N_C}$

 # Add the row $x^{(1)}$ *to the matrix* $X^{(1)}$

$X^{(1)} = \begin{bmatrix} X^{(1)} \\ x^{(1)} \end{bmatrix}$

endfor

$N_{L_1}, W^{(1)} = TrainingGraphSAGENet1\left(G^{(1)}, X^{(1)}\right)$

$w = GenerateEmbedding\left(G^{(1)}, X^{(1)}, N_{L_1}, W^{(1)}\right)$

return w

End

Thuật toán 2.2. GenerateEmbedding$\left(G^{(1)}, X^{(1)}, N_{L_1}, W^{(1)}\right)$

Đầu vào: đồ thị $G^{(1)} = \left(V^{(1)}, A^{(1)}\right)$, véc-tơ đặc *trưng* $X^{(1)} = \left\{x_v^{(1)}, \forall v \in V^{(1)}\right\}$, số tầng N_{L_1}, ma trận trọng số $W^{(l,1)}, \forall l = \overline{1, N_{L_1}}$, hàm kích hoạt phi tuyến σ.

Đầu ra: véc-tơ embedding $z_v, \forall v \in V^{(1)}$.

Begin

$h_v^{(0)} \leftarrow x_v^{(1)}, \forall v \in V^{(1)}$

for $l = \overline{1..N_L}$ *do*

for $v \in V^{(1)}$ ***do***

$h_{N(v)}^{(l)} = f_{agg}\left(\left\{h_u^{(l-1)}, \forall u \in N(v)\right\}\right);$

$h_v^{(l)} = \sigma\left(W^{(l,1)}.\left[h_v^{(l-1)}, h_{N(v)}^{(l)}\right]\right)$

endfor

$h_v^{(l)} \leftarrow h_v^{(l)} / \left\| h_v^{(l)} \right\|_2, \forall v \in V^{(1)}$

endfor

$z_v \leftarrow h_v^{(N_L)}, \forall v \in V^{(1)}$

return z_v

End

Thuật toán 2.2 tính toán embedding cho các đỉnh trong đồ thị LC-Graph thông qua nhiều tầng GraphSAGE. Mỗi đỉnh bắt đầu với một véc-tơ đặc trưng ban đầu $x_v^{(1)}$, và qua mỗi tầng, thuật toán tổng hợp đặc trưng từ các đỉnh liền kề, sau đó kết hợp thông tin này với đặc trưng hiện tại của đỉnh thông qua một ma trận trọng số và một hàm kích hoạt phi tuyến σ. Đặc trưng này sau đó được chuẩn hóa trước khi chuyển sang tầng tiếp theo. Sau khi đi qua tất cả các tầng, đặc trưng cuối cùng của mỗi đỉnh $\left(h_v^{(N_L)}\right)$ được sử dụng làm véc-tơ embedding đại diện của đỉnh. Embedding z_v là biểu diễn ngữ cảnh của nhãn đối tượng v, có cùng kích thước đặc trưng đối với tất cả các đỉnh và được chuẩn hóa để đảm bảo tính ổn định trong các bước xử lý tiếp theo của mô hình.

b) Giai đoạn 2: Hiệu chỉnh kết quả phát hiện đối tượng

Dựa trên véc-tơ embedding (hay còn gọi là hệ số điều chỉnh - adjustment weight), vốn mô tả mối quan hệ giữa các nhãn lớp đối tượng đã được học trong Giai đoạn 1, quá trình điều chỉnh lại kết quả phát hiện đối tượng cho từng ảnh đầu vào được thực hiện theo các bước cụ thể như sau:

Bước 1 (dectect objects): Ở bước này, các mô hình phát hiện đối tượng đã được huấn luyện trước được áp dụng để xác định và trích xuất các vùng chứa đối tượng trong ảnh đầu vào. Kết quả đầu ra là ma trận độ tin cậy Y.

Bước 2 : thực hiện phép toán $element - wise\ product$ giữa ma trận Y và véc-tơ trọng số điều chỉnh w được ký hiệu là $\odot$ với tác động như sau:

$$\hat{Y} = Y \odot \alpha w = \begin{bmatrix} y_{11} & \cdots & y_{1N_C} \\ \cdots & \cdots & \cdots \\ y_{N_B1} & \cdots & y_{N_BN_C} \end{bmatrix} \odot \alpha \left[w_1 \ldots w_{N_C} \right] = \begin{bmatrix} y_{11}\alpha w_1 & \cdots & y_{1N_C}\alpha w_{N_C} \\ \cdots & \cdots & \cdots \\ y_{N_B1}\alpha w_1 & \cdots & y_{N_BN_C}\alpha w_{N_C} \end{bmatrix} \quad (2.5)$$

Trong đó, $\hat{y}_{ij}$ biểu thị xác suất vùng đối tượng thứ i được gán cho nhãn lớp j sau khi áp dụng hệ số điều chỉnh α. Nếu trong (2.5) ta đặt $\alpha = 1$ và mọi thành phần của véc-tơ trọng số w đều bằng 1, thì ma trận độ tin cậy ban đầu được giữ nguyên, tức là không có sự hiệu chỉnh nào từ mạng tích chập đồ thị.

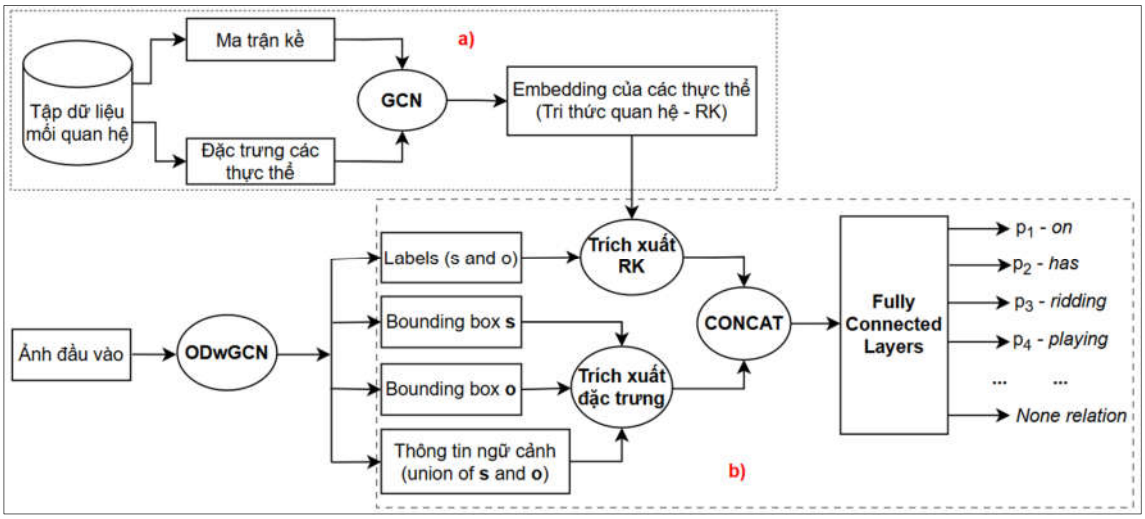
Khi đó, ma trận $\hat{Y}$ chính là đầu ra cuối cùng của mô hình ODwGCN cho bài toán phát hiện đối tượng.

2.2.2.2. Mô hình dự đoán mối quan hệ giữa các đối tượng

Mối quan hệ giữa các đối tượng trong ảnh góp phần quan trọng trong việc hiểu đầy đủ nội dung hình ảnh. Tuy nhiên, các công trình công bố gần đây có hạn chế là thường tập trung vào một số loại mối quan hệ cụ thể, chẳng hạn như quan hệ vị trí, hành động (*tương tác giữa các đối tượng*),…Hơn nữa, thường chỉ sử dụng đặc trưng

các vùng đối tượng [91, 92] hoặc kết hợp đặc trưng vùng đối tượng và đặc trưng ngữ cảnh mối quan hệ (*union của 2 vùng đối tượng*) [22, 93] mà chưa khai thác thông tin tri thức quan hệ vốn có trong tập dữ liệu dẫn đến độ chính xác chưa cao. Vì vậy, trong nghiên cứu này, mô hình dự đoán mối quan hệ được đề xuất có thể nhận ra nhiều loại mối quan hệ khác nhau, khai thác thông tin tri thức quan hệ giữa các thực thể trong tập dữ liệu nhằm nâng cao độ chính xác.

Mối quan hệ giữa các cặp đối tượng trong ảnh thường được biểu diễn dưới dạng bộ ba, còn gọi là triplet, <$subject$, $predicate$, $object$>. Trong đó, $predicate$ là từ được sử dụng để liên kết các cặp đối tượng, ví dụ như: <$woman$, $riding$, $motorcycle$>, <$flower$, in, $vase$>…Chú ý rằng, đây là mối quan hệ có hướng giữa một đối tượng ($subject$) và một đối tượng khác ($object$) thông qua một $predicate$. Trong chương này, mô hình dự đoán mối quan hệ giữa các đối tượng được mô tả trong *Hình 2.3,* gọi là VRP$^+$RK**,** gồm các bước chính: (a) học tri thức quan hệ bằng GCN và (b) phân lớp mối quan hệ giữa các đối tượng bằng mạng FC (*Fully connected*). Cụ thể, đầu vào là hai vùng đối tượng và thông tin ngữ cảnh của mối quan hệ (*vùng chứa cả* 2 đối tượng). Để tăng độ chính xác, tri thức quan hệ giữa hai đối tượng được thêm vào. Các đặc trưng này được ghép lại (CONCAT), sau đó đưa qua các tầng kết nối đầy đủ để tạo ra xác suất phân lớp trên N_{R+1} mối quan hệ (N_R mối quan hệ và lớp $none - relation$).



Hình 2.3. Kiến trúc mô hình dự đoán quan hệ giữa các đối tượng. (a) GCN được huấn luyện trên tập dữ liệu quan hệ để tạo embedding tri thức quan hệ (RK) giữa các thực thể; (b) ODwGCN trích xuất các cặp đối tượng, vùng đối tượng và thông tin ngữ cảnh, sau đó kết hợp đặc trưng thị giác và embedding RK để dự đoán nhãn quan hệ thông qua các tầng fully connected.

a) Học tri thức quan hệ

Để khai thác tri thức quan hệ giữa các đối tượng nhằm nâng cao độ chính xác cho việc dự đoán mối quan hệ, tập dữ liệu bộ ba biểu diễn đồ thị cảnh trong Visual Genome [79] được sử dụng. Các triplets trong tập huấn luyện được tổ chức thành một cơ sở tri thức K chứa N_E thực thể (*entity*), bao gồm các $subjects$, $objects$ và $predicates$. Trong nghiên cứu này, cơ sở tri thức K được biểu diễn dưới dạng một đồ thị thực thể, từ đó học mối quan hệ giữa các thực thể bằng mạng tích chập đồ thị.

Định nghĩa 2.2. Đồ thị thực thể **E-Graph** $G^{(2)} = \left(V^{(2)}, A^{(2)}\right)$ là đồ thị vô hướng, bao gồm:

– Tập đỉnh $V^{(2)} = \left\{v_i^{(2)} \in E, \forall i = \overline{1, N_E}\right\}$

– Ma trận kề $A^{(2)} = \left\{a_{ij}^{(2)} \in \{0,1\}, \forall i, j = \overline{1, N_E}\right\}$

Trong đó, E là tập thực thể trong K, ma trận kề nhị phân $A^{(2)}$ được xây dựng để biểu diễn mối quan hệ giữa các thực thể, cụ thể là cho biết có hay không mối quan hệ giữa các $subjecs/objects$ và $predicates$. Ví dụ: K có chứa triple <$child$, has, tie> thì các phần tử $a_{child,has}^{(2)}$ và $a_{has,tie}^{(2)}$ trong ma trận kề $A^{(2)}$ có giá trị là 1. Đặc trưng của các thực thể (*đỉnh của đồ thị*) là véc-tơ embedding (*word embedding*) của các thực thể E.

Từ đồ thị E-Graph như **Định nghĩa** 2.2 và đặc trưng của tập đỉnh, mạng tích chập đồ thị GraphSAGE [90] được sử dụng để học tri thức quan hệ giữa các thực thể, ký hiệu là GraphSAGENet2. Với $X^{(2)} = \{x_i^{(2)} \in E, \forall i = \overline{1, N_E}\}$ là tập véc-tơ embedding của các thực thể. Trong phần này, mạng GraphSAGENet2 được huấn luyện theo phương pháp học không giám sát để cập nhật các embedding của đỉnh bằng cách tối ưu hàm mất mát dựa trên độ tương tự giữa các đỉnh và các đỉnh liền kề trong đồ thị (contrastive loss) với hàm mất mát L_3.

Sau khi GraphSAGENet2 đã được huấn luyện và học được các ma trận trọng số $W^{(l,2)}, \forall l = \overline{1, N_{L_2}}$ (N_{L_2} là số layer của mạng). Khi đó, embedding các đỉnh của đồ thị, hay còn gọi là tri thức quan hệ của các thực thể được phát sinh bằng cách áp dụng **Thuật toán 2.2**.

$$Z^{(K)} \leftarrow GenerateEmbedding\left(G^{(2)}, X^{(2)}, N_{L_2}, W^{(2)}\right) \tag{2.6}$$

Trong (2.6), $Z^{(K)} \in R^{N_E \times N_D}$ là tri thức quan hệ của tập thực thể E (N_D là số chiều véc-tơ embedding của các thực thể). Sau đó, đặc trưng tri thức quan hệ của các nhãn lớp của đối tượng từ $Z^{(K)}$ có thể được trích xuất để bổ sung cho đầu vào của mạng fully connected nhằm tăng độ chính xác phân lớp mối quan hệ.

b) Phân lớp mối quan hệ giữa các đối tượng

Trong nghiên cứu này, việc dự đoán mối quan hệ được thực hiện như là một tác vụ phân lớp với đầu vào là hai vùng đối tượng trong ảnh (b_1 và b_2) và tri thức quan hệ của các nhãn đối tượng. Đầu ra là một trong số N_{R+1} mối quan hệ, gồm N_R mối quan hệ giữa hai đối tượng và "$none\ relation$". Quá trình dự đoán mối quan hệ giữa hai đối tượng trong ảnh được mô tả như trong **Thuật toán 2.3**.

Thuật toán 2.3 thực hiện dự đoán mối quan hệ giữa hai vùng đối tượng dựa trên sự kết hợp giữa đặc trưng thị giác và tri thức quan hệ đã được học trước đó. Trước hết, vùng union của hai bounding box được xác định nhằm trích xuất thông tin ngữ cảnh không gian của cặp đối tượng. Tuy nhiên, cần lưu ý rằng vùng union chỉ đóng vai trò cung cấp ngữ cảnh bổ sung, không phải là tiêu chí duy nhất để suy luận quan hệ giữa hai đối tượng. Tiếp theo, đặc trưng thị giác của từng bounding box và của vùng union được trích xuất thông qua mạng ResNet-101. Các đặc trưng này được tổng hợp để tạo thành biểu diễn thị giác chung cho cặp đối tượng. Đồng thời, tri thức quan hệ của hai nhãn lớp tương ứng được truy xuất từ ma trận embedding $Z^{(\mathcal{K})}$, là kết quả của quá trình học trên đồ thị thực thể E-Graph bằng GraphSAGE (2.6). Mỗi vector trong $Z^{(\mathcal{K})}$ biểu diễn quan hệ ngữ nghĩa của một thực thể trong không gian embedding. Cuối cùng, biểu diễn thị giác và embedding tri thức quan hệ được ghép nối và đưa vào mạng fully connected để thực hiện phân lớp trên tập N_{R+1} quan hệ. Nhờ sự kết hợp này, việc dự đoán quan hệ không chỉ dựa trên thông tin hình học hoặc ngữ cảnh cục bộ, mà còn khai thác cấu trúc ngữ nghĩa toàn cục từ cơ sở tri thức.

Thuật toán 2.3. **PredictRelationship**$\left(b_1, b_{2,}, Z^{(K)}\right)$

Đầu vào: bounding box b_1, bounding box b_2, $Z^{(K)}$
Đầu ra: r - quan hệ giữa 2 vùng đối tượng
Begin

$b_u \leftarrow UoI(b_1, b_2)$
$f_1 = ResNet - 101(b_1)$
$f_2 = ResNet - 101(b_2)$
$f_u = ResNet - 101(b_u)$
$f_{avg} \leftarrow \frac{1}{3}(f_1 + f_2 + f_u)$
$k_{b1} \leftarrow z_1^{(K)} \in Z^{(K)}$
$k_{b2} \leftarrow z_2^{(K)} \in Z^{(K)}$
$x \leftarrow f_{avg} \parallel k_{b1} \parallel k_{b2}$
Predict the relationship
$r \leftarrow FCN\left(x, W^{(3)}\right)$
return r

End

2.2.2.3. Đồ thị quan hệ

Đồ thị quan hệ là một công cụ mạnh mẽ để mô hình hóa các mối quan hệ phức tạp giữa các đối tượng trong thế giới thực. Trong cấu trúc này, các đối tượng được biểu diễn dưới dạng các đỉnh, còn các mối quan hệ giữa chúng được thể hiện dưới dạng các cạnh. Phần này định nghĩa khái niệm đồ thị quan hệ, đồng thời trình bày cách xây dựng và biểu diễn đồ thị này.

Trong nghiên cứu này, đồ thị quan hệ của một hình ảnh được định nghĩa như sau:

Định nghĩa 2.3. Đồ thị quan hệ của hình ảnh **R-Graph** $G = (V, E)$ là đồ thị có hướng bao gồm:

– Tập đỉnh $V = \{v_i \in B, \forall i = \overline{1, N_B}\}$

– Tập cạnh $E = \{e_{ij} = (v_i, v_j, r_{ij}) \in R, \forall i, j = \overline{1, N_B}, i \neq j\}$

a) Cách tạo đồ thị quan hệ

Sau khi bộ phân lớp quan hệ được huấn luyện trên tập dữ liệu quan hệ thị giác, mô hình này được kết hợp với ODwGCN để xây dựng đồ thị quan hệ cho ảnh đầu vào. Trước hết, từ ảnh đầu vào I, ODwGCN phát hiện N_B vùng đối tượng $\{b_i\}_{i=1}^{N_B}$; mỗi vùng đối tượng tương ứng với một đỉnh v_i trong đồ thị. Khi đó, tập đỉnh của đồ thị quan hệ được xác định là $V = \{v_i | b_i \in B\}$.

Tiếp theo, các vùng đối tượng trong ảnh được xem xét theo từng cặp (b_i, b_j) với $i \neq j$, tạo thành $N_B(N_B - 1)$ cặp vùng đối tượng. Đối với mỗi cặp (b_i, b_j), mô hình dự đoán mối quan hệ được áp dụng để ước lượng phân phối xác suất trên $\mathrm{N_R} + 1$ lớp quan hệ, bao gồm $\mathrm{N_R}$ quan hệ ngữ nghĩa và một lớp "$none\ relation$". Nếu xác suất của lớp "$none\ relation$" nhỏ hơn ngưỡng θ cho trước, mô hình giả định rằng giữa hai đối tượng tồn tại một quan hệ có ý nghĩa; khi đó, một cạnh được thêm vào đồ thị, liên kết hai đỉnh tương ứng v_i và v_j, với nhãn cạnh là lớp quan hệ có xác suất dự đoán cao nhất.

Quá trình này được áp dụng cho toàn bộ các cặp đối tượng trong ảnh, từ đó hình thành đồ thị quan hệ $G = (V, E)$, trong đó các đỉnh biểu diễn các đối tượng được phát hiện và các cạnh mô tả mối quan hệ ngữ nghĩa giữa chúng. Hình 2.4(c) minh họa một ví dụ đồ thị quan hệ thu được từ ảnh đầu vào, trong đó:

– Tập đỉnh $V = \{child, tie, water\ bowl\}$;

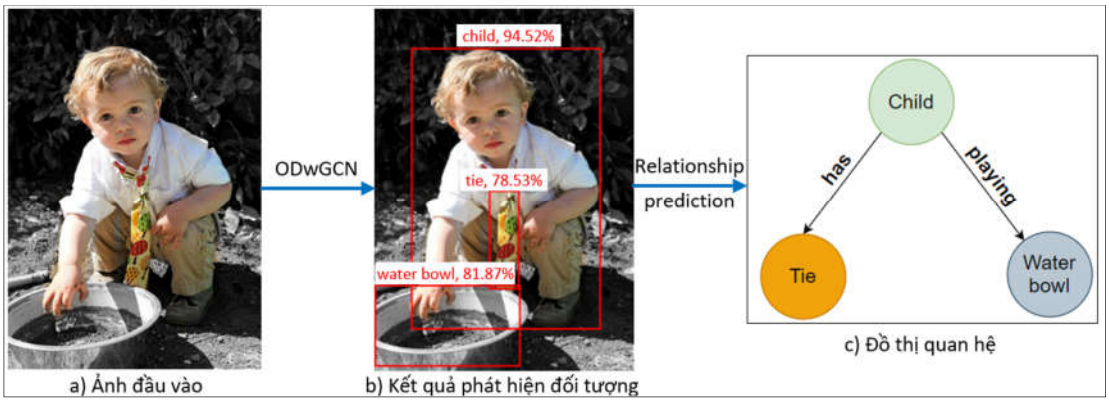
– Tập cạnh $E = \{< child, tie, has >, < child,\ water\ bowl, playing >\}$.

b) Cách biểu diễn và embedding đồ thị quan hệ

Mặc dù đồ thị quan hệ có khả năng mô tả thông tin trong ảnh một cách toàn diện và chính xác, song cấu trúc không đồng nhất của nó khiến cho việc sử dụng trực tiếp làm đầu vào cho hầu hết các thuật toán ngữ nghĩa trở nên không phù hợp [94]. Vì lý do đó, cần thiết phải chuyển đổi đồ thị quan hệ sang dạng tuyến tính, vừa đảm bảo giữ nguyên thông tin quan trọng, vừa có thể tích hợp dễ dàng vào các kiến trúc học sâu, chẳng hạn như khi đưa vào mô hình ngôn ngữ để sinh chú thích cho ảnh. Ngoài ra, để khai thác ngữ nghĩa của nhãn lớp đối tượng cũng như mối quan hệ giữa các đối tượng, thay vì chỉ dựa trên đặc trưng vùng và cấu trúc đồ thị như nhiều nghiên cứu trước, luận án này đề xuất mở rộng đồ thị quan hệ của hình ảnh thành đồ thị quan hệ mở rộng được ký hiệu là **R-Graph*** $G^* = (V^*, E^*)$:

– Tập đỉnh $V^* = \{v_i^* \in L, \forall i = \overline{1, N_L}\}$

– Tập cạnh $E^* = \{e_{ij}^* \in \{0,1\}, \forall i, j = \overline{1, N_L}, i \neq j\}$



Hình 2.4. Quá trình tạo đồ thị quan hệ từ ảnh đầu vào: a) là ảnh đầu vào, b) là kết quả sau khi thực hiện mô hình phát hiện đối tượng cải tiến ODwGCN, c) là đồ thị quan hệ thu được sau khi thực hiện dự đoán mối quan hệ giữa các đối tượng.

Trong đó, Tập đỉnh L gồm có hai loại đỉnh: đỉnh là nhãn của các vùng đối tượng và đỉnh là nhãn quan hệ giữa các đối tượng ($predicates$). Tập cạnh E^* được xây dựng theo quy tắc: Nếu có một cạnh $e_{ij} = (v_i, v_j, r_{ij}) \in E$ thì tạo hai cạnh có hướng lần lượt là cạnh từ đỉnh v_i đến đỉnh nhãn r_{ij} và cạnh từ đỉnh nhãn r_{ij} đến đỉnh v_j. Ví dụ như đồ thị quan hệ $G = (V, E)$ ở Hình 2.4c) có các đối tượng: $child, tie, water\ bowl$; các mối quan hệ giữa các đối tượng: $< child,\ tie,\ has >$, $< child,\ water\ bowl,\ playing >$ thì được chuyển thành đồ thị $G^* = (V^*, E^*)$ gồm 5 đỉnh và 4 cạnh như Hình 2.5.

Quá trình học biểu diễn các đỉnh của đồ thị R-Graph* được thực hiện bằng mạng GraphSAGE, trong đó sử dụng phương pháp học không giám sát với mục tiêu tối ưu hàm mất mát contrastive loss L_3, dựa trên mức độ tương đồng giữa một đỉnh và các

đỉnh kề trong cấu trúc đồ thị. Việc tạo ra véc-tơ embedding cho các đỉnh được tiến hành theo **Thuật toán 2.4**, bao gồm các bước sau:

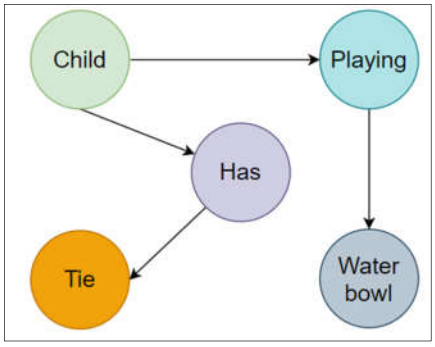
– **Bước 1:** Khởi tạo véc-tơ đặc trưng cho nhãn đỉnh bằng kỹ thuật word2vec.

– **Bước 2**: Chia tập đỉnh kề thành hai nhóm: tập đỉnh vào $(N-(v))$ gồm các đỉnh có cạnh đi đến v, và tập đỉnh ra $(N+(v))$ gồm các đỉnh mà v trỏ tới.

– **Bước 3**: Kết hợp thông tin của $(N-(v))$, $(N+(v))$ cùng với đặc trưng hiện tại của chính đỉnh v, qua đó hình thành véc-tơ trung gian, $h_{v-}^{(N_L)}, h_{v+}^{(N_L)}$.

– **Bước 4**: Nối hai véc-tơ $h_{v-}^{(N_L)}$ và $h_{v+}^{(N_L)}$ để thu được biểu diễn cuối cùng của đỉnh v, $z_v^* = [h_{v-}^{(N_L)}, h_{v+}^{(N_L)}], \forall v \in V^*$.

– **Bước 5**: Lặp lại các thao tác từ bước từ 2 đến 4 N_L vòng lặp, kết quả cuối cùng là véc-tơ biểu diễn ổn định cho đỉnh v.



*Hình 2.5. Kết quả khi chuyển đổi đồ thị quan hệ R-Graph ở Hình 4c) thành đồ thị R-Graph**

Dựa trên kết quả phát hiện đối tượng từ mô hình ODwGCN cùng với mô hình dự đoán quan hệ giữa các đối tượng đã trình bày ở Mục 2.2.2.1 và 2.2.2.2, các vùng đối tượng trong ảnh được xác định và từ đó hình thành đồ thị quan hệ của ảnh. Sau bước này, một mạng CNN đã được huấn luyện trước được áp dụng để trích xuất đặc trưng của từng vùng đối tượng, ký hiệu là $F = \{f_i | \forall i = \overline{1, N_B}\}$. Tiếp theo, đồ thị quan hệ được ánh xạ thành các véc-tơ đặc trưng theo quy trình mô tả tại Mục 2.2.2.3, thu được tập các véc-tơ biểu diễn của các đỉnh $z_v^*, \forall v \in V^*$. Những véc-tơ đặc trưng này sau đó được kết hợp với các trọng số chú ý trong cơ chế chú ý kép, và toàn bộ thông tin được đưa vào mạng LSTM để sinh ra chú thích cho ảnh đầu vào.

Thuật toán 2.4 thực hiện học biểu diễn cho các đỉnh trong đồ thị R-Graph* bằng cách khai thác tính có hướng của cấu trúc quan hệ. Việc tách tập đỉnh kề thành hai nhóm $N-(v)$ (các đỉnh đi vào) và $N+(v)$ (các đỉnh đi ra) cho phép mô hình phân biệt vai trò ngữ nghĩa của một thực thể khi nó đóng vai trò chủ thể $(subject)$ hoặc đối tượng $(object)$ trong các quan hệ. Quá trình tổng hợp thông tin theo hai hướng

riêng biệt giúp embedding cuối cùng z_v^* không chỉ phản ánh sự xuất hiện của các quan hệ, mà còn bảo toàn hướng tương tác giữa các thực thể trong đồ thị. Việc nối hai biểu diễn $h_{v-}^{(N_L)}$ và $h_{v+}^{(N_L)}$ tạo thành một embedding giàu thông tin cấu trúc hơn so với cách tổng hợp đơn hướng, từ đó tăng khả năng khai thác ngữ nghĩa khi đưa vào cơ chế chú ý trong bộ giải mã.

Thuật toán 2.4. GenerateRGraphNodeEmbedding$\left(G^*, X, N_{L_3}, W\right)$

Đầu vào: Đồ thị $G^* = (V^*, E^*)$, véc-tơ khởi tạo $X = \{x_v, \forall v \in V^*\}$, số tầng N_{L3}, ma trận trọng số $W^{(l)}, \forall l = \overline{1, N_{L3}}$.

Đầu ra: node embedding véc-tơ $z_v^*, \forall v \in V^*$.

Begin

$h_{v-}^{(0)} \leftarrow x_v, \forall v \in V^*;$

$h_{v+}^{(0)} \leftarrow x_v, \forall v \in V^*;$

for $l = \overline{1..N_L}$ *do*

  for $v \in V^*$ *do*

    $h_{N-(v)}^{(l)} = f_{agg}\left(\left\{h_u^{(l-1)}, \forall u \in N-(v)\right\}\right);$

    $h_{v-}^{(l)} = \sigma\left(W^{(l)}.\left[h_v^{(l-1)}, h_{N-(v)}^{(l)}\right]\right);$

    $h_{N+(v)}^{(l)} = f_{agg}\left(\left\{h_u^{(l-1)}, \forall u \in N+(v)\right\}\right);$

    $h_{v+}^{(l)} = \sigma\left(W_1^{(l)}.\left[h_v^{(l-1)}, h_{N+(v)}^{(l)}\right]\right);$

  endfor

endfor

$z_v^* \leftarrow [h_{v-}^{(N_L)}, h_{v+}^{(N_L)}], \forall v \in V^*;$

End

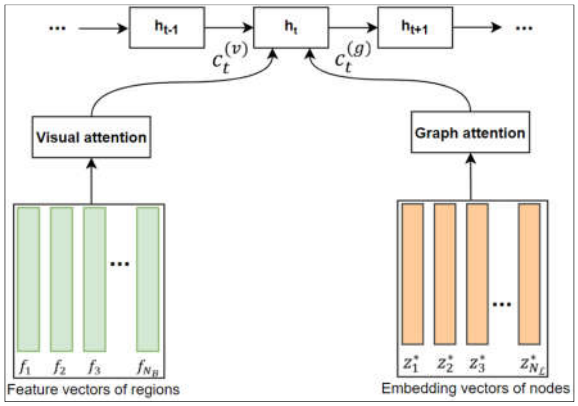
2.2.3. Bộ giải mã ngôn ngữ

Bộ giải mã ngôn ngữ được thiết kế với hai thành phần chính: (iv) Một cơ chế chú ý kép (dual attention), có nhiệm vụ xác định trọng số động cho các vùng thông tin quan trọng trong ảnh khi sinh ra từng từ trong câu mô tả; (v) Một mạng LSTM, được tích hợp với cơ chế chú ý kép này, để tạo ra chuỗi chú thích hoàn chỉnh cho hình ảnh.

2.2.3.1. Cơ chế chú ý kép

Trong bài toán sinh chú thích cho ảnh, cơ chế chú ý cho phép mô hình xác định trọng số thích nghi đối với các vùng quan trọng của ảnh tại từng bước giải mã. Các trọng số này được sử dụng để tổng hợp thành một véc-tơ ngữ cảnh (context véc-tơ), ký hiệu là c_t, đóng vai trò làm đầu vào cho bộ giải mã trong quá trình phát sinh mô tả. Trong nghiên cứu này, cơ chế chú ý kép (dual attention) được xây dựng với hai thành phần hoạt động độc lập: visual attention và graph attention (**Hình 2.6**). Visual attention tập trung vào việc tính toán mức độ liên kết giữa trạng thái ẩn của bộ giải

mã và đặc trưng các vùng đối tượng trong ảnh, trong khi graph attention khai thác đặc trưng của các đỉnh trong đồ thị quan hệ mở rộng. Cách kết hợp này cho phép mô hình tận dụng đồng thời cả thông tin thị giác lẫn ngữ nghĩa cấu trúc. Kết quả, hai véc-tơ ngữ cảnh riêng biệt là visual context véc-tơ $\left(c_t^{(v)}\right)$ và graph context véc-tơ $\left(c_t^{(g)}\right)$ được tạo ra theo **Thuật toán 2.5** và sau đó tích hợp vào bộ giải mã để sinh chú thích.



Hình 2.6. Cơ chế chú ý kép trong mô hình chú thích ảnh. Mạng LSTM tại mỗi bước thời gian kết hợp hai ngữ cảnh: visual attention trên các véc-tơ đặc trưng vùng ảnh và graph attention trên các véc-tơ embedding của các đỉnh trong đồ thị quan hệ.

Thuật toán 2.5. **CreateContextVector**(F, Z^*, h_{t-1})

Đầu vào: Tập đặc trưng vùng đối tượng F, tập véc-tơ embedding của đồ thị Z^*, trạng thái ẩn trước đó của decoder h_{t-1}

Đầu ra: Véc-tơ ngữ cảnh tại thời điểm t $\left(c_t^{(v)}, c_t^{(g)}\right)$

Begin

 # Calculate alignment score

$e_{ti}^{(v)} = f_{att}(f_i, h_{t-1}), \forall i = \overline{1, N_B}$; $e_{ti}^{(g)} = f_{att}(z_i, s_{t-1}), \forall i = \overline{1, N_L}$;

Calculate attention weight

$\rho_{ti}^{(v)} = \frac{exp\left(e_{ti}^{(v)}\right)}{\sum_{k=1}^{N_B} exp\left(e_{tk}^{(v)}\right)}$, $\sum_i \rho_{ti}^{(v)} = 1,\ 0 < \rho_{ti}^{(v)} < 1,\ \forall i = \overline{1, N_B}$;

$\rho_{ti}^{(g)} = \frac{exp\left(e_{ti}^{(g)}\right)}{\sum_{k=1}^{N_L} exp\left(e_{tk}^{(g)}\right)}$, $\sum_i \rho_{ti}^{(g)} = 1,\ 0 < \rho_{ti}^{(g)} < 1,\ \forall i = \overline{1, N_L}$;

 # Calculate context vector

$c_t^{(v)} = \sum_{i=1}^{N_B} f_i \rho_{ti}^{(v)}$; $c_t^{(g)} = \sum_{i=1}^{N_L} z_i \rho_{ti}^{(g)}$;

return $c_t^{(v)}, c_t^{(g)}$;

End

Trong **Thuật toán 2.5**, hệ số chú ý (*attention score*) trước hết được tính thông qua một phép biến đổi tuyến tính f_{att}, nhằm ước lượng mức độ liên quan giữa trạng thái ẩn của bộ giải mã tại thời điểm hiện tại và từng thành phần đặc trưng đầu vào. Quá trình này được thực hiện độc lập trên hai không gian đặc trưng: đặc trưng thị giác của

các vùng đối tượng và embedding của các đỉnh trong đồ thị quan hệ. Các hệ số chú ý sau đó được chuẩn hóa bằng hàm $softmax$ để thu được trọng số chú ý (*attention weight*). Tiếp theo, hai véc-tơ ngữ cảnh được tính bằng tổng có trọng số của các đặc trưng tương ứng, gồm véc-tơ ngữ cảnh thị giác $\left(c_t^{(v)}\right)$ và véc-tơ ngữ cảnh đồ thị $\left(c_t^{(g)}\right)$. Cơ chế chú ý kép này cho phép bộ giải mã đồng thời khai thác thông tin trực quan từ ảnh và thông tin cấu trúc ngữ nghĩa từ đồ thị quan hệ, qua đó tăng khả năng lựa chọn từ phù hợp và duy trì tính nhất quán ngữ nghĩa trong câu mô tả sinh ra.

2.2.3.2. Mạng LSTM phát sinh chú thích

Mạng với kiến trúc lưu trữ dài và ngắn hạn (LSTM) [95] được phát triển nhằm khắc phục hạn chế của mạng nơ-ron hồi quy (RNN) [96], đặc biệt là hiện tượng triệt tiêu đạo hàm khi xử lý các chuỗi có quan hệ phụ thuộc dài. Trong chương này, LSTM được sử dụng như một mô hình ngôn ngữ kết hợp với cơ chế chú ý kép để sinh mô tả cho hình ảnh. Tại mỗi bước t, LSTM tiếp nhận đầu vào gồm: nhúng từ (x_t), trạng thái ẩn trước đó (h_{t-1}), cùng với hai véc-tơ ngữ cảnh $c_t^{(v)}$, và $c_t^{(g)}$ sinh ra từ cơ chế chú ý kép. Dựa trên các thông tin này, các thành phần của LSTM bao gồm cổng vào (i_t), cổng quên (f_t), cổng đầu ra (o_t), cùng với trạng thái bộ nhớ (C_t) được cập nhật tại thời điểm t theo các công thức tính toán chuẩn của LSTM.

$$\begin{cases} i_t = \sigma\left(W_{x_i}x_t + W_{h_i}h_{t-1} + W_{i_c}\left[c_t^{(v)}; c_t^{(g)}\right] + b_i\right) \\ f_t = \sigma\left(W_{x_f}x_t + W_{h_f}h_{t-1} + W_{f_c}\left[c_t^{(v)}; c_t^{(g)}\right] + b_f\right) \\ o_t = \sigma\left(W_{x_o}x_t + W_{h_o}h_{t-1} + W_{o_c}\left[c_t^{(v)}; c_t^{(g)}\right] + b_o\right) \\ \widetilde{C_t} = tanh\left(W_{x_c}x_t + W_{h_c}h_{t-1} + W_{c_c}\left[c_t^{(v)}; c_t^{(g)}\right] + b_c\right) \\ C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \\ h_t = o_t \odot tanh\left(\widetilde{C_t}\right) \end{cases} \quad (2.7)$$

Trong các (2.7), ký hiệu δ biểu thị hàm sigmoid, còn W và b là các tham số cần học của mô hình. Ký hiệu $\odot$ thể hiện phép nhân theo từng phần tử (element-wise). Ở bước cuối, trạng thái ẩn h_t được sử dụng để suy ra từ cần dự đoán bằng cách sinh ra phân phối xác suất p_t trên từ hiện tại y_t thông qua hàm $softmax$, được mô tả bởi:

$$y_t \sim p_t = softmax\left(W_p h_t + b_p\right) \quad (2.8)$$

Trong (2.8), W_b và b_p là các trọng số được mô hình tối ưu trong quá trình huấn luyện. Toàn bộ mạng LSTM được huấn luyện bằng thuật toán lan truyền ngược theo thời gian (BackPropagation Through Time), với hàm mất mát lựa chọn là L_1 [97].

2.3. Thực nghiệm

Từ cơ sở lý thuyết và mô hình đã giới thiệu, phần này cài đặt thực nghiệm trên các tập dữ liệu và đánh giá tính hiệu quả của phương pháp đề xuất với các độ đo được sử dụng phổ biến trong bài toán phát hiện đối tượng, dự đoán mối quan hệ giữa các đối tượng và chú thích ảnh.

2.3.1. Dữ liệu và cấu hình thực nghiệm

Mục này trình bày chi tiết về tập dữ liệu được sử dụng trong quá trình thực nghiệm, các tham số huấn luyện cũng như thiết lập cấu hình của mô hình đề xuất. Ngoài ra, các độ đo dùng để đánh giá hiệu quả của mô hình cũng được giới thiệu trong phần này.

2.3.1.1. Dữ liệu thực nghiệm

Để bảo đảm tính nhất quán với khung đã nêu tại Mục 1.4.1, phần này trình bày cụ thể dữ liệu được sử dụng cho hai mục đích: (i) học quan hệ đối tượng để xây dựng đồ thị quan hệ, và (ii) phát hiện đối tượng cùng sinh chú thích trên MS COCO.

(i) Dữ liệu cho mô hình dự đoán quan hệ (Visual Genome). Tập ảnh Visual Genome [79] có 108.077 ảnh, mỗi ảnh có trung bình 35 đối tượng và 21 mối quan hệ cặp đôi giữa các đối tượng (triplet) được sử dụng cho mô hình dự đoán mối quan hệ. Tuy nhiên, nhiều mô tả trong tập dữ liệu này có chất lượng thấp và các vùng đối tượng trùng lắp, nhập nhằng tên các đối tượng. Điều này dẫn đến mô hình khó có thể học được các thông tin một cách hiệu quả. Do đó, bước tiền xử lý được áp dụng bằng cách lọc các mô tả chất lượng thấp và các vùng đối tượng trùng lắp theo phương pháp của Xu và cộng sự [98]. Sau khi tiền xử lý, 150 lớp đối tượng (*subjects*/*objects*) và 50 mối quan hệ (*predicates*) phổ biến nhất dùng cho mô hình dự đoán mối quan hệ giữa các đối tượng trong chương này. Tập dữ liệu mới gồm 95.998 ảnh được chia thành hai phần, 70% cho huấn luyện và 30% cho kiểm thử. Tập dữ liệu đã được tiền xử lý theo cách này cũng được sử dụng rộng rãi trong các công trình khác [98-100].

(ii) Trong các thực nghiệm liên quan đến phát hiện đối tượng (ODwGCN) và sinh mô tả ảnh, nghiên cứu này sử dụng bộ dữ liệu MS COCO [101]. Bộ dữ liệu bao gồm 82.783 ảnh dùng để huấn luyện và 40.504 ảnh cho giai đoạn kiểm định. Nhằm đảm bảo tính tương đồng và khả năng so sánh với các nghiên cứu trước, phép chia dữ liệu theo chuẩn Karpathy được áp dụng, với 82.783 ảnh huấn luyện, 5.000 ảnh dùng để kiểm định, và 5.000 ảnh kiểm tra. Mỗi ảnh trong tập dữ liệu được gán năm câu mô tả được tạo thủ công bởi con người. Ngoài ra, tập từ vựng gồm 10.010 từ cùng với giới hạn độ dài tối đa của chú thích là 16 từ được giữ nguyên như mô tả tại Mục 1.4.1 nhằm đảm bảo tính nhất quán trong toàn bộ quá trình thực nghiệm.

2.3.1.2. Chi tiết cài đặt

Các thực nghiệm của mô hình đề xuất được cài đặt bằng ngôn ngữ Python 3.6 trên nền tảng học sâu PyTorch 2.0 kết hợp với thư viện torch-geometric, và được triển khai trong môi trường Google Colab Pro. Cấu hình mô hình và các siêu tham số được thiết lập cụ thể như sau:

(1) **Thiết lập chung**: Hai mô hình Faster R-CNN_wGCN và ResNet101 được sử dụng cho nhiệm vụ phát hiện và trích xuất đặc trưng của các vùng đối tượng trong ảnh, với véc-tơ đặc trưng có kích thước 2.048 chiều. Các nhãn lớp, từ trong chú thích chuẩn, cũng như đối tượng và quan hệ trong đồ thị đều được biểu diễn bằng kỹ thuật GloVe với kích thước embedding 512, trong khi mạng GCN được cấu hình gồm hai tầng lan truyền.

(2) **Mô hình phát hiện đối tượng kết hợp GCN**: Thành phần này chỉ huấn luyện GCN để học biểu diễn (embedding) cho các đỉnh của đồ thị quan hệ. Kiến trúc GCN gồm hai lớp tích chập, với kích thước đầu ra lần lượt (1, c) → (4, c) → (1, c)**,** trong đó c là tổng số lớp đối tượng trong tập dữ liệu. Đối với MS COCO, giá trị c = 81, bao gồm 80 lớp đối tượng thực và một lớp nền (background)

(3) **Mô hình dự đoán quan hệ**: Mỗi đối tượng được ánh xạ thành véc-tơ đặc trưng có chiều 512, sau đó đi qua mạng fully connected được huấn luyện với hàm mất mát cross-entropy và bộ tối ưu Adam. Các tham số huấn luyện gồm: tốc độ học (learning rate) 0.0003, số vòng lặp tối đa 20.000, kích thước batch 64, và tỷ lệ dropout 0.5 nhằm tránh hiện tượng overfitting.

(4) **Mô hình sinh chú thích ảnh**: Thành phần giải mã ngôn ngữ sử dụng mạng LSTM với kích thước trạng thái ẩn 1.024, và độ dài tối đa của câu sinh ra là 16 từ. Quá trình huấn luyện cũng sử dụng hàm mất mát cross-entropy cùng với bộ tối ưu Adam**,** learning rate 0.0001, và batch-size 32.

2.3.2. Độ đo đánh giá

Trong các thực nghiệm ở này, hiệu quả của mô hình được đánh giá theo ba nhóm nhiệm vụ khác nhau:

(i) **Phát hiện đối tượng**: Hiệu suất được đo bằng độ chính xác trung bình mAP (Mean Average Precision) và $mAP@0.5$, $mAP@0.75$.

(ii) **Dự đoán mối quan hệ**: Hiệu suất được đánh giá bằng Recall@K với *K=50,100*. Chỉ số này cho biết tỷ lệ bộ ba quan hệ đúng ($subject - predicate - object$) xuất hiện trong K dự đoán hàng đầu của mô hình cho mỗi ảnh so với tập bộ ba tham chiếu.

(iii) **Phát sinh chú thích**: Đối với nhiệm vụ chú thích ảnh, các độ đo chuẩn đã được giới thiệu chi tiết ở Mục 1.4.2 được sử dụng, bao gồm BLEU, METEOR, ROUGE-L, CIDEr và SPICE. Các độ đo này phản ánh mức độ tương đồng giữa chú thích sinh ra bởi mô hình và chú thích chuẩn dưới nhiều góc độ khác nhau (n-gram, ngữ nghĩa, chuỗi con chung dài nhất, đồng thuận thống kê và cấu trúc ngữ nghĩa).

Tất cả các độ đo đều được tính toán ở mức corpus và giá trị cao hơn biểu thị hiệu suất tốt hơn. Nhờ sử dụng kết hợp nhiều độ đo, chất lượng chú thích được đánh giá toàn diện cả ở mức độ trùng khớp bề mặt lẫn mức độ cấu trúc ngữ nghĩa.

2.3.3. Chi phí tính toán và thời gian thực hiện

Ngoài hiệu quả về độ chính xác, phần này phân tích chi phí tính toán của mô hình OD-VR-Cap nhằm làm rõ khả năng triển khai thực tế của pipeline chú thích ảnh gồm nhiều mô-đun xử lý liên tiếp trong điều kiện phần cứng phổ biến. Toàn bộ quá trình huấn luyện được thực hiện trên GPU NVIDIA Tesla T4 (16 GB VRAM) và được chia thành ba giai đoạn tương ứng với các thành phần cấu thành của mô hình. Cụ thể, mô hình phát hiện đối tượng kết hợp mạng tích chập đồ thị (ODwGCN) được huấn luyện trong khoảng 8 giờ cho 20 epoch; tiếp theo, mô hình dự đoán quan hệ giữa các đối tượng (VRP+RK) được tối ưu trong 6 giờ cho 15 epoch; và cuối cùng, thành phần sinh chú thích ảnh OD-VR-Cap sử dụng bộ giải mã LSTM với cơ chế chú ý kép được huấn luyện trong 14 giờ cho 25 epoch trên tập dữ liệu MS COCO.

Ở pha suy luận, thời gian sinh chú thích trung bình đạt khoảng 0.10 giây/ảnh với batch size = 32. Kết quả này cho thấy mặc dù mô hình bao gồm nhiều mô-đun xử lý độc lập, chi phí tính toán vẫn được kiểm soát ở mức phù hợp, đáp ứng yêu cầu triển khai trong các hệ thống thực tế mà không đòi hỏi hạ tầng phần cứng chuyên biệt.

2.3.4. Kết quả thực nghiệm

Các kết quả thực nghiệm, bàn luận và so sánh với các công trình baseline hoặc công trình công bố gần đây được trình bày trong phần này nhằm làm rõ ưu khuyết điểm của phương pháp đề xuất. Cụ thể gồm: kết quả thực nghiệm phát hiện đối tượng, thực nghiệm dự đoán mối quan hệ giữa các đối tượng và thực nghiệm chú thích ảnh.

2.3.4.1. Kết quả thực nghiệm phát hiện đối tượng

Kết quả đánh giá hiệu suất của các mô hình phát hiện đối tượng được huấn luyện trước - bao gồm SSD, Faster R-CNN và YOLOX - cùng với mô hình ODwGCN được trình bày trong **Bảng 2.1**, thông qua các chỉ số mAP, mAP@0.5 và mAP@0.75. Trong các tên mô hình, hậu tố **GCN** biểu thị việc áp dụng mạng tích chập đồ thị để điều chỉnh lại kết quả phát hiện đối tượng dựa trên mối quan hệ đồng xuất hiện giữa các nhãn lớp. Kết quả thực nghiệm cho thấy, việc bổ sung GCN giúp nâng cao độ

chính xác trung bình (mAP) cho toàn bộ các mô hình cơ sở. Cụ thể, mức cải thiện dao động từ 0.9 đến 3.2 phần trăm, tùy thuộc vào kiến trúc backbone. Trong số đó, SSD sử dụng ResNet101-FPN ghi nhận mức cải thiện lớn nhất, với mAP tăng 2.2 và mAP@0.5 tăng 3.2, trong khi YOLOX kết hợp DarkNet-53 có mức tăng thấp nhất, lần lượt 1.2 (mAP), 0.9 (mAP@0.5) và 1.3 (mAP@0.75).

Mặc dù mức cải thiện của YOLOX nhỏ hơn, nhưng do nền tảng ban đầu đã có độ chính xác cao, mô hình YOLOX+GCN vẫn đạt kết quả tốt nhất trong toàn bộ thử nghiệm. Điều này khẳng định tính hiệu quả của việc bổ sung thông tin ngữ cảnh giữa các đối tượng thông qua GCN, giúp mô hình phát hiện ổn định hơn trong các cảnh phức tạp. Đáng chú ý, mô hình ODwGCN có thể được tích hợp linh hoạt với bất kỳ kiến trúc phát hiện đối tượng hiện có nào nhằm tăng cường khả năng nhận diện và giảm lỗi phân loại sai trong các hệ thống thị giác máy tính.

Bảng 2.1. So sánh kết quả phát hiện đối tượng giữa các mô hình huấn luyện trước và ODwGCN trên tập dữ liệu MS COCO

Mô hình	Bộ trích xuất đặc trưng	Avg. Precision (mAP)	mAP@IoU=0.5	mAP@IoU=0.75
SSD	VGG16	28.8	48.5	30.3
SSD+**GCN**	VGG16	30.9	49.4	32.8
SSD	ResNet-101-FPN	31.2	50.4	33.3
SSD+**GCN**	ResNet-101-FPN	33.4	53.6	35.4
Faster R-CNN	ResNet-101-FPN	36.2	59.1	39.0
Faster R-CNN+**GCN**	ResNet-101-FPN	37.5	60.4	41.8
Faster R-CNN	Inception_ResNet_v2	34.7	55.1	36.7
Faster R-CNN+**GCN**	Inception_ResNet_v2	36.2	56.9	38.6
YOLOX	DarkNet-53	47.4	67.3	52.1
YOLOX+**GCN**	DarkNet-53	**48.2**	**68.2**	**53.4**

2.3.4.2. Kết quả thực nghiệm dự đoán mối quan hệ giữa các đối tượng

Bảng 2.2. So sánh độ chính xác của các phương pháp dự đoán mối quan hệ trên tập dữ liệu thực nghiệm

Phương pháp	Recall@50 (%)	Recall@100 (%)
VRD [99]	27.9	35.0
Message Passing [98]	44.8	53.0
MSTG [100]	52.5	58.4
VRP$^+$RK	**54.7**	**60.3**

Kết quả dự đoán mối quan hệ giữa các đối tượng của phương pháp đề xuất có giá trị lần lượt là *Recall@50* = 54.7 và *Recall@100* = 60.3. So sánh với kết quả của các phương pháp khác được mô tả trong **Bảng 2.2** cho thấy phương pháp đề xuất có độ

chính xác cao hơn. Cụ thể, so với phương pháp baseline (VRD) là cao hơn 26.8% cho độ đo Recall@50 và 25.3% cho Recall@100. Hai phương pháp còn lại cao hơn từ 2.2 đến 2.9 với độ đo Recall@50 và từ 1.9 đến 7.3 cho độ đo Recall@100. Từ đó cho thấy thông tin ngữ cảnh của mối quan hệ, và đặc biệt là tri thức quan hệ vốn có giữa các đối tượng có đóng góp quan trọng trong việc phát hiện mối quan hệ giữa các đối tượng trong ảnh.

2.3.4.3. Kết quả thực nghiệm chú thích ảnh

Để đánh giá hiệu quả của mô hình OD-VR-Cap, luận án tiến hành thực nghiệm trên tập dữ liệu chuẩn MS COCO Karpathy và so sánh với nhiều phương pháp chú thích ảnh đã công bố. Các độ đo được sử dụng gồm BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr và SPICE, trong đó các độ đo BLEU và METEOR phản ánh mức độ trùng khớp n-gram, ROUGE-L đánh giá tính mạch lạc, CIDEr đo lường sự đồng thuận với chú thích tham chiếu, còn SPICE tập trung vào tính đúng đắn ngữ nghĩa của các chú thích.

Bảng 2.3. So sánh độ chính xác chú thích ảnh của các phương pháp trên tập kiểm tra MSCOCO Karpathy. Giá trị in đậm biểu thị điểm số cao nhất trong mỗi cột độ đo.

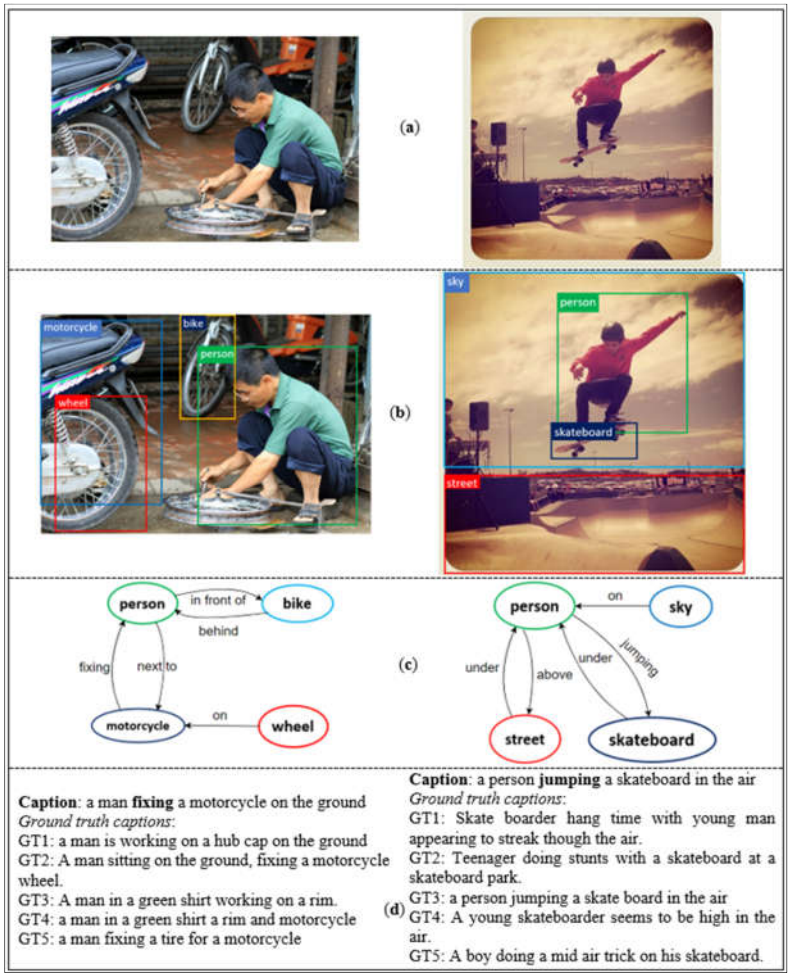
Phương pháp	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Show, attend and tell (Hard-ATT) [18]	71.8	25.0	23.0	-	-	-
Show, attend and tell (Soft-ATT) [18]	70.7	24.3	23.9	-	-	-
Dense_Soft-ATT [42]	68.3	22.9	22.6	53.0	74.3	-
En-De-Cap [43]	70.6	24.3	-	-	-	-
Bi-LS-AttM [46]	68.8	25.2	21.5	-	41.2	-
Image+SceneGraph [51]	67.2	26.1	22.3	-	76.0	-
G-LSTM+att [54]	67.7	20.7	23.9	-	-	-
OD-VR-Cap	**72.6**	**28.3**	**24.8**	**53.4**	**85.1**	17.6

Kết quả trong **Bảng 2.3** cho thấy mô hình OD-VR-Cap đạt hiệu suất vượt trội trên hầu hết các độ đo so với các phương pháp đối sánh. Xét theo nhóm độ đo n-gram, OD-VR-Cap đạt BLEU-1 = 72.6 và BLEU-4 = 28.3, lần lượt cao hơn 0.8 và 3.3 điểm so với Hard-ATT [18], đồng thời cải thiện 4.0 điểm BLEU-4 so với Soft-ATT. Mức cải thiện đáng kể ở BLEU-4 cho thấy mô hình không chỉ dự đoán chính xác các từ đơn lẻ mà còn sinh được các cụm từ dài có tính nhất quán cao hơn.

Đối với METEOR và ROUGE-L, OD-VR-Cap đạt 24.8 và 53.4, cao hơn các phương pháp Dense_Soft-ATT [42] và Image+SceneGraph [51]. Điều này phản ánh

khả năng bao phủ nội dung và duy trì tính mạch lạc trong chú thích được cải thiện nhờ khai thác quan hệ giữa các đối tượng thay vì chỉ dựa vào đặc trưng vùng độc lập.

Đáng chú ý nhất là chỉ số CIDEr đạt 85.1, cao hơn 9.1 điểm so với Image+SceneGraph [51]. Vì CIDEr đánh giá mức độ tương đồng với nhiều chú thích tham chiếu của con người, kết quả này cho thấy việc tích hợp đồ thị quan hệ và cơ chế tinh chỉnh phát hiện đối tượng đã giúp mô hình sinh chú thích gần với diễn đạt tự nhiên hơn. Chỉ số SPICE đạt 17.6, phản ánh sự cải thiện về tính đúng đắn ngữ nghĩa ở mức cấu trúc cảnh.



Hình 2.7. Ví dụ từ tập ảnh kiểm tra minh họa hiệu quả của phương pháp đề xuất. (a) và (b): Ảnh đầu vào cùng với các đối tượng được phát hiện (chỉ hiển thị 4 đối tượng có xác suất cao nhất); (c): Đồ thị quan hệ được trích xuất từ ảnh (chỉ hiển thị 5 quan hệ có xác suất cao nhất);(d): Chú thích được sinh tự động và các chú thích chuẩn tương ứng.

Bên cạnh kết quả định lượng, Hình 2.7 minh họa trực quan hiệu quả của mô hình trên hai ảnh trong tập kiểm tra. Mô hình phát hiện chính xác các đối tượng chính và xây dựng đồ thị quan hệ ngữ nghĩa giàu thông tin, từ đó sinh ra các chú thích gần gũi với chú thích chuẩn. Ví dụ, trong ảnh thứ nhất, mô hình nắm bắt đúng ngữ cảnh *"a man fixing a motorcycle on the ground"*, thể hiện sự liên kết chặt chẽ giữa đối tượng

(*person*, *motorcycle*, *wheel*) và hành động. Tương tự, trong ảnh thứ hai, mô hình mô tả chính xác hành động *"a person jumping a skateboard in the air"*, phù hợp với các chú thích chuẩn.

Những minh chứng định lượng và định tính này cho thấy OD-VR-Cap không chỉ khắc phục được hạn chế của các mô hình CNN-LSTM, mà còn vượt qua các phương pháp dựa trên đồ thị đã công bố, khẳng định tính hiệu quả của việc mô hình hóa quan hệ giữa các đối tượng trong việc nâng cao độ chính xác và tính giàu ngữ nghĩa của chú thích ảnh.

Sự cải thiện hiệu suất của OD-VR-Cap đến từ việc chuyển hướng biểu diễn hình ảnh từ dạng phẳng sang cấu trúc đồ thị, giúp mô hình không chỉ nhận diện từng đối tượng riêng lẻ mà còn học được cách chúng tương tác với nhau trong ngữ cảnh tổng thể. Việc kết hợp mạng GCN vào quy trình phát hiện và mã hóa quan hệ cho phép lan truyền thông tin giữa các đối tượng liên quan, giúp bộ giải mã sinh ra chú thích có cấu trúc logic và sát với nội dung thực tế hơn. Đây chính là đóng góp học thuật chính của mô hình - đưa tri thức quan hệ vào khung chú thích ảnh, tạo nền tảng cho các mô hình kế tiếp mở rộng sang biểu diễn ngữ nghĩa sâu hơn.

2.4. Kết chương

Trong chương này, mô hình OD-VR-Cap đã được đề xuất nhằm khắc phục những hạn chế của các phương pháp chú thích ảnh truyền thống - vốn chủ yếu dựa trên đặc trưng toàn cục hoặc đặc trưng riêng lẻ của từng vùng đối tượng mà chưa khai thác đầy đủ thông tin về mối quan hệ giữa các đối tượng trong ảnh. Mô hình được thiết kế theo hướng tích hợp ba thành phần chính: phát hiện đối tượng bằng ODwGCN, xây dựng và nhúng (embedding) đồ thị quan hệ, và giải mã sinh chú thích thông qua mạng LSTM kết hợp cơ chế chú ý kép. Đây là một hướng tiếp cận có cấu trúc rõ ràng, cho phép mô hình học được cả thông tin trực quan và các quan hệ ngữ nghĩa giữa các thành phần trong ảnh.

Thực nghiệm trên tập dữ liệu MS COCO cho thấy mô hình OD-VR-Cap đạt hiệu quả cao hơn so với các phương pháp gần đây theo nhiều độ đo đánh giá khác nhau. Đặc biệt, các thành phần như bộ phát hiện đối tượng ODwGCN và bộ dự đoán quan hệ cũng chứng minh được hiệu quả riêng biệt. Mô hình này không chỉ thể hiện ưu thế về mặt định lượng, mà còn mở ra một hướng tiếp cận khả thi trong việc mô hình hóa ngữ nghĩa hình ảnh thông qua cấu trúc đồ thị. Trong chương tiếp theo, mô hình RGTranCNet được giới thiệu như một sự kế thừa và mở rộng mô hình hiện tại, với việc tích hợp kiến trúc giải mã dựa trên Transformer cùng tri thức ngữ nghĩa từ ConceptNet nhằm tiếp tục nâng cao khả năng hiểu và sinh chú thích giàu ý nghĩa.

CHƯƠNG 3. CHÚ THÍCH ẢNH SỬ DỤNG TRANSFORMER VÀ TRI THỨC TỪ CONCEPTNET

Bài toán chú thích ảnh đòi hỏi mô hình vừa hiểu nội dung thị giác vừa biểu đạt ngôn ngữ một cách tự nhiên và giàu ngữ nghĩa. Chương 2 đã trình bày mô hình OD-VR-Cap, khai thác quan hệ giữa các đối tượng thông qua đồ thị quan hệ và cơ chế chú ý kép. Tuy nhiên, kiến trúc dựa trên LSTM vẫn còn hạn chế trong việc mô hình hóa phụ thuộc dài hạn giữa các từ và chưa tận dụng được tri thức ngữ nghĩa ngoài ảnh để hỗ trợ quá trình sinh chú thích. Xuất phát từ những hạn chế này, Chương này giới thiệu mô hình RGTranCNet như một bước phát triển tiếp theo trong chuỗi nghiên cứu của luận án. Mô hình được xây dựng trên nền tảng Transformer Decoder, giúp tăng cường khả năng học các quan hệ ngữ nghĩa dài hạn, đồng thời tích hợp tri thức ngữ nghĩa từ ConceptNet nhằm cải thiện tính chính xác và mạch lạc của chú thích được sinh ra.

Đóng góp nổi bật của RGTranCNet nằm ở cơ chế tích hợp đa nguồn thông tin. Cụ thể, cross-attention được sử dụng để kết hợp đặc trưng thị giác và đặc trưng quan hệ từ đồ thị quan hệ, qua đó giúp mô hình nắm bắt cấu trúc cảnh một cách đầy đủ và có hệ thống hơn. Song song với đó, tri thức ngữ nghĩa từ ConceptNet được đưa trực tiếp vào bộ giải mã, bổ sung các liên kết ngữ nghĩa không có trong tập huấn luyện và tăng cường khả năng xử lý các đối tượng hiếm hoặc mới. Nhờ vậy, RGTranCNet mở rộng đáng kể phạm vi hiểu ngữ nghĩa của mô hình so với OD-VR-Cap và tạo nền tảng vững chắc cho các phương pháp nâng cao được trình bày ở Chương 4. Các nội dung chính của chương đã được công bố trong [CT4] và một phần trong [CT3].

Nội dung của Chương được cấu trúc gồm bốn phần chính: (3.1) giới thiệu, (3.2) mô hình đề xuất, (3.3) thực nghiệm và kết quả, và (3.4) kết luận.

3.1. Giới thiệu

Trên cơ sở các nghiên cứu đã phân tích ở Chương 2, việc khai thác mối quan hệ giữa các đối tượng thông qua biểu diễn đồ thị cho thấy tiềm năng rõ rệt trong việc nâng cao khả năng hiểu ngữ cảnh và cấu trúc nội dung ảnh. Tuy nhiên, trong nhiều mô hình hiện có, quá trình giải mã ngôn ngữ vẫn chủ yếu dựa trên mạng LSTM kết hợp các cơ chế chú ý truyền thống, dẫn đến những hạn chế về khả năng mô hình hóa phụ thuộc dài hạn và hiệu quả huấn luyện do tính toán tuần tự.

Sự ra đời của kiến trúc Transformer [56] trong lĩnh vực xử lý ngôn ngữ tự nhiên đã mở ra hướng phát triển mới cho bài toán chú thích ảnh. Transformer cho phép huấn luyện song song, giảm đáng kể thời gian tính toán và đạt hiệu suất cao hơn nhờ khả năng học phụ thuộc dài hạn thông qua các cơ chế tự chú ý (self-attention) và

chú ý chéo (cross-attention). Nhờ vậy, trong những năm gần đây, các mô hình chú thích ảnh đã dần thay thế LSTM bằng Transformer trong vai trò bộ giải mã, mang lại khả năng nắm bắt ngữ cảnh toàn cục và liên kết ngữ nghĩa giữa các đối tượng hiệu quả hơn so với các cơ chế chú ý truyền thống.

Bên cạnh đó, các phương pháp chú thích ảnh công bố gần đây đều được huấn luyện trên tập dữ liệu gồm các cặp ảnh-câu chú thích (*paired image-captions*). Tuy nhiên, các tập dữ liệu này thường chỉ cung cấp một số lượng rất hạn chế chú thích cho mỗi ảnh (thường từ 1 đến 5 câu). Vì vậy, các mô hình này thiếu thông tin cần thiết để mô tả các đối tượng mới không xuất hiện trong tập huấn luyện, hoặc các khía cạnh không được thể hiện rõ ràng trong ảnh. Vấn đề này có thể được giải quyết bằng cách tích hợp thông tin từ các nguồn dữ liệu bên ngoài vào quá trình phát sinh chú thích. Nhiều công trình đã khai thác thông tin từ các nguồn dữ liệu bên ngoài tập dữ liệu huấn luyện, chẳng hạn như: khai thác tri thức đối tượng từ tập dữ liệu nhận dạng đối tượng và văn bản bên ngoài tập dữ liệu để tạo chú thích cho các đối tượng mới [102], sử dụng đồ thị tri thức để nâng cao hiệu quả cho chú thích ảnh [62], biểu diễn ngữ nghĩa (sử dụng cơ sở tri thức ConceptNet) và mạng chú ý cho chú thích ảnh [21], … Việc sử dụng tri thức bên ngoài tập dữ liệu huấn luyện, cụ thể là cơ sở tri thức ConceptNet để nâng cao hiệu quả cho mô hình chú thích ảnh là cần thiết, khả thi và hiệu quả.

Do đó, nghiên cứu này đề xuất một phương pháp mới có tên **RGTranCNet**, được cấu thành từ ba thành phần chính: **RG** (Relationship Graph) biểu thị đồ thị quan hệ, **Tran** đại diện cho bộ giải mã Transformer, và **CNet** thể hiện sự tích hợp tri thức ngữ nghĩa từ ConceptNet. Trong mô hình này, (i) đặc trưng vùng đối tượng và đồ thị quan hệ của ảnh được hợp nhất trong một khối chú ý chéo duy nhất thay vì sử dụng hai cơ chế chú ý độc lập, từ đó nâng cao hiệu quả chú ý; (ii) tri thức ngoài từ ConceptNet được tích hợp một cách liền mạch vào quá trình giải mã để tinh chỉnh câu chú thích được sinh ra, giúp xử lý hiệu quả hơn các đối tượng không xuất hiện trong tập huấn luyện - một hạn chế phổ biến trong các mô hình chú thích ảnh hiện nay - qua đó cải thiện khả năng tổng quát hóa và làm phong phú nội dung ngữ nghĩa của chú thích; và (iii) chỉ thành phần decoder được huấn luyện, trong khi các mô-đun chịu trách nhiệm trích xuất đặc trưng vùng đối tượng, xây dựng và biểu diễn đồ thị quan hệ, cũng như truy xuất tri thức ngữ nghĩa từ ConceptNet được giữ nguyên không thay đổi. Chiến lược này giúp giảm đáng kể chi phí huấn luyện mà vẫn duy trì độ chính xác và khả năng mở rộng của mô hình.

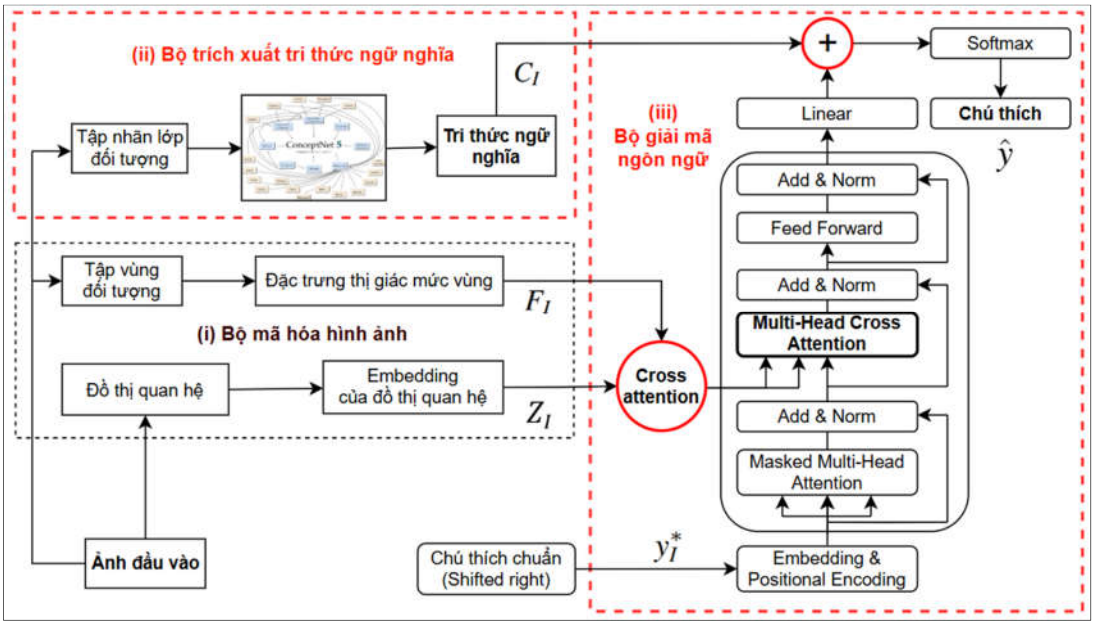
Đóng góp chính của chương này gồm:

- Cải thiện hiệu quả chú thích ảnh bằng cách dùng Transformer decoder làm mô hình ngôn ngữ thay cho mạng LSTM và cơ chế cross-attention thay cho các cơ chế chú ý truyền thống để kết hợp thông tin đa phương thức giữa encoder và decoder.
- Tích hợp tri thức ngữ nghĩa từ cơ sở tri thức ConceptNet vào decoder để khai thác tri thức bên ngoài tập dữ liệu huấn luyện nhằm nâng cao độ chính xác cho chú thích phát sinh, đặc biệt là các đối tượng mới. Phương pháp này có thể dễ dàng áp dụng vào các mô hình chú thích ảnh khác.
- Thực nghiệm trên bộ dữ liệu chuẩn MS COCO cho thấy mô hình dùng transformer decoder với cross-attention có độ chính xác cao hơn LSTM với các cơ chế chú ý truyền thống, và việc bổ sung tri thức ngữ nghĩa của các đối tượng trích xuất từ ConceptNet vào decoder cũng góp phần nâng cao độ chính xác cho chú thích được phát sinh.

3.2. Phương pháp chú thích ảnh đề xuất

Phần này trình bày chi tiết phương pháp chú thích ảnh đề xuất RGTranCNet, bao gồm kiến trúc tổng thể của mô hình và các thành phần chính cấu thành hệ thống, từ bộ mã hóa hình ảnh, bộ trích xuất tri thức ngữ nghĩa đến bộ giải mã ngôn ngữ dựa trên Transformer.

3.2.1. Kiến trúc tổng thể của mô hình RGTranCNet



Hình 3.1. Kiến trúc tổng thể của mô hình RGTranCNet. Bộ mã hóa ảnh tạo đặc trưng vùng đối tượng F_I và embedding đồ thị quan hệ Z_I. Bộ trích xuất tri thức truy xuất ConceptNet để thu thập tri thức C_I. Bộ giải mã Transformer tích hợp F_I và Z_I thông qua multi-head cross-attention và sử dụng C_I để tăng cường dự đoán từ khi sinh chú thích.

Mô hình RGTranCNet được xây dựng theo khung encoder-decoder như minh họa trong Hình 3.1, gồm ba khối chức năng chính: (i) Bộ mã hóa hình ảnh (Image Encoder) trích xuất biểu diễn thị giác và cấu trúc quan hệ của ảnh; (ii) Bộ trích xuất tri thức ngữ nghĩa (Semantic Knowledge Extractor) truy xuất tri thức ngữ nghĩa liên quan từ ConceptNet dựa trên nhãn lớp đối tượng; và (iii) Bộ giải mã ngôn ngữ (Language Decoder) dựa trên Transformer để phát sinh chú thích.

Trong ba khối chức năng trên, Image Encoder kế thừa trực tiếp thiết kế đã trình bày ở Chương 2 nhằm đảm bảo tính nhất quán về quy trình trích xuất đặc trưng vùng và xây dựng đồ thị quan hệ. Hai biểu diễn đầu ra của encoder gồm: ma trận đặc trưng vùng đối tượng F_I và ma trận embedding đồ thị quan hệ Z_I. Trên nền tảng đó, chương này tập trung vào hai cơ chế mở rộng tại phía decoder.

Thứ nhất, mô hình sử dụng Transformer Decoder thay cho LSTM và khai thác cơ chế multi-head cross-attention để hợp nhất thông tin thị giác F_I và thông tin quan hệ Z_I theo ngữ cảnh ngôn ngữ ở từng bước sinh từ.

Thứ hai, tri thức ngoài từ ConceptNet được đưa vào quá trình giải mã thông qua bộ trích xuất tri thức và cơ chế điều chỉnh phân phối dự đoán từ; cụ thể, với mỗi ảnh, tập tri thức C_I được truy xuất theo các nhãn đối tượng và được sử dụng để tăng cường xác suất cho các khái niệm liên quan khi tính phân phối dự đoán.

Trong mô hình RGTranCNet, các mô-đun trích xuất đặc trưng vùng đối tượng, xây dựng và biểu diễn đồ thị quan hệ, cũng như truy xuất tri thức ngữ nghĩa từ ConceptNet được xem là các bước tiền xử lý hoặc đã được huấn luyện trước và được giữ cố định trong suốt quá trình huấn luyện. Chỉ các tham số của bộ giải mã Transformer, ký hiệu là φ, được tối ưu hóa nhằm học cách kết hợp thông tin thị giác-quan hệ và tri thức ngữ nghĩa trong quá trình sinh chú thích.

3.2.2. Bộ mã hóa hình ảnh

Quá trình mã hóa hình ảnh trong mô hình RGTranCNet kế thừa trực tiếp kiến trúc bộ mã hóa đã được trình bày chi tiết trong Mục 2.2.2 của Chương 2. Cụ thể, mô hình tiếp tục sử dụng hai thành phần chính: (i) bộ phát hiện đối tượng tích hợp GCN (ODwGCN) nhằm cải thiện độ chính xác phát hiện thông qua việc bổ sung mối quan hệ đồng xuất hiện giữa các đối tượng trong ảnh, và (ii) khối xây dựng đồ thị quan hệ kết hợp mô hình dự đoán quan hệ (VRP+RK), từ đó xây dựng đồ thị R-Graph có cấu trúc rõ ràng, phản ánh mối quan hệ giữa các đối tượng trong ảnh.

Sau khi thu được đồ thị quan hệ, đặc trưng ngữ nghĩa của từng đỉnh được làm giàu bằng mạng nơ-ron tích chập trên đồ thị hai tầng. Với một ảnh đầu vào I, kết quả của bộ mã hóa hình ảnh là hai ma trận đầu ra F_I, Z_I. Ma trận F_I lưu trữ đặc trưng trực

tiếp từ các vùng ảnh, trong khi Z_I chứa các véc-tơ đặc trưng sau khi lan truyền ngữ nghĩa trên đồ thị quan hệ. Hai ma trận này được sử dụng làm đầu vào cho tầng chú ý tại bộ giải mã Transformer.

Bộ mã hóa hình ảnh trong RGTranCNet được giữ nguyên theo thiết kế đã trình bày tại Mục 2.2.2; do đó luận án không lặp lại các công thức và thuật toán liên quan.

3.2.3. Bộ trích xuất tri thức ngữ nghĩa đối tượng

ConceptNet là một cơ sở tri thức đa ngôn ngữ, mô tả các từ, cụm từ mà con người thường sử dụng và các mối quan hệ thông thường giữa chúng. Tri thức trong ConceptNet được thu thập từ nhiều nguồn khác nhau, gồm các nguồn tài nguyên được đóng góp từ cộng đồng như Wiktionary và Open Mind Common Sense, các ngồn tài nguyên do chuyên gia tạo ra như WordNet và JMDict và nhiều nguồn dữ liệu mở khác [103]. Từ đó tạo ra một cơ sở tri thức liên kết các khái niệm với nhau thông qua các mối quan hệ ngữ nghĩa như như "IsA" (là một), "$PartOf$" (là một phần của), "$UsedFor$" (dùng để), và nhiều mối quan hệ khác. Các mối quan hệ này cho phép mô hình hiểu rõ hơn về bối cảnh và liên kết ngữ nghĩa giữa các từ, cụm từ. Từ đó, hỗ trợ các hệ thống trí tuệ nhân tạo trong việc hiểu ngữ cảnh, suy luận và tương tác với con người. Trong nghiên cứu này, cơ sở tri thức ConceptNet K được định nghĩa dưới dạng đồ thị như sau:

Định nghĩa 3.1. Đồ thị **CK-Graph** $K = (V, E, W)$ là đồ thị tri thức bao gồm:

– Tập đỉnh $V = \{v_i \in C, \forall i = \overline{1, N_C}\}$, trong đó C là tập khái niệm trong ConceptNet,

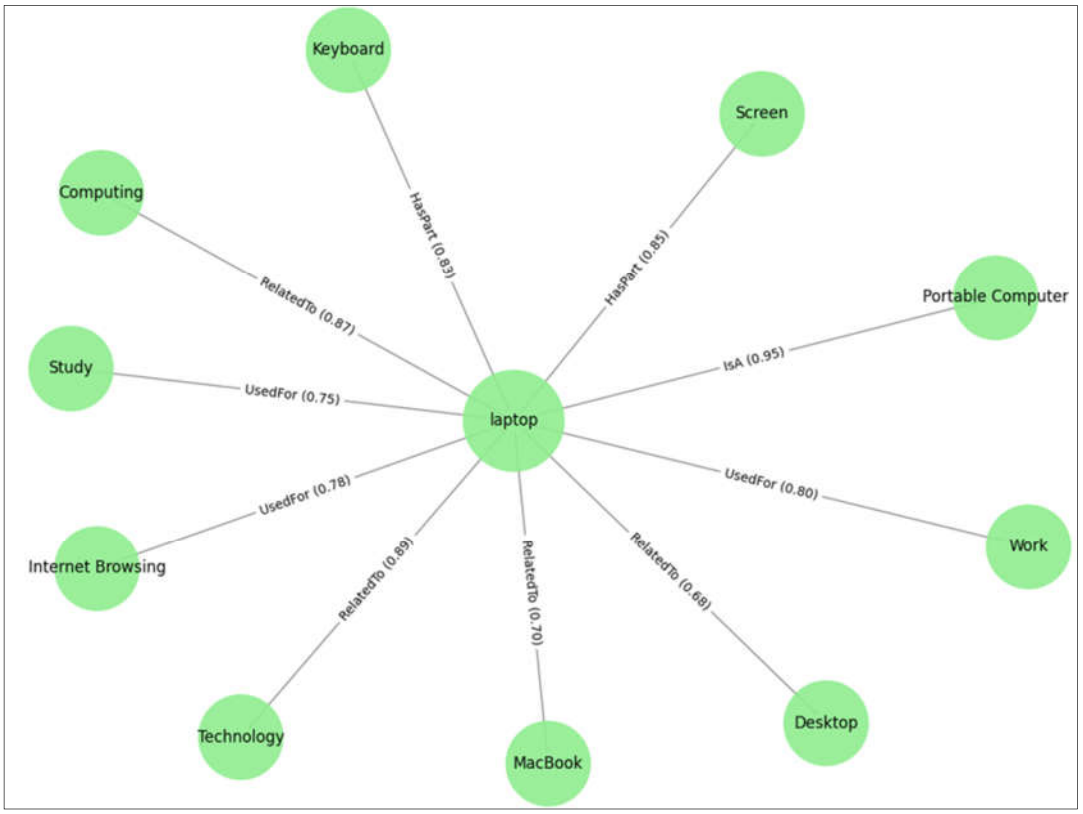
– Tập cạnh $E = \{e_{ij} = (v_i, v_j, w), \forall i, j = \overline{1, N_C}, i \neq j\}$, biểu diễn mối liên hệ ngữ nghĩa giữa các cặp khái niệm;

– Tập trọng số $W = \{w_i \in R^+, \forall i = \overline{1, N_E}\}$, trong đó N_E là số lượng cạnh trong tập E, và mỗi trọng số phản ánh mức độ liên quan ngữ nghĩa giữa hai khái niệm tương ứng.

Để tích hợp tri thức ngữ nghĩa từ ConceptNet vào quá trình sinh chú thích nhằm cải thiện độ chính xác, đặc biệt trong việc mô tả các đối tượng mới chưa xuất hiện trong tập dữ liệu huấn luyện, ConceptNet được biểu diễn dưới dạng một đồ thị tri thức $K = (V, E, W)$, như được định nghĩa trong Định nghĩa 3.1. Sau đó, các nhãn lớp đối tượng trong ảnh được sử dụng để truy vấn các tri thức có ý nghĩa tương tự từ đồ thị này. Hình 3.2 minh họa kết quả truy vấn đối với đối tượng "*laptop*" từ đồ thị K. Mỗi khái niệm liên quan được gán một giá trị xác suất, phản ánh mức độ tương đồng hoặc liên quan ngữ nghĩa giữa khái niệm đó và đối tượng được truy vấn.

Trong nghiên cứu này, *top*-k đối tượng liên quan đến mỗi đối tượng phát hiện trong ảnh được sử dụng để tăng cường thông tin cho decoder trong quá trình phát sinh chú thích ảnh. Tập này, ký hiệu là C, được sử dụng khi phát sinh từ tiếp theo của bộ giải mã. Quá trình trích xuất tri thức của các đối tượng liên quan được mô tả trong **Thuật toán 3.1**.

Thuật toán 3.1 thực hiện việc trích xuất các đối tượng liên quan từ cơ sở tri thức ConceptNet dựa vào nhãn lớp của các đối tượng trong ảnh. Thuật toán bắt đầu bằng việc khởi tạo một tập rỗng để chứa các đối tượng liên quan cùng với giá trị trọng số tương ứng của chúng. Đối với mỗi nhãn lớp đối tượng trong ảnh, thuật toán truy xuất các cạnh từ ConceptNet có nguồn là nhãn đó, từ đó thu được tập các cạnh tương ứng. Các đối tượng liên quan của nhãn đối tượng được xác định từ các cạnh này, và tập kết quả được cập nhật bằng cách kết hợp chúng vào tập đối tượng liên quan. Kết quả cuối cùng của thuật toán là danh sách các đối tượng liên quan cùng với trọng số. Như vậy, từ một ảnh đầu vào I gồm tập các nhãn lớp của các đối tượng đã phát hiện được trong ảnh L_I, qua **Thuật toán 3.1** thu được kết quả là tri thức ngữ nghĩa của các đối tượng liên quan C_I gồm tập đối tượng và giá trị trọng số tương ứng.



Hình 3.2. Minh hoạ kết quả trích xuất tri thức từ ConceptNet của nhãn lớp đối tượng "laptop". Mỗi cạnh được gán nhãn với loại quan hệ (IsA, UsedFor, PartOf, CapableOf) và điểm tin cậy tương ứng.

Thuật toán 3.1. **ExtractRelatedObjectCNet**(L_I, K)

Đầu vào: Tập nhãn đối tượng của ảnh I, $L_I = \{l_1, l_2, .., l_{N_I}\}$, cơ sở tri thức ConceptNet K

Đầu ra: Danh sách các đối tượng liên quan và trọng số tương ứng C_I

Begin

```
    # Khởi tạo tập C rỗng
    C_I ← ∅;
    foreach l_i ∈ L_I do
        # Truy xuất tập cạnh E_i ∈ E từ ConceptNet K
        E_i = {(v_s, v_t, w) | v_s = l_i}
        # Tập đối tượng liên quan của l_i
        c_i = {(v_t, w) | (l_i, v_t, w) ∈ E_i}
        # Cập nhật C_I
        C_I ← C_I ∪ c_i
    end
```

End

3.2.4. Bộ giải mã ngôn ngữ

Trong nghiên cứu này, bộ giải mã (decoder) của kiến trúc Transformer được khai thác như một mô hình ngôn ngữ có khả năng tạo sinh mô tả cho hình ảnh. Các đặc trưng hình ảnh được trích xuất từ bộ mã hóa (encoder) - bao gồm biểu diễn của các vùng đối tượng cùng với embedding của đồ thị quan hệ - được kết hợp và đưa vào bộ giải mã thông qua cơ chế cross-attention. Mô hình được huấn luyện để sinh chú thích ảnh, đồng thời tích hợp thêm tri thức ngữ nghĩa rút trích từ ConceptNet nhằm cải thiện tính chính xác và độ bao quát ngữ nghĩa của kết quả. Toàn bộ quy trình huấn luyện và sinh mô tả ảnh của mô hình được trình bày qua hai thuật toán: **Thuật toán 3.2**: mô tả quá trình huấn luyện bộ giải mã Transformer có tích hợp tri thức ngữ nghĩa để tạo ra mô hình sinh mô tả hình ảnh; **Thuật toán 3.3**: mô tả quá trình sinh chú thích cho ảnh đầu vào bằng cách sử dụng mô hình đã được huấn luyện ở **Thuật toán 3.2**.

Nghiên cứu tập trung vào hai cải tiến chính: (1) cơ chế Multi-Head Cross-Attention để kết hợp thông tin giữa đặc trưng vùng đối tượng và embedding đồ thị quan hệ; và (2) tích hợp tri thức ngữ nghĩa từ ConceptNet nhằm điều chỉnh xác suất dự đoán từ phát sinh, giúp mô hình tạo ra chú thích chính xác, đa dạng và giàu ý nghĩa hơn, đặc biệt với các đối tượng chưa xuất hiện trong dữ liệu huấn luyện. Các tầng khác của Transformer Decoder như Masked Multi-Head Attention, Feed-Forward Layer và Add & Norm không được trình bày chi tiết vì được giữ nguyên theo thiết kế gốc.

Thuật toán 3.2. **TrainingTransDecCNet**(D, φ)

Đầu vào: Tập dữ liệu huấn luyện $D = \{(F_i, Z_i, S_i, C_i), \forall i = \overline{1, N_T}\}$

Đầu ra: Tham số của mô hình φ đã được tối ưu

Begin

$\varphi \leftarrow$ Random Initialization

for $i = 1$ *to* N_T *do*

$Loss \leftarrow 0$

$H_{init} = Embedding(S_i)$

$Q_{masked} = H_{init} W_q^{masked}, K_{masked} = H_{init} W_k^{masked},$

$V_{masked} = H_{init} W_v^{masked}$

$Att_{masked} = Softmax\left(\frac{Q_{masked} K_{masked}^T}{\sqrt{d}}\right) V_{masked}$

$H_{masked} = Add\&Norm(H_{init} + Att_{masked})$

$\mathrm{Q_F} = \mathrm{H_{masked}} \mathrm{W_q^F}, \mathrm{K_F} = \mathrm{F_i} \mathrm{W_k^F}, \mathrm{V_F} = \mathrm{F_i} \mathrm{W_v^F}$

$\mathrm{Att_F} = \mathrm{softmax}\left(\frac{\mathrm{Q_F K_F^T}}{\sqrt{\mathrm{d}}}\right) \mathrm{V_F}$

$\mathrm{Q_Z} = \mathrm{H_{masked}} \mathrm{W_q^Z}, \mathrm{K_Z} = \mathrm{ZW_k^F}, \mathrm{V_Z} = \mathrm{Z_i} \mathrm{W_v^Z}$

$\mathrm{Att_Z} = \mathrm{softmax}\left(\frac{\mathrm{Q_Z K_Z^T}}{\sqrt{\mathrm{d}}}\right) \mathrm{V_Z}$

$\mathrm{CombinedAtt} = \alpha . \mathrm{Att_F} + (1 - \alpha) . \mathrm{Att_Z}$

$H_{cross} = Add\&Norm(H_{masked} + CombinedAtt)$

$H_{final} = FeedForward(H_{cross})$

$Logits = W_o H_{final}$

foreach $(l, w) \in C_i$ *do*

$Logits'[l] = Logits[l] + \beta w$

endfor

$P = Softmax(Logits')$

$Loss_i = -\sum_{t=1}^{Ns_i} logP\left(s_t^i | S_{<t}^i, F_i, Z_i, C_i\right)$

$\varphi \leftarrow \varphi - \eta \frac{\partial Loss_i}{\partial \varphi}$

endfor

End

Trong **Thuật toán 3.2**, N_T là số mẫu dữ liệu trong tập dữ liệu huấn luyện, F_i, Z_i, S_i và C_i lần lượt là tập đặc trưng vùng đối tượng, embedding các đỉnh của đồ thị quan hệ, chú thích chuẩn và tập đối tượng tri thức liên quan của mẫu dữ liệu (ảnh) thứ i. Chú thích chuẩn của mỗi ảnh $S = \{s_1, s_2, \ldots, s_{N_S}\}$, trong đó s_i biểu diễn từ thứ i trong câu, $\forall i = \overline{1, N_S}$, với N_S là số từ trong câu S. H là trạng thái ẩn, W_Q, W_K, W_V và W_O là các ma trận trọng số của Transformer decoder, các ma trận trọng số này được khởi tạo ngẫu nhiên, và được học để cập nhật giá trị trong quá trình huấn luyện.

Thuật toán 3.2 thực hiện huấn luyện Transformer Decoder để phát sinh chú thích ảnh bằng cách kết hợp đặc trưng thị giác từ các vùng đối tượng và đặc trưng ngữ nghĩa từ embedding đồ thị quan hệ thông qua cơ chế cross-attention. Đồng thời, tri thức từ ConceptNet được tích hợp ở tầng đầu ra nhằm cải thiện tính chính xác và độ bao quát ngữ nghĩa của chú thích dự đoán. Cụ thể, tại mỗi mẫu dữ liệu, quá trình huấn luyện bắt đầu với Masked Multi-Head Self-Attention để đảm bảo tính nhân quả, tức là mô hình chỉ sử dụng các từ trước đó để dự đoán từ tiếp theo. Tiếp theo, đặc trưng thị giác (từ các vùng đối tượng) và embedding ngữ nghĩa (từ đồ thị quan hệ) được kết hợp trong cơ chế Multi-Head Cross-Attention. Hệ số kết hợp α được sử dụng để điều chỉnh mức độ đóng góp tương đối giữa hai nguồn thông tin này, qua đó kiểm soát sự cân bằng giữa tín hiệu thị giác và cấu trúc quan hệ trong quá trình suy diễn ngữ nghĩa. Biểu diễn cuối cùng của decoder được ánh xạ qua một tầng tuyến tính để tính điểm số (logits) cho toàn bộ từ vựng. Các logits này sau đó được điều chỉnh trực tiếp bằng cách cộng thêm giá trị βw tương ứng với các từ xuất hiện trong tập tri thức ConceptNet của ảnh. Bước điều chỉnh này làm thay đổi phân phối đầu ra trước khi áp dụng $softmax$, qua đó gia tăng xác suất của các từ có liên hệ ngữ nghĩa với nội dung ảnh, đặc biệt là các đối tượng hiếm hoặc ít xuất hiện trong dữ liệu huấn luyện.

Sau khi áp dụng hàm $softmax$, thu được phân phối xác suất $P\left(s_t^i | S_{<t}^i, F_i, Z_i, C_i\right)$, từ đó tính giá trị hàm mất mát Cross-Entropy cho từng bước thời gian. Ở giai đoạn đầu huấn luyện, do các tham số được khởi tạo ngẫu nhiên, xác suất dự đoán của từ mục tiêu thường thấp, dẫn đến giá trị Loss lớn. Thông qua cơ chế lan truyền ngược (backpropagation), các tham số attention và các ma trận chiếu được cập nhật dần nhằm gia tăng xác suất của từ đúng, làm cho Loss giảm theo các epoch. Việc tích hợp ConceptNet tác động trực tiếp lên $logits$ trước $softmax$, qua đó thúc đẩy quá trình tối ưu đối với các thực thể quan trọng và đóng vai trò như một dạng điều chuẩn ngữ nghĩa ở tầng đầu ra.

Quá trình huấn luyện được lặp lại trên toàn bộ N_T mẫu dữ liệu. Kết quả thực nghiệm cho thấy hàm Loss giảm ổn định và hội tụ sau một số epoch nhất định, cho thấy việc tích hợp tri thức ngoài không làm mất ổn định quá trình tối ưu mà còn hỗ trợ cải thiện khả năng dự đoán các quan hệ và thực thể trong chú thích ảnh.

Trong **Thuật toán 3.3**, quá trình tạo chú thích cho ảnh đầu vào được thực hiện bằng bộ giải mã Transformer (φ) đã được huấn luyện trước đó. Dựa trên đặc trưng vùng đối tượng cùng biểu diễn embedding của đồ thị quan hệ tương ứng với ảnh đầu vào, chuỗi mô tả được khởi tạo bằng token bắt đầu <$start$>. Tại mỗi vòng lặp, chuỗi hiện thời được mã hoá thông qua cơ chế masked multi-head attention, sau đó kết hợp với đặc trưng hình ảnh thông qua multi-head cross-attention để nắm bắt mối liên hệ

giữa ngữ cảnh ngôn ngữ và thông tin thị giác. Biểu diễn kết hợp này được chiếu qua tầng tuyến tính để thu được véc-tơ logits trên toàn bộ từ vựng; trong đó các logits tương ứng với các khái niệm truy xuất từ ConceptNet được điều chỉnh nhằm tăng cường xác suất sinh các từ liên quan về ngữ nghĩa. Phân phối xác suất từ tiếp theo được tính bằng hàm $softmax$. Từ có xác suất lớn nhất được chọn làm từ kế tiếp trong câu. Quy trình này lặp lại, trong đó mỗi từ mới sinh ra được thêm vào chuỗi hiện tại, cho đến khi token kết thúc <end> xuất hiện hoặc đạt độ dài tối đa của câu. Chú thích cuối cùng thu được là một mô tả ảnh chính xác, mạch lạc và có ý nghĩa, được sinh ra dựa trên các đặc trưng biểu diễn của ảnh đầu vào.

Thuật toán 3.3. GenerateCaptionCNet(F_I, Z_I, C_I, φ)

Đầu vào: $F_I,\ Z_I, C_I$ mô hình transformer đã huấn luyện φ
Đầu ra: Chú thích của ảnh $\hat{S}_I$
Begin

$\widehat{S_I} = [< start >]$
while $(last\ token\ \neq < end >)\ and\ \left(length\ of\ \hat{S}_I < max\ length\right)$ **do**
    $H_{init} = Embedding(\hat{S}_I)$
    $H_{masked} = MaskedMultiHeadAttention(H_{init})$
    $H_{cross} = MultiHeadCrossAttention(H_{masked}, F_I, Z_I)$
    $Logits(s_t) = W_o.FeedForward(H_{cross})$
    $Logits' \leftarrow Logits$
    foreach $(l, w) \in C_I$ do
        $Logits'[l] = Logits[l] + \beta w$
    endfor
    $P(s_t) = Softmax(Logits'(s_t))$
    $s_t = argmax_{s \in Vocab} P(s_t)$
    $\hat{S}_I = \hat{S}_I \cup s_t$
endwhile
End

3.3. Thực nghiệm và kết quả

Từ cơ sở lý thuyết và mô hình đã đề xuất, phần này cài đặt thực nghiệm và đánh giá tính hiệu quả của mô hình với các độ đo được sử dụng phổ biến trong bài chú thích ảnh. Các kết quả thực nghiệm, bàn luận và so sánh với các công trình công bố gần đây cũng được trình bày trong phần này nhằm làm rõ ưu khuyết điểm của phương pháp đề xuất.

3.3.1. Dữ liệu và thiết lập thực nghiệm

Phần này trình bày nguồn dữ liệu được sử dụng trong quá trình thử nghiệm, cùng với các tham số huấn luyện và cấu hình hệ thống áp dụng cho phương pháp

được đề xuất. Ngoài ra, các độ đo đánh giá hiệu quả mô hình cũng được mô tả chi tiết trong mục này.

3.3.1.1. Dữ liệu thực nghiệm

Trong chương này, mô hình RGTranCNet được huấn luyện và đánh giá trên tập MS COCO [101] theo đúng các quy tắc chuẩn hoá văn bản/hình ảnh, xây dựng tập từ vựng và chia tách dữ liệu đã trình bày tại Mục 1.4.1. Ngoài ra, cơ sở tri thức ConceptNet [103] được sử dụng như nguồn tri thức ngữ nghĩa ngoài hỗ trợ quá trình giải mã. Việc tích hợp ConceptNet không làm thay đổi cấu hình dữ liệu/đánh giá nêu ở Mục 1.4.1.

3.3.1.2. Chi tiết cài đặt

Việc triển khai mô hình được thực hiện bằng ngôn ngữ Python (phiên bản 3.9) cùng với thư viện học sâu PyTorch (phiên bản 2.0), và toàn bộ quá trình huấn luyện, kiểm thử được thực thi trong môi trường Google Colab Pro. Các cấu hình và siêu tham số chính được thiết lập như sau:

- **Phần tạo và embedding đồ thị quan hệ:** Quy trình này được giữ nguyên theo thiết lập trong OD-VR-Cap, sử dụng các đặc trưng và đồ thị quan hệ đã được huấn luyện sẵn ở Chương 2 nhằm đảm bảo tính nhất quán giữa các mô hình.

- **Thành phần Transformer Decoder:** Kiến trúc giải mã bao gồm 6 khối ($N = 6$) với 8 đầu chú ý (*heads*), kích thước véc-tơ biểu diễn từ được đặt là 512 chiều. Quá trình huấn luyện áp dụng bộ tối ưu Adam, tốc độ học (*learning rate*) = 0.00004 và kích thước lô (*batch size*) = 32.

- **Phần ConceptNet:** Cơ sở tri thức ConceptNet 5.7 được sử dụng để truy xuất các tri thức ngữ nghĩa liên quan đến các đối tượng trong ảnh thông qua REST API tại *api.conceptnet.io*. Với mỗi ảnh, 5 đối tượng có xác suất cao nhất được chọn; đối với mỗi đối tượng, 10 tri thức ngữ nghĩa có độ liên quan lớn nhất được thu nhận và đưa vào bộ giải mã nhằm tăng cường khả năng hiểu ngữ cảnh của mô hình.

3.3.2. Độ đo đánh giá

Hiệu quả của mô hình chú thích ảnh đề xuất (RGTranCNet) được đánh giá bằng các độ đo chuẩn được sử dụng rộng rãi trong bài toán chú thích ảnh, bao gồm BLEU, METEOR, ROUGE-L, CIDEr và SPICE. Các định nghĩa, công thức tính toán và nguyên tắc sử dụng của các độ đo này đã được trình bày chi tiết tại Mục 1.4.2 trong Chương 1.

Tất cả các giá trị đánh giá đều được tính ở mức corpus, và giá trị cao hơn thể hiện hiệu suất tốt hơn của mô hình.

3.3.3. Chi phí tính toán và thời gian thực hiện

Trong bối cảnh chuyển từ bộ giải mã LSTM sang Transformer, việc đánh giá chi phí tính toán là cần thiết nhằm xem xét tính khả thi của mô hình RGTranCNet trong các môi trường huấn luyện tiêu chuẩn. Trong chương này, bộ mã hóa hình ảnh được giữ cố định theo thiết lập huấn luyện trước; do đó, chi phí huấn luyện tập trung chủ yếu vào Transformer decoder và cơ chế tích hợp tri thức ngữ nghĩa từ ConceptNet.

Các thực nghiệm được triển khai trên nền tảng Google Colab Pro với GPU NVIDIA Tesla T4 (16 GB VRAM). Kết quả cho thấy quá trình huấn luyện kéo dài khoảng 16 giờ cho 30 epoch với batch size = 32. So với mô hình sử dụng LSTM ở Chương 2, chi phí huấn luyện tăng lên do số lượng tham số lớn hơn và cơ chế self-attention của Transformer; tuy nhiên, mức chi phí này vẫn nằm trong phạm vi chấp nhận được đối với các hệ thống huấn luyện hiện nay.

Ở pha suy luận, thời gian sinh chú thích trung bình đạt khoảng 0.12 giây/ảnh với batch size = 32. Điều này cho thấy việc áp dụng Transformer decoder giúp cải thiện chất lượng chú thích mà không làm gia tăng đáng kể chi phí suy luận, qua đó đảm bảo khả năng ứng dụng của mô hình trong thực tế.

3.3.4. Kết quả và bàn luận

Kết quả thực nghiệm của phương pháp đề xuất trên tập kiểm tra MSCOCO Karpathy được trình bày trong **Bảng 3.1**. Bảng này nhằm phân tích tác động của việc tích hợp tri thức ngữ nghĩa từ ConceptNet trong kiến trúc Transformer. Với mô hình RGTran (không tích hợp ConceptNet), các chỉ số BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr và SPICE lần lượt đạt 77.5, 34.9, 28.3, 55.3, 98.4 và 18.7. Khi bổ sung thành phần tri thức ngữ nghĩa ngoài để hình thành RGTranCNet, toàn bộ các độ đo đều tăng, đạt 79.8, 36.3, 35.6, 57.2, 107.8 và 20.5. Sự cải thiện đồng thời trên tất cả các độ đo cho thấy việc tích hợp ConceptNet không chỉ mang lại lợi ích cục bộ mà có tác động toàn diện đến chất lượng chú thích. Đáng chú ý, METEOR tăng 7.3 điểm (từ 28.3 lên 35.6), phản ánh khả năng xử lý tốt hơn các biến thể ngôn ngữ và từ đồng nghĩa. CIDEr tăng 9.4 điểm (từ 98.4 lên 107.8), cho thấy mức độ tương đồng cao hơn với các mô tả tham chiếu của con người. Đồng thời, SPICE tăng từ 18.7 lên 20.5, xác nhận rằng tri thức ngoài góp phần cải thiện tính nhất quán ngữ nghĩa và khả năng biểu diễn quan hệ giữa các thực thể trong ảnh. Những kết quả này chứng minh rằng việc tích hợp tri thức ngữ nghĩa vào bộ giải mã Transformer giúp mô hình không chỉ sinh câu đúng về mặt hình thức mà còn chính xác hơn ở mức cấu trúc và ý nghĩa.

Bảng 3.1. Hiệu suất chú thích ảnh của phương pháp đề xuất trên tập kiểm tra MSCOCO Karpathy. Giá trị in đậm biểu thị điểm số cao nhất trong mỗi cột độ đo.

Phương pháp	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
RGTran (**without ConceptNet**)	77.5	34.9	28.3	55.3	98.4	18.7
RGTranCNet (**with ConceptNet**)	**79.8**	**36.3**	**35.6**	**57.2**	**107.8**	**20.5**

Bảng 3.2 mở rộng phân tích bằng cách đặt RGTranCNet trong tương quan với các hướng tiếp cận tiêu biểu trên cùng tập dữ liệu. Trước hết, so với mô hình **OD-VR-Cap** ở Chương 2 – vốn sử dụng LSTM và cơ chế chú ý kép – RGTranCNet thể hiện bước tiến rõ rệt trên toàn bộ các độ đo. Sự cải thiện này phản ánh lợi thế của kiến trúc Transformer với cơ chế self-attention và cross-attention, cho phép tích hợp linh hoạt đặc trưng thị giác, biểu diễn quan hệ và tri thức ngữ nghĩa trong cùng một không gian đặc trưng, thay vì xử lý tuần tự như LSTM. Điều này đặc biệt quan trọng trong việc mô hình hóa các phụ thuộc dài và cấu trúc phức tạp trong câu mô tả.

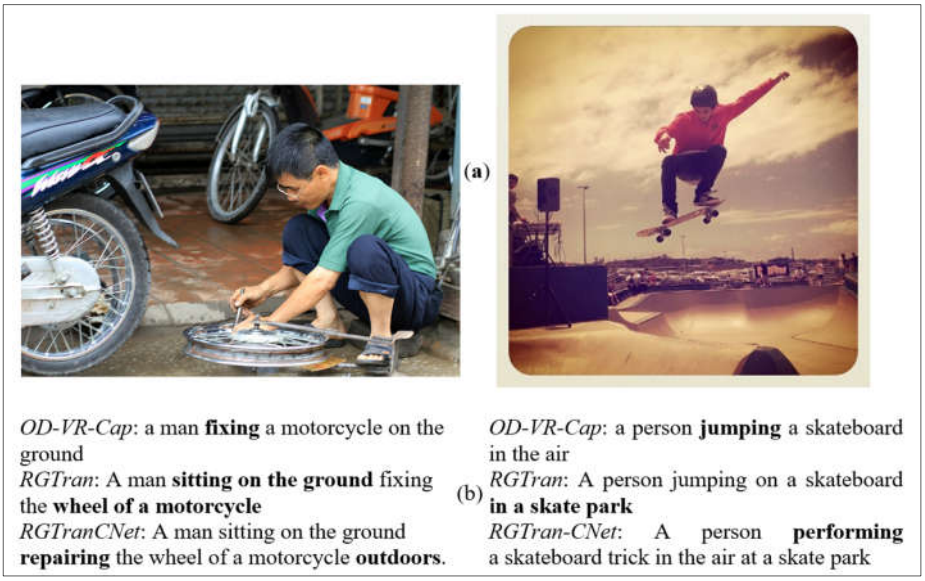
Khi so sánh với các phương pháp RNN truyền thống như Bi-LS-AttM và G-LSTM+att, RGTranCNet vượt trội rõ rệt, cho thấy sự chuyển dịch từ kiến trúc hồi quy sang Transformer là xu hướng tất yếu trong bài toán chú thích ảnh. Đối với CNet-NIC – một phương pháp cũng tích hợp ConceptNet nhưng dựa trên khung NIC – RGTranCNet đạt kết quả cao hơn, khẳng định rằng hiệu quả của tri thức ngoài phụ thuộc mạnh vào cơ chế tích hợp trong kiến trúc giải mã.

Ở một hướng khác, các phương pháp như Image+SceneGraph hay ConvNeXt tập trung tăng cường biểu diễn thị giác thông qua đồ thị cảnh hoặc encoder mạnh. Tuy nhiên, kết quả của RGTranCNet cho thấy việc cải tiến encoder thị giác là chưa đủ nếu thiếu cơ chế tích hợp ngữ nghĩa ở phía giải mã. Hiệu suất cao đồng thời trên BLEU, METEOR, CIDEr và SPICE chứng minh rằng mô hình đề xuất đạt được sự cân bằng giữa độ chính xác bề mặt và tính nhất quán ngữ nghĩa.

Bảng 3.2. So sánh độ chính xác chú thích ảnh của các phương pháp trên tập kiểm tra MSCOCO Karpathy.

Phương pháp	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
CNet-NIC [62]	73.1	29.9	25.6	53.9	107.2	-
Bi-LS-AttM [46]	68.8	25.2	21.5	-	41.2	-
Image+SceneGraph [51]	67.2	26.1	23.3	-	76.0	-
G-LSTM+att [54]	67.7	20.7	23.9	-	-	-
ConvNeXt [104]	74.8	34.8	-	-	-	-
OD-VR-Cap	72.6	28.1	24.8	53.4	85.1	17.6
RGTranCNet	**79.8**	**36.3**	**35.6**	**57.2**	**107.8**	**20.5**

Để làm rõ hơn tính hiệu quả của **RGTranCNet**, **Bảng 3.2** trình bày so sánh với các nghiên cứu gần đây. Kết quả cho thấy RGTranCNet đạt hiệu suất cao nhất trên toàn bộ các độ đo. So với CNet-NIC - một phương pháp cũng tích hợp ConceptNet nhưng dựa trên kiến trúc NIC truyền thống - RGTranCNet vượt trội rõ rệt ở BLEU-4 (+6.4), METEOR (+10.0) và CIDEr (+0.6), khẳng định lợi thế của việc kết hợp tri thức ngoài với kiến trúc Transformer thay vì RNN. Khi so với ConvNeXt, vốn sử dụng kiến trúc mã hóa thị giác tiên tiến, RGTranCNet vẫn đạt hiệu quả vượt trội ở BLEU-1 và BLEU-4, chứng minh rằng không chỉ encoder mà cả cơ chế giải mã và tích hợp tri thức ngữ nghĩa mới đóng vai trò then chốt trong việc nâng cao hiệu suất.



Hình 3.3. Các ví dụ định tính so sánh giữa các mô hình sinh chú thích ảnh. (a) và (b) là hai ảnh thử nghiệm với các chú thích được tạo bởi ba mô hình: OD-VR-Cap, RGTran, và RGTranCNet.

Sau các phân tích định lượng, Hình 3.3 cung cấp minh chứng định tính về chất lượng chú thích. Có thể thấy, OD-VR-Cap chủ yếu dừng lại ở việc mô tả đối tượng và hành động chính (*"a man fixing a motorcycle"*, *"a person jumping a skateboard"*) nhưng bỏ qua các chi tiết ngữ cảnh. RGTran cải thiện hơn khi bổ sung thông tin cụ thể hơn về vị trí hoặc bối cảnh (*"sitting on the ground"*, *"in a skate park"*), tuy nhiên vẫn hạn chế trong việc khai thác chi tiết và quan hệ ngữ nghĩa phức tạp. Trong khi đó, RGTranCNet không chỉ giữ được các yếu tố cốt lõi mà còn sinh ra các mô tả tự nhiên, chi tiết và giàu ngữ nghĩa hơn, chẳng hạn *"repairing the wheel of a motorcycle outdoors"* hoặc *"performing a skateboard trick in the air at a skate park"*. Sự khác biệt này cho thấy tri thức ngoài từ ConceptNet đã giúp mô hình bổ sung thông tin mà dữ liệu huấn luyện chưa bao phủ, từ đó tạo ra chú thích phong phú và gần gũi hơn với ngôn ngữ tự nhiên.

Tóm lại, các kết quả định lượng và định tính đều khẳng định rằng RGTranCNet vượt trội hơn so với các phương pháp CNN-LSTM truyền thống, các mô hình dựa trên đồ thị quan hệ, cũng như các hướng tích hợp tri thức trước đây. Việc kết hợp đồ thị quan hệ, Transformer decoder với cross-attention, và tri thức ngữ nghĩa từ ConceptNet đã chứng minh là một hướng tiếp cận hiệu quả, giúp mô hình không chỉ mô tả chính xác các đối tượng và hành động trong ảnh mà còn tạo ra các chú thích tự nhiên, giàu thông tin và có tính ngữ nghĩa cao.

Sự cải thiện hiệu suất của RGTranCNet phản ánh bước tiến quan trọng trong hướng tiếp cận chú thích ảnh dựa trên tri thức. Việc kết hợp tri thức ngữ nghĩa từ ConceptNet với cơ chế *cross-attention* của Transformer đã mở rộng không gian biểu diễn ngữ nghĩa, giúp mô hình không chỉ "nhìn" được các quan hệ trực tiếp giữa đối tượng mà còn "hiểu" được ý nghĩa khái quát và mối liên hệ ngữ cảnh giữa chúng. Nhờ đó, mô hình có thể diễn đạt linh hoạt hơn, giảm các lỗi mô tả rời rạc và tiến gần hơn đến khả năng diễn đạt tự nhiên của con người. Một đóng góp chính của RGTranCNet là việc chứng minh tính hiệu quả của tích hợp tri thức ngoài vào khung mô hình Transformer, đặt nền tảng cho việc học biểu diễn ngữ nghĩa sâu hơn trong các chương kế tiếp.

Mặc dù đạt kết quả tích cực trên MS COCO, phương pháp vẫn tồn tại một số hạn chế. Thứ nhất, chất lượng chú thích phụ thuộc vào độ chính xác của các mô-đun tiền xử lý như phát hiện đối tượng và dự đoán quan hệ; các sai lệch ở giai đoạn mã hóa có thể lan truyền và ảnh hưởng đến quá trình giải mã. Thứ hai, tri thức truy xuất từ ConceptNet mang tính thường thức tổng quát; trong một số trường hợp, các khái niệm liên quan theo nghĩa phổ quát nhưng không phù hợp với ngữ cảnh cụ thể của ảnh có thể gây nhiễu nếu thiếu cơ chế điều hướng tri thức theo ngữ cảnh mạnh hơn. Thứ ba, cơ chế tích hợp tri thức hiện tại chủ yếu tác động ở mức tăng cường xác suất từ vựng, nên khả năng biểu diễn và kiểm soát cấu trúc ngữ nghĩa ở cấp sự kiện - vai trò ngữ nghĩa vẫn còn hạn chế trong các tình huống cần suy luận trừu tượng.

Những hạn chế này gợi mở nhu cầu đưa vào một biểu diễn ngữ nghĩa có cấu trúc chặt chẽ hơn để điều khiển quá trình sinh câu theo hướng nhất quán ở mức quan hệ và vai nghĩa. Trên cơ sở đó, Chương 4 tiếp tục mở rộng mô hình theo hướng tích hợp Abstract Meaning Representation (AMR) nhằm tăng cường khả năng biểu diễn ngữ nghĩa trừu tượng và cải thiện tính nhất quán ngữ nghĩa của chú thích.

3.4. Kết chương

Trong chương này, mô hình RGTranCNet đã được đề xuất như một hướng phát triển mở rộng từ mô hình OD-VR-Cap nhằm tận dụng ưu điểm của kiến trúc

Transformer và khả năng khai thác tri thức ngữ nghĩa từ nguồn ngoài để nâng cao hiệu quả sinh chú thích ảnh. Cụ thể, mô hình sử dụng bộ giải mã Transformer với cơ chế cross-attention để kết hợp biểu diễn thị giác và ngữ nghĩa từ đồ thị quan hệ; đồng thời, tri thức từ ConceptNet được tích hợp trực tiếp vào quá trình giải mã nhằm tăng cường khả năng hiểu ngữ cảnh và biểu đạt chính xác nội dung hình ảnh.

Kết quả thực nghiệm đã chứng minh rằng kiến trúc RGTranCNet vượt trội so với các phương pháp trước đó trên nhiều độ đo chuẩn như BLEU, METEOR, ROUGE, CIDEr và SPICE khi đánh giá trên tập dữ liệu MS COCO, đồng thời thể hiện tính khả thi trong việc mở rộng áp dụng cho các tập dữ liệu khác như Flickr8k và Flickr30k. Việc tích hợp tri thức ngoài vốn không phụ thuộc vào đặc trưng trực tiếp của ảnh cũng góp phần duy trì tính tổng quát của mô hình, giảm thiểu hiện tượng lệ thuộc vào cấu trúc huấn luyện của từng tập dữ liệu.

Bên cạnh hiệu quả định lượng, RGTranCNet còn mở ra triển vọng ứng dụng thực tiễn trong các hệ thống hỗ trợ người khiếm thị, tìm kiếm ảnh theo ngữ nghĩa, hoặc các trợ lý đa phương thức. Tuy nhiên, một số hạn chế vẫn còn tồn tại, đặc biệt là khả năng trừu tượng hóa và suy luận ngữ nghĩa ở mức sâu. Do đó, trong chương tiếp theo, luận án đề xuất mô hình AMR-GT&RG, kết hợp biểu diễn ngữ nghĩa trừu tượng Abstract Meaning Representation (AMR) với đồ thị quan hệ nhằm đạt được khả năng hiểu ngữ cảnh và biểu diễn ngữ nghĩa toàn diện hơn trong quá trình sinh chú thích ảnh.

CHƯƠNG 4. TÍCH HỢP BIỂU DIỄN AMR VÀO TRANSFORMER TRONG VIỆC CHÚ THÍCH ẢNH

Mặc dù các mô hình chú thích ảnh gần đây đã đạt được những tiến bộ đáng kể, khả năng nắm bắt ngữ nghĩa trừu tượng và hiểu sâu cấu trúc khái niệm trong ảnh vẫn còn hạn chế. Chương 3 trình bày mô hình RGTranCNet, tích hợp thành công đặc trưng quan hệ và tri thức ngữ nghĩa từ ConceptNet, nhưng biểu diễn ngữ nghĩa của mô hình vẫn chủ yếu dựa trên các khái niệm cụ thể, mang tính cục bộ và chưa phản ánh được các cấu trúc ý nghĩa trừu tượng ở cấp độ khái niệm - hành động - quan hệ. Đây là yếu tố quan trọng giúp mô hình sinh ra các mô tả linh hoạt, khái quát và gần với cách con người diễn đạt. Xuất phát từ những giới hạn, Chương này đưa mô hình AMR-GT&RG, mô hình này được xây dựng nhằm tích hợp biểu diễn ngữ nghĩa trừu tượng dựa trên Abstract Meaning Representation (AMR) vào tiến trình chú thích ảnh. Cách tiếp cận này cho phép mô hình kết nối các quan hệ ngữ nghĩa sâu hơn, vượt ra ngoài những gì có thể học trực tiếp từ dữ liệu hình ảnh.

Đóng góp trọng tâm của chương nằm ở việc thiết kế một cơ chế hợp nhất tri thức đa tầng, kết nối đặc trưng thị giác, quan hệ đối tượng và biểu diễn ngữ nghĩa trừu tượng trong cùng một pipeline. AMR được khai thác từ hai nguồn bổ sung - chú thích chuẩn (AMR-GT) và đồ thị quan hệ ảnh (AMR-like) - tạo thành hai dạng biểu diễn được tích hợp vào bộ giải mã Transformer, giúp mô hình hiểu sâu sắc hơn cấu trúc ngữ nghĩa và tăng cường tính tự nhiên, nhất quán của mô tả sinh ra. Nội dung chính của chương đã được công bố trong [CT5] và kế thừa một phần từ các kết quả trong [CT2] và [CT4]. Nội dung của Chương được tổ chức gồm năm phần: (4.1) giới thiệu, (4.2) mô hình đề xuất, (4.3) thực nghiệm và kết quả, và (4.4) kết luận.

4.1. Giới thiệu

Trong khoảng một thập kỷ qua, các mô hình chú thích ảnh chủ yếu được phát triển dựa trên kiến trúc encoder-decoder kết hợp với học sâu, với sự tiến hóa từ các mô hình CNN-RNN/LSTM [3, 6, 18, 30], sang các hướng khai thác đồ thị quan hệ/đồ thị cảnh [23, 49, 50, 93, 105] và gần đây là các kiến trúc Transformer với cơ chế self-attention và cross-attention [104, 106-108]. Đồng thời, một số nghiên cứu đã bắt đầu tích hợp tri thức từ các nguồn dữ liệu bên ngoài tập huấn luyện nhằm nâng cao khả năng tổng quát hóa của mô hình [20, 21, 70, 71, 109]. Các hướng tiếp cận này đã được tổng quan và phân tích chi tiết trong Chương 2 và Chương 3 của luận án.

Mặc dù đạt được nhiều kết quả khả quan, các nghiên cứu gần đây cho thấy bài toán chú thích ảnh vẫn còn tồn tại nhiều thách thức. Cụ thể, phần lớn các phương pháp chỉ khai thác đặc trưng thị giác bề mặt mà chưa tận dụng triệt để ngữ nghĩa trừu

tượng. Mặc dù đồ thị quan hệ hoặc đồ thị cảnh có thể biểu diễn tốt mối quan hệ giữa các đối tượng trong ảnh, nhưng chưa thể hiện đầy đủ cấu trúc ngữ nghĩa sâu của ngôn ngữ. Trong các ảnh có cấu trúc phức tạp, việc chỉ khai thác đặc trưng thị giác hoặc đồ thị quan hệ tường minh thường chưa đủ để mô hình nắm bắt đầy đủ ý nghĩa ngữ cảnh. Đặc biệt, các phương pháp dựa trên trùng khớp bề mặt hoặc quan hệ cục bộ gặp khó khăn khi số lượng đối tượng và tương tác trong ảnh gia tăng.

Trong bối cảnh đó, Abstract Meaning Representation (AMR) [72] nổi lên như một khuôn khổ biểu diễn ngữ nghĩa trừu tượng hiệu quả, cho phép chuẩn hóa ý nghĩa của câu theo dạng *khái niệm - quan hệ - vai trò ngữ nghĩa*, độc lập với biểu thức cú pháp bề mặt. AMR đã được ứng dụng thành công trong nhiều bài toán xử lý ngôn ngữ tự nhiên như dịch máy, trả lời câu hỏi và tóm tắt văn bản, nhờ khả năng làm nổi bật cấu trúc ngữ nghĩa cốt lõi của câu [110]. Tuy nhiên, trong bài toán chú thích ảnh, AMR vẫn chưa được khai thác một cách hệ thống, đặc biệt trong việc kết nối ngữ nghĩa ngôn ngữ trừu tượng với nội dung thị giác của ảnh.

Bên cạnh đó, các chú thích chuẩn trong tập dữ liệu là nguồn thông tin giàu ngữ nghĩa, nhưng phần lớn các nghiên cứu trước đây chưa khai thác hoặc mới chỉ khai thác chúng ở mức chuỗi từ hoặc n-gram, mà chưa tận dụng cấu trúc ngữ nghĩa tiềm ẩn bên trong. Ngoài ra, các độ đo đánh giá phổ biến như BLEU, METEOR, ROUGE, CIDEr và SPICE chủ yếu dựa trên so khớp n-gram, nên chưa phản ánh đầy đủ chất lượng ngữ nghĩa của chú thích được sinh ra [74].

Xuất phát từ những hạn chế nêu trên, chương này đề xuất một mô hình chú thích ảnh mới tích hợp biểu diễn ngữ nghĩa trừu tượng AMR nhằm nâng cao độ chính xác và chiều sâu ngữ nghĩa của chú thích sinh ra. Cụ thể, AMR được khai thác từ hai nguồn bổ sung: (i) AMR trích xuất từ chú thích chuẩn (AMR-GT), đóng vai trò tín hiệu học ngữ nghĩa toàn cục trong giai đoạn huấn luyện; và (ii) biểu diễn AMR-like suy diễn từ đồ thị quan hệ ảnh, cung cấp ràng buộc ngữ nghĩa trực tiếp từ nội dung thị giác trong cả huấn luyện và suy diễn. Hai nguồn biểu diễn này được tích hợp vào Transformer Decoder thông qua các cơ chế Cross-Modal Attention và Masked Multi-Head Attention, đồng thời kế thừa cơ chế tích hợp tri thức từ ConceptNet của mô hình RGTranCNet (Chương 3) nhằm cải thiện khả năng xử lý các đối tượng hiếm.

Bên cạnh kiến trúc mô hình, chương này còn đề xuất độ đo SCS nhằm đánh giá trực tiếp mức độ tương đồng ngữ nghĩa giữa chú thích sinh ra và chú thích chuẩn trong không gian embedding, qua đó khắc phục hạn chế của các độ đo dựa trên n-gram. Các thực nghiệm được tiến hành trên hai tập dữ liệu chuẩn MS COCO và Flickr30K cho thấy phương pháp đề xuất đạt hiệu suất vượt trội, đặc biệt trên các đánh giá mang tính ngữ nghĩa, qua đó khẳng định tính hiệu quả và khả năng tổng quát

hóa của mô hình.

Đóng góp chính của chương này bao gồm:

- Tích hợp ngữ nghĩa trừu tượng từ AMR vào Transformer Decoder: Đề xuất một phương pháp kết hợp ngữ nghĩa trừu tượng từ biểu diễn AMR vào Transformer Decoder thông qua cơ chế Masked Multi-Head Attention và Cross-Modal Attention. Cách tiếp cận này giúp mô hình tạo ra các chú thích chính xác, phong phú về ý nghĩa và phù hợp với nội dung hình ảnh.

- Chuyển đổi đồ thị quan hệ thành đồ thị AMR-like: Giới thiệu một phương pháp chuyển đổi đồ thị quan hệ thành đồ thị AMR-like, đảm bảo giữ nguyên thông tin ngữ nghĩa và cung cấp ngữ cảnh trừu tượng cho quá trình giải mã. Phương pháp này cho phép mô hình khai thác hiệu quả thông tin ngữ nghĩa sâu từ mối quan hệ giữa các đối tượng, từ đó cải thiện khả năng mô tả nội dung hình ảnh một cách chi tiết và tự nhiên hơn.

- Đề xuất độ đo Semantic Consistence Score (SCS): Để vượt qua hạn chế của các độ đo truyền thống dựa trên n-gram, luận án phát triển độ đo SCS. SCS sử dụng mô hình Sentence-BERT huấn luyện trước để trích xuất embedding ngữ nghĩa toàn cục, từ đó đánh giá mức độ tương đồng ngữ nghĩa giữa chú thích do mô hình tạo ra và chú thích chuẩn. Độ đo này phản ánh chính xác chất lượng ngữ nghĩa của chú thích, ngay cả khi cách diễn đạt ngôn ngữ có sự khác biệt.

- Xây dựng mô hình chú thích ảnh toàn diện: mô hình chú thích ảnh hoàn chỉnh được thiết kế, tích hợp nhiều nguồn thông tin bao gồm đặc trưng vùng đối tượng, embedding đồ thị quan hệ, AMR từ chú thích chuẩn, đồ thị AMR-like và tri thức bổ sung từ ConceptNet. Mô hình được thử nghiệm trên hai tập dữ liệu chuẩn MS COCO và Flickr30k, đạt hiệu suất vượt trội trên các độ đo truyền thống (như BLEU, CIDEr) và độ đo SCS. Kết quả thực nghiệm chứng minh tính hiệu quả và khả năng tổng quát hóa cao của mô hình trong việc tạo chú thích ảnh.

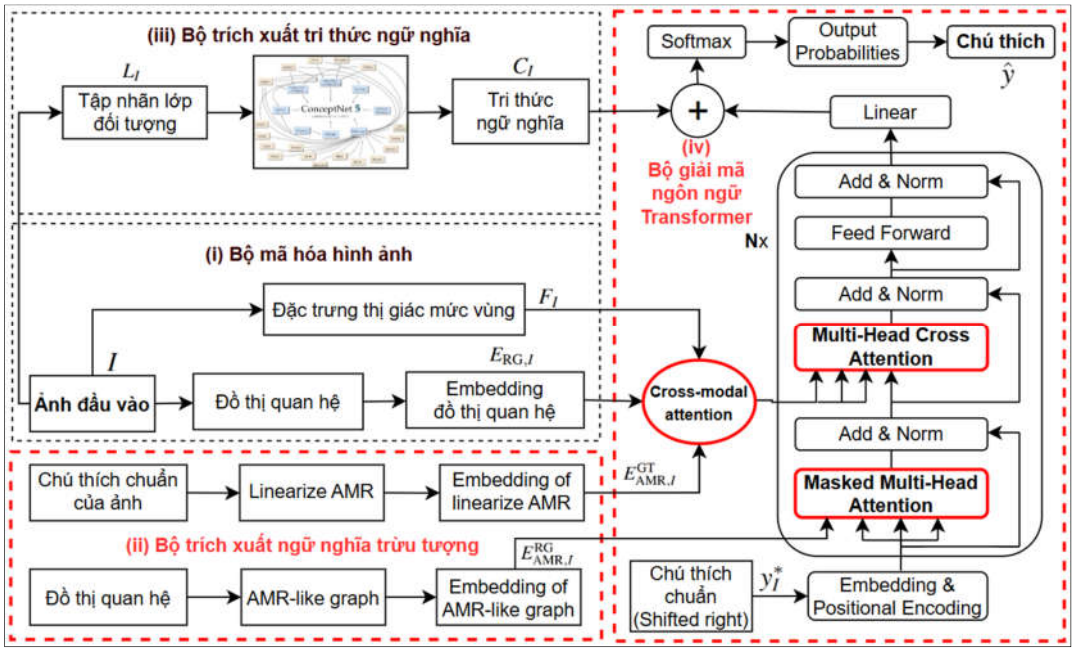
4.2. Phương pháp chú thích ảnh đề xuất

Phần này trình bày chi tiết phương pháp chú thích ảnh đề xuất AMR-GT&RG, bao gồm kiến trúc tổng thể của mô hình và các thành phần chính cấu thành hệ thống, từ bộ mã hóa hình ảnh, bộ trích xuất ngữ nghĩa trừu tượng và tri thức ngữ nghĩa, đến bộ giải mã ngôn ngữ dựa trên Transformer.

4.2.1. Kiến trúc tổng thể của mô hình AMR-GT&RG

Trong chương này, mô hình **AMR-GT&RG** được triển khai theo khung kiến trúc encoder-decoder như minh họa trong Hình 4.1, bao gồm bốn khối chính: (i) bộ mã hóa hình ảnh (Image Encoder), (ii) bộ trích xuất ngữ nghĩa trừu tượng (Abstract

Semantic Extractor), (iii) bộ trích xuất tri thức ngữ nghĩa (Semantic Knowledge Extractor), và (iv) bộ giải mã ngôn ngữ dựa trên Transformer (Transformer Language Decoder). Trong đó, bộ mã hóa hình ảnh được kế thừa trực tiếp từ mô hình OD-VR-Cap trình bày ở Chương 2 nhằm đảm bảo tính nhất quán trong biểu diễn thị giác và đồ thị quan hệ; bộ trích xuất tri thức ngữ nghĩa từ ConceptNet được kế thừa từ mô hình RGTranCNet ở Chương 3; trong khi bộ trích xuất ngữ nghĩa trừu tượng dựa trên AMR và cơ chế tích hợp AMR-like là các thành phần cải tiến cốt lõi được đề xuất trong chương này.



Hình 4.1. Kiến trúc tổng thể của mô hình chú thích ảnh đề xuất. (i) Bộ mã hóa ảnh trích xuất đặc trưng vùng và đồ thị quan hệ; (ii) Bộ trích xuất ngữ nghĩa trừu tượng biểu diễn embedding của đồ thị AMR và AMR-like; (iii) Bộ trích xuất tri thức ngữ nghĩa cung cấp các tri thức ngữ nghĩa từ ConceptNet; (iv) Bộ giải mã Transformer được cải tiến bằng cơ chế MHA, attention đa phương thức, và điều chỉnh điểm số dựa trên ConceptNet nhằm tích hợp tri thức thị giác, quan hệ và tri thức ngoài để tạo chú thích cho ảnh.

Về phương diện kiến trúc, mô hình AMR-GT&RG hướng tới việc tích hợp ba tầng thông tin bổ sung: (i) đặc trưng thị giác mức vùng và quan hệ giữa các đối tượng, (ii) biểu diễn ngữ nghĩa cấu trúc trừu tượng dưới dạng AMR và AMR-like, và (iii) tri thức ngữ nghĩa bên ngoài từ ConceptNet. Các nguồn thông tin này được tích hợp theo cơ chế phân tầng trong bộ giải mã Transformer thay vì đưa vào như các đặc trưng độc lập. Cụ thể, đặc trưng thị giác và embedding đồ thị quan hệ được khai thác trong cơ chế Cross-Modal Attention nhằm cung cấp ngữ cảnh đa phương thức tại mỗi bước giải mã; trong khi embedding AMR-like được tích hợp vào tầng Masked Multi-Head Attention để bổ sung tín hiệu cấu trúc ngữ nghĩa song song với chuỗi từ đã sinh. Cách tổ chức này cho phép mô hình học sự tương thích giữa cấu trúc thị giác và cấu trúc

ngữ nghĩa ở mức sâu hơn so với các phương pháp chỉ dựa trên đặc trưng vùng hoặc đồ thị quan hệ thuần túy.

Trên cơ sở kiến trúc đã minh họa, toàn bộ quy trình được tổ chức thành hai pha chính: huấn luyện và suy diễn. Trong pha huấn luyện, biểu diễn ngữ nghĩa trừu tượng được khai thác từ hai nguồn thông tin bổ sung nhằm cung cấp tín hiệu học giàu ngữ nghĩa cho bộ giải mã. Cụ thể, **Thuật toán 4.1** (EmbeddingGroundTruthAMR) trích xuất embedding AMR từ chú thích chuẩn, đóng vai trò như một chuẩn ngữ nghĩa ở mức trừu tượng để định hướng quá trình học. Song song với đó, **Thuật toán 4.2** (ConvertRGtoAMR) chuyển đổi đồ thị quan hệ của ảnh sang dạng biểu diễn AMR-like, qua đó phản ánh cấu trúc quan hệ ngữ nghĩa nội tại của nội dung thị giác. Trên cơ sở các đặc trưng này, **Thuật toán 4.3** (TrainTransformerDecoder) tiến hành huấn luyện bộ giải mã Transformer với đầu vào đa nguồn, bao gồm đặc trưng thị giác, embedding đồ thị quan hệ, các biểu diễn AMR và tập tri thức ngữ nghĩa.

Trong pha suy diễn, mô hình sinh chú thích theo cơ chế tự hồi quy thông qua **Thuật toán 4.4** (GenerateCaptionAMR). Ở giai đoạn này, bộ giải mã chỉ sử dụng các đặc trưng có thể trích xuất trực tiếp từ ảnh đầu vào, trong đó biểu diễn AMR-like đóng vai trò cung cấp tín hiệu ngữ nghĩa trừu tượng thay thế. Biểu diễn AMR trích xuất từ chú thích chuẩn không được sử dụng nhằm đảm bảo tính nhất quán giữa huấn luyện và suy diễn, đồng thời tránh hiện tượng rò rỉ thông tin từ dữ liệu huấn luyện.

Sau khi mô tả quy trình tổng thể, các khối của mô hình AMR-GT&RG được mô tả chi tiết theo thứ tự từ khâu mã hóa đến giải mã; trước hết là bộ mã hóa hình ảnh, kế thừa từ Chương 2, đóng vai trò cung cấp các biểu diễn thị giác và quan hệ ban đầu cho toàn bộ hệ thống.

4.2.2. Bộ mã hóa hình ảnh

Bộ mã hóa hình ảnh trong mô hình AMR-GT&RG kế thừa kiến trúc đã được trình bày trong Mục 3.2.2 (Chương 3), và gián tiếp từ mô hình OD-VR-Cap tại Mục 2.2.2 (Chương 2). Thành phần này thực hiện ba giai đoạn chính: (i) phát hiện và trích xuất đặc trưng các vùng đối tượng trong ảnh, (ii) xây dựng đồ thị quan hệ giữa các đối tượng, và (iii) embedding đồ thị quan hệ bằng GCN. Hai ma trận đặc trưng đầu ra bao gồm: ma trận đặc trưng thị giác F_I và ma trận đặc trưng ngữ nghĩa Z_I.

Đối với tập dữ liệu MS COCO, mô hình sử dụng lại toàn bộ pipeline ODwGCN để phát hiện đối tượng, kết hợp với các bước xây dựng đồ thị quan hệ (dựa trên mô hình dự đoán quan hệ VRP+RK) và embedding đồ thị bằng GCN hai tầng. Trong khi đó, đối với tập dữ liệu Flickr30K, do không có đủ thông tin để áp dụng mô hình ODwGCN, thành phần phát hiện đối tượng được thay thế bằng mô hình Faster

R-CNN. Các bước tiếp theo như xây dựng đồ thị quan hệ và embedding bằng GCN vẫn được giữ nguyên như đối với tập dữ liệu MS COCO. Điều này bảo đảm sự nhất quán về cấu trúc mô hình, đồng thời duy trì khả năng khai thác mối quan hệ giữa các đối tượng trong ảnh.

Toàn bộ chi tiết về kiến trúc, thuật toán và thiết lập của bộ mã hóa hình ảnh đã được trình bày trong Mục 2.2.2 và Mục 3.2.2.

4.2.3. Bộ trích xuất ngữ nghĩa trừu tượng

Trong phần này, bộ trích xuất ngữ nghĩa trừu tượng (Abstract Semantic Extractor) được xây dựng nhằm khai thác các đặc trưng ngữ nghĩa trừu tượng từ cả chú thích chuẩn (ground-truth caption) và đồ thị quan hệ của ảnh, thông qua biểu diễn AMR. Bộ phận này bao gồm hai hướng xử lý chính: (4.2.3.1) Trích xuất đặc trưng ngữ nghĩa từ các câu chú thích chuẩn thông qua AMR; (4.2.3.2) Khai thác ngữ nghĩa trừu tượng xuất phát từ đồ thị quan hệ của ảnh, được ánh xạ và mở rộng thông qua biểu diễn AMR.

Hai quy trình trên giúp bổ sung thông tin ngữ nghĩa phong phú, tạo nền tảng quan trọng cho bộ giải mã trong quá trình sinh chú thích ảnh trở nên chính xác và tự nhiên hơn.

4.2.3.1. Trích xuất đặc trưng ngữ nghĩa từ chú thích chuẩn thông qua AMR

Các chú thích chuẩn (ground-truth captions) mô tả hình ảnh đóng vai trò là nguồn tri thức ngữ nghĩa giàu thông tin, phản ánh chi tiết về nội dung và ngữ cảnh của hình ảnh. Trong phần này, các chú thích chuẩn được khai thác để xây dựng biểu diễn ngữ nghĩa trừu tượng dưới dạng véc-tơ embedding nhằm tăng cường khả năng nắm bắt nội dung hình ảnh của mô hình một cách sâu sắc hơn. Quá trình này bao gồm ba giai đoạn chính: (i) chuyển chú thích chuẩn thành đồ thị AMR, (ii) tuyến tính hóa đồ thị AMR, và (iii) trích xuất embedding từ linearized AMR.

a) Chuyển chú thích chuẩn thành đồ thị AMR

Trong giai đoạn đầu tiên, việc sinh đồ thị AMR từ văn bản mô tả được thực hiện bằng mô hình NeuralAMR [111], một trong những công cụ tiên tiến và được sử dụng phổ biến cho nhiệm vụ phân tích cú pháp ngữ nghĩa (semantic parsing). Mô hình này đã đạt độ chính xác SMATCH 62.1, thể hiện hiệu quả cao trong cả hai hướng Text-to-AMR (từ văn bản sang đồ thị AMR) và AMR-to-Text (từ đồ thị AMR tái tạo lại văn bản).

Các câu chú thích chuẩn được đưa vào trình phân tích AMR parser để tạo thành đồ thị AMR, trong đó mỗi đồ thị thể hiện cấu trúc ngữ nghĩa nội tại của một hoặc nhiều chú thích tương đồng về ý nghĩa. Mỗi đồ thị AMR bao gồm các khái niệm

(concepts) và quan hệ ngữ nghĩa (semantic relations). Ví dụ, với câu chú thích *"A child is playing with a water bowl"*, đồ thị AMR tương ứng được biểu diễn như sau:

$$(p \;/\; play\text{-}01$$
$$:ARG0\ (c \;/\; child)$$
$$:ARG1\ (b \;/\; bowl$$
$$:mod\ (w \;/\; water)))$$

Trong đó:

- $p/play\text{-}01$: Hành động chính "*play*" với định nghĩa từ PropBank.
- $:ARG0$: Quan hệ chỉ tác nhân hành động, nối với "*child*".
- $:ARG1$: Quan hệ chỉ đối tượng hành động, nối với "*bowl*", được bổ nghĩa bởi "*water*".

b) Tuyến tính hóa AMR graph

Sử dụng định dạng PENMAN để chuyển đổi đồ thị AMR thành chuỗi văn bản tuyến tính, duy trì đầy đủ thông tin ngữ nghĩa của đồ thị. Ví dụ: dạng tuyến tính của đồ thị trên là:

$$(p \;/\; play-01 : ARG0\ (c \;/\; child) : ARG1\ (b \;/\; bowl : mod\ (w \;/\; water)))$$

Chuỗi tuyến tính hóa này được chuẩn hóa để loại bỏ các ký hiệu hoặc thông tin không cần thiết, đảm bảo phù hợp làm đầu vào cho mô hình ngôn ngữ.

c) Trích xuất embedding từ Linearized AMR

Chuỗi tuyến tính sau đó được tiền xử lý, làm đầu vào cho các mô hình ngôn ngữ pretrained để trích xuất embedding, biểu diễn ngữ nghĩa trừu tượng và mối liên kết ngữ nghĩa của hình ảnh. Trong thực nghiệm của nghiên cứu này, mô hình pretrained BERT [112] được sử dụng để trích xuất đặc trưng của các *token*.

Việc chuyển chú thích chuẩn thành embedding thông qua AMR và biểu diễn tuyến tính (linearized AMR) không chỉ giúp giảm sự dư thừa do các biến thể diễn đạt giữa nhiều câu mô tả cùng một hình ảnh, mà còn cho phép trích xuất cấu trúc ngữ nghĩa trừu tượng ở mức sâu hơn, bao gồm các khái niệm cốt lõi (concepts) và vai trò ngữ nghĩa (semantic roles) giữa chúng. Nhờ đó, thông tin ngữ nghĩa được chuẩn hóa theo cấu trúc đồ thị thay vì phụ thuộc vào trật tự từ hay cú pháp bề mặt của câu.

Biểu diễn embedding $E^{GT}_{AMR,I}$ thu được từ quá trình này đóng vai trò như một tín hiệu ngữ nghĩa chuẩn trong pha huấn luyện, cung cấp thông tin bổ sung giàu cấu trúc nhằm hỗ trợ bộ giải mã học được sự tương thích giữa nội dung thị giác và quan hệ ngữ nghĩa trừu tượng. Điều này giúp mô hình biểu diễn nội dung hình ảnh một cách toàn diện và chính xác hơn, đặc biệt ở mức quan hệ giữa các đối tượng.

Thuật toán 4.1. **EmbeddingGroundTruthAMR**$(C_I, M_{AMR}, M_{BERT}, E^{GT}_{AMR,I})$

Đầu vào:

$C_I = \{c_1, c_2 \dots, c_{N_C}\}$: tập chú thích chuẩn của ảnh I, với $N_C = |C|$

M_{AMR}: Mô hình AMR parser (text-to-AMR)

M_{LM}: Mô hình ngôn ngữ huấn luyện trước (BERT)

Đầu ra: $E^{GT}_{AMR,I}$: Biểu diễn embedding ngữ nghĩa trừu tượng của ảnh I.

Begin

1 # Converting Ground-Truth captions into AMR

2 $G_{AMR} = M_{AMR}(C_I)$, *với* $G_{AMR} = \{G_1, G_2, \dots, G_{N_{AMR}}\}, N_{AMR} = |G_{AMR}|$

3 **if** $N_{AMR} > 1$ then

4 $G^* = \underset{G \in G_{AMR}}{argmax} |V_G|$

5 **else**

6 $G^* = G_1$

7 **endif**

8 # Linearizing the graph G* in PENMAN format

9 $S = Linearize(G^*)$

10 # Trích xuất embedding từ Linearized AMR

11 $E^{GT}_{AMR,I} = M_{LM}(S)$

12 **return** $E^{GT}_{AMR,I}$

End

Quy trình trên được hệ thống hóa trong **Thuật toán 4.1**. Cụ thể, tập chú thích chuẩn của ảnh trước hết được phân tích bởi AMR Parser để tạo thành tập các đồ thị AMR tương ứng. Trong trường hợp tồn tại nhiều đồ thị, thuật toán lựa chọn đồ thị có số nút lớn nhất nhằm giữ lại cấu trúc ngữ nghĩa phong phú nhất, dựa trên giả định rằng số lượng nút phản ánh mức độ bao quát các đối tượng và quan hệ trong mô tả. Đồ thị được chọn sau đó được tuyến tính hóa theo định dạng PENMAN để bảo toàn thông tin cấu trúc trong khi chuyển sang dạng chuỗi tuần tự. Cuối cùng, chuỗi tuyến tính này được đưa vào mô hình ngôn ngữ pretrained (BERT) để trích xuất embedding ngữ nghĩa toàn cục, tạo thành biểu diễn $E^{GT}_{AMR,I}$ phục vụ cho các bước huấn luyện tiếp theo của mô hình.

4.2.3.2. Khai thác ngữ nghĩa trừu tượng từ đồ thị quan hệ thông qua AMR

Trong các hệ thống chú thích ảnh, mối quan tâm chính là làm sao để mô hình có thể tạo ra các mô tả chính xác và mang tính ngữ nghĩa sâu từ hình ảnh đầu vào. Đồ thị quan hệ $G = (V, E)$, một biểu diễn cấu trúc phổ biến, cung cấp thông tin về các đối tượng (V) và mối quan hệ giữa chúng (E). Tuy nhiên, đồ thị quan hệ thường

chứa các thông tin cụ thể như vị trí hoặc quan hệ hình học mà thiếu các biểu diễn ngữ nghĩa trừu tượng cần thiết, khiến mô hình gặp khó khăn trong việc phát sinh các chú thích giàu ý nghĩa. AMR, một chuẩn biểu diễn ngữ nghĩa trừu tượng, được thiết kế để nắm bắt ý nghĩa cốt lõi của câu văn thông qua các khái niệm và mối quan hệ trừu tượng [72]. Lấy cảm hứng từ AMR, nghiên cứu này đề xuất chuyển đổi đồ thị quan hệ sang đồ thị AMR-like, một biểu diễn đồ thị trừu tượng hóa nhằm tích hợp thông tin ngữ nghĩa cấp cao vào mô hình chú thích ảnh. Việc embedding đồ thị AMR-like và tích hợp nó vào decoder bổ sung thông tin ngữ nghĩa phong phú, giúp cải thiện hiệu quả của hệ thống chú thích ảnh.

Định nghĩa 4.1: Đồ thị AMR-like $G' = (V', E')$ được định nghĩa như sau:

• V': Tập các đỉnh đại diện cho các khái niệm, được ánh xạ từ tập đỉnh V của đồ thị quan hệ.

• E': Tập các cạnh, đại diện cho các quan hệ ngữ nghĩa giữa các khái niệm, được ánh xạ từ tập cạnh E của đồ thị quan hệ.

Phần này trình bày: (1) Chuyển đổi đồ thị quan hệ thành đồ thị *AMR-like*: Ánh xạ các đối tượng và mối quan hệ trong đồ thị quan hệ sang các khái niệm và nhãn ngữ nghĩa trừu tượng, thông qua hai bảng ánh xạ M_V và M_E. (2) Biểu diễn và embedding đồ thị AMR-like: Sử dụng GraphSAGE để embedding đồ thị AMR-like nhằm trích xuất đặc trưng ngữ nghĩa.

a) Việc chuyển đồ thị quan hệ thành đồ thị AMR-like

Đồ thị quan hệ $G = (V, E)$ là một biểu diễn cấu trúc trong đó tập đỉnh V tương ứng với các đối tượng trong ảnh, và tập cạnh $E \subseteq V \times R \times V$ biểu diễn các quan hệ giữa các đối tượng đó. Các đối tượng trong V được phân loại thành 12 loại chính, bao gồm *person*, *vehicle*, *food*, *furniture*, *clothing*, *animal*, *plant*, *location*, *building*, *structure*, *artifact*, *part*. Mỗi loại chứa nhiều đối tượng cụ thể, ví dụ như loại *vehicle* bao gồm *car*, *bicycle*, *bus*, loại *animal* bao gồm *dog*, *cat*, *horse*… Tập hợp các quan hệ R trong đồ thị quan hệ được chia thành bốn loại chính, phản ánh các mối quan hệ không gian, sở hữu, ngữ nghĩa và quan hệ khác giữa các thực thể. Cụ thể:

−*Geometric relations*: mô tả vị trí tương đối giữa hai đối tượng, bao gồm *above*, *behind*, *under*, *near*…

−*Possessive relations*: thể hiện sự sở hữu hoặc thành phần của đối tượng, bao gồm *has*, *part of*, *wearing*…

−*Semantic relations*: mô tả hành động hoặc tác động giữa hai đối tượng, bao gồm *carrying*, *eating*, *using*…

–*Miscellaneous relations*: bao gồm các quan hệ không thuộc ba nhóm trên, chẳng hạn như *for*, *from*, *made of*...

Danh mục đối tượng và mối quan hệ trên dựa vào phân loại trong tập dữ liệu Visual Genome, MS COCO, Flickr30k, và các nghiên cứu liên quan đến phát sinh đồ thị quan hệ và scene graph. Đây cũng chính là các đối tượng và mối quan hệ tồn tại trong đồ thị quan hệ. Trong đó, đáng chú ý, có 25.2% đối tượng loại parts và 90.9% mối quan hệ là *geometric* hoặc *possessive* [92].

Dựa vào sự tương tự về cấu trúc giữa đồ thị quan hệ và AMR, cùng với danh mục đối tượng V và mối quan hệ E của đồ thị quan hệ đã đề cập. Quá trình chuyển đổi từ đồ thị quan hệ $G = (V, E)$ sang cấu trúc AMR tương ứng, gọi là đồ thị AMR-like $G' = (V', E')$, được thực hiện dựa trên việc ánh xạ tập đỉnh gồm các đối tượng (V) và tập cạnh gồm các mối quan hệ (E) sang tập khái niệm V' và nhãn quan hệ ngữ nghĩa E' trong AMR.

Đồ thị AMR-like $G' = (V', E')$ được định nghĩa như sau:

- V': Tập các đỉnh, đại diện cho các khái niệm, được ánh xạ từ tập đỉnh V của đồ thị quan hệ.
- E': Tập các cạnh, đại diện cho các mối quan hệ ngữ nghĩa giữa các khái niệm, được ánh xạ từ tập cạnh E trong đồ thị quan hệ.

Quá trình chuyển đồ thị quan hệ $G = (V, E)$ thành đồ thị AMR-like $G' = (V', E')$ được thực hiện như **Thuật toán 4.2**, sử dụng quy tắc 1 và 2 như sau:

- **Quy tắc 1**: Ánh xạ tập đỉnh $(V \rightarrow V')$

Xây dựng hàm M_V, ánh xạ một đối tượng $v_i \in V$ thành một khái niệm AMR $c_i \in V'$, với thứ tự ưu tiên như sau:

$$M_V(v_i) = \begin{cases} c_i^g, & if\ \exists c_i^g \in D_{AMR}, c_i^g = type(v_i) \\ c_i^s, & if\ \nexists c_i^g, \exists c_i^s \in D_{A\,MR}, c_i^s = v_i \\ v_i, & otherwise \end{cases} \quad (4.1)$$

Trong đó, c_i^g, c_i^s, D_{AMR} lần lượt là khái niệm tổng quát, khái niệm cụ thể và tập khái niệm trong AMR [72], $type(v_i)$ cho biết loại đối tượng của v_i.

- **Quy tắc 2**: Chuyển đổi tập cạnh $(E \rightarrow E')$:

Xây dựng hàm M_E, xác định quan hệ AMR tương ứng cho quan hệ $r \in E'$, với quy tắc chuyển đổi dựa trên 4 loại mối quan hệ đã đề cập.

$$M_E((c_1, r, c_2)) = \begin{cases} (c_1, :location, r), (r, :opt1, c_2), & if\ r \in R_G \\ (c_1, :poss, c_2), & if\ r \in R_P \\ (r, :ARG0, c_1), (r, :ARG1, c_2), & if\ r \in R_S \\ (c_1, r, c_2), & otherwise \end{cases} \quad (4.2)$$

Trong đó: R_G, R_P, R_S, R_M lần lượt là 4 loại mối quan hệ trong đồ thị quan hệ (*geometric*, *possessive*, *semantic* và *miscellaneous*); c_1, c_2 là *subject* và *object* của mối quan hệ, cũng là các các khái niệm trong AMR đã được ánh xạ từ các đối tượng trong đồ thị quan hệ; $ARG0, AGR1, location, opt1$ là các quan hệ ngữ nghĩa được định nghĩa trong AMR [72].

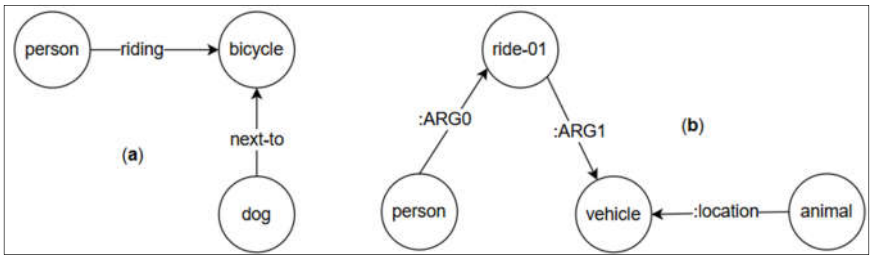
Thuật toán 4.2 chuyển đổi đồ thị quan hệ $G = (V, E)$ thành biểu diễn AMR $G' = (V', E')$ thông qua hai giai đoạn: ánh xạ tập đỉnh và ánh xạ tập cạnh, sử dụng các hàm ánh xạ M_V và M_E. Trước tiên, mỗi đối tượng $v_i \in V$ được ánh xạ sang một khái niệm AMR c_i theo quy tắc ưu tiên trong M_V, đảm bảo lựa chọn khái niệm tổng quát nếu có, hoặc giữ nguyên nhãn nếu không tìm thấy ánh xạ phù hợp. Tập khái niệm V' được xây dựng từ các đối tượng đã được ánh xạ. Tiếp theo, mỗi quan hệ $(c_i, r_{ij}, c_j) \in E$ được ánh xạ bởi M_E thành một hoặc nhiều quan hệ AMR r'_{ij}, tương ứng với bốn nhóm quan hệ chính trong AMR. Tập quan hệ E' chứa các quan hệ đã chuyển đổi. Kết quả là đồ thị AMR-like $G' = (V', E')$, duy trì tính nhất quán về mặt ngữ nghĩa.

Thuật toán 4.2. **ConvertRGtoAMR** (G, M_V, M_E)

Đầu vào: $G = (V, E)$, M_V, M_E
Đầu ra: $G' = (V', E')$
Begin
1 $V' \leftarrow \emptyset$, $E' \leftarrow \emptyset$
2 **for** $v_i \in V$ **do**
3 $c_i \leftarrow M_V(v_i)$
4 $V' \leftarrow V' \cup \{c_i\}$
5 **endfor**
6 **for** $(c_i, r_{ij}, c_j) \in E$ **do**
7 $r'_{ij} \leftarrow M_E[(c_i, r_{ij}, c_j)]$
8 $E' \leftarrow E' \cup \{r'_{ij}\}$
9 **end**
10 **return** $G' = (V', E')$
End

Để minh họa quá trình chuyển đổi từ đồ thị quan hệ sang đồ thị *AMR-like*, xét đồ thị $G = (V, E)$ với tập đỉnh và tập cạnh như sau:

- Tập đỉnh $V = \{person, bicycle, dog\}$,
- Tập cạnh $E = \{(person, riding, bicyle), (dog, next\text{-}to, bicycle)\}$



Hình 4.2. Chuyển đổi từ đồ thị quan hệ sang đồ thị AMR-like, trong đó (a) biểu diễn đồ thị quan hệ đầu vào với các đối tượng và quan hệ giữa chúng, và (b) là đồ thị AMR-like sau chuyển đổi, chuẩn hóa hành động và quan hệ theo cấu trúc AMR.

Hình 4.2 (a) minh họa đồ thị quan hệ đầu vào, trong đó các đối tượng được liên kết bằng các quan hệ trực tiếp giữa chúng. Áp dụng thuật toán ConvertRGtoAMR qua hai bước như sau. Đầu tiên, ánh xạ tập đỉnh V qua M_V, do vehicle là thực thể tổng quát của $bicycle$ và $animal$ là thực thể tổng quát của dog, tập đỉnh thu được là:

$$V' = \{person, vehicle, aminal\}$$

Sau đó, ánh xạ tập cạnh E qua M_E, mỗi quan hệ trong E được chuyển thành nhãn AMR tương ứng. Cụ thể, $riding$ được chuyển thành $ride\text{-}01$ với các tham số ngữ nghĩa là $:ARG0$ và $:ARG1$, $next\text{-}to$ được chuyển thành $:location$. Khi đó, tập cạnh sau khi ánh xạ:

$$E' = \{(ride\text{-}01, :AGR0, person), (ride\text{-}01, :ARG1, vehicle), (vehicle, :location, animal)\}$$

Hình 4.2(b) minh họa đồ thị AMR-like thu được sau khi chuyển đổi. Trong đó, $ride\text{-}01$ được biểu diễn dưới dạng hành động chính, với $person$ là tác nhân ($:ARG0$) và $vehicle$ là đối tượng bị tác động ($:ARG1$). Đồng thời, $next\text{-}to$ được ánh xạ thành $:\text{location}$, giúp biểu diễn mỗi quan hệ ngữ nghĩa giữa $vehicle$ và $animal$ theo cấu trúc AMR.

b) Embedding đồ thị AMR-like

Sau khi chuyển đổi từ đồ thị quan hệ sang đồ thị *AMR-like*, bước tiếp theo là biểu diễn đồ thị này trong không gian đặc trưng nhằm trích xuất thông tin ngữ nghĩa phục vụ cho quá trình sinh chú thích ảnh. Do đồ thị *AMR-like* có cấu trúc tương đồng với đồ thị quan hệ mở rộng đã được trình bày tại Mục 2.2.1.3 (Chương 2), phương pháp embedding sử dụng GraphSAGE, vốn đã được mô tả chi tiết trong Mục 2.2.1.3, cũng được áp dụng để học biểu diễn các đỉnh trong đồ thị *AMR-like*. Phương pháp này dựa trên cơ chế lan truyền thông tin từ các đỉnh lân cận, giúp kết hợp đặc trưng

nội tại của từng đỉnh với thông tin ngữ nghĩa từ các cạnh, qua đó cải thiện hiệu quả biểu diễn ngữ nghĩa trừu tượng. Do kỹ thuật embedding bằng GraphSAGE đã được trình bày đầy đủ tại Mục 2.2.1.3, nội dung hiện tại chỉ nêu rõ bối cảnh áp dụng cho đồ thị AMR-like và không lặp lại chi tiết kỹ thuật nhằm đảm bảo tính nhất quán và tránh trùng lặp.

Đối với mỗi ảnh đầu vào I, embedding của đồ thị dạng AMR tương ứng, ký hiệu là $E_{AMR,I}^{RG}$, được tích hợp vào bộ giải mã Transformer thông qua cơ chế masked multi-head attention nhằm tăng cường khả năng hiểu ngữ nghĩa trong quá trình sinh chú thích.

4.2.4. Bộ trích xuất tri thức ngữ nghĩa

Thành phần bộ trích xuất tri thức ngữ nghĩa trong mô hình AMR-GT&RG được giữ nguyên so với mô hình RGTranCNet đã trình bày trong Mục 3.2.3. Cụ thể, mô hình tiếp tục sử dụng tri thức từ cơ sở dữ liệu ConceptNet để truy xuất các khái niệm ngữ nghĩa liên quan đến các nhãn lớp đối tượng phát hiện trong ảnh. Quá trình truy vấn, ánh xạ và tích hợp tri thức vào bộ giải mã được thực hiện theo **Thuật toán 3.1** - ExtractRelatedObjectCNet, cũng đã được mô tả chi tiết trong Chương 3.

Do không có thay đổi về kiến trúc mô hình hoặc phương thức xử lý tri thức ngữ nghĩa, luận án không lặp lại các định nghĩa, công thức hay mô tả thuật toán liên quan. Quy trình xây dựng và tích hợp tri thức ngữ nghĩa từ ConceptNet đã được trình bày đầy đủ tại Mục 3.2.3.

4.2.5. Bộ giải mã ngôn ngữ Transformer

Decoder đóng vai trò then chốt trong mô hình chú thích ảnh theo khung encoder-decoder, chịu trách nhiệm chuyển đổi thông tin hình ảnh đã được mã hóa thành văn bản mô tả một cách tự nhiên và chính xác về ngữ nghĩa. Trong nghiên cứu này, bộ giải mã ngôn ngữ được triển khai trên kiến trúc Transformer, với mục tiêu khai thác hiệu quả các nguồn thông tin thị giác và ngôn ngữ, ngữ nghĩa. Cụ thể, mô hình tập trung vào các cải tiến chính trong 3 thành phần: (4.2.5.1) Tích hợp ngữ nghĩa trừu tượng của ảnh thông qua AMR, ngữ nghĩa trừu tượng từ đồ thị *AMR-like* (được chuyển từ đồ thị quan hệ) được tích hợp vào Decoder thông qua Masked Multi-Head Attention; (4.2.5.2) Cross-Model Attention, kết hợp thông tin từ đặc trưng vùng đối tượng (F_I), embedding đồ thị quan hệ ($E_{R,I}$) và embedding AMR từ chú thích chuẩn ($E_{AMR,I}^{GT}$) trong một cơ chế chú ý; (4.2.5.3) Tích hợp tri thức ngữ nghĩa từ ConceptNet, kế thừa cơ chế đã công bố trong công trình RGTranCNet, xác suất dự đoán từ được hiệu chỉnh dựa trên tri thức ngữ nghĩa liên quan từ ConceptNet, giúp mô hình tạo chú thích chính xác và giàu ý nghĩa hơn, đặc biệt là các đối tượng ít hoặc chưa có trong

tập dữ liệu. Bên cạnh đó, chi tiết các tầng khác trong Transformer Decoder được bỏ qua, chẳng hạn như Feed-Forward Network, Residual, Layer Normalization, vì những thành phần này được giữ nguyên như thiết kế gốc.

4.2.5.1. Tích hợp ngữ nghĩa trừu tượng của ảnh thông qua AMR

Ngữ nghĩa trừu tượng của ảnh được biểu diễn dưới dạng đồ thị *AMR-like* $G' = (V', E')$, được chuyển đổi từ đồ thị quan hệ của ảnh bằng các quy tắc ánh xạ dựa trên cấu trúc của đồ thị quan hệ và AMR. Thông tin này được đưa vào Transformer Decoder thông qua cơ chế Masked Multi-Head Self-Attention, giúp mô hình học cách kết nối các đối tượng và quan hệ một cách có ngữ nghĩa, thay vì chỉ dựa trên thông tin bề mặt của hình ảnh.

Quy trình xử lý trong Masked Multi-Head Attention được tính toán như sau:

$$Q_{mask} \leftarrow H_{init} W_Q^A, \qquad K_{mask} \leftarrow E_{AMR,I}^{RG} W_K^A, \qquad V_{mask} \leftarrow E_{AMR,I}^{RG} W_V^A$$

$$H_{masked} = softmax\left(\frac{Q_{mask} K_{mask}^T}{\sqrt{d_k}}\right) V_{mask} + H_{init}$$

Trong đó, H_{init} là embedding của các từ đã được sinh trước đó trong câu chú thích, W_Q^A, W_K^A, W_V^A là các ma trận trọng số học được của mô hình. Nhờ đó, mô hình có thể học cách tổ chức ngữ nghĩa dựa trên thông tin trừu tượng thay vì chỉ dựa vào sự xuất hiện của từ trong tập dữ liệu huấn luyện. Mặt nạ (mask) bảo đảm mô hình không nhìn thấy các token ở tương lai, tuân thủ nguyên tắc autoregressive.

4.2.5.2. Cơ chế chú ý Cross-Modal

Bên cạnh cơ chế Masked Multi-Head Attention, thay vì nối các nguồn đặc trưng thành một véc-tơ chung, Cross-Modal Attention được thiết kế nhằm kết hợp ba nguồn thông tin chính từ ảnh theo trọng số: F_I - đặc trưng vùng đối tượng của ảnh; $E_{RG,I}$ - embedding của đồ thị quan hệ, biểu diễn thông tin cấu trúc của ảnh và $E_{AMR,I}^{GT}$ - embedding của AMR từ chú thích chuẩn, giúp bổ sung thông tin ngữ nghĩa đã có trong dữ liệu huấn luyện. Các bước tính toán trong Cross-Modal Attention được thực hiện như sau:

Đầu ra của Masked Multi-Head Attention tại bước t, ký hiệu H_{masked}, được sử dụng để tính toán Query (Q_{cross}), trong khi Key (K_{cross}) và Value (V_{cross}) được tạo bằng cách kết hợp thông tin từ các nguồn khác nhau:

$$Q_{cross} \leftarrow H_{masked} W_Q^C,$$

$$K_{cross} \leftarrow \left[F_I W_K^F; E_{RG,I} W_K^{RG}; E_{AMR,I}^{GT} W_K^{GT}\right],$$

$$V_{cross} \leftarrow \left[F_I W_V^F; E_{RG,I} W_V^{RG}; E_{AMR,I}^{GT} W_V^{GT}\right]$$

Trọng số attention được tính toán thông qua $softmax$ với phép toán dot-product giữa Q_{cross} và K_{cross}:

$$\gamma_{cross} = softmax\left(\frac{Q_{cross}K_{cross}^{T}}{\sqrt{d_k}}\right)V_{cross}$$

Đặc trưng kết hợp được tính bằng cách nhân trọng số attention với *Value*, sau đó thực hiện cơ chế residual connection với H_{masked}:

$$H_{cross} = \gamma_{cross}V_{cross} + H_{masked}$$

Trong đó, $W_Q^C, W_K^F, W_K^{RG}, W_K^{GT}, W_V^F, W_V^{RG}, W_V^{GT}$ là các ma trận trọng số học được của mô hình, trọng số attention xác định mức độ quan trọng của từng nguồn thông tin đối với quá trình phát sinh chú thích. Kết quả của Cross-Modal Attention được sử dụng làm đầu vào cho Feed-Forward Network. Cơ chế này giúp mô hình học cách tập trung vào các thông tin quan trọng từ cả hình ảnh và tri thức ngữ nghĩa trong quá trình phát sinh chú thích. Bằng cách kết hợp các nguồn thông tin khác nhau, Cross-Modal Attention giúp cải thiện tính chính xác và sự liên kết ngữ nghĩa giữa hình ảnh và văn bản đầu ra.

4.2.5.3. Tích hợp tri thức ngữ nghĩa từ ConceptNet

Ngoài hai cơ chế trên, kỹ thuật tích hợp tri thức từ ConceptNet trong công trình RGTranCNet ở Mục 3.2.3 (Chương 3) được áp dụng nhằm điều chỉnh xác suất của từ phát sinh dựa trên các đối tượng có liên quan trong ảnh nhằm cải thiện độ chính xác cho các đối tượng hiếm hoặc chưa xuất hiện trong tập dữ liệu huấn luyện. Cụ thể, tại mỗi bước t, sau khi tính toán phân phối xác suất của từ phát sinh.

$$z_t = H_{final}W_{vocab} \tag{4.3}$$

Các nhãn lớp đối tượng phát hiện trong ảnh được sử dụng để truy vấn trong ConceptNet, từ đó thu thập tập C_I gồm các đối tượng có liên kết ngữ nghĩa mạnh với nội dung ảnh và giá trị trọng số tương ứng. Tập C_I được sử dụng để điều chỉnh xác suất sinh từ:

$$z_t^{(c)} = \begin{cases} z_t[k] + \beta w_k, & if\ k \in C_I \\ z_t[k], & otherwise \end{cases} \tag{4.4}$$

$$P(y_t) = softmax\left(z_t^{(c)}\right) \tag{4.5}$$

Trong đó, β là hệ số điều chỉnh, w_k là trọng số của từ k trong ConceptNet. Nhờ vậy, các từ hiếm hoặc mới được ConceptNet gợi ý có xác suất cao hơn, giúp mô hình tạo ra các chú thích chính xác và hợp lý.

Thuật toán 4.3. **TrainTransformerDecoder** (D, N, φ, η)

Đầu vào: $D = \{(X_I, y_I^*) | X_I = (F_I, E_{R,I}, E_{AMR,I}^{GT}, E_{AMR,I}^{RG}, C_I), y_I^*$: chú thích chuẩn$\}$,
N: số epoch, φ: tham số của mô hình, η tốc độ học.

Output: φ: tham số mô hình đã huấn luyện

Begin

Khởi tạo tham số φ
for $epoch = 1$ to N **do**
  shuffle(D)
  for $(X_I, y^*) \in D$ **do**
    $F_I, E_{RG,I}, E_{AMR,I}^{GT}, E_{AMR,I}^{RG} \leftarrow X_I$
    $y_{<1} \leftarrow$ BOS, $L \leftarrow 0$
    for $i = 1$ to $|y^*|$ **do**
      $H_{init} \leftarrow Embedding(y_{<t})$
      $H_{masked} \leftarrow MaskedMultiHeadAttention(H_{init}, E_{AMR,I}^{RG}, \varphi)$
      $H_{cross} \leftarrow CrossModalAttention(H_{masked}, F_I, E_{RG,I}, E_{AMR,I}^{GT}, \varphi)$
      $H_{final} \leftarrow FeedForwardNetwork(H_{cross}, \varphi)$
      $z_t \leftarrow H_{final} W_{vocab}$
      $z_t^{(c)} \leftarrow AdjustWithConceptNet(z_t, C_I)$ (theo (4.4))
      $P(y_t) \leftarrow softmax(z_t^{(c)})$ (theo (4.5))
      $L \leftarrow L - logP(y_t^* | y_{<t}, X_I, \varphi)$
      $y_{<t+1} \leftarrow y_{<t} \cup \{y_t^*\}$
    endfor
    $\varphi \leftarrow \varphi - \eta \nabla_\varphi L$
  endfor
endfor
return φ

End

Thuật toán 4.3 trình bày quy trình huấn luyện mô hình Transformer Decoder để phát sinh chú thích tự động cho hình ảnh, tận dụng đồng thời các đặc trưng thị giác và tri thức ngữ nghĩa. Cụ thể, đầu vào của thuật toán bao gồm tập đặc trưng thị giác của ảnh (F_I), embedding của đồ thị quan hệ $(E_{R,I})$, embedding AMR từ chú thích chuẩn $(E_{AMR,I}^{GT})$, embedding AMR được chuyển đổi từ đồ thị quan hệ $(E_{AMR,I}^{RG})$, và tập tri thức ngữ nghĩa từ ConceptNet (C_I). Trong mỗi vòng lặp huấn luyện, dữ liệu được xáo trộn để tăng tính tổng quát hóa của mô hình. Với mỗi ảnh đầu vào, các đặc trưng được chuyển vào Decoder để dự đoán các từ trong chú thích. Cơ chế Masked Multi-Head Attention được sử dụng để tích hợp embedding đồ thị AMR-like với chú thích

chuẩn tại từng bước thời gian, trong khi cơ chế Cross-Modal Attention kết hợp các đặc trưng thị giác và ngữ nghĩa để cung cấp ngữ cảnh toàn diện. Trước khi áp dụng hàm $softmax$, điểm số logits của các từ trong từ vựng được điều chỉnh dựa trên tri thức ngữ nghĩa từ ConceptNet, giúp mô hình ưu tiên các từ liên quan đến ngữ cảnh của ảnh. Xác suất của từ đúng được tính từ phân phối $softmax$ và sử dụng để tối ưu hóa hàm mất mát *cross-entropy* thông qua thuật toán gradient descent. Teacher Forcing được áp dụng trong quá trình huấn luyện bằng cách sử dụng token đúng ở bước hiện tại làm đầu vào cho bước tiếp theo, đảm bảo sự ổn định và tốc độ hội tụ của mô hình. Thuật toán này tận dụng triệt để các thông tin từ ảnh và tri thức ngữ nghĩa để nâng cao hiệu quả phát sinh chú thích, tạo ra các mô tả chính xác, ngữ nghĩa phong phú và sát với ngữ cảnh của hình ảnh.

Thuật toán 4.4. **GenerateCaptionAMR**(X_I, φ)

Đầu vào: $X_I = \left(F_I, E_{R,I}, E^R_{AMR,I}, C_I\right), \varphi$: mô hình đã huấn luyện, T_{max}: chiều dài tối đa của câu chú thích.

Đầu ra: $\hat{y}$- câu chú thích cho ảnh đầu vào I.

Begin

$y_{<t} \leftarrow \{start\}, t \leftarrow 1$

while $t < T_{max}$ and $y_t \neq$ <end> **do**

  $H_{init} \leftarrow Embedding(y_{<t})$

  $H_{mask} \leftarrow MaskedMultiHeadAttention\left(H_{init}, E^R_{AMR,I}, \varphi\right)$

  $H_{cross} \leftarrow CrossModalAttention\left(H_{mask}, F_I, E_{R,I}, \varphi\right)$

  $H_{final} \leftarrow FeedForwardNetwork(H_{cross}, \varphi)$

  $z_t \leftarrow Linear\left(H_{final}, \varphi_{vocab}\right)$

  $z_t^{(c)} \leftarrow AdjustWithConceptNet(z_t, C_I, \beta)$

  $y_t \leftarrow softmax\left(z_t^{(c)}\right)$

  $y_t \leftarrow argmax\ P(y_t)$

  $y_{<t+1} \leftarrow y_t \cup \{y_t\}$

  $t \leftarrow t + 1$

endwhile

return $\hat{y} = y_{<t}$

End

Sau khi hoàn tất giai đoạn huấn luyện mô hình Transformer Decoder trong **Thuật toán 4.3**, mô hình đã huấn luyện được sử dụng để phát sinh chú thích ảnh (inference) theo cơ chế autoregressive decoding như **Thuật toán 4.4**. Trong giai đoạn này, do không có sẵn chú thích chuẩn, embedding AMR từ chú thích chuẩn $\left(E^{GT}_{AMR,I}\right)$ không được sử dụng. Thay vào đó, mô hình tận dụng các đặc trưng trích xuất từ ảnh, bao gồm đặc trưng thị giác (F_I), embedding đồ thị quan hệ $(E_{RG,I})$, embedding AMR

được chuyển đổi từ đồ thị quan hệ $\left(E_{AMR,I}^{RG}\right)$, và tri thức ngữ nghĩa từ ConceptNet (C_I). Những đặc trưng này cung cấp đầy đủ thông tin trực quan và ngữ nghĩa, cho phép mô hình phát sinh các chú thích chính xác, giàu ngữ nghĩa, đồng thời xử lý hiệu quả các đối tượng hoặc mối quan hệ phức tạp trong hình ảnh. Quá trình phát sinh bắt đầu với token khởi tạo (<$start$>) và tiếp tục từng bước cho đến khi đạt token kết thúc (<end>) hoặc độ dài tối đa của câu chú thích. Tại mỗi bước t, embedding của chuỗi từ đã sinh được kết hợp với embedding của đồ thị AMR-like thông qua Masked Multi-Head Attention để tích hợp thông tin ngữ nghĩa trừu tượng. Tiếp theo, đặc trưng thị giác và embedding đồ thị quan hệ được kết hợp qua Cross-Modal Attention, giúp mô hình liên kết thông tin giữa các miền dữ liệu. Đầu ra từ các bước này được xử lý bởi Feed-Forward Network (FFN) để tạo biểu diễn cuối cùng. Trước khi áp dụng $softmax$, xác suất dự đoán từ được điều chỉnh bằng tri thức từ ConceptNet để tăng cường khả năng sinh từ phù hợp với ngữ nghĩa ảnh. Từ có xác suất cao nhất được chọn và nối vào chuỗi đã sinh, quá trình lặp lại cho đến khi hoàn tất, tạo ra câu chú thích mô tả chính xác nội dung hình ảnh.

Tóm lại, mô hình đề xuất khai thác đồng thời thông tin thị giác, ngữ nghĩa trừu tượng và tri thức ngữ nghĩa bên ngoài để tạo ra các chú thích chính xác và giàu ý nghĩa. Đặc trưng vùng đối tượng và đồ thị quan hệ được trích xuất từ ảnh, embedding AMR từ chú thích chuẩn và đồ thị AMR-like được sử dụng để bổ sung thông tin ngữ nghĩa trừu tượng, và tri thức từ ConceptNet được tích hợp để điều chỉnh xác suất phát sinh từ. Nhờ các cải tiến này, mô hình có khả năng sinh chú thích không chỉ phù hợp với nội dung ảnh mà còn nhất quán về mặt ngữ nghĩa.

4.3. Thực nghiệm và kết quả

Phần thực nghiệm nhằm đánh giá hiệu quả của mô hình chú thích ảnh đề xuất, với sự tích hợp giữa biểu diễn ngữ nghĩa trừu tượng AMR, đồ thị quan hệ và tri thức ngữ nghĩa từ ConceptNet. Nội dung phần này bao gồm: dữ liệu thực nghiệm (4.3.1), chi tiết cài đặt (4.3.2), độ đo đánh giá (4.3.3), các phương pháp so sánh (4.3.4) và cuối cùng là kết quả cùng với phân thích, bàn luận (4.3.5).

4.3.1. Dữ liệu thực nghiệm

Trong Chương này, các thực nghiệm của mô hình AMR-GT&RG được triển khai trên hai tập dữ liệu MS COCO và Flickr30K theo đúng thiết lập đã trình bày tại Mục 1.4.1 (Karpathy split, chuẩn hoá văn bản/ảnh, xây dựng từ vựng, giới hạn độ dài). Các thông số và quy tắc tiền xử lý ở đây được giữ nguyên nhằm bảo đảm tính nhất quán và khả năng so sánh công bằng với các công trình liên quan.

Trong khuôn khổ các thực nghiệm của Chương này, quy trình xây dựng đồ thị quan hệ được kế thừa từ Chương 2, trong khi cơ chế khai thác ConceptNet như nguồn tri thức ngữ nghĩa ngoài tập dữ liệu được kế thừa từ Chương 3. Hai thành phần này được tích hợp vào pipeline thực nghiệm của chương mà không làm thay đổi cấu hình dữ liệu đã nêu tại Mục 1.4.1.

4.3.2. Chi tiết cài đặt

Phần này mô tả các tham số cài đặt chi tiết cho từng thành phần trong mô hình chú thích ảnh đề xuất, bao gồm: Cấu hình của bộ mã hóa hình ảnh, cài đặt của bộ trích xuất ngữ nghĩa trừu tượng, chi tiết của bộ trích xuất tri thức ngữ nghĩa, kiến trúc bộ giải mã ngôn ngữ Transformer và thiết lập huấn luyện.

4.3.2.1. Cấu hình của bộ mã hóa hình ảnh

Đặc trưng vùng đối tượng F_I được trích xuất từ ảnh đầu vào bằng các mô hình phát hiện đối tượng huấn luyện trước. Đối với MS COCO, ODwGCN được sử dụng để trích xuất đặc trưng vùng đối tượng, kết hợp các mối quan hệ giữa các đối tượng để cải thiện độ chính xác. Trong khi đó, đối với Flickr30k, do hạn chế về dữ liệu quan hệ, Faster R-CNN với ResNet-101 làm backbone được sử dụng để phát hiện các vùng đối tượng. Mỗi đối tượng trong ảnh được biểu diễn bằng một véc-tơ đặc trưng chiều 2048.

Embedding đồ thị quan hệ $E_{RG,I}$: Embedding của các đỉnh trong đồ thị quan hệ được học bằng GraphSAGE như trong mô hình OD-VR-Cap. Embedding của mỗi đỉnh có kích thước 512 và được sử dụng để hỗ trợ quá trình kết hợp ngữ nghĩa trong bộ giải mã.

4.3.2.2. Cài đặt của bộ trích xuất ngữ nghĩa trừu tượng

Embedding AMR từ chú thích chuẩn $E_{AMR,I}^{GT}$: Chú thích chuẩn được chuyển đổi thành đồ thị AMR bằng mô hình NeuralAMR [111]. Đồ thị này được tuyến tính hóa thông qua biểu diễn PENMAN và mã hóa bằng mô hình ngôn ngữ BERT. Kích thước embedding của mỗi token là 768.

Embedding AMR-like từ đồ thị quan hệ $E_{AMR,I}^{RG}$: Đồ thị quan hệ của ảnh được chuyển đổi thành đồ thị AMR-like thông qua các quy tắc ngữ nghĩa. Các embedding của đỉnh trong đồ thị AMR-like được học bằng GraphSAGE với kích thước embedding là 512.

4.3.2.3. Chi tiết của bộ trích xuất tri thức ngữ nghĩa

Cơ chế trích xuất và khai thác tri thức ngữ nghĩa từ ConceptNet trong chương này được kế thừa trực tiếp từ thiết kế đã trình bày chi tiết ở Chương 3. Cụ thể, hệ

thống sử dụng ConceptNet 5.7 để truy xuất các tri thức liên quan đến các đối tượng được phát hiện trong ảnh thông qua REST API, sau đó lựa chọn một tập con các khái niệm và quan hệ ngữ nghĩa có mức độ liên quan cao để tích hợp vào bộ giải mã ngôn ngữ. Các tham số lựa chọn tri thức và hệ số điều chỉnh được giữ nguyên theo thiết lập ở Chương 3 nhằm bảo đảm tính nhất quán và khả năng so sánh công bằng giữa các mô hình.

4.3.2.4. Kiến trúc bộ giải mã ngôn ngữ Transformer và thiết lập huấn luyện

Phần giải mã ngôn ngữ được xây dựng dựa trên kiến trúc Transformer với 6 lớp (layers), trong đó mỗi lớp bao gồm Masked Multi-Head Self-Attention, Cross-Modal Attention, và Feed-Forward Network. Kích thước embedding trong Language Decoder được đặt là 768, với 8 đầu (heads) trong cơ chế Multi-Head Attention. Các tham số được tối ưu hóa thông qua Adam optimizer với tốc độ học $\eta = 0.0001$, gradient clipping được giới hạn ở giá trị 1.0.

4.3.3. Độ đo đánh giá

Trong Chương 4, hiệu quả của mô hình AMR-GT&RG được đánh giá dựa trên các độ đo chuẩn phổ biến trong lĩnh vực chú thích ảnh, bao gồm BLEU, METEOR, ROUGE-L, CIDEr và SPICE. Các định nghĩa, công thức và nguyên tắc tính toán chi tiết của các độ đo này đã được trình bày tại Mục 1.4.2 trong Chương 1.

Mỗi độ đo phản ánh chất lượng chú thích ở một khía cạnh riêng và có cách tính toán khác nhau, song đều có đặc điểm chung là giá trị cao hơn thể hiện hiệu suất mô hình tốt hơn. Phần lớn các độ đo này dựa trên mức độ khớp n-gram giữa chú thích sinh và tham chiếu, chỉ khác nhau ở cơ chế điều chỉnh. Cụ thể, BLEU sử dụng precision n-gram có trọng số nhưng chưa xét đến tính mạch lạc hay ngữ nghĩa; METEOR bổ sung phép khớp đồng nghĩa và biến thể từ vựng song vẫn dựa nhiều vào n-gram; CIDEr dùng trọng số TF-IDF để nhấn mạnh các từ quan trọng, tuy nhiên vẫn thiên về tần suất từ phổ biến hơn là ngữ nghĩa sâu; ROUGE-L dựa trên chuỗi con chung dài nhất (LCS); trong khi SPICE chuyển chú thích thành đồ thị ngữ nghĩa, qua đó phản ánh tốt hơn mối quan hệ giữa các thực thể.

Tuy nhiên, hạn chế chung của các độ đo trên là chưa đánh giá đầy đủ tính tương đồng ngữ nghĩa. Các phương pháp dựa trên n-gram có thể chấm điểm thấp với những câu diễn đạt khác nhau nhưng giữ nguyên ý nghĩa, hoặc ngược lại cho điểm cao với các câu chứa nhiều từ trùng lặp nhưng diễn đạt sai nội dung. Thực tế, chỉ có CIDEr và SPICE được thiết kế riêng cho nhiệm vụ chú thích ảnh; các độ đo còn lại như BLEU, METEOR và ROUGE-L vốn phát triển cho dịch máy hoặc tóm tắt văn bản [74], nên chưa hoàn toàn phù hợp để đánh giá chất lượng mô hình chú thích ảnh

theo nghĩa rộng.

Do đó, nhằm khắc phục hạn chế của các độ đo truyền thống vốn chỉ tập trung vào so khớp n-gram, nghiên cứu này đề xuất độ đo SCS nhằm phản ánh tốt hơn sự tương đồng về ngữ nghĩa giữa chú thích phát sinh và chú thích chuẩn. SCS đo lường mức độ tương đồng ngữ nghĩa giữa câu chú thích phát sinh bởi mô hình chú thích ảnh S_{gen} và câu chú thích chuẩn S_{ref} bằng cách so sánh biểu diễn ngữ nghĩa toàn cục (embedding) của hai câu. Để đảm bảo tính chính xác và ổn định, SCS sử dụng mô hình ngôn ngữ lớn (LLM) Sentence-BERT để trích xuất embedding ngữ nghĩa của mỗi câu, sau đó tính toán độ tương đồng giữa các véc-tơ biểu diễn bằng hàm *cosine similarity*.

$$\mathrm{SCS}(\mathrm{S_{gen}}, \mathrm{S_{ref}}) = \frac{\mathrm{E}(\mathrm{S_{gen}}) \cdot \mathrm{E}(\mathrm{S_{ref}})}{\|\mathrm{E}(\mathrm{S_{gen}})\| \times \|\mathrm{E}(\mathrm{S_{ref}})\|} \tag{4.6}$$

Trong đó, $E(S_{gen})$, $E(S_{ref})$ lần lượt là embedding ngữ nghĩa của của S_{gen} và S_{ref}, $\cdot$ là phép toán tích vô hướng, $\|.\|$ là chuẩn L2 (norm-2). Giá trị SCS nằm trong đoạn $[0,1]$, với: $\mathrm{SCS} = 1$, nếu hai câu hoàn toàn nhất quán về ngữ nghĩa, ngược lại ($\mathrm{SCS} = 0$) thì hai câu không có sự tương đồng về ngữ nghĩa.

Độ đo này không chỉ giúp đánh giá chính xác hơn sự phù hợp ngữ nghĩa giữa câu chú thích và nội dung ảnh mà còn giúp khắc phục hạn chế của các phương pháp dựa trên n-gram, đặc biệt là trong các trường hợp có cách diễn đạt linh hoạt nhưng vẫn giữ nguyên ý nghĩa chính của câu. Trong thực nghiệm của chương này, mô hình Sentence-BERT [113] được sử dụng để trích xuất embedding ngữ nghĩa toàn cục của câu chú thích.

4.3.4. Chi phí tính toán và thời gian thực hiện

Bên cạnh cải thiện về chiều sâu ngữ nghĩa, chương này xem xét chi phí tính toán của mô hình AMR-GT&RG, đặc biệt tập trung vào ảnh hưởng của việc tích hợp biểu diễn ngữ nghĩa trừu tượng AMR trong các giai đoạn huấn luyện và suy luận. Các thực nghiệm được thực hiện trên nền tảng Google Colab Pro với GPU NVIDIA Tesla T4 (16 GB VRAM), sử dụng batch size = 32.

Mô hình được huấn luyện trong 20 epoch trên tập MS COCO và 30 epoch trên tập Flickr30K. Thời gian huấn luyện trung bình ghi nhận khoảng 22 giờ đối với MS COCO và 10 giờ đối với Flickr30K. So với mô hình RGTranCNet ở Chương 3, chi phí huấn luyện tăng lên do việc bổ sung các thành phần xử lý AMR, bao gồm AMR trích xuất từ chú thích chuẩn trong giai đoạn huấn luyện và AMR-like suy diễn từ đồ thị quan hệ ảnh.

Ở pha suy luận, thời gian suy luận trung bình của mô hình AMR-GT&RG đạt khoảng 0.14 giây/ảnh với batch size = 32, chỉ tăng nhẹ so với mô hình ở Chương 3. Cần lưu ý rằng AMR trích xuất từ chú thích chuẩn (AMR-GT) chỉ được sử dụng trong huấn luyện và không tham gia vào pha suy luận; do đó, chi phí suy luận của mô hình vẫn được duy trì ở mức hợp lý, phù hợp với các hệ thống chú thích ảnh trong thực tế.

4.3.5. Kết quả và bàn luận

Trong phần này, hiệu quả của phương pháp đề xuất AMR-GT&RG được phân tích và đánh giá chi tiết theo ba khía cạnh chính. Trước tiên, phương pháp được so sánh với các nghiên cứu công bố gần đây trên hai tập dữ liệu benchmark MS COCO và Flickr30K nhằm xác định mức độ hiệu quả của mô hình trong bối cảnh nghiên cứu hiện tại (4.3.5.1). Tiếp theo, tiến hành phân tích ảnh hưởng của từng thành phần trong mô hình để làm rõ vai trò của AMR thu được từ tập chú thích chuẩn và AMR trích xuất từ đồ thị quan hệ (4.3.5.2). Cuối cùng, phần đánh giá định tính được trình bày thông qua các ví dụ minh họa, qua đó nêu bật những điểm mạnh của mô hình và giải thích sự khác biệt giữa độ đo ngữ nghĩa SCS với các độ đo truyền thống (4.3.5.3).

4.3.5.1. So sánh với các phương pháp công bố gần đây

Để đánh giá một cách khách quan và toàn diện hiệu quả của phương pháp đề xuất (AMR-GT&RG), nghiên cứu tiến hành so sánh kết quả với các công trình hiện tại trên hai tập dữ liệu benchmark trong bài toán chú thích ảnh, bao gồm MS COCO và Flickr30K. Quá trình so sánh được thực hiện dựa trên các độ đo chuẩn trong bài toán này, bao gồm BLEU-1 (B@1), BLEU-4 (B@4), METEOR (M), ROUGE-L (R), CIDEr (C) và SPICE (S). Ngoài ra, độ đo SCS cũng được sử dụng - đây là một độ đo được đề xuất trong nghiên cứu này nhằm đánh giá mức độ nhất quán về ngữ nghĩa giữa câu chú thích sinh ra và chú thích chuẩn của hình ảnh. Các kết quả so sánh được trình bày chi tiết trong **Bảng 4.1** và **Bảng 4.2**.

Kết quả trong **Bảng 4.1** cho thấy rằng phương pháp AMR-GT&RG đạt hiệu suất vượt trội hơn hầu hết các phương pháp trước đây trên đa số các độ đo quan trọng. Cụ thể, trên độ đo BLEU-4, phương pháp đề xuất đạt giá trị 39.5, cao hơn COS-Net (39.1), X-Transformer (37.0), JCRR (37.7) và SGAE (36.9). BLEU-4 đo lường mức độ trùng khớp của cụm từ bốn-gram giữa câu chú thích sinh ra và câu tham chiếu. Kết quả này phản ánh rằng mô hình đề xuất có khả năng tạo ra câu chú thích có mức độ trùng khớp với dữ liệu tham chiếu cao hơn so với đa số các phương pháp trước đây, dù vẫn dựa trên chiến lược tiếp cận khác biệt. Bên cạnh đó, phương pháp AMR-GT&RG đạt 37.2 trên độ đo METEOR, vượt trội hơn tất cả các phương pháp khác, bao gồm COS-Net (29.5), MLA-LRN (29.4), X-Transformer (28.7) và SGAE (27.7).

Khác với BLEU, METEOR không chỉ dựa trên n-gram mà còn xem xét sự liên kết giữa các từ đồng nghĩa và biến thể từ vựng. Giá trị METEOR cao hơn cho thấy rằng mô hình đề xuất có khả năng sinh ra câu chú thích có sự liên kết ngữ nghĩa mạnh mẽ hơn với câu tham chiếu, thay vì chỉ tối ưu hóa trùng khớp từ vựng.

Bảng 4.1. So sánh độ chính xác chú thích ảnh của các phương pháp khác nhau trên tập kiểm tra MSCOCO Karpathy. Giá trị in đậm biểu thị điểm số cao nhất trong mỗi cột độ đo.

Phương pháp	B@1	B@4	M	R	C	S	SCS
Bi-LS-AttM (2023) [46]	68.8	25.2	21.5	-	41.2	-	-
DRL-Attention (2023) [47]	75.2	34.4	28.9	58.8	106.6	-	-
X-Transformer (2020) [107]	77.3	37.0	28.7	57.5	120.0	21.8	-
COS-Net (2022) [108]	78.8	39.1	29.5	58.7	127.4	22.8	-
SGAE (2019) [49]	77.6	36.9	27.7	57.2	116.7	20.9	
MLA-LRN (2023) [114]	79.9	**42.4**	29.4	59.7	125.7	23.2	-
GIC-SSF (2025) [23]	**81.6**	40.1	29.7	59.5	135.4	23.2	
CNet-NIC (2019) [20]	73.1	29.9	25.6	53.9	107.2	-	-
JCRR (2020) [109]	-	37.7	28.2	-	120.1	21.6	-
SemanticGuideAtt (2023) [21]	78.6	36.0	27.6	57.7	120.9	-	-
OD-VR-Cap	72.6	28.1	24.8	53.4	85.1	17.6	75.8
RGTranCNet	79.8	36.3	35.6	57.2	107.8	20.5	82.3
AMR-GT&RG	81.2	39.5	**37.2**	**59.9**	**136.7**	**25.1**	**89.1**

Một trong những điểm mạnh đáng kể của phương pháp đề xuất thể hiện ở độ đo CIDEr, với giá trị 136.7, cao hơn GIC-SSF (135.4) và MLA-LRN (125.7). CIDEr được thiết kế để đánh giá mức độ phù hợp của câu chú thích với tập dữ liệu tham chiếu bằng cách gán trọng số cao hơn cho các từ quan trọng. Kết quả này cho thấy rằng AMR-GT&RG có khả năng sinh ra các câu mô tả ảnh có ý nghĩa gần gũi với tập chú thích chuẩn hơn so với các phương pháp khác, nhờ vào việc tận dụng thông tin từ đồ thị AMR. Ngoài ra, trên độ đo SPICE, phương pháp AMR-GT&RG đạt 25.1, cao hơn COS-Net (22.8), MLA-LRN (23.2) và GIC-SSF (23.2). SPICE là độ đo đặc biệt quan trọng vì nó không chỉ xét đến trùng khớp từ mà còn đánh giá mối quan hệ giữa các thực thể trong câu chú thích bằng cách sử dụng đồ thị cú pháp. Kết quả này phản ánh rằng mô hình đề xuất có khả năng nắm bắt và biểu diễn tốt hơn các mối quan hệ ngữ nghĩa giữa các đối tượng trong ảnh. Mặc dù BLEU-1 (81.2) và BLEU-4 (39.5) của AMR-GT&RG thấp hơn một chút so với GIC-SSF (BLEU-1: 81.6, BLEU-4: 40.1), phương pháp đề xuất vẫn vượt trội trên các độ đo phản ánh tính chính xác ngữ nghĩa, bao gồm METEOR, ROUGE-L, CIDEr và SPICE. Điều này cho thấy rằng mặc dù BLEU có thể đánh giá cao các câu có trùng khớp từ vựng cao, nó không phản ánh đầy đủ khả năng biểu diễn ngữ nghĩa của mô hình.

Trên tập Flickr30K (**Bảng 4.2**), phương pháp AMR-GT&RG tiếp tục thể hiện hiệu suất cao nhất trên hầu hết các độ đo quan trọng. Đặc biệt, chỉ số CIDEr đạt 94.5, cao hơn đáng kể so với GIC-SSF (80.6), MLA-LRN (65.6) và DRL-Attention (92.1). Điều này cho thấy phương pháp đề xuất có khả năng sinh ra câu chú thích phù hợp với tập chú thích chuẩn trên Flickr30K tốt hơn so với các phương pháp so sánh.

Ngoài ra, độ đo METEOR đạt giá trị 35.6, tiếp tục cao hơn các phương pháp đối sánh gồm GIC-SSF (27.3), MLA-LRN (23.5) và DRL-Attention (22.2). Kết quả này củng cố nhận định rằng mô hình đề xuất có khả năng tạo ra câu chú thích có liên kết ngữ nghĩa chặt chẽ hơn với câu tham chiếu, ngay cả trên Flickr30K - một tập dữ liệu có mức độ đa dạng ngôn ngữ cao hơn MS COCO. Tuy nhiên, phương pháp AMR-GT&RG có chỉ số BLEU-1 (79.1) và BLEU-4 (36.4) thấp hơn đôi chút so với GIC-SSF (79.9 và 38.1). Nguyên nhân là do mô hình tập trung khai thác thông tin ngữ nghĩa sâu thông qua AMR thay vì tối ưu chủ yếu theo mức độ trùng khớp n-gram. Việc tối ưu hóa dựa trên n-gram có thể cải thiện BLEU, nhưng không đảm bảo rằng câu chú thích sinh ra phù hợp về ngữ nghĩa với nội dung hình ảnh. Trong khi đó, các chỉ số SCS và SPICE cho thấy tính chính xác ngữ nghĩa đã được cải thiện đáng kể - một khía cạnh mà BLEU không phản ánh đầy đủ.

Bảng 4.2. So sánh độ chính xác chú thích ảnh của các phương pháp khác nhau trên tập kiểm tra Flickr30K Karpathy.

Phương pháp	B@1	B@4	M	R	C	S	SCS
DRL-Attention (2023) [47]	73.8	33.5	22.2	50.4	92.1	-	-
VisualCaptionNet (2024) [48]	74.8	-	20.7	52.8	-	-	-
MLA-LRN (2023) [114]	74.3	31.2	23.5	51.5	65.6	17.2	-
GIC-SSF (2025) [23]	**79.9**	**38.1**	27.3	55.0	80.6	-	-
AMR-GT&RG	79.1	36.4	**35.6**	**56.7**	**94.5**	**22.7**	**87.2**

Bên cạnh việc so sánh với các mô hình tiên tiến gần đây, hiệu suất của phương pháp đề xuất AMR-GT&RG cũng được đối chiếu với hai công trình đã công bố trước đó trong Chương 2 và Chương 3, gồm OD-VR-Cap và RGTranCNet, cả hai đều được đánh giá trên tập dữ liệu MS COCO. RGTranCNet là phiên bản cải tiến của OD-VR-Cap, trong đó bộ giải mã LSTM và cơ chế chú ý kép (dual-attention) được thay thế bằng bộ giải mã dựa trên Transformer và cơ chế chú ý chéo (cross-attention), đồng thời tích hợp tri thức ngữ nghĩa từ nguồn ngoài (ConceptNet). Dựa trên RGTranCNet, mô hình AMR-GT&RG tiếp tục giới thiệu biểu diễn ngữ nghĩa trừu tượng dựa trên AMR được tạo từ cả chú thích chuẩn và đồ thị quan hệ. Như được trình bày trong **Bảng 4.1**, phương pháp AMR-GT&RG vượt trội đáng kể so với cả OD-VR-Cap và RGTranCNet trên tất cả các độ đo đánh giá. So với OD-VR-Cap, AMR-GT&RG cải thiện BLEU-4 +11.4 điểm (39.5 so với 28.1), CIDEr +51.6 điểm (136.7 so với 85.1)

và SCS +13.3 điểm (89.1 so với 75.8). Trong khi RGTranCNet đã mang lại những cải tiến đáng kể so với OD-VR-Cap (ví dụ: CIDEr: 107.8, SCS: 82.3) thông qua nâng cấp kiến trúc và tích hợp tri thức ngữ nghĩa, thì AMR-GT&RG tiếp tục nâng cao hiệu suất (ví dụ: CIDEr: +28.9, SCS: +6.8 so với RGTranCNet). Những kết quả này khẳng định hiệu quả của việc tích hợp biểu diễn ngữ nghĩa trừu tượng dựa trên AMR trong việc nắm bắt ý nghĩa toàn cục và ngữ cảnh quan hệ sâu hơn trong ảnh. Điểm METEOR cũng tăng từ 35.6 (RGTranCNet) lên 37.2 (AMR-GT&RG), cho thấy sự cải thiện về độ phong phú ngữ cảnh và tính đa dạng biểu đạt trong các câu chú thích được sinh ra.

Những kết quả này cho thấy rằng mặc dù BLEU có thể hữu ích trong việc đánh giá mức độ trùng khớp từ vựng, các độ đo như CIDEr, SPICE và đặc biệt là SCS phản ánh chính xác hơn mức độ phù hợp của câu chú thích với nội dung hình ảnh. Điều này chứng minh rằng việc sử dụng AMR để biểu diễn thông tin ngữ nghĩa trong mô hình AMR-GT&RG không chỉ giúp cải thiện tính chính xác của câu chú thích mà còn giúp mô hình tổng quát hóa tốt hơn trên nhiều tập dữ liệu khác nhau.

4.3.5.2. Ảnh hưởng của từng thành phần trong phương pháp đề xuất

Để đánh giá tác động của từng thành phần trong phương pháp AMR-GT&RG, nghiên cứu tiến hành thực nghiệm với hai biến thể: AMR-GT only (chỉ sử dụng AMR từ chú thích chuẩn) và AMR-RG only (chỉ sử dụng AMR chuyển đổi từ đồ thị quan hệ của ảnh). Kết quả so sánh giữa ba mô hình này trên tập dữ liệu MS COCO và Flickr30K được trình bày trong **Bảng 4.3** và **Bảng 4.4**.

Bảng 4.3. Phân tích ảnh hưởng của các thành phần trên tập kiểm tra MS COCO Karpathy. AMR-GT chỉ sử dụng AMR tuyến tính hóa từ chú thích chuẩn; AMR-RG chỉ sử dụng các đồ thị dạng AMR được chuyển đổi từ đồ thị quan hệ; AMR-GT&RG kết hợp cả hai nguồn.

Phương pháp	B@1	B@4	M	R	C	S	SCS
AMR-GT only	80.5	37.8	36.8	59.4	132.5	24.6	87.3
AMR-RG only	79.3	36.9	36.2	58.8	129.8	23.9	84.5
AMR-GT&RG	**81.2**	**39.5**	**37.2**	**59.9**	**136.7**	**25.1**	**89.1**

Kết quả trong **Bảng 4.3** cho thấy rằng phương pháp AMR-GT&RG, khi kết hợp cả AMR từ tập chú thích chuẩn (GT) và AMR từ đồ thị quan hệ (RG), đạt hiệu suất cao nhất trên tất cả các độ đo. Cụ thể, chỉ số CIDEr đạt 136.7, cao hơn đáng kể so với AMR-GT only (132.5) và AMR-RG only (129.8). CIDEr là độ đo quan trọng trong bài toán chú thích ảnh vì nó phản ánh mức độ phù hợp của câu chú thích với tập chú thích chuẩn. Kết quả này chứng minh rằng việc kết hợp cả hai nguồn thông tin AMR giúp mô hình khai thác toàn diện hơn các đặc trưng ngữ nghĩa và tạo ra câu chú thích có mức độ tương đồng cao hơn với chú thích chuẩn. Tương tự, chỉ số SPICE

của AMR-GT&RG cũng đạt 25.1, cao hơn so với AMR-GT only (24.6) và AMR-RG only (23.9). SPICE là một trong những độ đo quan trọng nhất vì nó phản ánh khả năng mô hình nắm bắt các quan hệ ngữ nghĩa giữa các thực thể trong ảnh. Việc kết hợp thông tin từ AMR đồ thị quan hệ đã giúp mô hình tăng cường khả năng biểu diễn mối quan hệ giữa các đối tượng trong ảnh, từ đó cải thiện đáng kể chất lượng ngữ nghĩa của câu chú thích sinh ra.

Bên cạnh đó, chỉ số SCS của phương pháp AMR-GT&RG đạt 89.1, cao hơn đáng kể so với AMR-GT only (87.3) và AMR-RG only (84.5). SCS đo lường mức độ nhất quán về ngữ nghĩa giữa câu chú thích và nội dung hình ảnh, phản ánh rằng phương pháp kết hợp hai loại AMR đã giúp mô hình tạo ra các câu chú thích phù hợp hơn với bối cảnh hình ảnh, thay vì chỉ tập trung vào sự trùng khớp từ vựng như các độ đo truyền thống dựa trên n-gram.

Đáng chú ý, kết quả phân tích ảnh hưởng từng thành phần cho thấy lợi ích của AMR-GT&RG không chỉ thể hiện ở các độ đo dựa trên trùng khớp n-gram mà còn nổi bật hơn ở các độ đo ngữ nghĩa như SPICE và SCS. Về mặt cơ chế, biểu diễn AMR trích xuất từ chú thích chuẩn (AMR-GT) đóng vai trò như một tín hiệu ngữ nghĩa chuẩn hóa, giúp bộ giải mã học được các mẫu cấu trúc khái niệm và vai trò ngữ nghĩa ổn định trong giai đoạn huấn luyện. Trong khi đó, biểu diễn AMR-RG, được suy diễn trực tiếp từ đồ thị quan hệ của ảnh, cung cấp ngữ cảnh ngữ nghĩa trừu tượng có thể khai thác trong giai đoạn suy diễn khi không còn chú thích chuẩn. Sự kết hợp này mang lại lợi ích kép: mô hình vừa học được cấu trúc ngữ nghĩa giàu thông tin từ dữ liệu huấn luyện, vừa duy trì được sự nhất quán ngữ nghĩa giữa các đối tượng và quan hệ khi sinh chú thích ở pha suy diễn.

Bảng 4.4. Phân tích ảnh hưởng của các thành phần trên tập kiểm tra Flickr30K. AMR-GT chỉ sử dụng AMR tuyến tính hóa từ các chú thích chuẩn; AMR-RG chỉ sử dụng các đồ thị dạng AMR được chuyển đổi từ đồ thị quan hệ; AMR-GT&RG kết hợp cả hai nguồn.

Phương pháp	B@1	B@4	M	R	C	S	SCS
AMR-GT only	78.6	35.7	35.2	56.1	90.2	21.8	85.7
AMR-RG only	77.2	34.9	34.7	55.6	88.6	21.3	83.2
AMR-GT&RG	**79.1**	**36.4**	**35.6**	**56.7**	**94.5**	**22.7**	**87.2**

Kết quả trong **Bảng 4.4** trên tập Flickr30K tiếp tục khẳng định tính hiệu quả của phương pháp đề xuất. Cụ thể, chỉ số CIDEr của AMR-GT&RG đạt 94.5, cao hơn so với AMR-GT only (90.2) và AMR-RG only (88.6). Flickr30K là tập dữ liệu có nội dung mô tả đa dạng hơn so với MS COCO, do đó việc mô hình đạt CIDEr cao nhất chứng tỏ rằng sự kết hợp giữa AMR từ tập dữ liệu gốc và AMR từ đồ thị quan hệ giúp mô hình thích nghi tốt hơn với ngữ cảnh ảnh có nội dung phong phú hơn. Tương

tự, SPICE của AMR-GT&RG đạt 22.7, tiếp tục cao hơn AMR-GT only (21.8) và AMR-RG only (21.3). SPICE trên Flickr30K có giá trị thấp hơn so với trên MS COCO do tính đa dạng và độ phức tạp của mô tả trong Flickr30K. Tuy nhiên, phương pháp AMR-GT&RG vẫn đạt kết quả cao nhất, khẳng định rằng mô hình khai thác tốt thông tin quan hệ giữa các đối tượng, ngay cả trong trường hợp câu chú thích có cấu trúc mô tả linh hoạt hơn.

Đặc biệt là, SCS của AMR-GT&RG đạt 87.2, cao hơn so với AMR-GT only (85.7) và AMR-RG only (83.2). Điều này chứng tỏ rằng phương pháp đề xuất giúp cải thiện sự nhất quán về ngữ nghĩa giữa câu chú thích và nội dung ảnh, một yếu tố quan trọng để đảm bảo chất lượng chú thích ảnh trong các tình huống thực tế. BLEU-1 và BLEU-4 trên Flickr30K cũng cho thấy sự cải thiện khi kết hợp hai nguồn AMR (BLEU-1 = 79.1, BLEU-4 = 36.4), cao hơn so với AMR-GT only (78.6/35.7) và AMR-RG only (77.2/34.9). Điều này tiếp tục củng cố nhận định rằng mô hình không chỉ cải thiện độ chính xác ngữ nghĩa mà còn tăng cường mức độ trùng khớp từ vựng với câu tham chiếu.

Nhìn chung, các kết quả ablation trên cả hai tập dữ liệu cho thấy việc chỉ sử dụng AMR-GT hoặc AMR-RG riêng lẻ đều mang lại hiệu quả nhất định nhưng vẫn tồn tại hạn chế. AMR-GT only khai thác tốt thông tin từ dữ liệu huấn luyện nhưng thiếu khả năng ràng buộc ngữ nghĩa trực tiếp với cấu trúc ảnh trong pha suy diễn, trong khi AMR-RG only phản ánh tốt quan hệ giữa các đối tượng trong ảnh nhưng thiếu tín hiệu ngữ nghĩa chuẩn hóa từ chú thích. Phương pháp AMR-GT&RG khắc phục được các hạn chế này bằng cách kết hợp hai nguồn AMR bổ sung lẫn nhau, từ đó giúp mô hình đạt cải thiện đồng thời trên các độ đo CIDEr, SPICE, SCS và METEOR, khẳng định tính hiệu quả của phương pháp đề xuất.

4.3.5.3. Đánh giá định tính

Bên cạnh các phân tích định lượng ở trên, nghiên cứu tiếp tục đánh giá định tính nhằm làm rõ hơn tính hiệu quả của phương pháp đề xuất (AMR-GT&RG) thông qua một số ví dụ minh họa điển hình trên tập dữ liệu MS COCO. Nội dung phần này tập trung vào hai khía cạnh chính: (1) So sánh câu chú thích sinh ra bởi phương pháp đề xuất cùng các biến thể (AMR-GT only**,** AMR-RG only) với phương pháp RGTranCNet; (2) Phân tích trường hợp độ đo SCS thể hiện tốt hơn các độ đo truyền thống dựa trên n-gram.

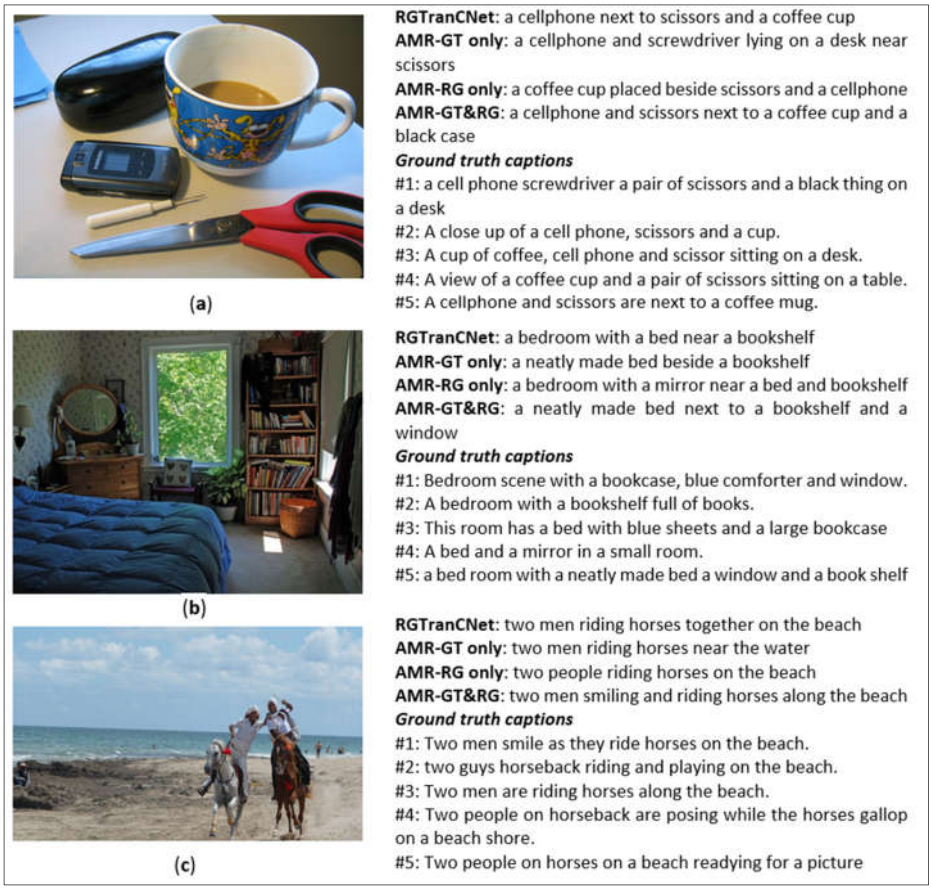
Thứ nhất, để minh chứng cho sự cải tiến của phương pháp đề xuất và các biến thể thành phần (AMR-GT only và AMR-RG only) so với phương pháp nền tảng (RGTranCNet), ba hình ảnh đại diện đã được lựa chọn với mức độ đa dạng nội dung

gồm: nhiều đồ vật (a), nội thất có cấu trúc không gian phức tạp (b) và hoạt động ngoài trời thể hiện tương tác giữa các đối tượng (c). Kết quả chú thích sinh ra bởi các mô hình được trình bày như trong Hình 4.3. Ở Hình 4.3(a), câu chú thích do RGTranCNet tạo ra đề cập đến các đối tượng chính nhưng thiếu chi tiết về *screwdriver* và vật thể màu đen trên bàn. Phương pháp AMR-GT only đã bổ sung thêm thông tin về *screwdriver* nhưng vẫn chưa đầy đủ. AMR-RG only tập trung vào quan hệ giữa các đối tượng nhưng bỏ sót *screwdriver*. Trong khi đó, phương pháp đề xuất AMR-GT&RG khai thác đầy đủ các đối tượng trong ảnh, tạo ra một mô tả tự nhiên và chính xác hơn. Với Hình 4.3(b), tương tự như trên, RGTranCNet mô tả cơ bản về phòng ngủ nhưng thiếu chi tiết về cửa sổ. AMR-GT only tập trung vào giường ngủ, AMR-RG only bổ sung chi tiết về gương nhưng vẫn chưa đầy đủ. Phương pháp AMR-GT&RG kết hợp cả ba yếu tố giường, tủ sách và cửa sổ, tạo ra mô tả toàn diện hơn về bối cảnh hình ảnh. Ở ví dụ Hình 4.3(c), RGTranCNet mô tả hai người đàn ông cưỡi ngựa trên bãi biển nhưng không đề cập đến các yếu tố cảm xúc. Câu chú thích của AMR-GT only bổ sung thêm thông tin về nước biển, trong khi AMR-RG only nhấn mạnh đến quan hệ giữa người và ngựa. Phương pháp AMR-GT&RG kết hợp tất cả các yếu tố quan trọng, tạo ra câu chú thích đầy đủ nhất khi đề cập đến cả hành động và cảm xúc của đối tượng.

Thứ hai, nhằm làm nổi bật ưu điểm của độ đo SCS được đề xuất so với các độ đo đánh giá truyền thống dựa trên n-gram, hai ví dụ minh họa được trình bày trong Hình 4.4. Trong các trường hợp này, câu chú thích được tạo bởi mô hình đề xuất đã phản ánh chính xác ngữ nghĩa của nội dung ảnh, tuy nhiên lại khác biệt về cách diễn đạt, dẫn đến điểm số thấp từ các độ đo n-gram, trong khi điểm SCS nhận giá trị cao - thể hiện rõ hơn độ chính xác về ngữ nghĩa và mức độ phù hợp ngữ cảnh của câu mô tả. Cụ thể, trong Hình 4.4 (a), do sự khác biệt về cách dùng từ vựng (ví dụ: “striped animals” thay vì “zebras”, “habitat” thay vì “enclosure”), điểm BLEU-4 chỉ đạt 7.0. Tuy nhiên, điểm SCS đạt tới 80.6, cho thấy mức độ phù hợp ngữ nghĩa cao với nội dung hình ảnh. Tương tự, trong Hình 4.4(b), câu chú thích được tạo bởi mô hình AMR-GT&RG truyền tải cùng một ý nghĩa với chú thích chuẩn, nhưng sử dụng cách diễn đạt khác (“arranged with various foods” thay vì “filled with food”), dẫn đến điểm BLEU-4 chỉ là 17.3. Tuy vậy, điểm SCS đạt 94.3, phản ánh chính xác mức độ tương thích ngữ nghĩa giữa câu được sinh và câu tham chiếu.

Phương pháp AMR-GT&RG thể hiện ưu điểm trong việc tạo ra câu chú thích chi tiết, chính xác và phù hợp ngữ cảnh hơn so với phương pháp RGTranCNet. Bằng cách tích hợp AMR, mô hình đề xuất có thể mô tả các đối tượng và mối quan hệ trong ảnh tốt hơn, dẫn đến câu chú thích giàu thông tin hơn. Ngoài ra, phân tích về độ đo

SCS khẳng định rằng các độ đo truyền thống như BLEU có thể đánh giá thấp những câu mô tả đúng nhưng có cách diễn đạt khác biệt. Độ đo SCS giúp khắc phục hạn chế này bằng cách tập trung vào sự tương đồng ngữ nghĩa, mang lại đánh giá toàn diện và chính xác hơn về chất lượng chú thích ảnh.



Hình 4.3. Ví dụ định tính trên ba ảnh kiểm tra (a)-(c). Đối với mỗi ảnh, chú thích được tạo bởi mô hình RGTranCNet, ba biến thể của mô hình đề xuất - AMR-GT only, AMR-RG only và AMR-GT&RG - được liệt kê lần lượt, tiếp theo là năm chú thích chuẩn.

Sự cải thiện hiệu suất của AMR-GT&RG bắt nguồn từ việc mô hình không chỉ xử lý các đặc trưng thị giác bề mặt mà còn học được cấu trúc ngữ nghĩa tiềm ẩn của cảnh vật thông qua biểu diễn AMR. Cơ chế kết hợp song song giữa AMR từ chú thích chuẩn và AMR chuyển đổi từ đồ thị quan hệ đã giúp mô hình đồng thời nắm bắt được tri thức ngôn ngữ trừu tượng và tri thức cấu trúc cụ thể, từ đó hiểu sâu hơn mối liên hệ giữa các thực thể trong ảnh. Đóng góp nổi bật của mô hình là minh chứng rằng việc tích hợp biểu diễn ngữ nghĩa trừu tượng vào khung Transformer có thể thu hẹp khoảng cách giữa hiểu biết thị giác và ngôn ngữ, giúp mô hình sinh ra các mô tả mang tính khái quát, tự nhiên và gần với cách diễn đạt của con người hơn.

Bên cạnh những cải thiện về ngữ nghĩa, phương pháp AMR-GT&RG vẫn tồn tại một số đánh đổi nhất định. Thứ nhất, nhánh AMR-GT làm tăng chi phí tiền xử lý

do phụ thuộc vào chú thích chuẩn và quá trình phân tích AMR. Thứ hai, chất lượng biểu diễn ngữ nghĩa chịu ảnh hưởng bởi độ chính xác của AMR parser, trong đó các sai lệch phân tích có thể lan truyền sang bộ giải mã. Thứ ba, trong một số trường hợp, các độ đo dựa trên n-gram như BLEU không tăng tương ứng do mô hình ưu tiên biểu diễn ngữ nghĩa sâu hơn thay vì tối ưu trùng khớp từ vựng. Những hạn chế này gợi mở các hướng cải tiến tiếp theo như giảm phụ thuộc vào AMR parser hoặc thay thế nhánh AMR-GT bằng các biểu diễn ngữ nghĩa được học trực tiếp trong quá trình huấn luyện.

(a)

AMR-GT&RG: Two striped animals eating grass in their habitat.

Ground truth captions
#1: Two zebras graze on grass inside of an enclosure.
#2: Two zebras at the zoo eating some grass.
#3: Two zebras grazing on the green grass in an enclosure.
#4: A pair of zebras grazing on grass in a zoo.
#5: A couple of zebras are grazing in a field

(b)

AMR-GT&RG: A dining table arranged with various foods and a cake covered with berries.

Ground truth captions
#1: A plate of finger foods next to a blue and raspberry topped cake.
#2: A nicely set dining table filled with food and a cake topped with berries.
#3: a close up of a table with many plates of food
#4: A table topped with a cake covered in berries next to a plate of sandwiches.
#5: A white cake topped with berries and a plate of fruit and cheeses.

Hình 4.4. Ví dụ về chú thích phát sinh của mô hình đề xuất - trường hợp chú thích phù hợp với ngữ nghĩa của nội dung ảnh nhưng hiệu suất trên các độ đo dựa trên n-gram thấp.

4.4. Kết chương

Trong chương này, luận án đã đề xuất mô hình AMR-GT&RG, một kiến trúc chú thích ảnh mở rộng từ RGTranCNet nhằm khai thác biểu diễn ngữ nghĩa trừu tượng dựa trên AMR kết hợp với đồ thị quan hệ ảnh, qua đó nâng cao chất lượng và chiều sâu ngữ nghĩa của chú thích sinh ra. Mô hình tích hợp đồng thời ba nguồn tri thức bổ sung: (i) AMR trích xuất từ chú thích chuẩn để cung cấp tín hiệu học ngữ nghĩa có cấu trúc; (ii) biểu diễn AMR-like suy diễn từ đồ thị quan hệ nhằm bổ sung ngữ cảnh ngữ nghĩa trực tiếp từ nội dung thị giác; và (iii) tri thức ngoài từ ConceptNet để cải thiện khả năng xử lý các đối tượng hiếm. Trên nền kiến trúc Transformer, hai cơ chế Masked Multi-Head Attention và Cross-Modal Attention được thiết kế nhằm hợp nhất hiệu quả thông tin thị giác, quan hệ và ngữ nghĩa trừu tượng trong quá trình giải mã.

Kết quả thực nghiệm trên hai tập dữ liệu chuẩn MS COCO và Flickr30K cho thấy AMR-GT&RG đạt hiệu suất vượt trội trên các độ đo phản ánh chất lượng ngữ nghĩa như METEOR, CIDEr, SPICE và SCS, đồng thời thể hiện khả năng mô hình hóa tốt hơn các quan hệ và bối cảnh phức tạp trong ảnh. Mặc dù các độ đo dựa trên

trùng khớp n-gram như BLEU không luôn đạt giá trị cao nhất, các kết quả ngữ nghĩa cho thấy mô hình đề xuất tạo ra các chú thích nhất quán, giàu ý nghĩa và gần với cách diễn đạt của con người hơn.

Bên cạnh những cải thiện về mặt ngữ nghĩa, việc tích hợp biểu diễn AMR cũng kéo theo một số đánh đổi nhất định về chi phí tính toán. Cụ thể, chi phí bổ sung chủ yếu phát sinh từ bước chuyển đổi đồ thị quan hệ sang biểu diễn AMR-like và quá trình embedding ngữ nghĩa trừu tượng. Tuy nhiên, các bước này được thực hiện chủ yếu trong pha huấn luyện; ở pha suy luận, mô hình chỉ sử dụng lại các tham số đã được học và không yêu cầu trích xuất AMR từ chú thích chuẩn, do đó độ trễ suy diễn không tăng đáng kể so với các mô hình chú thích ảnh dựa trên đồ thị quan hệ. Nhìn chung, chi phí suy diễn của AMR-GT&RG về cơ bản tương đương với các mô hình cùng họ, và hoàn toàn có thể tiếp tục tối ưu thông qua các kỹ thuật như nén mô hình hoặc sử dụng các mô-đun phát hiện đối tượng nhẹ hơn trong triển khai thực tế.

Trên cơ sở các kết quả đạt được, Chương 5 tiếp tục phát triển hướng nghiên cứu này theo hướng hợp nhất ngữ nghĩa đa phương thức, kết hợp đặc trưng thị giác-ngôn ngữ từ CLIP, biểu diễn đồ thị quan hệ và AMR/AMR-like, đồng thời giới thiệu các cơ chế Adaptive Attention và GPT-based Re-ranking nhằm cải thiện thêm tính tự nhiên, mạch lạc và khả năng kiểm soát chất lượng của chú thích ảnh sinh ra.

CHƯƠNG 5. MÔ HÌNH CHÚ THÍCH ẢNH DỰA TRÊN HỢP NHẤT NGỮ NGHĨA ĐA PHƯƠNG THỨC VÀ GPT RE-RANKING

Chương 2, Chương 3 và Chương 4 phát triển các hướng tiếp cận nhằm nâng cao chất lượng chú thích ảnh, bao gồm mô hình hóa quan hệ giữa các đối tượng (OD-VR-Cap), tích hợp tri thức ngữ nghĩa ngoài tập huấn luyện (RGTranCNet), và khai thác biểu diễn ngữ nghĩa trừu tượng dựa trên AMR (AMR-GT&RG). Mặc dù các mô hình này đã giải quyết nhiều thách thức cốt lõi, các phương pháp chú thích ảnh hiện nay vẫn còn hai hạn chế nổi bật. Thứ nhất, chưa có cơ chế hợp nhất linh hoạt giữa các nguồn tri thức khác nhau - đặc trưng thị giác, cấu trúc quan hệ và ngữ nghĩa trừu tượng - dẫn đến việc khai thác thông tin bổ trợ giữa các miền chưa tối ưu. Thứ hai, chất lượng ngôn ngữ đầu ra đôi khi vẫn thiếu mạch lạc và tự nhiên, cho thấy cần có cơ chế kiểm soát và tinh chỉnh câu chú thích hiệu quả hơn.

Để khắc phục những hạn chế nêu ra, Chương này giới thiệu mô hình CLIP-AMR-GPT như bước phát triển hoàn thiện trong chuỗi nghiên cứu của luận án. Mô hình hợp nhất đặc trưng thị giác-ngôn ngữ từ CLIP, thông tin cấu trúc từ đồ thị quan hệ, và ngữ nghĩa trừu tượng từ biểu diễn AMR trong một kiến trúc encoder-decoder đa nguồn. Hai cơ chế bổ trợ - Adaptive Attention và GPT-based Re-ranking - được thiết kế nhằm điều tiết linh hoạt các nguồn tri thức trong quá trình giải mã, đồng thời nâng cao tính tự nhiên và độ chính xác ngữ nghĩa của chú thích sinh ra. Đóng góp trọng tâm của chương nằm ở việc xây dựng một cơ chế hợp nhất tri thức đa nguồn kết hợp hiệu quả thông tin thị giác, cấu trúc và ngữ nghĩa, đồng thời bổ sung bước tinh chỉnh ngôn ngữ để cải thiện đáng kể sự mạch lạc và tự nhiên của câu chú thích. Cách tiếp cận này hướng đến xây dựng một khuôn khổ hợp nhất tri thức đa tầng, tối ưu đồng thời chất lượng ngữ nghĩa và chất lượng ngôn ngữ. Nội dung chính của chương này đã được công bố trong [CT6]. Cấu trúc chương gồm bốn phần: (5.1) giới thiệu bối cảnh và hướng tiếp cận; (5.2) mô tả chi tiết mô hình CLIP-AMR-GPT; (5.3) thực nghiệm và kết quả; và (5.4) kết luận chương.

5.1. Giới thiệu

Mặc dù đã đạt được nhiều tiến bộ đáng kể, các mô hình chú thích ảnh hiện nay vẫn tồn tại một số hạn chế chính cần được khắc phục. Thứ nhất, hầu hết các phương pháp chưa khai thác đầy đủ tri thức ngữ nghĩa kết hợp giữa thị giác và ngôn ngữ. Cụ thể, chúng chủ yếu dựa vào đặc trưng cục bộ từ các mô hình phát hiện đối tượng như Faster R-CNN hoặc đặc trưng toàn cục từ các mạng CNN huấn luyện trước, mà chưa tận dụng hiệu quả các biểu diễn cấu trúc ngữ nghĩa sâu hơn như Abstract Meaning Representation và đồ thị quan hệ giữa các đối tượng trong ảnh [72, 115]. Thứ hai, độ

tự nhiên và mạch lạc trong ngôn ngữ của các câu chú thích sinh ra vẫn còn hạn chế; một số câu chú thích có thể chứa lỗi ngữ pháp hoặc thiếu tính liên kết, cho thấy nhu cầu cấp thiết về một cơ chế hậu xử lý nhằm kiểm soát chất lượng ngôn ngữ ở mức cao hơn [31]. Thứ ba, vẫn còn thiếu một cơ chế hợp nhất linh hoạt và hiệu quả giữa các nguồn đặc trưng khác nhau, đặc biệt là trong việc kết hợp các biểu diễn toàn cục và cục bộ từ mô hình thị giác-ngôn ngữ huấn luyện trước như CLIP-ViT với tri thức ngữ nghĩa trừu tượng như AMR và thông tin cấu trúc từ đồ thị quan hệ [39].

Những hạn chế này đặt ra yêu cầu cấp thiết về một kiến trúc chú thích ảnh có khả năng tích hợp linh hoạt tri thức ngữ nghĩa đa nguồn, vừa tận dụng hiệu quả tri thức thị giác - ngôn ngữ huấn luyện trước, vừa khai thác thông tin cấu trúc và trừu tượng để sinh ra các mô tả tự nhiên, chính xác và ngữ nghĩa hơn.

Để giải quyết các thách thức nêu trên, trong nghiên cứu này, một mô hình chú thích ảnh mới được đề xuất- gọi là CLIP-AMR-GPT - kết hợp các đặc trưng thị giác-ngôn ngữ từ mô hình CLIP, biểu diễn cấu trúc từ đồ thị quan hệ, ngữ nghĩa trừu tượng từ AMR, và tri thức ngôn ngữ từ mô hình ngôn ngữ lớn GPT thông qua cơ chế Re-ranking. Mô hình được thiết kế theo kiến trúc encoder-decoder, trong đó encoder tích hợp ba nguồn tri thức: đặc trưng CLIP, embedding từ đồ thị quan hệ, và embedding từ đồ thị AMR của chú thích chuẩn; còn decoder sử dụng Transformer với cơ chế adaptive attention để tích hợp embedding AMR-like được tạo từ đồ thị quan hệ của ảnh đầu vào. Đầu ra của decoder được cải thiện thêm bởi một bước Re-ranking sử dụng GPT để tăng tính tự nhiên và ngữ nghĩa của chú thích. Các thực nghiệm được thực hiện trên hai tập dữ liệu chuẩn là MS COCO và Flickr30K với các độ đo phổ biến trong bài toán chú thích ảnh. Kết quả cho thấy mô hình đề xuất vượt trội hơn so với các phương pháp công bố gần đây, bao gồm cả các mô hình dựa trên Transformer, mô hình tích hợp CLIP, và mô hình sử dụng biểu diễn AMR. Đặc biệt, phần phân tích ảnh hưởng của các thành phần (ablation study) đã minh chứng cho vai trò quan trọng và đóng góp riêng biệt của từng thành phần trong kiến trúc đề xuất.

Đóng góp chính của chương này gồm:

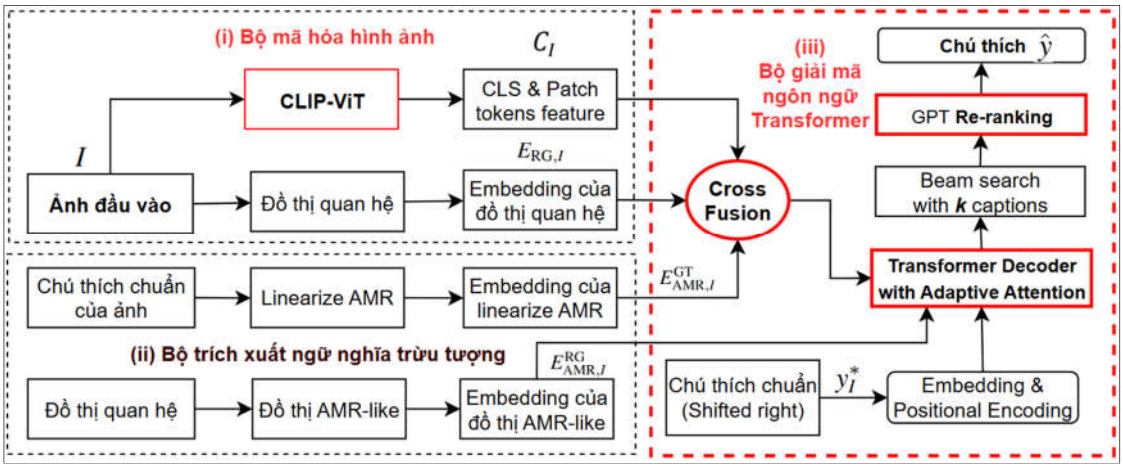
- **Đề xuất khối hợp nhất tri thức Cross-Fusion đa nguồn mới**: Nghiên cứu đề xuất một kiến trúc *Cross-Fusion* mới, cho phép kết hợp đồng thời ba loại đặc trưng ngữ nghĩa: (i) biểu diễn hình ảnh từ mô hình CLIP, (ii) embedding từ đồ thị quan hệ, và (iii) embedding AMR trích xuất từ chú thích chuẩn. Thiết kế này giúp mô hình tận dụng hiệu quả tri thức thị giác-ngôn ngữ cũng như cấu trúc quan hệ giữa các đối tượng để cung cấp biểu diễn ngữ nghĩa phong phú cho giai đoạn giải mã.

- **Tích hợp embedding AMR-like thông qua cơ chế Adaptive Attention**: Khác với các nghiên cứu trước vốn tích hợp toàn bộ embedding đồ thị vào mỗi bước sinh chú thích, nghiên cứu này đề xuất cơ chế *Adaptive Attention* để điều tiết mức độ ảnh hưởng của từng đỉnh trong đồ thị AMR-like một cách linh hoạt tại mỗi thời điểm sinh từ. Thiết kế này cho phép mô hình khai thác hiệu quả thông tin cấu trúc khi cần thiết, đồng thời hạn chế ảnh hưởng tiêu cực từ nhiễu ngữ nghĩa không liên quan.

- **Đưa ra mô-đun re-ranking dựa trên GPT để tối ưu chất lượng chú thích phát sinh**: Một mô-đun hậu xử lý sử dụng mô hình ngôn ngữ lớn GPT-2 được xây dựng nhằm đánh giá và xếp hạng các câu mô tả ứng viên theo xác suất ngôn ngữ (GPTScore). Câu có điểm cao nhất được chọn làm đầu ra cuối cùng, góp phần nâng cao độ trôi chảy, tính ngữ pháp và mức độ tự nhiên của chú thích.

- **Xây dựng và thực nghiệm mô hình chú thích ảnh tích hợp tri thức đa nguồn**: Mô hình được huấn luyện và đánh giá trên hai tập dữ liệu chuẩn MS COCO và Flickr30K, cho kết quả vượt trội so với các phương pháp trước đó, khẳng định hiệu quả của chiến lược tích hợp tri thức đa nguồn trong cải thiện độ chính xác và tính ngữ nghĩa của chú thích.

5.2. Phương pháp chú thích ảnh đề xuất



Hình 5.1. Kiến trúc tổng quát của mô hình đề xuất. Mô hình gồm 3 khối: (i) Bộ mã hóa hình ảnh - *đặc trưng trích xuất từ CLIP-ViT và embedding đồ thị quan hệ; (ii)* Bộ trích xuất ngữ nghĩa trừu tượng - *Embedding AMR của chú thích chuẩn và AMR-like của đồ thị quan hệ; (iii)* Bộ giải mã ngôn ngữ Transformer - *Cross Fusion hợp nhất các nguồn đặc trưng, giải mã với Adaptive Attention, sau đó beam-search và GPT re-ranking chọn chú thích cuối cùng.*

Phần này trình bày kiến trúc mô hình đề xuất **CLIP-AMR-GPT**, được thiết kế theo hướng hợp nhất tri thức đa nguồn nhằm nâng cao chất lượng chú thích ảnh. Mô hình tuân theo kiến trúc encoder-decoder và kế thừa có chọn lọc từ các mô hình trước trong luận án, gồm ba thành phần chính: (i) Bộ mã hóa hình ảnh (Image

Encoder), khai thác đặc trưng thị giác từ CLIP kết hợp biểu diễn đồ thị quan hệ giữa các đối tượng kế thừa từ mô hình **OD-VR-Cap** (Chương 2); (ii) Bộ trích xuất ngữ nghĩa trừu tượng (Abstract Semantic Extractor), kế thừa từ mô hình **AMR-GT&RG** (Chương 4), thực hiện trích xuất và ánh xạ ngữ nghĩa trừu tượng AMR từ chú thích và ảnh; và (iii) Bộ giải mã ngôn ngữ Transformer (Transformer Language Decoder), tích hợp đặc trưng đa nguồn thông qua cơ chế chú ý thích ứng và tái xếp hạng chú thích bằng GPT. Kiến trúc tổng thể của mô hình được minh họa trong Hình 5.1, trong khi chi tiết từng thành phần được trình bày tại các Mục 5.2.1-5.2.3.

5.2.1. Bộ mã hóa hình ảnh

Bộ mã hóa hình ảnh đảm nhiệm vai trò trích xuất hai loại đặc trưng quan trọng từ ảnh đầu vào: (i) đặc trưng ngôn ngữ-thị giác từ mô hình CLIP-ViT, và (ii) embedding cấu trúc từ đồ thị quan hệ giữa các đối tượng. Hai loại đặc trưng này là cơ sở để bổ sung tri thức nền phong phú cho bộ giải mã ở giai đoạn sau.

5.2.1.1. Trích xuất đặc trưng thị giác - ngôn ngữ từ CLIP

Trong mô hình chú thích ảnh đề xuất, mô hình CLIP (Contrastive Language-Image Pretraining) đóng vai trò là một thành phần quan trọng trong việc trích xuất đặc trưng ngôn ngữ - thị giác từ ảnh đầu vào. CLIP là một mô hình học sâu đa phương thức do OpenAI phát triển, được huấn luyện trên một tập dữ liệu lớn gồm hơn 400 triệu cặp ảnh-văn bản. Quá trình huấn luyện dựa trên mục tiêu tương phản (contrastive learning) nhằm học một không gian biểu diễn chung cho ngôn ngữ và hình ảnh, trong đó các cặp tương ứng được đưa lại gần nhau, còn các cặp không tương ứng bị đẩy ra xa.

Trong phần này, ảnh đầu vào I được xử lý bởi backbone thị giác của CLIP với kiến trúc Vision Transformer (ViT-L/14). Cụ thể, ảnh được chia thành các patch có kích thước 14×14, kết quả thu được 256 vùng không chồng lấn khi áp dụng cho ảnh chuẩn hóa kích thước 224×224. Sau đó, một token đặc biệt $[CLS]$ được thêm vào đầu chuỗi nhằm tóm lược thông tin tổng thể của ảnh. Kết quả đầu ra của CLIP gồm:

- Một véc-tơ toàn cục $G_I \in R^{768}$ tương ứng với token $[CLS]$, đại diện cho đặc trưng tổng quát của toàn bộ ảnh;
- Một tập véc-tơ patch tokens $P_I = \{p_1, p_2, \dots, p_{256}\}$, trong đó mỗi véc-tơ $p_i \in R^{768}$ mã hóa thông tin ngữ nghĩa của một vùng ảnh cụ thể.

Tập đặc trưng $M_I = G_I \cup P_I$ sau đó được đưa vào một nhánh Cross-Fusion trong bộ giải mã Transformer nhằm truyền tải tri thức thị giác-ngôn ngữ từ mô hình CLIP đến quá trình sinh câu mô tả. Việc sử dụng đồng thời cả véc-tơ toàn cục và patch tokens cho phép mô hình tận dụng đầy đủ thông tin từ cấp độ tổng thể đến chi

tiết, góp phần tăng cường khả năng biểu đạt ngữ nghĩa và cấu trúc khi phát sinh chú thích ảnh.

5.2.1.2. Tạo và embedding đồ thị quan hệ của hình ảnh

Trong mô hình CLIP-AMR-GPT, thông tin cấu trúc giữa các đối tượng trong ảnh tiếp tục được biểu diễn dưới dạng đồ thị quan hệ (R-Graph), kế thừa trực tiếp phương pháp đã trình bày ở Chương 2 (OD-VR-Cap) và được sử dụng lại trong Chương 3 (RGTranCNet) và Chương 4 (AMR-GT&RG). Cụ thể, các đối tượng được phát hiện từ ảnh đầu vào và kết nối thông qua các quan hệ ngữ nghĩa được dự đoán từ tập Visual Genome, tạo thành một đồ thị R-Graph. Biểu diễn của đồ thị này được tính toán bằng GraphSAGE với lan truyền hai chiều nhằm thu được embedding giàu ngữ nghĩa cho từng đỉnh. Tập hợp embedding của các đỉnh tạo thành biểu diễn $E_{RG,I}$, được sử dụng như một trong ba nguồn tri thức chính trong cơ chế Cross-Fusion của bộ giải mã.

Cách tiếp cận này đã được chứng minh là hiệu quả trong việc mô hình hóa cấu trúc quan hệ ngữ nghĩa giữa các đối tượng trong ảnh. Trong Chương 5, đồ thị quan hệ tiếp tục được khai thác như một nguồn tri thức bổ sung bên cạnh đặc trưng CLIP và embedding AMR nhằm nâng cao chất lượng và chiều sâu ngữ nghĩa của các mô tả được sinh ra.

5.2.2. Bộ trích xuất ngữ nghĩa trừu tượng

Để tăng cường khả năng hiểu ngữ nghĩa của mô hình, nghiên cứu này kế thừa phương pháp trích xuất ngữ nghĩa trừu tượng đã được trình bày trong Chương 4 (mô hình AMR-GT&RG), trong đó sử dụng Abstract Meaning Representation (AMR) nhằm thu nhận các khái niệm ngữ nghĩa trừu tượng từ hai nguồn: (i) các câu chú thích chuẩn và (ii) đồ thị quan hệ của ảnh. Cụ thể, các câu chú thích chuẩn được chuyển đổi thành đồ thị AMR thông qua mô hình NeuralAMR, sau đó tuyến tính hóa theo định dạng PENMAN và trích xuất embedding bằng mô hình ngôn ngữ huấn luyện trước (BERT). Kết quả thu được là véc-tơ biểu diễn ngữ nghĩa trừu tượng $E^{GT}_{AMR,I}$. Song song đó, đồ thị quan hệ giữa các đối tượng trong ảnh được ánh xạ sang cấu trúc AMR-like bằng cách gán nhãn khái niệm và quan hệ ngữ nghĩa theo chuẩn AMR. Biểu diễn này được embedding bằng GraphSAGE để thu được véc-tơ $E^{RG}_{AMR,I}$.

Trong kiến trúc CLIP-AMR-GPT, hai nguồn embedding này được tích hợp với vai trò bổ sung: $E^{GT}_{AMR,I}$ được đưa vào tầng hợp nhất tri thức (Fusion Layer) nhằm cung cấp nguồn tri thức ngữ nghĩa chuẩn từ dữ liệu huấn luyện, trong khi $E^{RG}_{AMR,I}$ được tích hợp vào tầng Masked Multi-Head Attention của Transformer Decoder thông qua

cơ chế Adaptive Attention, cho phép điều tiết ảnh hưởng của thông tin cấu trúc một cách linh hoạt trong quá trình sinh chú thích.

5.2.3. Bộ giải mã ngôn ngữ Transformer

Bộ giải mã ngôn ngữ đóng vai trò then chốt trong quá trình sinh chú thích từ các biểu diễn đặc trưng trích xuất từ ảnh, đồ thị quan hệ, và khung ngữ nghĩa trừu tượng. Trong mô hình đề xuất, bộ giải mã được xây dựng dựa trên kiến trúc Transformer decoder với hai cơ chế cải tiến chính nhằm khai thác hiệu quả các nguồn tri thức đầu vào: (i) Cơ chế Cross Fusion Attention kết hợp trực tiếp toàn bộ véc-tơ đặc trưng từ ba miền (thị giác-ngôn ngữ từ CLIP, cấu trúc quan hệ từ đồ thị quan hệ, và ngữ nghĩa trừu tượng AMR từ chú thích chuẩn), và (ii) Cơ chế Adaptive Attention điều khiển mức độ đóng góp của embedding AMR-like trong mỗi bước giải mã.

5.2.3.1. Cơ chế chú ý Cross Fusion giữa ba nguồn đặc trưng

Thay vì tính trung bình các véc-tơ đặc trưng như một số công trình khác, phương pháp đề xuất giữ nguyên tất cả các véc-tơ thành phần của ba miền đầu vào nhằm bảo toàn thông tin ngữ nghĩa ở mức độ chi tiết. Cụ thể:

- Đặc trưng CLIP gồm véc-tơ $[CLS]$ và 256 patch tokens được biểu diễn dưới dạng ma trận M_I.
- Embedding đồ thị quan hệ của ảnh, $E_{RG,I}$;
- Embedding của AMR từ chú thích chuẩn, $E^{GT}_{AMR,I}$.

Ba nguồn đặc trưng này được chiếu tuyến tính sang không gian *Key-Value* riêng biệt rồi hợp nhất lại để tạo thành tập khóa K và giá trị V cho lớp Cross-Attention:

$$K_{multi} = \left[C_I W_K^C; E_{RG,I} W_K^{RG}; E^{GT}_{AMR,I} W_K^{GT}\right], V_{multi} = \left[C_I W_V^C; E_{RG,I} W_V^{RG}; E^{GT}_{AMR,I} W_V^{GT}\right]$$

Với mỗi bước giải mã, véc-tơ truy vấn Q được tạo từ token hiện tại trong quá trình phát sinh, véc-tơ đầu ra được tính thông qua cơ chế Multi-Head Attention:

$$f_{fused} = MultiHeadAttention(Q, K_{multi}, V_{multi})$$

Cách kết hợp này đảm bảo mô hình có thể học được các trọng số tương ứng cho từng thành phần đặc trưng, từ đó khai thác đầy đủ thông tin thị giác-ngôn ngữ, cấu trúc đồ thị và biểu diễn nghĩa trừu tượng mà không mất mát thông tin do phép trung bình.

5.2.3.2. Cơ chế chú ý Adaptive cho embedding đồ thị AMR-like

Để tăng cường khả năng tích hợp thông tin ngữ nghĩa cấu trúc từ đồ thị AMR-like vào quá trình sinh mô tả, mô hình đề xuất một cơ chế Adaptive Attention, được thực hiện tại tầng Masked Multi-Head Attention trong Transformer Decoder. Khác

với các phương pháp tích hợp cố định toàn bộ embedding tại mọi bước sinh từ, phương pháp này cho phép mô hình tự động học cách lựa chọn mức độ đóng góp của từng nút trong đồ thị tại mỗi thời điểm sinh token.

Cụ thể, giả sử tại thời điểm t, véc-tơ trạng thái ẩn trước đó của decoder là h_{t-1}. Mô hình sử dụng véc-tơ này để tính toán một véc-tơ cổng điều khiển g_t thông qua một hàm sigmoid:

$$g_t = \sigma(W_g \cdot h_{t-1} + b_g) \quad (5.1)$$

Trong đó, W_g là ma trận trọng số học được, và σ là hàm sigmoid áp dụng theo từng chiều.

Giả sử embedding của đồ thị AMR-like được biểu diễn bằng tập véc-tơ $E^{RG}_{AMR,I} = [e_1, e_2, \dots, e_k]$, với mỗi e_i là embedding của một đỉnh trong đồ thị. Cơ chế adaptive attention điều chỉnh từng véc-tơ e_i theo véc-tơ điều khiển g_t như sau:

$$\tilde{e}_1^{(t)} = g_t \odot e_i, \text{ với } i = 1, 2, \dots k \quad (5.2)$$

Toàn bộ tập véc-tơ điều chỉnh $\tilde{E}^{RG(t)}_{AMR,I} = \left[\tilde{e}_1^{(t)}, \dots, \tilde{e}_k^{(t)}\right]$ được sử dụng làm Key và $Value$ trong lớp Masked Multi-Head Attention tại thời điểm t, trong khi h_{t-1} được sử dụng làm $Query$.

Cơ chế này giúp mô hình học được cách chọn lọc linh hoạt các đỉnh trong đồ thị AMR-like tùy theo ngữ cảnh đang sinh từ, thay vì sử dụng đồng đều hoặc toàn bộ embedding như các phương pháp trước đây. Nhờ đó, mô hình có thể tập trung vào các khía cạnh ngữ nghĩa trừu tượng phù hợp với nội dung đang được tạo ra, cải thiện cả tính mạch lạc cú pháp và chiều sâu ngữ nghĩa của câu mô tả sinh ra.

Cơ chế này không chỉ giúp giữ nguyên cấu trúc đầy đủ của đồ thị, mà còn tạo điều kiện cho Decoder tự học chiến lược tham chiếu phù hợp với nội dung ngữ nghĩa đang được phát sinh. Việc này góp phần nâng cao tính linh hoạt và khả năng biểu đạt của mô hình khi sinh ra các câu chú thích giàu ngữ nghĩa hơn.

5.2.3.3. Tái xếp hạng qua GPT

Sau khi mô hình sinh ngôn ngữ tạo ra một tập các câu chú thích ứng viên cho một ảnh đầu vào, chất lượng cuối cùng của mô hình chú thích ảnh phụ thuộc đáng kể vào khả năng lựa chọn được chú thích phù hợp và tự nhiên nhất. Nhằm tăng cường tính mạch lạc ngữ nghĩa và độ trôi chảy của chú thích, nghiên cứu này đề xuất tích hợp một bước tái xếp hạng (re-ranking) sử dụng mô hình ngôn ngữ lớn GPT, đóng vai trò như một bộ đánh giá ngôn ngữ tổng quát (language scorer).

Cụ thể, sau khi mô hình decoder sinh ra một tập các câu chú thích ứng viên $Y_I = \{\hat{y}_I^{(1)}, \hat{y}_I^{(2)}, \dots \hat{y}_I^{(K)}\}$ thông qua thuật toán beam search kích thước K, mỗi câu $\hat{y}_I^{(k)}$ được đánh giá bởi mô hình GPT để tính điểm ngữ nghĩa theo công thức:

$$GPTScore(\hat{y}_I^{(k)}) = \frac{1}{|\hat{y}_I^{(k)}|} \sum_{t=1}^{|\hat{y}_I^{(k)}|} log\, P_{GPT}(\hat{y}_{I,t}^{(k)} | \hat{y}_{I,<t}^{(k)}) \quad (5.3)$$

Trong đó: $\hat{y}_I^{(k)}$ là token thứ t trong câu ứng viên thứ k, $|\hat{y}_I^{(k)}|$ là độ dài của chú thích $\hat{y}_I^{(k)}$, $P_{GPT}(.|.)$ là xác suất có điều kiện được tính từ mô hình GPT đã huấn luyện.

Chú thích cuối cùng $\hat{y}_I$ được chọn là chú thích có GPTScore cao nhất:

$$\hat{y}_I = argmax_{\hat{y}_I \in Y_I} GPTScore(\hat{y}) \quad (5.4)$$

Việc sử dụng mô hình ngôn ngữ huấn luyện trước như GPT trong giai đoạn suy luận (inference) giúp tăng độ tự nhiên và khả năng biểu đạt ngữ nghĩa mà không làm thay đổi quá trình huấn luyện chính của mô hình phát sinh mô tả.

5.2.3.4. Thuật toán huấn luyện và phát sinh chú thích

Quá trình huấn luyện được mô tả trong **Thuật toán 5.1**, với đầu vào là cặp ảnh và chú thích chuẩn. Mô hình sử dụng cơ chế teacher forcing để cập nhật trọng số qua hàm mất mát cross-entropy. Sau khi huấn luyện, quá trình suy luận được thực hiện theo **Thuật toán 5.2**, sử dụng beam search để sinh ra nhiều câu chú thích ứng viên. Sau đó, bước re-ranking bằng mô hình GPT lựa chọn câu có giá trị GPTScore cao nhất làm đầu ra cuối cùng.

Thuật toán 5.1 trình bày quy trình huấn luyện mô hình sinh chú thích ảnh dựa trên kiến trúc Transformer Decoder kết hợp với cơ chế tích hợp tri thức ngữ nghĩa. Đầu vào của thuật toán là tập dữ liệu D, trong đó mỗi mẫu dữ liệu bao gồm các đặc trưng hình ảnh X_I và chú thích chuẩn y_I^*. Tại mỗi epoch huấn luyện, tập dữ liệu được xáo trộn ngẫu nhiên để tăng khả năng tổng quát hóa của mô hình. Với mỗi mẫu, mô hình tiến hành sinh tuần tự từng token trong câu chú thích bằng cách khởi tạo chuỗi với token bắt đầu ($start$). Tại mỗi bước thời gian t, chuỗi đầu vào hiện tại y_t được ánh xạ thành véc-tơ nhúng H_{init}. Sau đó, biểu diễn này được điều chỉnh thông qua mô-đun AdaptiveAttention, cho phép mô hình học cách tham chiếu linh hoạt đến các véc-tơ embedding từ đồ thị AMR-like $E_{AMR,I}^{RG}$. Tiếp theo, các đặc trưng ngữ nghĩa từ ảnh - bao gồm đặc trưng từ CLIP M_I, đồ thị quan hệ $E_{RG,I}$ và AMR từ chú thích chuẩn $E_{AMR,I}^{GT}$ - được kết hợp cùng H_{adp} thông qua mô-đun CrossFusion, tạo ra đầu vào tổng hợp H_{cross} cho tầng Feed-Forward Network. Kết quả từ tầng FFN được dùng để tính phân phối xác suất trên không gian từ vựng thông qua hàm $softmax$. Mất mát huấn

luyện được tích lũy theo hàm log-likelihood giữa từ mục tiêu y_t^* và xác suất dự đoán $P(y_t)$. Sau khi hoàn thành việc sinh toàn bộ chuỗi, tham số mô hình φ được cập nhật bằng phương pháp gradient descent với tốc độ học η.

Thuật toán 5.1. TrainCaptioningModel (D, N, φ, η)

Đầu vào: Tập dữ liệu $D = \{(X_I, y_I^*)\}$, $X_I = \left(E_{RG,I}, E_{AMR,I}^{GT}, E_{AMR,I}^{RG}, M_I\right)$ và y_I^* là chú thích cho trước; số epoch N_T, hệ số học η; tham số ban đầu φ.
Đầu ra: Tham số đã huấn luyện φ.

Begin

Khởi tạo φ
for $epoch = 1$ to N_T **do**
  shuffle(D)
  for all $(X_I, y^*) \in D$ **do**
    Decompose X_I into $\left(E_{RG,I}, E_{AMR,I}^{GT}, E_{AMR,I}^{RG}, M_I\right)$
    $y_{<1} \leftarrow start$, $L \leftarrow 0$
    for $i = 1$ to $|y^*|$ **do**
      $H_{init} \leftarrow Embedding(y_{<t})$
      $H_{adp} \leftarrow AdaptiveAttention\left(H_{init}, E_{AMR,I}^{GT}, \varphi\right)$
      $H_{cross} \leftarrow CrossFusion\left(H_{adp}, \{C_I, E_{RG,I}, E_{AMR,I}^{GT}\}, \varphi\right)$
      $H_{final} \leftarrow FFN(H_{cross}, \varphi)$
      $P(y_t) = softmax\left(WH_{final} + b\right)$
      $L \leftarrow L - logP(y_t^*|y_{<t}, X_I, \varphi)$
      $y_{<t+1} \leftarrow y_{<t} \cup \{y_t^*\}$
    endfor
    $\varphi \leftarrow \varphi - \eta \nabla_\varphi L$
  endfor
endfor
return φ

End

Thuật toán 5.2 thực hiện suy luận để sinh chú thích ảnh tối ưu. Với mỗi ảnh đầu vào, mô hình sinh ra k chú thích ứng viên bằng cơ chế giải mã tuần tự: tại mỗi bước, embedding chuỗi hiện tại được xử lý qua Adaptive Attention (tích hợp embedding AMR-like), sau đó Cross-Fusion (kết hợp đặc trưng CLIP và embedding đồ thị quan hệ), rồi qua Feed-Forward Network để dự đoán token tiếp theo. Quá trình lặp lại cho đến khi đạt T_{max} hoặc gặp token <end>. Sau khi tạo xong tập k chú thích ứng viên, mô hình GPT-2 được dùng để đánh giá và xếp hạng dựa trên điểm ngôn ngữ (GPTScore). Chú thích có điểm cao nhất được chọn làm đầu ra cuối cùng, đảm bảo vừa chính xác ngữ nghĩa vừa tự nhiên về ngôn ngữ.

Thuật toán 5.2. GenerateCaptionWithGPTReRanking($X_I, \varphi, \mathcal{M}, k$)

Đầu vào: $X_I = \left(E_{RG,I}, E_{AMR,I}^{RG}, M_I\right)$, mô hình đã huấn luyện φ, độ dài tối đa của câu chú thích T_{max}, mô hình GPT sử dụng $\mathcal{M}$, số câu ứng viên k.

Đầu ra: Chú thích được sinh ra $\hat{y}$ cho ảnh đầu vào I (câu chú thích sau khi re-ranking).

Begin

$Y \leftarrow \emptyset$

for $i = 1$ **to** k **do**

  $y_{<1}^{(i)} \leftarrow \{start\}, t \leftarrow 1$

  while $t < T_{max}$ and $y_t \neq$ **<***end***>** **do**

    $H_{init}^{(t)} \leftarrow Embedding\left(y_{<1}^{(i)}\right)$

    $H_{adp}^{(t)} \leftarrow AdaptiveAttention\left(H_{init}^{(t)}, E_{AMR,I}^{GT}, \varphi\right)$

    $H_{cross}^{(t)} \leftarrow CrossFusion\left(H_{adp}^{(t)}, \{M_I, E_{RG,I}\}, \varphi\right)$

    $H_{final}^{(t)} \leftarrow FFN\left(H_{cross}^{(t)}, \varphi\right)$

    $P^{(t)} \leftarrow softmax\left(WH_{final}^{(t)} + b\right)$

    $y_t^{(i)} \sim P^{(t)}$

    $y_{<t+1}^{(i)} \leftarrow y_{<t}^{(i)} \cup \left\{y_t^{(i)}\right\}$

    $t \leftarrow t + 1$

  endwhile

  $Y \leftarrow Y \cup \left\{y^{(i)} = y_{<t}^{(i)}\right\}$

endfor

$S \leftarrow GPT\_ReRanking(Y, M)$

$\hat{y} \leftarrow argmax_{i \in \{1,\ldots,k\}} S^{(i)}$

return $\hat{y}$

End

5.3. Thực nghiệm và kết quả

Phần này trình bày quy trình thiết lập thực nghiệm và kết quả đánh giá hiệu quả của mô hình đề xuất trên hai tập dữ liệu chuẩn trong bài toán chú thích ảnh. Các nội dung bao gồm: mô tả dữ liệu sử dụng, các độ đo đánh giá, chi tiết cấu hình huấn luyện, so sánh với các phương pháp hiện tại và phân tích thành phần mô hình nhằm làm rõ vai trò đóng góp của từng thành phần trong kiến trúc đề xuất.

5.3.1. Thiết lập thực nghiệm và đánh giá

Để đánh giá một cách khách quan và toàn diện, các thí nghiệm được tiến hành trên hai tập dữ liệu phổ biến là MS COCO và Flickr30K, sử dụng các độ đo đánh giá tiêu chuẩn. Ngoài ra, cấu hình huấn luyện được giữ cố định giữa các mô hình nhằm đảm bảo tính nhất quán và độ tin cậy trong so sánh. Mục này trình bày chi tiết các yếu tố liên quan đến dữ liệu, độ đo đánh giá và thiết lập cài đặt mô hình.

5.3.1.1. Dữ liệu thực nghiệm

Trong nghiên cứu này, hai bộ dữ liệu chuẩn MS COCO và Flickr30K tiếp tục được sử dụng để đánh giá hiệu quả của mô hình CLIP-AMR-GPT. Đặc điểm chi tiết của hai bộ dữ liệu này, bao gồm quy mô, cấu hình chia tập và cách thức tiền xử lý, đã được trình bày trong Mục 4.4.1 của Chương 4.

Việc sử dụng đồng thời cả hai tập dữ liệu nhằm kiểm chứng khả năng khái quát hóa và mức độ thích ứng của mô hình trên các tập có quy mô và độ phức tạp khác nhau. Trong phạm vi Chương 5, các thiết lập chia tập và tiền xử lý được giữ nguyên như trong Chương 4 để đảm bảo tính thống nhất và khả năng so sánh trực tiếp với các nghiên cứu trước đây.

5.3.1.2. Độ đo đánh giá

Các độ đo đánh giá được sử dụng trong nghiên cứu này bao gồm BLEU, METEOR, ROUGE-L, CIDEr, và SPICE, vốn là những độ đo chuẩn mực trong bài toán chú thích ảnh. Bên cạnh đó, nghiên cứu này còn tiếp tục sử dụng SCS - một độ đo đã được đề xuất và trình bày chi tiết trong Mục 4.3.3 của Chương 4 - nhằm đánh giá mức độ nhất quán về ngữ nghĩa giữa chú thích sinh ra và nội dung ảnh.

Việc kết hợp đồng thời các độ đo truyền thống với SCS cho phép đánh giá toàn diện hơn hiệu quả của mô hình CLIP-AMR-GPT, không chỉ ở khía cạnh chính xác về cú pháp và *n-gram*, mà còn ở khả năng duy trì tính mạch lạc và nhất quán ngữ nghĩa trong các câu mô tả.

5.3.1.3. Chi tiết cài đặt

Mô hình thực nghiệm của phương pháp đề xuất được cài đặt bằng ngôn ngữ lập trình Python (*3.9*) và framework học sâu Pytorch (*2.0*), thực thi trên Google Colab Pro với cấu hình và tham số cụ thể như sau:

- Phần tạo và embedding đồ thị quan hệ: thực hiện theo thiết lập trong nghiên cứu OD-VR-Cap (Chương 2).
- Phần Embedding của AMR tạo từ chú thích chuẩn và AMR-like tạo từ đồ thị quan hệ được thực hiện theo thiết lập trong nghiên cứu AMR-GT&RG (Chương 4).
- CLIP-ViT: Sử dụng phiên bản ViT-L/14 từ mô hình OpenAI CLIP, trích xuất token toàn cục [CLS] và 256 patch tokens từ ảnh đầu vào.
- Phần Transformer Decoder: với số block $N = 6$, số $head = 8$, số chiều cho véc-tơ biểu diễn từ là 512, sử dụng bộ tối ưu Adam, hệ số học là 0.00004, batch size là 32.

▪ Re-ranking: sau khi sinh ra 5 chú thích ứng viên cho mỗi ảnh bằng beam search (beam width = 5), một mô hình GPT-2 medium (345M tham số, từ thư viện Huggingface) được sử dụng để đánh giá xác suất ngôn ngữ (GPTScore). Chú thích có log-probability cao nhất được chọn làm đầu ra cuối cùng.

5.3.2. Chi phí tính toán và thời gian thực hiện

Toàn bộ quá trình huấn luyện được thực hiện trên nền tảng Google Colab Pro, sử dụng GPU NVIDIA Tesla T4 (16 GB VRAM). Trong chương này, mô hình CLIP được sử dụng như một bộ trích xuất đặc trưng thị giác-ngôn ngữ huấn luyện trước và được giữ cố định trong suốt quá trình huấn luyện. Khác với các chương trước, mô hình không sử dụng đặc trưng vùng đối tượng (region-level features), qua đó loại bỏ chi phí liên quan đến phát hiện đối tượng và trích xuất đặc trưng vùng.

Chi phí huấn luyện của mô hình vì vậy tập trung chủ yếu vào các mô-đun được đề xuất trong Chương 5, bao gồm cơ chế hợp nhất đa nguồn đặc trưng, Transformer decoder, và Adaptive Attention trên embedding AMR-like. Mô hình được huấn luyện trong khoảng 24 giờ trên tập MS COCO và 12 giờ trên tập Flickr30K, phản ánh chi phí tăng lên do việc tích hợp đồng thời biểu diễn CLIP, đồ thị quan hệ và ngữ nghĩa trừu tượng AMR, thay vì do mở rộng các mô-đun xử lý thị giác cục bộ.

Trong giai đoạn suy luận, mô hình áp dụng beam search với beam size = 5 để sinh tập các câu chú thích ứng viên, sau đó sử dụng GPT Re-ranking nhằm lựa chọn câu mô tả có tính tự nhiên và phù hợp ngữ cảnh cao nhất. Với batch size = 32, thời gian suy luận trung bình đạt khoảng 0.15 giây/ảnh, cho thấy chi phí phát sinh từ bước re-ranking được kiểm soát hiệu quả do chỉ áp dụng trên một tập nhỏ các câu ứng viên.

Nhìn chung, mặc dù Chương 5 giới thiệu thêm các thành phần mới như CLIP và GPT, việc loại bỏ đặc trưng vùng đối tượng giúp cân bằng chi phí tính toán, cho phép mô hình CLIP-AMR-GPT duy trì khả năng triển khai thực tế trên hạ tầng GPU phổ thông trong khi đạt hiệu năng cao về mặt ngữ nghĩa và chất lượng chú thích.

5.3.3. Kết quả và bàn luận

Phần này tổng hợp và phân tích các kết quả thực nghiệm đạt được bởi mô hình đề xuất khi so sánh với các phương pháp chú thích ảnh hiện có. Ngoài ra, một phân tích thành phần (ablation study) được tiến hành nhằm đánh giá ảnh hưởng của từng thành phần chính trong kiến trúc mô hình. Việc kết hợp các đặc trưng đa nguồn như CLIP, đồ thị quan hệ, AMR và cơ chế re-ranking bằng GPT cho thấy những cải tiến đáng kể về độ chính xác và tính ngữ nghĩa của câu mô tả sinh ra.

5.3.3.1. So sánh với các công trình công bố gần đây

Bảng 5.1 trình bày kết quả so sánh giữa mô hình đề xuất **CLIP-AMR-GPT** với các phương pháp chú thích ảnh tiên tiến trên tập kiểm tra Karpathy của MSCOCO, sử dụng các độ đo chuẩn gồm BLEU-1 (B@1), BLEU-4 (B@4), METEOR (M), ROUGE-L (R), CIDEr (C), SPICE (S) và SCS. Các mô hình đối sánh được chọn bao gồm các mô hình Transformer hiện đại như *X-Transformer*, các mô hình tích hợp tri thức ngữ nghĩa như *COS-Net*, *MLA-LRN*, cũng như các phương pháp gần đây có liên quan như *GIC-SSF*, *RGTranCNet*, *AMR-GT&RG* và *CLIP-Captioner*.

Bảng 5.1. So sánh độ chính xác của các phương pháp chú thích ảnh đề xuất trên tập kiểm tra MSCOCO Karpathy. Giá trị in đậm biểu thị điểm số cao nhất trong mỗi cột độ đo.

Phương pháp	B@1	B@4	M	R	C	S	SCS
X-Transformer (2020) [107]	77.3	37.0	28.7	57.5	120.0	21.8	-
COS-Net (2022) [108]	78.8	39.1	29.5	58.7	127.4	22.8	-
MLA-LRN (2023) [114]	79.9	**42.4**	29.4	59.7	125.7	23.2	-
GIC-SSF (2025) [23]	81.6	40.1	29.7	59.5	135.4	23.2	-
CLIP-Captioner(2022) [68]	-	40.6	30.0	59.9	139.4	23.9	-
BCEFT (2025) [69]	76.5	35.2	27.4	56.0	128.6	-	-
OD-VR-Cap	72.6	28.1	24.8	53.4	85.1	-	75.8
RGTranCNet	79.8	36.3	35.6	57.2	107.8	-	82.3
AMR-GT&RG	81.2	39.5	37.2	59.9	136.7	25.1	89.1
CLIP-AMR-GPT	**82.9**	41.4	**38.5**	**61.8**	**144.2**	**26.7**	**91.7**

Kết quả thực nghiệm cho thấy mô hình CLIP-AMR-GPT đạt hiệu suất vượt trội trên hầu hết các độ đo. Cụ thể:

- Về độ chính xác n-gram, mô hình đạt B@1 = 82.9 và B@4 = 41.4, nằm trong nhóm điểm số cao nhất trong số các phương pháp được so sánh. Kết quả này cho thấy mô hình không chỉ tạo ra các từ vựng đúng mà còn tạo ra các cụm từ liên tục phù hợp với câu tham chiếu.

- Về độ đo ngữ nghĩa, mô hình đạt METEOR = 38.5 và SPICE = 26.7, đều vượt qua mô hình AMR-GT&RG (METEOR = 37.2, SPICE = 25.1) và CLIP-Captioner (METEOR = 30.0, SPICE = 23.9), cho thấy hiệu quả của việc tích hợp biểu diễn ngữ nghĩa AMR và đặc trưng thị giác-ngôn ngữ CLIP trong việc tạo ra mô tả ngôn ngữ tự nhiên giàu ngữ nghĩa và nhất quán với nội dung ảnh.

- Độ đo CIDEr - độ đo phản ánh sự đồng thuận với các mô tả của con người - đạt 144.2, cao nhất trong tất cả các mô hình, vượt qua cả CLIP-Captioner (139.4) và AMR-GT&RG (136.7), chứng tỏ mô hình sinh ra các chú thích gần sát nhất với nhận thức ngôn ngữ của con người.

▪ ROUGE-L đạt 61.8, cao hơn đáng kể so với các phương pháp khác, cho thấy mô hình không chỉ chú ý đến n-gram mà còn duy trì cấu trúc câu tốt hơn.

So với mô hình AMR-GT&RG - mô hình cơ sở gần nhất - mô hình đề xuất cải thiện đáng kể trên tất cả các độ đo, đặc biệt là các độ đo đánh giá chiều sâu ngữ nghĩa như METEOR, SPICE và CIDEr. Điều này xác nhận rằng việc bổ sung đặc trưng ngữ nghĩa từ CLIP, tích hợp biểu diễn AMR-like qua adaptive attention, và sử dụng GPT-based Re-ranking đã mang lại sự cải thiện toàn diện về chất lượng chú thích. Ngoài ra, khi so sánh với các phương pháp dùng Transformer như X-Transformer hay các phương pháp dùng tri thức như COS-Net, mô hình cũng thể hiện độ vượt trội rõ rệt, khẳng định tính hiệu quả của việc kết hợp tri thức thị giác và tri thức cấu trúc ngôn ngữ trong mô hình hóa đa phương thức.

Bảng 5.2. So sánh độ chính xác của các phương pháp chú thích ảnh đề xuất trên tập kiểm tra Flickr30K Karpathy

Phương pháp	B@1	B@4	M	R	C	S	SCS
DRL-Attention (2023) [47]	73.8	33.5	22.2	50.4	92.1	-	-
VisualCaptionNet (2024) [48]	74.8	-	20.7	52.8	-	-	-
MLA-LRN (2023) [114]	74.3	31.2	23.5	51.5	65.6	17.2	-
GIC-SSF (2025) [23]	79.9	38.1	27.3	55.0	80.6	-	-
AMR-GT&RG	79.1	36.4	35.6	56.7	94.5	22.7	87.2
CLIP-AMR-GPT	**80.5**	**38.2**	**36.9**	**58.2**	**102.8**	**24.0**	**89.4**

Tiếp nối kết quả trên tập MSCOCO, **Bảng 5.2** trình bày hiệu quả của mô hình CLIP-AMR-GPT trên tập Flickr30K nhằm kiểm chứng khả năng tổng quát hóa của phương pháp đề xuất trên một tập dữ liệu có quy mô nhỏ hơn và có tính đa dạng cao hơn về cách diễn đạt ngôn ngữ. Mô hình CLIP-AMR-GPT tiếp tục duy trì ưu thế vượt trội so với các phương pháp so sánh. Cụ thể, mô hình đạt BLEU-1 = 80.5 và BLEU-4 = 38.2, cải thiện nhẹ so với AMR-GT&RG (79.1; 36.4), cho thấy khả năng mô phỏng chính xác từ vựng và cấu trúc cụm từ trong câu mô tả. Đáng chú ý, độ đo METEOR = 36.9 và SPICE = 24.0 đều vượt trội so với các phương pháp gần nhất (AMR-GT&RG: 35.6 và 22.7), cho thấy mô hình không chỉ tái hiện nội dung thị giác mà còn sinh ra các mô tả giàu ngữ nghĩa và phù hợp ngôn ngữ tự nhiên. Độ đo CIDEr = 102.8 và ROUGE-L = 58.2 đạt mức cao nhất trong tất cả các phương pháp được so sánh, phản ánh mức độ đồng thuận cao giữa các câu sinh ra và tập chú thích chuẩn được gán bởi con người.

So với tập dữ liệu MS COCO, các mức cải thiện trên Flickr30K nhìn chung vẫn được duy trì một cách ổn định, cho thấy mô hình đề xuất không chỉ phù hợp với các tập dữ liệu quy mô lớn mà còn thích ứng hiệu quả với những tập có cấu trúc biểu đạt đa dạng hơn. Kết quả này khẳng định tính hiệu quả và khả năng tổng quát hóa của

mô hình trong nhiều bối cảnh ngôn ngữ khác nhau. Bên cạnh đó, mặc dù một số mô hình gần đây như *DRL-Attention*, *MLA-LRN*, hay *GIC-SSF* tuy đã đạt được những cải tiến nhất định, hiệu năng của chúng vẫn chưa đạt tới mức độ chính xác và ổn định như mô hình đề xuất.

Tổng thể, kết quả trên cả hai tập MSCOCO và Flickr30K cho thấy mô hình CLIP-AMR-GPT không chỉ cải thiện đáng kể độ chính xác mô tả hình ảnh, mà còn nâng cao khả năng biểu đạt ngữ nghĩa sâu sắc, nhờ vào sự kết hợp hiệu quả giữa đặc trưng thị giác-ngôn ngữ từ CLIP, tri thức cấu trúc từ đồ thị AMR-like, cơ chế adaptive attention và bước tái xếp hạng bằng GPT.

5.3.3.2. Đánh giá ảnh hưởng của từng thành phần trong mô hình đề xuất

Để đánh giá vai trò và mức độ đóng góp của từng thành phần trong kiến trúc đề xuất, phân tích thành phần (ablation study) được thực hiện trên hai tập dữ liệu MS COCO và Flickr30K. Cụ thể, ba biến thể của mô hình được xây dựng bằng cách loại bỏ lần lượt từng thành phần chính: (i) không sử dụng đặc trưng từ CLIP (*w/o CLIP features*), (ii) không áp dụng cơ chế adaptive attention cho embedding đồ thị AMR-like (*w/o adaptive attention*), và (iii) không sử dụng mô-đun GPT Re-ranking (*w/o GPT Re-ranking*). Các biến thể này được huấn luyện và đánh giá với cùng cấu hình như mô hình đầy đủ nhằm đảm bảo tính so sánh công bằng.

Bảng 5.3. Kết quả phân tích ảnh hưởng từng thành phần trên tập dữ liệu MSCOCO

Mô hình	**BLUE-4**	**METEOR**	**CIDEr**	**SPICE**
w/o CLIP features	39.1	36.7	134.5	25.2
w/o adaptive attention	39.8	37.3	137.1	25.7
w/o GPT Re-ranking	40.2	37.8	139.6	26.0
Proposed model (Full)	**41.4**	**38.5**	**144.2**	**26.7**

Kết quả trên tập dữ liệu MS COCO được trình bày trong **Bảng 5.3** cho thấy việc loại bỏ từng thành phần đều làm giảm hiệu suất của mô hình. Cụ thể, khi không sử dụng đặc trưng từ CLIP, điểm CIDEr giảm từ 144.2 xuống còn 134.5 và SPICE giảm từ 26.7 còn 25.2, phản ánh tầm quan trọng của thông tin ngữ nghĩa trực quan mà CLIP mang lại. Việc bỏ qua cơ chế adaptive attention cũng khiến mô hình kém linh hoạt hơn trong việc điều hướng đóng góp của từng nút trong đồ thị AMR-like tại mỗi bước sinh, dẫn đến giảm hiệu quả tổng thể (CIDEr: 137.1, SPICE: 25.7). Đặc biệt, không sử dụng GPT Re-ranking khiến kết quả giảm rõ rệt ở các độ đo đánh giá tính mạch lạc ngữ nghĩa (CIDEr: 139.6, SPICE: 26.0).

Tương tự, trên tập Flickr30K (**Bảng 5.4**), các xu hướng trên tiếp tục được duy trì. Việc loại bỏ đặc trưng từ CLIP làm điểm BLEU-4 giảm từ 38.2 xuống còn 36.2,

và CIDEr giảm 7.8 điểm. Mô hình thiếu adaptive attention ghi nhận kết quả CIDEr là 97.6, thấp hơn đáng kể so với mô hình đầy đủ (102.8). Tác động của việc không sử dụng GPT Re-ranking cũng đáng kể với SPICE giảm từ 24.0 xuống 23.5.

Đóng góp chính của mô hình CLIP-AMR-GPT nằm ở việc xây dựng một khung hợp nhất tri thức đa nguồn theo hướng bổ sung lẫn nhau giữa ba miền thông tin: thị giác, ngữ nghĩa cấu trúc và ngôn ngữ tự nhiên. Cụ thể, CLIP đảm nhiệm vai trò căn chỉnh thị giác-ngôn ngữ ở mức khái niệm, giúp mô hình nắm bắt chính xác các đối tượng và ngữ cảnh hình ảnh; AMR/AMR-like cung cấp cấu trúc ngữ nghĩa dạng vai trò - hành động - thực thể, hỗ trợ biểu diễn quan hệ và ràng buộc ngữ nghĩa giữa các thành phần; trong khi Adaptive Attention điều tiết động mức đóng góp của từng nguồn tri thức theo từng bước sinh từ, đảm bảo quá trình tổng hợp thông tin diễn ra cân bằng và ngữ cảnh phụ thuộc được duy trì xuyên suốt. Trên nền tảng đó, GPT Re-ranking hoạt động như một cơ chế hiệu chỉnh hậu giải mã, đánh giá và lựa chọn các phương án chú thích có độ mạch lạc, tự nhiên và phù hợp ngữ cảnh cao nhất trong không gian ứng viên.

Bảng 5.4. Kết quả phân tích ảnh hưởng từng thành phần trên tập dữ liệu Flickr30K

Mô hình	BLUE-4	METEOR	CIDEr	SPICE
w/o CLIP features	36.2	35.3	95.0	22.8
w/o adaptive attention	36.9	35.9	97.6	23.1
w/o GPT Re-ranking	37.5	36.3	100.1	23.5
Proposed model (Full)	**38.2**	**36.9**	**102.8**	**24.0**

Sự cải thiện hiệu suất của CLIP-AMR-GPT bắt nguồn từ khả năng kết hợp có kiểm soát giữa hai khía cạnh cốt lõi của quá trình sinh ngôn ngữ: độ hiểu ngữ nghĩa (semantic grounding) và độ trôi chảy ngôn ngữ (linguistic fluency). CLIP giúp mô hình nhận diện chính xác hơn các đối tượng có tần suất xuất hiện thấp và các ngữ cảnh phức tạp, trong khi AMR/AMR-like giảm thiểu lỗi cú pháp và quan hệ thiếu hụt, qua đó nâng cao các độ đo phản ánh chiều sâu ngữ nghĩa như SPICE và CIDEr. Cơ chế Adaptive Attention cho phép mô hình điều chỉnh mức độ tập trung lên các thành phần ngữ nghĩa khác nhau trong embedding AMR-like theo từng bước giải mã, duy trì tính linh hoạt ngữ cảnh, còn GPT Re-ranking tinh chỉnh cấu trúc ngữ pháp và lựa chọn mô tả tối ưu về ngữ nghĩa, dẫn đến cải thiện nhất quán trên các độ đo BLEU-4, METEOR và ROUGE-L. Kết quả của phân tích ảnh hưởng từng thành phần xác nhận rằng mỗi thành phần đều đóng góp tích cực một cách độc lập, trong khi việc kết hợp đồng thời các thành phần này tạo nên hiệu ứng cộng hưởng, giúp mô hình đạt hiệu suất cao nhất trên cả hai tập dữ liệu chuẩn.

Những kết quả này khẳng định rằng cả ba thành phần - đặc trưng từ CLIP, cơ chế adaptive attention cho embedding AMR-like, và GPT Re-ranking - đều đóng vai trò quan trọng trong việc nâng cao chất lượng sinh chú thích. Đặc biệt, việc kết hợp đồng thời ba thành phần mang lại hiệu quả tốt nhất trên cả hai tập dữ liệu.

5.3.3.3. Đánh giá định tính

Bên cạnh các kết quả định lượng và phân tích ảnh hưởng các thành phần đã trình bày, phần này tiếp tục thực hiện đánh giá định tính nhằm làm rõ hơn tính hiệu quả của mô hình CLIP-AMR-GPT so với các phương pháp trước đó. Cụ thể, chú thích sinh ra từ mô hình CLIP-AMR-GPT được đặt trong tương quan so sánh với hai mô hình nền tảng AMR-GT&RG và RGTranCNet, dựa trên ba hình ảnh đại diện của tập kiểm tra MS COCO (Hình 5.2). Ba ví dụ này lần lượt phản ánh các bối cảnh: (a) cảnh tĩnh với nhiều đồ vật, (b) cảnh nội thất có cấu trúc không gian phức tạp, và (c) cảnh động thể hiện hành động và tương tác giữa các đối tượng.

Thứ nhất - Cảnh đồ vật (a): Ở ví dụ này, RGTranCNet sinh câu mô tả "*a cellphone next to scissors and a coffee cup*", nêu được các đối tượng chính nhưng còn thiếu chi tiết ngữ cảnh và chưa tự nhiên về mặt ngôn ngữ. Mô hình AMR-GT&RG cải thiện đáng kể với mô tả "*a cellphone and scissors next to a coffee cup and a black case*", phản ánh tốt hơn cấu trúc vật thể trong không gian. Trong khi đó, *CLIP-AMR-GPT* sinh câu "*a cup of coffee a cellphone and a pair of scissors on a desk*", thể hiện rõ ngữ cảnh "*on a desk*" và trôi chảy hơn về mặt ngôn ngữ. Kết quả này cho thấy lợi thế của việc kết hợp đặc trưng CLIP, giúp mô hình nắm bắt bố cục hình ảnh tổng thể, và cơ chế GPT Re-ranking, giúp tăng tính tự nhiên trong diễn đạt.

Thứ hai - Cảnh nội thất (b): Trong ví dụ này, RGTranCNet sinh mô tả *"a bedroom with a bed near a bookshelf"*, thể hiện được các đối tượng chính nhưng bỏ sót các yếu tố không gian như cửa sổ hay gương. AMR-GT&RG bổ sung được thông tin về bối cảnh với mô tả "*a neatly made bed next to a bookshelf and a window*" giúp mô hình nắm bắt tốt hơn bố trí tổng thể của căn phòng. Trong khi đó, CLIP-AMR-GPT sinh câu "*a neatly made bedroom with a mirror near a bed and bookshelf*", thể hiện đầy đủ các đối tượng chính và quan hệ không gian, đồng thời duy trì được tính tự nhiên và mạch lạc trong ngôn ngữ. Kết quả này cho thấy biểu diễn AMR-like kết hợp với CLIP giúp mô hình nhận diện và diễn đạt chính xác hơn các mối quan hệ các đối tượng trong không gian.

Thứ ba - Cảnh hành động (c): Trong ví dụ này, RGTranCNet sinh mô tả *"two men riding horses together on the beach"*, thể hiện đúng chủ thể và hành động chính. Mô hình AMR-GT&RG mở rộng mô tả thành *"two men smiling and riding horses*

along the beach", bổ sung chi tiết cảm xúc "smiling" giúp câu mô tả sinh động hơn, tuy nhiên lại thêm vào yếu tố không có trong một số chú thích chuẩn. Ngược lại, mô hình CLIP-AMR-GPT sinh câu *"two men riding horses on a beach shore"*, ngắn gọn, mạch lạc và bám sát nội dung ảnh, thể hiện rõ hành động và không gian mà không tạo chi tiết thừa. Hiệu quả này đến từ bước GPT Re-ranking, giúp lựa chọn phương án chú thích có xác suất ngữ nghĩa cao nhất, đảm bảo sự cân bằng giữa độ chính xác ngữ nghĩa và độ tự nhiên ngôn ngữ.

Hình 5.2. Các ví dụ định tính trên ba ảnh kiểm tra (a)-(c) thuộc tập MS COCO. Mỗi ảnh minh họa chú thích sinh ra bởi ba mô hình RGTranCNet, AMR-GT&RG và CLIP-AMR-GPT, kèm theo năm chú thích chuẩn (ground truth).

Từ ba ví dụ điển hình trên có thể nhận thấy mô hình CLIP-AMR-GPT tạo ra các câu mô tả vừa chính xác về mặt ngữ nghĩa, vừa trôi chảy về ngôn ngữ, vượt trội hơn so với các phương pháp trước đó. Sự phối hợp của ba nguồn tri thức - đặc trưng thị giác-ngôn ngữ từ CLIP, biểu diễn ngữ nghĩa AMR/AMR-like, và bước tái xếp hạng bằng GPT - giúp mô hình đạt được sự cân bằng giữa khả năng hiểu ngữ nghĩa và khả năng biểu đạt ngôn ngữ tự nhiên. Nhờ đó, các câu chú thích sinh ra không chỉ phản ánh chính xác nội dung và mối quan hệ trong ảnh, mà còn thể hiện tính tự nhiên, mạch lạc và phù hợp hơn với phong cách ngôn ngữ của con người.

5.4. Kết chương

Trong chương này, một mô hình chú thích ảnh mới có tên CLIP-AMR-GPT đã được đề xuất, kết hợp hiệu quả giữa các đặc trưng thị giác-ngôn ngữ từ CLIP, biểu diễn ngữ nghĩa trừu tượng từ AMR, cơ chế adaptive attention, và bước re-ranking bằng mô hình ngôn ngữ GPT. Khác với các phương pháp trước đó chỉ tập trung vào một nguồn đặc trưng hoặc bỏ qua các tri thức ngữ nghĩa sâu, mô hình CLIP-AMR-GPT tích hợp đa tầng tri thức nhằm tăng cường khả năng hiểu nội dung ảnh và sinh mô tả tự nhiên, giàu ngữ nghĩa hơn.

Kết quả thực nghiệm trên hai tập dữ liệu chuẩn là MS COCO và Flickr30K cho thấy mô hình đề xuất đạt hiệu suất vượt trội so với nhiều phương pháp tiên tiến hiện nay, đặc biệt là trên các độ đo đánh giá ngữ nghĩa như CIDEr, METEOR và SPICE. Phân tích ảnh hưởng từng thành phần trong mô hình cũng khẳng định vai trò quan trọng của chúng, trong đó các đặc trưng CLIP đóng vai trò cung cấp ngữ nghĩa thị giác chính xác, biểu diễn AMR mang lại cấu trúc ngôn ngữ sâu sắc, còn GPT re-ranking giúp cải thiện độ mạch lạc và phù hợp ngữ cảnh cho câu sinh ra. Hướng nghiên cứu tiếp theo có thể bao gồm: (1) mở rộng mô hình sang các ngôn ngữ khác ngoài tiếng Anh, ví dụ như tiếng Việt; và (2) tích hợp các mô hình ngôn ngữ lớn (LLMs) như GPT-4 hoặc Gemini để nâng cao chất lượng sinh chú thích trong các bối cảnh phức tạp hơn.

Trong phần tiếp theo của luận án, phần Kết luận tổng kết toàn bộ quá trình nghiên cứu, phân tích đóng góp khoa học, và đề xuất các hướng phát triển khả thi cho bài toán chú thích ảnh dựa trên mạng học sâu.

KẾT LUẬN

Phần này tổng kết toàn bộ nội dung nghiên cứu đã được triển khai trong luận án, bắt đầu từ phần Mở đầu, nơi xác lập tính cấp thiết, mục tiêu, phạm vi và phương pháp tiếp cận của đề tài, cho đến Chương 1, nơi trình bày một cách hệ thống các hướng nghiên cứu hiện tại, cũng như phân tích khoảng trống nghiên cứu cần giải quyết. Trên nền tảng đó, các Chương 2, 3, 4 và 5 lần lượt đề xuất bốn mô hình chú thích ảnh theo hướng kế thừa và mở rộng, với mức độ tích hợp tri thức ngữ nghĩa và biểu diễn đa phương thức ngày càng sâu sắc hơn. Phần kết luận này tổng hợp các kết quả đạt được, khẳng định những đóng góp khoa học và thực tiễn của luận án, đồng thời chỉ ra những vấn đề còn bỏ ngỏ và đề xuất các định hướng nghiên cứu tiếp theo trong lĩnh vực chú thích ảnh dựa trên học sâu đa phương thức.

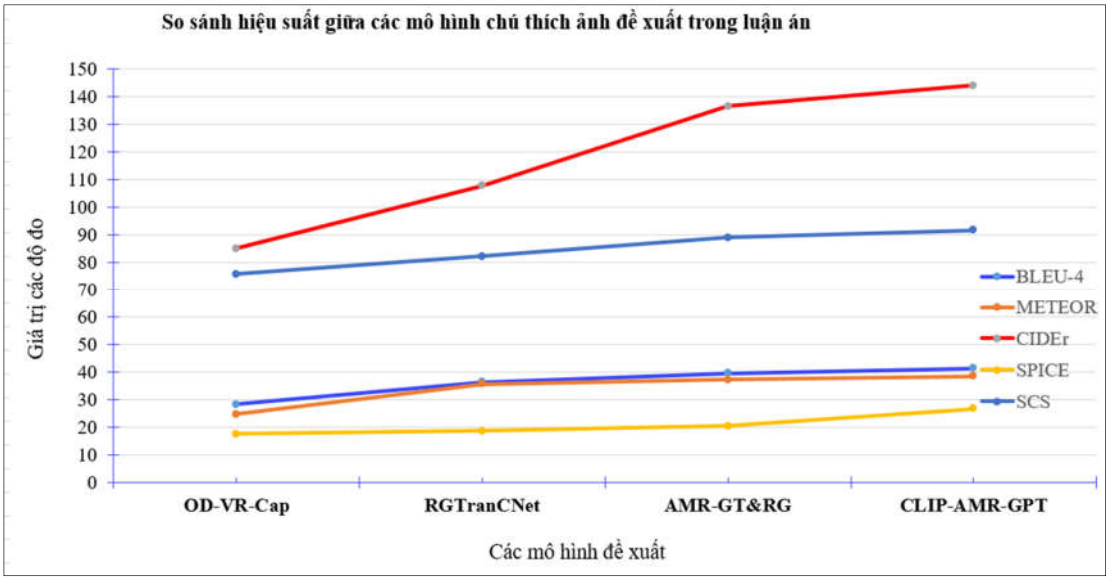
1. Tổng kết nội dung nghiên cứu và kết quả đạt được

Luận án tập trung giải quyết bài toán chú thích ảnh tự động thông qua việc đề xuất và phát triển bốn mô hình học sâu theo hướng kế thừa và cải tiến lũy tiến, với mức độ tích hợp tri thức ngữ nghĩa và đặc trưng đa phương thức ngày càng sâu sắc hơn: (i) mô hình OD-VR-Cap (Chương 2) khai thác đặc trưng thị giác và quan hệ giữa các đối tượng trong ảnh qua đồ thị quan hệ; (ii) mô hình RGTranCNet (Chương 3) tích hợp tri thức ngữ nghĩa từ ConceptNet nhằm tăng cường khả năng diễn đạt; (iii) mô hình AMR-GT&RG (Chương 4) kết hợp biểu diễn ngữ nghĩa trừu tượng dựa trên AMR với đồ thị quan hệ để cải thiện chiều sâu ngữ nghĩa; và (iv) mô hình CLIP-AMR-GPT (Chương 5) hợp nhất đặc trưng đa phương thức kết hợp cơ chế GPT Re-ranking nhằm nâng cao tính tự nhiên, mạch lạc và phù hợp ngữ cảnh của chú thích sinh ra. Các mô hình được đánh giá trên các tập dữ liệu chuẩn MS COCO và Flickr30K, sử dụng các độ đo phổ biến của bài toán chú thích ảnh (BLEU, METEOR, ROUGE, CIDEr, SPICE) cùng độ đo mới SCS do luận án đề xuất nhằm phản ánh mức độ phù hợp ngữ nghĩa giữa chú thích sinh ra và chú thích chuẩn ở mức embedding câu.

Chuỗi mô hình đề xuất cho thấy luận án đã hoàn thành đầy đủ mục tiêu nghiên cứu nêu trong phần Mở đầu. Mục tiêu khảo sát, phân tích và nhận diện khoảng trống nghiên cứu được thực hiện trong Chương 1. Các mục tiêu tiếp theo liên quan đến nâng cao khả năng biểu diễn nội dung hình ảnh, khai thác tri thức ngoài, biểu diễn ngữ nghĩa trừu tượng và hợp nhất tri thức đa nguồn đều được hiện thực hóa thông qua từng mô hình cụ thể, kèm theo các kiểm chứng thực nghiệm định lượng và định tính. Ngoài ra, việc đề xuất độ đo SCS giúp bổ sung một góc nhìn toàn diện hơn về mức độ phù hợp ngữ nghĩa giữa chú thích sinh ra và chú thích chuẩn.

Kết quả thực nghiệm trên MS COCO cho thấy bốn mô hình thể hiện sự cải thiện liên tục qua từng giai đoạn phát triển. Mô hình OD-VR-Cap khẳng định hiệu quả của việc khai thác cấu trúc quan hệ đối tượng; RGTranCNet mở rộng khả năng hiểu ngữ nghĩa nhờ tri thức ngoài; AMR-GT&RG vượt trội hơn đáng kể nhờ biểu diễn ngữ nghĩa trừu tượng; và CLIP-AMR-GPT đạt hiệu suất cao nhất nhờ hợp nhất đặc trưng đa phương thức và sử dụng GPT Re-ranking để chọn ra câu chú thích tối ưu. Các thực nghiệm mở rộng trên Flickr30K tiếp tục khẳng định khả năng tổng quát hóa của hai mô hình AMR-GT&RG và CLIP-AMR-GPT, cho thấy hiệu quả của hướng tiếp cận dựa trên hợp nhất tri thức đa tầng.

Hình KL.1 minh họa trực quan sự cải thiện hiệu suất giữa bốn mô hình đề xuất trên tập dữ liệu MS COCO. Biểu đồ cho thấy xu hướng tăng trưởng ổn định qua từng giai đoạn, trong đó CLIP-AMR-GPT đạt giá trị cao nhất trên tất cả các độ đo, đặc biệt ở CIDEr và SCS, phản ánh hiệu quả nổi bật của việc kết hợp đặc trưng đa phương thức (CLIP + AMR + đồ thị quan hệ) cùng GPT Re-ranking trong việc nâng cao chất lượng sinh chú thích.



Hình KL. 1 So sánh hiệu suất giữa các mô hình chú thích ảnh đề xuất trong luận án. Mô hình CLIP-AMR-GPT đạt kết quả cao nhất, khẳng định hiệu quả của việc hợp nhất đặc trưng đa phương thức và cơ chế GPT Re-ranking..

Tóm lại, chuỗi bốn mô hình do luận án đề xuất thể hiện một tiến trình nghiên cứu có hệ thống, kế thừa và mở rộng qua từng giai đoạn: từ khai thác quan hệ giữa các đối tượng, tích hợp tri thức ngữ nghĩa ngoài ảnh, kết hợp biểu diễn ngữ nghĩa trừu tượng dựa trên AMR, đến hợp nhất tri thức đa phương thức với sự hỗ trợ của mô hình ngôn ngữ lớn. Sự phát triển liên tục này cho thấy tính khả thi và hiệu quả của hướng tiếp cận mà luận án theo đuổi, đồng thời khẳng định việc hoàn thành đầy đủ các mục

tiêu nghiên cứu đã đề ra. Những kết quả đạt được không chỉ chứng minh tính đúng đắn của các mô hình mà còn mở ra triển vọng ứng dụng trong các hệ thống thị giác-ngôn ngữ hiện đại, đáp ứng yêu cầu cao về độ chính xác, chiều sâu ngữ nghĩa và khả năng biểu đạt tự nhiên của ngôn ngữ.

Đặc biệt, xét dưới góc độ khả năng tổng quát hóa và mở rộng ứng dụng, các mô hình được đề xuất trong luận án được thiết kế theo kiến trúc độc lập ngôn ngữ, trong đó các cơ chế hợp nhất đặc trưng thị giác, quan hệ ngữ nghĩa và biểu diễn trừu tượng không phụ thuộc chặt chẽ vào một ngôn ngữ cụ thể. Do đó, giới hạn về ngôn ngữ trong các thực nghiệm của luận án mang tính dữ liệu và điều kiện thực nghiệm, không phản ánh giới hạn của phương pháp đề xuất, và các mô hình hoàn toàn có khả năng mở rộng sang các ngôn ngữ khác, bao gồm tiếng Việt, khi có sẵn tài nguyên ngôn ngữ phù hợp.

2. Những vấn đề còn bỏ ngỏ

Mặc dù lĩnh vực chú thích ảnh tự động dựa trên mạng học sâu đã đạt được nhiều thành tựu đáng kể, bài toán này vẫn còn tồn tại nhiều thách thức và khoảng trống học thuật cần tiếp tục nghiên cứu. Các vấn đề còn bỏ ngỏ được trình bày đan xen với các hướng phát triển tương ứng nhằm gợi mở định hướng cho các công trình tiếp theo. Cụ thể, các vấn đề đáng chú ý bao gồm:

(i) **Hiểu ngữ nghĩa sâu và suy luận ngữ cảnh**: Mặc dù các mô hình hiện tại, bao gồm cả những kiến trúc tiên tiến như Transformer và các mô hình ngôn ngữ lớn (LLMs), chủ yếu học từ tương quan giữa đặc trưng hình ảnh và văn bản mà chưa thực sự nắm bắt được ngữ nghĩa sâu hay khả năng suy luận theo logic nhân quả. Việc sinh chú thích cho các tình huống đòi hỏi kiến thức nền, hiểu được mục đích hành động hoặc nhận diện hành vi ngụ ý vẫn còn là thách thức lớn. Việc tích hợp các mô hình ngôn ngữ lớn (LLMs) đã được huấn luyện trên lượng tri thức khổng lồ để hỗ trợ suy luận đa tầng, đồng thời áp dụng các kỹ thuật như causal reasoning hoặc multi-hop inference nhằm nâng cao khả năng hiểu hành vi và logic ngữ cảnh là cần thiết [116, 117].

(ii) **Xử lý đối tượng hiếm hoặc chưa từng xuất hiện** (novel/zero-shot object captioning): Các mô hình học sâu thường yêu cầu khối lượng lớn dữ liệu huấn luyện. Trong khi đó, thực tế chứa nhiều đối tượng hiếm gặp hoặc mới xuất hiện mà mô hình chưa từng quan sát. Việc sinh mô tả chính xác cho các ảnh có chứa các đối tượng như vậy đòi hỏi các cơ chế nhận thức linh hoạt, kết hợp với khả năng suy diễn từ tri thức ngoài tập dữ liệu, điều mà các mô hình hiện nay vẫn chưa giải quyết hiệu quả. Việc khai thác tri thức ngoài (ví dụ: ConceptNet, Wikidata) để biểu diễn khái niệm dạng

ngữ nghĩa, hoặc sử dụng tri thức từ mô hình pretrained như CLIP [64], đồng thời áp dụng các kỹ thuật zero-shot learning, semantic generalization hoặc knowledge distillation để mở rộng khả năng của mô hình mà không cần gán nhãn trực tiếp, hoặc sử dụng tri thức từ mô hình pretrained như CLIP cũng là vấn đề đặt ra

(iii) **Thiếu chuẩn đánh giá thống nhất về chất lượng ngữ nghĩa**: Các độ đo hiện hành như BLEU, METEOR, CIDEr, ROUGE hay SPICE chủ yếu đánh giá sự tương đồng về mặt ngôn ngữ bề mặt mà chưa phản ánh đầy đủ khía cạnh ngữ nghĩa, sự phù hợp ngữ cảnh hoặc mức độ tự nhiên trong biểu đạt. Bên cạnh đó, chưa tồn tại một chuẩn đánh giá thống nhất được cộng đồng nghiên cứu chấp nhận rộng rãi, gây khó khăn trong việc so sánh và tái lập kết quả. Chuẩn hóa các độ đo đánh giá ngữ nghĩa, chẳng hạn như SCS do luận án đề xuất, hoặc kết hợp giữa đánh giá tự động với phản hồi từ người dùng (human-in-the-loop evaluation) để ghi nhận tính chính xác ngữ nghĩa và tự nhiên của chú thích sinh ra cũng nên được xem xét [74].

(iv) **Hạn chế trong xử lý đa ngôn ngữ và dữ liệu ít tài nguyên**: Hầu hết các nghiên cứu hiện nay tập trung vào tiếng Anh, trong khi các ngôn ngữ khác - đặc biệt là nhóm ngôn ngữ ít tài nguyên - chưa nhận được sự quan tâm tương xứng. Việc thiếu các tập dữ liệu song ngữ được gán nhãn chất lượng cao và công cụ xử lý ngôn ngữ phù hợp dẫn đến việc khó áp dụng các mô hình hiện có trong môi trường đa ngữ thực tế. Xây dựng các mô hình chú thích ảnh đa ngôn ngữ bằng cách khai thác mô hình ngôn ngữ lớn đa ngữ (multilingual LLMs), kết hợp với kỹ thuật machine translation, self-supervised learning hoặc pretraining đa miền dữ liệu để huấn luyện mô hình hiệu quả cho các ngôn ngữ ít tài nguyên như tiếng Việt rất cần thiết [118].

(v) **Tích hợp tri thức ngữ nghĩa ngoài vẫn còn sơ khai**: Mặc dù một số hướng tiếp cận gần đây đã bước đầu tích hợp tri thức từ các nguồn như ConceptNet, WordNet, Wikidata hoặc AMR vào mô hình học sâu, nhưng vẫn tồn tại nhiều vấn đề như: khó khăn trong việc biểu diễn tri thức theo dạng tương thích với mô hình, thiếu cơ chế lọc và đánh giá mức độ liên quan của tri thức, và hạn chế trong khả năng huấn luyện toàn bộ hệ thống theo cơ chế từ đầu đến cuối (end-to-end) [119]. Nên thiết kế kiến trúc linh hoạt, có thể tiếp nhận tri thức dưới dạng đồ thị, embedding hoặc ngôn ngữ tự nhiên, đồng thời phát triển các cơ chế chú ý có điều kiện (conditional attention) hoặc học chọn lọc tri thức (knowledge selection) nhằm khai thác hiệu quả những thông tin có liên quan và tin cậy trong quá trình sinh chú thích [28, 120].

(vi) **Thiếu khả năng giải thích và minh bạch trong mô hình**: Đa phần các mô hình hiện hành được xem như "hộp đen", thiếu khả năng giải thích rõ ràng cho việc sinh ra một chú thích cụ thể. Điều này làm giảm độ tin cậy trong các ứng dụng quan trọng như trợ lý hỗ trợ người khiếm thị, y tế hoặc giáo dục. Việc phát triển các

mô hình có khả năng giải thích và cung cấp thông tin minh bạch về cơ chế hoạt động là một yêu cầu cấp thiết. Có thể phát triển các mô hình chú thích có khả năng giải thích, ví dụ: sử dụng attention visualization, hoặc hệ thống giải thích hậu nghiệm (post-hoc explanation) như LIME hoặc SHAP trong các mô hình thị giác-ngôn ngữ để tăng tính minh bạch và đáng tin cậy nhằm cung cấp thông tin trực quan về cơ sở sinh chú thích [121-123].

(vii) **Khả năng thích ứng với dữ liệu thực tế và môi trường biến động**: Các mô hình chú thích ảnh hiện tại thường được huấn luyện và đánh giá trên tập dữ liệu tĩnh, được xử lý và gán nhãn đầy đủ. Trong khi đó, dữ liệu thực tế thường có tính biến động cao, chứa nhiều yếu tố nhiễu như đối tượng bị che khuất, điều kiện ánh sáng không thuận lợi hoặc thành phần ảnh không đầy đủ. Việc mở rộng mô hình để xử lý hiệu quả trong các điều kiện môi trường thực tế vẫn là một hướng nghiên cứu còn bỏ ngỏ. Nên tập trung xây dựng mô hình chịu lỗi tốt, có khả năng hoạt động ổn định trong điều kiện ảnh động, thời gian thực, hoặc ảnh có chất lượng kém. Ngoài ra, cần đánh giá hiệu suất trên các ứng dụng cụ thể như chú thích video, trợ lý thị giác, hoặc mô tả hình ảnh trong giáo dục và y tế [124].

Bên cạnh những vấn đề học thuật nêu trên, các kết quả của luận án cũng mở ra khả năng ứng dụng thực tiễn cao tại Việt Nam. Các mô hình chú thích ảnh thông minh có thể được triển khai trong các hệ thống hỗ trợ người khiếm thị (ví dụ: kính thông minh mô tả cảnh vật bằng giọng nói tiếng Việt), trong giáo dục (các nền tảng học tập trực quan, tạo mô tả tự động cho hình ảnh minh họa), hoặc trong lưu trữ và truyền thông số (gắn nhãn tự động cho kho dữ liệu ảnh, báo chí, thư viện số). Ngoài ra, các kỹ thuật được phát triển trong luận án còn có thể mở rộng cho các ứng dụng giám sát thị giác thông minh, phân tích hành vi, và tìm kiếm đa phương thức trong môi trường công nghiệp và chính phủ điện tử. Tuy nhiên, khi triển khai trong thực tế tại Việt Nam, vẫn tồn tại một số thách thức đáng kể, như sự thiếu hụt dữ liệu gán nhãn tiếng Việt, hạn chế về tài nguyên tính toán và khả năng huấn luyện các mô hình quy mô lớn, cũng như yêu cầu về độ tin cậy, minh bạch và kiểm định đầu ra của các hệ thống AI trong các lĩnh vực nhạy cảm như y tế, giáo dục và an sinh xã hội. Do đó, hướng nghiên cứu tiếp theo cần tập trung vào việc phát triển các mô hình gọn nhẹ (lightweight models), kỹ thuật nén mô hình (model compression), hoặc học chưng cất tri thức (knowledge distillation), để có thể triển khai hiệu quả trên hạ tầng phần cứng trong nước mà vẫn đảm bảo hiệu năng và độ an toàn cao.

Nhìn chung, các định hướng trên không chỉ phản ánh xu thế phát triển của lĩnh vực chú thích ảnh tự động mà còn mở ra tiềm năng ứng dụng rộng rãi trong bối cảnh Việt Nam. Việc kết hợp tri thức ngữ nghĩa, khai thác mô hình ngôn ngữ lớn và tối ưu

hóa trên hạ tầng hạn chế là bước quan trọng để chuyển các mô hình thị giác - ngôn ngữ từ nghiên cứu sang thực tiễn. Để làm được điều đó, cần đồng thời phát triển dữ liệu mở và hệ sinh thái ngữ liệu tiếng Việt, tạo nền tảng cho việc ứng dụng kết quả luận án vào các lĩnh vực như giáo dục, y tế, giao thông thông minh và hỗ trợ người khiếm thị.

3. Tổng kết

Từ những kết quả đạt được và các phân tích ở trên, có thể khẳng định rằng luận án đã hoàn thành toàn bộ mục tiêu và nội dung nghiên cứu đặt ra trong phần Mở đầu, cả về phương diện lý thuyết lẫn thực nghiệm. Chuỗi bốn mô hình được đề xuất trong luận án - từ OD-VR-Cap, RGTranCNet, AMR-GT&RG đến CLIP-AMR-GPT - đã thể hiện một tiến trình phát triển có tính kế thừa, mở rộng và hội tụ, hướng đến mục tiêu chung là phát triển phương pháp chú thích ảnh dựa trên mạng học sâu có khả năng hiểu ngữ nghĩa và sinh ngôn ngữ tự nhiên ở mức cao.

Các kết quả đạt được không chỉ khẳng định tính đúng đắn và hiệu quả của hướng tiếp cận mà còn đóng góp thiết thực vào tri thức khoa học trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo đa phương thức. Về mặt học thuật, luận án đã mở rộng phạm vi nghiên cứu từ khai thác quan hệ giữa các đối tượng, đến tích hợp tri thức ngữ nghĩa ngoài và biểu diễn nghĩa trừu tượng. Về mặt thực tiễn, các kết quả này đặt nền móng cho việc phát triển các hệ thống ứng dụng thông minh trong điều kiện Việt Nam, nơi mà dữ liệu, ngôn ngữ và tài nguyên tính toán còn hạn chế.

Như vậy, luận án không chỉ góp phần củng cố cơ sở lý luận cho hướng nghiên cứu chú thích ảnh dựa trên mạng học sâu mà còn mở ra triển vọng áp dụng trong nhiều lĩnh vực khác nhau của đời sống, từ giáo dục, y tế, truyền thông đến công nghệ hỗ trợ con người. Việc tiếp tục mở rộng nghiên cứu theo các hướng đã nêu trong Mục 2 giúp hoàn thiện hơn khả năng biểu diễn ngữ nghĩa, mở rộng sang đa ngôn ngữ, và nâng cao tính thích ứng của mô hình với dữ liệu và ngữ cảnh thực tế.

DANH MỤC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

[CT1] **Nguyen Van Thinh**, Tran Van Lang, Van The Thanh (2024). *OD-VR-Cap: Image captioning based on detecting and predicting relationships between objects*. Journal of Computer Science and Cybernetics, 40(4), 326-345, DOI: 10.15625/1813-9663/20929.

[CT2] **Nguyễn Văn Thịnh**, Trần Văn Lăng, Văn Thế Thành, Trần Hữu Quốc Thư, Lê Thị Vĩnh Thanh (2024). *ViT-Trans-AMR: Nâng cao hiệu quả chú thích ảnh với đồ thị ngữ nghĩa AMR và mạng Transformer*. Hội nghị khoa học quốc gia về nghiên cứu cơ bản và ứng dụng công nghệ thông tin, FAIR'2024, Hà Nội, 8-9/8/2024, NXB Khoa học tự nhiên và Công nghệ, *DOI: 10.15625/vap.2024.0289,* pp.878-888.

[CT3] **Nguyễn Văn Thịnh,** Trần Văn Lăng, Trần Hữu Quốc Thư, Nguyễn Thị Ngọc Hoa (2024). *Nâng cao hiệu quả chú thích ảnh sử dụng mạng Transformer và cơ sở tri thức ConceptNet*. Hội thảo quốc gia lần thứ XXVII về một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông, VNICT2024, Nha Trang, 11-12/10/2024, ISBN: 978-604-67-3029-3, NXB Khoa học và Kỹ thuật, pp.404-409.

[CT4] **N. V. Thinh**, T. V. Lang, and V. T. Thanh (2026). *RGTranCNet: Effective Image Captioning Model using Cross-Attention and Semantic Knowledge,* Vietnam Journal of Science and Technology, vol. 64, no. 1, 2026, DOI: https://doi.org/10.15625/2525-2518/22381 (Chỉ mục: **Scopus**; thuộc nhóm **Q4** theo SJR).

[CT5] **N. V. Thinh**, T. V. Lang and V. T. Thanh (202). *Integrating Abstract Meaning Representation to Enhance Transformer-Based Image Captioning*, in IEEE Access, vol. 13, pp. 112528-112551, 2025, DOI: 10.1109/ACCESS.2025.3584128 (Chỉ mục: **SCIE**; thuộc nhóm **Q1** theo JCR/SJR).

[CT6] **N. V. Thinh**, T. V. Lang, and N. M. Hai (2026). *CLIP-AMR-GPT: Enhancing Image Captioning via Cross-Modal Semantics Fusion and GPT-Based Re-Ranking*. Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2025. Lecture Notes in Computer Science, vol 16354. Springer, Singapore. https://doi.org/10.1007/978-981-95-4960-3_17 (Chỉ mục: **Scopus**; LNCS thuộc nhóm **Q2** theo SJR).

DANH MỤC TÀI LIỆU THAM KHẢO

1. Hussain, A., et al. *Foundation Models: From Current Developments, Challenges, and Risks to Future Opportunities*. in *2025 27th International Conference on Advanced Communications Technology (ICACT)*. 2025. IEEE.
2. Baltrušaitis, T., C. Ahuja, and L.-P. Morency, *Multimodal machine learning: A survey and taxonomy.* IEEE transactions on pattern analysis and machine intelligence, 2018. **41**(2): p. 423-443.
3. Vinyals, O., et al. *Show and tell: A neural image caption generator*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
4. Al-Shamayleh, A.S., et al., *A comprehensive literature review on image captioning methods and metrics based on deep learning technique.* Multimedia Tools and Applications, 2024. **83**(12): p. 34219-34268.
5. Stefanini, M., et al., *From show to tell: A survey on deep learning-based image captioning.* IEEE transactions on pattern analysis and machine intelligence, 2022. **45**(1): p. 539-559.
6. Agarwal, L. and B. Verma, *From methods to datasets: A survey on Image-Caption Generators.* Multimedia Tools and Applications, 2024. **83**(9): p. 28077-28123.
7. Sharma, A., H. Singh, and M. Pant, *Pixels to Prose: A Comprehensive Survey of Image Captioning Techniques with Deep Learning and Generative Artificial Intelligence.* Neurocomputing, 2025: p. 132385.
8. Oluwasammi, A., et al., *Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning.* Complexity, 2021. **2021**(1): p. 5538927.
9. Ghandi, T., H. Pourreza, and H. Mahyar, *Deep learning approaches on image captioning: A review.* ACM Computing Surveys, 2023. **56**(3): p. 1-39.
10. Arystanbekov, B., et al. *Image captioning for the visually impaired and blind: a recipe for low-resource languages*. in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2023. IEEE.
11. Yin, C., et al. *Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network*. in *2019 IEEE international conference on data mining (ICDM)*. 2019. IEEE.

12. Ayesha, H., et al., *Automatic medical image interpretation: State of the art and future directions.* Pattern Recognition, 2021. **114**: p. 107856.

13. Lee, D.I., et al., *Crop Disease Diagnosis with Deep Learning-Based Image Captioning and Object Detection.* Applied Sciences, 2023. **13**(5): p. 3148.

14. Li, W., et al., *The traffic scene understanding and prediction based on image captioning.* IEEE Access, 2020. **9**: p. 1420-1427.

15. Hossain, M.Z., et al., *A comprehensive survey of deep learning for image captioning.* ACM Computing Surveys (CsUR), 2019. **51**(6): p. 1-36.

16. Kearns, L., *Content moderation assistance through image caption generation.* Intelligent Systems with Applications, 2025. **25**: p. 200489.

17. Tsai, W.-L., et al., *Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention.* Automation in Construction, 2025. **169**: p. 105863.

18. Xu, K., et al. *Show, attend and tell: Neural image caption generation with visual attention*. in *International conference on machine learning*. 2015. PMLR.

19. Cornia, M., et al. *Meshed-memory transformer for image captioning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

20. Zhou, Y., Y. Sun, and V. Honavar. *Improving image captioning by leveraging knowledge graphs*. in *2019 IEEE winter conference on applications of computer vision (WACV)*. 2019. IEEE.

21. Hafeth, D.A., S. Kollias, and M. Ghafoor, *Semantic representations with attention networks for boosting image captioning.* IEEE Access, 2023. **11**: p. 40230-40239.

22. Yao, T., et al. *Exploring visual relationship for image captioning*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.

23. Parseh, M.J. and S. Ghadiri, *Graph-based image captioning with semantic and spatial features.* Signal Processing: Image Communication, 2025. **133**: p. 117273.

24. Bhattacharyya, A., M. Palmer, and C. Heckman. *ReCAP: Semantic Role Enhanced Caption Generation*. in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024.

25. Basioti, K., et al. *CIC-BART-SSA: Controllable Image Captioning with Structured Semantic Augmentation*. in *European Conference on Computer Vision*. 2024. Springer.

26. Kim, J., et al. *De-Bias Using Abstract Meaning Representation for Image Captioning*. in *2024 IEEE International Conference on Consumer Electronics (ICCE)*. 2024. IEEE.

27. Li, Y., et al. *Pointing novel objects in image captioning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

28. Zhou, L., et al. *Unified vision-language pre-training for image captioning and vqa*. in *Proceedings of the AAAI conference on artificial intelligence*. 2020.

29. Song, Z., et al. *Direction relation transformer for image captioning*. in *Proceedings of the 29th ACM International Conference on Multimedia*. 2021.

30. Verma, A., et al., *Automatic image caption generation using deep learning.* Multimedia Tools and Applications, 2024. **83**(2): p. 5309-5325.

31. Abdulgalil, H.D. and O.A. Basir, *Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models.* Natural Language Processing Journal, 2025: p. 100159.

32. Sarto, S., M. Cornia, and R. Cucchiara. *Image Captioning Evaluation in the Age of Multimodal LLMs: Challenges and Future Perspectives*. in *Proceedings of the 34th International Joint Conference on Artificial Intelligence*. 2025.

33. Yuan, Y., Z. Li, and B. Zhao, *A survey of multimodal learning: Methods, applications, and future.* ACM Computing Surveys, 2025. **57**(7): p. 1-34.

34. Farkh, R., G. Oudinet, and Y. Foued, *Image Captioning Using Multimodal Deep Learning Approach.* Computers, Materials & Continua, 2024. **81**(3).

35. Kim, T., et al. *Vipcap: Retrieval text-based visual prompts for lightweight image captioning*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2025.

36. Lai, Z., et al. *Revisit Large-Scale Image-Caption Data in Pre-training Multimodal Foundation Models*. in *The Thirteenth International Conference on Learning Representations*.

37. Albadarneh, I.A., B.H. Hammo, and O.S. Al-Kadi, *Attention-based transformer models for image captioning across languages: An in-depth survey and evaluation.* Computer Science Review, 2025. **58**: p. 100766.

38. Santiesteban, S.S., et al. *Improved Image Captioning Via Knowledge Graph-Augmented Models*. in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024. IEEE.

39. Salgotra, G., P. Abrol, and A. Selwal, *A survey on automatic image captioning approaches: Contemporary trends and future perspectives.* Archives of Computational Methods in Engineering, 2025. **32**(3): p. 1459-1497.

40. Ming, Y., et al., *Visuals to text: A comprehensive review on automatic image captioning.* IEEE/CAA Journal of Automatica Sinica, 2022. **9**(8): p. 1339-1365.

41. Jiang, W., et al. *Recurrent fusion network for image captioning*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

42. Hossain, M.Z., et al. *Attention-based image captioning using DenseNet features*. in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part V 26*. 2019. Springer.

43. Patwari, N. and D. Naik. *En-de-cap: An encoder decoder model for image captioning*. in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. 2021. IEEE.

44. Azhar, I., I. Afyouni, and A. Elnagar. *Facilitated deep learning models for image captioning*. in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*. 2021. IEEE.

45. Poddar, A.K. and R. Rani, *Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language.* Procedia Computer Science, 2023. **218**: p. 686-696.

46. Xie, T., et al., *Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning.* Applied Sciences, 2023. **13**(13): p. 7916.

47. Bai, T., et al., *An image caption model based on attention mechanism and deep reinforcement learning.* Frontiers in Neuroscience, 2023. **17**: p. 1270850.

48. Abinaya, S., M. Deepak, and A. Sherly Alphonse, *Enhanced Image Captioning Using Bahdanau Attention Mechanism and Heuristic Beam Search Algorithm.* IEEE Access, 2024. **12**: p. 100991-101003.

49. Yang, X., et al. *Auto-encoding scene graphs for image captioning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

50. Song, Z. and X. Zhou. *Exploring Explicit And Implicit Visual Relationships For Image Captioning*. in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. 2021. IEEE Computer Society.

51. Gao, L., B. Wang, and W. Wang. *Image captioning with scene-graph based semantic concepts*. in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. 2018.

52. Chen, S., et al. *Say as you wish: Fine-grained control of image caption generation with abstract scene graphs*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

53. Yan, J., et al., *Caption TLSTMs: combining transformer with LSTMs for image captioning.* International Journal of Multimedia Information Retrieval, 2022. **11**(2): p. 111-121.

54. Mozes, M., et al., *Scene graph generation for better image captioning?* arXiv preprint arXiv:2109.11398, 2021.

55. Monti, F., K. Otness, and M.M. Bronstein. *Motifnet: a motif-based graph convolutional network for directed graphs*. in *2018 IEEE data science workshop (DSW)*. 2018. IEEE.

56. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.

57. Dosovitskiy, A., et al., *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.

58. Liu, Z., et al. *Swin transformer: Hierarchical vision transformer using shifted windows*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

59. Liu, Z., et al. *A convnet for the 2020s*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

60. Wang, Y., J. Xu, and Y. Sun. *End-to-end transformer based model for image captioning*. in *Proceedings of the AAAI conference on artificial intelligence*. 2022.

61. Yang, X., et al., *Context-aware transformer for image captioning.* Neurocomputing, 2023. **549**: p. 126440.

62. Zhou, Y., Y. Sun, and V.G. Honavar, *Improving Image Captioning by Leveraging Knowledge Graphs. IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. 2019, IEEE.

63. Li, Z., Q. Su, and T. Chen, *External knowledge-assisted Transformer for image captioning.* Image and Vision Computing, 2023. **140**: p. 104864.

64. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. 2021. PmLR.

65. Li, J., et al. *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*. in *International conference on machine learning*. 2022. PMLR.

66. Li, J., et al. *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*. in *International conference on machine learning*. 2023. PMLR.

67. Alayrac, J.-B., et al., *Flamingo: a visual language model for few-shot learning.* Advances in neural information processing systems, 2022. **35**: p. 23716-23736.

68. Barraco, M., et al. *The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis*. in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

69. Guo, J., et al., *Based-CLIP early fusion transformer for image caption.* Signal, Image and Video Processing, 2025. **19**(2): p. 112.

70. Mishra, S., et al. *Image caption generation using vision transformer and gpt architecture*. in *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. 2024. IEEE.

71. Song, R., B. Zhao, and L. Yu, *Enhanced CLIP-GPT Framework for Cross-Lingual Remote Sensing Image Captioning.* IEEE Access, 2024.

72. Banarescu, L., et al. *Abstract meaning representation for sembanking*. in *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*. 2013.

73. Neto, A.M.A., H.M. Caseli, and T.A. Almeida. *Dense Captioning Using Abstract Meaning Representation*. in *Brazilian Conference on Intelligent Systems*. 2020. Springer.

74. González-Chávez, O., et al., *Are metrics measuring what they should? An evaluation of Image Captioning task metrics.* Signal Processing: Image Communication, 2024. **120**: p. 117071.

75. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. 2014. Springer.

76. Karpathy, A. and L. Fei-Fei. *Deep visual-semantic alignments for generating image descriptions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

77. Young, P., et al., *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.* Transactions of the association for computational linguistics, 2014. **2**: p. 67-78.

78. Karpathy, A., A. Joulin, and L.F. Fei-Fei, *Deep fragment embeddings for bidirectional image sentence mapping.* Advances in neural information processing systems, 2014. **27**.

79. Krishna, R., et al., *Visual genome: Connecting language and vision using crowdsourced dense image annotations.* International journal of computer vision, 2017. **123**(1): p. 32-73.

80. Papineni, K., et al. *Bleu: a method for automatic evaluation of machine translation*. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.

81. Banerjee, S. and A. Lavie. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.

82. Vedantam, R., C. Lawrence Zitnick, and D. Parikh. *Cider: Consensus-based image description evaluation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

83. Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*. in *Text summarization branches out*. 2004.

84. Anderson, P., et al. *Spice: Semantic propositional image caption evaluation*. in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. 2016. Springer.

85. Zohourianshahzadi, Z. and J.K. Kalita, *Neural attention for image captioning: review of outstanding methods.* Artificial Intelligence Review, 2022. **55**(5): p. 3833-3862.

86. You, Q., et al. *Image captioning with semantic attention*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

87. Jia, J., et al., *Image captioning based on scene graphs: A survey.* Expert Systems with Applications, 2023: p. 120698.

88. Chen, Z.-M., et al. *Multi-label image recognition with graph convolutional networks*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

89. Kipf, T.N. and M. Welling, *Semi-supervised classification with graph convolutional networks.* arXiv preprint arXiv:1609.02907, 2016.

90. Hamilton, W., Z. Ying, and J. Leskovec, *Inductive representation learning on large graphs.* Advances in neural information processing systems, 2017. **30**.

91. Yang, J., et al. *Graph r-cnn for scene graph generation*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.

92. Zellers, R., et al. *Neural motifs: Scene graph parsing with global context*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

93. Xu, N., et al., *Scene graph captioner: Image captioning based on structural visual representation.* Journal of Visual Communication and Image Representation, 2019. **58**: p. 477-485.

94. Kumar, V., et al. *A Novel Approach to Scene Graph Vectorization*. in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. 2021. IEEE.

95. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

96. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*. 1985, Institute for Cognitive Science, University of California, San Diego La ….

97. Werbos, P.J., *Backpropagation through time: what it does and how to do it.* Proceedings of the IEEE, 1990. **78**(10): p. 1550-1560.

98. Xu, D., et al. *Scene graph generation by iterative message passing*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

99. Lu, C., et al. *Visual relationship detection with language priors*. in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. 2016. Springer.

100. Tian, P., H. Mo, and L. Jiang, *Scene graph generation by multi-level semantic tasks.* Applied Intelligence, 2021: p. 1-13.

101. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *European conference on computer vision*. 2014. Springer.

102. Hendricks, L.A., et al. *Deep compositional captioning: Describing novel object categories without paired training data*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

103. Speer, R., J. Chin, and C. Havasi. *Conceptnet 5.5: An open multilingual graph of general knowledge*. in *Proceedings of the AAAI conference on artificial intelligence*. 2017.

104. Ramos, L., et al., *A study of convnext architectures for enhanced image captioning.* IEEE Access, 2024.

105. Li, Z., et al., *Modeling graph-structured contexts for image captioning.* Image and Vision Computing, 2023. **129**: p. 104591.

106. He, S., et al. *Image captioning through image transformer*. in *Proceedings of the Asian conference on computer vision*. 2020.

107. Pan, Y., et al. *X-linear attention networks for image captioning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

108. Li, Y., et al. *Comprehending and ordering semantics for image captioning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

109. Hou, J., et al. *Joint commonsense and relation reasoning for image and video captioning*. in *Proceedings of the AAAI conference on artificial intelligence*. 2020.

110. Wein, S. and J. Opitz. *A survey of AMR applications*. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

111. Konstas, I., et al. *Neural AMR: Sequence-to-Sequence Models for Parsing and Generation*. in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017.

112. Devlin, J., et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*. in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.

113. Reimers, N. and I. Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019. Association for Computational Linguistics.

114. Sharma, H. and S. Srivastava, *Multilevel attention and relation network based image captioning model.* Multimedia Tools and Applications, 2023. **82**(7): p. 10981-11003.

115. Ren, S., et al., *Faster r-cnn: Towards real-time object detection with region proposal networks.* Advances in neural information processing systems, 2015. **28**.

116. Shinde, G., et al., *A Survey on Efficient Vision-Language Models.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2025. **15**(3): p. e70036.

117. Zhang, J., et al., *Vision-language models for vision tasks: A survey.* IEEE transactions on pattern analysis and machine intelligence, 2024. **46**(8): p. 5625-5644.

118. Ramos, R., B. Martins, and D. Elliott. *LMCAP: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting*. in *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. 2023. Association for Computational Linguistics (ACL).

119. Wei, H., et al., *Integrating scene semantic knowledge into image captioning.* ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021. **17**(2): p. 1-22.

120. Peng, Y., T. Bonald, and M. Alam. *Refining wikidata taxonomy using large language models*. in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024.

121. Wu, Q. and J. Xiao, *An explainable fully attentional network for multivariate time-series forecasting.* Knowledge-Based Systems, 2025: p. 113780.

122. Elguendouze, S., et al., *Explainability in image captioning based on the latent space.* Neurocomputing, 2023. **546**: p. 126319.

123. Elguendouze, S., et al. *Towards explainable deep learning for image captioning through representation space perturbation*. in *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022. IEEE.

124. Xiao, H., et al., *A comprehensive survey of large language models and multimodal large language models in medicine.* Information Fusion, 2025. **117**: p. 102888.