TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 22, Số 2 (2025): 247-259
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 22, No. 2 (2025): 247-259
https://doi.org/10.54607/hcmue.js.22.2.4328(2025)

Bài báo nghiên cứu

MÔ HÌNH CHÚ Ý NGỮ CẢNH ĐA TẦM NHÌN CẢI TIẾN CHO BÀI TOÁN TRẢ LỜI CÂU HỎI DỰA TRÊN HÌNH ẢNH BẰNG TIẾNG VIỆT

Bùi Anh Đài*, Nguyễn Quốc Trung, Trần Thanh Nhã, Nguyễn Viết Hưng

Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam

**Tác giả liên hệ: Bùi Anh Đài – Email: buianhdai1412@gmail.com*

Ngày nhận bài: 11-6-2024; ngày nhận bài sửa: 25-11-2024; ngày duyệt đăng: 19-12-2024

TÓM TẮT

Bài toán trả lời câu hỏi dựa trên hình ảnh là một bài toán tiêu biểu cho sự giao thoa giữa hai lĩnh vực thị giác máy tính (Computer Vision) và xử lí ngôn ngữ tự nhiên (Natural Language Processing). Bài toán này không chỉ có giá trị khoa học mà còn có giá trị to lớn trong thực tiễn cuộc sống. Việc tích hợp mô hình VQA vào các thiết bị di động có thể hỗ trợ người mù và người khiếm thị trong việc tiếp cận và hiểu nội dung hình ảnh. Phương pháp tiếp cận phổ biến hiện nay là rút trích đặc trưng từ từng vùng trong hình ảnh, giúp mô hình nắm bắt bối cảnh cục bộ. Tuy nhiên, phương pháp này thường bỏ qua bối cảnh toàn cục, ảnh hưởng đến khả năng tổng hợp thông tin và suy luận của mô hình. Các phương pháp hiện nay sử dụng Vision Transformer để rút trích đặc trưng toàn cục và cục bộ từ hình ảnh giúp cải thiện hiệu suất mô hình. Thêm vào đó, cơ chế chú ý đa phương thức (multimodal attention) cũng được áp dụng nhằm tối ưu hóa quá trình kết hợp thông tin giữa hình ảnh và câu hỏi, giúp mô hình có khả năng hiểu được ngữ cảnh và chú ý vào các đặc trưng quan trọng. Hiện nay, nhiều mô hình VQA được tối ưu cho dữ liệu tiếng Anh và một số mô hình được tối ưu cho ngôn ngữ tiếng Việt (ViVQA) đã được công bố. Bài báo này đề xuất một mô hình cải tiến từ mô hình Multi-vision Contextual Attention và đạt được độ chính xác là 62,41% so với mô hình gốc là 60% trên tập dữ liệu ViVQA.

Từ khóa: đa phương thức; ngôn ngữ tiếng Việt; ngôn ngữ tự nhiên; PhoBERT; ResNet; Swin Transformer; trả lời câu hỏi qua hình ảnh

1. Giới thiệu

Trong thập kỉ vừa qua, lĩnh vực Thị giác máy tính (Computer Vision - CV) và Xử lí ngôn ngữ tự nhiên (Natural Language Processing - NLP) đã đạt được những bước tiến vượt bậc, đặc biệt là với sự xuất hiện của cơ chế chú ý (Bahdanau et al., 2014). Cơ chế chú ý cùng với các mạng nơ-ron truyền thống như Kiến trúc Mạng Nơ-ron Tích chập (Convolutional Neural Network - CNN) (LeCun et al., 1989) trong lĩnh vực CV và mô hình Bộ nhớ Ngắn hạn Dài (Long Short-Term Memory - LSTM) (Hochreiter & Schmidhuber, 1997) trong NLP,

Cite this article as: Bui Anh Dai, Nguyen Quoc Trung, Tran Thanh Nha, & Nguyen Viet Hung (2025). An improved multi-vision contextual attention model for Vietnamese visual-based question answering. *Ho Chi Minh City University of Education Journal of Science, 22*(2), 247-259.

đã cải thiện đáng kể hiệu suất xử lí trong nhiều nhiệm vụ quan trọng. Các nhiệm vụ này bao gồm nhận dạng khuôn mặt (Lagorio et al., 2013), nhận dạng biển số xe xác định sản phẩm qua ảnh (Jallouli et al., 2016), phân loại văn bản dịch thuật tự động (Bar-Hillel, 1960) và phát hiện Đối tượng và Dịch máy (Vaswani et al., 2017). Những tiến bộ này không chỉ giúp giải quyết các thách thức cơ bản mà còn mở rộng khả năng của các nhà nghiên cứu trong việc giải quyết các vấn đề phức tạp đòi hỏi sự kết hợp sâu sắc giữa thị giác và ngôn ngữ.

Bài toán hỏi đáp hình ảnh (Visual Question Answering- VQA) đòi hỏi hệ thống phải hiểu và trả lời các câu hỏi mở về nội dung của một hình ảnh. Đầu vào của hệ thống là một cặp hình ảnh và câu hỏi liên quan đến hình ảnh đó, còn đầu ra là một câu trả lời chính xác và phù hợp. Bài toán có thể được phát biểu thông qua thuật toán:

Đặt:

I: Hình ảnh (đầu vào của mô hình) (1)

Q: Câu hỏi (đầu vào của mô hình)

Θ: Tập hợp các câu trả lời tiềm năng

A: Câu trả lời được mô hình đưa ra

Input:

Cặp câu hỏi và hình ảnh: (Q,I)

Output:

Mô hình sẽ đưa ra dự đoán A với $A \in \Theta$

Bài toán Visual Question Answering (VQA) đại diện cho một lĩnh vực sáng giá và nhiều thách thức trong trí tuệ nhân tạo (AI), nơi sự kết hợp giữa thị giác máy tính (CV) và xử lí ngôn ngữ tự nhiên (NLP) được triển khai để phát triển các hệ thống AI có khả năng trả lời các câu hỏi dựa trên nội dung của hình ảnh. Mục tiêu của bài toán này là tạo ra một mô hình AI hiểu được và tổng hợp thông tin từ hai nguồn dữ liệu đa dạng: hình ảnh và ngôn ngữ, để đưa ra câu trả lời chính xác và phù hợp. Bài toán VQA không chỉ có khả năng cải tiến và tạo ra các hệ thống thông minh giúp tương tác tốt hơn với người dùng mà còn có ứng dụng rộng rãi trong các ngành như y tế, giáo dục và tự động hóa. Qua đó, VQA giúp tối ưu hóa các quy trình và nâng cao hiệu quả công việc. Sự phát triển của các phương pháp và mô hình mới trong VQA (Yu et al., 2019) không chỉ làm phong phú thêm kho tàng kiến thức của môi trường nghiên cứu mà còn mở ra cơ hội áp dụng thực tiễn trong nhiều bối cảnh khác nhau.

Đa số các mô hình VQA trên tiếng Việt được xây dựng gồm ba thành phần chính. Thành phần thứ nhất là hiểu hình ảnh, mô hình được áp dụng các kĩ thuật tiên tiến để rút trích đặc trưng thị giác từ hình ảnh. Điều này bao gồm việc sử dụng kiến trúc mạng nơ-ron tích chập như CNN với các mô hình phổ biến như Xception (Chollet, 2017), Efficientnet (Tan & Le, 2019), VGGNet (Simonyan & Zisserman, 2014) để phân tích và hiểu các chi tiết về ngữ cảnh của hình ảnh được truy vấn. Thành phần thứ hai là hiểu câu hỏi, trong đó các mô hình sử dụng các kiến trúc NLP như PhoBERT (Nguyen & Nguyen, 2020) để xử lí và rút trích ý nghĩa từ câu hỏi. Việc hiểu câu hỏi này cho phép mô hình nắm bắt được bối cảnh và nội dung của câu hỏi, từ đó liên kết chính xác hơn với các đặc trưng được rút ra từ hình

ảnh. Cuối cùng, sự kết hợp các đặc trưng câu hỏi và hình ảnh thành một đại diện đặc trưng chung là bước quan trọng cuối cùng, nơi các thông tin được tổng hợp để dự đoán câu trả lời.

Mô hình VQA cơ sở mà chúng tôi chọn để cải tiến là mô hình chú ý ngữ cảnh đa tầm nhìn (multi-vision contextual attention model) được đề xuất bởi (Nguyen et al., 2022). Mô hình này sử dụng đặc trưng toàn cục và cục bộ từ hình ảnh sau đó kết hợp với đặc trưng ngữ nghĩa từ câu hỏi bằng cơ chế chú ý có hướng dẫn (guide attention). Chúng tôi đề xuất bổ sung 2 khối hợp nhất đa mô hình (Multimodel Fusion Module) và khối đa tự chú ý (Multiple Self – Attention Module). Bên cạnh đó, mô hình cải tiến mà chúng tôi đề xuất được huấn luyện trên hai hàm mất mát là: Cross Entropy Loss và Normalized Temperature-scaled Cross Entropy Loss. Mô hình đề xuất được thực hiện trên tập ViVQA và so sánh với mô hình cơ sở và các phương pháp hiện có.

Tran et al. (2021) đã đề xuất một hệ thống sử dụng Mô hình Hierarchical Co-Attention để xác định câu trả lời cho mỗi câu hỏi dựa trên nội dung hình ảnh. Co-Attention là cơ chế chú ý lẫn nhau giữa hai luồng thông tin khác loại, trong trường hợp này là hình ảnh và ngôn ngữ. Mô hình Hierarchical Co-Attention khai thác thông tin từ các điểm hình ảnh và các từ trong câu hỏi để xác định những phần quan trọng cần tập trung, từ đó cải thiện khả năng trả lời câu hỏi. Hệ thống được thử nghiệm trên bộ dữ liệu ViVQA và đạt được Accuracy là 34,96%, WUPS 0.9 là 45,13%.

Tran et al. (2022) đã xây dựng mô hình Bidirectional Cross-Attention. Mô hình này tận dụng sức mạnh của các mô hình đã được tiền huấn luyện (pre-trained models) để tối ưu hóa việc trích xuất đặc trưng từ hình ảnh và văn bản. Cụ thể, đặc trưng hình ảnh được trích xuất bằng cách sử dụng mô hình Vision Transformer tiền huấn luyện, và đặc trưng câu hỏi thì sử dụng mô hình PhoBERT tiền huấn luyện dành riêng cho tiếng Việt. Sau đó, cấu trúc Bi-directional Cross-Attention được áp dụng để học các mối quan hệ giữa đặc trưng hình ảnh và văn bản, sử dụng đặc trưng đã học đó để phân loại câu trả lời. Mô hình đạt được kết quả accuracy là 51,3% trên tập dữ liệu ViVQA.

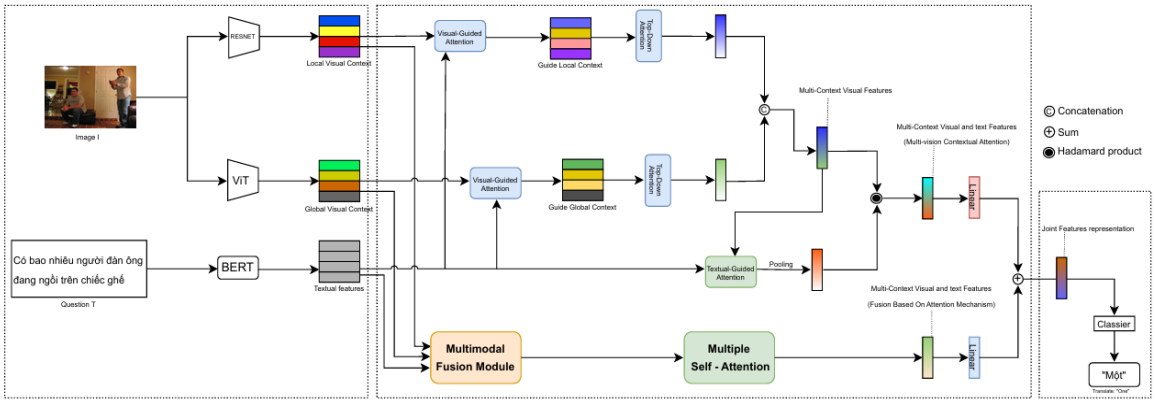
Antol et al. (2015) đã đề xuất bài toán trả lời câu hỏi hình ảnh vào năm 2015 trong nghiên cứu VQA. Đây là nền tảng khởi đầu cho hệ thống VQA với sự kết hợp các lĩnh vực quan trọng là Thị giác máy tính - Computer Vision (CV) cùng xử lí ngôn ngữ tự nhiên (NLP), kết hợp với bộ dữ liệu bao gồm 614,163 câu hỏi và 7.984.199 câu trả lời cho 204,721 hình ảnh từ bộ ảnh Microsoft COCO. Độ chính xác của mô hình tốt nhất (LSTM Q+I được chọn dựa trên độ chính xác của VQA test-dev) trên VQA test-standard là 54,06%.

2. Mô hình đề xuất

2.1. Mô hình cơ sở: mô hình chú ý ngữ cảnh đa tầm nhìn (multi-vision contextual attention model)

Mô hình chú ý ngữ cảnh đa tầm nhìn (multi-vision contextual attention model) là một phương pháp mới được Nguyen et al. (2022) đề xuất để giải quyết bài toán VQA cho tiếng Việt (gọi tắt là ViVQA), kiến trúc mô hình được minh họa qua Hình 1. Mô hình kết hợp hai phương pháp trích xuất đặc trưng hình ảnh: sử dụng ResNet để nắm bắt thông tin ngữ cảnh

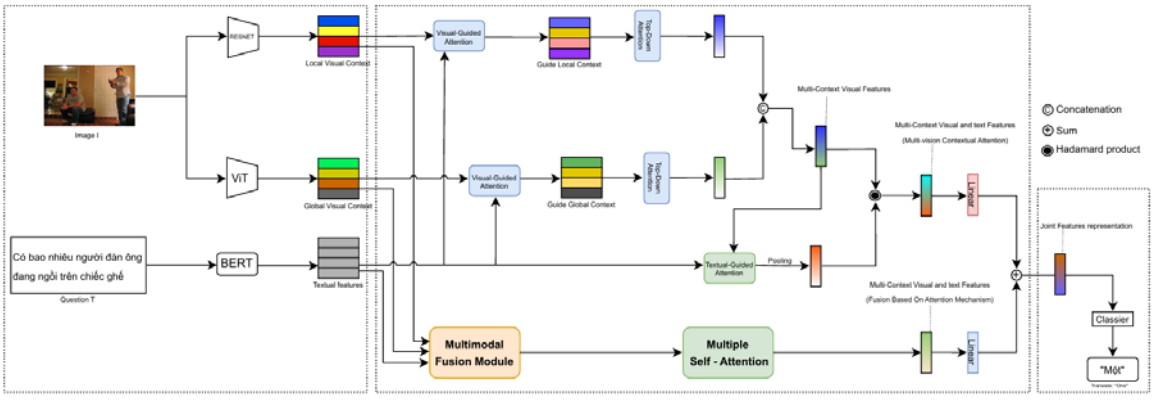
cục bộ (local) và Vision Transformer (ViT) để nắm bắt thông tin ngữ cảnh toàn cục (global) của hình ảnh. Đối với xử lí câu hỏi đầu vào, mô hình sử dụng PhoBERT- một biến thể của BERT được huấn luyện trên dữ liệu tiếng Việt. Đặc biệt, mô hình đề xuất một cơ chế chú ý đa nhánh để tích hợp thông tin từ cả hình ảnh và câu hỏi một cách hiệu quả. Kết quả thực nghiệm trên bộ dữ liệu ViVQA cho thấy mô hình đạt độ chính xác 60,76%, vượt trội so với các phương pháp cơ sở trước đó cho bài toán VQA tiếng Việt.



Hình 1. *Minh họa cho mô hình cơ sở dùng để cải tiến (mô hình* Multi-vision Contextual Attention*)*

2.2. Mô hình cải tiến dựa trên mô hình cơ sở

Cụ thể, chúng tôi giới thiệu một nhánh kết hợp mới mang tên Fusion Based on Attention Mechanism, trong đó tận dụng khả năng kết hợp và khai thác thông tin của hai khối chính: Multimodal Fusion Module và Multiple Self-Attention. Trong mô hình cải tiến, đặc trưng của nhánh Cơ chế chú ý dựa trên sự hợp nhất (Fusion Based on Attention Mechanism) và đặc trưng của nhánh Chú ý theo ngữ cảnh đa tầm nhìn (Multi-vision Contextual Attention) được kết hợp với nhau nhằm cải thiện hiệu suất so với mô hình cơ sở, kiến trúc mô hình được minh họa ở Hình 2.



Hình 2. *Minh họa mô hình đã được cải tiến dựa trên mô hình* Multi-vision Contextual Attention

- **Khối Multimodel Fusion Module**

Khối Multimodal Fusion Module chịu trách nhiệm nối và điều chỉnh đặc trưng hình ảnh và câu hỏi vào không gian chung, đảm bảo rằng thông tin từ cả hai nguồn được biểu diễn trong cùng một miền đặc trưng với cùng kích thước. Sau đó, tiếp tục xử lí biểu diễn này bằng cách áp dụng

cơ chế Nhiều đầu tự chú ý (Multi-head self-attention), cho phép mô hình học được sự tương quan phức tạp giữa các thành phần hình ảnh và câu hỏi ở nhiều mức độ khác nhau.

Đầu tiên, đặc trưng hình ảnh được trích xuất dưới hai dạng: đặc trưng cục bộ, phản ánh thông tin chi tiết của từng vùng trong ảnh, và đặc trưng toàn cục, mô tả bối cảnh tổng thể của hình ảnh. Đồng thời, đặc trưng của câu hỏi cũng được trích xuất để biểu diễn thông tin ngữ nghĩa của văn bản. Tất cả các đặc trưng trên sau đó được đưa vào khối Multimodal Fusion Module. Tại đây, đặc trưng hình ảnh cục bộ và toàn cục được nối lại (2) và giảm chiều về 1024 bằng lớp tuyến tính. Đặc trưng câu hỏi cũng được chuyển đổi về cùng số chiều (kích thước là 1024) , trước khi kết hợp với đặc trưng hình ảnh để tạo ra một biểu diễn hợp nhất (3). Tiếp theo, biểu diễn này được đưa vào khối Multi-Head Attention (4), nơi các đặc trưng được chia thành nhiều đầu để thực hiện tính toán tự chú ý độc lập. Mỗi đầu học được một biểu diễn khác nhau về sự tương tác giữa hình ảnh và văn bản, giúp mô hình tập trung vào các phần quan trọng của cả hai nguồn thông tin. Kết quả từ tất cả các đầu sau đó được nối lại và chuyển đổi tuyến tính để tạo ra biểu diễn hợp nhất cuối cùng.

$$V_{concat\_visual} = concatenation(V_{global}, V_{local}) \quad (2)$$

$$V_{q,t} = concatenation[\, Linear(\, V_{concat\_visual}), Linear(\, V_{text})] \quad (3)$$

$$V_{mha} = Self - MHA(V_{q,t}) = MHA\ (V_{q,t}, V_{q,t}, V_{q,t}) \quad (4)$$

trong đó

V_{local} vector đặc trưng cục bộ của hình ảnh.

V_{global} là đặc trưng toàn cục của hình ảnh.

V_{concat_visual} là đặc trưng đa tầm nhìn của hình ảnh.

V_{text} là vector đặc trưng của câu hỏi văn bản.

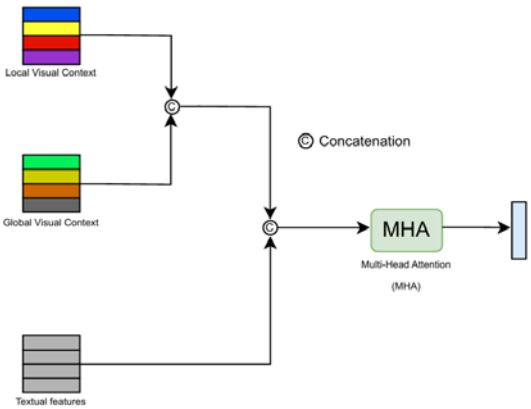
Linear là lớp tuyến tính trong mô hình mạng nơ-ron.

$V_{q,t}$ là vector đa ngữ cảnh hình ảnh và câu hỏi.

V_{mha} là vector đa ngữ cảnh hình ảnh và câu hỏi đã được xử lí qua MHA.

$Self - MHA(V_{q,t})$ là lớp Multi-Self Attention.

$MHA\ (V_{q,t}, V_{q,t}, V_{q,t})$ là lớp Multi-Head Attention.



Hình 3. *Minh họa khối hợp nhất đa mô hình (Multimodel Fusion Module)*

Multi-Head Attention (MHA) là một cơ chế quan trọng trong các mô hình Transformer, giúp mô hình có khả năng tập trung vào nhiều vùng thông tin khác nhau của dữ liệu đồng thời. Thay vì sử dụng một đầu self-attention duy nhất, MHA bao gồm nhiều đầu attention hoạt động song song, mỗi đầu học một cách biểu diễn khác nhau của dữ liệu, giúp mô hình hiểu thông tin một cách phong phú và đa chiều hơn. Mỗi đầu attention được tính dựa trên ba thành phần: Query (Q) đại diện cho thông tin cần tìm, Key (K) đóng vai trò so sánh để xác định mức độ liên quan, và Value (V) chứa thông tin đầu vào cần trích xuất. Cơ chế attention được tính theo công thức chuẩn hóa softmax của tích vô hướng giữa Q và K, giúp xác định trọng số chú ý, sau đó nhân với V để tạo ra đầu ra có trọng số phù hợp. Trong MHA, các đầu attention này được ghép lại và đưa qua một phép chiếu tuyến tính để giữ nguyên kích thước ban đầu của dữ liệu. Việc sử dụng nhiều đầu attention giúp mô hình nắm bắt các mối quan hệ từ nhiều góc độ khác nhau, cải thiện khả năng hiểu ngữ cảnh, đặc biệt hữu ích trong các bài toán như VQA, nơi mô hình cần phân tích đồng thời thông tin từ cả hình ảnh và văn bản để đưa ra câu trả lời chính xác.

$$MHA(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

trong đó:

Q (Query): là vector truy vấn, đại diện cho thông tin tìm kiếm.

K (Key): là vector khóa, đại diện cho thông tin có sẵn để so sánh với truy vấn.

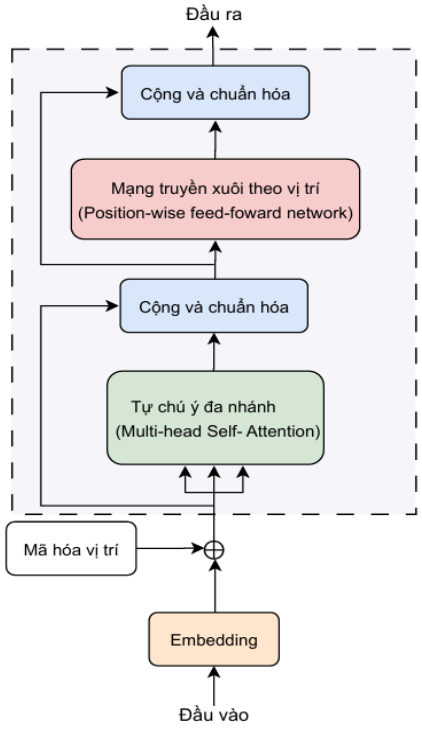
V (Value): là vector giá trị, chứa thông tin mà ta muốn lấy ra dựa trên mức độ tương đồng giữa Q và K.

Việc sử dụng Multimodel Fusion Module giúp cải thiện đáng kể khả năng xử lí của mô hình, cho phép nó tự động điều chỉnh trọng tâm vào các thông tin quan trọng từ cả hình ảnh và văn bản. Cơ chế này đặc biệt hữu ích trong việc xử lí các câu hỏi yêu cầu phân tích mối quan hệ phức tạp giữa các đối tượng trong ảnh.

- **Khối Multiple Self - Attention**

Kết quả đầu ra của khối Multimodal Fusion Module tiếp tục được đưa vào khối Multiple Self-Attention, bao gồm hai lớp self-attention, đây chính là khối encoder trong kiến trúc transformer. Mỗi lớp self-attention trong khối này đảm nhận nhiệm vụ khai thác các mối quan hệ dài hạn và tinh chỉnh biểu diễn thông tin, giúp mô hình tập trung tốt hơn vào các vùng quan trọng của hình ảnh theo ngữ cảnh của câu hỏi.

Khối Encoder trong Transformer bao gồm nhiều lớp (layers) giống nhau, mỗi lớp chứa hai thành phần chính: Multi-Head Self-Attention và Feed-Forward Network (FFN). Đầu vào của Encoder được chuyển thành vector nhúng (embedding) và kết hợp với positional encoding để giữ thông tin thứ tự. Sau đó, qua Multi-Head Self-Attention, mô hình học mối quan hệ giữa các phần khác nhau trong dữ liệu. Tiếp theo, đầu ra được đưa qua Feed-Forward Network, giúp tăng khả năng biểu diễn phi tuyến. Cả hai thành phần này đều có Layer Normalization và Residual Connection để ổn định quá trình huấn luyện.



Hình 4. *Minh họa khối Multiple Self - Attention (encoder trong kiến trúc transformer)*

Nhờ vào thiết kế này, mô hình không chỉ có khả năng tự động điều chỉnh trọng tâm vào các thông tin quan trọng trong cả hình ảnh và văn bản mà còn cải thiện hiệu suất phân tích các câu hỏi yêu cầu suy luận phức tạp về mối quan hệ giữa các đối tượng. Cách tiếp cận này mang lại sự cải thiện đáng kể về độ chính xác và tính linh hoạt, khiến mô hình trở nên phù hợp hơn với các ứng dụng VQA trong thực tế.

2.3. Thành phần kết hợp đặc trưng Multi-Context Visual and text Features từ Multi-vision Contextual Attention (mô hình cơ sở) và Fusion Based on Attention Mechanism (nhánh được cải tiến từ mô hình cơ sở)

Trong nghiên cứu này, chúng tôi đề xuất một mô hình cải tiến cho bài toán trả lời câu hỏi trực quan (VQA) trên ngôn ngữ tiếng Việt dựa trên mô hình Multi-vision Contextual Attention. Mô hình bao gồm hai nhánh chính:

- Nhánh của mô hình cơ sở (Multi-vision Contextual Attention Model), trong đó đặc trưng hình ảnh và câu hỏi được kết hợp bằng cơ chế chú ý có hướng dẫn.
- Nhánh cải tiến (Fusion Based on Attention Mechanism), nơi chúng tôi áp dụng một cách tiếp cận hợp nhất mới với hai khối chính: Multimodal Fusion Module và Multiple Self-Attention.

$$J_{FBOAM} = Linear(V_{mha}) \quad (6)$$

$$J_{MVCAM} = Linear(J_{MVCAM}) \quad (7)$$

$$J = J_{FBOAM} + J_{MVCAM} \quad (8)$$

trong đó:

V_{mha} là vector đa ngữ cảnh hình ảnh và câu hỏi của nhánh cải tiến (Fusion Based on Attention Mechanism).

Linear phép biến đổi tuyến tính trong mạng nơ-ron.

J_{FBOAM} là vector đa ngữ cảnh hình ảnh và câu hỏi của nhánh cải tiến (Fusion Based on Attention Mechanism) sau khi đi qua lớp linear.

J_{MVCAM} là vector đa ngữ cảnh hình ảnh và câu hỏi của nhánh mô hình cơ sở (Multi-vision Contextual Attention Model).

J vector tổng của cả 2 nhánh cơ sở và nhánh cải tiến.

Sau khi đi qua khối Multimodal Fusion Module và khối Multiple Self-Attention, chúng tôi thu được Multi-Context Visual and Text Features từ nhánh Fusion Based on Attention Mechanism. Để tận dụng đầy đủ thông tin từ cả hai nhánh, chúng tôi tiếp tục kết hợp đặc trưng này với Multi-Context Visual and Text Features từ nhánh của mô hình cơ sở (Multi-vision Contextual Attention Model) thông qua phép cộng vector.

2.4. Hàm mất mát

Trong nghiên cứu này, chúng tôi huấn luyện mô hình cải tiến bằng cách sử dụng hàm sử dụng 2 hàm mất mát là Cross-Entropy Loss và Normalized Temperature-scaled Cross Entropy Loss. Cụ thể hàm mất mát tổng được định nghĩa như sau:

$$L_{total} = L_{cross-entropy} + L_{i,j} \quad (9)$$

trong đó: $L_{cross-entropy}$ và $L_{i,j}$ lần lượt là hàm mất mát Cross-Entropy Loss và Normalized Temperature-scaled Cross Entropy Loss.

- **Cross-Entropy Loss**

Cross-Entropy Loss là một hàm mất mát phổ biến trong bài toán phân loại, đặc biệt khi sử dụng mô hình mạng nơ-ron nhân tạo. Hàm mất mát này đo lường sự khác biệt giữa phân phối xác suất dự đoán của mô hình và nhãn thực tế.

$$L_{cross-entropy} = -\sum_{i=1}^{c} y_i \log(p_i) \quad (10)$$

trong đó: C là tổng số câu trả lời (tổng số nhãn), xác suất đầu ra của mô hình cho mỗi lớp i là p_i, và nhãn thực tế được biểu diễn bằng một vector one-hot y_i.

- **Normalized Temperature-scaled Cross Entropy Loss**

Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent Loss) là một biến thể của Cross-Entropy Loss, được sử dụng phổ biến trong học tự giám sát (Self-Supervised Learning), đặc biệt trong Contrastive Learning (như trong mô hình SimCLR). NT-Xent Loss giúp tối ưu hóa mô hình bằng cách đưa các điểm dữ liệu tương đồng lại gần nhau trong không gian nhúng và đẩy các điểm khác biệt ra xa.

$$L_{i,j} = -\log \frac{\exp(\text{cosine}(z_i, z_j)/T)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{cosine}(z_i, z_j)/T)} \quad (11)$$

trong đó: z_i và z_j là hai vector nhúng (embedding) tương ứng câu trả lời đúng và câu trả lời sai, $\text{cosine}(z_i, z_j)/T$ là hàm đo độ tương đồng cosine giữa hai vector, T (temperature scaling) là một hệ số nhiệt độ điều chỉnh độ sắc nét của phân phối xác suất, $[k \neq i]$ đại diện cho các câu trả lời sai trong một lô dữ liệu (1 batch).

2.5. Thành phần dự đoán câu trả lời

Nhiệm vụ trả lời câu hỏi dựa trên hình ảnh sử dụng ngôn ngữ tiếng Việt, có thể tiếp cận bài toán này dưới góc độ của một tác vụ phân loại (Tsoumakas & Katakis, 2008). Cách tiếp cận này cho phép khai thác hiệu quả các đặc trưng kết hợp đa phương thức, được kí hiệu là J, thu được từ quá trình tích hợp thông tin thị giác và ngôn ngữ. Đặc trưng kết hợp này sau đó được đưa qua một cấu trúc mạng nơ-ron bao gồm hai lớp kết nối đầy đủ (Fully Connected Layers - FCL) và một hàm softmax ở lớp cuối cùng, nhằm mục đích tính toán và đưa ra xác suất dự đoán cho mỗi phương án trả lời có thể $\hat{a}$ trong một tập hợp các câu trả lời có sẵn A.

Cụ thể, quá trình dự đoán này được thực hiện thông qua việc áp dụng hàm softmax trên kết quả đầu ra của chuỗi hai mạng kết nối đầy đủ, biểu diễn qua công thức dưới đây:

$$p(\hat{a}) = soft\max\big(FFN(J)\big) \quad (12)$$

trong đó, FFN(J) đại diện cho kết quả đầu ra của quá trình xử lí đặc trưng J thông qua mạng kết nối đầy đủ. Hàm softmax sau đó được sử dụng để chuyển đổi kết quả này thành một phân phối xác suất trên toàn bộ tập câu trả lời có sẵn, giúp xác định câu trả lời cuối cùng mà mô hình dự đoán là phù hợp nhất với câu hỏi và hình ảnh đầu vào.

3. Thực nghiệm và so sánh

3.1. Môi trường thực nghiệm

Mô hình được triển khai trên Python 3.10 và thư viện: Pytorch (phiên bản 2.2). Chúng tôi cũng sử dụng một số thư viện hỗ trợ như: Huggingface (bản 4.26), Timm (bản 0.9.16. Các thư viện khác liên quan: numpy, pandas, scikit-learn, matplotlib, pytorch-grad-cam…

Để đạt được hiệu suất tốt cho bài toán trả lời câu hỏi trực quan trên tiếng Việt, việc lựa chọn các siêu tham số một cách cẩn thận là điều quan trọng. Các siêu tham số chính được sử dụng trong quá trình huấn luyện mô hình bao gồm: số lượng epoch là 40, kích thước batch là 32, tham số học khởi đầu (learning rate) là 1e-4, trọng số suy giảm (weight decay) là 1e-5, và tỉ lệ dropout là 0.3.

3.2. Dữ liệu đầu vào

Bộ dữ liệu: chúng tôi sử dụng bộ dữ liệu ViVQA để thực nghiệm mô hình trong bài toán trả lời câu hỏi trực quan (VQA) dành cho tiếng Việt, bộ dữ liệu này được xây dựng bởi Tran et al. (2021). Bộ dữ liệu gồm 15.000 dòng câu hỏi, câu trả lời và id tấm ảnh tương ứng với 10.328 hình ảnh được chia thành 80% cho tập huấn luyện (train) và 20% cho tập kiểm tra (test) để đánh giá mô hình.

Chỉ số đánh giá: Độ chính xác (accuracy) được sử dụng làm thước đo chính để đánh giá hiệu suất mô hình, với bài toán VQA được xem như một bài toán phân loại.

3.3. Kết quả thực nghiệm

Sau quá trình huấn luyện và đánh giá mô hình dựa trên các chỉ số đánh giá đã được xác định và tiến hành so sánh kết quả của mô hình này với các công trình nghiên cứu khác đã được công bố gần đây. Để có một cái nhìn toàn diện và chi tiết hơn, các mô hình từ nghiên cứu của Tran et al. (2021) đã được lựa chọn, đã đề xuất mô hình cơ sở trong bài báo nghiên cứu về hệ thống ViVQA. Ngoài ra, mô hình Bidirectional Cross-Attention, một phát kiến mới nhất của Tran et al. (2021) cũng được lựa chọn. Mục tiêu của việc so sánh này không chỉ để đánh giá hiệu quả của mô hình đã phát triển, mà còn nhằm mục đích khám phá các tiềm năng và hướng cải tiến cho các mô hình tương lai trong lĩnh vực tương tự.

Bảng 1. *So sánh kết quả thực nghiệm với các mô hình khác*

Mô hình	Độ chính xác (%)
LSTM+PhoW2Vec	33,85
Bi-LSTM+PhoW2Vec	33,97
Hier-Co-Att+PhoW2Vec	34,96
Bideractional Cross-Att	51,30
Multi-vision Contextual Attention	60,76
Mô hình đề xuất	62,41

Trong Bảng 1, thể hiện kết quả so sánh hiệu suất giữa mô hình đề xuất với các nghiên cứu trước đây cho thấy sự cải thiện đáng kể của mô hình đề xuất so với mô hình cơ sở và các mô hình được công bố trước đó trong việc xử lí bài toán trả lời câu hỏi trực quan (VQA) trên ngôn ngữ tiếng Việt. Các nghiên cứu trước đây sử dụng các phương pháp như LSTM, Bi-LSTM, Hier-Co-Att hay Bidirectional Cross-Att, mặc dù đạt được hiệu quả nhất định, nhưng vẫn còn hạn chế trong việc tích hợp chặt chẽ giữa thông tin hình ảnh và ngôn ngữ. Cụ thể mô hình đề xuất đạt được độ chính xác 62,41% so với mô hình cơ sở Multi-vision Contextual Attention (60,76%) và vượt trội so với các mô hình trước đó: LSTM+PhoW2Vec (33,85%), Bi-LSTM+PhoW2Vec (33.97%), Hier-Co-Att+PhoW2Vec (34,96%) và Bideractional Cross-Att (51,30%).

Kết quả dự đoán trên các mô hình khác nhau

Để tối ưu cho việc triển khai mô hình VQA trên các nền tảng, chúng tôi thực nghiệm trên hai mô hình có kích thước lớn và một mô hình có kích thước nhỏ thường được triển khai trên các thiết bị di động để rút trích đặc trưng hình ảnh. Ba mô hình này đã được được lần lượt được thử nghiệm để trích xuất đặc trưng toàn cục của hình dùng để đánh giá độ chính xác trong bài toán trả lời câu hỏi qua hình ảnh (VQA): Vision Transformer, Swin Transformer (Liu et al., 2021) và Mobinet (Wang et al., 2020). Kết quả thu được từ từng mô hình giúp xác định mức độ hiệu quả của chúng trong việc trích xuất các đặc trưng từ hình ảnh. Bằng cách này, có thể tìm ra phương pháp tối ưu nhất để áp dụng vào hệ thống VQA (Bảng 2) qua đó nâng cao khả năng nhận diện và trả lời chính xác các câu hỏi dựa trên hình ảnh được trình bày. Kết quả cho thấy việc triển khai mô hình trên các thiết bị di động là khả thi với mô hình nhỏ Mobinet đạt độ chính xác 60,2% – không chênh lệch quá nhiều so với việc áp dụng hai mô hình lớn còn lại.

Bảng 2. *So sánh các mô hình trích xuất đặc trưng hình ảnh*

Mô hình	Độ chính xác (%)
Vision Transformer	62.41
Swin Transformer	59,97
MobileNets	60,2

4. Kết luận và hướng phát triển

Trong bài báo này, chúng tôi đề xuất mô hình chú ý ngữ cảnh đa tầm nhìn cải tiến cho bài toán trả lời câu hỏi hình ảnh (VQA) trên ngữ liệu tiếng Việt. Cụ thể, chúng tôi cải tiếp bằng cách bổ sung một nhánh kết hợp mới mang tên Fusion Based on Attention Mechanism, trong đó tận dụng khả năng kết hợp và khai thác thông tin của hai khối chính: Multimodal Fusion Module và Multiple Self-Attention. Kết quả thực nghiệm trên bộ dữ liệu ViVQA cho thấy mô hình của chúng tôi đã cải thiện đáng kể độ chính xác và vượt trội so với một số phương pháp trước đây. Cụ thể mô hình đề xuất đạt được độ chính xác 62,41% so với mô hình cơ sở Multi-vision Contextual Attention (60,76%) và vượt trội so với các mô hình được công bố trước đó. Kết quả thực nghiệm cho thấy hiệu hiệu suất của mô hình cải tiến đã được cải thiện rõ rệt so với mô hình cơ sở và phần đề xuất cải tiến của chúng tôi là hiệu quả và có ý nghĩa.

Tuy nhiên, nghiên cứu này vẫn còn tồn tại một số hạn chế. Kích thước và sự đa dạng của bộ dữ liệu ViVQA còn khá khiêm tốn so với các bộ dữ liệu VQA tiếng Anh, điều này có thể ảnh hưởng đến khả năng mở rộng và tổng quát hóa của mô hình trong các tình huống thực tế. Bên cạnh đó, kiến trúc mô hình mặc dù đã đạt kết quả tích cực nhưng vẫn còn tiềm năng để tối ưu hóa hơn nữa, đặc biệt là trong việc cân bằng giữa hiệu suất và chi phí tính toán.

Trong tương lai, chúng tôi hướng đến việc mở rộng bộ dữ liệu VQA tiếng Việt để đảm bảo sự phong phú và đa dạng hơn, giúp mô hình có thể học được các ngữ cảnh phức tạp hơn. Đồng thời, việc cải tiến kiến trúc mô hình để tăng cường khả năng kết hợp thông tin giữa đặc trưng hình ảnh và ngôn ngữ sẽ là một trong những trọng tâm nghiên cứu tiếp theo.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425-2433. https://doi.org/10.1109/ICCV.2015.279

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate (arXiv:1409.0473). *arXiv*. https://arxiv.org/abs/1409.0473

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, *1*, 91-163.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp.1251-1258). https://doi.org/10.1109/CVPR.2017.195

Duan, T. D., Du, T. H., Phuoc, T. V., & Hoang, N. V. (2005, February). Building an automatic vehicle license plate recognition system. In *Proceedings of the International Conference on Computer Science RIVF* (Vol. 1, pp.59-63).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jallouli, N., Elghniji, K., Hentati, O., Ribeiro, A. R., Silva, A. M., & Ksibi, M. (2016). UV and solar photo-degradation of naproxen: TiO_2 catalyst effect, reaction kinetics, products identification and toxicity assessment. *Journal of Hazardous Materials, 304*, 329-336. https://doi.org/10.1016/j.jhazmat.2015.10.045

Lagorio, A., Tistarelli, M., Cadoni, M., Fookes, C., & Sridharan, S. (2013, April). *Liveness detection based on 3D face shape analysis. In 2013 International Workshop on Biometrics and Forensics* (IWBF) (pp.1-4). IEEE. https://doi.org/10.1109/IWBF.2013.6547310

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541-551. https://doi.org/10.1162/neco.1989.1.4.541

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp.10012-10022). https://doi.org/10.1109/ICCV48922.2021.00986

Nguyen, A. D., Le, T., & Nguyen, H. T. (2022, November). Combining multi-vision embedding in contextual attention for Vietnamese visual question answering. *In Pacific-Rim Symposium on Image and Video Technology* (pp.172185). Springer. https://doi.org/10.1007/978-3-031-26431-3_14

Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese (arXiv:2003.00744). *arXiv*. https://doi.org/10.48550/arXiv.2003.00744

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*. https://arxiv.org/abs/1409.1556

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds*.), Proceedings of the 36th International Conference on Machine Learning* (ICML 2019) (pp.6105-6114). PMLR.

Tran, D. M. N., Le, T., Nguyen, M. L., & Nguyen, H. T. (2022, October). Bi-directional cross-attention network on Vietnamese visual question answering. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation* (pp.834-841).

Tran, K. Q., Nguyen, A. T., Le, A. T. H., & Van Nguyen, K. (2021). ViVQA: Vietnamese visual question answering. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation* (pp.683-691).

Tsoumakas, G., & Katakis, I. (2008). Multi-label classification: An overview. *In Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp.64-74). https://doi.org/10.4018/978-1-59904-951-9.ch005

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

Wang, W., Li, Y., Zou, T., Wang, X., You, J., & Luo, Y. (2020). A novel image classification approach via dense-MobileNet models. *Mobile Information Systems, 2020*(1), Article 7602384. https://doi.org/10.1155/2020/7602384

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp.6281-6290). https://doi.org/10.48550/arXiv.1906.10770

AN IMPROVED MULTI-VISION CONTEXTUAL ATTENTION MODEL FOR VIETNAMESE VISUAL-BASED QUESTION ANSWERING

Bui Anh Dai*, Nguyen Quoc Trung, Tran Thanh Nha, Nguyen Viet Hung

Ho Chi Minh City University of Education, Vietnam

**Corresponding author: Bui Anh Dai – Email: buianhdai1412@gmail.com*

Received: June 11, 2024; Revised: November 25, 2024; Accepted: December 19, 2024

ABSTRACT

Visual Question Answering (VQA) represents the intersection of Computer Vision and Natural Language Processing, offering both scientific significance and practical applications. Integrating VQA models into mobile devices can assist blind and visually impaired individuals in accessing and understanding image content. A common approach involves extracting features from different image regions to capture local context. However, this method often overlooks the global context, which affects the model's ability to aggregate information and make accurate inferences. Recent methods leverage Vision Transformer to extract both global and local features from images, enhancing model performance. Additionally, multimodal attention mechanisms are applied to optimize the integration of image and question features, allowing the model to focus on key features and better understand the context. While most VQA models are designed for English datasets, research on Vietnamese VQA (ViVQA) remains limited. In this paper, we propose an improved model based on Multi-Vision Contextual Attention, achieving an accuracy of 62.41%, a significant improvement over the original model's 60% on the ViVQA dataset.

Keywords: multimodal; natural language; PhoBERT; ResNet; Swin Transformer; Vietnamese language; visual question answering (VQA)