Phương pháp diễn giải cho mô hình hỏi đáp hình ảnh tiếng Việt

Nguyễn Thanh Tân , Trần Thị Phương Linh, Lê Thanh Tùng, Nguyễn Tiến Huy
Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

Tác giả liên hệ: Nguyễn Tiến Huy. Email: ntienhuy@fit.hcmus.edu.vn
Ngày nhận bài: 14/10/2024, ngày sửa chữa: 25/11/2024, ngày duyệt đăng: 28/11/2024
DOI: 10.32913/mic-ict-research-vn.v2024.n2.1338

Tóm tắt: Diễn giải trong mô hình hỏi đáp hình ảnh là chủ đề được quan tâm bởi vì tính minh bạch của mô hình sẽ giúp người sử dụng tin tưởng cũng như hiểu về điểm mạnh và hạn chế của mô hình. Các mô hình đa phương thức như hỏi đáp hình ảnh sử dụng cả hai loại dữ liệu hình ảnh và văn bản, do đó gây thách thức cho các kỹ thuật giải thích thường chỉ áp dụng trên một loại dữ liệu. Trong công trình này, chúng tôi đề xuất kết hợp phương pháp giải thích hình ảnh Grad-CAM và phương pháp giải thích văn bản Transformer để phân tích hành vi của mô hình. Các kết quả thu được trên tập dữ liệu hỏi đáp tiếng Việt giúp chúng tôi rút ra các nhận xét như sau: i) Transformer thị giác trích xuất đặc trưng toàn cục trong khi ResNet trích xuất đặc trưng cục bộ; ii) mô hình suy luận không tốt với ảnh có nhiều yếu tố gây nhiễu; iii) phương thức văn bản có ít tác động hơn phương thức hình ảnh; iv) mô-đun PhoBERT có sự thiên vị một số câu hỏi nhất định.

Từ khóa: *trí tuệ nhân tạo khả diễn, hỏi đáp hình ảnh*

Title: **An XAI Method for Vietnamese Visual Question Answering Models**

Abstract: Interpretability in Visual Question Answering (VQA) models is a topic of interest because the transparency of the model can help users trust and understand its strengths and limitations. Multimodal models like VQA utilise both image and text data, posing challenges for interpretability techniques that are typically designed for a single type of data. In this work, we propose a combination of the Grad-CAM image explanation method and the Transformer-based text explanation method to analyse the behaviour of the model. The results obtained from the Vietnamese VQA dataset lead us to the following observations: i) the visual Transformer extracts global features, while ResNet extracts local features; ii) the model struggles with images containing many distracting elements; iii) the text modality has less impact compared to the image modality; iv) the PhoBERT module shows bias toward certain types of questions.

Keywords: *XAI, visual question answering*

I. GIỚI THIỆU

Diễn giải mô hình trí tuệ nhân tạo (AI) là một quá trình giúp con người hiểu rõ hơn về cách mà mô hình hoạt động, tại sao nó đưa ra các quyết định hoặc dự đoán nhất định. Quá trình này không chỉ giúp người sử dụng tin tưởng vào quyết định của ứng dụng trí tuệ nhân tạo mà còn giúp tìm ra điểm mạnh, điểm yếu, cũng như các điểm lỗi của mô hình [1].

Các nghiên cứu về diễn giải mô hình trí tuệ nhân tạo ngày nay cũng đã đạt được nhiều bước tiến quan trọng. Nổi bật có thể kể đến như nhóm Sevalraju và các cộng sự [2] đề xuất sử dụng Gradient của một lớp theo một tầng tích chập để tính ra một biểu đồ thể hiện sự quan trọng của các điểm ảnh hoặc đặc trưng được trích xuất của tầng tích chập đang xét. Nhóm Chefer và các cộng sự [3] sử dụng các luật kết hợp giữa quá trình lan truyền tới (forward) và lan truyền ngược (backward) để tạo ra biểu đồ liên quan giữa các đặc trưng ảnh hoặc văn bản để diễn giải cho mô hình Transformer.

Tuy nhiên, hầu hết các nghiên cứu về trí tuệ nhân tạo khả diễn thường tập trung vào các mô hình đơn phương thức. Trong khi đó, với sự phát triển của các hệ thống trí tuệ nhân tạo gần đây, học máy đa phương thức đang trở thành đề tài nghiên cứu quan trọng, nhận được sự quan tâm của rất nhiều nhà khoa học. Mô hình đa phương thức có thể xử lý đồng thời nhiều nguồn dữ liệu khác nhau như âm thanh, hình ảnh, văn bản... Việc kết hợp các thông tin đa dạng này là vô cùng cần thiết và đóng vai trò quan trọng trong sự bùng nổ không ngừng của các dữ liệu đa phương tiện ngày nay. Trong số đó, các nghiên cứu đa phương thức có

sự kết hợp giữa ảnh và văn bản đang được tập trung ngày càng nhiều nhằm giải quyết các bài toán đầy thử thách như chú thích ảnh, hỏi đáp trên ảnh, sinh ảnh từ văn bản... Nổi bật trong số đó có thể kể đến các bài toán thực tế như hỏi đáp hình ảnh giúp ích cho người khiếm thị của Le và các cộng sự [4]. Nghiên cứu này đề xuất hệ thống đa phương thức ngôn ngữ - thị giác giải quyết vấn đề của tập dữ liệu chất lượng thấp chẳng hạn như thiếu vắng các vật thể. Cũng trong chủ đề này, Nguyen và các cộng sự [5] đã đề xuất các mô hình đa phương thức dựa trên cơ chế attention và các biến thể như attention có hướng dẫn để tăng độ hiệu quả. Với sự phát triển không ngừng này, việc diễn giải các mô hình đa phương thức, cụ thể là với bài toán hỏi đáp hình ảnh, là nhu cầu thiết yếu và vô cùng cần thiết.

Mặc dù đã có nhiều nghiên cứu về diễn giải mô hình hỏi đáp hình ảnh, nhưng hầu hết đều tập trung vào các mô hình được huấn luyện trên dữ liệu tiếng Anh. Chưa có nhiều các công trình tập trung giải thích hành vi của mô hình đa phương thức trong tiếng Việt. Do đó, nghiên cứu của chúng tôi tập trung vào việc khảo sát và chọn lựa phương pháp giải thích phổ biến và hiệu quả để áp dụng cho mô hình hỏi đáp hình ảnh tiếng Việt. Cụ thể, công trình đề xuất sử dụng kết hợp giữa Grad-CAM [2] và diễn giải Transformer [3] để làm rõ hơn quá trình ra quyết định của mô hình hỏi đáp. Dựa trên các giải thích, chúng tôi tiến hành phân tích định tính để tìm hiểu những khuôn mẫu, hành vi cũng như điểm mạnh, điểm thiếu sót của mô hình hỏi đáp hình ảnh. Từ đó, công trình rút ra bốn kết luận chính:

1) Khối Transformer thị giác có xu hướng trích xuất đặc trưng toàn cục, trong khi khối ResNet chủ yếu tập trung vào việc trích xuất đặc trưng cục bộ.
2) Mô hình không thực hiện suy luận tốt khi xử lý các hình ảnh chứa nhiều vật thể.
3) Thông tin văn bản ít ảnh hưởng đến kết quả suy luận hơn so với thông tin thị giác.
4) Khối PhoBERT thể hiện sự thiên vị nhất định trong quá trình suy luận.

Tiếp theo, trong phần II, chúng tôi sẽ trình bày các kỹ thuật cơ bản trong diễn giải. Sau đó, phương pháp diễn giải cho mô hình hỏi đáp sẽ được đề xuất ở phần III. Hai phần IV và V sẽ trình bày các thực nghiệm cũng như thảo luận, phân tích về các kết quả này.

II. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

Trong phần này, chúng tôi trình bày một số nghiên cứu liên quan về phương pháp diễn giải cho mô hình trí tuệ nhân tạo.

1. Diễn giải bằng phân rã đa phương thức (DIME)

Xét trên bài toán hỏi đáp hình ảnh, DIME [6] phân rã một mô hình đa phương thức thành hai thành tố: đóng góp của từng phương thức và tương tác của các phương thức.

Sau khi phân rã mô hình, DIME tạo ra diễn giải bằng cách đánh giá mô hình, xem xét sự thay đổi của logit trên một lớp dự báo c, bằng cách tạo ra một tập dữ liệu mới như sau: thực hiện n lần việc thay đổi ngẫu nhiên một vùng trên ảnh so với ảnh ban đầu hoặc thực hiện n lần việc thay đổi ngẫu nhiên một chữ trong câu hỏi so với câu hỏi ban đầu. Trong môi trường đa phương thức, một tập dữ liệu mới chỉ nên chứa các thay đổi của một phương thức (chỉ ảnh hoặc chỉ văn bản), phương thức còn lại giữ cố định. Cuối cùng, thực hiện khớp một hàm tuyến tính trên dữ liệu logit của lớp c. Kết quả thu được là trọng số của các vùng trên ảnh dùng để làm biểu đồ nhiệt ảnh và histogram của các chữ trong câu hỏi.

DIME trực quan trên từng phương thức (câu hỏi và hình ảnh) để đưa ra biểu đồ nhiệt thay vì trực quan tại từng mô-đun của mô hình. Giả sử như người lập trình muốn biết độ hiệu quả (khả năng trích chọn đặc trưng) của từng mô-đun trong mô hình để cân nhắc mở rộng, thay thế hoặc sửa lỗi, họ không có một cách trực tiếp để đưa ra quyết định. Một vấn đề khác chúng tôi lưu tâm là quy trình của DIME tốn khá nhiều thời gian, bởi vì phải lặp lại quá trình suy diễn trên tập dữ liệu mới lớn hơn nhiều so với dữ liệu ban đầu, đây có thể là nơi "nghẽn cổ chai" của phương pháp này.

2. Diễn giải bằng văn bản kèm biểu đồ nhiệt

Trong bài toán hỏi đáp hình ảnh, nhóm tác giả Park và các cộng sự [7] đề xuất huấn luyện một mô hình có khả năng trả lời câu hỏi và tạo ra diễn giải bằng văn bản lẫn diễn giải bằng biểu đồ nhiệt trên hình ảnh.

Để thực hiện điều này, tác giả đã đề xuất hai tập dữ liệu mới chứa chân trị của văn bản diễn giải và biểu đồ nhiệt. Về kiến trúc của mô hình, tác giả gộp cả hai phần tạo ra câu trả lời và tạo ra diễn giải thành một kiến trúc thống nhất. Đầu vào là câu hỏi và hình ảnh để tạo ra câu trả lời; sau đó dùng cả câu hỏi, hình ảnh và câu trả lời để tạo ra biểu đồ nhiệt ảnh, từ đó tạo các đặc trưng attention để đưa ra diễn giải văn bản.

Mặc dù phương pháp này tạo ra diễn giải rất thân thiện với người dùng, nhưng việc tạo ra chân trị của văn bản thường do con người thực hiện, và phải trải qua quy trình sàng lọc để thu được dữ liệu chất lượng cao, do đó, rất tốn kém để thực hiện điều này.

3. Phương pháp biểu đồ lớp kích hoạt sử dụng Gradient (Grad-CAM)

a) Phương pháp biểu đồ lớp kích hoạt

Biểu đồ lớp kích hoạt (CAM) [8] là một phương pháp trực quan bằng cách sử dụng biểu đồ nhiệt để chỉ ra vùng ảnh đặc biệt mà mô hình CNN sử dụng để xác định nhãn của một nhãn nhất định trong tập các nhãn đầu ra. Ở tầng CNN cuối cùng, trước khi áp dụng một hàm kích hoạt để xác định đầu ra (ví dụ như softmax cho việc phân lớp), người ta dùng phép gộp trung bình toàn cục (GAP) các biểu đồ đặc trưng tích chập, sau đó cho qua một tầng kết nối dày đặc để đưa ra kết quả. Kết quả của việc áp dụng GAP là tạo ra một vector, tổ hợp tuyến tính phần tử thứ i của vector này với biểu đồ đặc trưng thứ i tương ứng trong tầng CNN cuối cùng sẽ cho ra một biểu đồ duy nhất được gọi là biểu đồ lớp kích hoạt. Nhược điểm của CAM nằm ở biểu đồ đặc trưng. Bởi vì cần thay đổi kiến trúc mô hình, cụ thể là thêm phép trung bình toàn cục vào biểu đồ đặc trưng cuối cùng, nên hiệu quả đã bị giảm sút ở một số tác vụ như phân lớp và không thể ứng dụng cho các tác vụ kết hợp ảnh và văn bản.

b) Phương pháp biểu đồ lớp kích hoạt sử dụng Gradient

Biểu đồ lớp kích hoạt sử dụng Gradient [2] là một giải pháp cải tiến của CAM nhằm loại bỏ những hạn chế đã nêu. Bằng cách sử dụng Gradient, phương pháp này không can thiệp vào kiến trúc của mô hình và hoàn toàn khả dụng cho các tác vụ kết hợp ảnh và văn bản. Ở tầng kết nối dày đặc cuối cùng của mô hình, ta thu được logit, y^c của nhãn c nào đó (đối với bài toán phân lớp), Grad-CAM sử dụng Gradient của y^c theo một hàm kích hoạt A^k trong lớp tích chập, hay nói cách khác là $\frac{\partial y^c}{\partial A^k}$. Áp dụng phép gộp trung bình toàn cục các luồng đạo hàm theo từng điểm ảnh của ảnh để cho ra trọng số quan trọng của điểm ảnh, gọi là α_k^c.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

Đây là phương pháp sẽ được sử dụng để giải thích các khối thị giác trong quy trình diễn giải của nghiên cứu này.

4. Phương pháp diễn giải cho kiến trúc Transformer

a) Sử dụng attention thô

Như đã biết, ở trong self-attention thì đầu ra chính là tổng có trọng số của đầu vào.

$$A = softmax(\frac{Q \cdot K^T}{\sqrt{d_h}}) \tag{2}$$

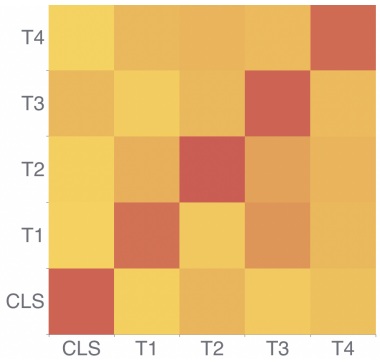
$$O = A \cdot V \tag{3}$$

Cụ thể hơn, mỗi đầu ra y_i ứng với token i sẽ là tổng có trọng số của các vector giá trị v của tất cả các token:

$$y_i = \sum_j \alpha_{i,j} v_j \tag{4}$$

trong đó, trọng số $\alpha_{i,j}$ là phần tử thứ i, j của ma trận attention A.

Thế nên, bằng cách thuần túy nhất, ta có thể sử dụng giá trị $\alpha_{i,j}$ để ước lượng mức độ *chú ý* của token j với token i. Ma trận attention A trong self-attention dùng để đo mức độ đóng góp của tất cả các token đầu vào đối với token đầu ra và mỗi dòng trong ma trận A biểu thị cho sự đóng góp của tất cả các token đối với một token (token đang xét tại dòng đó) (Hình 1).



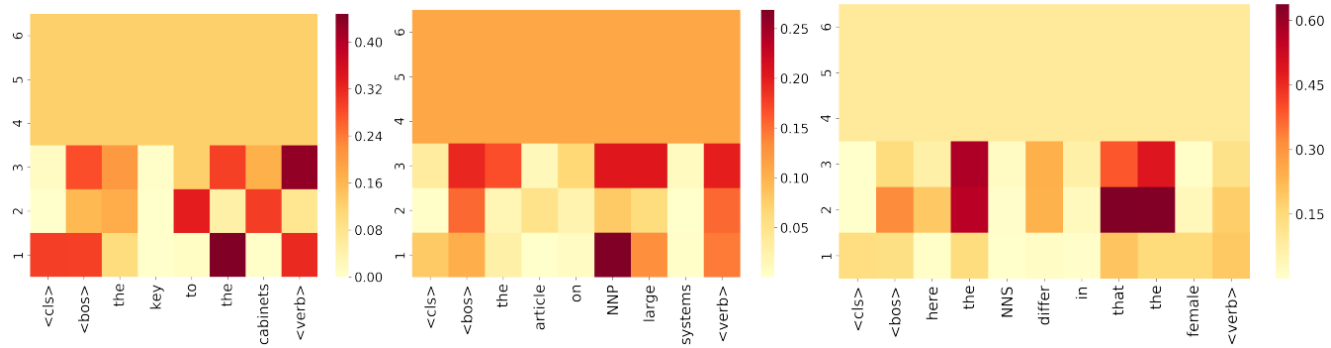
Hình 1. Mô tả ma trận attention của một câu với 4 token.

Vì token <CLS> chứa thông tin phân lớp toàn cục của đặc trưng để đưa ra dự đoán phân lớp cuối cùng, nên nó sẽ được sử dụng để đưa ra diễn giải. Thông thường, ta sẽ chỉ xét dòng tương ứng với token <CLS> có trong ma trận A đó, cũng có nghĩa là ta tập trung vào các đóng góp của tất cả các token có trong đầu vào lên token <CLS>. Tuy nhiên, ta cũng cần phải loại bỏ token đầu tiên (cũng chính là token <CLS>) vì việc lấy thông tin đóng góp của token <CLS> lên chính nó là hoàn toàn vô nghĩa.

Hạn chế: Mặc dù phương pháp này khá đơn giản và nhanh chóng nhưng [9] chỉ ra rằng đối với các mô hình Transformer có quá nhiều lớp attention, càng đi sâu hơn thì sự diễn giải trông khá "đồng đều" (Hình 2). Theo quan sát của tác giả, càng đi sâu hơn thì các embedding càng được *ngữ cảnh hóa*, làm cho các thông tin embedding mang ý nghĩa tương tự nhau.

Vấn đề cần giải quyết: Phương hướng khả thi nhất là tận dụng và phát huy càng nhiều thông tin trong Transformer càng tốt. Ví dụ như Transformer có nhiều block và head, mỗi cái đều chứa ma trận attention, chúng ta có thể "tổng hợp" các thông tin này lại. Tất nhiên cũng sẽ phát sinh các thách thức sau đây:

- ***Nhiều đầu attention:*** Một số head có thể sẽ không mang ý nghĩa quan trọng đối với đầu ra, cho dù có bỏ qua những head này thì mô hình vẫn đưa ra dự đoán



Hình 2. Biểu đồ attention thô cho token <CLS> tại các lớp khác nhau [9].

tốt. Vậy nên ta không thể xét các ma trận trong từng head một cách "bình đẳng" được.

- ***Nhiều lớp attention:*** Thông tin token được truyền giữa các lớp (khối) attention như thế nào và làm sao để tính toán được chúng.

Các phương pháp dưới đây đã được đề xuất để giải quyết những thách thức được nêu trên.

b) Attention Rollout

Attention Rollout [9] sử dụng đệ quy để tính attention của từng lớp trong Transformer từ giá trị embedding đầu vào.

- **Đối với việc giải quyết nhiều đầu attention:** Việc phân tích thông tin của từng đầu đòi hỏi phải xét đến việc các thông tin giữa chúng bị hòa lẫn với nhau. Trong bài báo gốc, để cho đơn giản thì tác giả chỉ tổng hợp các đầu trong một lớp lại thành một đầu duy nhất để tính toán bằng cách lấy trung bình attention các đầu.

$$E_h(A^{(b)}) = \frac{1}{M}\sum_m A_m^{(b)} \tag{5}$$

trong đó, b đại diện cho lớp (block) thứ b.

- **Đối với việc lan truyền thông tin giữa các lớp attention:** Để tính toán được cách thông tin được lan truyền giữa các lớp, đầu tiên, ta không thể không xét đến các kết nối phần dư có trong mô hình. Những kết nối phần dư này đóng vai trò rất quan trọng trong việc liên kết các vị trí tương ứng nhau giữa các lớp. Vậy nên, tác giả đề xuất việc chèn thêm trọng số để biểu diễn cho các kết nối phần dư, cụ thể là cộng thêm một ma trận đơn vị với ma trận attention: $A = I + E_h(A^{(b)})$.

Để tính toán attention từ lớp thứ i (l_i) truyền đến lớp thứ j (l_j), ta nhân các ma trận trọng số attention theo kiểu đệ quy:

$$\tilde{A}(l_i) = \begin{cases} A(l_i)\tilde{A}(l_{i-1}) & \text{nếu } i > j \\ A(l_i) & \text{nếu } i = j \end{cases} \tag{6}$$

Với công thức như trên, để tính toán attention từ đầu vào (lớp đầu tiên), ta có công thức $\tilde{A}(l_i) = A(l_1) \cdot A(l_2) \cdot \ldots \cdot A(l_i)$ (attention từ lớp thứ i đến lớp đầu vào). Sau khi thực hiện Rollout, ta có thể tính được "mức độ đóng góp" của lớp đầu tiên tới lớp đầu ra (với i là lớp cuối cùng). Phương pháp này hiệu quả hơn nhiều so với việc sử dụng attention thô vì attention thô chỉ tính được "mức độ đóng góp" của lớp gần cuối đối với lớp cuối.

c) Phương pháp được đề xuất trong bài báo Transformer Interpretability Beyond Attention Visualization

Bài báo [3] chỉ ra rằng phương pháp Rollout vẫn còn quá đơn giản. Thế nên, tác giả đã mở rộng và cải thiện phương pháp này với các ý tưởng sau đây:

- **Tổng hợp các đầu attention bằng relevance và Gradient:** Bởi vì mỗi đầu sẽ thể hiện các thông tin khác nhau nên "mức độ đóng góp" của từng đầu đối với dự đoán cũng phải khác nhau. Nếu ta chỉ sử dụng cách lấy trung bình như phương pháp Rollout thì sẽ không thể hiện được đặc điểm này. Do đó, ta cần phải xem xét tới việc đặt trọng số cho từng đầu. Trong bài báo này, tác giả sử dụng Gradient để làm trọng số, Gradient cũng cho biết thông tin liên quan đến phân lớp. Ngoài ra, thay vì sử dụng thuần attention, tác giả lại sử dụng relevance dựa vào phương pháp LRP [10] để biểu diễn cho attention của mỗi đầu trong từng lớp.

$$E_h(A^{(b)}) = \frac{1}{M}\sum_m \nabla A_m^{(b)} \odot R_m^{(n_b)} \tag{7}$$

với $\nabla A_m^{(b)} = \frac{\partial y_c}{\partial A_m^{(b)}}$, y_c là giá trị đầu ra của lớp c mà chúng ta cần diễn giải.

Tuy nhiên, vì ta chỉ cần tập trung vào phần nào của đặc trưng đầu vào có ý nghĩa đối với dự đoán, nên ta sẽ chỉ xét tới các đóng góp dương và loại bỏ các đóng góp âm ra khỏi ma trận Relevance:

$$E_h(A^{(b)}) = \frac{1}{M}\sum_m \nabla A_m^{(b)} \odot R_m^{(n_b)^+} \tag{8}$$

-**Lan truyền thông tin giữa các lớp attention bằng nhân ma trận:** Ý tưởng vẫn giữ nguyên giống như phương pháp Rollout:

$$\bar{A}^{(b)} = I + E_h(A^{(b)}) \tag{9}$$

$$\mathbf{C} = \bar{A}^{(1)} \cdot \bar{A}^{(2)} \cdot \ldots \cdot \bar{A}^{(B)} \tag{10}$$

Đa số tập dữ liệu về lĩnh vực hỏi đáp hình ảnh thiếu vắng diễn giải chân trị, do đó, đề xuất của chúng tôi là sử dụng các phương pháp không cần đến diễn giải chân trị, chẳng hạn như biểu đồ nhiệt và histogram. Mục tiêu của chúng tôi thiên về diễn giải mô hình hoạt động hơn là diễn giải dễ hiểu cho người dùng cuối, bởi lẽ biểu đồ nhiệt và histogram cũng không quá thân thiện với người dùng cuối; do đó, chúng tôi sẽ giải thích mô-đun trong mô hình mà không nhấn mạnh vào tác động của từng phương thức. Đặc biệt hơn, chúng tôi sử dụng mô hình được chứng minh có độ hiệu quả tốt cho tác vụ hỏi đáp hình ảnh tiếng Việt để phân tích nhằm đưa ra hướng cải thiện cho mô hình.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Trước tiên, chúng tôi giới thiệu mô hình được sử dụng trong thực nghiệm, bởi vì quy trình diễn giải đề xuất gắn liền với một mô hình học đa phương thức cụ thể.

1. Mô hình hỏi đáp hình ảnh

Mô hình hỏi đáp hình ảnh tiếng Việt của nhóm Nguyen và các cộng sự [5] được chọn để phân tích vì độ chính xác đạt 60,76% với kiến trúc kết hợp giữa attention và mạng nơ-ron tích chập. Đây là những khối kiến trúc phổ biến trong các tác vụ hỏi đáp hình ảnh.

Kiến trúc của mô hình được thể hiện ở Hình 4, cụ thể như sau:

1) Đầu vào của mô hình gồm một hình ảnh và một câu hỏi.
2) Hình ảnh đi qua song song hai mạng gồm: i) ResNet để trích xuất đặc trưng thị giác cục bộ; ii) Transformer thị giác (ViT) để trích xuất đặc trưng thị giác toàn cục, trong khi câu hỏi sẽ đi qua PhoBERT và tầng self-attention đa đầu (MHSA) để cho ra đặc trưng ngôn ngữ. Như vậy có ba đặc trưng được tạo ra, gồm hai đặc trưng thị giác và một đặc trưng ngôn ngữ.
3) Mỗi đặc trưng thị giác sẽ đi qua tầng attention có hướng dẫn tương ứng. Trong tầng này, đặc trưng thị giác sẽ làm truy vấn và đặc trưng ngôn ngữ sẽ làm khóa và giá trị. Kết quả là cho ra hai đặc trưng thị giác có hướng dẫn tương ứng với từng loại. Tầng attention có hướng dẫn của đặc trưng ngôn ngữ sẽ được trình bày sau.
4) Từng đặc trưng thị giác có hướng dẫn sẽ đi qua một tầng top-down attention để cho ra một vector đặc trưng thị giác. Hai vector đặc trưng thị giác được tạo ra bởi quá trình này sẽ nối lại để trở thành đặc trưng thị giác đa ngữ nghĩa.
5) Quay lại với đặc trưng ngôn ngữ. Sau khi được tạo ra, đặc trưng ngôn ngữ sẽ đi qua attention có hướng dẫn với đặc trưng ngôn ngữ là truy vấn và đặc trưng thị giác đa ngữ nghĩa làm khóa và giá trị. Kết hợp với phép gộp thì cuối cùng cho ra đặc trưng ngôn ngữ có hướng dẫn.
6) Đặc trưng thị giác đa ngữ nghĩa và đặc trưng ngôn ngữ có hướng dẫn sẽ đi qua tầng kết nối đầy đủ và áp dụng tích Hadamard lên chúng để cho ra đặc trưng đa phương thức.
7) Cuối cùng, đặc trưng đa phương thức đi qua lớp phân loại để đưa ra nhãn.

Mô hình này được huấn luyện và đánh giá trên tập dữ liệu cho tác vụ hỏi đáp hình ảnh tiếng Việt ViVQA [11] với độ chính xác là 60.76% và điểm F1 là 59.86%.

Chúng tôi đề xuất thử nghiệm những phương pháp đã được áp dụng thành công cho tập dữ liệu tiếng Anh. Chúng tôi tập trung vào giải thích ba mô-đun: mạng nơ-ron tích chập, kiến trúc Transformer và cơ chế attention có hướng dẫn.

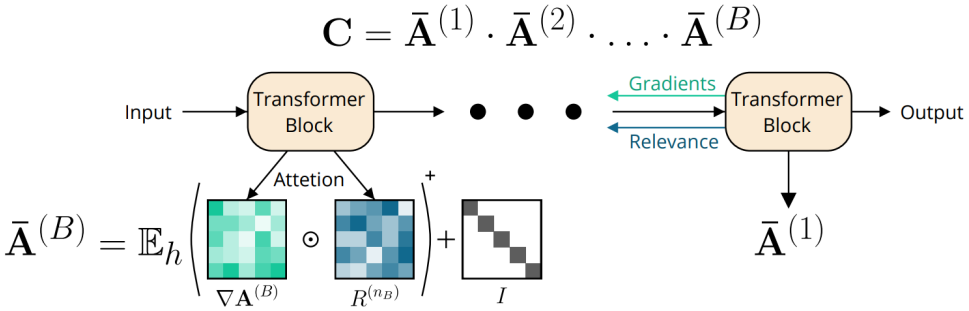
2. Phương pháp

a) Giải thích mạng nơ-ron tích chập bằng Grad-CAM

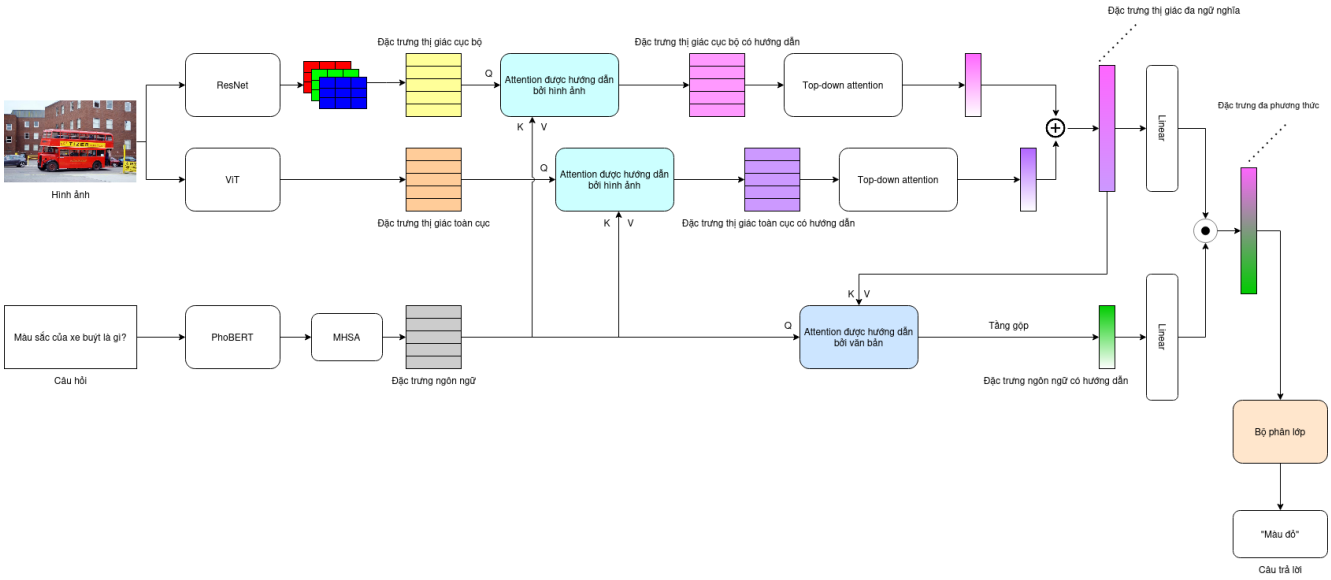
Nghiên cứu về Grad-CAM [2] đã cung cấp đầy đủ lý thuyết về phương pháp này. Gọi l là nhãn tại đầu ra của mô hình, y^l là logit tại tầng phân lớp (trước khi đi qua softmax). Tại một tầng tích chập A cuối cùng trong kiến trúc, ta dễ dàng tính được Gradient $\mathbf{G} = \frac{\partial y^l}{\partial A}$. Ngoài ra, ta cũng lưu lại giá trị trong quá trình lan truyền tiến (forward) của mô hình tại tầng tích chập A này, gọi là $\mathbf{C}$. $\mathbf{G}$ và $\mathbf{C}$ đều là tensor 4 chiều (đánh số từ 0 đến 3) có kích thước (b, c, n, n) với b là kích thước một batch đầu vào; c là số kênh sau khi đi qua tầng tích chập; n là kích thước của ma trận đặc trưng thu được, ở đây ta chỉ xét ma trận đặc trưng vuông.

Quá trình tìm biểu đồ nhiệt như sau:

1) Tính vector Gradient gộp của $\mathbf{G}$ (gọi là $\mathbf{g_p}$) bằng cách tính trung bình $\mathbf{G}$ theo chiều thứ 0, 2 và 3, tức là b và hai kích thước n của ma trận đặc trưng. Vector gộp của Gradient $\mathbf{g_p}$ có số chiều bằng với kích thước c của $\mathbf{G}$.
2) Tại mỗi kênh thứ i của ma trận $\mathbf{C}$, ta nhân chúng với phần tử thứ i của vector gộp $\mathbf{g_p}$.
3) Tính trung bình theo chiều thứ 1 của $\mathbf{C}$, tức là trung bình theo toàn bộ kênh. Kết quả thu được là một



Hình 3. Mô tả phương pháp. Gradient và Relevance được truyền từ đầu ra về lại đầu vào. Nguồn: Từ bài báo "Transformer Interpretability Beyond Attention Visualization" [3].



Hình 4. Kiến trúc của mô hình hỏi đáp hình ảnh được sử dụng trong nghiên cứu của nhóm Nguyen và các cộng sự [5].

tensor 3 chiều (gồm 1 chiều chứa số batch ảnh và 2 chiều chứa biểu đồ nhiệt), chứa ma trận có thể được trực quan hoá thành biểu đồ nhiệt.

b) Giải thích self-attention

Chúng tôi sử dụng phương pháp được đề xuất trong nghiên cứu của nhóm Chefer và các cộng sự [3] về diễn giải cơ chế self-attention thông qua biểu đồ Relevance.

Đầu tiên, khởi tạo ma trận Relevance **R**. Gọi i là số lượng token đầu vào của mô-đun self-attention. Do mô hình này chỉ sử dụng cơ chế self-attention cho nên ta sẽ khởi tạo ma trận Relevance là một ma trận đơn vị.

$$\mathbf{R} = \mathbb{I}^{i \times i} \tag{11}$$

Tiếp theo, thực hiện phép trung bình toàn bộ đầu của mô-đun self-attention sau khi nhân từng phần tử với đạo hàm của chúng. Đặt A là một tầng attention (chứa mô-đun self-attention) nào đó trong mô hình.

$$\bar{A} = E_h(A) = \frac{1}{M} \sum_m (\nabla A_m \odot A_m)^+ \tag{12}$$

Cuối cùng, cập nhật ma trận Relevance.

$$R = R + \bar{A}R \tag{13}$$

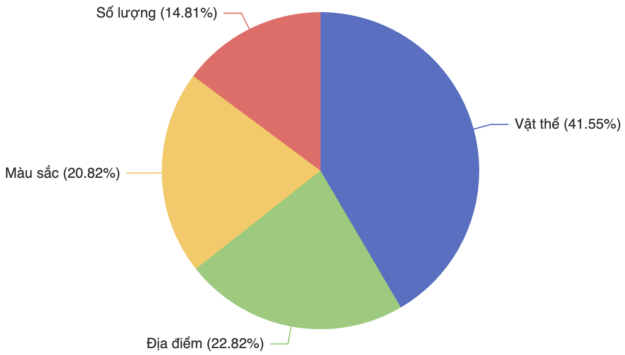
Ma trận R được dùng để trực quan thành biểu đồ nhiệt. Với kiến trúc được sử dụng để xử lý văn bản (ví dụ: PhoBERT), người ta sẽ đặt phần tử Relevance của token `<CLS>` bằng 0.

IV. THỰC NGHIỆM VÀ PHÂN TÍCH

1. Tập dữ liệu

Tập dữ liệu được chúng tôi sử dụng để thực nghiệm là tập ViVQA ([11]) bao gồm 10.328 hình ảnh cùng 15.000 cặp câu hỏi-trả lời bằng tiếng Việt có liên quan đến các

hình ảnh đó. Tác giả chia tập dữ liệu thành các tập huấn luyện và tập kiểm thử với tỷ lệ 8:2.



Hình 5. Phân bổ các câu hỏi trong tập dữ liệu.

Các dạng câu hỏi có trong tập dữ liệu được chia làm bốn loại chính, bao gồm: Vật thể, số lượng, màu sắc, địa điểm. Phân bổ các dạng câu hỏi được trình bày trong hình 5, lưu ý rằng phân bổ này ở cả hai tập dữ liệu là như nhau.

2. Phân tích

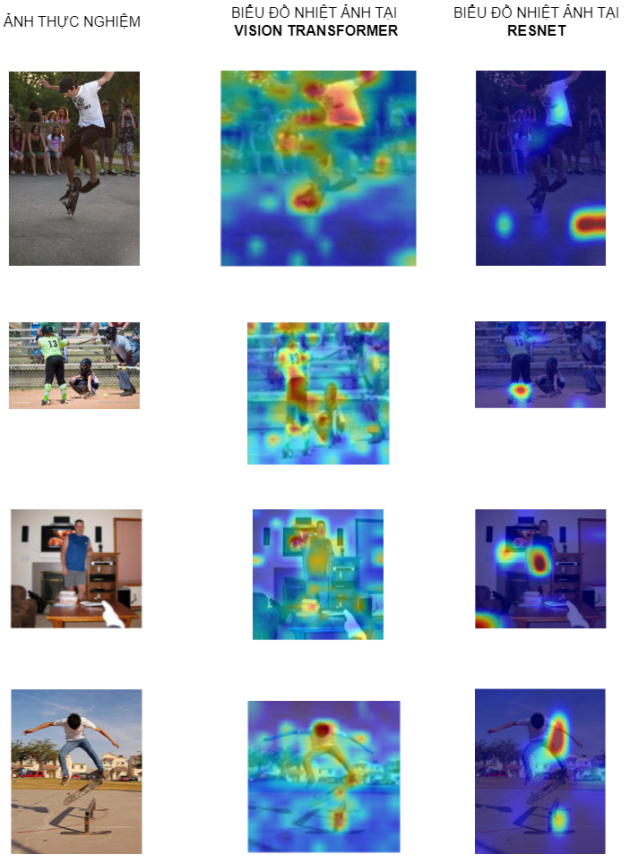
Nghiên cứu tập trung phân tích định tính để đưa ra những nhận định về hành vi của mô hình. Những hành vi này phụ thuộc vào hai thông tin đầu vào là hình ảnh và câu hỏi. Khi thực hiện thay đổi các đầu vào này, chúng tôi quan sát để tìm ra những khuôn mẫu nhất định, hoặc hạn chế của mô hình hỏi đáp hình ảnh. Do đó, chúng tôi thực hiện các thí nghiệm để quan sát sự thay đổi hành vi trả lời như sau:

1) Thực nghiệm trên câu hỏi gốc và hình ảnh gốc của tập dữ liệu.
2) Dựa trên câu hỏi gốc, tạo ra các câu hỏi làm thay đổi ý nghĩa của câu.
3) Dựa trên câu hỏi gốc, tạo ra các câu hỏi có ý nghĩa tương đương.
4) Dựa trên hình ảnh gốc, sử dụng phép biến đổi để làm dịch chuyển một phần thông tin của hình ảnh.

Dựa trên các thí nghiệm trên, nghiên cứu đã rút ra được một số nhận xét như sau:

a) Khối Transformer thị giác có xu hướng trích xuất đặc trưng toàn cục, trong khi khối ResNet chủ yếu tập trung vào việc trích xuất đặc trưng cục bộ.

Như được minh họa trong hình 6, hầu như toàn bộ đối tượng trong ảnh đều được khối *Vision Transformer* chú ý đến, nhất là con người và động vật. Trong khi đó, ở khối *ResNet* lại thể hiện sự tập trung vào một vùng nhất định trong ảnh.



Hình 6. Thực nghiệm trên câu hỏi gốc và hình ảnh gốc để tìm hiểu về xu hướng tập trung của mô hình. Những bức ảnh có nhiều vật thể, Vision Transformer thường tập trung vào toàn bộ vật thể tồn tại trong hình (tính toàn cục); trong khi đó, ResNet chỉ tập trung vào một vùng nhất định (tính cục bộ).

b) Mô hình không thực hiện suy luận tốt khi xử lý các hình ảnh chứa nhiều vật thể.

Chúng tôi phát hiện mô hình thường bỏ qua thông tin đến từ các giới từ như “trong”, “trên”, “dưới”,... vốn ám chỉ đến mối quan hệ giữa các chủ thể trong câu hỏi. Ví dụ ở hình 7, trong câu hỏi *Con gì nằm dưới bàn?*, giới từ *dưới* không được mô hình quá quan tâm ảnh hưởng đến kết quả đưa ra của mô hình là *laptop* - một vật thể nằm phía trên bàn - chứ không phải là *con chó* đang nằm dưới bàn.

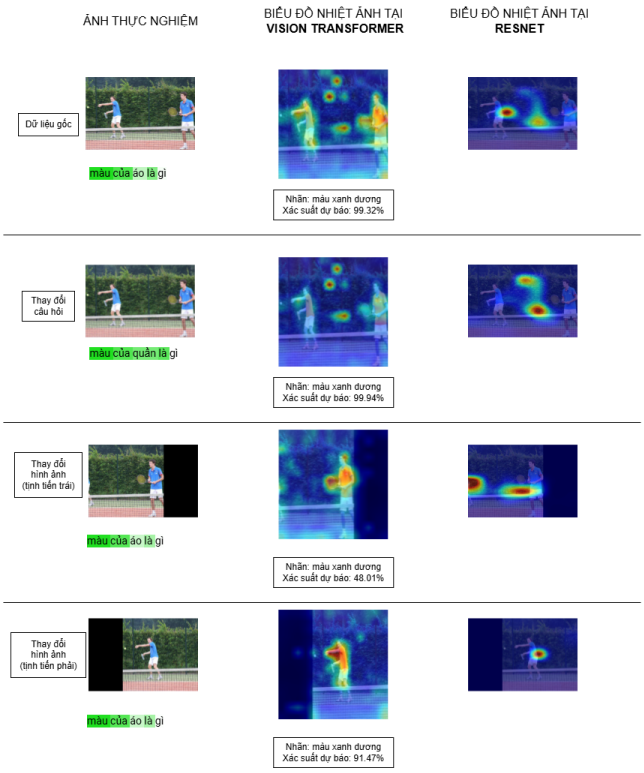
Một phát hiện khá thú vị nữa là mô hình dễ bị đánh lừa bởi những từ khóa gây nhiễu, điển hình như trong câu hỏi *Cậu bé đang ngồi xem cái gì?*, mô hình tập trung nhiều vào chữ *ngồi* mà bỏ qua chữ quan trọng hơn trong ngữ cảnh là *xem*, dẫn đến câu trả lời của mô hình là *cái ghế* với xác suất khá cao. Đây là trường hợp rất dễ gặp trong tiếng Việt khi mà một câu có cấu trúc ngữ pháp phức tạp và chứa nhiều từ "gây nhiễu".

Ngược lại, với những hình ảnh ít vật thể, mô hình trích xuất hợp lý hơn. Trong hình 8, câu hỏi *Màu của áo là gì?*, mô hình tập trung vào người đàn ông và hai chú chó trong

Hình 7. Thực nghiệm trên câu hỏi gốc và hình ảnh gốc (ảnh trên); câu hỏi được thay đổi khác ngữ nghĩa (ảnh dưới) để tìm hiểu về xu hướng tập trung trên câu hỏi và tác động của các chữ được tập trung. Ở ảnh trên, chữ *ngồi xem* được tập trung nhưng vì ngữ nghĩa phức tạp của từ này dẫn đến mô hình hiểu nhầm và cho rằng từ *ngồi xem* cùng nghĩa với từ *ngồi*. Ở ảnh dưới, chữ *dưới* ít được quan tâm dẫn đến mô hình tập trung sai vị trí.

Hình 8. Thực nghiệm trên câu hỏi gốc và hình ảnh gốc để xem xét về khả năng tập trung trên ảnh ít vật thể và ảnh nhiều vật thể.

ảnh đầu tiên; người mặc áo trắng đang trượt ván trong ảnh phía dưới. Mặc dù vậy, mô-đun ResNet đôi khi trích xuất đặc trưng không phù hợp trong ngữ cảnh, ví dụ như mô-đun này tập trung vào hai chú chó mặc dù tâm điểm chúng ta cần là người đàn ông.



Hình 9. Từ trên xuống dưới là các thực nghiệm: Dữ liệu gốc; Thay đổi câu hỏi; Tịnh tiến ảnh sang trái; Tịnh tiến ảnh sang phải.

c) Thông tin văn bản ít ảnh hưởng đến kết quả suy luận hơn so với thông tin thị giác.

Chúng tôi thực hiện thí nghiệm với đầu vào gồm bốn tình huống được minh họa ở hình 9:

1) Sử dụng hình ảnh và câu hỏi gốc.
2) Sử dụng hình ảnh gốc và thay đổi câu hỏi khác nghĩa với câu hỏi gốc.
3) Sử dụng câu hỏi gốc và tịnh tiến ảnh sang trái.
4) Sử dụng câu hỏi gốc và tịnh tiến ảnh sang phải.

So với biểu đồ nhiệt với đầu vào là hình ảnh gốc và câu hỏi gốc, thí nghiệm thay đổi câu hỏi cho thấy sự chuyển dịch vùng tập trung của biểu đồ nhiệt không sâu sắc bằng thí nghiệm thay đổi hình ảnh.

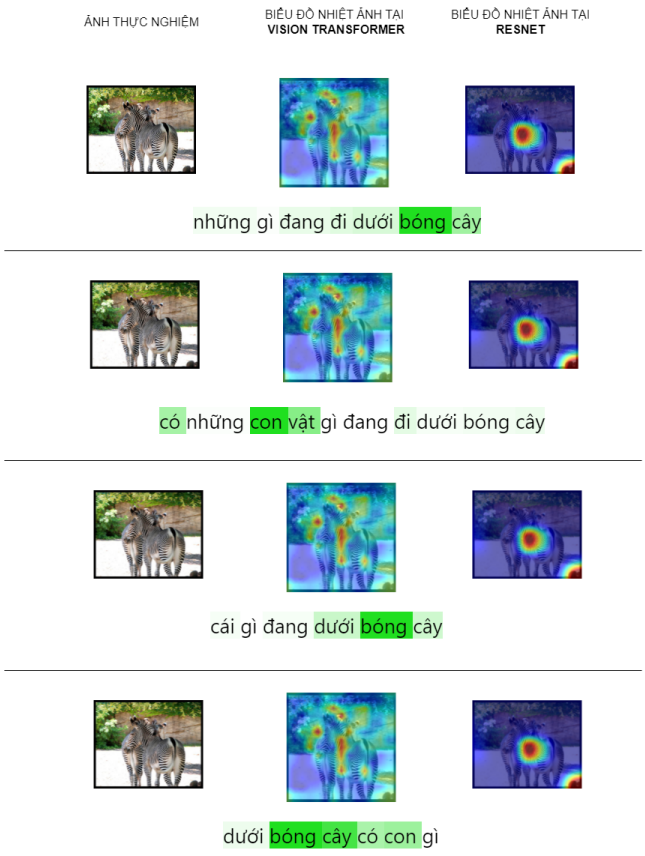
Cụ thể, đối với thí nghiệm thay đổi câu hỏi, biểu đồ nhiệt ảnh có sự thay đổi nhẹ ở độ đậm nhạt của vùng tập trung. Khi xem xét đến sự thay đổi của xác suất phân lớp trong thí nghiệm này so với câu hỏi gốc, kết quả cho thấy không có sự thay đổi đáng kể.

Sự suy giảm xác suất dự báo trong hai thí nghiệm tịnh tiến ảnh rõ rệt hơn thí nghiệm thay đổi câu hỏi. Bên cạnh đó, tịnh tiến ảnh sang trái cho thấy rõ rệt nhất. Ngoài ra, biểu đồ nhiệt ảnh trong thí nghiệm thay đổi câu hỏi chỉ thay đổi cường độ, tức là độ đậm nhạt, của vùng tập trung chứ không tạo thêm hay bỏ đi vùng nào.

Điều này cũng dễ hiểu vì hình ảnh là nguồn cung cấp ngữ cảnh trong tác vụ hỏi đáp hình ảnh. Nếu không trích xuất đặc trưng câu hỏi thì mô hình vẫn có thể dự đoán được với xác suất dự báo tương đối cao dựa vào vật thể trong ảnh; còn nếu không trích xuất đặc trưng hình ảnh thì sẽ gây khó khăn cho việc dự đoán của mô hình.

d) Khối PhoBERT thể hiện sự thiên vị nhất định trong quá trình suy luận.

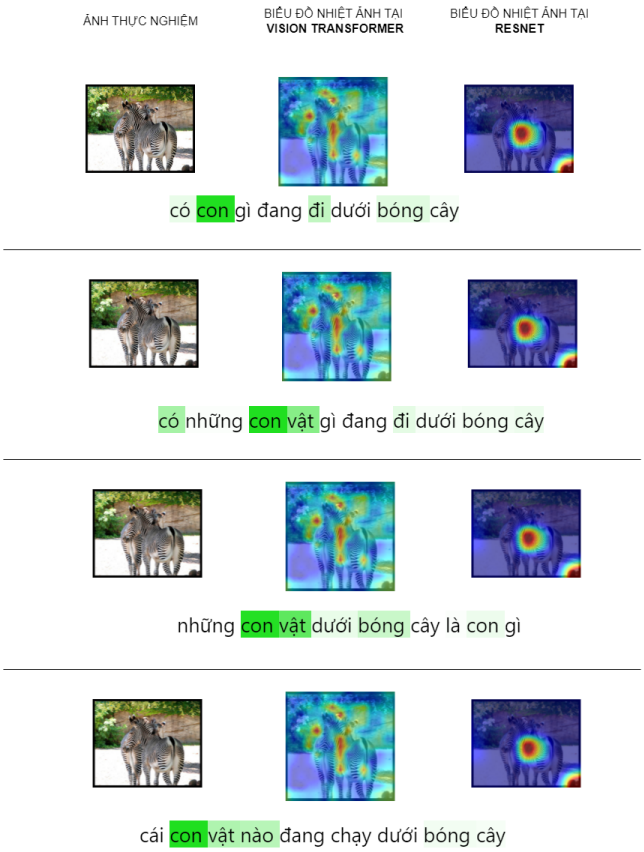
Trong thí nghiệm này, chúng tôi tập trung vào việc thay đổi câu hỏi như trong hình 10. Khi thay đổi câu hỏi thì hầu như biểu đồ nhiệt ảnh không thay đổi và các danh từ (bóng cây, con vật) được quan tâm nhiều hơn. Khi thay đổi sang dạng câu hỏi được thể hiện ở Hình 11 thì hầu như sự tập trung của mô hình dồn về chữ *con*. Từ đó cho thấy, khối PhoBERT khi được huấn luyện hoặc tinh chỉnh đã có thiên vị với một khuôn mẫu câu hỏi nhất định.



Hình 10. Thực nghiệm sử dụng một hình ảnh cùng với nhiều câu hỏi có cùng ngữ nghĩa để tìm hiểu sự tập trung của mô hình trên phương diện câu hỏi. Dễ nhận thấy một số danh từ như *bóng cây* hay *con vật* được tập trung nhiều.

V. KẾT LUẬN

Nghiên cứu đã triển khai các phương pháp diễn giải cho ba khối trong mô hình hỏi đáp hình ảnh, cung cấp các diễn



Hình 11. Thực nghiệm sử dụng một hình ảnh cùng với nhiều câu hỏi có cùng ngữ nghĩa có chữ *con* ở đầu câu để tìm hiểu xem mô hình có thiên vị với dạng câu hỏi này không. Với dạng câu chứa chữ *con* ở phía đầu câu này, khối PhoBERT luôn cho rằng chữ này quan trọng nhất.

giải trực quan cho phương thức hình ảnh và văn bản. Đồng thời, chúng tôi cũng đã thử nghiệm và khám phá ra một số đặc điểm trong hành vi của mô hình, đưa ra được một số điểm yếu mà mô hình thường gặp phải dẫn đến việc đưa ra những phán đoán nhầm lẫn.

Tuy nhiên, việc đưa ra biểu đồ nhiệt có thể chưa phản ánh hết tính chất của mô hình và bản thân của mô hình không có khả năng tự diễn giải và phản hồi để tự cải thiện. Nghiên cứu có thể phát triển tiếp để cung cấp khả năng diễn giải bằng ngôn ngữ tự nhiên, dễ hiểu hơn so với biểu đồ nhiệt. Ngoài ra, việc đánh giá độ hiệu quả của diễn giải vẫn còn là bài toán mở cần nhiều nghiên cứu hơn.

TÀI LIỆU THAM KHẢO

[1] A. Weller, *Transparency: Motivations and Challenges*. Cham: Springer International Publishing, 2019.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[3] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, 2020.
[4] T. Le, K. Pho, T. Bui, H. T. Nguyen, and M. L. Nguyen, "Object-less vision-language model on visual question classification for blind people," in *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pp. 180–187, SciTePress, Feb. 2022. ICAART 2022, February 3–5, 2022.
[5] A. D. Nguyen, T. Le, and H. T. Nguyen, "Combining multi-vision embedding in contextual attention for vietnamese visual question answering," in *Image and Video Technology*, (Cham), pp. 172–185, Springer International Publishing, 2023.
[6] Y. Lyu, P. P. Liang, Z. Deng, R. Salakhutdinov, and L.-P. Morency, "DIME: Fine-grained interpretations of multimodal models via disentangled local explanations," 2022. arXiv preprint.
[7] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788, 2018.
[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.
[9] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 4190–4197, Association for Computational Linguistics, July 2020.
[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, 2015.
[11] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and K. V. Nguyen, "Vivqa: Vietnamese visual question answering," in *Pacific Asia Conference on Language, Information and Computation*, 2021.

Tran Thi Phuong Linh received her B.S. degree in Computer Science from the University of Science, Ho Chi Minh City, Vietnam (VNU-HCM) in 2024. She research interests center on deep learning models to make their predictions more understandable and trustworthy.
Email: linh.ttphuong@fit.hcmus.edu.vn

Nguyen Thanh Tan received his B.S. degree in Software Technology in 2024 from the University of Science, Ho Chi Minh City, Vietnam (VNU-HCM). He is currently working as a Software Engineer in the industry. His research interests center on Deep Learning, with a particular focus on strategies for optimizing the cost and computational efficiency of Large Language Models (LLMs) for enterprise applications.
Email: tan.nt@fit.hcmus.edu.vn

Le Thanh Tung earned his B.S. degree in Computer Science from the Honour Program at the University of Science, Vietnam National University, Ho Chi Minh City, in 2012, and his M.S. degree in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2018. He completed his Ph.D. in Information Science at JAIST in 2021. Now, he is a Lecturer of Information Technology at the University of Science, Ho Chi Minh City, Vietnam (HCMUS). His research interests include information retrieval, natural language processing, visual question answering, and multi-modal systems
Email: tung.lt@fit.hcmus.edu.vn

Nguyen Tien Huy received his B.S. degree in Software Technology in 2010 and his M.S. degree in Computer Science in 2015, both from the University of Science, Ho Chi Minh City, Vietnam (VNU-HCM). He earned his Ph.D. in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2019. He is currently a lecturer in the Department of Computer Science, Faculty of Information Technology, VNU-HCM. His research interests include Natural Language Processing and Deep Learning, with a particular focus on intelligent systems and multimodal data analysis.
Email: ntienhuy@fit.hcmus.edu.vn