# CHÚ THÍCH HÌNH ẢNH BẰNG TIẾNG VIỆT VỚI BỘ DỮ LIỆU UIT-ViIC

Chi-Thanh Dang[1,2*], Thuy-Hong T.Dang[1,2†] and Tien-Duong Pham[1,2†]

[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam.
[2]Vietnam National University, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): 20520761@gm.uit.edu.vn;
Contributing authors: 20520523@gm.uit.edu.vn;
20521222@gm.uit.edu.vn;
†These authors contributed equally to this work.

**Abstract**

Với sự thành công của các mô hình học sâu trong nhiều lĩnh vực, Image Captioning trở thành một bài toán thu hút nhiều sự quan tâm của các nhà nghiên cứu trên khắp thế giới. Trong nghiên cứu này, chúng tôi giải quyết bài toán Image Captioning bằng cách thử nghiệm nhiều nhóm mô hình học sâu trên bộ dữ liệu UIT-ViIC bao gồm các hình ảnh được chú thích tiếng việt. Bên cạnh đó, chúng tôi kết hợp giải quyết bài toán Machine Translation dịch những chú thích của ảnh vừa được trích xuất. Sau đó, chúng tôi tiến hành đánh giá các nhóm mô hình đã thử nghiệm và nhận được kết quả khá khả quan. Từ đó, chúng tôi tiếp tục phát triển bộ dữ liệu cũng như các mô hình phục vụ cho các nghiên cứu trong tương lai.

**Keywords:** Image Captioning, Machine Translation, Deep Neural Network

## 1 Introduction

Image Captioning, bài toán tự động tạo ra các chú thích cho hình ảnh, đã thu hút sự quan tâm của các nhà nghiên cứu trong nhiều lĩnh vực khoa học máy tính, thị giác máy tính, xử lý ngôn ngữ tự nhiên và học máy trong những năm

gần đây. Bài toán này có thể giúp con người xử lý các công việc thực tế, chẳng hạn như xây dựng hệ thống mô tả tự động cho người khiếm thị hoặc xây dựng các tính năng thông minh cho các ứng dụng trợ lý ảo và mạng xã hội.

Để đáp ứng nhu cầu của cộng đồng Xử lý Ngôn ngữ Tự nhiên trên thế giới, đã có rất nhiều bộ dữ liệu về chủ đề này ra đời kể từ năm 2013 trở đi. Một trong số đó có thể kể đến như bộ dữ liệu Flickr8k với 8092 hình ảnh do Hodosh và cộng sự xuất bản vào năm 2013[1], bộ dữ liệu Flickr30k với hơn 30000 hình ảnh do Young và cộng sự xuất bản năm 2014[2], đến năm 2015 Chen và cộng sự đã cho ra đời bộ dữ liệu Microsoft COCO với 123000 hình ảnh[3]. Như vậy, đến năm 2015 thế giới đã có 3 bộ dữ liệu lớn chất lượng phục vụ cho bài toán Image Captioning và các bộ dữ liệu lớn này cũng đã mở đường cho sự ra đời của các mô hình học sâu trong lĩnh vực này. Tuy nhiên, hầu hết các bộ dữ liệu này đều có chú thích mô tả hình ảnh bằng tiếng Anh và không có phần nào được sử dụng bằng tiếng Việt.

Cho đến năm 2020, sự ra đời của bộ dữ liệu UIT-ViIC do Lâm và cộng sự[4] xuất bản đã tạo ra một bước tiến mới trong việc xây dựng và đánh giá bài toán với phụ đề hình ảnh tiếng Việt và UIT-ViIC là bộ dữ liệu chúng tôi sử dụng để thực hiện đề tài của mình. Trong đề tài lần này, chúng tôi sử dụng một số mô hình học sâu để trích xuất chú thích hình ảnh, chúng tôi sẽ nói rõ hơn về những mô hình ở phần sau.

## 2 Bộ dữ liệu

Trong nghiên cứu này, chúng tôi tiến hành đánh giá trên bộ dữ liệu UIT-ViIC, bộ dữ liệu này bao gồm 3.850 hình ảnh liên quan đến các môn thể thao chơi bóng từ phiên bản 2017 của bộ dữ liệu Microsoft COCO. Tương tự như Microsoft COCO, nó cung cấp 5 chú thích tiếng Việt cho mỗi hình ảnh, tổng cộng có tới 19.250 chú thích. Đáng chú ý, những chú thích này được gán bằng tay, với tham chiếu từ chú thích gốc của các hình ảnh tương ứng trên MSCOCO. Mặc dù, bộ dữ liệu UIT-ViIC có một số hạn chế như tổng số hình ảnh so với nguồn gốc của MSCOCO. Tuy nhiên, nó vẫn được coi là một bộ dữ liệu chất lượng vì tính phức tạp trong việc mô tả hành vi của những người năng động, đặc biệt là những người chơi thể thao bằng tiếng việt phong phú.

## 3 Phương pháp thực hiện

Trong đề tài này, chúng tôi thực hiện trích xuất chú thích của hình ảnh trên bộ dữ liệu UIT-VilC, sau đó chúng tôi tiến hành dịch chú thích đó sang tiếng Anh.
Ở bài toán **Image Captioning** (trích xuất chú thích hình ảnh), chúng tôi tiến hành:

- Tiền xử lý dữ liệu: Làm sạch dữ liệu bằng cách loại bỏ kí tự số, kí tự đặc biệt, chỉ giữ lại chữ cái từ a - z; Thêm kí tự (token) startseq, endseq; Mapping dữ liệu: nhóm dữ liệu lại thành dạng dictionary, json; Chia bộ dữ liệu thành tập Train, Val, Test

- Tokenize dữ liệu: tạo từ điển dữ liệu (dữ liệu ở đây gồm các từ), mỗi từ được đánh 1 số theo thứ tự. Tổng số lượng các từ là 1351 từ.
- Load các mô hình đã train sẵn nhằm trích xuất đặc trưng của ảnh (còn gọi là pre-trained model, kĩ thuật này là Transfer Learning). Các model trích xuất đặc trưng [5] gồm VGG-16, Inception-V3, ResNet-50, VGG-19, EfficientNetV2L, DenseNet-201, Inception-ResNet-V2.
- Tạo mô hình Data generator (mô hình sản sinh dữ liệu). Tạo các mô hình CNN-LSTM và CNN-GRU[6] và bắt đầu huấn luyện dựa trên data generator và các đặc trưng đã trích xuất.
- Cuối cùng, cho mô hình dự đoán trên tập test.

Ở phần **Machine Translation** (dịch máy), chúng tôi tiến hành:

- Parsing data: đưa các câu tiếng anh cần dịch về dạng: [start] this is sentence [end].
- Vector hóa kí tự sử dụng Keras TextVectorization layer.
- Xây dựng mô hình Sequence-to-sequence learning với GRU và LSTM
- Và cuối cùng, dịch một câu chú thích mới vừa được trích xuất với RNN encoder and decoder.

## 3.1 Long short-term memory (LSTM)

Được giới thiệu lần đầu tiên bởi Hochreiter và Schmidhuber đã tạo nên nhiều cải tiến cho các mô hình mạng RNN truyền thống. Khi sử dụng LSTM, cấu tạo các cổng (gate) sẽ có thể được điều chỉnh khác nhau tùy mục đích để có được các kết quả thử nghiệm khác nhau với mục tiêu xây dựng được mô hình tốt nhất.

Cấu tạo của LSTM cơ bản gồm có các cổng là update gate, forget gate và output gate với các chức năng tương ứng: chọn lọc những thông tin được thêm vào bộ nhớ, loại bỏ thông tin không cần thiết, xác định những thông tin được chọn làm đầu ra.

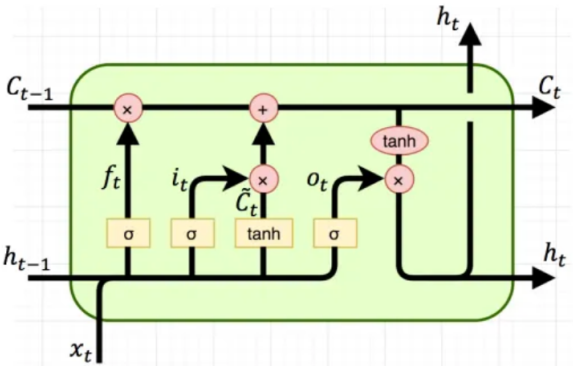


**Figure 1** Cấu trúc mạng LSTM

## 3.2 Convolutional Neural Networks (CNN)

CNN là một trong những mô hình Deep Learning phổ biến nhất và có ảnh hưởng nhiều nhất trong cộng đồng Computer Vision. CNN được dùng trong trong nhiều bài toán như nhận dạng ảnh, phân tích video, ảnh MRI, hoặc cho bài các bài của lĩnh vự xử lý ngôn ngữ tự nhiên,và hầu hết đều giải quyết tốt các bài toán này.

Mạng CNN là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số trong các node. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo.

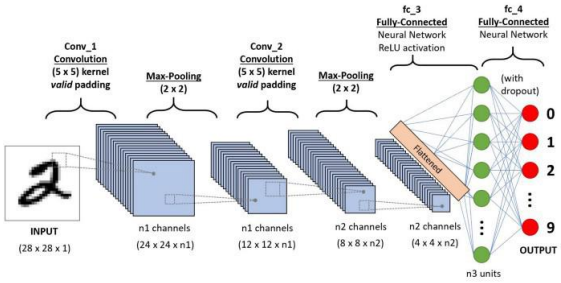


**Figure 2** Cấu trúc mạng CNN

## 3.3 Gated recurrent unit (GRU)

Là một kiến trúc tương tự như LSTM được giới thiệu để khắc phục những hạn chế của mạng RNN truyền thống. GRU cũng có cấu tạo là các cổng như LSTM tuy nhiên hạn chế hơn là chỉ có hai cổng là update gate và relevance gate vì vậy việc tính toán trong GRU cũng đơn giản hơn. Ở GRU có cổng relevance với mục đích tìm ra thông tin phía trước có quan trọng hay không để loại bỏ.

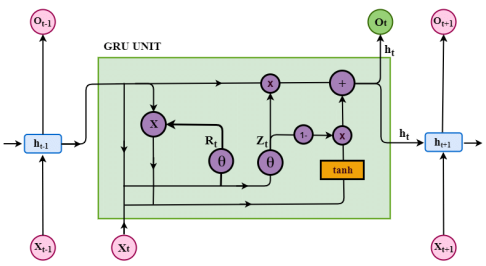


**Figure 3** Cấu trúc mạng GRU

## 3.4 Resnet-50

ResNet là viết tắt của Residual Network và là một loại mạng thần kinh tích chập 50 lớp (48 lớp tích chập, một lớp MaxPool và một lớp nhóm trung bình). Residual neural networks hình thành mạng bằng cách xếp chồng các khối lại với nhau, nhưng có một điểm khác biệt quan trọng. ResNet-50 lớp sử dụng thiết kế cổ chai cho khối xây dựng. Một khối còn lại của nút cổ chai sử dụng các tích chập 1×1, làm giảm số lượng tham số và phép nhân ma trận. Điều này cho phép đào tạo từng lớp nhanh hơn nhiều.

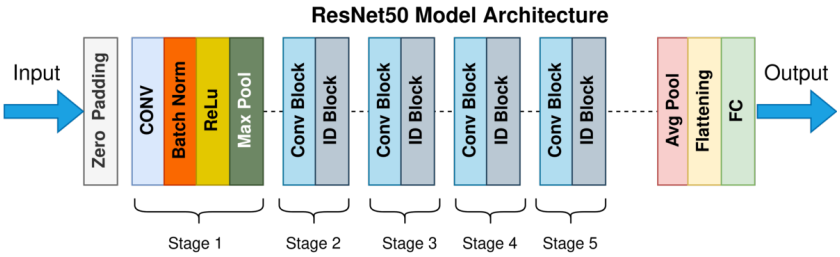


**Figure 4** Cấu trúc mạng Resnet-50

## 3.5 VGG-16 và VGG-19

VGG-16 là một kiến trúc CNN đã giành chiến thắng trong cuộc thi "2014 ILSVR (ImageNet)". Cấu hình mạng VGG-16 là hình ảnh 224 x 224 pixel với ba kênh (R, G và B). VGG-16 có 16 lớp, tuân theo sự sắp xếp gồm 13 lớp tích chập, 3 lớp fully-connected và các lớp gộp tối đa giúp giảm kích thước và chức năng kích hoạt softmax, tiếp theo là lớp fully-connected cuối cùng. Kiến trúc VGG-19 là một biến thể của mô hình VGG, bao gồm 16 mạng thần kinh tích chập, 3 lớp FC, 5 lớp MaxPool và 1 lớp SoftMax. Hình ảnh đầu vào có kích thước cố định là 224 x 224 pixel với ba kênh (R, G và B) có nghĩa là ma trận có hình dạng (224,224,3).

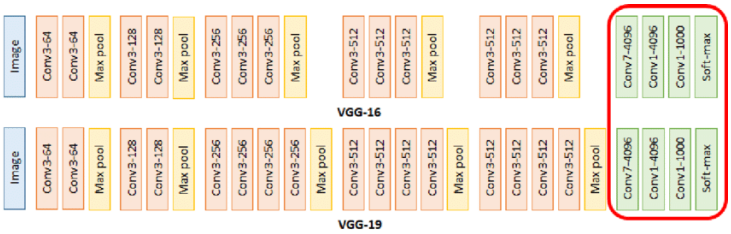


**Figure 5** Cấu trúc mạng VGG-16 và VGG-19

## 3.6 EfficientNetV2L

EfficientNetV2 là một loại mạng thần kinh tích chập có tốc độ đào tạo nhanh hơn và hiệu quả tham số tốt hơn so với các mô hình trước đó. Để phát triển các mô hình này, các tác giả sử dụng kết hợp training-aware neural architecture search và scaling, để tối ưu hóa tốc độ đào tạo.

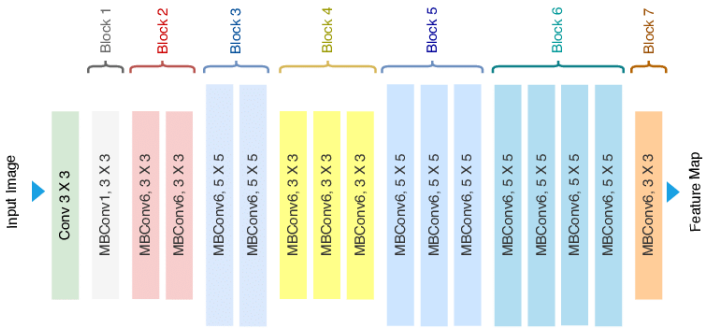


**Figure 6** Cấu trúc mạng EfficientNet

## 3.7 DenseNet-201

DenseNet là mạng tích chập được kết nối dày đặc. Nó rất giống với ResNet nhưng có một số khác biệt cơ bản. ResNet đang sử dụng một phương pháp bổ sung có nghĩa là lấy đầu ra lớp trước đó làm đầu vào cho lớp tiếp theo và trong khi DenseNet lấy tất cả đầu ra trước đó làm đầu vào cho lớp tiếp theo. Vì vậy, DenseNet được phát triển đặc biệt để cải thiện độ chính xác do vanishing gradient trong các mạng neural-network nhiều lớp dẫn đến khoảng cách xa giữa các lớp đầu vào, đầu ra làm cho thông tin biến mất trước khi đến đích.

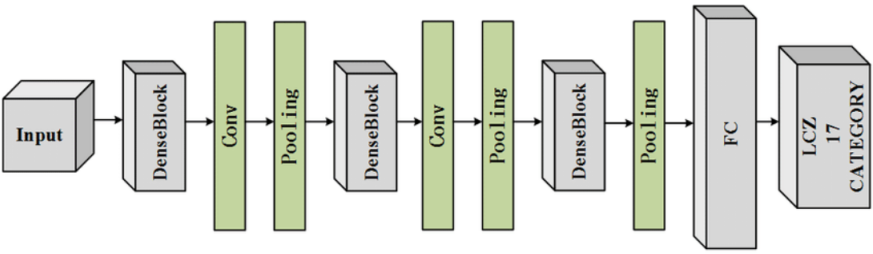


**Figure 7** Cấu trúc mạng DenseNet-201

## 3.8 Inception-V3

Inception-V3 là một mô hình học sâu dựa trên Convolutional Neural Networks, được sử dụng để phân loại hình ảnh. Inception-V3 Model có tổng cộng 42

lớp và tỷ lệ lỗi thấp hơn so với các mô hình trước đó. Inception-V3 là phiên bản tốt hơn của Inception-V1 được giới thiệu với tên gọi GoogLeNet vào năm 2014. Đúng như tên gọi, nó được phát triển bởi một nhóm tại Google.

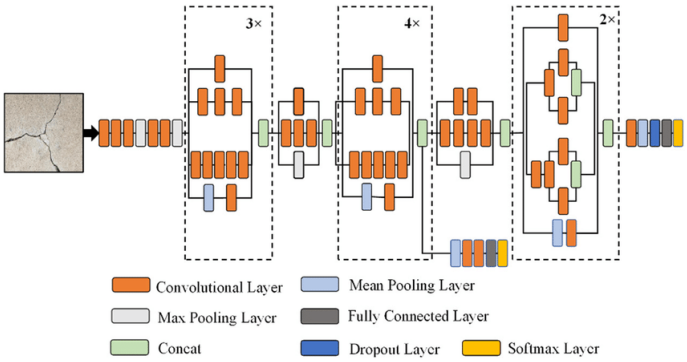


**Figure 8** Cấu trúc mạng Inception-V3

## 3.9 Inception-ResNet-V2

Inception-ResNet-V2 là mạng thần kinh tích chập được đào tạo trên hơn một triệu hình ảnh từ cơ sở dữ liệu ImageNet. Mạng có 164 lớp và có thể phân loại hình ảnh thành 1000 loại đối tượng, chẳng hạn như bàn phím, chuột, bút chì và nhiều loài động vật. Kết quả là, mạng đã học được các biểu diễn tính năng phong phú cho nhiều loại hình ảnh. Mạng có kích thước đầu vào hình ảnh là 299 x 299 và đầu ra là danh sách các xác suất ước tính của lớp.

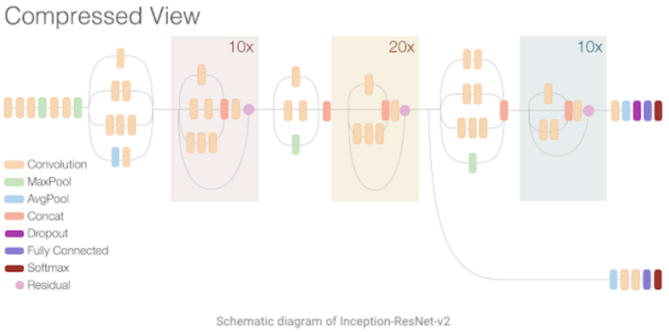


**Figure 9** Cấu trúc mạng Inception-Resnet-V2

# 4 Đánh giá

Để đánh giá các mô hình, chúng tôi sử dụng độ đo BLEU-Score. BLEU được thiết kế để sử dụng trong dịch máy (Machine Translation), đồng thời, phép đo này cũng được sử dụng trong các nhiệm vụ như tóm tắt văn bản, nhận dạng giọng nói, sinh nhãn ảnh (Image Captioning).

## 4.1 Image Captioning

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| LSTM – VGG16 | 0.73 | 0.58 | 0.48 | 0.39 |
| GRU – VGG16 | 0.71 | 0.57 | 0.47 | 0.38 |
| LSTM – InceptionV3 | 0.74 | 0.59 | 0.50 | 0.41 |
| GRU – InceptionV3 | 0.74 | 0.59 | 0.50 | 0.42 |
| LSTM – ResNet50 | 0.75 | 0.62 | 0.53 | 0.45 |
| GRU – ResNet50 | 0.75 | 0.61 | 0.52 | 0.44 |
| LSTM – EfficientNetV2L | 0.79 | 0.66 | 0.56 | 0.49 |
| GRU – EfficientNetV2L | 0.77 | 0.64 | 0.55 | 0.47 |
| LSTM – VGG19 | 0.73 | 0.59 | 0.49 | 0.41 |
| GRU – VGG19 | 0.72 | 0.58 | 0.48 | 0.40 |
| LSTM – DenseNet201 | 0.74 | 0.60 | 0.51 | 0.42 |
| GRU – DenseNet201 | 0.74 | 0.61 | 0.51 | 0.43 |
| LSTM – InceptionResNetV2 | 0.73 | 0.59 | 0.49 | 0.41 |
| GRU – InceptionResNetV2 | 0.74 | 0.60 | 0.50 | 0.42 |

**Table 1** Đánh giá mô hình Image Captioning

**Nhận xét:**

- LSTM – EfficientNetV2L và GRU – EfficientNetV2L là 2 nhóm mô hình thể hiện kết quả tốt nhất. Trong đó, LSTM – EfficientNetV2L có sự nhỉnh hơn đôi chút về các chỉ số.
- LSTM – VGG16 và GRU – VGG16 là 2 nhóm mô hình thể hiện kết quả không quá khả quan, trong đó, GRU – VGG16 có thông số thấp hơn.
- Qua đó thấy được, VGG-16 là pre-trained model không khả quan với bộ dữ liệu UITViIC, ngược lại, EfficientNetV2L, như được kỳ vọng, đã trích xuất được các đặc trưng quan trọng của hình ảnh, nhờ đó nâng cao độ hiệu quả của việc dự đoán câu mô tả.

## 4.2 Machine Translation

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| LSTM | 0.49 | 0.37 | 0.27 | 0.18 |
| GRU | 0.5 | 0.38 | 0.28 | 0.2 |
| Google API | 0.54 | 0.42 | 0.31 | 0.21 |

**Table 2** Đánh giá mô hình Machine Translation

**Nhận xét:**

- Ở phần này, chúng tôi tiến hành so sánh độ chính xác của các mô hình dịch máy so với mô hình Google Translation API.
- Khi tiến hành chạy các mô hình dịch máy, so sánh giữa mô hình GRU và LSTM, GRU có sự nhỉnh hơn đôi chút về các chỉ số so với LSTM.
- Khi so sánh các chỉ số giữa các mô hình học máy với mô hình Google Translation API, Google Translation AP đạt độ chính xác cao hơn.
- Chúng tôi nhận thấy mô hình Google Translation API là mô hình dịch tốt nhất trên bộ dữ liệu. Bên cạnh đó, chúng tôi cần phải cải thiện các mô hình học sâu để có thể thu được độ chính xác cao hơn.

# 5 Phân tích lỗi

## 5.1 Image Captioning

**Lỗi 1:** Số lượng dữ liệu về từng môn thể thao được phân bố không đều nhau, điển hình là môn bóng đá và bóng rổ. Điều này dẫn đến mô hình dự đoán chỉ dừng lại ở bước là nhận diện môn thể chứ chưa nhận diện dược hành động, bối cảnh trong bức hình. Lỗi được trực quan ở một số hình dưới đây:


**Figure 10** Một số ví dụ chú thích được trích xuất từ ảnh

**Lỗi 2:** Deep learning là thuật toán luôn "đói dữ liệu", tuy nhiên số lượng ảnh trong bộ dữ liệu UIT-ViIC vẫn còn khá ít (3850 ảnh), do đó, các mô hình trích xuất đặc trưng và mô hình dự đoán vẫn hoạt động chưa tốt, khiến một số câu mô tả được tạo ra không thực sự chính xác. Lỗi đươc trực quan ở một số hình bên dưới:


**Figure 11** Một số ví dụ chú thích được trích xuất từ ảnh

## 5.2 Machine Translation

Sau khi chạy mô hình và dịch trên tập Image Captioning. Chúng tôi nhận thấy có vài câu xuất hiện [UNK]. Giải thích cho điều này là bởi số lượng dữ liệu train không nhiều, dẫn đến tập từ vựng không đủ lớn, các từ vựng không có trong tập sau khi fit model và encode sẽ được chuyển thành [UNK] như bảng sau:

| Tiếng Anh | Tiếng Việt |
|---|---|
| the men are [UNK] to the frisbee | Những người đàn ông đang luyện tập ném dĩa |
| two children are playing on a [UNK] life | Hai đứa trẻ đang chơi trên chiếc phao |
| a [UNK] [UNK] is tennis player is riding a ball to his toy | Một nam vận động viên tennis đang vung vợt để đánh bóng |
| a [UNK] [UNK] is the tennis player to his ball | Một nam vận động viên tennis đang nhún chân để đỡ bóng |

**Table 3** Một số lỗi khi chạy mô hình dịch máy

# 6 Kết luận

Trong báo cáo này, chúng tôi đã sử dụng bộ dữ liệu UIT-ViIC, bộ dữ liệu này bao gồm 3.850 hình ảnh và 19.250 chú thích bằng tiếng Việt liên quan đến các môn thể thao chơi bóng từ phiên bản 2017 của bộ dữ liệu Microsoft COCO. Chúng tôi tiến hành chạy một số mô hình học sâu trên bộ dữ liệu để trích xuất chú thích cho hình ảnh và tiến hành dịch các chú thích đã trích xuất sang tiếng anh bằng các mô hình học sâu. Cuối cùng, chúng tôi sử dụng

độ đo BLEU để đánh giá các mô hình học sâu.

Nhìn chung, chúng ta có thể thấy các mô hình thể hiện khá tốt ở thang đo BLEU-1. Bên cạnh đó, EfficientNetV2L là mô hình trích xuất được các đặc trưng quan trọng của hình ảnh, nhờ đó nâng cao độ hiệu quả của việc dự đoán câu mô tả. Mặt khác, khi so sánh các chú thích được dịch bằng cách chạy các mô hình học sâu trên tập dữ liệu với chú thích dịch bởi mô hình Google Translation API, kết quả đánh giá và ví dụ đầu ra cho thấy rằng mô hình Google Translation API có thể hoạt động ở mức chấp nhận được, tuy nhiên một số phụ đề được dịch không hoàn toàn và sát nghĩa.

Trong tương lai, chúng tôi sẽ mở rộng bộ từ điển tiếng Việt để có thể giảm thiểu tối đa mức độ lỗi khi thử nghiệm các mô hình trên bộ dữ liệu. Bên cạnh đó, chúng tôi sẽ mở rộng bộ dữ liệu về số lượng, chủ đề và thử nghiệm các mô hình trên nhiều bộ dữ liệu để có thể cải thiện, cho ra kết quả tốt hơn nữa. Đặc biệt, chúng tôi sẽ tiếp tục hoàn thiện và phát triển bộ dữ liệu trên nhiều ngôn ngữ khác để tạo điều kiện phục vụ trong lĩnh vực Image Captioning. Với đề tài này, chúng tôi hy vọng sẽ cung cấp thêm cho cộng đồng nghiên cứu Việt Nam những đánh giá trong bài toán Image Captioning. Từ đó, bài toán này có thể được cải thiện và phát huy để sớm phục vụ các mục đích thiết thực tại Việt Nam.

## References

[1] Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013). https://doi.org/10.1613/jair.3994

[2] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014). https://doi.org/10.1162/tacl_a_00166

[3] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

[4] Lam, Q.H., Le, Q.D., Nguyen, V.K., Nguyen, N.L.-T.: Uit-viic: A dataset for the first evaluation on vietnamese image captioning, 730–742 (2020). https://doi.org/10.1007/978-3-030-63007-2_57

[5] Predić, B., Manić, D., Saračević, M., Karabašević, D., Stanujkić, D.: Automatic image caption generation based on some machine learning algorithms. Mathematical Problems in Engineering **2022** (2022)

[6] Wang, H., Zhang, Y., Yu, X.: An overview of image caption generation methods. Computational intelligence and neuroscience **2020** (2020)