RESEARCH ARTICLE

CHÚ THÍCH ẢNH CHO NGƯỜI KHIẾM THỊ SỬ DỤNG TRANSFORMER

Nguyễn Thiên Khiêm, Đỗ Đức Đạt, Nguyễn Ngọc Hoài Trí, Võ Văn Thịnh, Văn Thành Thích, Nguyễn Văn Thịnh*

Khoa Công nghệ thông tin, Trường Đại học Sư phạm TP.HCM

{4801104070, 4801104024, 4801104140, 4801104129, 4801104127}@student.hcmue.edu.vn, thinhnv@hcmue.edu.vn

TÓM TẮT— Suy giảm thị lực ảnh hưởng đến hàng triệu người trên thế giới, gây khó khăn trong việc tiếp cận thông tin trực quan. Với sự phát triển của thiết bị di động, đặc biệt là nền tảng Android, các giải pháp chuyển đổi hình ảnh thành mô tả âm thanh đang trở nên phổ biến. Tuy nhiên, việc tạo ra các mô tả chính xác, giàu ngữ cảnh và phù hợp với thiết bị di động vẫn là một thách thức. Nghiên cứu này đề xuất một mô hình chú thích ảnh dựa trên kiến trúc encoder–decoder, kết hợp Swin Transformer để trích xuất đặc trưng hình ảnh và Transformer Decoder để sinh chú thích. Mô hình được huấn luyện trên các tập dữ liệu chuẩn MS COCO, Flickr30k và tinh chỉnh với dữ liệu tiếng Việt KTVIC. Thực nghiệm cho thấy mô hình đạt hiệu suất cao trên các độ đo BLEU, METEOR và CIDEr. Bên cạnh mô hình, chúng tôi cũng xây dựng một ứng dụng Android tích hợp chức năng chú thích ảnh và chuyển văn bản thành giọng nói, hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh theo thời gian thực, hoạt động ổn định với tốc độ xử lý phù hợp trong điều kiện sử dụng thực tế. Kết quả này cho thấy tiềm năng trong việc cải thiện khả năng tiếp cận thông tin trực quan cho cộng đồng người khiếm thị.

Từ khóa— Chú thích ảnh tự động, mô hình Encoder-Decoder, Swin Transformer, Transformer Decode, ứng dụng hỗ trợ người khiếm thị.

I. GIỚI THIỆU

Theo thống kê của Tổ chức Y tế Thế giới (WHO) vào tháng 8 năm 2023, có khoảng 2.2 tỷ người trên toàn cầu bị suy giảm thị lực hoặc mù lòa, gây ảnh hưởng nghiêm trọng đến khả năng tiếp cận thông tin, giao tiếp và hòa nhập xã hội của họ [1]. Trong một thế giới ngày càng phụ thuộc vào thông tin trực quan, việc thiếu khả năng nhận thức hình ảnh trở thành rào cản lớn đối với người khiếm thị trong học tập, làm việc và sinh hoạt hàng ngày. Điều này thúc đẩy nhu cầu phát triển các công nghệ, hệ thống hỗ trợ giúp chuyển đổi thông tin thị giác sang các định dạng dễ tiếp cận hơn như văn bản hoặc âm thanh. Một trong những hướng tiếp cận để giải quyết vấn đề này là bài toán chú thích ảnh (Image Captioning)-bài toán kết hợp giữa thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural Language Processing), với mục tiêu tự động sinh ra các mô tả bằng văn bản ngắn gọn, chính xác và mạch lạc từ nội dung hình ảnh đầu vào [2]. Với khả năng mô tả nội dung hình ảnh bằng ngôn ngữ tự nhiên, chú thích ảnh có nhiều ứng dụng thực tế như hỗ trợ cải tiến hệ thống xe tự lái [3], phân tích video giám sát [4], hỗ trợ chẩn đoán và kê đơn thuốc tự động từ ảnh y khoa [5], phân tích dữ liệu giao thông [6], chẩn đoán bệnh cây trồng [7], kiểm duyệt nội dung [8], an toàn lao động [9], đặc biệt là hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh thông qua hệ thống chuyển đổi hình ảnh thành văn bản và giọng nói [10, 11].

Gần đây, nhiều mô hình chú thích ảnh được xây dựng theo khung encoder–decoder. Trong đó, encoder thường sử dụng mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng hình ảnh. Phần decoder sử dụng mạng RNN hoặc LSTM để tạo ra câu mô tả [12] [13]. CNN chủ yếu tập trung vào đặc trưng cục bộ thông qua các lớp tích chập. Do đó, mô hình khó học được quan hệ không gian hoặc ngữ cảnh tổng thể. Điều này làm giảm khả năng nắm bắt quan hệ phức tạp giữa các đối tượng trong ảnh. Hạn chế này có thể ảnh hưởng đến độ chính xác của chú thích. Bên cạnh đó, LSTM cần xử lý tuần tự nên tốc độ huấn luyện khá chậm. Mô hình cũng dễ gặp vấn đề triệt tiêu hoặc bùng nổ đạo hàm. Ngoài ra, LSTM thiếu cơ chế tập trung vào vùng ảnh quan trọng. Hệ quả là mô hình dễ tạo ra chú thích thiếu chính xác hoặc thiếu chi tiết.

Transformer [14] đã tạo ra bước đột phá lớn trong xử lý ngôn ngữ tự nhiên. Từ đó, nhiều nghiên cứu chú thích ảnh đã thay thế LSTM bằng Transformer Decoder trong vai trò bộ giải mã. Nguyên nhân là vì Transformer hỗ trợ huấn luyện song song và đạt hiệu suất tốt. Khả năng này đặc biệt hữu ích khi làm việc với tập ảnh có quy mô lớn. Bên cạnh đó, các mô hình phân loại ảnh dựa trên Transformer như Vision Transformer (ViT) [15] và Swin Transformer [16] được đề xuất. Những kiến trúc này giúp tăng cường khả năng trích xuất đặc trưng hình ảnh. Chúng cho phép mô hình học quan hệ toàn cục giữa các vùng ảnh khác nhau. Nhờ vậy, nhiều công trình đã áp dụng Transformer vào bài toán sinh mô tả ảnh. Cách tiếp cận này nhằm khắc phục các hạn chế của mô hình CNN–LSTM truyền thống.

Từ những hạn chế đã nêu, một mô hình chú thích ảnh theo khung encoder–decoder được đề xuất. Trong mô hình này, Swin Transformer đảm nhiệm vai trò mã hóa đặc trưng từ ảnh đầu vào. Transformer Decoder được sử dụng để sinh mô tả hình ảnh một cách chính xác và mạch Swin Transformer là bản mở rộng từ Vision Transformer với thiết kế chia cửa sổ dịch chuyển. Thiết kế này giúp mô hình nhận diện chi tiết tốt hơn ở các mức độ không gian

khác nhau. Transformer Decoder sử dụng self-attention để nắm bắt thông tin ngữ cảnh trong quá trình sinh câu. Mô hình được huấn luyện trên ba bộ dữ liệu: MS COCO [17], Flickr30k [18] và KTVIC [19]. KTVIC là tập dữ liệu tiếng Việt cho mục tiêu ứng dụng người Việt Nam. Sau khi huấn luyện, mô hình được tích hợp vào một ứng dụng Android hỗ trợ người khiếm thị. Người dùng có thể chụp ảnh và nhận mô tả bằng âm thanh qua tính năng chuyển văn bản thành giọng nói. Toàn bộ quá trình diễn ra nhanh chóng, thuận tiện và phù hợp với điều kiện sử dụng thực tế.

Đóng góp chính của bài báo này bao gồm:

- Đề xuất mô hình kết hợp Swin Transformer và Transformer Decoder nhằm nâng cao độ chính xác chú thích ảnh, tận dụng cơ chế tự chú ý (seft-attention) mạnh mẽ của kiến trúc Transformer.
- Thực nghiệm trên tập dữ liệu chuẩn MS COCO và Flickr 30K cho thấy mô hình chú thích ảnh đề xuất có độ chính xác vượt trội trên các độ đo phổ biến so với các công trình công bố gần đây.
- Phát triển ứng dụng di động Android tích hợp mô hình chú thích ảnh và công nghệ Text-to-Speech với giao diện thân thiện, hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh một cách thuận tiện và hiệu quả.

Phần còn lại của bài báo được tổ chức như sau: phần 2 trình bày các công trình nghiên cứu liên quan; phần 3 mô tả chi tiết phương pháp đề xuất; phần 4 trình bày thực nghiệm và phân tích kết quả; cuối cùng, phần 5 đưa ra kết luận và định hướng nghiên cứu trong tương lai.

II. CÔNG TRÌNH LIÊN QUAN

Các nghiên cứu ban đầu về chú thích ảnh theo tiếp cận học sâu thường dựa trên việc trích xuất đặc trưng bằng CNN. CNN được dùng để trích xuất thông tin hình ảnh, sau đó đưa vào mô hình ngôn ngữ như RNN hoặc LSTM để phát sinh chú thích. Show and Tell [12] là một ví dụ tiêu biểu, kết hợp CNN và RNN trong mô hình encoder–decoder. Phương pháp này giúp tạo mô tả tự động cho ảnh đầu vào. Show, Attend and Tell [20] mở rộng mô hình với cơ chế attention để tập trung vào vùng ảnh quan trọng. Azhar và cộng sự [21] sử dụng CNN để trích xuất đặc trưng toàn cục của hình ảnh. Thêm vào đó, mạng phát hiện đối tượng được kết hợp để lấy thông tin vùng ảnh chi tiết. Sau đó, đặc trưng toàn cục và vùng được đưa vào LSTM với cơ chế attention để sinh chú thích. Tiếp nối hướng này, Patwari và cộng sự [22] đề xuất kiến trúc encoder–decoder khác. CNN Inception-v3 được dùng làm encoder, GRU làm decoder với tích hợp attention. Mô hình được huấn luyện trên MS COCO và đánh giá bằng BLEU từ 1 đến 4. Kết quả cho thấy hiệu quả của phương pháp so với các mô hình trước đó. Xie và cộng sự [23] kết hợp Faster R-CNN để trích xuất đặc trưng vùng đối tượng. Đặc trưng vùng được đưa vào Bidirectional LSTM để sinh mô tả ảnh. Mô hình được thử nghiệm trên Flickr30k và MS COCO với độ chính xác cao. Kết quả vượt một số mô hình baseline đã công bố trước đó. Tuy vậy, phần lớn các mô hình trên chỉ dùng CNN huấn luyện trước để trích xuất đặc trưng. Điều này khiến mô hình khó học sâu mối quan hệ giữa các vùng trong ảnh. Do đó, mô hình gặp khó khăn khi cần biểu diễn ngữ nghĩa hình ảnh một cách toàn diện. LSTM cũng tồn tại hạn chế trong huấn luyện tuần tự và gặp vấn đề về gradient. Việc triệt tiêu hoặc bùng nổ đạo hàm làm giảm hiệu quả của quá trình sinh chú thích.

Do các mô hình kết hợp CNN và LSTM còn tồn tại nhiều hạn chế, nhiều nghiên cứu đã chuyển sang sử dụng Transformer. Transformer được xem là một lựa chọn tiềm năng cho bài toán chú thích ảnh. Wang và cộng sự [24] đã phát triển một mô hình sử dụng Transformer thay cho kiến trúc CNN–LSTM. Mục tiêu của mô hình là cải thiện độ chính xác và khả năng tổng quát hóa. Yang và cộng sự [25] đề xuất Transformer có khả năng nhận biết ngữ cảnh để tăng hiệu suất sinh mô tả. Cornia và cộng sự [26] giới thiệu Meshed-Memory Transformer với các kết nối dạng lưới tích hợp bộ nhớ. Cấu trúc này giúp mô hình học được quan hệ phức tạp giữa các vùng trong hình ảnh. Một số công trình gần đây đã áp dụng Transformer Decoder vào ứng dụng hỗ trợ người khiếm thị. Abubeker và cộng sự [27] sử dụng Transformer Decoder kết hợp encoder là mạng Inception. Harshitha và cộng sự [28] thay thế encoder bằng EfficientNet để trích xuất đặc trưng đầu vào. Những mô hình này cho thấy tiềm năng trong các hệ thống hỗ trợ tiếp cận hình ảnh.

Manay và cộng sự [10] đã phát triển một mô hình chú thích ảnh dựa trên GRU. Mô hình này được triển khai thành ứng dụng Android hỗ trợ người khiếm thị nhận biết môi trường xung quanh. Để sử dụng, người dùng phải thực hiện nhiều thao tác thủ công. Các bước bao gồm mở khóa điện thoại, bật Google Assistant và gọi lệnh để khởi chạy ứng dụng. Sau đó, người dùng cần chụp ảnh để mô hình xử lý và tạo chú thích. Quy trình yêu cầu nhiều lệnh điều khiển, gây bất tiện cho người khiếm thị. Trong một nghiên cứu khác, Kavitha và cộng sự [11] cũng đề xuất mô hình hỗ trợ tương tự. Ứng dụng này chụp ảnh qua camera và tạo mô tả bằng mô hình học sâu. Chú thích sinh ra được chuyển thành giọng nói và phát lại cho người dùng. Mô hình được huấn luyện trên tập dữ liệu MS COCO gồm ảnh và chú thích. Đặc trưng ảnh được trích xuất bằng EfficientNet-B3 và đưa vào mạng RNN. Tuy nhiên, mô hình vẫn gặp hạn chế do sử dụng kiến trúc CNN–LSTM. Ngoài ra, ứng dụng chỉ hỗ trợ ngôn ngữ tiếng Anh nên không phù hợp cho đa số người Việt.

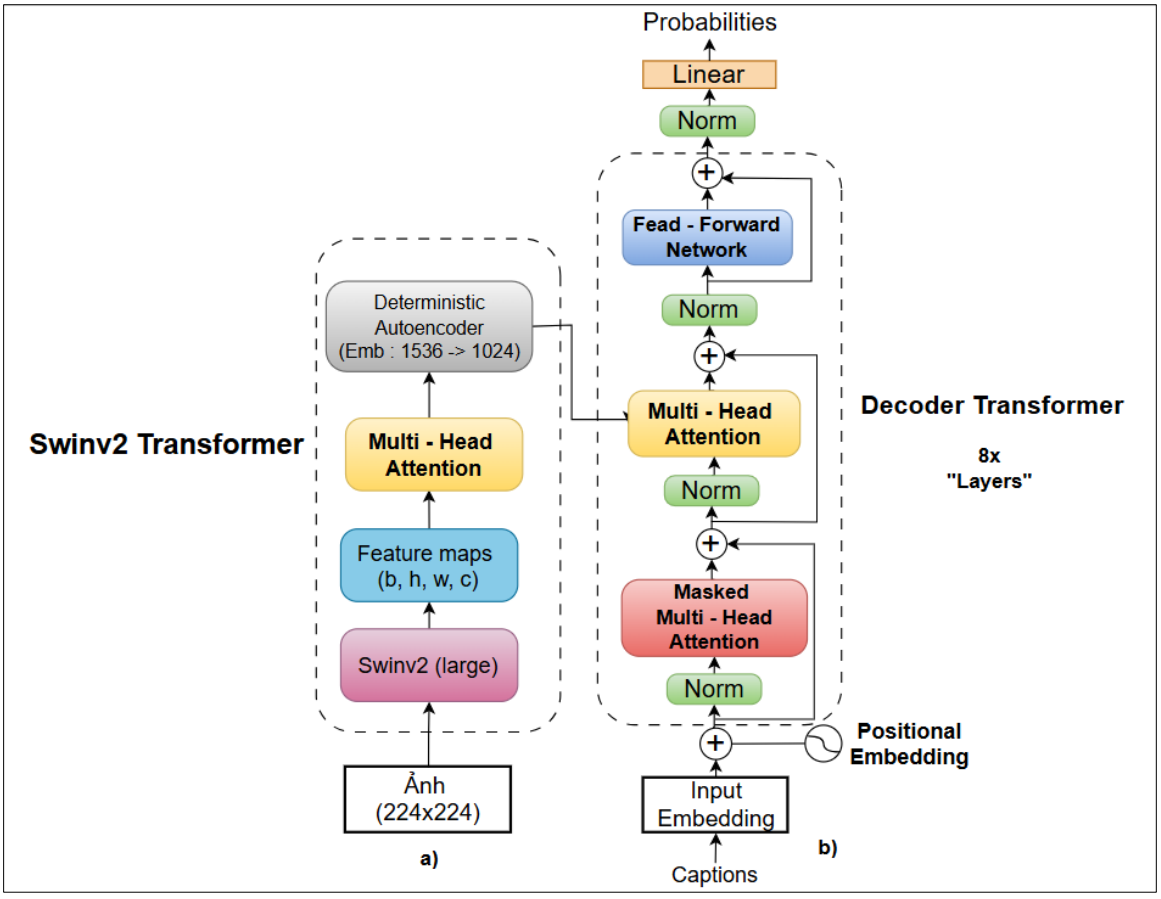
Qua quá trình phân tích các công trình công bố trước đây, có thể nhận thấy nhiều mô hình đã đạt được những bước tiến đáng kể. Các kiến trúc học sâu giúp nâng cao chất lượng mô tả hình ảnh theo thời gian. Tuy vậy, nhiều phương pháp vẫn còn những hạn chế nhất định trong việc nắm bắt ngữ nghĩa. Cụ thể, các phương pháp dựa trên CNN và LSTM còn gặp khó khăn trong việc biểu diễn ngữ nghĩa toàn cục và duy trì thông tin ngữ cảnh dài hạn. Ngược lại, Transformer cho thấy khả năng vượt trội về biểu diễn và học quan hệ dài hạn. Tuy nhiên, các nghiên cứu này chủ yếu được triển khai trong môi trường thử nghiệm. Mặt khác, các ứng dụng dành cho người khiếm thị chưa thân thiện trong thao tác sử dụng. Xuất phát từ những hạn chế đó, nghiên cứu này đề xuất một mô hình mới. Mô hình kết hợp Swin Transformer và Transformer Decoder để sinh mô tả ảnh. Đặc biệt, mô hình được tích hợp vào ứng dụng Android hỗ trợ người khiếm thị. Người dùng có thể chụp ảnh và nhận mô tả bằng âm thanh một cách thuận tiện. Giải pháp đề xuất hướng tới khả năng ứng dụng thực tế cao và cải thiện trải nghiệm người dùng.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Nghiên cứu này đề xuất một phương pháp xây dựng mô hình chú thích ảnh dựa trên kiến trúc học sâu. Đồng thời, một ứng dụng hỗ trợ người khiếm thị tiếp cận hình ảnh cũng được phát triển. Ứng dụng kết hợp mô tả ảnh tự động và chuyển văn bản thành âm thanh để phát lại. Phương pháp gồm hai thành phần chính hoạt động bổ trợ lẫn nhau. Thành phần thứ nhất là (A) mô hình sinh chú thích ảnh sử dụng kiến trúc Transformer. Thành phần thứ hai là (B) ứng dụng Android hỗ trợ người khiếm thị mô tả nội dung hình ảnh.

A. MÔ HÌNH CHÚ THÍCH ẢNH SỬ DỤNG TRANSFORMER

Kiến trúc Swin Transformer sử dụng trong nghiên cứu này bao gồm các thành phần chính như SwinV2-Large, cơ chế Multi-Head Self-Attention và Deterministic Autoencoder nhằm trích xuất đặc trưng thị giác hiệu quả theo cấu trúc phân cấp. Trong khi đó, Transformer Decoder được cấu thành từ các khối Masked Multi-Head Attention, Cross-Attention, Feed-Forward Network, Normalization và lớp tuyến tính, đảm nhiệm việc sinh chú thích từ đặc trưng ảnh. Việc lựa chọn kiến trúc này nhằm khai thác tối ưu khả năng của Transformer trong việc xử lý đồng thời thông tin thị giác và ngôn ngữ, từ đó tạo ra các mô tả chính xác và ngữ nghĩa mạch lạc.



Hình 1. *Sơ đồ tổng quát của mô hình chú thích ảnh đề xuất*

1. BỘ MÃ HÓA HÌNH ẢNH

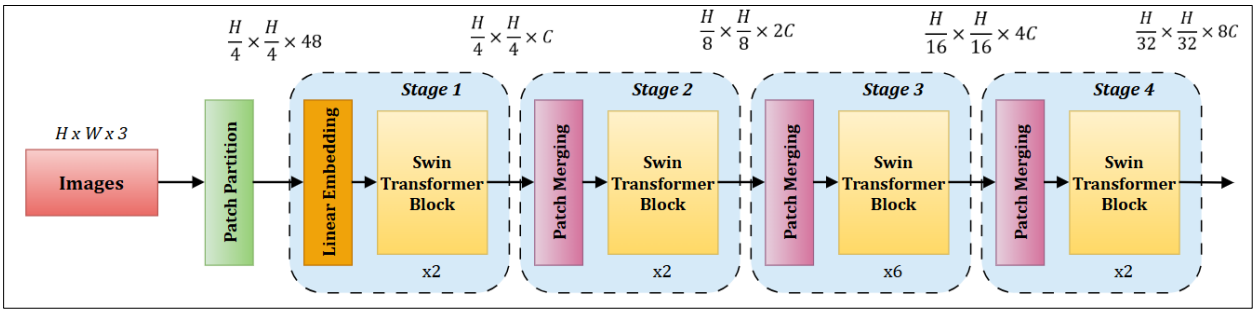
Swin Transformer là một kiến trúc Transformer phân cấp được đề xuất bởi Liu và cộng sự [16], nhằm khắc phục các hạn chế của Vision Transformer trong xử lý ảnh có độ phân giải cao, như chi phí tính toán lớn và thiếu khả năng khai thác đặc trưng cục bộ. Thay vì tính attention toàn cục, mô hình sử dụng cơ chế self-attention trong các cửa sổ không chồng lắp, kết hợp với cửa sổ dịch chuyển (shifted windows) giữa các tầng liên tiếp để đảm bảo khả năng trao đổi thông tin giữa các vùng ảnh lân cận. Ngoài ra, Swin Transformer còn áp dụng kiến trúc phân cấp, trong đó độ phân giải không gian giảm dần và chiều đặc trưng tăng dần qua từng giai đoạn, tương tự như kiến trúc CNN.

Hình 2 minh họa tổng quan kiến trúc Swin Transformer, bao gồm bốn giai đoạn xử lý. Cụ thể, với ảnh đầu vào $I \in \mathbb{R}^{H\times W\times 3}$ (H là chiều cao, W là chiều rộng của ảnh), mô hình chia ảnh thành các patch không chồng lắp kích thước 4×4, sau đó ánh xạ mỗi patch thành vector nhúng qua lớp chiếu tuyến tính, tạo ra ma trận đặc trưng đầu vào $X \in \mathbb{R}^{N\times D}$, trong đó $N = \frac{H\times W}{16}$ và D là số chiều không gian nhúng. Qua bốn giai đoạn xử lý, số lượng patch giảm bốn lần tại mỗi tầng, trong khi chiều đặc trưng tăng lên.

Đầu ra cuối cùng của bộ mã hóa là ma trận đặc trưng:

$$F_I = SwinTranformer(I) \in \mathbb{R}^{N'\times D'},$$

Trong đó, N'(nhỏ hơn rất nhiều so với N) là số patch sau khi giảm độ phân giải (downsample) thông qua các tầng patch merging, D' là số chiều của véc-tơ đặc trưng cuối cùng. Các đặc trưng này sau đó được đưa vào bộ giải mã mã ngôn ngữ để phát sinh chú thích cho hình ảnh.



Hình 2. *Kiến trúc tổng quát của Swin Transformer. Mỗi stage gồm một số khối Swin Transformer Block, được chèn giữa các lớp gộp patch (Patch Merging) để giảm kích thước không gian và tăng chiều đặc trưng.*

2. BỘ GIẢI MÃ TRANSFORMER

Sau khi mã hóa hình ảnh thành vector đặc trưng F_I, Transformer Decoder được sử dụng để phát sinh chú thích. Decoder hoạt động theo cơ chế sinh tuần tự có điều kiện (auto-regressive), trong đó từng từ được sinh ra dựa trên các từ trước đó và đặc trưng ảnh đã mã hóa. Mô hình áp dụng cơ chế Multi-Head Attention và Cross-Attention để tối ưu hóa việc kết hợp thông tin hình ảnh và văn bản.

Bộ giải vã nhận hai đầu vào chính: (1) tập đặc trưng của ảnh đầu vào từ bộ mã hóa $F_I \in \mathbb{R}^{N'\times D'}$ (đóng vai trò Key và $Value$ trong Cross-Attention), và (2) chuỗi các $token$ đã được sinh ra trước đó $Y = (y_1, y_2, \dots, y_{t-1})$, chuỗi này được đưa qua lớp nhúng từ vựng và mã hóa vị trí.

Kiến trúc bộ giải mã gồm 3 thành phần chính: (i) Lớp nhúng từ vựng (Token Embedding) để ánh xạ các token vào không gian vector nhúng (ii) Lớp mã hóa vị trí (Positional Encoding) có chức năng thêm thông tin vị trí vào từng token nhúng, giúp mô hình nắm bắt thứ tự từ trong chuỗi, và (iii) Các khối Transformer Decoder, mỗi khối gồm ba cơ chế chính:

- **Masked Multi-Head Self-Attention**: Áp dụng attention giữa các từ đã sinh, đồng thời sử dụng mặt nạ để đảm bảo mô hình không nhìn thấy các từ tương lai.

$$MaskedMultiHeadAttention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}} + M\right)V$$

  Trong đó, Q, K, V lần lượt là các ma trận Query, Key, Value được sinh ra từ các $token$ đã có; M là ma trận mặt nạ (masked), che các từ phía sau để đảm bảo mô hình không nhìn thấy; d_K là số chiều của véc-tơc key, được sử dụng để chuẩn hóa dot-product nhằm ổn định giá trị đầu vào cho hàm softmax.
- **Cross-Attention**: liên kết đặc trưng hình ảnh từ bộ mã hóa với các token đang được sinh ra.

$$CrossAttention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

Với, Q là chuỗi token đang sinh ra, K,V từ đặt trưng hình ảnh F_I.

- **Feed-Forward Network (FFN)**: tăng cường khả năng biểu diễn phi tuyến của mô hình:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2$$

Với W_1, W_2 là các ma trận trọng số; x_1, x_2 là các hệ số điều chỉnh (bias), x là đầu vào từ Attention.

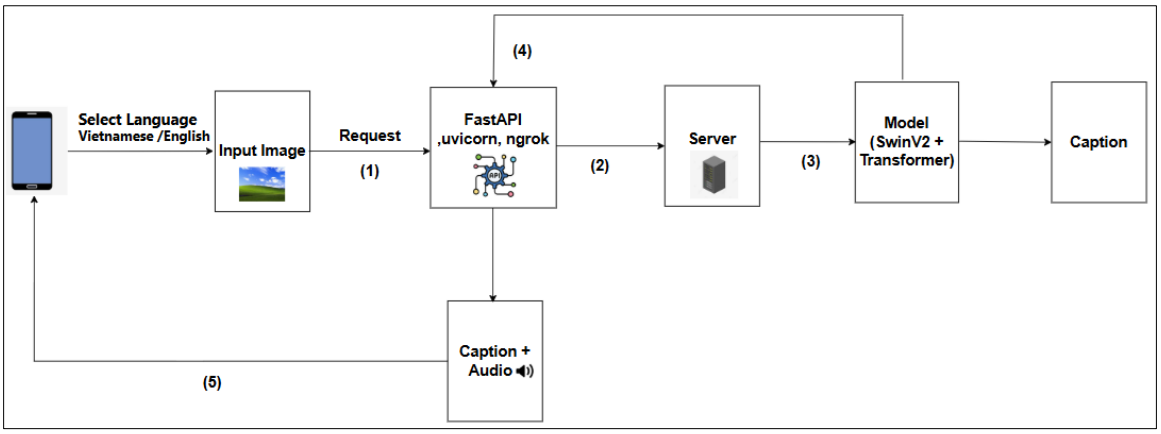
Tại mỗi thời điểm t, bộ giả mã sinh ra một phân phối xác suất trên toàn bộ tập từ vựng, ký hiệu là:

$$P(y_t|y_1, \dots y_{t-1}, F_I)$$

Trong đó, mỗi giá trị trong phân phối thể hiện xác suất của một từ cụ thể được chọn làm từ kế tiếp. Từ có xác suất cao nhất sẽ được chọn để sinh ra tại bước tiếp theo. Quá trình sinh chú thích kết thúc khi mô hình tạo ra token kết thúc câu.

B. XÂY DỰNG ỨNG DỤNG CHÚ THÍCH ẢNH

Sau khi huấn luyện mô hình chú thích ảnh sử dụng kiến trúc Swin Transformer V2 kết hợp Transformer Decoder, chúng tôi triển khai mô hình vào một ứng dụng di động nhằm hỗ trợ người khiếm thị trong việc nhận diện và tiếp cận thông tin hình ảnh thông qua mô tả ngôn ngữ. Kiến trúc tổng thể của hệ thống được minh họa trong Hình 3. Hệ thống ứng dụng bao gồm hai thành phần chính: (i) ứng dụng frontend chạy trên thiết bị Android và (ii) máy chủ backend đảm nhiệm xử lý và sinh chú thích. Ảnh được chụp trực tiếp từ thiết bị di động sẽ được gửi đến máy chủ thông qua API được xây dựng bằng FastAPI, trong đó ngrok và uvicorn được sử dụng để đảm bảo kết nối linh hoạt trong môi trường phát triển. Tại backend, ảnh đầu vào được đưa vào mô hình đã huấn luyện để sinh chú thích dạng văn bản, sau đó được trả về thiết bị người dùng và chuyển đổi thành âm thanh thông qua thư viện flutter_tts tích hợp trong Android.



Hình 3. *Kiến trúc ứng dụng chú thích ảnh hỗ trợ người khiếm thị*

Với mục tiêu tối ưu khả năng tiếp cận cho người khiếm thị, giao diện người dùng được thiết kế tối giản, sử dụng nền tối và chữ sáng nhằm tăng độ tương phản, đồng thời hỗ trợ điều hướng bằng âm thanh và thao tác chạm đơn giản. Ngay khi khởi động, ứng dụng tự động phát hướng dẫn sử dụng bằng giọng nói. Các chức năng chính như chụp ảnh, thay đổi ngôn ngữ (tiếng Việt/ tiếng Anh) và nhận mô tả ảnh đều có thể thực hiện thông qua thao tác một hoặc hai lần chạm, không yêu cầu thao tác thị giác. Người dùng có thể thay đổi ngôn ngữ bằng cách giữ tay trên màn hình trong vài giây. Cách thao tác này phù hợp khi không có giao diện hiển thị trực quan. Toàn bộ quá trình xử lý được thực hiện theo thời gian thực, không có độ trễ đáng kể. Hệ thống phản hồi nhanh và chính xác, đảm bảo trải nghiệm sử dụng ổn định. Kết quả cho thấy mô hình học sâu có thể được tích hợp hiệu quả vào ứng dụng thực tế. Giải pháp này hứa hẹn hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh dễ dàng hơn.

IV. THỰC NGHIỆM

Dựa trên mô hình đã đề xuất, phần này trình bày quá trình triển khai thực nghiệm nhằm đánh giá hiệu quả của phương pháp thông qua các độ đo phổ biến trong bài toán chú thích ảnh. Các kết quả định lượng và phân tích so sánh với các phương pháp liên quan được trình bày chi tiết, bên cạnh đó là một số minh họa trực quan từ ứng dụng di động Android được phát triển nhằm kiểm chứng tính khả thi của hệ thống trong điều kiện sử dụng thực tế.

A. DỮ LIỆU VÀ THIẾT LẬP THỰC NGHIỆM

Phần này trình bày chi tiết các tập dữ liệu được sử dụng để huấn luyện và đánh giá mô hình, cùng với các thông số cài đặt trong quá trình thực nghiệm.

1. DỮ LIỆU THỰC NGHIỆM

Trong nghiên cứu này, hai tập dữ liệu chuẩn là MS COCO [17] và Flickr30k [29] được sử dụng để đánh giá mô hình. MS COCO gồm khoảng 123.000 ảnh, còn Flickr30k chứa 31.000 ảnh. Mỗi ảnh trong hai tập dữ liệu đều được gán kèm 5 câu mô tả khác nhau. Để đảm bảo tính nhất quán khi so sánh với các nghiên cứu trước, dữ liệu được chia theo cách chuẩn hóa. MS COCO được tách thành ba phần theo đề xuất của Karpathy [30]. Cụ thể gồm 82.783 ảnh huấn luyện, 5.000 ảnh kiểm tra và 5.000 ảnh dùng để xác nhận. Với Flickr30k, việc chia tách tuân theo thiết lập của [31] trong các nghiên cứu gần đây. Tập này gồm 1.000 ảnh validation, 1.000 ảnh test, phần còn lại dùng để huấn luyện mô hình. Dữ liệu văn bản được tiền xử lý để loại bỏ các từ ít xuất hiện. Các từ xuất hiện dưới 5 lần đều bị loại khỏi tập từ vựng. Sau xử lý, tập từ vựng của MS COCO còn 10.010 từ. Flickr30k có 7.414 từ sau khi áp dụng cùng tiêu chí loại bỏ. Chiều dài tối đa của một câu mô tả trong cả hai tập là 16 từ.

Bên cạnh đó, bộ dữ liệu KTVIC [19] bằng tiếng Việt cũng được sử dụng nhằm đánh giá khả năng tổng quát hóa của mô hình trong ngữ cảnh ngôn ngữ bản địa. KTVIC bao gồm 4.327 ảnh và 21.635 câu chú thích được tạo thủ công, phản ánh các tình huống sinh hoạt đời thường trong môi trường thực tế. Đặc biệt, bộ dữ liệu này còn được sử dụng để thử nghiệm hiệu quả của mô hình khi tích hợp vào ứng dụng di động hỗ trợ người khiếm thị, qua đó đánh giá tính khả dụng và hiệu suất hoạt động của hệ thống trong điều kiện sử dụng thực tiễn.

2. CHI TIẾT CÀI ĐẶT

Để đảm bảo hiệu quả biểu diễn và khả năng phát sinh chú thích từ ảnh đầu vào, mô hình được thiết kế với cấu trúc encoder–decoder tối ưu hóa dựa trên kiến trúc Transformer hiện đại. Thành phần mã hóa và giải mã được cấu hình chi tiết như sau:

Bộ mã hóa hình ảnh trong mô hình sử dụng biến thể *swinv2-large-window12-192*, một kiến trúc hierarchical transformer huấn luyện trước trên ImageNet-22k và triển khai từ thư viện *timm của pytorch*. Mô hình này áp dụng cơ chế local self-attention với cửa sổ trượt 12×12 trên ảnh đầu vào 192×192 pixel. Đầu ra có chiều sâu 1536 được sắp xếp lại và tinh chỉnh bằng multi-head self-attention để tăng cường biểu diễn không gian toàn cục.

Bộ giải mã ngôn ngữ Transformer Decoder: với số *block* = 8, số *head* = 16, số chiều cho véc-tơ nhúng từ là 1024, số chiều lớp FFN là 4096, dropout 0.1, hàm kích hoạt GeLU và sử dụng chiến lược norm-first. Đặc trưng ảnh trước khi đưa vào decoder được giảm chiều bằng Deterministic Autoencoder (DAEProjection) nhằm nén thông tin trọng yếu, tối ưu theo hàm mất mát MSELoss. Quá trình huấn luyện sử dụng AdamW với learning rate 3×10^{-4}, weight decay 0.01 và chiến lược điều chỉnh CosineAnnealingWarmRestarts. Mô hình kết hợp mixed precision training với GradScaler, và tối ưu bởi tổng hợp hai hàm mất mát: CrossEntropyLoss (cho sinh ngôn ngữ) và MSELoss (cho DAE) cùng với các kỹ thuật như gradient clipping, khởi tạo trọng số có kiểm soát, và DataParallel để tối ưu hiệu suất huấn luyện.

B. ĐỘ ĐO ĐÁNH GIÁ

Hiệu quả của mô hình chú thích ảnh được đánh giá thông qua các độ đo chuẩn được sử dụng rộng rãi trong bài toán này, bao gồm BLEU [32], METEOR [33] và CIDEr [34]. BLEU (Bilingual Evaluation Understudy Score) đo lường mức độ khớp n-gram giữa câu sinh ra và câu tham chiếu, thường được tính theo BLEU-1, BLEU-2, BLEU-3 và BLEU-4 tương ứng với độ dài n-gram từ 1 đến 4. METEOR (Metric for Evaluation of Translation with Explicit ORdering) là một cải tiến so với BLEU khi xem xét các phép ánh xạ linh hoạt hơn như đồng nghĩa, dạng gốc (stem) và hoán vị từ. CIDEr (Consensus-based Image Description Evaluation) sử dụng TF-IDF của n-gram và tính điểm dựa trên sự đồng thuận giữa các chú thích tham chiếu. Mỗi độ đo có cách tính toán và đánh giá kết quả theo các khía cạnh khác nhau, tuy nhiên tất cả đều có đặc điểm chung là giá trị càng cao cho biết hiệu suất sinh chú thích càng tốt.

C. KẾT QUẢ THỰC NGHIỆM VÀ BÀN LUẬN

Hiệu quả của mô hình chú thích ảnh đề xuất được đánh giá trên ba tập dữ liệu gồm MS COCO, Flickr30k và KTVIC, với các độ đo phổ biến BLEU-1, BLEU-4, METEOR và CIDEr. Kết quả được trình bày trong Bảng 1 cho thấy mô hình đạt hiệu suất khá cao và ổn định trên cả ba tập dữ liệu, trong đó kết quả trên tập MS COCO là tốt nhất với BLEU-1 đạt 72.5, BLEU-4 đạt 30.3, METEOR đạt 27.5 và CIDEr đạt 85.2. Trên tập Flickr30k, mô hình đạt BLEU-1 là 67.3, BLEU-4 là 27.9, METEOR là 26.2 và CIDEr là 70.3. Đối với tập dữ liệu tiếng Việt KTVIC, mô hình vẫn cho thấy khả năng tổng quát hóa tốt với các kết quả BLEU-1 là 65.2, BLEU-4 là 21.5, METEOR là 26.9 và CIDEr là 58.5, mặc dù tập dữ liệu này có quy mô nhỏ hơn và mang tính bản địa hóa cao hơn.

Sự khác biệt hiệu suất của mô hình SwinV2–Transformer Decoder trên ba tập dữ liệu (Bảng 1) phản ánh ảnh hưởng trực tiếp của quy mô, chất lượng chú thích, cũng như tính chất ngôn ngữ và ngữ cảnh trong từng tập. Cụ

thể, mô hình đạt kết quả cao nhất trên MSCOCO với BLEU-4 = 30.3 và CIDEr = 85.2, do tập dữ liệu này có quy mô lớn (82.783 ảnh huấn luyện và 5.000 ảnh kiểm tra), chú thích được chuẩn hóa cao và thống nhất về phong cách ngôn ngữ, giúp mô hình học được phân bố ngữ nghĩa rõ ràng và phản hồi tốt hơn với các biểu đạt quen thuộc. Trên Flickr30k, hiệu suất tuy thấp hơn một chút (BLEU-4 = 27.9, CIDEr = 70.3) nhưng vẫn giữ ở mức cao, nhờ nội dung ảnh đa dạng và chú thích mang phong cách diễn đạt tự nhiên hơn, từ đó đánh giá được khả năng mô hình thích nghi với sự linh hoạt ngôn ngữ trong thực tế. Trong khi đó, trên KTVIC tập dữ liệu tiếng Việt có quy mô nhỏ hơn nhiều (4.327 ảnh) và tính địa phương hóa cao, hiệu suất của mô hình giảm nhẹ (BLEU-4 = 21.5, CIDEr = 58.5), chủ yếu do hạn chế về số lượng mẫu huấn luyện, cũng như các đặc điểm ngữ pháp, từ vựng và biểu đạt văn hóa đặc thù của tiếng Việt. Dù vậy, mô hình vẫn duy trì hiệu suất ổn định, cho thấy khả năng tổng quát hóa tốt của kiến trúc SwinV2 kết hợp Transformer Decoder khi chuyển đổi giữa các miền ngôn ngữ và thị giác khác nhau. Nhìn chung, sự chênh lệch hiệu suất giữa các tập là hợp lý về mặt thống kê và phản ánh tác động của đặc điểm ngôn ngữ, chất lượng dữ liệu và phạm vi biểu đạt. Mô hình đề xuất không chỉ cho thấy hiệu quả nổi bật trên các tập chuẩn tiếng Anh mà còn khẳng định tiềm năng mở rộng ứng dụng sang các ngôn ngữ khác.

Bảng 2 so sánh hiệu suất giữa mô hình đề xuất và một số phương pháp chú thích ảnh baseline và tiên tiến trên tập MS COCO. Kết quả cho thấy mô hình của chúng tôi vượt trội hơn so với các phương pháp baseline như Show and Tell [12], Show, Attend and Tell (Soft/Hard Attention) [13] cũng như các phương pháp gần đây như En-De-Cap [22] và Bi-LS-AttM [23]. Cụ thể, mô hình đạt BLEU-4 là 30.3, cao hơn đáng kể so với mức 25.2 của Bi-LS-AttM và 24.3 của En-De-Cap. Đồng thời, điểm METEOR đạt 27.5 và CIDEr là 85.7, đều cao hơn so với các phương pháp so sánh. Những kết quả này không chỉ xác nhận ưu thế của mô hình đề xuất, mà còn cho thấy hiệu quả thực sự của việc sử dụng Swin Transformer làm backbone trong bài toán chú thích ảnh. Việc kết hợp cấu trúc attention đa cấp của Swin với Transformer Decoder giúp mô hình học được thông tin không gian ngữ cảnh sâu hơn, từ đó sinh ra các chú thích rõ ràng, chính xác và giàu ngữ nghĩa hơn các phương pháp so sánh trong bảng 2.

Bảng 1. *Hiệu suất của phương pháp chú thích ảnh đề xuất trên các tập dữ liệu thực nghiệm*

Tập dữ liệu	BLEU-1	BLEU-4	METEOR	CIDEr
KTVIC	65.2	21.5	26.9	58.5
FLICKr30k	67.3	27.9	26.2	70.3
MS COCO	**72.5**	**30.3**	**27.5**	**85.2**

Bảng 2. *So sánh hiệu suất giữa các phương pháp chú thích ảnh trên tập dữ liệu MS COCO*

Phương pháp	BLEU1	BLEU4	METEOR	CIDEr
Show and Tell (2015) [12]	-	27.7	23.7	85.5
Show, attend and tell (Soft-ATT) (2015) [13]	70.7	24.3	23.9	-
Show, attend and tell (Hard-ATT) (2015) [13]	71.8	25	23.0	-
En-De-Cap (2021) [22]	70.6	24.3	-	-
Bi-LS-AttM (2023) [23]	68.8	25.2	21.5	41.2
Đề xuất của chúng tôi	**72.5**	**30.3**	**27.5**	**85.7**

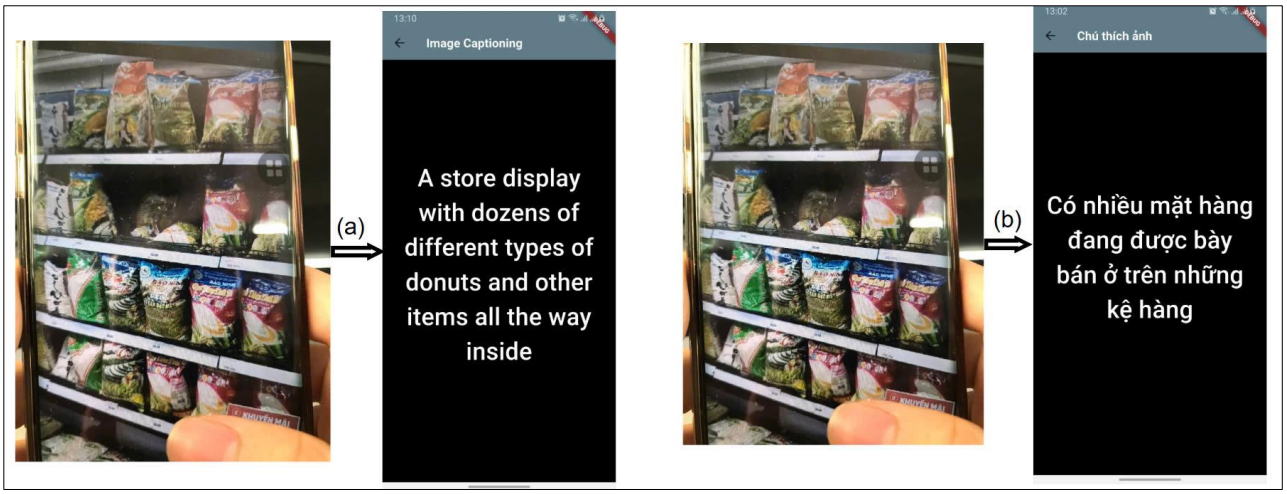
Trên tập Flickr30k, như trình bày trong Bảng 3, mô hình đề xuất tiếp tục đạt kết quả cao hơn rõ rệt so với các phương pháp baseline và gần đây. Cụ thể, mô hình đạt BLEU-1 là 67.3 và BLEU-4 là 27.9, trong khi các phương pháp như Soft-Attention và Hard-Attention chỉ đạt BLEU-4 lần lượt là 19.1 và 19.9. Tương tự, điểm METEOR của mô hình chúng tôi là 26.2, cao hơn đáng kể so với mức dưới 20 của các phương pháp so sánh. Đặc biệt, trên độ đo CIDEr, mô hình đề xuất đạt 70.3, vượt xa các phương pháp như Xception và Xception với importance factor [35], vốn chỉ đạt lần lượt là 14.8 và 15. Các kết quả này củng cố nhận định rằng việc sử dụng Swin Transformer giúp khai thác thông tin không gian trong ảnh hiệu quả hơn, từ đó giúp Transformer Decoder tạo ra các chú thích chính xác và ngữ nghĩa hơn ngay cả trong bối cảnh tập dữ liệu đa dạng như Flickr30k.

Bảng 3. *So sánh hiệu suất giữa các phương pháp chú thích ảnh trên tập dữ liệu Flickr30k*

Phương pháp	BLEU1	BLEU4	METEOR	CIDEr
Show and Tell (2015) [12]	66.3	18.3	-	-
Show, attend and tell (Soft-ATT) (2015) [13]	66.7	19.1	19.49	-
Show, attend and tell (Hard-ATT) (2015) [13]	66.9	19.9	18.46	-
Bi-LS-AttM (2023) [23]	64.5	20.2	18.6	
Xception (2022) [35]	39.9	6.2	12.3	14.8
Xception + importance factor (2022) [35]	39.8	6.1	12.9	15.0
Đề xuất của chúng tôi	**67.3**	**27.9**	**26.2**	**70.3**

Từ các kết quả thực nghiệm cho thấy mô hình chú thích ảnh sử dụng Swin Transformer và Transformer Decoder có khả năng sinh mô tả chính xác, mạch lạc và giàu ngữ nghĩa. Mô hình không chỉ đạt hiệu suất cao trên các tập dữ liệu tiếng Anh chuẩn mà còn cho thấy khả năng thích ứng tốt với dữ liệu tiếng Việt, minh chứng cho tính linh hoạt và tiềm năng ứng dụng thực tiễn của phương pháp đề xuất.

Bên cạnh việc đánh giá hiệu quả mô hình trên các tập dữ liệu chuẩn, chúng tôi cũng tiến hành thử nghiệm tích hợp mô hình vào ứng dụng di động chạy trên thiết bị Android nhằm kiểm chứng khả năng ứng dụng thực tế trong bối cảnh hỗ trợ người khiếm thị. Ứng dụng được triển khai và đánh giá trên thiết bị Samsung Galaxy Note 9 với hệ điều hành Android 10, bộ vi xử lý Exynos 9810 và 6GB RAM.



Hình 5. *Minh họa kết quả tạo chú thích ảnh trên ứng dụng Android: (a) Chú thích được sinh bởi mô hình đề xuất huấn luyện trên tập dữ liệu MS COCO, và (b) chú thích sinh bởi mô hình tương tự nhưng huấn luyện trên tập dữ liệu tiếng Việt KTVIC.*

Hình 5 minh họa kết quả sinh chú thích ảnh trên ứng dụng Android trong hai cấu hình khác nhau: (a) sử dụng mô hình huấn luyện trên tập dữ liệu MS COCO và (b) sử dụng mô hình huấn luyện trên tập dữ liệu tiếng Việt KTVIC. Trong cả hai trường hợp, mô hình đều nhận diện được ngữ cảnh chính là các sản phẩm trưng bày trên kệ hàng, song chú thích bằng tiếng Việt trong (b) thể hiện tính khái quát và phù hợp hơn với ngữ cảnh sử dụng tại Việt Nam. Ứng dụng cho thấy khả năng hoạt động ổn định, với thời gian sinh chú thích trung bình từ 3 đến 6 giây, phụ thuộc vào chất lượng kết nối Internet, cho thấy độ trễ ở mức chấp nhận được trong môi trường sử dụng thực tế. Tuy nhiên, trong một số trường hợp ánh sáng yếu hoặc hình ảnh bị mờ, mô hình vẫn gặp khó khăn trong việc nhận diện chi tiết chính xác, dẫn đến sai lệch nhẹ trong mô tả. Dù vậy, kết quả thử nghiệm bước đầu đã chứng minh tính khả thi và hiệu quả ứng dụng của mô hình trong các tình huống thực tế.

Để nâng cao hiệu suất trong tương lai, một số hướng cải tiến bao gồm: (i) tối ưu mô hình cho các nền tảng học máy di động để cải thiện tốc độ xử lý, (ii) tinh chỉnh kiến trúc để giảm thời gian phản hồi, và (iii) tích hợp các phiên bản mô hình nhẹ nhằm giảm chi phí tính toán và tăng khả năng triển khai trong môi trường tài nguyên hạn chế.

V. KẾT LUẬN

Bài báo này đã trình bày một phương pháp chú thích ảnh hiệu quả dựa trên kiến trúc encoder–decoder, trong đó Swin Transformer được sử dụng làm bộ mã hóa để trích xuất đặc trưng hình ảnh, còn Transformer Decoder đảm nhiệm sinh mô tả ngôn ngữ. Mô hình được huấn luyện và đánh giá trên ba tập dữ liệu MS COCO, Flickr30k và KTVIC, cho kết quả vượt trội so với các phương pháp chú thích ảnh kinh điển và hiện đại trên cả bốn độ đo BLEU-1, BLEU-4, METEOR và CIDEr. Đặc biệt, hiệu suất cao trên tập dữ liệu tiếng Việt KTVIC cho thấy khả năng tổng quát hóa tốt của mô hình trong ngữ cảnh ngôn ngữ bản địa. Bên cạnh đánh giá định lượng, mô hình còn được triển khai thực tế vào một ứng dụng di động Android nhằm hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh thông qua mô tả bằng giọng nói. Kết quả thử nghiệm trên thiết bị thực cho thấy ứng dụng hoạt động ổn định và cho phản hồi nhanh trong các điều kiện sử dụng tiêu chuẩn, cho thấy tính khả thi cao của giải pháp trong bối cảnh thực tiễn. Trong tương lai, các hướng nghiên cứu mở bao gồm: tối ưu hóa mô hình cho nền tảng di động để cải thiện tốc độ xử lý, tích hợp các mô hình nhẹ hơn nhằm giảm thiểu chi phí tính toán, và kết hợp thêm thông tin ngữ nghĩa (semantic knowledge) để nâng cao chất lượng mô tả trong các tình huống phức tạp hơn.

VI. LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Nguồn ngân sách khoa học và công nghệ Trường Đại học Sư phạm Thành phố Hồ Chí Minh trong đề tài NCKH của sinh viên năm học 2024 – 2025.

VII. TÀI LIỆU THAM KHẢO

[1] WHO. (2023, January 3). *Blindness and vision impairment*. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[2] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys,* vol. 56, pp. 1-39, 2023.

[3] B. Jin, X. Liu, Y. Zheng, P. Li, H. Zhao, T. Zhang*, et al.*, "Adapt: Action-aware driving caption transformer," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7554-7561.

[4] Y. Mishra and K. Senapati, "Deep Learning-based Image Caption Generator for Real-time Monitoring and Predictive Control of Concentration of Polluting gases," 2025.

[5] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li*, et al.*, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *2019 IEEE international conference on data mining (ICDM)*, 2019, pp. 728-737.

[6] H. Ayesha, S. Iqbal, M. Tariq, M. Abrar, M. Sanaullah, I. Abbas*, et al.*, "Automatic medical image interpretation: State of the art and future directions," *Pattern Recognition,* vol. 114, p. 107856, 2021.

[7] D. I. Lee, J. H. Lee, S. H. Jang, S. J. Oh, and I. C. Doo, "Crop Disease Diagnosis with Deep Learning-Based Image Captioning and Object Detection," *Applied Sciences,* vol. 13, p. 3148, 2023.

[8] L. Kearns, "Content moderation assistance through image caption generation," *Intelligent Systems with Applications,* vol. 25, p. 200489, 2025.

[9] W.-L. Tsai, P.-L. Le, W.-F. Ho, N.-W. Chi, J. J. Lin, S. Tang*, et al.*, "Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention," *Automation in Construction,* vol. 169, p. 105863, 2025.

[10] S. P. Manay, S. A. Yaligar, Y. Thathva Sri Sai Reddy, and N. J. Saunshimath, "Image captioning for the visually impaired," in *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2020, Volume 1*, 2022, pp. 511-522.

[11] R. Kavitha, S. S. Sandhya, P. Betes, P. Rajalakshmi, and E. Sarubala, "Deep learning-based image captioning for visually impaired people," in *E3S Web of Conferences*, 2023, p. 04005.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov*, et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048-2057.

[14] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems,* 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner*, et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang*, et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan*, 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, pp. 740-755.

[18] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641-2649.
[19] A.-C. Pham, V.-Q. Nguyen, T.-H. Vuong, and Q.-T. Ha, "KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain," *arXiv preprint arXiv:2401.08100,* 2024.
[20] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044,* 2015.
[21] I. Azhar, I. Afyouni, and A. Elnagar, "Facilitated deep learning models for image captioning," in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021, pp. 1-6.
[22] N. Patwari 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, pp. 1192-1196.
[23] T. Xie, W. Ding, J. Zhang, X. Wan, and J. Wang, "Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning," *Applied Sciences,* vol. 13, p. 7916, 2023.
[24] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2585-2594.
[25] X. Yang, Y. Wang, H. Chen, J. Li, and T. Huang, "Context-aware transformer for image captioning," *Neurocomputing,* vol. 549, p. 126440, 2023.
[26] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10578-10587.
[27] A. K. Muhammed Kunju, S. Baskar, S. Zafar, B. AR, and R. S, "A transformer based real-time photo captioning framework for visually impaired people with visual attention," *Multimedia Tools and Applications,* pp. 1-20, 2024.
[28] V. Krishnamurthy, "Transeffivisnet–an image captioning architecture for auditory assistance for the visually impaired," *Multimedia Tools and Applications,* pp. 1-20, 2024.
[29] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics,* vol. 2, pp. 67-78, 2014.
[30] A. Karpathy 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, pp. 3128-3137.
[31] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems,* vol. 27, 2014.
[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.
[33] S. Banerjee 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, pp. 65-72.
[34] R. Vedantam, 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, pp. 4566-4575.
[35] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data,* vol. 9, p. 20, 2022.

IMAGE CAPTIONING FOR THE VISUALLY IMPAIRED USING TRANSFORMER

Nguyen Thien Khiem, Do Duc Dat, Nguyen Ngoc Hoai Tri, Vo Van Thinh, Van Thanh Thich, Nguyen Van Thinh

ABSTRACT— Visual impairment affects millions worldwide, posing significant challenges in accessing visual information. With the rapid development of mobile devices, particularly the Android platform, image-to-audio description applications are becoming increasingly popular. However, generating accurate, context-rich descriptions that are compatible with mobile deployment remains a considerable challenge. This study proposes an image captioning model based on the encoder-decoder architecture, in which the Swin Transformer is employed to extract hierarchical visual features and a Transformer Decoder is used to generate textual descriptions. The model is trained on standard datasets such as MS COCO and Flickr30k and fine-tuned on the Vietnamese-language KTVIC dataset to enhance its applicability in local contexts. Experimental results show that the model performs well on standard evaluation metrics including BLEU, METEOR, and CIDEr. In addition to the model, we developed an Android application that integrates image captioning and text-to-speech functionality, enabling real-time spoken descriptions to assist visually impaired users in accessing image content. The application demonstrates stable performance and responsive inference time under practical conditions. These results highlight the potential of the proposed approach in improving visual information accessibility for the visually impaired community.

Keywords: Automatic Image Captioning, Encoder–Decoder Model, Swin Transformer, Transformer Decoder, Applications for Assisting the Visually Impaired

Nguyễn Thiên Khiêm là sinh viên ngành Công nghệ thông tin, trường Đại học Sư Phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm về các nghiên cứu trong lĩnh vực học máy và ứng dụng.

Văn Thành Thích là sinh viên ngành Công nghệ thông tin thuộc trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện tại đang quan tâm đến các nghiên cứu trong lĩnh vực Trí tuệ nhân tạo và ứng dụng

Nguyễn Ngọc Hoài Trí là sinh viên năm ba ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm và theo đuổi các nghiên cứu trong lĩnh vực Học máy.

Đỗ Đức Đạt là sinh viên ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm và theo đuổi các nghiên cứu trong lĩnh vực Trí tuệ nhân tạo và ứng dụng

Võ Văn Thịnh là sinh viên ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm và theo đuổi các nghiên cứu trong lĩnh vực Học máy.

Nguyễn Văn Thịnh nhận bằng thạc sĩ ngành Hệ thống thông tin tại Trường Đại học Khoa học Tự nhiên – Đại học Quốc gia TP. Hồ Chí Minh năm 2012. Hiện nay, ông là nghiên cứu sinh ngành Khoa học máy tính tại Học viện Khoa học và Công nghệ – Viện Hàn lâm Khoa học và Công nghệ Việt Nam, đồng thời là giảng viên tại Khoa Công nghệ thông tin, Trường Đại học Sư phạm TP. Hồ Chí Minh. Nghiên cứu của ông tập trung vào trí tuệ nhân tạo và các ứng dụng, đặc biệt là học sâu đa phương thức và các mô hình ngôn ngữ–thị giác, hướng đến các bài toán như truy vấn thông tin, chú thích ảnh tự động và hỏi-đáp dựa trên hình ảnh.