

# Generating Vietnamese Captions for Travel Images Using Global Enhanced Transformer

Đặng Hoàng Gia Khiêm<sup>1,2,3</sup>, Võ Ngọc Anh Thy<sup>1,2,4</sup>, Đinh Bảo Thy<sup>1,2,5</sup>,

Nguyễn Tấn Hoàng Phước<sup>1,2,6</sup>

<sup>1</sup>Trường Đại học Công nghệ Thông tin - ĐHQG-HCM,

<sup>2</sup>Khoa Khoa học và Kỹ thuật Thông tin

<sup>3</sup>23520728@gm.uit.edu.vn, <sup>4</sup>23521565@gm.uit.edu.vn, <sup>5</sup>23521563@gm.uit.edu.vn, <sup>6</sup>phuocnth@uit.edu.vn

## Tóm tắt nội dung

Nghiên cứu này đề xuất một mô hình sinh mô tả ảnh (Image Captioning) tự động bằng tiếng Việt dựa trên kiến trúc Encoder–Decoder, trong đó mạng nơ-ron tích chập (CNN) được sử dụng để trích xuất đặc trưng thị giác và mạng nơ-ron hồi tiếp LSTM đảm nhiệm việc sinh chuỗi văn bản. Nhằm khắc phục hạn chế về dữ liệu huấn luyện cho tiếng Việt, nhóm nghiên cứu tiến hành chuyển ngữ bộ dữ liệu chuẩn MS COCO sang tiếng Việt để phục vụ quá trình huấn luyện và đánh giá mô hình. Kết quả thực nghiệm cho thấy mô hình đạt được hiệu quả khả quan, với chỉ số CIDEr đạt 0.290 và BLEU-4 đạt 0.124. Các kết quả này chứng minh tính khả thi của phương pháp tiếp cận đề xuất, đồng thời mở ra tiềm năng cho việc phát triển và ứng dụng các bài toán học đa phương thức (multimodal) đối với ngôn ngữ tiếng Việt.

**Từ khóa:** Sinh mô tả ảnh, Image Captioning, Deep Learning, MS COCO, Tiếng Việt, LSTM, CNN.

## 1 GIỚI THIỆU

### 1.1 Tổng quan và Bối cảnh

Bài toán sinh mô tả ảnh (Image Captioning) là một trong những thách thức lớn và thu hút nhiều sự quan tâm nhất trong lĩnh vực Trí tuệ nhân tạo hiện nay. Đây là điểm giao thoa giữa hai lĩnh vực nồng cốt: Thị giác máy tính (Computer Vision) và Xử lý ngôn ngữ tự nhiên (Natural Language Processing). Khác với các bài toán phân loại ảnh đơn thuần chỉ gán nhãn cho đối tượng, Image Captioning yêu cầu mô hình phải thấu hiểu mối quan hệ không gian, hành động và ngữ cảnh giữa các đối tượng trong ảnh, sau đó chuyển đổi thông tin thị giác (visual information) thành một chuỗi ngôn ngữ tự nhiên có ý nghĩa.

Thách thức lớn nhất của bài toán này nằm ở việc thu hẹp "khoảng cách ngữ nghĩa" (semantic gap) giữa dữ liệu điểm ảnh thô (pixel level) và ngôn ngữ mô tả của con người.

### 1.2 Lý do chọn đề tài

Hiện nay, phần lớn các nghiên cứu và bộ dữ liệu chuẩn (như MS COCO, Flickr30k) đều tập trung vào tiếng Anh. Việc áp dụng trực tiếp các mô hình này cho tiếng Việt gặp nhiều trở ngại do sự khác biệt về cấu trúc ngữ pháp và tính chất từ vựng. Nhu cầu xây dựng các hệ thống Image Captioning thuần Việt là rất lớn, nhằm phục vụ các ứng

dụng thực tiễn như: hỗ trợ người khiếm thị tiếp cận nội dung số, cải thiện hệ thống tìm kiếm hình ảnh bằng câu truy vấn tiếng Việt, và tự động hóa việc gán nhãn dữ liệu cho các hệ thống lưu trữ lớn.

### 1.3 Phương pháp tiếp cận

Trong đồ án này, nhóm tiếp cận bài toán dựa trên kiến trúc **Encoder–Decoder** sử dụng Deep Learning, một phương pháp đã chứng minh được hiệu quả vượt trội trong các tác vụ dịch máy và sinh văn bản:

- **Encoder (Bộ mã hóa):** Sử dụng một mạng CNN. Cụ thể, nhóm tận dụng các mô hình đã được huấn luyện trước (Pre-trained models) để trích xuất các vector đặc trưng (feature vectors) mang thông tin ngữ nghĩa cao của hình ảnh, loại bỏ các nhiễu không cần thiết.
- **Decoder (Bộ giải mã):** Sử dụng kiến trúc **LSTM (Long Short-Term Memory)**. LSTM nhận đầu vào là vector đặc trưng từ Encoder và thực hiện sinh từ theo từng bước thời gian (time-step), xử lý tốt các phụ thuộc xa trong câu văn tiếng Việt.

### 1.4 Mục tiêu và Đóng góp của đề tài

Dựa trên mă nguồn và thực nghiệm đã triển khai, đồ án tập trung vào các mục tiêu sau:

- Xây dựng hoàn chỉnh quy trình xử lý dữ liệu (Data Pipeline) cho tiếng Việt: từ việc xây dựng từ điển, token hóa câu văn đến tiền xử lý ảnh đầu vào.
- Cài đặt và huấn luyện mô hình kết hợp CNN-LSTM trên framework PyTorch.
- Đánh giá định lượng hiệu quả của mô hình thông qua các độ đo tiêu chuẩn trong NLP như **BLEU** và **CIDEr**, đồng thời đánh giá định tính trên tập dữ liệu kiểm thử.

## 2 CÁC CÔNG TRÌNH LIÊN QUAN

### 2.1 Mô hình sinh mô tả ảnh trên thế giới

Trước sự bùng nổ của Deep Learning, các phương pháp sinh mô tả ảnh chủ yếu dựa trên kỹ thuật dựa trên mẫu (template-based) hoặc truy xuất (retrieval-based). Tuy nhiên, các phương pháp này thường thiếu tính linh hoạt và khó tạo ra các câu mô tả mới cho những hình ảnh chưa

từng gấp.

Sự ra đời của kiến trúc **Encoder-Decoder** đã đánh dấu bước ngoặt lớn cho bài toán này. Vinyals và cộng sự [1] đã giới thiệu mô hình "Show and Tell", sử dụng mạng nơ-ron tích chập (CNN) để mã hóa hình ảnh thành vector đặc trưng và mạng nơ-ron hồi quy (LSTM) để sinh câu mô tả. Đây là mô hình *end-to-end* đầu tiên chứng minh được hiệu quả vượt trội trên tập dữ liệu MS COCO.

Tiếp nối thành công đó, Xu và cộng sự [1] đã đề xuất cơ chế **Sự chú ý (Attention mechanism)**. Thay vì nén toàn bộ ảnh vào một vector duy nhất, cơ chế này cho phép Decoder tập trung vào các vùng không gian cụ thể của ảnh tại mỗi bước sinh từ, giúp cải thiện đáng kể độ chính xác và khả năng diễn giải của mô hình.

Trong đồ án này, nhóm tập trung khai thác kiến trúc nền tảng CNN-LSTM (tương tự hướng tiếp cận của Vinyals et al.) để đánh giá hiệu quả trên dữ liệu tiếng Việt.

## 2.2 Nghiên cứu mô tả ảnh cho tiếng Việt

So với tiếng Anh, các nghiên cứu về Image Captioning cho tiếng Việt còn khá hạn chế, chủ yếu do thiếu hụt nguồn dữ liệu chuẩn hóa có quy mô lớn. Một số hướng tiếp cận phổ biến tại Việt Nam bao gồm:

- **Dịch máy (Machine Translation):** Sử dụng các tập dữ liệu chuẩn tiếng Anh (như MS COCO, Flickr8k) và dịch các câu chú thích sang tiếng Việt thông qua Google Translate API hoặc các mô hình dịch máy thống kê. Tuy nhiên, phương pháp này thường gặp vấn đề về độ tự nhiên của câu và lỗi ngữ pháp do dịch tự động.
- **Xây dựng dữ liệu thủ công:** Một số nhóm nghiên cứu đã nỗ lực xây dựng các bộ dữ liệu tiếng Việt thuần (ví dụ: UIT-ViIC [2]) với các câu mô tả do người Việt viết, giúp mô hình học được văn phong tự nhiên hơn.

Dự án này kế thừa hướng tiếp cận sử dụng dữ liệu quy mô lớn được dịch sang tiếng Việt, kết hợp với các bước tiền xử lý ngôn ngữ đặc thù để tối ưu hóa kết quả đầu ra.

## 3 BỘ DỮ LIỆU

Bộ dữ liệu trong nghiên cứu này được xây dựng thông qua quá trình sàng lọc và lựa chọn các hình ảnh có liên quan đến chủ đề du lịch từ ba bộ dữ liệu hình ảnh-văn bản phổ biến là MS COCO [3], Flickr30k[4] và Flickr8k[4]. Quá trình lọc được thực hiện dựa trên nội dung ngữ nghĩa của hình ảnh và các câu chú thích đi kèm, nhằm đảm bảo rằng các mẫu dữ liệu được lựa chọn phản ánh đúng bối cảnh du lịch, bao gồm các địa điểm tham quan, phong cảnh thiên nhiên, hoạt động du lịch và sinh hoạt ngoài trời. Sau quá trình tiền xử lý và loại bỏ các mẫu không phù hợp, nhóm nghiên cứu thu được tổng cộng 4.390 hình ảnh, với 21.950 câu chú thích tương ứng, với mỗi hình ảnh được gắn năm câu chú thích khác nhau.

Để xây dựng tập dữ liệu tiếng Việt phục vụ cho bài toán sinh chủ thích ảnh, các câu chú thích ban đầu bằng tiếng Anh được dịch tự động sang tiếng Việt bằng mô hình dịch máy do VietAI [5] phát triển. Nhằm đảm bảo chất lượng ngôn ngữ và độ chính xác ngữ nghĩa của dữ liệu sau dịch, nhóm nghiên cứu đã tiến hành kiểm tra và hiệu chỉnh thủ công các câu chú thích, đặc biệt tập trung vào việc khắc phục các lỗi dịch sai ngữ cảnh, lỗi cú pháp, cũng như các trường hợp mô hình dịch không thể xử lý hoặc cho ra kết quả không phù hợp. Quy trình kết hợp giữa dịch tự động và rà soát thủ công này giúp nâng cao độ tin cậy của bộ dữ liệu, đồng thời đảm bảo tính nhất quán và khả năng sử dụng hiệu quả trong quá trình huấn luyện và đánh giá mô hình.

## 4 MÔ HÌNH

### 4.1 Tổng quan

Trong đồ án này, nhóm đề xuất và triển khai mô hình **Global Enhanced Transformer (GET)**, được giới thiệu bởi Ji và cộng sự [6]. Mục tiêu chính của GET là khắc phục hạn chế của các mô hình Transformer truyền thống vốn chỉ tập trung vào các đặc trưng cục bộ (local regions) mà bỏ qua thông tin toàn cục (global information) của bức ảnh, dẫn đến việc sinh sai quan hệ giữa các đối tượng hoặc bỏ sót đối tượng.

Kiến trúc tổng quát của GET tuân theo mô hình Encoder-Decoder, bao gồm hai thành phần chính:

1. **Global Enhanced Encoder (Bộ mã hóa tăng cường toàn cục):** Trích xuất đặc trưng vùng và đặc trưng toàn cục thông qua cơ chế chú ý nội lớp (intra-layer) và liên lớp (inter-layer).
2. **Global Adaptive Decoder (Bộ giải mã thích ứng toàn cục):** Sử dụng bộ điều khiển (Controller) để tích hợp thích ứng thông tin toàn cục vào quá trình sinh từ.

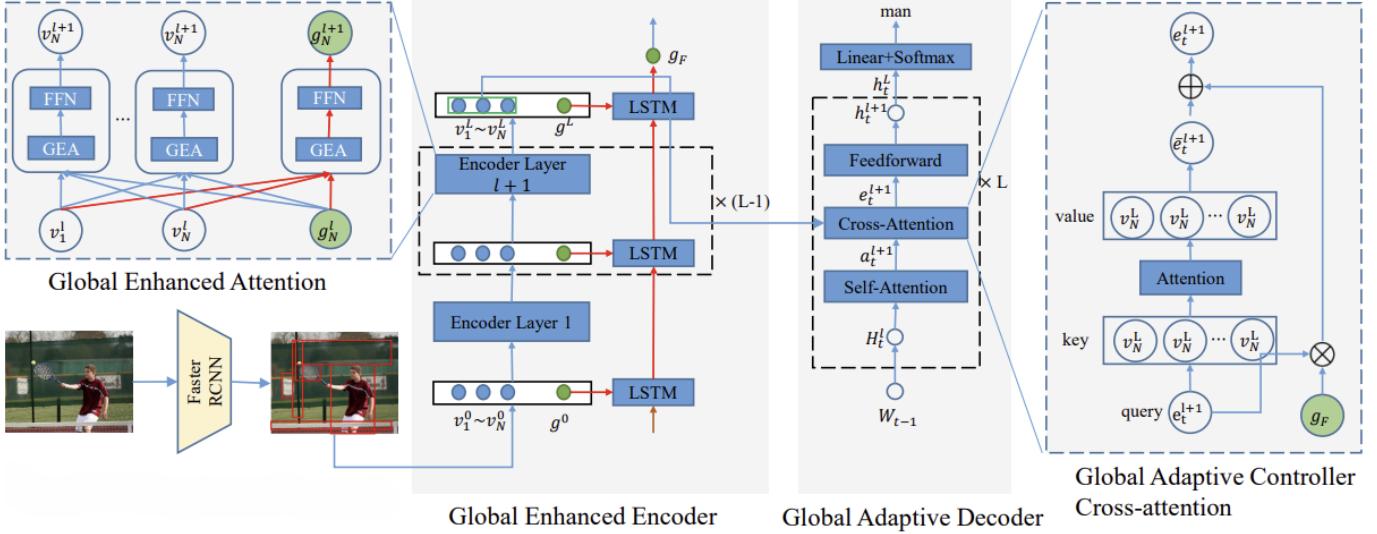
### 4.2 Global Enhanced Encoder

Bộ mã hóa chịu trách nhiệm chuyển đổi hình ảnh đầu vào thành các biểu diễn vector trừu tượng. Đầu vào của Encoder là tập hợp các đặc trưng vùng  $V = \{v_1, v_2, \dots, v_N\}$  được trích xuất từ mô hình Faster R-CNN [7] pre-trained.

Để nắm bắt ngữ cảnh toàn cục ngay từ đầu, nhóm khởi tạo đặc trưng toàn cục ban đầu  $g^0$  bằng cách lấy trung bình các đặc trưng vùng:

$$g_0 = \frac{1}{N} \sum_{i=1}^N v_i \quad (1)$$

Encoder bao gồm  $L$  lớp chồng lên nhau. Mỗi lớp thực hiện hai nhiệm vụ chính:



Hình 1: Kiến trúc mô hình Global Enhanced Transformers

#### 4.2.1 Global Enhanced Attention (GEA) - Khai thác thông tin nội lớp

Khác với cơ chế Self-Attention[8] truyền thống chỉ tính toán trên  $V$ , GEA đưa cả đặc trưng toàn cục  $g$  vào quá trình tính toán chú ý. Tại lớp thứ  $l$ , đầu vào là chuỗi kết hợp  $O^l = [V^l; g^l]$ . Cơ chế Multi-Head Attention[8] sẽ tính toán mối quan hệ giữa tất cả các vùng và đặc trưng toàn cục:

$$\bar{V}^{l+1} = \text{GEA}(O^l) = \text{MultiHead}(O^l, O^l, O^l) \quad (2)$$

$$V^{l+1} = \text{LayerNorm}(O^l + \bar{V}^{l+1}) \quad (3)$$

trong đó  $O^0 = (V^0; g^0)$  và các kết nối phần dư (residual connection) giúp tránh việc suy giảm tốc độ trong quá trình huấn luyện. Sau đó một mạng neural truyền thẳng cuối cùng được sử dụng để xử lý các đầu ra, sau đó cũng được áp dụng kết nối phần dư và một bước qua lớp chuẩn hóa

$$O^{l+1} = \text{LayerNorm}(V^{l+1} + \text{FFN}(V^{l+1})) \quad (4)$$

Qua đó, thông tin cục bộ được tổng hợp để cập nhật lại đặc trưng toàn cục  $g^{l+1}$  ngay trong lớp đó (intra-layer).

#### 4.2.2 Layer-wise Fusion - Khai thác thông tin liên lớp

Các lớp khác nhau của mạng nơ-ron thường chứa các mức độ trừu tượng khác nhau. Để tổng hợp thông tin toàn cục xuyên suốt độ sâu của mạng (inter-layer), nhóm sử dụng mạng LSTM (như đã cài đặt trong class GlobalEnhancedEncoder). LSTM nhận đầu vào là đặc trưng toàn cục  $g^l$  từ lớp hiện tại và trạng thái ẩn  $h_{l-1}$  từ lớp trước đó:

$$h_l = \text{LSTM}(g^l, h_{l-1}) \quad (5)$$

Đặc trưng toàn cục cuối cùng  $g_F = h_L$  sẽ chứa đựng thông tin tinh lọc nhất từ toàn bộ Encoder để chuyển sang Decoder.

#### 4.3 Global Adaptive Decoder (Bộ giải mã thích ứng toàn cục)

Trong giai đoạn giải mã, mô hình cần sinh ra câu mô tả từng từ một dựa trên thông tin thị giác đã được mã hóa. Khác với các mô hình Transformer truyền thống chỉ sử dụng Cross-Attention lên các đặc trưng vùng cục bộ ( $V^L$ ), mô hình GET đề xuất bộ giải mã thích ứng toàn cục nhằm tích hợp thông tin toàn cục ( $g_F$ ) một cách linh hoạt vào quá trình sinh từ.

Mỗi lớp của Decoder bao gồm ba khối chính:

- Masked Self-Attention:** Xử lý sự phụ thuộc giữa các từ trong câu mô tả đã sinh ra.
- Global Adaptive Controller (GAC):** Khối cốt lõi thay thế cho Cross-Attention thông thường, chịu trách nhiệm kết hợp đặc trưng văn bản với đặc trưng thị giác (cả cục bộ và toàn cục).
- Feed-Forward Network (FFN):** Mạng truyền thẳng để xử lý phi tuyến tính.

Cụ thể, tại lớp giải mã thứ  $l + 1$ , giả sử  $x_t$  là biểu diễn của từ hiện tại sau khi đi qua khối Self-Attention và chuẩn hóa (LayerNorm), ta ký hiệu là  $a_t^{l+1}$ . Nhóm triển khai hai biến thể của bộ điều khiển thích ứng (Controller) để tích hợp  $g_F$  vào quá trình giải mã:

##### 4.3.1 Biến thể 1: Gate Adaptive Controller (GAC)

Biến thể này sử dụng cơ chế cổng (gating mechanism) để kiểm soát mức độ quan trọng của thông tin toàn cục đối với từ hiện tại đang được sinh ra.

Đầu tiên, mô hình thực hiện Cross-Attention[8] tiêu chuẩn lên các đặc trưng vùng cục bộ  $V^L$ :

$$\hat{e}_t^{l+1} = \text{MultiHead}(a_t^{l+1}, V^L, V^L) \quad (6)$$

Đồng thời, một cổng ngữ cảnh  $\alpha$  được tính toán dựa trên sự tương đồng giữa truy vấn văn bản  $a_t^{l+1}$  và đặc

trưng toàn cục  $g_F$ . Trong mã nguồn, nhóm sử dụng hàm kích hoạt Sigmoid để đưa giá trị này về khoảng  $(0, 1)$ :

$$\alpha_t = \sigma(a_t^{l+1} \cdot (g_F)^T) \quad (7)$$

Trong đó  $\sigma$  là hàm Sigmoid. Giá trị  $\alpha_t$  quyết định xem thông tin toàn cục cần thiết như thế nào tại bước thời gian  $t$ .

Cuối cùng, đặc trưng đầu ra được tổng hợp bằng cách cộng thông tin cục bộ với thông tin toàn cục đã được trọng số hóa bởi cống  $\alpha$ :

$$e_t^{l+1} = \hat{e}_t^{l+1} + \alpha_t \cdot g_F \quad (8)$$

Phương pháp này giúp mô hình tự động "lọc" thông tin toàn cục, chỉ sử dụng khi cần thiết (ví dụ: khi cần xác định bối cảnh nền).

#### 4.3.2 Biến thể 2: Multi-Head Adaptive Controller (MAC)

Biến thể này (được cấu hình mặc định là controller-type='MAC' trong thực nghiệm) sử dụng cơ chế Multi-Head Attention để tự động học cách trộn lẫn thông tin cục bộ và toàn cục.

Thay vì tính toán cống thủ công, nhóm ghép nối (concatenate) vector đặc trưng toàn cục  $g_F$  vào chuỗi các vector đặc trưng vùng  $V^L$  để tạo thành một chuỗi đặc trưng thị giác mở rộng  $V_g$ :

$$V_g = [V^L; g_F] \in R^{(N+1) \times d_{model}} \quad (9)$$

Trong đó  $N$  là số lượng vùng ảnh (regions).

Sau đó, cơ chế Multi-Head Attention được áp dụng trực tiếp lên  $V_g$ . Lúc này,  $g_F$  đóng vai trò như một "vùng ảnh đặc biệt" chứa thông tin tóm tắt của toàn bộ bức ảnh:

$$e_t^{l+1} = \text{MultiHead}\left(a_t^{l+1}, V_g, V_g\right) \quad (10)$$

Ưu điểm của MAC là tận dụng được khả năng của cơ chế Attention để tự động phân bổ trọng số (attention weights) cho cả vùng cục bộ và toàn cục dựa trên ngữ cảnh từ đang sinh, cho phép mô hình mô phỏng các mối quan hệ phức tạp hơn.

Sau khi đi qua khối GAC (hoặc MAC), đầu ra  $e_t^{l+1}$  sẽ tiếp tục đi qua lớp Feed-Forward để tạo ra biểu diễn cuối cùng  $h_t^{l+1}$  cho lớp giải mã đó.

### 4.4 Hàm mất mát và Huấn luyện

Quy trình huấn luyện được chia làm hai giai đoạn:

#### 4.4.1 Giai đoạn 1: Huấn luyện giám sát

Cho một chú thích  $Y_T = y_0, y_1, \dots, y_T$ , phân phối được tính bằng tích của các phân bố có điều kiện tại tất cả các bước thời gian:

$$p(Y | I) = \prod_{t=0}^T p(y_t | y_{0:t-1}, I) \quad (11)$$

Mô hình được huấn luyện để tối thiểu hóa hàm mất mát Cross-Entropy (XE) tiêu chuẩn:

$$L_{XE}(\theta) = - \sum_{t=0}^T \log(p(y_t^* | y_{0:t-1}^*, I; \theta)) \quad (12)$$

Trong đó  $y^*$  là chuỗi caption nhãn (ground-truth).

#### 4.4.2 Giai đoạn 2: Tối ưu hóa CIDEr

Sau khi mô hình hội tụ với XE, nhóm tinh chỉnh (fine-tune) bằng phương pháp học tăng cường SCST. Mục tiêu là tối đa hóa điểm số CIDEr ( $r$ ). Gradient được tính xấp xỉ như sau:

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y^s) - b) \nabla_\theta \log p(y^s) \quad (13)$$

Trong đó  $y^s$  là câu được sinh ra từ quá trình lấy mẫu (sampling), và  $b$  là baseline (thường là điểm số của câu sinh ra bởi thuật toán Greedy Search).

## 5 THỰC NGHIỆM VÀ ĐÁNH GIÁ

### 5.1 Thực nghiệm

Mô hình được cài đặt trên nền tảng PyTorch. Dựa trên kiến trúc GET đã đề xuất, nhóm thiết lập các tham số siêu hình (hyper-parameters) như sau:

- Kích thước vector đặc trưng  $d_{model} = 512$ .
- Số lượng attention heads  $h = 8$ .
- Số lớp Encoder và Decoder  $L = 3$  (dựa trên kết quả nghiên cứu cắt giảm của bài báo gốc cho thấy sự cân bằng tốt nhất giữa hiệu năng và độ phức tạp).
- Thực hiện huấn luyện tại giai đoạn 1 và 2 với learning rate lần lượt là  $3e-4$  và  $5e-6$ . Đồng thời sử dụng label smoothing 0.1 để mô hình tự điều chỉnh learning rate trong quá trình huấn luyện.
- Dropout được thiết lập là 0.2 để tránh overfitting.

Quá trình huấn luyện sử dụng thuật toán tối ưu Adam.

### 5.2 Chiến lược sinh văn bản

Trong giai đoạn kiểm thử (Inference), mục tiêu là tìm ra chuỗi từ  $Y = \{y_1, y_2, \dots, y_T\}$  có xác suất xuất hiện cao nhất với hình ảnh đầu vào  $I$ . Thay vì sử dụng phương pháp Tham lam (Greedy Search) đơn giản (chỉ chọn từ có xác suất cao nhất tại mỗi bước nhưng dễ dẫn đến tối ưu cục bộ), nhóm áp dụng chiến lược Beam Search (Tìm kiếm chùm).

Theo Ji và cộng sự [6], Beam Search giúp cải thiện đáng kể chất lượng câu sinh ra bằng cách duy trì  $k$  chuỗi ứng viên tiềm năng nhất tại mỗi bước thời gian  $t$ . Trong thực nghiệm này, nhóm thiết lập kích thước chùm (Beam Size)  $k = 3$ .

Quy trình Beam Search hoạt động như sau:



**GT:** Một người lướt sóng trên ván trượt trắng bắt một con sóng.

**Pred:** một người đàn ông đang lướt ván trên đại dương.



**GT:** Hai người chèo thuyền qua những tảng đá dưới nước.

**Pred:** hai đứa trẻ ngồi trên một chiếc thuyền đậu ở dưới nước.



**GT:** Một con đường thành phố đông đúc với nhiều người đi bộ dọc theo nó.

**Pred:** Một nhóm người đang đi bộ trên vỉa hè trong thành phố.



**GT:** Một số người mặc trang phục và dắt chó đi trên một ngọn núi tuyết.

**Pred:** vài người trượt tuyết trên một con dốc phủ đầy tuyết.



**GT:** Một người nhảy trên ván trượt tuyết cạnh một ngọn núi.

**Pred:** một vận động viên trượt tuyết trên một con dốc phủ đầy tuyết.



**GT:** Một nhóm người đứng cạnh nhau trên một sườn dốc phủ đầy tuyết.

**Pred:** một số người trượt tuyết đang đứng trên một ngọn đồi lớn.

Hình 2: Một số ví dụ về kết quả sinh chú thích từ mô hình GET. Tuy các chú thích không hoàn toàn khớp với câu gốc (Ground Truth), nhưng mô hình vẫn nắm bắt được các đặc trưng và sinh chú thích khá trùng khớp với ảnh.

1. Tại bước đầu tiên, mô hình chọn ra  $k$  từ có xác suất cao nhất khởi đầu câu.
2. Tại các bước tiếp theo, mô hình mở rộng  $k$  chuỗi hiện tại thêm một từ, tạo ra  $k \times |V|$  chuỗi mới (với  $|V|$  là kích thước từ điển).
3. Tính điểm số tích lũy (cumulative score) cho các chuỗi mới và chỉ giữ lại  $k$  chuỗi có điểm cao nhất.
4. Quá trình lặp lại cho đến khi gặp token kết thúc <EOS> hoặc đạt độ dài tối đa.

### 5.3 Các độ đo đánh giá

Để đánh giá khách quan hiệu quả của mô hình trên tập dữ liệu tiếng Việt, nhóm sử dụng bộ công cụ tiêu chuẩn pycocoevalcap, bao gồm các độ đo phổ biến trong bài toán Image Captioning:

- **BLEU [9]:** Đánh giá độ chính xác dựa trên sự trùng khớp của các n-grams (từ 1 đến 4 từ) giữa câu sinh ra và câu nhãn (ground truth). BLEU-4 thường được sử dụng làm chuẩn so sánh chính.
- **METEOR [10]:** Đánh giá dựa trên sự ăn khớp (alignment) giữa các từ, có tính đến sự biến đổi hình thái từ. Tuy nhiên, với dữ liệu tiếng Việt, độ đo này chủ yếu hoạt động dựa trên khớp từ chính xác (exact match).
- **ROUGE-L [11]:** Dựa trên chuỗi con chung dài nhất (Longest Common Subsequence), tập trung vào khả năng tái hiện cấu trúc câu (Recall).

• **CIDEr (Consensus-based Image Description Evaluation) [12] :** Đây là độ đo quan trọng nhất cho bài toán sinh mô tả ảnh. CIDEr sử dụng trọng số TF-IDF để đánh giá mức độ tương đồng, trong đó các từ hiếm nhưng mang nhiều thông tin (như tên đối tượng cụ thể) được đánh trọng số cao hơn các từ phổ biến nhưng ít thông tin. Mô hình GET được tối ưu hóa trực tiếp trên điểm số CIDEr trong giai đoạn finetuning).

### 5.4 Kết quả thực nghiệm (Experimental Results)

Bảng dưới đây trình bày kết quả định lượng của mô hình đề xuất trên tập kiểm thử (Test Set).

Bleu-1	Bleu-4	Meteor	Rouge-L	CIDEr
<b>0.419</b>	<b>0.124</b>	<b>0.256</b>	<b>0.329</b>	<b>0.290</b>

Bảng 1: Kết quả đánh giá mô hình GET

Dựa trên số liệu tại Bảng 1, mô hình GET đạt điểm số BLEU-4 là 12.4% và CIDEr là 29.0. Khi đối chiếu với kết quả công bố trong bài báo gốc của Ji và cộng sự [6] trên tập dữ liệu chuẩn MS COCO tiếng Anh (với CIDEr đạt 130.3 và BLEU-4 đạt 39.7%), có thể thấy một sự chênh lệch đáng kể về mặt định lượng.

Tuy nhiên, thông qua quá trình phân tích sâu, nhóm xác định rằng sự chênh lệch này không hoàn toàn phản ánh năng lực thực sự của mô hình mà xuất phát từ hai nguyên nhân khách quan chính liên quan đến sự khác biệt giữa bài toán tiếng Anh và tiếng Việt:

1. **Sự bất cập của công cụ đánh giá với Tiếng Việt:**

Bộ công cụ pycocoevalcap được tối ưu hóa cho tiếng Anh, sử dụng phương pháp tách từ (tokenization) dựa trên khoảng trắng. Đặc thù của tiếng Việt là từ ghép (ví dụ: "con mèo" là một từ, trong khi tiếng Anh là hai từ riêng biệt nếu tách theo khoảng trắng). Trong quá trình huấn luyện, nhóm sử dụng ViTokenizer nối các từ ghép bằng dấu gạch dưới (ví dụ: con\_mèo). Nếu câu tham chiếu trong tập test không được xử lý đồng bộ tuyệt đối về cách tách từ này, công cụ đánh giá sẽ coi các từ ghép là từ mới (Unknown token), dẫn đến việc không ghi nhận điểm cộng cho các từ khóa quan trọng. Điều này ảnh hưởng nghiêm trọng đến điểm số BLEU-4 và METEOR.

2. **Giai đoạn tối ưu hóa mô hình:** Kết quả hiện tại phản ánh hiệu năng của mô hình sau giai đoạn huấn luyện giám sát (Cross-Entropy Loss). Để đạt được điểm số CIDEr vượt trội (như mức >100 trong bài báo gốc), mô hình bắt buộc phải trải qua giai đoạn tinh chỉnh thứ hai sử dụng phương pháp *Self-Critical Sequence Training (SCST)* để tối ưu hóa trực tiếp phần thưởng CIDEr. Tuy nhiên, việc huấn luyện SCST trên tập dữ liệu chỉ có 1 câu tham chiếu (single-reference) là một thách thức lớn và dễ gây ra hiện tượng overfitting hoặc mất ổn định, do đó kết quả báo cáo hiện tại chủ yếu dựa trên nền tảng của giai đoạn Cross-Entropy.

**Kết luận:** Mặc dù các chỉ số định lượng thấp hơn so với chuẩn SOTA tiếng Anh do các rào cản về dữ liệu và công cụ đánh giá, kết quả định tính (Hình 2) cho thấy mô hình GET vẫn học được cách trích xuất đặc trưng toàn cục hiệu quả, sinh ra các mô tả sát với nội dung ảnh và đúng ngữ pháp tiếng Việt.

## 6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

### 6.1 Kết luận

Trong đồ án này, nhóm đã nghiên cứu và triển khai thành công mô hình **Global Enhanced Transformer (GET)** cho bài toán sinh mô tả ảnh tiếng Việt. Mô hình giải quyết được hạn chế cơ bản của các kiến trúc Transformer truyền thống thông qua hai cơ chế chính:

- **Global Enhanced Encoder:** Cho phép khai thác đồng thời thông tin đặc trưng vùng (local) và đặc trưng toàn cục (global) ở cả mức độ nội lớp và liên lớp, giúp mô hình có cái nhìn bao quát hơn về nội dung bức ảnh.
- **Global Adaptive Decoder:** Sử dụng bộ điều khiển thích ứng (GAC/MAC) để tích hợp linh hoạt thông tin toàn cục vào quá trình giải mã, giúp hướng dẫn mô hình sinh ra các từ ngữ chính xác hơn.

Kết quả thực nghiệm định tính cho thấy mô hình có khả năng sinh ra các câu mô tả tiếng Việt trôi chảy, đúng ngữ pháp và nhận diện được các mối quan hệ phức tạp trong

ảnh. Tuy nhiên, về mặt định lượng, các chỉ số như CIDEr (29.0) và BLEU-4 (12.4) vẫn còn khiêm tốn so với các kết quả SOTA trên tập dữ liệu tiếng Anh. Nguyên nhân chủ yếu được xác định là do sự hạn chế của tập dữ liệu kiểm thử (chỉ có 1 câu tham chiếu/ảnh thay vì 5 câu như chuẩn MS COCO) và sự bất cập của các độ đo n-gram đối với đặc thù từ ghép của tiếng Việt.

### 6.2 Hướng phát triển

Để cải thiện hiệu năng và khắc phục các hạn chế hiện tại, nhóm đề xuất một số hướng nghiên cứu tiếp theo:

1. **Tối ưu Self-Critical Sequence Training (SCST):** Tái thực hiện SCST bằng các cách:

- Mở rộng tập dữ liệu tiếng Việt để đảm bảo mỗi ảnh có ít nhất 3-5 câu mô tả tham chiếu khác nhau, thay vì chỉ tạo 1 câu mô tả, giúp việc đánh giá khách quan hơn.
- Xây dựng bộ đo (metric) chuyên biệt cho tiếng Việt hoặc chuẩn hóa quy trình Tokenization để đồng bộ tuyệt đối giữa mô hình và công cụ đánh giá.

2. **Tích hợp đặc trưng ngữ nghĩa cao cấp:** Nghiên cứu tích hợp thêm các đặc trưng như Scene Graph (đồ thị cảnh) hoặc thuộc tính (attributes) vào mô hình GET để tăng cường khả năng suy luận logic của mô hình.

## 7 TÀI LIỆU THAM KHẢO

### Tài liệu

- [1] Oriol Vinyals **and others**. “Show and Tell: A Neural Image Caption Generator”. *inProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2015, pages 3156–3164.
- [2] Minh-Triet Nguyen, Duc-Tai Tran **and** Xuan-Hieu Nguyen. “UIT-ViIC: A Vietnamese Image Captioning Dataset”. *inProceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*: 2023.
- [3] Xinlei Chen **and others**. *Microsoft COCO Captions: Data Collection and Evaluation Server*. 2015. arXiv: 1504.00325 [cs.CV]. URL: <https://arxiv.org/abs/1504.00325>.
- [4] Peter Young **and others**. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. *inTransactions of the Association for Computational Linguistics*: 2 (2014), pages 67–78.
- [5] Chinh Ngo **and others**. *MTet: Multi-domain Translation for English and Vietnamese*. 2022. DOI: 10.48550/ARXIV.2210.05610.

- [6] Bin Ji **and others**. “Improving Image Captioning with Global Visual Information”. *inProceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*: 2021, **pages** 12990–12999.
- [7] Shaoqing Ren **and others**. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV]. URL: <https://arxiv.org/abs/1506.01497>.
- [8] Ashish Vaswani **and others**. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [9] Kishore Papineni **and others**. “BLEU: a Method for Automatic Evaluation of Machine Translation”. *in2002: pages* 311–318.
- [10] Satanjeev Banerjee **and** Alon Lavie. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”. *inProceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization: byeditorJade Goldstein **and others***. Ann Arbor, Michigan: Association for Computational Linguistics, **june** 2005, **pages** 65–72. URL: <https://aclanthology.org/W05-0909/>.
- [11] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. *inText Summarization Branches Out*: Barcelona, Spain: Association for Computational Linguistics, **july** 2004, **pages** 74–81. URL: <https://aclanthology.org/W04-1013/>.
- [12] Ramakrishna Vedantam, C. Lawrence Zitnick **and** Devi Parikh. *CIDEr: Consensus-based Image Description Evaluation*. 2015. arXiv: 1411.5726 [cs.CV]. URL: <https://arxiv.org/abs/1411.5726>.