HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN KHÁNH SƠN

ĐỀ CƯƠNG
LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI-2023

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN KHÁNH SƠN

SINH MÔ TẢ TIẾNG VIỆT CHO ẢNH SỬ DỤNG MÔ HÌNH MÃ HÓA - GIẢI MÃ

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 8.48.01.01

ĐỀ CƯƠNG LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. NGÔ XUÂN BÁCH

HÀ NỘI-2023.

I. MỞ ĐẦU

1. Lý do chọn đề tài:

Theo dòng chảy của cuộc cách mạng 4.0, trí tuệ nhân tạo ngày càng được phổ biến và ứng dụng rộng rãi trong mọi lĩnh vực của cuộc sống, mặc dù được John McCarthy – nhà khoa học máy tính người Mỹ đề cập lần đầu tiên vào những năm 1950 nhưng đến ngày nay thuật ngữ trí tuệ nhân tạo mới thực sự được biết đến rộng rãi và được các "ông lớn" của làng công nghệ chạy đua phát triển. Những năm gần đây, chúng ta đã thấy một bước nhảy vọt đáng kể trong cách Trí tuệ nhân tạo (AI) đang trở thành một phần không thể thiếu trong cuộc sống. Hành trình chuyển đổi số đã được khởi động và do tình hình đại dịch càng khiến chúng ta chứng kiến rõ sự đổi mới đáng kể trong lĩnh vực công nghệ mà trong đó AI đã giúp tạo ra rất nhiều bước tiến đột phá. Các ứng dụng của AI đang tạo ra một phần tác động lớn đối với trải nghiệm người dùng và tính thương mại hóa của các công ty. Đặc biệt hơn, các thuật toán AI ứng dụng Deep learning hiện nay đang được phát triển vô cùng mạnh mẽ và chóng mặt giúp giải quyết được nhiều bài toán lớn trong cuộc sống.

Trong đó, bài toán sinh mô tả cho ảnh hay chú thích hình ảnh (Image captioning) cũng dần thu hút sự quan tâm của nhiều nhà nghiên cứu trong lĩnh vực trí tuệ nhân tạo và trở thành một bài toán thú vị nhưng cũng gặp rất nhiều thách thức. Sinh mô tả ảnh giúp tự động tạo mô tả ngôn ngữ tự nhiên theo nội dung được quan sát trong hình ảnh, đây là kết hợp thú vị giữa 2 lĩnh vực lớn của AI là thị giác máy tính và xử lý ngôn ngữ tự nhiên.

Một trong những ứng dụng nổi bật của Image captioning là giúp những người bị các bệnh về mắt hay khiếm thị có thể nhận biết được môi trường xung quanh qua việc sinh mô tả ảnh sang dạng text và sau đó phát lên âm thanh thông báo cho họ. Ngoài ra nó còn được google search ứng dụng để tìm kiếm được hình ảnh dựa vào mô tả hay camera an ninh trong công nghệ tự động giám sát…

Tuy là ứng dụng của Image captioning là vô cùng phong phú, nhưng hiện nay có rất ít nghiên cứu về vấn đề này đối với tiếng Việt do nguồn dữ liệu còn hạn chế. Đa phần các nghiên cứu cho tiếng Việt hiện nay đều dựa trên các phương pháp sinh mô tả đạt hiệu quả cao cho tiếng Anh rồi triển khai nghiên cứu dành riêng cho tiếng Việt.

Với những lý do trên, học viên chọn đề tài ***"Sinh mô tả tiếng Việt cho ảnh sử dụng mô hình mã hóa - giải mã"*** làm luận văn tốt nghiệp cao học.

2. Tổng quan về vấn đề nghiên cứu:

- Giới thiệu bài toán: Với đầu vào là một bức ảnh bất kỳ, Sinh mô tả cho ảnh hay chú thích hình ảnh (Image captioning) là việc sinh ra các mô tả bằng ngôn ngữ tự nhiên theo nội dung quan sát được trong hình ảnh. Một bức ảnh có thể có nhiều mô tả và một mô tả tốt sẽ giúp nắm bắt được toàn bộ nội dung chính của toàn bức ảnh. Bài toán này chính là một thách thức trong việc hiểu ngữ cảnh và là sự kết hợp thú vị giữa hai lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên. Tổng quát, một bài toán sinh mô tả cho ảnh sẽ có đầu vào là một bức ảnh bất kỳ và đầu ra là một mô tả tương ứng với nội dung của bức ảnh. Chúng ta theo dõi ví dụ sau đây:

Các cầu thủ bóng rổ đang thi đấu trên sân.

Hình 1. 1. Ví dụ về bài toán sinh mô tả cho ảnh.

- Các giải pháp cho bài toán: Các thuật toán sinh mô tả cho ảnh được chia thành 2 phương pháp chính là phương pháp dựa trên phương pháp khuôn mẫu và phương pháp dựa trên kiến trúc bộ mã hóa- giải mã, trong phạm luận văn này, chúng ta đi sâu nghiên cứu vào phương pháp dựa trên kiến trúc bộ mã hóa – giải mã.

- Hầu hết các hệ thống chú thích hình ảnh hiện nay đều sử dụng kiến trúc bộ mã hóa- giải mã. Một mô hình sinh mô tả ảnh theo kiến trúc này bao gồm 3 thành phần chính sau: bộ mã hóa đặc trưng của ảnh (Image Feature Encoder), bộ giải mã trình tự (Sequence Decoder) và trình tạo câu (Sentence Generator).

    ● Bộ mã hóa đặc trưng của ảnh: đây là một kiến trúc mạng giúp trích rút những đặc trưng của hình ảnh. Các kiến trúc nổi bật nhất khi nói đến tác vụ này là CNN và các biến thể của mạng này.

    ● Bộ giải mã trình tự: tác dụng của khối này là sử dụng các đặc trưng ảnh đã được trích rút đem huấn luyện qua một mạng tuần tự giúp xuất ra mô tả cho ảnh. Các kiến trúc mạng sử dụng cho khối này đa dạng hơn so với khối mã hóa như RNN, LSTM, BiLSTM... Luận văn sử dụng 2 kiến trúc là LSTM thông thường và LSTM có sử dụng cơ chế chú ý để giải quyết bài toán.

    ● Trình tạo câu: đây là quá trình tạo ra mô tả cho ảnh theo ngôn ngữ yêu cầu. Nó diễn ra theo trình tự, các từ (token) trong mô tả được sinh ra phía sau sẽ dựa vào đặc trưng ảnh, các từ đã được sinh ra ở phía trước và một số quy tắc xác định trước

- Mô hình NIC (Neural Image Caption) sử dụng kiến trúc encoder- decoder cơ bản hay còn gọi là "Inject" giúp kết nối trực tiếp bộ mã hóa đặc trưng của ảnh CNN làm đầu vào cho bộ giải mã tuần tự LSTM, cuối cùng là bộ tạo câu giúp sinh ra mô tả như đã mô tả ở phần kiến trúc tổng quát.
- Mạng CNN [1],[4] (Mạng nơ ron tích chập) có kiến trúc khá giống với mạng nơ-ron truyền thống. Chúng được cấu tạo từ rất nhiều nơ-ron và có khả năng học thông qua tính toán chênh lệch và tối ưu tham số. Do đó cách học và huấn luyện không có nhiều thay đổi là sử dụng hàm mất mát để đánh giá độ chính xác của mô hình tại các lớp đầu ra cuối cùng. Điểm đặc biệt ở đây là dữ liệu đầu vào có thể là dạng ảnh, tức là dữ liệu dạng ma trận và áp dụng phép tích chập để trích xuất tính năng tự động từ ảnh đầu vào. Ngoài ra, CNN cũng hiệu quả về mặt tính toán. Nó sử dụng các phép toán tích hợp và thực hiện chia sẻ tham số. CNN có thể xử lý được dữ liệu ảnh mà vẫn giữ nguyên cấu trúc của chúng và không phải duỗi ảnh về mảng một chiều. Cụ

thể hơn các nơ-ron trong CNN được sắp xếp trong không gian 3 chiều: chiều cao, chiều rộng, chiều sâu và các lớp ẩn không kết nối với toàn bộ với các nơ-ron của lớp trước mà chỉ kết nối một phần. Nhìn chung, sau khi đi qua mỗi lớp, dữ liệu giảm đi về kích thước chiều rộng, chiều cao và tăng lên về chiều sâu. Có 3 lớp chính để xây dựng một mạng CNN: lớp tích chập (Convolutional layer), lớp pooling (Pooling layer), lớp kết nối toàn bộ (Fully- connected layer).

- Mạng LSTM (Long Short-term memory) [2] là một kiến trúc đặc biệt của RNN, LSTM được thiết kế để tránh được vấn đề phụ thuộc xa. Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó có thể ghi nhớ được mà không cần bất kỳ can thiệp nào. Vì vậy LSTM thường được áp dụng cho các đối tượng văn bản hoặc âm thanh và giải quyết được vấn đề phụ thuộc xa mà RNN không làm được. Tuy nhiên qua phân tích nguyên lý hoạt động ta thấy tốc độ nó cho ra kết quả thường sẽ chậm hơn mạng RNN.
- Mô hình CNN-LSTM trong luận văn sử dụng có kiến trúc 2 khối encoder và decoder tương tự như mô hình NIC. Điểm khác ở đây là đầu ra của 2 khối sẽ được kết hợp độc lập với nhau qua phép tính cộng đặc trưng

3. Mục đích nghiên cứu:

- Cung cấp các kiến thức, khái niệm làm nền tảng cho các phương pháp tiếp cận giải quyết bài toán dựa trên mô hình mã hóa - giải mã

- Khảo sát các kiến trúc mô hình khác nhau nhằm giải quyết bài toán sinh mô tả ảnh cho tiếng Việt.

- Thực nghiệm các kiến trúc mô hình sinh mô tả ảnh cho tập dữ liệu tiếng Việt. Từ đó đưa ra đánh giá, nhận xét về kết quả đạt được và đưa ra hướng cải tiến cho bài toán.

- Xây dựng hệ thống sinh mô tả ảnh cho tiếng Việt

4. Phạm vi nghiên cứu:

- Luận văn này sẽ tập trung vào khảo sát, nghiên cứu bài toán kết hợp giữa hai lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên là bài toán sinh mô tả ảnh. Những phương pháp, mô hình phù hợp sẽ được triển khai và áp dụng cho tập dữ liệu hình ảnh- mô tả tiếng Việt.
- Sinh mô tả cho ảnh là một bài toán rộng và chứa nhiều tiềm năng trong tương lai cho các hệ thống thông minh nhưng để phát triển ở hiện tại vẫn gặp khá nhiều khó khăn. Do đó, tập dữ liệu sử dụng sẽ chỉ bao quát thông tin ảnh- mô tả trong miền thể thao với mong muốn đạt được kết quả tốt và đưa ra những đánh giá trực quan về bài toán.

5. Phương pháp nghiên cứu:

- ***Về mặt lý thuyết:*** Thu thập, khảo sát, phân tích các tài liệu, bài báo và thông tin có liên quan tới bài toán sinh mô tả cho ảnh.

- ***Về mặt thực nghiệm:*** Thử nghiệm và đánh giá kết quả.

II. NỘI DUNG

Dự kiến cấu trúc nội dung chính của luận văn gồm 3 chương như sau:

CHƯƠNG 1. GIỚI THIỆU BÀI TOÁN SINH MÔ TẢ CHO ẢNH

- Trong chương 1, luận văn trình bày cái nhìn tổng quan về bài toán sinh mô tả cho ảnh, bao gồm: giới thiệu bài toán, ứng dụng, một số nghiên cứu liên quan cùng với phạm vi và đóng góp của luận văn
- Giới thiệu bài toán: Với đầu vào là một bức ảnh bất kỳ, Sinh mô tả cho ảnh hay chú thích hình ảnh (Image captioning) là việc sinh ra các mô tả bằng ngôn ngữ tự nhiên theo nội dung quan sát được trong hình ảnh. Một bức ảnh có thể có nhiều mô tả và một mô tả tốt sẽ giúp nắm bắt được toàn bộ nội dung chính của toàn bức ảnh. Bài toán này chính là một thách thức trong việc hiểu ngữ cảnh và là sự kết hợp thú vị giữa hai lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên. Tổng quát, một bài toán sinh mô tả cho ảnh sẽ có đầu vào là một bức ảnh bất kỳ và đầu ra là một mô tả tương ứng với nội dung của bức ảnh.
- Kết luận chương 1

CHƯƠNG 2. NGHIÊN CỨU SINH MÔ TẢ TIẾNG VIỆT CHO ẢNH SỬ DỤNG MÔ HÌNH MÃ HÓA - GIẢI MÃ

- Trong chương này, luận văn sẽ trình bày các kiến trúc mô hình phổ biến và một số kiến thức liên quan để giải quyết bài toán, bao gồm các phần sau: kiến trúc tổng quát, chi tiết về các mô hình NIC, CNN-LSTM được sử dụng:

  + Hầu hết các hệ thống chú thích hình ảnh hiện nay đều sử dụng kiến trúc bộ mã hóa- giải mã. Một mô hình sinh mô tả ảnh theo kiến trúc này bao gồm 3 thành phần chính sau: bộ mã hóa đặc trưng của ảnh (Image Feature Encoder), bộ giải mã trình tự (Sequence Decoder) và trình tạo câu (Sentence Generator).

  + Mô hình NIC (Neural Image Caption) sử dụng kiến trúc encoder- decoder

  + Mô hình CNN-LSTM trong luận văn sử dụng có kiến trúc 2 khối encoder và decoder tương tự như mô hình NIC. Điểm khác ở đây là đầu ra của 2 khối sẽ được kết hợp độc lập với nhau qua phép tính cộng đặc trưng

- Kết luận chương 2

CHƯƠNG 3. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

- Chương này trình bày về quy trình thực nghiệm bài toán sinh mô tả tiếng Việt cho ảnh với các mô hình đã khảo sát ở trên, bao gồm: trình bày về dữ liệu sử dụng, quy trình huấn luyện, thiết lập thực nghiệm, các chỉ số đánh giá cho bài toán và cuối cùng là đưa ra kết quả thực nghiệm và các đánh giá, phân tích. Các kết quả thực nghiệm trên các mô hình sẽ được thống kê và việc so sánh đánh giá các mô hình.

- Quy trình huấn luyện:

  + Tập dữ liệu sử dụng: UIT-ViIC [6] là tập dữ liệu cho bài toán sinh mô tả ảnh đầu tiên cho tiếng Việt được phát triển bởi nhóm nghiên cứu thuộc trường Đại học Công nghệ Thành phố Hồ Chí Minh (UIT). Bộ dữ liệu UIT-ViIC được tổ chức như sau:

    - 3 folder ảnh “train”, “dev”, “test” chứa lần lượt 2695, 924, 231 ảnh.

● 3 file json tương ứng với 3 tập ảnh. Nội dung trong mỗi file json bao gồm 2 phần: “images" và “annotations": Phần “images" chứa thông tin chi tiết của từng ảnh như id, file_name, height, width...; Phần “annotations" liên kết tới từng ảnh và chứa mô tả cho ảnh đó. Mỗi image sẽ có 5 annotation liên kết với nó.

+ Quy trình xử lý được xây dựng cho kiến trúc học sâu bao gồm 2 giai đoạn

● Đối với giai đoạn đầu tiên, các phương pháp học chuyển giao được xử dụng để xử lý trước các hình ảnh thô, ví dụ có thể sử dụng mạng CNN đã được đào tạo trước. Điều này lấy hình ảnh làm đầu vào và tạo ra các vector hình ảnh được mã hóa để nắm bắt các tính năng thiết yếu của hình ảnh. Phần mạng này sẽ không được đào tạo lại trong quá trình huấn luyện.

● Giai đoan thứ hai sử dụng các đặc trưng được trích rút ở bước trên thay vì hình ảnh thô. Mô hình tuần tự sẽ sử dụng chúng và học cách dự đoán mô tả phù hợp cho ảnh.

Khi đó, dữ liệu huấn luyện sẽ bao gồm 2 thành phần: các vector đặc trưng mã hóa cho hình ảnh và chú thích tương ứng.

- Thiết lập thực hiện, thu thập kết quả thực hiện và đánh giá: Luận văn thực hiện thử nghiệm 2 kiến trúc mô hình trên tập dữ liệu UIT-ViIC, bao gồm: NIC, CNN-LSTM. Các mô hình đều sử dụng pretrained phoBERT để tạo word embedding cho từng từ trong bộ từ điển. Tập dữ liệu huấn luyện (train set) giúp cập nhật bộ trọng số tốt nhất cho từng mô hình và sử dụng tập dữ liệu thẩm định (validattion set) giúp đánh giá độ tốt của các mô hình. Mô hình sau khi huấn luyện với bộ tham số đạt kết quả tốt nhất sẽ được sử dụng để đánh giá kết quả trên tập dữ liệu kiển tra (test set)
- Kết luận chương 3.

III. KẾT LUẬN

- Các kết quả đạt được của luận văn tốt nghiệp: Nghiên cứu các kiến trúc mô hình khác nhau nhằm giải quyết bài toán sinh mô tả ảnh cho tiếng Việt. Thực nghiệm các kiến trúc mô hình sinh mô tả ảnh cho tập dữ liệu tiếng Việt. Từ

đó đưa ra đánh giá, nhận xét về kết quả đạt được và đưa ra hướng cải tiến cho bài toán.

- Nhận xét, đề xuất, khuyến nghị và định hướng nghiên cứu phát triển tiếp theo.

IV. DANH MỤC TÀI LIỆU THAM KHẢO

[1] Chaoyang Wang & Ziwei Zhou1 & Liang Xu "An Integrative Review of Image Captioning Research" (2020)

[2] Shuang Liu & Liang Bai & Yanli Hu & Haoran Wang *"Image Captioning Based on Deep Neural Networks"* (2018)

[3] J. X. C. P. D. &. S. R. Lu, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 375-383, (2017)

[4] L. Z. H. X. J. N. L. S. J. L. W. &. C. T. S. Chen, "Sca-cnn: Spatial and channel- wise attention in convolutional networks for image captioning," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 5659-5667, (2017).

[5] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T, "Boosting image captioning with attributes," *Proceedings of the IEEE International Conference on Computer Vision,* pp. 4894-4902, (2017).

[6] Q. D. L. K. V. N. a. N. L.-T. N. Quan Hoang Lam, "UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning," *ICCCI Conference, (*2020)

[7] T.-H. D. &. V.-A. N. Ha Nguyen Tien, "Image Captioning in Vietnamese Language Based on Deep Learning Network," *International Conference on Computational Collective Intelligence, (*2020)

V. DỰ KIẾN KẾ HOẠCH THỰC HIỆN

TT	Nội dung	Dự kiến thời gian thực hiện
1	Nghiên cứu, chọn đề tài, xây dựng đề cương luận văn tốt nghiệp	Từ 01/12/2022 – 11/01/2023
2	Nộp đề cương luận văn tốt nghiệp	12/01/2023
3	Báo cáo đề cương, sửa chữa hoàn thiện, nộp đề cương sau báo cáo	Từ 01/02/2023 – 17/02/2023
4	Nghiên cứu, viết, hoàn thiện luận văn tốt nghiệp	Từ 18/02/2023 –21/05/2023
	Chương 1	
	Chương 2	
	Chương 3	
	Chỉnh sửa, hoàn thiện luận văn	
5	Nộp quyển luận văn tốt nghiệp và hồ sơ bảo vệ	Từ 25/05/2023 – 31/05/2023

Ý KIẾN CỦA GIÁO VIÊN HƯỚNG DẪN
(Ký ghi rõ họ tên)

NGƯỜI LẬP ĐỀ CƯƠNG
(Ký ghi rõ họ tên)

PGS.TS. NGÔ XUÂN BÁCH

NGUYỄN KHÁNH SƠN

DUYỆT CỦA TRƯỞNG TIỂU BAN CHẤM ĐỀ CƯƠNG
(Ký ghi rõ họ tên)