DOI: https://doi.org/10.62985/j.huit_ojs.vol25.no4.334

HỆ CHÚ THÍCH ẢNH TỰ ĐỘNG CHO NGƯỜI KHIẾM THỊ

Đinh Thị Mận[1], Nguyễn Văn Thịnh[2], Trần Hữu Quốc Thư[2], Nguyễn Hải Yến[1], Nguyễn Phương Hạc[1], Trần Thị Vân Anh[1*]

[1]Trường Đại học Công Thương Thành phố Hồ Chí Minh

[2]Trường Đại học Sư phạm Tp.HCM

*Email: *anhttv@huit.edu.vn*

Ngày nhận bài: 26/01/2024; Ngày nhận bài sửa: 27/5/2024; Ngày chấp nhận đăng: 31/5/2024

TÓM TẮT

Suy giảm thị lực khiến người khiếm thị gặp nhiều trở ngại trong việc nhận biết và tương tác với môi trường xung quanh. Nhằm hỗ trợ khắc phục vấn đề này, nghiên cứu đề xuất một hệ thống chú thích ảnh tự động hoạt động đa nền tảng. Mô hình được thiết kế theo kiến trúc mã hóa–giải mã, trong đó DenseNet đảm nhận vai trò trích xuất đặc trưng hình ảnh, còn LSTM kết hợp với cơ chế chú ý để tạo ra mô tả ngôn ngữ. Phương pháp được huấn luyện và đánh giá trên hai bộ dữ liệu chuẩn MS COCO và Flickr30K, với các độ đo phổ biến như BLEU và METEOR. Kết quả cho thấy hệ thống đạt độ chính xác cao hơn nhiều phương pháp công bố gần đây. Ngoài ra, một phiên bản ứng dụng chạy trên Desktop và thiết bị di động cũng được phát triển, cho phép sinh mô tả ảnh dưới dạng âm thanh, góp phần hỗ trợ người khiếm thị tiếp cận thông tin trực quan.

Từ khóa: Chú thích ảnh tự động, CNN, LSTM, cơ chế chú ý, người khiếm thị.

1. GIỚI THIỆU

Suy giảm hoặc mất thị lực do bệnh lý, tai nạn hay tuổi tác gây ra nhiều trở ngại cho người khiếm thị trong sinh hoạt hằng ngày. Việc không thể quan sát trực tiếp khiến họ gặp khó khăn trong di chuyển và giao tiếp, đồng thời có xu hướng mong muốn được tự lập thay vì phụ thuộc vào sự trợ giúp. Trong những năm gần đây, nhiều nghiên cứu đã tập trung vào việc nâng cao khả năng tiếp cận thông tin cho cộng đồng người khiếm thị. Ví dụ, Đinh Điền và cộng sự [1] đã giới thiệu một số công cụ hỗ trợ như từ điển nói CLC MATA hay phần mềm Happy Sun giúp luyện gõ phím và đọc văn bản. Các giải pháp này phần nào cải thiện chất lượng cuộc sống, song vẫn thiếu những công cụ cho phép nhận diện và tương tác trực tiếp với môi trường thực tế. Do đó, việc nghiên cứu và phát triển hệ thống chú thích ảnh tự động được đặt ra như một nhu cầu thiết yếu, nhằm mang lại khả năng nhận biết cảnh vật cho người khiếm thị một cách độc lập và tự tin hơn.

Chú thích ảnh là bài toán kết hợp giữa nhận dạng hình ảnh và sinh văn bản mô tả tương ứng [2]. Mục tiêu cốt lõi của bài toán là tạo ra các chú thích tự nhiên, ngắn gọn, đúng ngữ pháp và phản ánh chính xác nội dung hình ảnh cũng như mối quan hệ giữa các đối tượng. Đây là một hướng nghiên cứu đa ngành, gắn kết thị giác máy tính với xử lý ngôn ngữ tự nhiên [3]. Nhờ đặc tính này, các phương pháp chú thích ảnh đã được ứng dụng trong nhiều lĩnh vực khác nhau, chẳng hạn như hỗ trợ chẩn đoán y khoa [4], nâng cao khả năng tương tác của robot với môi trường [3], hay phục vụ nhận diện và giám sát trong nông nghiệp [5].

Các phương pháp giải quyết bài toán chú thích ảnh có thể chia thành hai nhóm chính: cách tiếp cận truyền thống và các phương pháp dựa trên học sâu [6]. Từ sau năm 2015, hướng tiếp cận dựa trên mạng nơ-ron học sâu dần chiếm ưu thế nhờ khả năng khắc phục nhiều hạn chế của phương pháp truyền thống [7–10]. Đặc biệt, các mạng CNN huấn luyện sẵn như AlexNet, VGGNet, Inception, ResNet hay DenseNet đã chứng minh hiệu quả trong việc trích xuất đặc trưng hình ảnh cho nhiều tác vụ khác nhau. Mỗi kiến trúc có ưu và nhược điểm riêng, song điểm chung là khi mạng đủ sâu, đặc trưng thu được thường mang tính khái quát và hữu ích hơn. Tuy nhiên, độ sâu quá lớn có thể gây ra hiện tượng triệt tiêu đạo hàm (vanishing gradient). DenseNet ra đời như một giải pháp khắc phục hạn chế này bằng cơ chế kết nối dày đặc, bảo đảm luồng thông tin liên tục giữa các tầng. Song song đó, sự xuất hiện của cơ

chế chú ý đã giúp bộ giải mã tập trung vào những vùng quan trọng trong ảnh, loại bỏ các phần dư thừa và nhờ vậy cải thiện chất lượng mô tả [11].

Dựa trên những cơ sở đã phân tích, nghiên cứu này đề xuất một mô hình chú thích ảnh tự động, trong đó DenseNet được sử dụng để mã hóa đặc trưng hình ảnh, kết hợp với mạng LSTM có tích hợp cơ chế chú ý ở giai đoạn giải mã nhằm cải thiện độ chính xác của câu mô tả.

Những đóng góp chính của bài báo gồm:

- Khai thác DenseNet để rút trích hiệu quả các đặc trưng đa dạng từ hình ảnh;
- Xây dựng và huấn luyện mạng LSTM tích hợp cơ chế chú ý, giúp tập trung vào các vùng quan trọng trong quá trình sinh câu chú thích;
- Phát triển ứng dụng đa nền tảng (máy tính và thiết bị di động) hỗ trợ người khiếm thị tiếp cận nội dung hình ảnh thông qua âm thanh theo thời gian thực.

Bố cục bài báo được trình bày như sau: Phần 2 thảo luận các nghiên cứu liên quan; Phần 3 mô tả mô hình đề xuất và kiến trúc ứng dụng; Phần 4 thảo luận quá trình thực nghiệm và kết quả đạt được; cuối cùng, kết luận được đưa ra ở Phần 5.

2. CÁC CÔNG TRÌNH LIÊN QUAN

Trong những năm gần đây, hướng nghiên cứu chú thích ảnh dựa trên học sâu với cấu trúc mã hóa – giải mã đã thu hút được nhiều sự quan tâm [3, 12]. Nhiều mô hình đã được đề xuất với các cách kết hợp khác nhau giữa mạng nơ-ron tích chập và mạng nơ-ron hồi quy. Một số công trình tiêu biểu có thể kể đến như: CNN trích xuất đặc trưng ảnh, sau đó RNN chuyển đổi đặc trưng này thành câu mô tả [13]; mô hình sử dụng CNN trong vai trò bộ mã hóa và LSTM cho giai đoạn giải mã [14–16]; phương pháp dựa trên CNN, LSTM và Ontology để sinh chú thích theo vùng [9]; kiến trúc CNN kết hợp LSTM hai tầng [17]; hay cách tiếp cận sử dụng CNN kết hợp Transformer trong bài toán chẩn đoán bệnh cây trồng [5].

Năm 2021, Nikhil Patwari và cộng sự đã giới thiệu một mô hình dựa trên CNN và LSTM, trong đó đặc trưng hình ảnh được rút trích bằng Inception-v3 rồi đưa vào GRU (phiên bản rút gọn của LSTM) để sinh chú thích với cơ chế chú ý. Thử nghiệm trên tập dữ liệu MS COCO cho thấy mô hình đạt kết quả khả quan qua các độ đo BLEU 1–4 [18]. Tuy nhiên, nghiên cứu này mới dừng ở mức xây dựng mô hình, chưa có ứng dụng triển khai thực tế.

Cùng thời điểm, Aditya Lumar Yadav và cộng sự đề xuất mô hình kết hợp R-CNN để phát hiện vùng ảnh và LSTM để tạo chú thích cho từng vùng [19]. Năm 2022, Smriti P. Manay cùng các cộng sự phát triển hệ thống dựa trên GRU và triển khai thành ứng dụng Android hỗ trợ người khiếm thị [20]. Mặc dù có tính thực tiễn, song việc sử dụng ứng dụng này vẫn đòi hỏi nhiều thao tác bằng lệnh, gây bất tiện cho người dùng. Cũng trong năm đó, Hiba Ahsan và cộng sự đưa ra mô hình chú thích ảnh đa phương thức, bổ sung thông tin văn bản xuất hiện trong ảnh vào quá trình sinh chú thích [21]. Mô hình được thử nghiệm trên tập VizWiz Captions, cho kết quả khả thi nhưng mới dừng ở mức mô hình thử nghiệm, chưa phát triển thành ứng dụng hoàn chỉnh.

Năm 2023, nghiên cứu của nhóm tác giả R. Kavitha tạo chú thích hình ảnh cho người khiếm thị dựa trên mạng học sâu [2]. Ảnh được chụp thông qua camera, sau đó sẽ được nhận dạng và phát sinh chú thích bởi mô hình học sâu, tiếp đó ứng dụng sẽ chuyển chú thích dạng văn bản thành âm thanh và trả về cho người sử dụng. Nghiên cứu thực nghiệm trên bộ dữ liệu ảnh MS-COCO gồm có ảnh và bộ chú thích đính kèm, trích xuất đặc trưng ảnh bằng EfficientNet-B3 làm đầu vào để huấn luyện mạng RNN. Sau quá trình huấn luyện, mô hình RNN được xây dựng có thể đưa ra chú thích cho một ảnh đầu vào mới, ứng dụng sử dụng chú dưới dạng âm thanh. Công trình này có hạn chế là mạng RNN dễ bị mất mát thông tin khi câu chú thích quá dài.

Tổng quan các nghiên cứu trên cho thấy chú thích ảnh bằng học sâu đã đạt được nhiều kết quả đáng tin cậy. Tuy nhiên, phần lớn công trình mới dừng ở mức mô hình hoặc chỉ phát triển ứng dụng di động, đồng thời nhiều nghiên cứu vẫn sử dụng bộ dữ liệu có quy mô hạn chế. Từ thực tế đó, nghiên cứu này đề xuất một mô hình chú thích ảnh theo khung mã hóa – giải mã, đồng thời phát triển ứng dụng đa nền tảng với giao diện thân thiện và thao tác đơn giản, nhằm hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh qua mô tả bằng âm thanh theo thời gian thực.

3. PHƯƠNG PHÁP ĐỀ XUẤT

Phần này trình bày mô hình chú thích ảnh tự động được xây dựng dựa trên DenseNet, LSTM và cơ chế chú ý. Trên cơ sở mô hình, một kiến trúc ứng dụng đa nền tảng cũng được đề xuất nhằm hỗ trợ người khiếm thị tiếp cận nội dung hình ảnh.

3.1. Mô hình tạo chú thích ảnh tự động

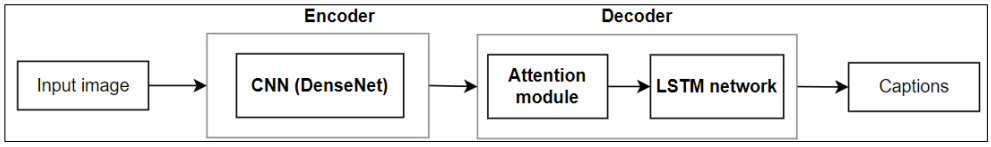
Bài toán được mô tả như sau: đầu vào là một ảnh I, đầu ra là một câu mô tả S. Câu S có thể xem như một chuỗi từ $\{w_t\}$, trong đó w_t là từ sinh ra ở bước thời gian t. Với tập huấn luyện gồm các cặp ảnh–chú thích, mô hình tham số θ được tối ưu bằng cách cực tiểu hóa hàm mất mát $cross\text{-}entropy$, biểu diễn ở công thức (1).

$$L(\theta) = -\sum_{t=1}^{N} \log P(w_t \mid I, w_{0,} w_1, ..., w_{t-1}; \theta) \quad (1)$$

Trong công thức (1), P là xác suất cho biết khả năng sinh ra từ tiếp theo w_t khi biết đặc trưng ảnh và các từ đã sinh ra ở thời điểm trước đó.

Trong bài báo này, kiến trúc đề xuất tuân theo khung mã hóa - giải mã (encoder - decoder) (Hình 1), gồm ba thành phần:

1. Bộ mã hóa ảnh (Image Encoder): sử dụng DenseNet để trích xuất đặc trưng hình ảnh.
2. Cơ chế chú ý (Attention): tính toán động trọng số cho các vùng quan trọng trong ảnh tại mỗi bước sinh từ.
3. Bộ giải mã ngôn ngữ (Language Decoder): LSTM kết hợp với thông tin ngữ cảnh để phát sinh chú thích.



Hình 1. Kiến trúc mô hình chú thích ảnh tự động, trong đó DenseNet được sử dụng như bộ mã hóa (encoder) để rút trích đặc trưng từ ảnh; các đặc trưng này sau đó được kết hợp với cơ chế chú ý (attention module) và đưa vào mạng LSTM ở giai đoạn giải mã (decoder) để tạo ra câu mô tả (caption) cho ảnh đầu vào.

3.1.1. Bộ mã hóa hình ảnh

Trong bài toán chú thích ảnh theo khung mã hóa - giải mã, việc trích xuất đặc trưng từ hình ảnh đóng vai trò quan trọng, vì đầu ra của nó sẽ trở thành đầu vào cho mô hình ngôn ngữ (decoder) nhằm sinh câu mô tả. DenseNet [1] là một kiến trúc mạng nơ-ron tích chập sâu, lần đầu tiên được giới thiệu bởi Gao Huang và cộng sự vào năm 2017. Điểm khác biệt của DenseNet nằm ở cơ chế kết nối dày đặc: trong một khối Dense, mỗi lớp không chỉ nhận dữ liệu từ lớp ngay trước đó mà còn từ tất cả các lớp trước đó. Nhờ vậy, mạng hình thành một cấu trúc liên kết dày đặc, trong đó đầu vào của mỗi lớp là sự tổng hợp của toàn bộ đặc trưng đã học được cho đến thời điểm hiện tại.

DenseNet đã chứng minh hiệu quả vượt trội trong nhiều tác vụ của thị giác máy tính, từ phân loại hình ảnh, nhận dạng đối tượng cho tới phân đoạn. Bên cạnh đó, kiến trúc này còn hạn chế việc tham số hóa dư thừa nhờ tái sử dụng đặc trưng, chỉ học thêm các thông tin cần thiết, đồng thời tăng cường khả năng chống hiện tượng mất gradient trong quá trình huấn luyện. Xuất phát từ những ưu điểm này, nghiên cứu sử dụng DenseNet để rút trích đặc trưng hình ảnh làm đầu vào cho mô hình ngôn ngữ, với mục tiêu cải thiện độ chính xác của câu chú thích.

Xét một ảnh đầu vào I, ký hiệu x_I^l là bản đồ đặc trưng đầu ra tầng thứ l của ảnh I, qua mạng DenseNet, kết quả x_I^l như ở công thức (2).

$$x_I^l = H_l([x_I^0, x_I^1, ..., x_I^{l-1}]) \quad (2)$$

Trong công thức (2), $[x_I^0, x_I^1, ..., x_I^{l-1}]$ đề cập đến sự kết hợp (*concatenation*) của các bản đồ đặc

trưng (*feature-maps*) của ảnh I được tạo ra ở các tầng $0,1,\dots l-1$, H_l là hàm tổng hợp bao gồm 3 hoạt động liên tiếp: chuẩn hóa batch (*Batch Normalization - BN),* kích hoạt phi tuyến ReLU và tích chập 3×3 (*Conv*).

3.1.2. Cơ chế chú ý

Cơ chế chú ý đóng vai trò xác định mức độ quan trọng của từng vùng ảnh tại mỗi bước giải mã. Thay vì coi tất cả các đặc trưng ảnh đều ngang nhau, mô hình sẽ tự động gán trọng số khác nhau cho các vùng, nhờ vậy tập trung nhiều hơn vào những chi tiết liên quan nhất đến từ cần sinh ra. Tổ hợp đặc trưng được gán trọng số này tạo thành một biểu diễn ngữ cảnh động (context vector), ký hiệu là $\widehat{c_t}$, và được cung cấp làm đầu vào cho bộ giải mã ở bước t. Trong nghiên cứu này, chúng tôi áp dụng cơ chế chú ý do Xu và cộng sự [6] đề xuất, được triển khai theo các bước sau:

– Tính điểm số liên kết (*alignment score*)

$$e_{t,i} = f_{att}(x_i, h_{t-1}) \tag{3}$$

Trong công thức (3), x_i là đặc trưng của vùng ảnh thứ i, và h_{t-1} là trạng thái ẩn của bộ giải mã tại thời điểm trước đó. Hàm f_{att} được tham số hóa bằng một phép biến đổi tuyến tính, nhằm ước lượng mức độ liên quan giữa vùng ảnh và từ cần dự đoán.

– Chuẩn hóa để thu được trọng số chú ý:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}\exp(e_{t,k})} \quad , \sum_i \alpha_{t,i} = 1;\ 0 < \alpha_{t,i} < 1 \tag{4}$$

Các giá trị $\alpha_{t,i}$ thể hiện xác suất phân bố chú ý trên toàn bộ các vùng ảnh.

– Tính vectơ ngữ cảnh.

$$c_t = \sum_{i=1} x_i \alpha_{t,i} \tag{5}$$

Vectơ ngữ cảnh $\widehat{c_t}$ tổng hợp thông tin từ các vùng quan trọng nhất, và sẽ được sử dụng song song với embedding từ và trạng thái ẩn trước đó trong quá trình sinh chú thích bằng LSTM.

3.1.3. Bộ giải mã ngôn ngữ

Để sinh ra chú thích từ chuỗi đặc trưng ảnh, nghiên cứu này sử dụng mạng LSTM [24] thay cho RNN truyền thống [25]. LSTM được lựa chọn vì có khả năng xử lý các quan hệ phụ thuộc dài nhờ cơ chế cổng, qua đó hạn chế hiện tượng triệt tiêu gradient thường gặp trong RNN.

Tại thời điểm t, đầu vào của LSTM bao gồm embedding từ hiện tại x_t, trạng thái ẩn của bước trước h_{t-1}, cùng với vectơ ngữ cảnh $\widehat{c_t}$ thu được từ cơ chế chú ý. Các cổng và bộ nhớ của LSTM được cập nhật theo công thức (6):

$$\begin{cases} i_t = \delta(W_{xi}x_t + W_{hi}h_{t-1} + W_{ic}\widehat{c_t} + b_i) \\ f_t = \delta(W_{xf}x_t + W_{hf}h_{t-1} + W_{fc}\widehat{c_t} + b_f) \\ o_t = \delta(W_{xo}x_t + W_{ho}h_{t-1} + W_{oc}\widehat{c_t} + b_o) \\ \tilde{C}_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}\widehat{c_t} + b_c) \\ C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ h_t = o_t \odot tanh(C_t) \end{cases} \tag{6}$$

Trong công thức (6), δ là hàm $sigmoid$, $\odot$ biểu diễn phép nhân từng phần tử, còn W và b là các tham số học được tối ưu trong quá trình huấn luyện.

Sau khi cập nhật trạng thái ẩn h_t, mô hình sinh từ tiếp theo dựa trên phân phối xác suất p_t trên từ hiện tại y_t với hàm $softmax$ như công thức (7).

$$y_t \sim p_t = soft\max(W_p h_t + b_p) \tag{7}$$

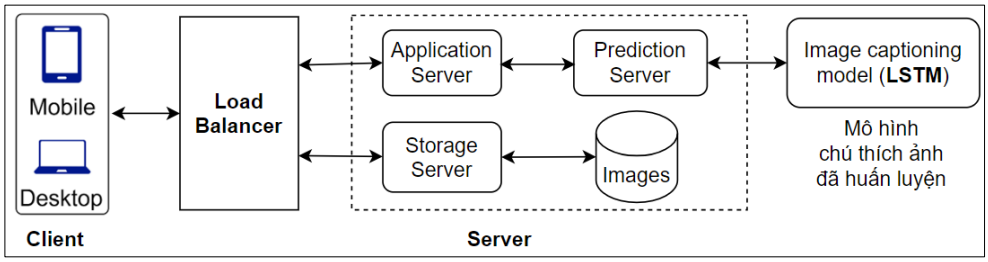
Quá trình tối ưu được thực hiện bằng thuật toán lan truyền ngược theo thời gian (Backpropagation Through Time – BPTT) [26]. Trạng thái ẩn h_t kết hợp với ngữ cảnh ảnh và thông tin từ chuỗi trước đó

giúp mô hình tạo ra chú thích nhất quán và phù hợp về mặt ngữ nghĩa.

3.2. Xây dựng ứng dụng hỗ trợ người khiếm thị chú thích ảnh

Dựa trên mô hình chú thích ảnh đã đề xuất, nhóm nghiên cứu phát triển một ứng dụng đa nền tảng nhằm hỗ trợ người khiếm thị tiếp cận nội dung hình ảnh. Ứng dụng này gồm hai phần chính: (1) phiên bản Desktop cho phép người dùng nghe lại chú thích của những hình ảnh có sẵn; (2) phiên bản di động (mobile) cho phép chụp ảnh trực tiếp bằng camera và ngay lập tức phát nội dung mô tả, giúp người khiếm thị nhận biết được khung cảnh xung quanh. Kiến trúc tổng thể của ứng dụng được minh họa trong Hình 2, bao gồm ba thành phần.

- **Client**: là ứng dụng phía người dùng, gồm 2 loại ứng dụng trên nền tảng Desktop và Mobile;
- **Load Balancer**: bộ phận cân bằng tải, thực hiện điều phối và định tuyến yêu cầu của người dùng (*client*) đến đúng máy chủ (*server*) phù hợp;
- **Server**: gồm các máy chủ thực hiện các chức năng tương ứng, lần lượt là: application server – thực hiện các chức năng nghiệp vụ của ứng dụng, trong đó, Google Translate API được sử dụng để dịch câu chú thích từ tiếng Anh sang tiếng Việt, text to speech của FPT AI được sử dụng để tạo file âm thanh từ câu chú thích; storage server – thực hiện lưu trữ hình ảnh, câu chú thích, cùng với file ghi âm của câu chú thích; prediction server – sử dụng mô hình chú thích ảnh đã huấn luyện ở phần trên để dự đoán chú thích cho ảnh đầu vào.



Hình 2. Kiến trúc ứng dụng hỗ trợ người khiếm thị chú thích hình ảnh

4. THỰC NGHIỆM VÀ KẾT QUẢ

Dựa trên mô hình đã trình bày, phần này mô tả chi tiết quá trình cài đặt, đồng thời báo cáo kết quả đánh giá trên các bộ dữ liệu chuẩn. Ứng dụng đa nền tảng hỗ trợ người khiếm thị cũng được kiểm chứng hiệu quả thông qua các tình huống thực tế.

4.1. Dữ liệu và thiết lập thực nghiệm

Qua khảo sát các công trình gần đây, có thể thấy dữ liệu chú thích ảnh tiếng Việt chưa phổ biến. Do đó, nghiên cứu này tiến hành huấn luyện và đánh giá trên hai bộ dữ liệu chuẩn là MS COCO [28] và Flickr30K.

Khảo sát các công trình trước đây cho thấy nguồn dữ liệu chú thích ảnh bằng tiếng Việt còn khá hạn chế. Vì vậy, nghiên cứu này tiến hành huấn luyện và đánh giá mô hình trên hai bộ dữ liệu phổ biến là MS COCO [28] và Flickr30K.

- MS COCO: tập dữ liệu này chứa 82.783 ảnh cho huấn luyện và 40.504 ảnh cho kiểm, mỗi ảnh có 5 câu mô tả do con người tạo thủ công. Theo cách chia dữ liệu của Karpathy và Li [28], bộ này được tách thành 82.783 ảnh dùng để huấn luyện, 5.000 ảnh cho xác thực và 5.000 ảnh cho kiểm thử. Sau bước tiền xử lý, từ điển còn 10,010 từ (loại bỏ các từ xuất hiện dưới 5 lần), và chiều dài tối đa của câu chú thích đặt là 16.
- Flickr30K: tập này có 31,783 ảnh, mỗi ảnh kèm 5 chú thích. Dữ liệu được chia thành 29,000 ảnh huấn luyện, 1,000 ảnh kiểm định và 1,000 ảnh kiểm tra theo chuẩn của [28].

Mô hình được triển khai bằng Python 3.9 và PyTorch 2.0 cho quá trình huấn luyện, kết hợp với C#/.NET 6, Xamarin và Qt để phát triển ứng dụng. Hệ thống chạy trên Google Colab, sử dụng máy chủ

ảo Droppet làm backend. Bộ mã hóa ảnh là DenseNet121 đã huấn luyện sẵn trên ImageNet, trích xuất đặc trưng từ tầng fc7 với vector 1,024 chiều. Đối với mô hình ngôn ngữ, LSTM có kích thước ẩn và embedding đều bằng 512, độ dài câu tối đa là 16. Huấn luyện sử dụng hàm mất mát cross-entropy, tối ưu bằng Adam [5] với learning rate 0,0001, batch size 16.

4.2. Độ đo đánh giá

Để đo lường chất lượng mô tả ảnh, nghiên cứu này áp dụng các độ đo phổ biến gồm BLEU [30], METEOR [31]. Các chỉ số này so sánh câu sinh ra với tập chú thích tham chiếu, phản ánh độ tương đồng ở nhiều mức độ khác nhau. Điểm số cao hơn thể hiện khả năng sinh chú thích chính xác và tự nhiên hơn.

4.3. Kết quả thực nghiệm

Kết quả trên tập MS COCO được thể hiện ở Bảng 1, cho thấy phương pháp đề xuất đạt BLEU-1 (B@1) đến BLEU-4 (B@4) lần lượt là 72,1, 51,2, 37,5, 26,4 và điểm METEOR (M) đạt 24,1. Trên tập Flickr30K, kết quả trình bày ở Bảng 2 với BLEU-1 (B@1) đến BLEU-4 (B@4) lần lượt là 67,6, 47,8, 33,7, 24,3 và METEOR (M) đạt 22,9.

Để minh chứng cho hiệu quả của mô hình, các kết quả này được so sánh với một số công trình tiêu biểu trên cùng bộ dữ liệu (Bảng 3). Trong đó, nhóm baseline gồm NIC, Hard-ATT, Soft-ATT và một phương pháp gần đây là En-De-Cap-2021. Có thể thấy, phương pháp đề xuất đạt kết quả tốt hơn ở hầu hết các độ đo. Sự vượt trội này có được nhờ DenseNet trích xuất đặc trưng phong phú và cơ chế chú ý tính toán động giúp tập trung vào các vùng ảnh quan trọng.

Ngoài ra, Hình 4 minh họa một số kết quả trên tập kiểm tra, trong khi Hình 5 và Hình 6 thể hiện giao diện ứng dụng trên Desktop và thiết bị di động. Các ví dụ cho thấy câu chú thích phản ánh chính xác đối tượng chính và mối quan hệ trong ảnh. Hệ thống còn tích hợp Google Translate API để dịch chú thích từ tiếng Anh sang tiếng Việt, mang lại tiện ích trực tiếp cho người khiếm thị trong quá trình sử dụng.

Bảng 1. Độ chính xác của phương pháp đề xuất trên tập dữ liệu MS COCO

Số lượng ảnh	Huấn luyện	Kiểm định	Kiểm tra	B@1	B@2	B@3	B@4	M
82.783	72.783	5.000	5.000	72,1	51,2	37,5	26,4	24,1

Bảng 2. Độ chính xác của phương pháp đề xuất trên tập dữ liệu Flickr30K

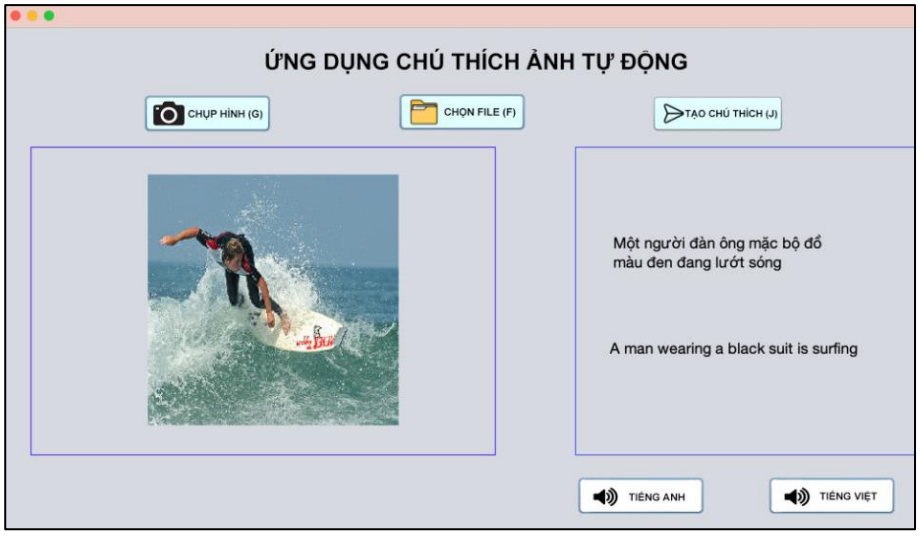
Số lượng ảnh	Huấn luyện	Kiểm định	Số lượng ảnh	B@1	B@2	B@3	B@4	M
31.783	29.000	1.000	1.000	67,6	47,8	33,7	24,3	22,9

Bảng 3. So sánh độ chính xác của phương pháp đề xuất với các phương pháp trên tập dữ liệu MS COCO

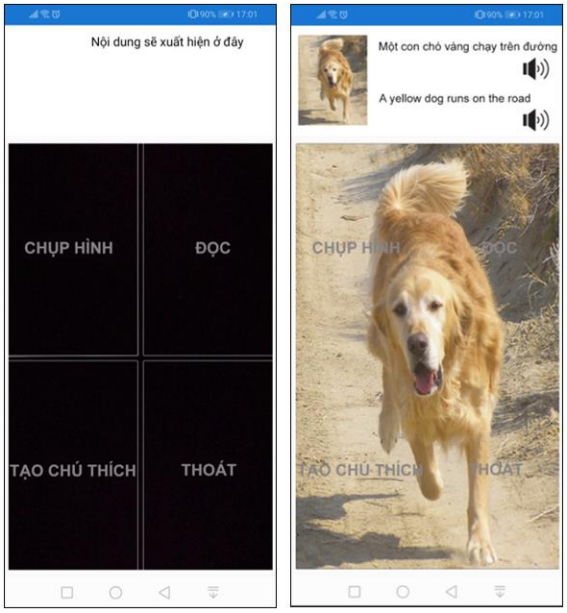
Phương pháp	B@1	B@2	B@3	B@4	M
Show and tell (NIC)-2015 [1]	66,6	46,1	32,9	24,6	-
Show, attend and tell (Hard-ATT)-2015 [2]	71,8	50,4	35,7	25,0	23,0
Show, attend and tell (Soft-ATT)-2015 [2]	70,7	49,2	34,4	24,3	23,9
En-De-Cap-2021 [3]	70,6	41,1	36,7	24,3	-
Đề xuất của bài báo	**72,1**	**51,2**	**37,5**	**26,4**	**24,1**

	**Caption**: a group of people standing around a tennis court *Ground truth (GT) captions*: GT1: The people are posing for a group photo. GT 2: a large family poses for picture on tennis court GT 3: A group of young children standing next to each other. GT 4: A group of people that are standing near a tennis net. GT 5: A group of kids posing for a picture on a tennis court.
	**Caption**: a piece of chocolate cake on a plate *Ground truth (GT) captions*: GT1: A piece of chocolate cake on a plate. GT2: a yellow and white plate with a piece of chocolate cake GT3: a piece of a chocolate cake on a plate GT4: A piece of chocolate pie sitting on top of a plate. GT5: A slice of chocolate cake is on a small plate.
	**Caption**: two giraffes standing next to a fence *Ground truth (GT) captions*: GT1: Giraffes in their wood and grass zoo enclosure GT2: A large giraffe looks at a smaller giraffe looking over a fence. GT3: Two giraffe stand in an enclosure, on giraffe leans over the wall. GT4: Pair of giraffes exploring a large, airy zoo enclosure. GT5: Two giraffes in a zoo enjoy a walk and a snack

Hình 4. Một số kết quả minh họa từ tập ảnh kiểm tra của phương pháp chú thích ảnh đề xuất



Hình 5. Giao diện và ví dụ về kết quả tạo chú thích ảnh trên ứng dụng Desktop



Hình 6. Giao diện và kết quả tạo chú thích ảnh trên thiết bị di động

5. KẾT LUẬN

Nghiên cứu này tập trung xây dựng một mô hình chú thích ảnh với mục tiêu hỗ trợ người khiếm thị nhận biết và hiểu được môi trường xung quanh. Phương pháp được đề xuất sử dụng DenseNet để rút trích đặc trưng hình ảnh, sau đó kết hợp với mạng LSTM có tích hợp cơ chế chú ý trong kiến trúc mã hóa–giải mã. Mô hình được huấn luyện và kiểm chứng trên hai bộ dữ liệu chuẩn là MS COCO và Flickr30K. Kết quả thực nghiệm chứng minh rằng phương pháp mang lại mức cải thiện đáng kể về độ chính xác so với các mô hình tham chiếu và một số nghiên cứu gần đây.

Dựa trên mô hình này, nhóm nghiên cứu cũng phát triển một ứng dụng đa nền tảng có khả năng mô tả hình ảnh theo thời gian thực từ camera hoặc xử lý ảnh có sẵn. Ứng dụng cung cấp giao diện thân thiện, đồng thời hỗ trợ chuyển đổi văn bản thành giọng nói, tạo điều kiện để người khiếm thị tiếp nhận thông tin dễ dàng hơn. Tuy nhiên, nghiên cứu vẫn tồn tại một số hạn chế, đặc biệt là sự phụ thuộc vào công cụ của bên thứ ba, cùng với những giới hạn về phần cứng khiến tốc độ xử lý chưa đạt mức tối ưu.

TÀI LIỆU THAM KHẢO

1. https://www.clc.hcmus.edu.vn/?page_id=1567
2. Kavitha R., Shree Sandhya S., Praveena B., Rajalakshmi P., Sarubala E. - Deep learning-based image captioning for visually impaired people. International Conference on Newer Engineering Concepts and Technology (2023). https://doi.org/10.1051/e3sconf/202339904005
3. Ming Y., Hu N., Fan C., Feng F., Zhou J., Yu H. - Visuals to text: A comprehensive review on automatic image captioning. IEEE/CAA Journal of Automatica Sinica **9** (8) (2022) 1339-1365. https://doi.org/10.1109/JAS.2022.105734
4. Pavlopoulos J., Kougia V., Androutsopoulos I. - A survey on biomedical image captioning, in Proceedings of the second workshop on shortcomings in vision and language (2019) 26–36. https://doi.org/10.18653/v1/W19-1803
5. Lee D.I., Lee J.H., Jang S.H., Oh S.J., Doo I.C. - Crop disease diagnosis with deep learning-based image captioning and object detection. Applied Sciences **13** (5) (2023) 3148. https://doi.org/10.3390/app13053148

6. Stefanini M., Cornia M., Baraldi L., Cascianelli S., Fiameni G., Cucchiara R. - From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence **45** (1) (2022) 539-559. https://doi.org/10.1109/TPAMI.2022.3148210
7. Dehaqi A.M., Seydi V., Madadi Y. - Adversarial image caption generator network. SN Computer Science **2** (2021) 1-14. https://doi.org/10.1007/s42979-021-00486-y
8. Han S.H., Choi H.J. - Domain-specific image caption generator with semantic ontology. In 2020 IEEE International Conference on Big Data and Smart Computing (2020) 526-530. https://doi.org/10.1109/BigComp48618.2020.00-12
9. Ke X., Zou J., Niu Y. - End-to-end automatic image annotation based on deep CNN and multi-label data augmentation. IEEE Transactions on Multimedia **21** (8) (2019) 2093-2106. https://doi.org/10.1109/TMM.2019.2895511
10. Ghandi T., Pourreza H., Mahyar H. - Deep learning approaches on image captioning: A Review. Computing Surveys **56.3** (2023) 1-39. https://doi.org/10.48550/arXiv.2201.12944
11. Zohourianshahzadi Z., Kalita J.K. - Neural attention for image captioning: Review of outstanding methods. Artificial Intelligence Review **55** (5) (2022) 3833-3862. https://doi.org/10.1007/s10462-021-10092-2
12. He S., Liao W., Tavakoli H.R., Yang M., Rosenhahn B., Pugeault N. - Image captioning through image transformer. in Proceedings of the Asian conference on computer vision (2020). https://doi.org/10.48550/arXiv.2004.14231
13. Bai C., Zheng A., Huang Y., Pan X., Chen N. - Boosting convolutional image captioning with semantic content and visual relationship. Displays **70** (2021) 102069. https://doi.org/10.1016/j.displa.2021.102069
14. Ding S., Qu S., Xi Y., Sangaiah A.K., Wan S. - Image caption generation with high level image features. Pattern Recognition Letters **123** (2019) 89-95. https://doi.org/10.1016/j.patrec.2019.03.021
15. Han M., Chen W., Moges A.D. - Fast image captioning using LSTM. Cluster Computing **22** (2019) 6143-6155. https://doi.org/10.1007/s10586-018-1885-9
16. Deng Z., Jiang Z., Lan R., Huang W., Luo X. - Image captioning using DenseNet network and adaptive attention. Signal Processing: Image Communication **85** (2020) 115836. https://doi.org/10.1016/j.image.2020.115836
17. Alzubi J.A., Jain R., Nagrath P., Satapathy S., Taneja S, Gupta P. - Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. Journal of Intelligent & Fuzzy Systems **40** (4) (2021) 5761-5769. https://doi.org/10.3233/JIFS-189415
18. Patwari, N., Naik D. - En-de-cap: An encoder decoder model for image captioning. International Conference on Computing Methodologies and Communication (2021). https://doi.org/10.1109/ICCMC51019.2021.9418414
19. Yadav A.K., Prakash.J. - Image captioning using R-CNN & LSTM deep learning model. International Journal of Innovative Science and Research Technology **5** (2021) 911-914
20. Manay S.P., Yaligar S.A., Thathva Sri Sai Reddy Y., Saunshimath N.J. - Image captioning for the visually impaired. Emerging Research in Computing, Information, Communication and Applications. Lecture Notes in Electrical Engineering **789** (2022). https://doi.org/10.1007/978-981-16-1338-8_43
21. Ahsan H., Bhalla N., Bhatt D., Shah K. - Multi-modal image captioning for the visually impaired. NAACL-HLT SRW (2021). https://doi.org/10.48550/arXiv.2105.08106
22. Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q. - Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition (2017) 2261-2269. https://doi.org/10.1109/CVPR.2017.243.
23. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R., Bengio Y. - Show, attend and tell: Neural image caption generation with visual attention. in International Conference on Machine Learning (2015). https://doi.org/10.48550/arXiv.1502.03044

24. Hochreiter S. Schmidhuber J. - Long short-term memory. Neural Computation **9** (1997) 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
25. Rumelhart D.E., Hinton G.E., Williams R.J. - Learning internal representations by error propagation. Institute for Cognitive Science, University of California, San Diego La. (1985)
26. Werbos P.J. - Backpropagation through time: what it does and how to do it. Proceedings of the IEEE **78** (10) (1990) 1550-1560. https://doi.org/10.1109/5.58337
27. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár & C. Lawrence Zitnick - Microsoft COCO: Common objects in context. in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. 2014. Springer. https://doi.org/10.1007/978-3-319-10602-1_48
28. Karpathy A., Fei-Fei L. - Deep visual-semantic alignments for generating image descriptions. in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. https://doi.org/10.48550/arXiv.1412.2306
29. Kingma D.P., Ba J. - Adam: A method for stochastic optimization. International Conference for Learning Representations, San Diego (2015). https://doi.org/10.48550/arXiv.1412.6980
30. Papineni K., Roukos S., Ward T., Zhu W. - Bleu: a method for automatic evaluation of machine translation. in Proceedings of the 40th annual meeting of the Association for Computational Linguistics (2002). https://doi.org/10.3115/1073083.1073135
31. Banerjee S., Lavie A.. - 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) 65-72.
32. Vinyals O., Toshev A., Bengio S., Erhan D. - Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (2015) 3156-3164. https://doi.org/10.48550/arXiv.1411.4555

ABSTRACT

AUTOMATIC IMAGE CAPTIONING SYSTEM FOR VISUALLY IMPAIRED PEOPLE

Dinh Thi Man[1], Nguyen Van Thinh[2], Tran Huu Quoc Thu[2],
Nguyen Hai Yen[1], Nguyen Phuong Hac[1], Tran Thi Van Anh[1*]

[1]*Ho Chi Minh City University of Industry and Trade*
[2]*Ho Chi Minh City University of Education*
*Email: *anhttv@huit.edu.vn*

Visual impairment poses significant challenges for visually impaired individuals in recognising and interacting with their surrounding environment. To address this issue, this study proposes a cross-platform automatic image captioning system. The model follows an encoder–decoder architecture, where DenseNet is used to extract visual features, while an LSTM network, combined with an attention mechanism, generates natural language descriptions. The proposed method is trained and evaluated on two benchmark datasets, MS COCO and Flickr30K, using widely adopted metrics such as BLEU and METEOR. Experimental results demonstrate that the system achieves higher accuracy compared to several recently published approaches. Furthermore, a practical application has been developed for both desktop and mobile platforms, enabling the production of audio descriptions for images, thereby enhancing accessibility to visual information for visually impaired users.

Keywords: Automatic Image Captioning, CNN, LSTM, attention mechanism, Visually Impaired.