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

TÓM TẮT LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Khoa học máy tính

Mã số: 9 48 01 01

Hà Nội - 2026

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:

1. Người hướng dẫn 1: PGS. TS. Trần Văn Lăng, Trường Đại học Ngoại ngữ - Tin học TP. Hồ Chí Minh
2. Người hướng dẫn 2: TS. Văn Thế Thành, Trường Đại học Sư phạm TP. Hồ Chí Minh

Phản biện 1: TS. Nguyễn Như Sơn

Phản biện 2: PGS. TS. Lê Kim Hùng

Phản biện 3: TS. Lê Quang Minh

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi 14h00, ngày 23 tháng 04 năm 2026.

Có thể tìm hiểu luận án tại:

1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

Chú thích ảnh (Image Captioning) là bài toán liên ngành kết nối thị giác máy tính và xử lý ngôn ngữ tự nhiên, đòi hỏi mô hình vừa "hiểu" nội dung ảnh vừa "diễn đạt" thành câu văn mạch lạc. Mặc dù học sâu đa phương thức đã đạt được những tiến bộ đáng kể, các hệ thống hiện nay vẫn hạn chế ở ba điểm: (i) khai thác đặc trưng hình ảnh chưa đủ sâu; (ii) biểu diễn và tận dụng quan hệ giữa các đối tượng còn yếu; (iii) tích hợp tri thức ngoài tập dữ liệu (ví dụ: ConceptNet, AMR, LLM) chưa hiệu quả, dẫn tới chú thích thiếu chiều sâu ngữ nghĩa. Do vậy, nghiên cứu phát triển phương pháp chú thích ảnh dựa trên mạng học sâu kết hợp đặc trưng thị giác, quan hệ và tri thức ngữ nghĩa là cần thiết cả về học thuật lẫn ứng dụng.

2. Mục tiêu nghiên cứu

Mục tiêu của luận án: Phát triển các phương pháp chú thích ảnh dựa trên mạng học sâu, kết hợp đặc trưng thị giác, quan hệ giữa các đối tượng, tri thức ngoài tập dữ liệu và biểu diễn ngữ nghĩa trừu tượng để nâng cao độ chính xác và tính tự nhiên của mô tả.

Luận án cụ thể hóa mục tiêu nghiên cứu thành các nội dung như sau:

(i) Khảo sát, phân tích các hướng chú thích ảnh hiện có, chỉ ra hạn chế trong biểu diễn quan hệ và tri thức ngữ nghĩa.

(ii) Đề xuất mô hình chú thích ảnh có khả năng mô hình hóa và khai thác quan hệ giữa các đối tượng trong ảnh.

(iii) Nghiên cứu cơ chế hợp nhất tri thức đa nguồn (đặc trưng thị giác–ngôn ngữ, ConceptNet, mô hình ngôn ngữ lớn) nhằm tăng chiều sâu ngữ nghĩa và độ tự nhiên của chú thích.

(iv) Khai thác biểu diễn ngữ nghĩa trừu tượng (AMR) để nâng cao khả năng hiểu nội dung ảnh ở mức khái niệm.

(v) Thực nghiệm và đánh giá toàn diện trên MS COCO và Flickr30K bằng các độ đo BLEU, METEOR, ROUGE-L, CIDEr, SPICE và SCS để kiểm chứng hiệu quả và khả năng tổng quát hóa của mô hình.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng: Đối tượng nghiên cứu của luận án là bài toán sinh chú thích ngôn ngữ tự nhiên cho ảnh theo khung encoder–decoder, với trọng tâm:

- Biểu diễn đặc trưng thị giác (đối tượng, vùng ảnh) kết hợp đồ thị quan hệ và đặc trưng thị giác–ngôn ngữ từ CLIP.
- Tích hợp tri thức ngữ nghĩa đa nguồn: tri thức ngoài (ConceptNet, LLM), tri thức trừu tượng (AMR) và tri thức ngữ cảnh trong ảnh.
- Thiết kế bộ giải mã LSTM/Transformer với chơ chế chú ý (dual, cross, masked attention) để sinh chú thích mạch lạc, đúng cú pháp và ngữ nghĩa.

Phạm vi:

- Dữ liệu: hai tập dữ liệu chuẩn MS COCO và Flickr30K.
- Mô hình: kiến trúc encoder–decoder, encoder trích xuất đặc trưng đối tượng, quan hệ, AMR và CLIP-ViT; decoder là LSTM/Transformer kết hợp attention; giai đoạn suy luận dùng GPT re-ranking để chọn chú thích cuối.
- Đánh giá: độ đo BLEU, METEOR, ROUGE-L, CIDEr, SPICE, SCS.

4. Phương pháp nghiên cứu

Luận án sử dụng ba nhóm phương pháp chính:

- Tổng quan tài liệu: Khảo sát có hệ thống các hướng chú thích ảnh để nhận diện hạn chế, xác định khoảng trống và đưa ra định hướng nghiên cứu.
- Thiết kế mô hình: Đề xuất chuỗi bốn mô hình encoder–decoder (OD-VR-Cap, RGTranCNet, AMR-GT&RG, CLIP-AMR-GPT), lần lượt bổ sung đồ thị quan hệ, tri thức ConceptNet, biểu diễn AMR/AMR-like và cơ chế hợp nhất đa nguồn (Cross-Fusion, Adaptive Attention, GPT re-ranking).
- Thực nghiệm và đánh giá: Huấn luyện và kiểm thử trên MS COCO và Flickr30K; đánh giá bằng BLEU, METEOR, ROUGE-L, CIDEr, SPICE và

độ đo mới Semantic Consistency Score (SCS), kết hợp so sánh với các baseline và phân tích một số ví dụ điển hình.

5. Những đóng góp chính của luận án

Luận án đề xuất chuỗi bốn mô hình có tính kế thừa:

(i) OD-VR-Cap: Kết hợp đặc trưng đối tượng và đồ thị quan hệ, sử dụng LSTM với chú ý kép trên vùng ảnh và embedding đồ thị.

(ii)RGTranCNet: Thay bộ giải mã LSTM bằng Transformer Decoder với cross-attention và tích hợp tri thức ConceptNet để hiệu chỉnh phân phối từ, qua đó tăng khả năng mô hình hóa quan hệ và xử lý đối tượng hiếm.

(iii) AMR-GT&RG: Tích hợp AMR từ chú thích chuẩn và AMR-like từ đồ thị quan hệ vào Transformer, đồng thời đề xuất độ đo SCS, giúp nâng đáng kể các độ đo ngữ nghĩa trên MS COCO và Flickr30K.

(iv) CLIP-AMR-GPT: Hợp nhất đặc trưng CLIP-ViT, đồ thị quan hệ và AMR/AMR-like bằng Cross-Fusion và Adaptive Attention, kết hợp GPT re-ranking khi suy luận; đạt hiệu năng cao nhất trong chuỗi, đặc biệt trên các độ đo phản ánh chiều sâu và tính nhất quán ngữ nghĩa.

6. Bố cục của luận án

Luận án gồm: Mở đầu, 5 chương nội dung, Kết luận và Tài liệu tham khảo:

- Mở đầu: Nêu tính cấp thiết, mục tiêu, đối tượng, phạm vi, phương pháp và đóng góp chính.
- Chương 1: Tổng quan chú thích ảnh dựa trên học sâu, phân tích khoảng trống và định hướng nghiên cứu.
- Chương 2–5: Trình bày bốn mô hình đề xuất theo hướng kế thừa: OD-VR-Cap, RGTranCNet, AMR-GT&RG, CLIP-AMR-GPT, kèm kiến trúc và kết quả thực nghiệm.
- Kết luận: Tổng hợp kết quả, nhấn mạnh đóng góp và đề xuất hướng nghiên cứu tiếp theo.

CHƯƠNG 1. CHÚ THÍCH ẢNH DỰA TRÊN MẠNG HỌC SÂU

1.1. Giới thiệu

Chú thích ảnh là một trong những bài toán điển hình kết hợp giữa thị giác máy tính (CV) và xử lý ngôn ngữ tự nhiên (NLP). Mục tiêu là sinh câu mô tả tự nhiên phản ánh đúng nội dung thị giác, đối tượng và quan hệ trong ảnh. Đây là bài toán học đa phương thức quan trọng, đóng vai trò nền tảng cho các ứng dụng như hỗ trợ người khiếm thị, tìm kiếm theo nội dung, y học hình ảnh, phân tích hành vi, và kiểm duyệt tự động.

Với sự phát triển của học sâu (Deep Learning), các phương pháp chú thích ảnh đã trải qua nhiều giai đoạn: từ CNN–LSTM cổ điển, đến Transformer, và gần đây là mô hình ngôn ngữ–thị giác tiền huấn luyện (VLMs). Tuy nhiên, hầu hết các phương pháp vẫn gặp hạn chế trong việc hiểu sâu mối quan hệ giữa các đối tượng và tích hợp tri thức ngữ nghĩa ngoài tập huấn luyện, khiến chú thích thiếu tự nhiên hoặc sai lệch ngữ nghĩa.

1.2. Các hướng tiếp cận trong chú thích ảnh

Các phương pháp chú thích ảnh có thể chia thành hai nhóm chính: (i) phương pháp truyền thống; và (ii) phương pháp hiện đại dựa trên học sâu.

1.2.1.Các phương pháp chú thích ảnh truyền thống

Các phương pháp truyền thống gồm hai nhóm chính:

- Dựa trên truy hồi (retrieval-based): So sánh ảnh đầu vào với tập dữ liệu ảnh–chú thích đã gán nhãn, truy xuất mô tả của ảnh tương tự nhất.
- Dựa trên mẫu dữ liệu (template-based): Trích xuất các yếu tố ngữ nghĩa như đối tượng, hành động, bối cảnh, rồi ghép các mẫu cú pháp đã học để sinh câu mô tả.

Các hướng này tuy dễ triển khai nhưng thiếu khả năng khái quát hóa, không thích ứng được với ảnh mới và phụ thuộc vào cấu trúc mẫu.

1.2.2.Các phương pháp chú thích ảnh dựa trên học sâu

Khung phổ biến là kiến trúc encoder–decoder: bộ mã hóa (CNN, object detection, ViT) trích xuất đặc trưng ảnh; bộ giải mã (LSTM, Transformer) sinh chú thích, thường kèm cơ chế chú ý. Các hướng chính gồm:

(i) Hướng CNN–LSTM: Các mô hình như *Show and Tell*, *Show, Attend and Tell* dùng CNN mã hóa ảnh và LSTM (kèm attention) giải mã; nhiều biến thể dùng CNN huấn luyện trước, attention theo vùng, đạt kết quả tốt trên MS COCO, Flickr30K. Hạn chế: chủ yếu xử lý đặc trưng toàn cục hoặc từng vùng, chưa mô hình hóa rõ cấu trúc quan hệ giữa đối tượng.

(ii) Hướng dựa trên đồ thị: Ảnh được mã hóa thành đồ thị (đối tượng, thuộc tính, quan hệ) và lan truyền bằng GCN/GraphSAGE; đặc trưng đồ thị kết hợp với decoder (thường LSTM có attention). Hướng này nắm bắt tốt hơn tương tác giữa đối tượng nhưng liên kết với ngôn ngữ còn rời rạc, chưa tận dụng hết cấu trúc ngữ nghĩa.

(iii) Hướng Transformer-based: Transformer mô hình hóa tốt phụ thuộc dài và tương tác token; nhiều mô hình dùng Transformer Decoder đạt hiệu quả cao trên các độ đo chuẩn. Tuy nhiên, Transformer thuần vẫn bị ràng buộc bởi phân bố dữ liệu huấn luyện, khó xử lý đối tượng hiếm, ngữ cảnh mới và thiếu suy luận dựa trên tri thức ngoài tập dữ liệu.

(iv) Hướng tích hợp tri thức ngoài: ConceptNet, WordNet và các mô hình ngôn ngữ lớn được dùng để bổ sung tri thức ngữ nghĩa và quan hệ (dưới dạng embedding, điều chỉnh trọng số hoặc tín hiệu chú ý). Cách tiếp cận này cải thiện độ chính xác ngữ nghĩa và mô tả đối tượng hiếm, nhưng dễ nhiễu nếu thiếu cơ chế chọn lọc và thường tốn chi phí tính toán.

(v) Hướng sử dụng AMR: AMR (Abstract Meaning Representation) biểu diễn câu dưới dạng đồ thị khái niệm–quan hệ, trừu tượng khỏi cú pháp bề mặt. Một số nghiên cứu dùng AMR hoặc vai trò ngữ nghĩa (SRL) để tăng

tính mạch lạc và nhất quán; tuy nhiên, ánh xạ từ thông tin thị giác sang AMR và tích hợp sâu vào pipeline chú thích ảnh vẫn còn rất hạn chế.

1.3. Khoảng trống và định hướng nghiên cứu

1.3.1. Khoảng trống nghiên cứu

Phân tích các hướng trên cho thấy những hạn chế chính:

- Quan hệ giữa các đối tượng chưa được biểu diễn và khai thác đầy đủ; đồ thị quan hệ chủ yếu đóng vai trò bổ trợ, chưa tích hợp chặt trong quá trình sinh chú thích.
- Khả năng mô tả đối tượng/khái niệm hiếm hoặc ngoài tập huấn luyện còn yếu; tích hợp tri thức ngoài (ConceptNet, LLMs, đồ thị tri thức) mới ở mức sơ khai, dễ gây nhiễu nếu không có cơ chế chọn lọc.
- Ứng dụng AMR trong chú thích ảnh ít được nghiên cứu; thiếu giải pháp nhất quán để chuyển từ đồ thị quan hệ ảnh sang AMR và tích hợp chúng vào kiến trúc học sâu.
- Đặc trưng thị giác truyền thống chưa đủ để biểu diễn ngữ nghĩa sâu; kết nối giữa thị giác, quan hệ và ngôn ngữ chưa thật sự thống nhất.
- Các độ đo (BLEU, METEOR, ROUGE, CIDEr, SPICE) thiên về trùng khớp bề mặt, chưa đánh giá tốt mức độ nhất quán ngữ nghĩa ở cấp câu.

1.3.2. Định hướng nghiên cứu

Để khắc phục các hạn chế trên, luận án định hướng:

- Xây dựng pipeline phát hiện đối tượng – dự đoán quan hệ – tạo đồ thị quan hệ, kết hợp GCN để làm giàu biểu diễn ảnh trước khi sinh chú thích.
- Tích hợp tri thức ngoài (ConceptNet, CLIP/BLIP-2…) bằng cơ chế fusion và cross-modal attention có kiểm soát, vừa hỗ trợ xử lý đối tượng hiếm, vừa hạn chế nhiễu.
- Kết hợp AMR/AMR-like từ chú thích chuẩn và từ đồ thị quan hệ ảnh, nhúng bằng mô hình đồ thị hoặc BERT, và đưa vào Transformer Decoder để tăng chiều sâu ngữ nghĩa trừu tượng.

- Sử dụng backbone hiện đại (ViT, CLIP), kết hợp học đa nhiệm và attention đa nguồn để thu được đặc trưng thị giác–ngữ nghĩa toàn diện hơn.
- Đề xuất độ đo ngữ nghĩa mới Semantic Consistency Score (SCS) dựa trên embedding câu, bổ sung cho các độ đo truyền thống nhằm phản ánh tốt hơn mức độ tương đồng ngữ nghĩa giữa chú thích sinh và chú thích chuẩn.

1.4. Phương pháp thực nghiệm và đánh giá

Phần này giới thiệu khung thực nghiệm thống nhất cho toàn bộ luận án:

- **Dữ liệu:** MS COCO và Flickr30K được sử dụng cho nhiệm vụ chú thích ảnh, với thiết lập chia tập theo Karpathy split; Visual Genome được dùng để huấn luyện/tinh chỉnh mô-đun dự đoán quan hệ giữa các đối tượng.
- **Độ đo đánh giá:** BLEU, METEOR, ROUGE-L, CIDEr, SPICE và độ đo ngữ nghĩa mới SCS, nhằm đánh giá cả trùng khớp bề mặt và mức độ nhất quán ngữ nghĩa giữa chú thích sinh và chú thích chuẩn.
- **Môi trường thực nghiệm:** Mô hình được cài đặt bằng Python, PyTorch và các thư viện liên quan, huấn luyện trên GPU (Google Colab/ T4), với cấu hình được điều chỉnh phù hợp cho từng mô hình.

1.5. Kết chương

Chương 1 đã cung cấp cái nhìn tổng quan, có hệ thống về chú thích ảnh, từ các phương pháp truyền thống đến các mô hình học sâu hiện đại dựa trên CNN–LSTM, đồ thị quan hệ, Transformer, tri thức ngoài tập dữ liệu và AMR. Qua đó, chương chỉ ra những khoảng trống nghiên cứu liên quan đến biểu diễn quan hệ, tích hợp tri thức ngữ nghĩa, khai thác AMR và đánh giá ngữ nghĩa. Những phân tích này đặt nền tảng lý thuyết và kỹ thuật cho toàn bộ luận án, đồng thời định hướng cho việc đề xuất bốn mô hình chú thích ảnh kế thừa và phát triển dần về chiều sâu ngữ nghĩa: OD-VR-Cap, RGTranCNet, AMR-GT&RG và CLIP-AMR-GPT, được trình bày tương ứng trong các Chương 2–5.

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

2.1. Giới thiệu

Đa số phương pháp chú thích ảnh dựa đặc trưng toàn cục hoặc vùng cục bộ nên chưa mô hình hoá đầy đủ quan hệ giữa các đối tượng, làm suy giảm khả năng nắm bắt ngữ cảnh và diễn đạt quan hệ ngữ nghĩa. Chương này đề xuất OD-VR-Cap: kết hợp biểu diễn đối tượng–quan hệ dưới dạng đồ thị quan hệ với LSTM có chú ý kép để hợp nhất thông tin thị giác và cấu trúc khi sinh chú thích. Kiến trúc tổng thể gồm: (i) ODwGCN phát hiện đối tượng; (ii) dự đoán quan hệ; (iii) mã hoá đồ thị; (iv) cơ chế chú ý kép (dual-attention); (v) LSTM decoder sinh mô tả.

Đóng góp chính gồm: (1) ODwGCN tăng độ chính xác phát hiện bằng tri thức đồng-xuất hiện nhãn; (2) Mô-đun dự đoán quan hệ khai thác tri thức quan hệ học từ đồ thị thực thể; (3) Biểu diễn R-Graph thành R-Graph* và embedding đồ thị bằng GraphSAGE; (4) Chú ý kép (trên thị giác và đồ thị) định hướng bộ giải mã LSTM sinh chú thích mạch lạc, giàu ngữ nghĩa.

2.2. Phương pháp đề xuất

2.2.1.Bộ mã hoá hình ảnh

(i) ODwGCN – Phát hiện đối tượng kết hợp GCN: ODwGCN tận dụng quan hệ đồng xuất hiện nhãn để hiệu chỉnh ma trận độ tin cậy của detector nền (SSD/Faster R-CNN/YOLOX). Cấu trúc 2 giai đoạn: (1) học đồ thị tương quan nhãn và embedding bằng GraphSAGE; (2) điều chỉnh xác suất phân lớp theo véc-tơ trọng số học được.

(ii)Nhiệm vụ phân lớp triplet <*subject*, *predicate*, *object*> trên N_{R+1} lớp (kể cả "*none*"). Mô-đun dùng đặc trưng của cặp vùng, hội của chúng (union) và tri thức quan hệ (embedding các thực thể/predicate) học từ đồ thị thực thể (E-Graph) bằng GraphSAGE;

(iii) Xây dựng và biểu diễn R-Graph / R-Graph*: Từ các cặp vùng dự đoán quan hệ, tạo R-Graph (đồ thị có hướng giữa các vùng/đối tượng). Để tương thích mô hình ngôn ngữ, chuyển sang R-Graph*: đỉnh gồm nhãn đối tượng và nhãn quan hệ, cạnh đi qua predicate để "chuẩn hoá" cấu trúc. Các đỉnh của R-Graph* được embedding bằng GraphSAGE (học không giám sát), tạo Z* dùng cho chú ý đồ thị.

2.2.2.Bộ giải mã ngôn ngữ

- Cơ chế chú ý kép: Tại mỗi bước t, tính đồng thời visual-attention trên tập đặc trưng vùng đối tượng và graph-attention trên các node embedding của R-Graph*, thu được hai vector ngữ cảnh $c_t^{(v)}$, và $c_t^{(g)}$. Hai véc-tơ ngữ cảnh này làm đầu vào cho LSTM để dự đoán từ kế tiếp, giúp câu sinh ra vừa bám vùng ảnh vừa đúng quan hệ ngữ nghĩa
- LSTM Decoder: LSTM nhận $\left[x_t, c_t^{(v)}, c_t^{(g)}\right]$, cập nhật trạng thái và sinh phân phối từ qua softmax; toàn bộ huấn luyện theo cross-entropy.

2.3. Thực nghiệm

2.3.1.Dữ liệu và cấu hình thực nghiệm

- Quan hệ: dùng Visual Genome để học tri thức quan hệ (tiền xử lý, rút gọn quan hệ phổ biến).
- Chú thích ảnh: dùng MS COCO theo Karpathy split cho huấn luyện/val/test; mỗi ảnh có 5 chú thích chuẩn.
- Môi trường & kiến trúc: backbone detector phổ biến; GraphSAGE cho hai đồ thị; LSTM decoder với dual-attention.

2.3.2.Độ đo đánh giá

- Phát hiện đối tượng: mAP, mAP@0.5, mAP@0.75.
- Dự đoán quan hệ: Recall@50/100.
- Chú thích ảnh: BLEU-1/4, METEOR, ROUGE-L, CIDEr, SPICE.

2.3.3.Kết quả thực nghiệm

- ODwGCN cải thiện mAP cho nhiều kiến trúc detector, xác nhận lợi ích của tri thức đồng-xuất hiện nhãn.
- VRP+RK đạt Recall@50/100 cao hơn các baseline/đối sánh cùng thiết lập nhờ tri thức quan hệ và đặc trưng union của cặp vùng.
- OD-VR-Cap vượt các CNN–LSTM có chú ý và một số phương pháp dựa scene-graph trên BLEU-4, METEOR, CIDEr (xem Bảng 2.1); các ví dụ định tính cho thấy mô tả đúng hành động/quan hệ hơn.

Bảng 2.1. So sánh độ chính xác chú thích ảnh của các phương pháp trên tập kiểm tra MSCOCO Karpathy

Phương pháp	B@1	B@4	M	R	C	S
Show, attend and tell (Hard-ATT)	71.8	25.0	23.0	-	-	-
Show, attend and tell (Soft-ATT)	70.7	24.3	23.9	-	-	-
Dense_Soft-ATT	68.3	22.9	22.6	53.0	74.3	-
En-De-Cap	70.6	24.3	-	-	-	-
Bi-LS-AttM	68.8	25.2	21.5	-	41.2	-
Image+SceneGraph	67.2	26.1	22.3	-	76.0	-
G-LSTM+att	67.7	20.7	23.9	-	-	-
OD-VR-Cap	**72.6**	**28.3**	**24.8**	**53.4**	**85.1**	17.6

2.4. Kết chương

Chương 2 đã giới thiệu OD-VR-Cap – một mô hình đặt nền cho chuỗi phương pháp của luận án: phát hiện đối tượng được hiệu chỉnh bằng GCN, xây đồ thị quan hệ và embedding, rồi giải mã bằng LSTM với chú ý kép. Kết quả trên MS COCO và nhiệm vụ quan hệ cho thấy tính hiệu quả và tính khả triển của hướng tiếp cận dựa tri thức cấu trúc này. Mô hình là nền tảng cho các mở rộng ở các chương sau.

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

3.1. Giới thiệu

Từ hạn chế của decoder LSTM (khó mô hình hoá phụ thuộc dài) và việc chưa tận dụng tri thức ngữ nghĩa ngoài ảnh, chương này đề xuất mô hình RGTranCNet: thay LSTM bằng Transformer Decoder để học phụ thuộc xa hiệu quả, đồng thời tích hợp tri thức từ ConceptNet nhằm tăng khả năng diễn đạt ngữ nghĩa và khái quát hoá với đối tượng hiếm/chưa thấy trong huấn luyện. Kiến trúc cốt lõi: hợp nhất đa nguồn (đặc trưng vùng, đồ thị quan hệ từ Chương 2, tri thức ConceptNet) qua cross-attention trong Transformer Decoder để tạo chú thích mạch lạc, tự nhiên, giàu ngữ nghĩa.

3.2. Phương pháp chú thích ảnh đề xuất

Trong chương này, mô hình chú thích ảnh được đề xuất thực hiện theo khung encoder- decoder gồm 3 phần chính: (i) Image encoder để học các đại diện của hình ảnh; (ii) Bộ trích xuất tri thức ngữ nghĩa đối tượng để tra cứu tri thức ngữ nghĩa của các đối tượng liên quan từ cơ sở tri thức ConceptNet; (iii) Bộ giải mã (*decoder*), là phần Decoder của mạng Transformer để phát sinh chú thích cho hình ảnh sử dụng đặc trưng thu được từ phần (i), đồng thời bổ sung tri thức ngữ nghĩa ở phần (ii) để nâng cao độ chính xác cho chú thích sinh ra.

3.2.1.Bộ mã hóa hình ảnh

Giữ nguyên pipeline: ODwGCN nâng mAP phát hiện; VRP+RK dự đoán quan hệ để xây dựng R-Graph; mã hoá thành embedding Z của đỉnh đồ thị. Đầu ra encoder cho mỗi ảnh I:

- F_I: đặc trưng vùng đối tượng,
- Z_I: embedding đồ thị quan hệ sau lan truyền ngữ nghĩa.

Hai ma trận này đưa vào decoder.

3.2.2.Bộ trích xuất tri thức ngữ nghĩa đối tượng

ConceptNet được mô hình hoá như đồ thị tri thức $K = (V, E, W)$; với mỗi nhãn đối tượng đã phát hiện, truy vấn $top\text{-}k$ khái niệm liên quan (quan hệ như *IsA, PartOf, UsedFor…*) cùng điểm tin cậy để tạo tập O (đối tượng liên quan cùng với trọng số). Tập O được dùng tại thời điểm sinh từ nhằm điều chỉnh phân phối xác suất, tăng xác suất cho các từ phù hợp ngữ cảnh/tri thức.

3.2.3.Bộ giải mã ngôn ngữ

- Multi-Head Cross-Attention hợp nhất một bước: query từ chuỗi đang sinh, key/value từ F_I (thị giác) và Z_I (đồ thị); đầu ra hai nhánh attention được kết hợp tuyến tính (hệ số điều tiết học được) để tạo biểu diễn ngữ cảnh hợp nhất.
- Điều chỉnh logits theo tri thức ConceptNet: sau tầng dự đoán, logits của những từ xuất hiện trong O được cộng βw (w là độ liên quan), giúp decoder ưu tiên các từ sát nghĩa với bối cảnh và kiến thức thường thức.

3.3. Thực nghiệm và kết quả

3.3.1.Dữ liệu và thiết lập thực nghiệm

Dữ liệu MS COCO (Karpathy split); đồ thị quan hệ và embedding kế thừa từ Chương 2; Transformer Decoder 6 khối/8 heads, d_model=512; tối ưu Adam. ConceptNet dùng để truy xuất tri thức liên quan theo nhãn đối tượng, đưa vào cho decoder trong lúc huấn luyện/suy luận.

3.3.2.Kết quả và bàn luận

So sánh nội bộ giữa hai biến thể (không/có ConceptNet) cho thấy RGTran (không ConceptNet) đạt BLEU-1 = 77.5, BLEU-4 = 34.9, METEOR = 28.3, ROUGE-L = 55.3, CIDEr = 98.4 và SPICE = 18.7; trong khi RGTranCNet (có ConceptNet) tăng đồng loạt lên 79.8, 36.3, 35.6, 57.2, 107.8 và 20.5 tương ứng (xem Bảng 3.1). Đặc biệt, METEOR tăng +7.3, CIDEr tăng +9.4 và SPICE tăng +1.8, cho thấy chú thích sinh ra chính xác hơn về ngữ nghĩa và gần gũi hơn với chú thích tham chiếu.

So với OD-VR-Cap (LSTM với dual attention), cả RGTran và RGTranCNet đều vượt trội trên mọi độ đo; riêng RGTran đã cao hơn OD-VR-Cap 13.3 điểm CIDEr nhờ bộ giải mã Transformer và cơ chế cross-attention hợp nhất (Bảng 3.1). Việc bổ sung ConceptNet trong RGTranCNet tiếp tục cải thiện khả năng khái quát với đối tượng hiếm và suy luận quan hệ chính xác hơn.

Bảng 3.1. So sánh độ chính xác chú thích ảnh của các phương pháp trên tập kiểm tra MSCOCO Karpathy.

Phương pháp	B@1	B@4	M	R	C	S
Show, attend and tell (Hard-ATT)	71.8	25.0	23.0	-	-	-
Show, attend and tell (Soft-ATT)	70.7	24.3	23.9	-	-	-
CNet-NIC	73.1	29.9	25.6	53.9	107.2	-
Bi-LS-AttM	68.8	25.2	21.5	-	41.2	-
Image+SceneGraph	67.2	49.2	23.3	-	76.0	-
G-LSTM+att	67.7	20.7	23.9	-	-	-
OD-VR-Cap	72.6	28.1	24.8	53.4	85.1	17.6
RGTran	77.5	34.9	28.3	55.3	98.4	18.7
RGTranCNet	**79.8**	**36.3**	**35.6**	**57.2**	**107.8**	**20.5**

3.4. Kết chương

RGTranCNet chứng minh rằng thay decoder LSTM bằng Transformer và tích hợp tri thức ConceptNet là hướng hiệu quả và khả triển cho chú thích ảnh: kết quả vượt OD-VR-Cap và nhiều baseline trên BLEU, METEOR, ROUGE-L, CIDEr, SPICE, đồng thời cải thiện ngữ nghĩa của câu sinh. Mô hình đặt nền cho Chương 4, nơi luận án tiếp tục nâng cao chiều sâu ngữ nghĩa bằng cách đưa AMR/AMR-like vào pipeline.

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

4.1. Giới thiệu

Các mô hình chú thích ảnh gần đây (đặc biệt dựa trên Transformer) đã được cải thiện đáng kể, song vẫn thiếu ngữ nghĩa trừu tượng và logic quan hệ sâu. Kế thừa RGTranCNet (Chương 3), chương này đề xuất AMR-GT&RG: tích hợp Abstract Meaning Representation (AMR) vào Transformer Decoder, hợp nhất ba tầng tri thức:

- Thị giác (đặc trưng vùng đối tượng);
- Cấu trúc quan hệ (đồ thị quan hệ chuyển thành AMR-like);
- Ngữ nghĩa trừu tượng (AMR từ chú thích chuẩn, AMR-GT) kèm tri thức ngoài (ConceptNet).

Khung hợp nhất này giúp mô hình hiểu–lý giải–diễn đạt ở mức khái niệm–hành động–quan hệ, tạo chú thích tự nhiên, nhất quán ngữ nghĩa và tổng quát tốt hơn.

4.2. Phương pháp chú thích ảnh đề xuất

4.2.1.Kiến trúc

(i) **Image Encoder**: Bộ mã hóa phát hiện các vùng đối tượng và trích xuất đặc trưng thị giác mức vùng F_I. Đồng thời, từ các đối tượng phát hiện được, hệ thống xây dựng đồ thị quan hệ và tính embedding đồ thị $E_{RG,I}$ bằng mạng nơ-ron tích chập đồ thị (GraphSAGE/GCN).

(ii) **Bộ trích xuất nghữ nghĩa trừu tượng** (Abstract Semantic Extractor)

- AMR-GT: Các chú thích chuẩn được chuyển đổi sang đồ thị AMR, tuyến tính hóa dưới dạng PENMAN và mã hóa bằng BERT để thu được embedding ngữ nghĩa trừu tượng $E_{AMR,I}^{GT}$. Thành phần này chỉ được sử dụng trong giai đoạn huấn luyện.
- AMR-like (AMR-RG): Đồ thị quan hệ ảnh được ánh xạ sang đồ thị AMR-like theo các quy tắc chuyển đổi khái niệm/quan hệ, sau đó được

nhúng bằng GraphSAGE để tạo embedding $E_{AMR,I}^{RG}$, cung cấp lớp ngữ nghĩa trừu tượng gắn với nội dung thị giác.

(iii) **Bộ trích xuất tri thức ngữ nghĩa** (Semantic Knowledge Extractor): Từ mỗi ảnh, hệ thống truy vấn ConceptNet để thu được tập khái niệm liên quan C_I kèm trọng số; các trọng số này sẽ được sử dụng để điều chỉnh xác suất sinh từ trong bước giải mã.

(iv) Bộ giải mã ngôn ngữ Transformer

- Masked Multi-Head Self-Attention: Trong huấn luyện, embedding $E_{AMR,I}^{RG}$ được đưa vào nhánh masked self-attention, giúp truyền tải ngữ nghĩa trừu tượng của ảnh vào quá trình sinh chú thích.
- Cross-Modal Attention hợp nhất 3 nguồn: F_I; $E_{RG,I}$; $E_{AMR,I}^{GT}$ (khi suy luận bỏ $E_{AMR,I}^{GT}$).
- Cross-Modal Attention hợp nhất đa nguồn: Mô-đun chú ý đa phương thức hợp nhất ba nguồn thông tin: đặc trưng thị giác F_I, embedding đồ thị quan hệ $E_{RG,I}$ và embedding AMR $E_{AMR,I}^{GT}$ (khi suy luận bỏ $E_{AMR,I}^{GT}$).
- Điều chỉnh logits dựa trên ConceptNet: Sau khi dự đoán phân phối từ, các logits tương ứng với những từ có lemma thuộc tập C_I được cộng thêm một lượng βw (với $w$là trọng số khái niệm), qua đó thiên lệch mô hình về các lựa chọn từ vựng phù hợp ngữ cảnh và tri thức thường thức.

4.2.2. Thuật toán huấn luyện và sinh chú thích

Trong mô hình AMR-GT&RG, quy trình huấn luyện và suy luận được tổ chức thành hai thuật toán chính:

- TrainTransformerDecoder (giai đoạn huấn luyện): Bộ giải mã Transformer được huấn luyện theo cơ chế *auto-regressive*, sử dụng *masked self-attention* kết hợp với thông tin AMR-like, *cross-attention* đa nguồn (giữa đặc trưng vùng ảnh và embedding đồ thị quan hệ/AMR), đồng thời điều chỉnh logits đầu ra dựa trên tri thức từ ConceptNet. Mô hình được tối ưu bằng hàm mất mát cross-entropy.

- GenerateCaptionAMR (giai đoạn suy luận): Khi suy luận, mô hình chỉ sử dụng các đặc trưng $\{F_I, E_{RG,I}, E_{AMR,I}^{RG}, C_I\}$, tương ứng với đặc trưng vùng ảnh, embedding đồ thị quan hệ, embedding AMR-like và tập khái niệm từ ConceptNet.

4.2.3.Độ đo ngữ nghĩa mới - SCS

Độ đo SCS được định nghĩa như sau:

$$SCS = cosine\left(SBERT(S_{gen}), SBERT(S_{ref})\right) \in [0,1]$$

Trong đó, S_{gen} là chú thích do mô hình sinh ra, S_{ref} là chú thích chuẩn, và SBERT là mô hình mã hóa câu thành véc-tơ ngữ nghĩa.

Mục tiêu của SCS là đo lường mức độ nhất quán ngữ nghĩa giữa hai câu trong không gian embedding, giúp khắc phục hạn chế của các độ đo dựa trên n-gram như BLEU, METEOR, ROUGE vốn chủ yếu phản ánh sự trùng khớp bề mặt.

4.3. Thực nghiệm và kết quả

4.3.1.Thiết lập thực nghiệm

- Dữ liệu: Mô hình được đánh giá trên hai bộ dữ liệu chuẩn MS COCO và Flickr30K, với cách chia tập huấn luyện/validation/kiểm thử tuân theo Karpathy split;
- Encoder: Trên MS COCO, luận án sử dụng ODwGCN để phát hiện đối tượng và xây dựng đồ thị quan hệ; trên Flickr30K, sử dụng Faster R-CNN kết hợp GCN/GraphSAGE để trích xuất đặc trưng vùng và embedding đồ thị.
- Decoder: Bộ giải mã là Transformer Decoder gồm 6 lớp, 8 đầu attention, kích thước ẩn $d_{model} = 768$; tối ưu bằng Adam với learning rate 1×10^{-4}, batch size 32; huấn luyện trên GPU T4 (Google Colab)..
- AMR: Ngữ nghĩa trừu tượng từ chú thích chuẩn được trích xuất bằng pipeline NeuralAMR → PENMAN → BERT để thu được embedding

$E^{GT}_{AMR,I}$; đồ thị quan hệ ảnh được chuyển sang AMR-like và nhúng bằng GraphSAGE để tạo $E^{RG}_{AMR,I}$.

- ConceptNet: Với mỗi ảnh, hệ thống truy vấn ConceptNet để lấy top-k khái niệm và liên kết liên quan nhất tới các đối tượng trong ảnh, dùng làm tri thức ngữ nghĩa bổ sung trong quá trình sinh chú thích.

4.3.2.Kết quả và bàn luận

Kết quả trên MS COCO: Như thể hiện trong Bảng 4.1, mô hình AMR-GT&RG đạt B@1 = 81,2; B@4 = 39,5; METEOR = 37,2; ROUGE-L = 59,9; CIDEr = 136,7; SPICE = 25,1 và SCS = 89,1. So với các phương pháp SOTA gần đây như COS-Net, X-Transformer, SGAE, MLA-LRN, GIC-SSF…, AMR-GT&RG đạt kết quả nổi trội trên các độ đo nhấn mạnh khía cạnh ngữ nghĩa và tính mạch lạc (METEOR, ROUGE-L, CIDEr, SPICE, SCS). Đặc biệt, so với RGTranCNet ở Chương 3, AMR-GT&RG cải thiện thêm 28,9 điểm CIDEr, 6,8 điểm SCS và 1,6 điểm METEOR (xem Bảng 4.1), cho thấy đóng góp đáng kể của tầng AMR trong việc củng cố tính nhất quán ngữ nghĩa và độ gần gũi với chú thích tham chiếu.

Kết quả trên Flickr30K: Bảng 4.2 cho thấy AMR-GT&RG đạt B@1 = 79,1; B@4 = 36,4; METEOR = 35,6; ROUGE-L = 56,7; CIDEr = 94,5; SPICE = 22,7 và SCS = 87,2. Mô hình tiếp tục thể hiện ưu thế trên các thước đo ngữ nghĩa (METEOR, CIDEr, SPICE, SCS) so với các phương pháp đối sánh, khẳng định khả năng tổng quát hóa khi chuyển miền từ MS COCO sang Flickr30K và duy trì chất lượng mô tả trong bối cảnh chú thích đa dạng.

Đánh giá ảnh hưởng từng thành phần của mô hình trên MS COCO: Phân tích thành phần cho thấy từng nguồn AMR đều đóng góp tích cực nhưng cấu hình kết hợp cho hiệu năng tốt nhất: sử dụng riêng AMR-GT cho CIDEr = 132,5; SPICE = 24,6; SCS = 87,3, trong khi chỉ dùng AMR-RG cho CIDEr = 129,8; SPICE = 23,9; SCS = 84,5. Kết hợp cả hai nguồn trong AMR-GT&RG giúp đạt giá trị cao nhất trên tất cả độ đo (CIDEr = 136,7; SPICE =

25,1; SCS = 89,1, xem Bảng 4.1), cho thấy AMR từ ngôn ngữ cung cấp tri thức khái quát, còn AMR-like từ ảnh bảo đảm sự đồng bộ với nội dung thị giác và cấu trúc quan hệ.

4.4. Kết chương

Trong chương này, luận án đề xuất mô hình AMR-GT&RG cho chú thích ảnh, mở rộng từ RGTranCNet bằng cách tích hợp ba nguồn tri thức: (i) AMR từ chú thích chuẩn (AMR-GT), (ii) AMR-like suy từ đồ thị quan hệ ảnh, và (iii) tri thức ngoài từ ConceptNet để điều chỉnh xác suất sinh từ. Các biểu diễn AMR/AMR-like được nhúng bằng GraphSAGE và đưa vào Transformer Decoder qua cơ chế chú ý đa nguồn, giúp tăng chiều sâu ngữ nghĩa và nắm bắt cấu trúc khái niệm–quan hệ trong ảnh.

Thực nghiệm trên MS COCO và Flickr30K cho thấy AMR-GT&RG vượt trội trên các độ đo nhấn mạnh ngữ nghĩa, phản ánh khả năng mô hình hóa tốt hơn quan hệ giữa các thực thể và bối cảnh, dù BLEU không luôn cao nhất. Hạn chế chính là phụ thuộc chú thích chuẩn để trích xuất AMR và chi phí tính toán cao khi xử lý/nhúng AMR. Các hướng mở gồm: tận dụng mô hình thị giác–ngôn ngữ tiền huấn luyện, mở rộng nguồn tri thức (Wikidata, WordNet, LLMs) và nghiên cứu khả năng tổng quát hóa đa ngôn ngữ.

Trên nền tảng này, Chương 5 tiếp tục xây dựng khung hợp nhất ngữ nghĩa đa phương thức với CLIP, đồ thị quan hệ và AMR/AMR-like, kết hợp Adaptive Attention và GPT-based Re-ranking để cải thiện tính chính xác, giàu ngữ nghĩa và sự tự nhiên của chú thích ảnh.

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

5.1. Giới thiệu

Chương 5 tập trung khắc phục các hạn chế còn lại của các mô hình ở Chương 2–4, cụ thể là: (i) chưa có cơ chế hợp nhất linh hoạt giữa các nguồn tri thức không đồng nhất (thị giác, cấu trúc quan hệ, ngữ nghĩa trừu tượng), (ii) độ tự nhiên và mạch lạc ngôn ngữ của chú thích còn chưa ổn định, và (iii) thiếu một bước kiểm soát chất lượng ngôn ngữ ở tầng cao. Để giải quyết các vấn đề này, luận án đề xuất mô hình CLIP-AMR-GPT theo khung encoder–decoder hợp nhất tri thức đa nguồn, trong đó encoder kết hợp đặc trưng thị giác–ngôn ngữ từ CLIP, embedding đồ thị quan hệ và biểu diễn AMR/AMR-like, còn decoder Transformer được trang bị cơ chế Cross-Fusion và Adaptive Attention để điều tiết đóng góp của từng nguồn tri thức trong quá trình giải mã. Ở giai đoạn suy luận, mô-đun GPT-based re-ranking được sử dụng để chọn chú thích cuối cùng có tính tự nhiên và mạch lạc cao nhất. Phương pháp chú thích ảnh đề xuất

5.1.1. Tổng quan kiến trúc

CLIP-AMR-GPT được xây dựng theo kiến trúc encoder–decoder hợp nhất tri thức đa nguồn. Ở encoder, mô hình đồng thời trích xuất: (a) đặc trưng CLIP-ViT (token [CLS] toàn cục và 256 patch tokens); (b) embedding đồ thị quan hệ R-Graph (đối tượng–quan hệ) bằng GraphSAGE; và (c) embedding AMR từ chú thích chuẩn (AMR-GT) qua pipeline NeuralAMR → PENMAN → BERT. Ở decoder, Transformer sinh chú thích với hai cơ chế: Cross-Fusion Attention để hợp nhất ba nguồn đặc trưng {*CLIP, R-Graph, AMR-GT*}, và Adaptive Attention để đưa AMR-like (nhúng từ R-Graph) vào tại các thời điểm giải mã phù hợp. Sau beam search, mô-đun GPT-based re-ranking chọn câu chú thích có GPTScore cao nhất làm đầu ra cuối cùng.

5.1.2.Bộ mã hóa hình ảnh

Ảnh qua CLIP ViT-L/14 cho véc-tơ tổng quát và patch-level, giúp mô hình vừa nắm bức tranh toàn cục, vừa giữ chi tiết cục bộ. Song song, mô hình phát hiện đối tượng và dự đoán quan hệ từ Visual Genome tạo R-Graph; GraphSAGE hai chiều học embedding nút/quan hệ thành $E_{RG,I}$, cung cấp cấu trúc ngữ nghĩa đối tượng–quan hệ.

5.1.3.Trích xuất ngữ nghĩa trừu tượng

Chú thích chuẩn được phân tích AMR và nhúng thành $E_{AMR,I}^{GT}$ - nguồn "chuẩn hoá ngữ nghĩa" ổn định trong huấn luyện. Từ R-Graph, chúng tôi ánh xạ sang AMR-like (khái niệm/quan hệ theo chuẩn AMR) và nhúng bằng GraphSAGE, thu được $E_{AMR,I}^{RG}$ - nguồn ngữ nghĩa gắn ảnh ở thời điểm suy luận. Hai nguồn AMR bổ trợ: AMR-GT giúp định hướng khái niệm, AMR-like đưa cấu trúc đúng ảnh vào từng bước sinh.

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

Thay vì cộng/trung bình, mô hình giữ nguyên toàn bộ token/đỉnh của ba miền. Mỗi miền được chiếu sang không gian Key–Value riêng, sau đó nối thành K_{multi}, V_{multi}. Truy vấn từ trạng thái decoder sẽ học trọng số lựa chọn thông tin phù hợp từ {*CLIP, R-Graph, AMR-GT*} theo ngữ cảnh sinh hiện thời, tránh mất mát chi tiết và cho phép bù trừ liên miền.

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

Luận án đề xuất cơ chế Adaptive Attention để tích hợp linh hoạt thông tin từ đồ thị AMR-like vào quá trình sinh chú thích. Ở mỗi bước t, trạng thái ẩn h_{t-1} được dùng tính véc-tơ cổng g_t, điều chỉnh trọng số các embedding nút AMR-like trước khi đưa vào lớp Masked Multi-Head Attention làm Key/Value. Nhờ đó, mô hình chọn lọc các nút phù hợp với ngữ cảnh hiện thời, cải thiện tính mạch lạc và chiều sâu ngữ nghĩa của câu chú thích.

5.1.6.Tái xếp hạng qua GPT

Decoder sinh k ứng viên bằng beam search. Mỗi câu được chấm GPTScore (log-probability trung bình theo GPT-2 medium); câu cao nhất được chọn. Mô-đun này đóng vai trò bộ lọc ngôn ngữ, cải thiện trôi chảy, ngữ pháp, liên kết mà không thay đổi mục tiêu huấn luyện chính.

5.1.7.Huấn luyện và triển khai

Mô hình huấn luyện bằng cross-entropy với teacher forcing; tối ưu Adam; decoder N=6 block, 8 heads, số chiều của véc-tơ biểu diễn từ là 512. Beam size = 5; re-ranking dùng GPT-2 medium. Thực nghiệm chạy ổn định trên GPU T4, dễ thay thế backbone (CLIP/LLMs), dễ mở rộng tri thức (Wikidata/WordNet/LLMs) và đa ngôn ngữ/đa miền.

5.2. Thực nghiệm và kết quả

5.2.1.Thiết lập thực nghiệm

Đánh giá trên MS COCO và Flickr30K (thiết lập chia tập/tiền xử lý kế thừa Chương 4). Độ đo: BLEU-1/4, METEOR, ROUGE-L, CIDEr, SPICE, và SCS (Semantic Consistency Score). Thực thi bằng PyTorch 2.0, Python 3.9, beam = 5, re-ranking bằng GPT-2 medium.

5.2.2.Kết quả và bàn luận

MS COCO (Karpathy split): CLIP-AMR-GPT đạt B@1 = 82.9, B@4 = 41.4, M = 38.5, R = 61.8, C = 144.2, S = 26.7, SCS = 91.7, đứng đầu trên hầu hết độ đo. So với AMR-GT&RG (81.2 / 39.5 / 37.2 / 59.9 / 136.7 / 25.1 / 89.1), mô hình cải thiện +1.7 B@1, +1.9 B@4, +1.3 METEOR, +1.9 ROUGE-L, +7.5 CIDEr, +1.6 SPICE và +2.6 SCS, đồng thời vượt CLIP-Captioner (C = 139.4; S = 23.9; M = 30.0). Điều này khẳng định hiệu quả của hợp nhất tri thức đa nguồn kết hợp điều tiết động và re-ranking bằng mô hình ngôn ngữ.

Flickr30K: CLIP-AMR-GPT đạt B@1 = 80.5, B@4 = 38.2, M = 36.9, R = 58.2, C = 102.8, S = 24.0, SCS = 89.4, tiếp tục vượt AMR-GT&RG (79.1

/ 36.4 / 35.6 / 56.7 / 94.5 / 22.7 / 87.2). Mức tăng +1.8 B@4, +1.3 METEOR, +1.5 ROUGE-L, +8.3 CIDEr, +1.3 SPICE và +2.2 SCS cho thấy khả năng tổng quát hoá tốt khi chuyển miền và xử lý phong cách mô tả đa dạng.

Tổng thể, ưu thế của CLIP-AMR-GPT đến từ sự kết hợp có kiểm soát giữa gắn kết ngữ nghĩa với ảnh và độ trôi chảy ngôn ngữ: CLIP cung cấp khái niệm thị giác giàu ý nghĩa, AMR/AMR-like bổ sung cấu trúc vai–hành động–thực thể, Adaptive Attention điều tiết đóng góp từng nguồn theo ngữ cảnh, còn GPT re-ranking tinh chỉnh ngữ pháp và tính tự nhiên. Sự cộng hưởng này giúp cải thiện đồng đều các thước đo ngữ nghĩa (CIDEr, SPICE, METEOR, SCS) và thước đo n-gram/cấu trúc (BLEU, ROUGE).

5.3. Kết luận

Trong Chương này, luận án đã đề xuất mô hình chú thích ảnh CLIP-AMR-GPT, một kiến trúc encoder–decoder hợp nhất đa nguồn tri thức, kết hợp đặc trưng thị giác–ngôn ngữ từ CLIP, biểu diễn ngữ nghĩa trừu tượng từ AMR/AMR-like, cơ chế Adaptive Attention và bước re-ranking bằng mô hình ngôn ngữ GPT. Thực nghiệm trên MS COCO và Flickr30K cho thấy mô hình đạt kết quả vượt trội so với nhiều phương pháp tiên tiến, đặc biệt trên các độ đo nhấn mạnh khía cạnh ngữ nghĩa như CIDEr, METEOR và SPICE; phân tích ablation khẳng định vai trò bổ trợ lẫn nhau của CLIP, AMR và GPT re-ranking trong việc nâng cao độ chính xác ngữ nghĩa và độ mạch lạc của chú thích.

Hướng phát triển tiếp theo bao gồm mở rộng mô hình sang đa ngôn ngữ (đặc biệt là tiếng Việt) và tích hợp các mô hình ngôn ngữ lớn mới (như GPT-4, Gemini) để xử lý các bối cảnh phức tạp hơn. Chương Kết luận và Kiến nghị sau đó sẽ tổng hợp toàn bộ kết quả nghiên cứu, nhấn mạnh đóng góp khoa học và đề xuất các hướng nghiên cứu tiếp nối cho bài toán chú thích ảnh dựa trên mạng nơ-ron sâu.

KẾT LUẬN

Phần kết luận tổng hợp toàn bộ quá trình nghiên cứu của luận án, từ việc xác lập tính cấp thiết, mục tiêu, phạm vi và phương pháp nghiên cứu, đến hệ thống hóa các hướng tiếp cận hiện có và đề xuất chuỗi bốn mô hình chú thích ảnh theo lộ trình kế thừa và mở rộng. Trên cơ sở đó, phần này khái quát các kết quả đạt được, chỉ ra một số hạn chế, trình bày các vấn đề còn bỏ ngỏ và định hướng nghiên cứu tiếp theo.

1. Tổng kết nội dung nghiên cứu và kết quả đạt được

Luận án tiếp cận bài toán chú thích ảnh thông qua bốn mô hình: OD-VR-Cap, RGTranCNet, AMR-GT&RG và CLIP-AMR-GPT. Các mô hình lần lượt khai thác đồ thị quan hệ, tri thức ConceptNet, biểu diễn AMR/AMR-like, đặc trưng CLIP và GPT re-ranking nhằm tăng cường khả năng biểu diễn ngữ nghĩa và nâng cao chất lượng chú thích ảnh. Các mô hình được đánh giá trên MS COCO và Flickr30K bằng BLEU, METEOR, ROUGE-L, CIDEr, SPICE và SCS.

Kết quả thực nghiệm cho thấy hiệu năng được cải thiện qua từng thế hệ mô hình. Trên MS COCO, CIDEr tăng từ 85.1 ở OD-VR-Cap lên 107.8 ở RGTranCNet, 136.7 ở AMR-GT&RG và 144.2 ở CLIP-AMR-GPT; SCS của CLIP-AMR-GPT đạt 91.7. Trên Flickr30K, CLIP-AMR-GPT đạt B@1 = 80.5, B@4 = 38.2, METEOR = 36.9, ROUGE-L = 58.2, CIDEr = 102.8, SPICE = 24.0 và SCS = 89.4, khẳng định hiệu quả của chiến lược hợp nhất tri thức đa nguồn và khả năng tổng quát hóa của mô hình.

2. Hạn chế của luận án

Bên cạnh các kết quả đạt được, luận án vẫn còn một số hạn chế. Các thực nghiệm chủ yếu được báo cáo theo một lần huấn luyện chính, chưa thực hiện đầy đủ nhiều lần huấn luyện độc lập để tính trung bình, độ lệch chuẩn và kiểm định ý nghĩa thống kê. Luận án mới đánh giá trên MS COCO và Flickr30K, chưa mở rộng sang NoCaps, dữ liệu out-of-domain, zero-shot

hoặc chú thích ảnh tiếng Việt. Ngoài ra, chất lượng biểu diễn AMR còn phụ thuộc vào công cụ AMR parser, trong khi các mô hình tích hợp nhiều nguồn tri thức có chi phí tính toán cao hơn các kiến trúc encoder–decoder đơn giản.

3. Những vấn đề còn bỏ ngỏ

Từ các kết quả và hạn chế của luận án, một số hướng nghiên cứu tiếp theo có thể tập trung vào: tăng cường khả năng suy luận ngữ nghĩa sâu, bao gồm mục đích hành động, quan hệ nhân quả và ngữ cảnh phức tạp; cải thiện khả năng xử lý đối tượng hiếm, dữ liệu ngoài miền và tình huống zero-shot thông qua CLIP, LLMs và tri thức ngoài; chuẩn hóa các độ đo đánh giá ngữ nghĩa như SCS, kết hợp đánh giá tự động với đánh giá của con người; mở rộng bài toán sang chú thích ảnh tiếng Việt và các ngữ cảnh ít tài nguyên; đồng thời nâng cao tính giải thích được, giảm chi phí tính toán và triển khai mô hình trong các môi trường thực tế như hỗ trợ người khiếm thị, giáo dục, lưu trữ số, tìm kiếm đa phương thức và trợ lý thị giác.

4. Tổng kết

Luận án đã xây dựng một lộ trình nghiên cứu nhất quán từ khai thác quan hệ thị giác, tích hợp tri thức ngoài, biểu diễn AMR đến hợp nhất ngữ nghĩa đa phương thức có điều tiết và GPT re-ranking. Các kết quả đạt được góp phần nâng cao độ chính xác, chiều sâu ngữ nghĩa và tính tự nhiên của chú thích ảnh, đồng thời làm rõ vai trò của từng thành phần trong kiến trúc đề xuất. Đây là cơ sở để tiếp tục phát triển các mô hình chú thích ảnh giàu ngữ nghĩa, có khả năng tổng quát hóa tốt hơn và phù hợp với các ứng dụng thực tiễn trong nước cũng như quốc tế.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

[1] **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.

[2] **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.

[3] **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.

[4] **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 (**Scopus-Q4**).

[5] **N. V. Thinh**, T. V. Lang and V. T. Thanh (2025). *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 (**SCIE-Q1**).

[6] **N. V. Thinh**, T. V. Lang, and N. M. Hai (2025). *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 (**Scopus, Q2**).