fit@hcmus

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

TỰ ĐỘNG MÔ TẢ NỘI DUNG ẢNH BẰNG MÔ HÌNH HỌC SÂU

(Deep learning models for generating image descriptions)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– ThS. Trần Trung Kiên (Khoa Công nghệ Thông tin)

Nhóm sinh viên thực hiện:

1. Trần Quốc Trình (MSSV: 1612743)
2. Trần Mạnh Thắng (MSSV: 1612892)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *1/2020* đến *6/2020*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Bài toán mô tả nội dung ảnh (hay phát sinh mô tả cho ảnh) được phát biểu như sau (minh họa ở hình 1):

- Cho input là một tấm ảnh.

- Yêu cầu: phát sinh ra (các) câu mô tả về nội dung của ảnh (tự động làm bằng máy).

Hình 1: Minh họa bài toán mô tả nội dung ảnh (nguồn: http://cocodataset.org/#captions-2015)

Nếu giải quyết được bài toán này thì một ứng dụng có thể có là xây dựng hệ thống hỗ trợ người khiếm thị: phát sinh ra các câu mô tả cho các ảnh thu được bằng camera, rồi đọc các câu này lên cho người khiếm thị. Đây là một bài toán không dễ vì ngoài việc phải xác định được các đối tượng trong ảnh, máy còn phải xác định được mối quan hệ giữa các đối tượng, và thể hiện ra bằng một (hoặc các) câu mô tả dưới dạng ngôn ngữ tự nhiên. Bài toán này là giao thoa giữa hai lĩnh vực lớn trong trí tuệ nhân tạo là thị giác máy tính và xử lý ngôn ngữ tự nhiên.

Trong thời gian gần đây, một hướng tiếp cận đạt được kết quả tốt trong bài toán mô tả nội dung ảnh là sử dụng máy học (machine learning), cụ thể hơn là dùng các mô hình học sâu (deep learning model). Và đây là hướng tiếp cận mà chúng em chọn để tìm hiểu.

2.2 Mục tiêu đề tài

- Hiểu rõ tình hình nghiên cứu của bài toán mô tả nội dung ảnh theo hướng tiếp cận học sâu (hiện nay, có các mô hình nào đã được đề xuất để giải quyết bài toán? ý tưởng, kết quả, ưu/nhược điểm của các mô hình này?). Từ đó,

chọn ra một mô hình tốt, có tiềm năng phát triển trong tương lai (và khả thi để có thể hoàn thành trong thời lượng của khóa luận) để tập trung tìm hiểu sâu.

- Hiểu rõ lý thuyết của mô hình đã chọn (trên cơ sở hiểu rõ lý thuyết nền tảng về máy học và học sâu).
- Cài đặt lại mô hình để ra được các kết quả trong bài báo tương ứng; có thể tiến hành thêm các thí nghiệm ngoài bài báo để thấy rõ hơn về ưu/nhược điểm của mô hình.
- Trên cơ sở đã hiểu rõ mô hình, nếu còn thời gian thì có thể xem xét các cải tiến có thể có (chẳng hạn như tăng tốc quá trình huấn luyện mô hình).
- Rèn luyện các kỹ năng: suy nghĩ rõ ràng, lên kế hoạch, làm việc nhóm, trình bày, ...

2.3 Phạm vi của đề tài

Đề tài làm với dữ liệu có các câu mô tả Tiếng Anh; cụ thể, chúng em dự kiến sẽ làm với 3 bộ dữ liệu thường được sử dụng trong bài toán mô tả nội dung ảnh là: Flickr8k, Flickr30k và MS COCO. Về cơ bản, đề tài chỉ tìm hiểu và cài đặt lại mô hình của một bài báo uy tín; ngoài ra, có thể có thêm các thí nghiệm ngoài bài báo nhằm thấy rõ hơn về ưu/nhược điểm của mô hình. Lý do chúng em giới hạn đề tài như vậy là vì: (i) chúng em thấy chỉ riêng việc hiểu rõ mô hình (và các kiến thức nền tảng bên dưới) và có thể tự cài đặt lại đã tốn rất nhiều thời gian, và (ii) chúng em xác định là chỉ trên cơ sở hiểu rõ mô hình (và các kiến thức nền tảng bên dưới) thì mới có thể có được các cải tiến thật sự trong tương lai, cũng như là mới có thể vận dụng được mô hình cho các bài toán khác. Tất nhiên, trong khóa luận, nếu có đủ thời gian thì chúng em cũng sẽ thử đề xuất và cài đặt các cải tiến; tuy nhiên, chúng em xác định đây không phải là mục đích chính.

2.4 Cách tiếp cận dự kiến

Dưới đây sẽ trình bày một số mô hình theo hướng tiếp cận học sâu để giải quyết bài toán mô tả nội dung ảnh mà chúng em đã tìm hiểu được cho đến thời điểm hiện tại, cũng như là mô hình mà chúng em dự kiến sẽ chọn để tập trung tìm hiểu sâu.

- Một bài báo nổi bật trong thời kỳ đầu của hướng tiếp cận học sâu cho bài toán mô tả nội dung ảnh là bài báo "Show and tell: A neural image caption generator" của nhóm tác giả ở Google, được công bố ở hội nghị CVPR2015 [1] (số lần trích dẫn của bài báo tính tới ngày 18/02/2020: 3245). Mô hình mà bài báo này đề xuất gồm 2 thành phần:
  - Mạng nơ-ron tích chập, gọi tắt là CNN (Convolutional Neural Network), dùng để rút trích các đặc trưng quan trọng của ảnh vào trong một véc-tơ có kích thước cố định.
  - Mạng nơ-ron hồi qui, gọi tắt là RNN (Recurrent Neural Network), dùng để phát sinh ra các từ của câu mô tả từ véc-tơ đặc trưng ảnh của CNN và từ các từ đã phát sinh ra.

  Toàn bộ mô hình này sẽ được huấn luyện với tập dữ liệu gồm các cặp (input, output), trong đó input là ảnh và output là câu mô tả đúng của ảnh đó. Ưu điểm của phương pháp này là đơn giản, chủ yếu là để máy tự động học từ dữ liệu, chứ không cần con người phải thiết kế cụ thể nhiều (như các phương pháp trước đó); và tuy đơn giản, phương pháp này lại cho kết quả tốt nhất ở thời điểm đó trên các tập dữ liệu được thử nghiệm.
- Sau đó không lâu, nhóm nghiên cứu của GS. Yoshua Bengio (một trong 3 người được nhận giải Turing vào năm 2018) công bố bài báo "Show, attend and tell: Neural image caption generation with visual attention" ở hội nghị ICML2015 [2] (số lần trích dẫn của bài báo tính tới ngày 18/02/2020: 4510). Mô hình được đề xuất trong bài báo khá tương tự với mô hình ở [1], nhưng có thêm cơ chế attention. Cơ chế attention dựa trên quan sát rằng: trong quá

trình phát sinh ra câu mô tả, để quyết định từ tiếp theo là gì thì không cần nhìn vào toàn bộ ảnh và chỉ cần tập trung (attend) vào một vùng ảnh nào đó (bên cạnh thông tin ảnh, để quyết định từ tiếp theo là gì thì ta cũng dựa vào các từ đã phát sinh trước đó). Như vậy, thay vì cố gắng gom tất cả thông tin của ảnh vào trong một véc-tơ đặc trưng có kích thước cố định (điều mà có thể sẽ khó thực hiện khi ảnh phức tạp, có nhiều đối tượng) và luôn dùng véc-tơ đặc trưng này ở tất các các bước - mỗi bước phát sinh ra một từ - trong quá trình phát sinh ra câu mô tả (điều mà không cần thiết hoàn toàn), cơ chế attention cho phép ở mỗi bước trong quá trình phát sinh ra câu mô tả sẽ chỉ tập trung vào thông tin của vùng ảnh cần thiết. Với cơ chế attention này, mô hình của bài báo [2] đã đạt được kết quả tốt nhất tại thời điểm lúc bấy giờ trên các tập dữ liệu được thử nghiệm. Ngoài độ chính xác, cơ chế attention còn giúp ta hiểu hơn về những gì mô hình đã học được bằng cách xem các vùng ảnh mà mô hình tập trung vào trong quá trình phát sinh ra câu mô tả.

- Sau bài báo trên thì có khá nhiều các bài báo theo sau đi theo hướng dùng cơ chế attention và tìm cách cải tiến, ví dụ như [3], [4], [5], [6]. Các cải tiến này tuy giúp kết quả tốt hơn nhưng cũng làm mô hình phức tạp hơn. Ngoài bài toán mô tả nội dung ảnh, cơ chế attention cũng giúp đạt được những kết quả tốt trong các bài toán khác như dịch máy.

Với những gì đã trình bày ở trên, trong khóa luận, chúng em dự kiến sẽ tập trung tìm hiểu và cài đặt mô hình học sâu sử dụng cơ chế attention được đề xuất trong bài báo [2]. Đây tuy không phải là mô hình đạt được kết quả tốt nhất hiện nay trong bài toán mô tả nội dung ảnh, nhưng là cơ sở cho các cải tiến sau này. Chúng em xác định là phải hiểu rõ mô hình này thì mới có thể hiểu được các cải tiến sau này. Mô hình này tuy đơn giản hơn so với các cải tiến sau này nhưng cũng đã phủ một lượng lớn các kiến thức (CNN, RNN, Attention) mà theo chúng em nghĩ là cũng không dễ để có thể thật sự hiểu rõ.

2.5 Kết quả dự kiến của đề tài

- Cài đặt lại được từ đầu mô hình được đề xuất trong bài báo [2].

- Có được các kết quả thí nghiệm để cho thấy mô hình tự cài đặt ra được các kết quả như trong bài báo.
- Có được các kết quả thí nghiệm để thấy rõ về ưu/nhược điểm của mô hình.
- Nếu có thời gian thì có thể cài đặt và thí nghiệm thêm các cải tiến.

2.6 Kế hoạch thực hiện

Công việc	**Thời gian**	**Người thực hiện**
Tìm hiểu về tình hình nghiên cứu của bài toán mô tả nội dung ảnh, chọn ra mô hình để tập trung tìm hiểu sâu	Tháng 01/2020 - tháng 02/2020	Trình, Thắng
Tìm hiểu về lý thuyết của mô hình đã chọn (bao gồm cả việc tìm hiểu về lý thuyết nền tảng bên dưới)	Tháng 03/2020	Trình, Thắng
Cài đặt lại từ đầu mô hình để ra được các kết quả giống như trong bài báo	Tháng 04/2020	Trình, Thắng
Tiến hành các thí nghiệm để thấy rõ về ưu/nhược điểm của mô hình; xem xét các cải tiến có thể có	Tháng 05/2020	Trình, Thắng
Viết cuốn và slide	Tháng 05/2020 - tháng 06/2020	Trình, Thắng

Bảng 1: *Bảng kế hoạch thực hiện khóa luận*

Tài liệu

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention", in *International conference on machine learning*, pp. 2048–2057, 2015.

[3] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, 2017.

[4] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, 2017.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[6] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning", in *International Conference on Computer Vision*, 2019.

TP. Hồ Chí Minh, ngày 19 tháng 02 năm 2020

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

ThS. Trần Trung Kiên

NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Trần Quốc Trình Trần Mạnh Thắng