BÁO CÁO ĐỒ ÁN MÔN HỌC
Hỏi Đáp Dựa Trên Ảnh
Visual Question Answering

Trần Thị Mỹ Linh[1,2], Dương Thị Hồng Hạnh[1,2], and Nguyễn Trọng Ân[1,2,]
Đỗ Văn Tiến[1,2,]
18520999@gm.uit.edu.vn, 18520711@gm.uit.edu.vn, 18520434@gm.uit.edu.vn,
tiendv@uit.edu.vn

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Tóm tắt. Hỏi đáp dựa trên hình ảnh là một chủ đề được sự quan tâm lớn thu hút sự chú ý của các nhà nghiên cứu từ các lĩnh vực như xử lý ngôn ngữ tự nhiên và thị giác máy tính trong nhiều năm gần đây. Trong bài này, chúng em đã tạo ra bộ dữ liệu bao gồm hình ảnh cùng với câu hỏi và trả lời từ bộ dữ liệu COCO-QA và MSCOCO. Đồng thời, chúng em áp dụng các mô hình trong lĩnh vực VQA thu được các kết quả nghiên cứu trên các độ đo là Accuracy, WUPS 0.9 và 0.0. Tiếp đến là xây dựng ứng dụng cho bài toán sử dụng Flask dựa trên mô hình đã học được. Cuối cùng, chúng em thảo luận về các hướng đi có thể có trong tương lai cho bài toán và công trình nghiên cứu với các mô hình để cải thiện các kết quả và mang tới những ứng dụng thiết thực hơn nữa.

1 Giới thiệu

Hỏi đáp dựa trên hình ảnh (VisualQA) là một lĩnh vực khá mới mẻ, dần trở nên sôi nổi và đạt được những tiến bộ lớn trong những năm gần đây, hỏi đáp trên ảnh tự động cũng là một trong những lĩnh vực nghiên cứu tiềm năng với sự kết hợp của xử lý ngôn ngữ tự nhiên và thị giác máy tính. Việc đưa ra một hình ảnh cùng với câu hỏi về nó, một hệ thống hỏi đáp có thể trích xuất được các thông tin cơ bản về ảnh và trả lời các câu hỏi liên quan, công việc tưởng chừng là điều đơn giản đối với con người chúng ta nhưng lại là một thách thức lớn đối với máy tính.

Hỏi đáp trên ảnh có tính ứng dụng vô cùng lớn đối với các doanh nghiệp và trong thực tế, bởi lẽ các hệ thống hỏi đáp được giao nhiệm vụ phân tích dữ liệu và hỗ trợ cho doanh nghiệp đưa ra các quyết định một cách đúng đắng. Một ứng dụng khác là chúng ta có thể tích hợp các hệ thống trên ảnh tự động vào nên tảng Chatbot để có thể giải đáp thắc mắc, và tìm kiếm thông tin của con người. Hệ thống hỏi đáp là cần thiết cho nhiều tình huống trong thế giới thực, bao gồm hỗ trợ khách hàng, khuyến nghị, trả lời câu hỏi, đối thoại và quản lý quan hệ khách hàng. Nó có tiềm năng đáng kinh ngạc cho các tình huống như giúp những người khiếm thị có khả năng nhận thức được các thông tin quan trọng và có ích từ môi trường xung quanh.

Từ đó, nắm bắt được tầm quan trọng của các hệ thống máy hỏi đáp trên ảnh đối với con người. chúng em thực hiện trích xuất một bộ dữ liệu chứa hơn 20000 bộ ảnh

và các cặp câu hỏi, câu trả lời tương ứng từ miền dữ liệu các hình ảnh MS COCO và COCO-QA. Cùng với đó, chúng em bước đầu áp dụng triển khai mô hình VisualQA với các phương pháp đã được đề xuất trước đó như BiLSTM, CNN để đánh giá chất lượng bộ dữ liệu, cũng như trình bày về các tinh chỉnh trong các mô hình được triển khai, nhằm tìm ra mô hình cho kết quả tốt nhất với bộ dữ liệu của chúng này. .

Trong báo cáo đồ án này, chúng em tập trung vào giới thiệu các thông tin liên quan đến bài toán hỏi đáp tự động trên ảnh với bộ dữ liệu mà chúng em trích xuất ra được. Trong mục 2, chúng em sẽ trình bày một số công trình nghiên cứu liên quan. Tiếp theo ở mục 3, chúng em trình bày chi tiết về quá trình xây dựng bộ dữ liệu. Trong mục 4, các giải pháp, mô hình được chúng em trình bày và đồng thời, kết quả thử nghiệm sẽ được đánh giá, phân tích ở mục 5. Cuối cùng, mục 6 sẽ là kết luận và hướng phát triển trong tương lai cho các bài toán hỏi đáp nói chung và các bài toán hỏi đáp tự động trên ảnh nói riêng.

2 Công trình liên quan

Việc xây dựng một hệ thống có thể trả lời tự động các câu hỏi từ hình ảnh ngẫu nhiên được coi là một mục tiêu đầy tham vọng. Thời gian gần đây, cùng với sự phát triển của các phương pháp học máy hiện đại và việc ứng dụng hàng loạt các nghiên cứu liên quan đã dẫn đến những bước tiến lớn trong việc giải quyết bài toán hỏi đáp dựa trên hình ảnh có thể kể đến như: Sự ra đời của bộ dữ liệu VQA (Antol et al.) [1] với 614.163 câu hỏi và 7.984.199 câu trả lời cho 204.721 ảnh từ bộ Microsoft COCO là nền tảng cho sự phát triển của các hệ thống hỏi đáp trên ảnh hiện tại. Sau đó, các công trình nghiên cứu khác trên nhiều ngôn ngữ dựa trên bộ dữ liệu COCO ra đời như bộ dữ liệu COCO-QA sử dụng 123.287 ảnh từ bộ COCO, và bộ dữ liệu COCO-VQA sử dụng 204,728 ảnh.

Table 1: Một số bộ dữ liệu nổi tiếng trong lĩnh vực VQA.

Bộ dữ liệu	**Công bố**	**Nguồn ảnh**	**Số ảnh**	**Cặp câu QA**
DAQUAR [3]	2014	NYUDv2	1,449	12,468
COCO-QA [5]	2015	COCO	123,287	117,684
COCO-VQA [1]	2015	COCO	204,721	614,163
Visual7W [10]	2016	COCO	47,300	327,939
Visual genome [2]	2014	COCO, YFCC	108,000	1,773,258

3 Bộ dữ liệu

3.1 Mô tả bài toán

Mục tiêu chính của đồ án này là xây dựng hệ thống hỏi đáp tự động dựa trên ảnh. Thông tin chi tiết về nhiệm vụ của chúng em được mô tả bên dưới.

Đầu vào: Hình ảnh và câu hỏi liên quan đến nội dung của hình ảnh.

Đầu ra: Câu trả lời tương ứng cho câu hỏi đặt ra.

Một số ví dụ được trích xuất từ dữ liệu được trình bày trong Hình 1.

Fig. 1: Ví dụ cho bài toán Visual question answering

3.2 Dataset Analysis

Bộ dữ liệu chúng em sử dụng có chứa 23011 ảnh và 29800 cặp câu hỏi và câu trả lời tương ứng thuộc bốn loại là màu sắc, vị trí, đối tượng và số lượng. Sau đó tiến hành chia train, test theo tỉ lệ 9:1. Bảng bên dưới là thống kê dữ liệu:

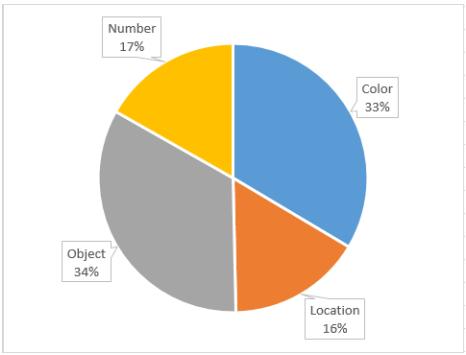

Fig. 2: Tỉ lệ các loại câu hỏi

Table 2: Thống kê số lượng theo các loại câu hỏi

Loại câu hỏi	Số lượng QA	Số lượng hình ảnh
Color	10000	8569
Location	4800	4163
Object	10000	9151
Number	5000	4299

4 Phương Pháp

Qua quá trình thực nghiệm, nhận thấy số lượng câu trả lời quá đa dạng, đặc biệt là dạng câu hỏi về object và location. Đồng thời, việc huấn luyện cùng lúc 4 loại câu hỏi khiến dữ liệu được học cho mỗi loại bị giảm đi ảnh hưởng đến việc học của mô hình. Do đó, ngoài việc huấn luyện chung 4 loại câu hỏi, chúng em tiến hành huấn luyện mô hình riêng cho từng loại câu hỏi, và tập trung tinh chỉnh với từng mô hình. Cuối cùng thông qua một mô hình phân loại câu hỏi để kết hợp 4 mô hình trên.

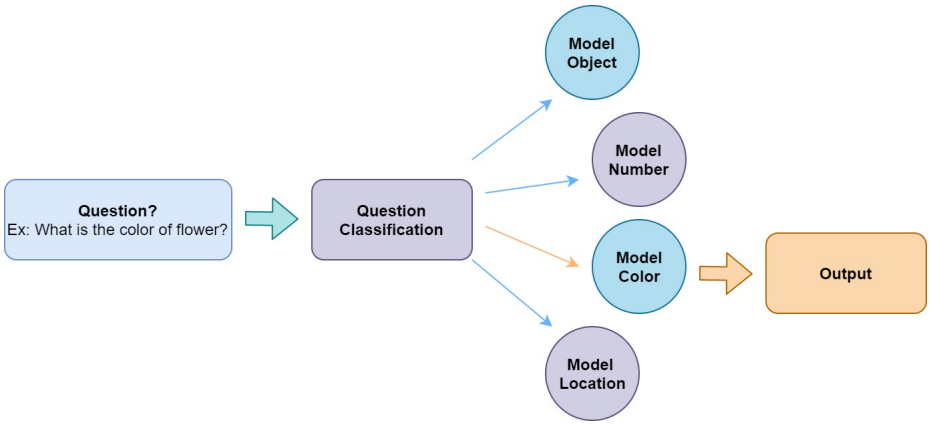

Fig. 3: Cách thức kết hợp mô hình cho bài toán

4.1 Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu là một bước quan trọng trong hầu hết các dự án Học máy hiện tại. Bằng cách thực hiện điều này, bộ dữ liệu sẽ được làm sạch và có khả năng trích xuất được nhiều thông tin hơn cho việc huấn luyện các mô hình, góp phần cải thiện các kết quả đạt được.

- Xử lý ảnh:
  - Định dạng lại kích thước toàn bộ các ảnh trong bộ dữ liệu vì sau nhiều lần huấn luyện mô hình chúng ta nhận thấy rằng việc huấn luyện với nhiều kích thước không đồng nhất sẽ dẫn đến độ chính xác của mô hình thấp. Chẳng hạn chúng em đã chuẩn hóa bộ dữ liệu ảnh với kích thước 64x64x3 cho các mô hình huấn luyện.
  - Thay vì sử dụng hàm img_to_array() để chuẩn hóa các ảnh về dạng mảng, tuy nhiên qua quá trình thực nghiệm, chúng em đã sử dụng preprocessing_input() được tích hợp trong efficientnet, và điều này góp phần cải thiện kết quả mô hình.
- Xử lý text:
  - Về phần dữ liệu dạng text, qua quá trình xem xét thì chúng em nhận thấy dữ liệu khá sạch, chỉ còn tồn tại một vài kí tự đặc biệt không gây ảnh hưởng đến mô hình, chữ đã được chuyển về dạng viết thường và việc loại bỏ stopword không mang lại kết quả tốt cho nên text sẽ giữ nguyên như ban đầu, sau đó

được đưa vào mô hình thông qua word embedding, điều này sẽ đượctrình bày ở phần tiếp theo.

- Ngoài ra, các câu trả lời được mã hóa thành các số tương ứng theo hàm to_categorical được tích hợp sẵn trong Keras.

4.2 Word embedding

– Word embedding là một kỹ thuật trong Xử lý ngôn ngữ tự nhiên (NLP), bằng cách ánh xạ các từ hoặc cụm từ từ nhóm từ vựng thành các vector số thực [4]. Nó giúp cải thiện độ chính xác của các mô hình ngôn ngữ tự nhiên khác nhau.
– Có nhiều kỹ thuật embedding khác nhau, qua thực nghiệm, nhóm chúng em lựa chọn text_to_sequences thuộc Tokenizer, lúc này mỗi từ trong câu sẽ được mã hóa thành một số duy nhất dựa theo tập từ của dữ liệu đầu vào.

4.3 Mô hình

Để giải quyết bài toán đặt ra, chúng em thực hiện xây dựng hai loại mô hình rút trích thông tin tương ứng cho từng loại đầu vào. Cụ thể như sau:

– Hình ảnh: sử dụng các kiến trúc mạng Convolutional Neural Network (CNN), mà cụ thể ở đây là VGG16 và MobileNetV2.
– Câu hỏi: sử dụng kiến trúc mạng Recurrent neural Network (RNN), cụ thể là sử dụng BiLSTM.

Sau cùng thực hiện kết hợp các thông tin rút trích được từ hai dữ liệu đầu vào để dự đoán ra câu trả lời tương ứng.

Quá trình trên được thể hiện trực quan thông qua hình bên dưới.

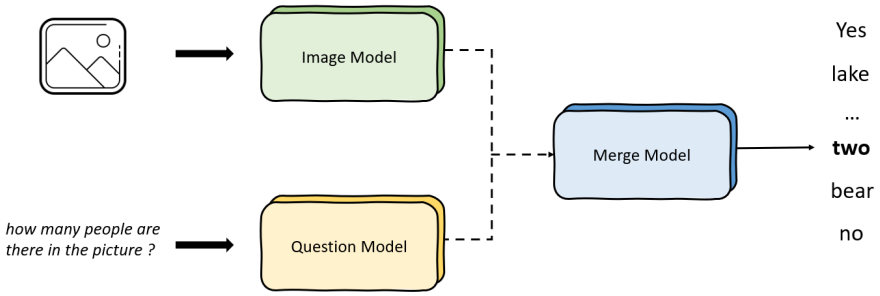

Fig. 4: Quy trình xây dựng mô hình để giải quyết bài toán

4.3.1 Trích xuất thuộc tính hình ảnh

1. VGG16
   VGG16 (hay còn gọi là OxfordNet) là một kiến trúc mạng CNN được đề xuất bởi Simonyan và Zisserman vào năm 2014 [8] và được xem là một trong những kiến trúc mô hình thị giác xuất sắc. Tại ILSVRC năm 2014 (The ImageNet Large Scale Visual Recognition Challenge), VGG16 là một trong những mô hình thuộc top đầu với độ chính xác 92.7% trên dữ liệu ImageNet gồm 14 triệu bức ảnh với 1000 phân lớp.
   VGG16 là một biến thể sâu hơn nhưng lại đơn giản hơn so với kiến trúc convolution (convolutional structure) thường thấy ở CNN, giúp mang lại độ chính xác cao. Thuật toán này được thiết kế với nguyên tắc gồm 2 hoặc 3 lớp layers Convolution (Conv) nối tiếp 1 layer MaxPooling 2D thay vì sử dụng các lớp Conv_Maxpooling liên tục như ở người tiền nhiệm LeNet, AlexNet. Nhờ đó mà VGG16 có thể giữ được nhiều thuộc tính của ảnh hơn, xử lý đặc biệt tốt với các bài toán nhận dạng đối tượng và đây cũng chính là lý do mà chúng em quyết định áp dụng mô hình này vào bài toán đặt ra.
   Dưới đây là hình ảnh thể hiện sơ lược về kiến trúc mạng VGG16:

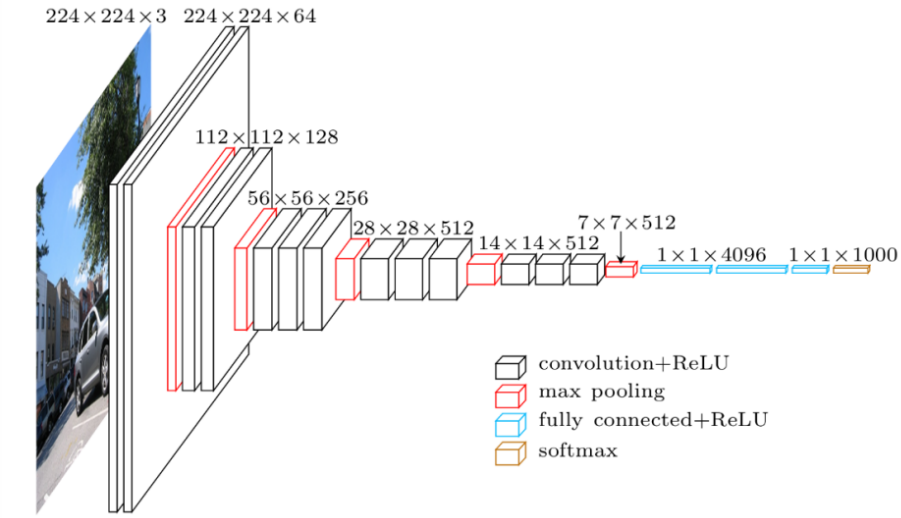

Fig. 5: Cách thức kết hợp mô hình cho bài toán

   Trong đó:
   – Gồm 16 layer :
     - 13 layer: 2 layer conv-conv-maxpooling, 3 layer conv-conv-conv-maxpooling.
     - 3 fully connection layer.
     - Có tổng cộng 138 triệu tham số.
     - Luôn sử dụng kernel 3x3 cho Convolution.
     - Maxpooling luôn có kích thước 2x2.
2. MobileNetV2:
   MobileNetV2 là 1 dạng kiến trúc CNN, đề xuất bởi Mark Sandler và cộng sự [6], được ưa chuộng bởi khả năng đảm bảo độ chính xác mô hình trong khi số lượng

tham số và số lượng các phép tính được tối ưu. Đây là 1 dạng kiến trúc inverted residual block, ở giữa các layer trong một block inverted residual block chúng ta cũng sử dụng những biến đổi tích chập tách biệt chiều sâu để giảm thiểu số lượng tham số của mô hình. Đồng thời, loại bỏ hàm phi tuyến tại layer input và output và thay bằng các phép chiếu tuyến tính.
Ngoài ra, khi Mark Sandler thực hiện huấn luyện trên bộ dữ liệu COCO (dữ liệu gốc của đồ án), MobileNetV2 có ít hơn 20 lần số lượng các phép tính và ít hơn 10 lần tham số so với YOLOv2. Do đó, trong phần này, chúng em quyết định áp dụng kỹ thuật transfer learning với mạng MobileNetV2 để trích xuất các thông tin từ ảnh đầu vào.

4.3.2 Trích xuất thuộc tính câu hỏi

- Bidirectional-Long Short Term Memory (Bi-LSTM) [7] là một biến thể hai chiều của Deeper Long Short Term Memory (LSTM) cho phép mô hình thực hiện học từ đầu vào hai lần theo hai hướng khác nhau.
- Cụ thể mô hình LSTM được đề xuất bởi S Antol et al. [1], theo tác giả thì mô hình này ra đời với mục đích giúp cho việc đào tạo và inference của mô hình nhanh hơn so với các mô hình trước đây. Cấu trúc mô hình bao gồm có hai phần là multi-layer perceptron (MLP) và mô hình LSTM dựa trên một softmax layer để tạo câu trả lời. Mô hình LSTM sử dụng one-hot encoding để mã hóa cho các từ câu trong câu hỏi, sau đó là phép biến đổi tuyến tính các đặc điểm của hình ảnh về kích thước phù hợp với bộ mã hóa LSTM của câu hỏi. Câu hỏi và hình ảnh được mã hóa dựa trên phép nhân ma trận (element- wise multiplication). Kiến trúc mạng của LSTM bao gồm các ô nhớ và cổng cho phép lưu trữ hoặc truy xuất thông tin.
- BiLSTM có thể được đào tạo bằng cách sử dụng tất cả thông tin đầu vào có sẵn trong quá khứ và trong tương lai đối với một khung thời gian đã chọn. Phương pháp nàycho phép tìm hiểu thêm thông tin ngữ cảnh được trích xuất từ hai hướng, do đó mạnh mẽ trong nhiều vấn đề khác nhau và hầu hết nó đều đạt được kết quả hiệu suất cao. Vì vậy, trong nhiệm vụ này, chúng em dự định chọn nó để thực nghiệm.

5 Thực nghiệm và Kết quả

5.1 Thực nghiệm

Các mô hình đều được huấn luyện trên tập huấn luyện và được đánh giá trên tập kiểm thử. Trong bài này, chúng em đánh giá các thí nghiệm thông qua hai mô hình đã chọn là VGG16 + BiLSTM, MobileNetV2 + BiLSTM:

• VGG16 + BiLSTM: mô hình được huấn luyện với 30 epochs, steps_per_epoch=100, validation_steps=100, callbacks được gọi với hàm EarlyStopping(monitor='val_loss', patience=3). VGG16 giữ nguyên kiến trúc ban đầu, tuy nhiên có thay lớp fully connected cuối cùng thành Dense(1024, activation='relu'). BiLSTM cũng được thêm lớp fully connected cuối cùng là Dense(1024, activation='tanh').

• MobileNetV2 + BiLSTM: mô hình được huấn luyện với 30 epochs, steps_per_epoch =100, validation_steps=100, callbacks được gọi với hàm EarlyStopping(monitor='val_loss', patience=3). MobileNetV2 giữ nguyên kiến trúc ban đầu, tuy nhiên có thay lớp fully connected cuối cùng thành Dense(1024, activation='relu'). BiLSTM được cài đặt tương tự như trên.

• Cuối cùng đầu ra của 2 mô hình trên được trả về sau khi qua 2 lớp dense, 1 lớp với size là 1000, activation là 'tanh', lớp cuối cùng với activation là 'softmax' trả về xác suất mỗi câu trả lời trong tập câu trả lời ban đầu.

5.2 Độ đo đánh giá

Để đánh giá hiệu suất của mô hình, chúng em đã sử dụng Accuracy và phép đo Wu-Palmer similarity (WUPS) [9] để tính toán sự chính xác giữa ground truth answer and a predicted answer với câu hỏi tương ứng. WUPS tính toán sự giống nhau giữa hai từ dựa trên dãy con chung dài nhất của chúng trong cây phân loại. Nếu độ tương đồng giữa hai từ nhỏ hơn một ngưỡng nhất định thì câu trả lời sẽ được coi là câu trả lời sai (0 điểm). Theo Malinowski và Fritz [3], chúng em đánh giá tất cả các mô hình dựa trên Accuracy, WUPS 0.9 và WUPS 0.0.

5.3 Kết quả thực nghiệm

Table 3: Kết quả thực nghiệm mô hình trên bộ dữ liệu

Model	Type QA	Acc	WUPS 0.9	WUPS 0.0
VGG16+BiLSTM	Color	**0.5569**	0.6240	0.9286
	Object	**0.3206**	0.4446	0.7726
	Number	0.5280	0.7501	0.9469
	Location	**0.3767**	0.4806	0.7984
	All	**0.3970**	0.8237	0.9411
MobileNetV2+BiLSTM	Color	0.4246	0.5104	0.9085
	Object	0.2825	0.4015	0.7535
	Number	**0.6620**	0.8171	0.9618
	Location	0.2900	0.6388	0.8631
	All	0.3232	0.4448	0.8159

▷ Nhận xét:

- Thực hiện huấn luyện cùng lúc 4 loại câu hỏi mang lại kết quả khá thấp, tốt nhất chỉ đạt 39.7% với mô hình VGG16+BiLSTM.
- Mô hình tốt nhất cho từng loại câu hỏi:
  - Color: VGG16+BiLSTM với độ chính xác 55.96%.
  - Object: VGG16+BiLSTM với độ chính xác 32.06%.
  - Number: MobileNetV2+BiLSTM với độ chính xác 66.2%.
  - Location: VGG16+BiLSTM với độ chính xác 39.7%.
- Sau khi huấn luyện riêng từng mô hình, có thể thấy các mô hình Color và Number đạt kết quả cao hơn hẳn 2 mô hình Location và Object, có thể giải thích điều này là do Location và Object có miền câu hỏi và trả lời đa dạng hơn, trong khi dữ liệu huấn luyện còn giới hạn.
- VGG16+BiLSTM hầu hết cho kết quả tốt trên các loại câu hỏi, tuy nhiên MobileNetV2+BiLSTM lại cho kết quả tốt hơn hẳn ở loại câu hỏi về Number, lên đến 66.2%.

– Khi quan tâm đến độ đo WUPS 0.9 và 0.0 của các mô hình, mặc dù các câu trả lời có thể không hoàn toàn chính xác, nhưng giữa dự đoán và thực tế có sự tương đồng khá cao.

6 Kết luận và Hướng phát triển

Hỏi đáp dựa trên hình ảnh là sự kết hợp giữa xử lý ngôn ngữ tự nhiên và thị giác máy tính hứa hẹn sẽ đem đến nhiều ứng dụng thực tế mà đặc biệt trong trí tuệ nhân tạo. Một lĩnh vực còn nhiều thách thức cũng như hứa hẹn nhiều tiềm năng khai thác về các phương pháp nghiên cứu và thực nghiệm.

Ở đồ án này, với bộ dữ liệu khá nhỏ do hạn chế về mặt tài nguyên, những kết quả thu được sau quá trình thực nhiệm chưa thật sự tốt để có thể so sánh với những kết quả được công bố trước đây.

Để có những cải tiến trong tương lai, chúng em dự định tiến hành phân tích và cải thiện chất lượng của các cặp câu hỏi và trả lời đồng thời tăng kích thước, độ đa dạng của tập dữ liệu khi có thể. Tiếp tục thực hiện tinh các chỉnh thông số mô hình để tìm kiếm mô hình mang lại hiệu suất cao hơn, cũng như tìm hiểu và thực hiện đề tài này trên Tiếng Việt.

7 Phân công và Demo

– Bảng phân chia nhiệm vụ các thành viên trong nhóm có thể xem tại đây.
– Link video demo về đề tài có thể xem tại đây.
– Xem chi tiết source code tại đây: Github.

References

1. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
2. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
3. Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *arXiv preprint arXiv:1410.0210*, 2014.
4. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
5. Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *arXiv preprint arXiv:1505.02074*, 2015.
6. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

7. Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
8. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
9. Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*, 1994.
10. Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.