p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

CẢI THIỆN HIỆU QUẢ CHÚ THÍCH ẢNH TIẾNG VIỆT DỰA TRÊN HỢP NHẤT ĐẶC TRƯNG BẰNG CƠ CHẾ CHÚ Ý

Hoàng Anh Công[1], Nguyễn Đình Công[2], Phạm Thế Anh[3]

TÓM TẮT

Bài báo đề xuất một phương pháp chú thích ảnh tiếng Việt dựa trên hợp nhất đặc trưng qua cơ chế chú ý (attention fusion). Trong đó đặc trưng hình ảnh được kết hợp với đặc trưng nhúng ngữ nghĩa văn bản sinh từ mô hình tiền huấn luyện. Cách tiếp cận này giúp mô hình tăng khả năng căn chỉnh ngữ nghĩa và sinh ra mô tả giàu thông tin hơn so với các mô hình cơ sở thường được sử dụng. Kết quả thực nghiệm trên hai bộ dữ liệu UIT-ViIC và KTVIC cho thấy phương pháp đề xuất giúp cải thiện chỉ số CIDEr khoảng 6%-10% so với các mô hình đối chứng, đồng thời đạt hiệu quả tốt trên các thước đo BLEU, METEOR, chứng minh hiệu quả và tính khả thi của hướng tiếp cận.

Từ khoá: *Chú thích ảnh, hợp nhất chú ý, biểu diễn văn bản.*

DOI: *https://doi.org/10.70117/hdujs.84.2.2026.1139*

1. ĐẶT VẤN ĐỀ

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của thị giác máy tính và xử lý ngôn ngữ tự nhiên, bài toán tạo chú thích ảnh (image captioning) đã trở thành một hướng nghiên cứu quan trọng trong các hệ thống đa phương thức. Mục tiêu của bài toán là tự động sinh ra mô tả ngôn ngữ tự nhiên phản ánh chính xác nội dung và ngữ cảnh của hình ảnh đầu vào. Bài toán này có nhiều ứng dụng thực tiễn như truy vấn ảnh theo ngữ nghĩa, hỗ trợ người khiếm thị, quản lý và lập chỉ mục nội dung đa phương tiện, cũng như các hệ thống trợ lý thông minh. Về bản chất, chú thích ảnh đòi hỏi mô hình phải căn chỉnh và hợp nhất hiệu quả giữa biểu diễn thị giác và biểu diễn ngôn ngữ, thường được hiện thực hóa thông qua các kiến trúc chú ý (attention) và transformer hiện đại [1] [2].

Phần lớn các nghiên cứu chú thích ảnh trước đây tập trung vào tiếng Anh và được kiểm chứng trên các bộ dữ liệu quy mô lớn như MS COCO Captions [3]. Tuy nhiên, khi chuyển sang các ngôn ngữ ít tài nguyên như tiếng Việt, bài toán trở nên thách thức hơn do hạn chế về dữ liệu gán nhãn, sự khác biệt về cấu trúc ngôn ngữ, hiện tượng đa nghĩa và tách từ, cũng như sự mất cân bằng trong phân bố từ vựng. Các yếu tố này khiến nhiều mô hình chú thích huấn luyện theo hướng truyền thống dễ sinh ra các mô tả chung chung, thiếu chi tiết hoặc không bám sát trọng tâm hình ảnh.

[1]*Trường Đại học Văn Hoá - Thể thao - Du lịch Thanh Hoá*
[2]*Khoa Kỹ thuật, Công nghệ và Truyền thông, Trường Đại học Hồng Đức; Email: nguyendinhcong@hdu.edu.vn*
[3]*Phòng Tổ chức - Hành chính - Quản trị, Trường Đại học Hồng Đức*

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Về tình hình nghiên cứu, các công trình gần đây cho thấy hiệu năng của chú thích ảnh không chỉ phụ thuộc vào bộ giải mã (decoder), mà còn chịu ảnh hưởng mạnh bởi chất lượng đặc trưng ngôn ngữ và cơ chế hợp nhất đa phương thức. Li et al. đề xuất một mô hình tạo chú thích ảnh dựa trên sự chú ý đa đầu để hợp nhất đặc trưng hình ảnh và văn bản, trong đó văn bản không chỉ đóng vai trò là đầu ra mà còn được khai thác như một nguồn đặc trưng ngữ nghĩa nhằm dẫn hướng cho quá trình sinh câu [4]. Kết quả thực nghiệm cho thấy việc tăng cường biểu diễn văn bản và hợp nhất chúng với đặc trưng thị giác giúp cải thiện đáng kể độ chính xác và tính mạch lạc của chú thích.

Song song với đó, một số nghiên cứu khác tập trung vào hướng làm giàu đặc trưng đa phương thức nhằm giảm phụ thuộc vào dữ liệu gán nhãn đầy đủ. Cheng et al. đề xuất Echo, một khung phương pháp nhấn mạnh rằng việc cải thiện và mở rộng không gian đặc trưng hình ảnh và đặc trưng ngôn ngữ có thể nâng cao chất lượng chú thích, đặc biệt trong bối cảnh dữ liệu hạn chế hoặc phân bố khái niệm không đầy đủ [5]. Mặc dù Echo được thiết kế cho bài toán học không cần huấn luyện (zero-shot learning) và miền ảnh viễn thám, tư tưởng cốt lõi của công trình cho thấy việc làm giàu đặc trưng giai đoạn giải mã là một hướng tiếp cận hiệu quả nhằm nâng cao chất lượng mô tả.

Đối với tiếng Việt, một số bộ dữ liệu chú thích đã được công bố trong những năm gần đây nhằm tạo nền tảng cho nghiên cứu và đánh giá. UIT-ViIC là bộ dữ liệu đầu tiên phục vụ đánh giá chú thích ảnh bằng tiếng Việt, được xây dựng dựa trên ảnh COCO với chú thích do con người gán nhãn [6]. KTVIC tiếp tục mở rộng theo miền đời sống, cung cấp thêm sự đa dạng về ngữ cảnh và cấu trúc mô tả [7]. Ngoài ra, VieCap4H hướng tới miền y tế, nhấn mạnh tính ứng dụng trong các kịch bản chuyên biệt [8]. Tuy nhiên, so với các bộ dữ liệu tiếng Anh, các tập dữ liệu tiếng Việt vẫn còn hạn chế về quy mô và độ phủ khái niệm, đặt ra yêu cầu về các phương pháp hiệu quả nhưng nhẹ, không phụ thuộc vào kiến trúc phức tạp hoặc dữ liệu huấn luyện quá lớn. Tiếng Việt là ngôn ngữ ít tài nguyên, có đặc thù về tách từ không rõ ràng và hiện tượng đa nghĩa phụ thuộc mạnh vào ngữ cảnh, trong khi các bộ dữ liệu chú thích ảnh thường có quy mô hạn chế. Những yếu tố này khiến mô hình dễ sinh câu thiếu nhất quán ngữ nghĩa hoặc không ổn định về cấu trúc. Trong bối cảnh đó, việc sử dụng biểu diễn nhúng ngữ nghĩa toàn cục của văn bản tiếng Việt đóng vai trò như một ngữ cảnh ngữ nghĩa cấp cao, giúp định hướng quá trình sinh câu, giảm nhiễu do mơ hồ ngôn ngữ và cải thiện tính nhất quán ngữ nghĩa của chú thích được tạo ra.

Tính thời sự của bài toán thể hiện ở một số khía cạnh. Thứ nhất, nhu cầu tự động sinh mô tả ảnh tiếng Việt ngày càng gia tăng trong các hệ thống nội dung số và trợ lý đa phương thức. Thứ hai, khoảng cách tài nguyên giữa tiếng Việt và tiếng Anh khiến việc huấn luyện các mô hình lớn từ đầu trở nên kém khả thi, thúc đẩy việc tận dụng các mô hình tiền huấn luyện đa phương thức như CLIP [9].

Từ các phân tích trên, bài báo này đặt ra vấn đề nghiên cứu: Liệu có thể cải thiện chất lượng tạo chú thích ảnh tiếng Việt bằng một cải tiến đơn giản, dựa trên việc tăng cường đặc trưng ngôn ngữ theo hướng trích xuất đặc trưng văn bản và sau đó hợp nhất đặc trưng hình ảnh với đặc trưng ngôn ngữ bằng cơ chế chú ý, mà không cần đến các kiến trúc sinh đặc trưng hoặc chiến lược huấn luyện phức tạp. Trên cơ sở đó, nghiên cứu tập trung đánh giá hiệu quả của việc làm giàu đặc trưng văn bản và hợp nhất cơ chế chú ý (attention fusion) trong bối cảnh chú thích ảnh bằng tiếng Việt, từ đó có cơ sở để phát triển các mô hình chú

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

thích tiên tiến trong tương lai. Khác với các hướng tiếp cận trên [4][5], bài báo này không thực hiện căn chỉnh trực tiếp giữa các đặc trưng ảnh và văn bản. Thay vào đó, biểu diễn nhúng văn bản tiếng Việt ở mức toàn cục được sử dụng như một ngữ cảnh ngữ nghĩa để điều hướng quá trình tổng hợp đặc trưng ảnh thông qua cơ chế chú ý. Cách thiết kế này cho phép tích hợp thông tin ngôn ngữ ở mức khái quát, giảm độ phức tạp mô hình, đồng thời phù hợp với bối cảnh dữ liệu hạn chế và tính đa dạng ngữ nghĩa của tiếng Việt.

2. PHƯƠNG PHÁP NGHIÊN CỨU

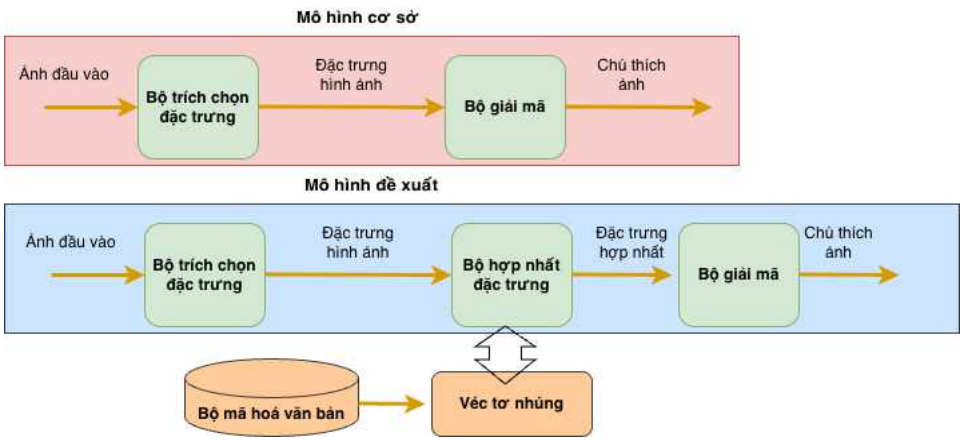
2.1. Mô hình đề xuất cho tạo chú thích ảnh tiếng Việt

Bài toán tạo chú thích ảnh nói chung và chú thích tiếng Việt nói riêng được mô hình hoá theo dạng sinh chuỗi có điều kiện, được thể hiện minh hoạ ở Hình 1. Với ảnh đầu vào I và chú thích tiếng Việt tương ứng Y = (y_1…. y_T) thì mô hình cần ước lượng theo phương trình (1) với θ là tham số mô hình.

$$\max_\theta p(Y|I;\theta) \quad (1)$$

Đặc trưng hình ảnh được trích xuất bởi bộ mã hoá ảnh theo phương trình (2).

$$\mathbf{V} = f_{img}(I) \in \mathbb{R}^{M\times d} \quad (2)$$



Hình 1. Hình minh hoạ mô tả sơ đồ của mô hình chú thích ảnh.

Trong đó, *M* là số lượng của đặc trưng hình ảnh và *d* là số chiều không gian đặc trưng. Song song, chú thích tiếng Việt Y được ánh xạ sang một véc tơ ngữ nghĩa toàn cục thông qua bộ mã hóa văn bản tiền huấn luyện (ví dụ CLIP text encoder), nhằm thu được biểu diễn giàu ngữ nghĩa hơn so với véc tơ nhúng học từ đầu như ở phương trình (3).

$$\mathbf{t} = f_{text}(Y) \in \mathbb{R}^{d_t} \quad (3)$$

Để đưa về cùng không gian với V, sử dụng phép chiếu tuyến tính như ở (4).

$$\mathbf{T} = \mathbf{W}_t\mathbf{t} \in \mathbb{R}^{\boldsymbol{d}} \quad (4)$$

Trong đó $\boldsymbol{t}$ là biểu diễn nhúng văn bản ban đầu và $\mathbf{W}_t$ là ma trận chiếu học được để ánh xạ biểu diễn này về cùng không gian đặc trưng với ảnh với $\mathbf{W}_t \in \mathbb{R}^{d\times d_t}$. Ở đây, $\mathbf{T}$ được hiểu là đại diện cho ngữ cảnh của ngôn ngữ toàn cục của câu tiếng Việt và được kỳ vọng giúp mô hình duy trì tính nhất quán ngữ nghĩa khi sinh câu. Trên cơ sở đó, cơ chế hợp

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

nhất được theo thiết theo hướng tập trung vào ngữ cảnh của văn bản và đặc trưng hình ảnh. Với đặc trưng hình ảnh đóng vai trò truy vấn (query - **Q**) còn đặc trưng văn bản tăng cường cung cấp key/value để điều hướng quá trình tập trung. Véc tơ **T** được mở rộng thành chuỗi $\tilde{\mathbf{T}}$ theo chiều M như ở (5). Lưu ý rằng, mặc dù biểu diễn nhúng văn bản toàn cục được lặp lại theo các vị trí không gian, cơ chế chú ý không bị suy biến vì các véc tơ truy vấn Q được sinh ra từ đặc trưng ảnh tại từng vị trí và do đó khác nhau theo không gian. Nhờ đó, mỗi vị trí ảnh vẫn tạo ra phân bố chú ý riêng, phản ánh mức độ phù hợp ngữ nghĩa giữa đặc trưng ảnh cục bộ và ngữ cảnh văn bản toàn cục.

$$\tilde{\mathbf{T}} = \text{Repeat}(\mathbf{T}, M) \in \mathbb{R}^{M \times d} \tag{5}$$

Sau đó quá trình hợp nhất chú ý được tính bởi (6).

$$\mathbf{Q} = \mathbf{V}\mathbf{W}_{\text{Q}}, \mathbf{K} = \tilde{\mathbf{T}}\mathbf{W}_{\text{K}}, \mathbf{U} = \tilde{\mathbf{T}}\mathbf{W}_{\text{V}} \tag{6}$$

$$\mathbf{A} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\text{T}}}{\sqrt{\mathbf{d}}}), \ \mathbf{F} = \mathbf{A}.\mathbf{U}$$

$\mathbf{W}_{\text{Q}}, \mathbf{W}_{\text{K}}, \mathbf{W}_{\text{V}}$ là các ma trận học được dùng để ánh xạ sang không gian truy vấn - khóa - giá trị theo thứ tự đó. **A** là ma trận trọng số chú ý sau chuẩn hóa softmax, mô tả mức độ liên hệ giữa mỗi token đặc trưng hình ảnh và đặc trưng ngữ nghĩa tương ứng. **F** là đặc trưng tổng hợp sau chú ý, thể hiện thông tin ngôn ngữ được đưa vào không gian thị giác để làm giàu ngữ nghĩa. Cuối cùng, đặc trưng hợp nhất được ổn định bằng kết nối dư và chuẩn hoá như ở (7).

$$\mathbb{Z} = \text{LayerNorm}(\mathbf{V} + \mathbf{F}) \tag{7}$$

Về cơ bản $\mathbb{Z}$ không chỉ mang thông tin ảnh, mà còn được điều chỉnh theo ngữ cảnh ngữ nghĩa của tiếng Việt, từ đó hỗ trợ mô hình sinh câu có nội dung sát ảnh hơn và giảm các lỗi mô tả mơ hồ.

2.2. Huấn luyện và suy luận sinh chú thích tiếng Việt

Trên biểu diễn đa phương thức $\mathbb{Z}$, mô hình sử dụng Transformer decoder để sinh chú thích theo cơ chế tự hồi quy. Ở thời điểm t, mô hình dự đoán phân bố xác suất của từ tiếp theo dựa trên chuỗi từ đã sinh và ngữ cảnh hợp nhất như ở (8), trong đó $g_{dec}(.)$ biểu thị hàm biến đổi của Transformer decoder.

$$p(y_t|y_{<t}, \text{I}) = \text{softmax}(g_{dec}(y_{<t}, \mathbb{Z})) \tag{8}$$

Trong quá trình huấn luyện hàm mất mát được sử dụng là hàm cross-entropy nhằm tối ưu khả năng dự đoán từ đúng tại mỗi bước như ở (9)

$$\mathcal{L}_{CE}(\theta) = -\sum_{t=1}^{T} \log p(y_t^*|y_{<t}^*, I;\ \theta) \tag{9}$$

với y_t^* là từ chuẩn của chú thích tiếng Việt. Ở giai đoạn suy luận, mô hình sinh chuỗi $\hat{Y}$ bằng cách lựa chọn chuỗi có xác xuất cao nhất theo nhãn tuần tự như ở (10).

$$\hat{Y} = \arg\max_{\mathbf{Y}} \ (p(Y|I;\theta)) \tag{10}$$

Trong giai đoạn huấn luyện, biểu diễn nhúng văn bản toàn cục T được trích xuất từ chú thích ảnh chuẩn bằng cơ chế teacher forcing nhằm cung cấp tín hiệu ngữ nghĩa ổn định cho mô-đun hợp nhất đặc trưng. Ở giai đoạn suy luận, khi không có chú thích chuẩn, T được sinh ra từ một gợi ý đầu vào (prompt), đảm bảo tính nhất quán của kiến trúc giữa huấn luyện và suy luận.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

2.3. Đánh giá hiệu năng trên bộ dữ liệu tiếng Việt

Đánh giá hiệu năng được thực hiện nhằm phản ánh toàn diện chất lượng chú thích sinh ra, không chỉ ở mức độ trùng khớp từ vựng mà còn về mức độ phù hợp ngữ nghĩa và tính tự nhiên. Vì vậy, nghiên cứu sử dụng các thước đo chuẩn trong chú thích ảnh gồm BLEU, METEOR và CIDEr. BLEU đo mức trùng khớp n-gram giữa chú thích sinh và chú thích tham chiếu, trong đó p_n là precision n-gram, w_n là trọng số và BP là hệ số phạt độ dài, ở (11). CIDEr tập trung đo tính đồng thuận giữa chú thích sinh và tập chú thích tham chiếu bằng TF-IDF, phản ánh tốt hơn mức tương đồng ngữ nghĩa trong các tình huống diễn đạt đa dạng, ở (12).

$$BLEU = BP.\exp\sum_{n=1}^{N} w_n \log p_n \tag{11}$$

$$CIDEr(c,s) = \frac{1}{N}\sum_{n=1}^{N}\frac{g_n(c).g_n(s)}{||g_n(c)||\,||g_n(s)||} \tag{12}$$

Ngoài ra, METEOR được sử dụng để đánh giá mức độ phù hợp ngữ nghĩa và tính tự nhiên của câu mô tả. METEOR kết hợp các dạng so khớp (chính xác, theo gốc từ, đồng nghĩa) và áp dụng hệ số phạt phân mảnh nhằm phản ánh độ trôi chảy của câu.

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Mô tả dữ liệu và phương thức đánh giá

Thực nghiệm được tiến hành trên một bộ dữ liệu chú thích ảnh tiếng Việt UIT-ViIC [6] và KTVIC [7], bao gồm các cặp *(I,Y)*, trong đó *I* là ảnh đầu vào và *Y* là câu mô tả tiếng Việt tương ứng. Dữ liệu được chia thành ba tập huấn luyện/kiểm định/kiểm tra theo tỷ lệ 70/10/20. Trước khi huấn luyện, các câu mô tả được chuẩn hóa văn bản (loại bỏ ký tự nhiễu, thống nhất dấu câu) và mã hóa thành chuỗi token để phục vụ quá trình giải mã tuần tự của mô hình.

Mô hình cơ sở được lựa chọn là kiến trúc mã hoá - giải mã tiêu chuẩn [4], trong đó ảnh được mã hóa bằng ResNet-101 [11] để thu nhận đặc trưng hình ảnh V, và quá trình sinh câu được thực hiện bởi một bộ giải mã [1] gồm 2 lớp. Mô hình này đại diện cho các hệ thống chú thích học sâu phổ biến hiện nay, đã được chứng minh hiệu quả trên các bộ dữ liệu chuẩn như MS COCO [3]. Trên nền tảng cơ sở này, chúng tôi tiến hành hai cải tiến:

(i) Tăng cường đặc trưng văn bản bằng véc tơ nhúng được sinh từ mô hình CLIP của OpenAI với kiến trúc ViT-B/32 [9].

(ii) Hợp nhất có điều hướng giữa đặc trưng thị giác và văn bản thông qua hợp nhất bằng cơ chế chú ý, như đã mô tả chi tiết tại Mục 2.1. Chúng tôi so sánh ba cấu hình sau:

Cơ sở: Chỉ sử dụng đặc trưng ảnh V, không có nhánh ngôn ngữ.

Cơ sở + Nhúng: Vector nhúng ngôn ngữ *T* từ CLIP, nối trực tiếp vào đặc trưng ảnh. Cách kết hợp này đóng vai trò là cơ sở mở rộng nhằm đánh giá hiệu quả của việc bổ sung thông tin ngữ nghĩa, đồng thời làm đối chứng trực tiếp với cơ chế hợp nhất đặc trưng có điều hướng bằng chú ý được đề xuất.

Đề xuất: Tăng cường đặc trưng ngôn ngữ và hợp nhất với đặc trưng hình ảnh.

Hiệu năng mô hình được đánh giá bằng các chỉ số BLEU-1, BLEU-4, METEOR và CIDEr [5], với trọng tâm là CIDEr thước đo phản ánh mức độ đồng thuận ngữ nghĩa giữa chú thích sinh và chú thích tham chiếu.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

3.2. Kết quả thực nghiệm

Kết quả trên UIT-ViIC

Bảng 1 trình bày kết quả trên bộ dữ liệu UIT-ViIC. Mô hình đề xuất cho thấy sự cải thiện rõ rệt trên tất cả các chỉ số so với mô hình cơ sở và biến thể nối đặc trưng đơn giản.

Bảng 1. So sánh kết quả trên UIT-ViIC

Phương pháp	BLEU-1 ↑	BLEU-4 ↑	METEOR ↑	CIDEr ↑
Cơ sở	0.642	0.248	0.215	0.782
Cơ sở + Nhúng	0.658	0.261	0.223	0.821
Đề xuất (Attention Fusion)	**0.679**	**0.278**	**0.236**	**0.874**

Sự cải thiện của mô hình đề xuất lên tới ~10% với CIDEr và ~3% với BLEU-4 so với mô hình cơ sở. Điều này khẳng định rằng nhúng ngữ nghĩa từ CLIP đóng vai trò như một hướng dẫn ngữ nghĩa, đặc biệt khi được hợp nhất thông qua chú ý có điều hướng, giúp mô hình sinh chú thích bám sát nội dung ảnh hơn và giàu thông tin hơn.

Kết quả trên KTVIC

Trên bộ dữ liệu KTVIC, vốn có phạm vi từ vựng rộng và ngữ cảnh đời sống đa dạng, xu hướng cải thiện vẫn được duy trì, như trình bày trong Bảng 2.

Bảng 2. So sánh kết quả trên KTVIC

Phương pháp	BLEU-1 ↑	BLEU-4 ↑	METEOR ↑	CIDEr ↑
Cơ sở	0.631	0.236	0.208	0.751
Cơ sở + Nhúng	0.647	0.249	0.217	0.792
Đề xuất (Attention Fusion)	**0.667**	**0.266**	**0.229**	**0.841**

Việc đạt mức CIDEr cao hơn ~9% so với mô hình cơ sở tiếp tục khẳng định hiệu quả của phương pháp. Trong thực tế, các mô tả sinh ra từ mô hình đề xuất thường rõ ràng hơn về hành động, vị trí và đối tượng chính trong ảnh. Điều này cho thấy cơ chế hợp nhất chú ý giữa đặc trưng hình ảnh và văn bản không chỉ bổ sung ngữ nghĩa, mà còn đóng vai trò lọc nhiễu ngôn ngữ, giúp mô hình tránh các câu mô tả chung chung, một vấn đề thường gặp khi chú thích ảnh tiếng Việt với dữ liệu huấn luyện hạn chế.

4. KẾT LUẬN

Bài báo đã trình bày một phương pháp chú thích ảnh tiếng Việt dựa trên mô hình hợp nhất đa phương thức, trong đó đặc trưng thị giác được kết hợp với đặc trưng ngữ nghĩa văn bản thông qua cơ chế hợp nhất chú ý. Khác với các mô hình cơ sở chỉ dựa trên đặc trưng ảnh, cách tiếp cận đề xuất tận dụng các véc tơ nhúng văn bản tiền huấn luyện để cung cấp tham chiếu ngữ nghĩa ổn định, đồng thời cho phép mô hình điều chỉnh trọng tâm mô tả theo từng vùng ảnh. Các kết quả thực nghiệm trên hai bộ dữ liệu UIT-ViIC và KTVIC cho thấy phương pháp này vượt trội hơn các đường cơ sở trên toàn bộ thước đo BLEU, METEOR và

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

CIDEr, trong đó mức cải thiện đáng kể ở CIDEr phản ánh khả năng sinh mô tả giàu ngữ nghĩa và gần với mô tả tham chiếu hơn. Điều này cho thấy hợp nhất chú ý là một hướng tiếp cận hiệu quả cho bài toán chú thích ảnh tiếng Việt, đặc biệt trong bối cảnh nguồn dữ liệu huấn luyện còn hạn chế. Mặc dù nghiên cứu này chưa thực hiện so sánh trực tiếp với các bộ mã hóa văn bản chuyên biệt cho tiếng Việt như PhoBERT hoặc ViT5, các mô hình này có tiềm năng cải thiện chất lượng biểu diễn ngữ nghĩa và có thể được tích hợp vào hệ thống đề xuất trong các nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017), *Attention is all you need*, Advances in neural information processing systems, 30.

[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018), *Bottom-up and top-down attention for image captioning and visual question answering*, In Proceedings of the IEEE CVPR (pp. 6077-6086).

[3] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015), *Microsoft coco captions: Data collection and evaluation server*. arXiv preprint arXiv:1504.00325.

[4] Li, L., Li, H., & Ren, P. (2025), *Underwater image captioning via attention mechanism based fusion of visual and textual information*, Information Fusion, 103269.

[5] Cheng, K., Liu, J., Mao, R., Wu, Z., & Cambria, E. (2025), *Echo: Generating cross-modal features for unseen classes in zero-shot remote sensing image captioning*, Information Fusion, 103952.

[6] Hoang Lam, Q., Duy Le, Q., Van Nguyen, K., & Luu-Thuy Nguyen, N. (2020), *UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning*, arXiv-2002.

[7] Pham, A. C., Nguyen, V. Q., Vuong, T. H., & Ha, Q. T. (2024), *Ktvic: A vietnamese image captioning dataset on the life domain*, arXiv preprint arXiv:2401.08100.

[8] Doanh, B. C., Truc, T. T. T., Thuan, N. T., Vu, N. D., & Vo, N. D. (2022), *Viecap4h challenge 2021: A transformer-based method for healthcare image captioning in vietnamese*, VNU Journal of Science: Computer Science and Communication Engineering, 38(2).

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021), *Learning transferable visual models from natural language supervision*, In International conference on machine learning (pp. 8748-8763). PmLR.

[10] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015), *Cider: Consensus-based image description evaluation*, In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016), *Deep residual learning for image recognition*, In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

ENHANCING VIETNAMESE IMAGE CAPTIONING PERFORMANCE THROUGH ATTENTION - GUIDED FEATURE FUSION

Hoang Anh Cong, Nguyen Dinh Cong, Pham The Anh

ABSTRACT

This paper proposes a Vietnamese image captioning method based on attention-guided feature fusion, in which visual representations are integrated with semantic text embeddings extracted from a pre-trained model through an Attention Fusion. This approach enhances semantic alignment and enables the model to generate more informative and contextually appropriate descriptions compared to traditional baseline methods. Experimental results on the UIT-ViIC and KTVIC datasets show that the proposed method improves the CIDEr score by approximately 6% - 10% over baseline models, while also achieving strong performance on BLEU and METEOR, demonstrating both the effectiveness and practical feasibility of the proposed approach.

Keywords: *Image captioning, attention fusion, text embedding.*

** Ngày nộp bài:29/01/2026; Ngày gửi phản biện:02/02/2026; Ngày duyệt đăng: 28/02/2026*