TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CHUYÊN NGÀNH

**TÌM HIỂU VỀ BÀI TOÁN**

**MÔ TẢ NỘI DUNG ẢNH**

**N**gành Chuyên Ngành Hệ Thông Tin

Sinh viên: **NGUYỄN ĐÌNH LỘC**

MSSV: 15110243

**PHẠM QUỐC BẢO**

MSSV: 15110161

TP. HỒ CHÍ MINH – 09/2018

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CHUYÊN NGÀNH

**TÌM HIỂU VỀ BÀI TOÁN**

**MÔ TẢ NỘI DUNG ẢNH**

Chuyên Nghành Hệ Thống Thông Tin

Sinh viên: **NGUYỄN ĐÌNH LỘC**

MSSV: 15110243

**PHẠM QUỐC BẢO**

MSSV: 15110161

GVHD**: TS. NGUYỄN THIÊN BẢO**

TP. HỒ CHÍ MINH – 09/2018

**LỜI CẢM ƠN**

**TÓM TẮT**

**MỤC LỤC**

[Danh mục hình viii](#_Toc407486274)

[Danh mục bảng ix](#_Toc407486275)

[Các từ viết tắt x](#_Toc407486276)

[Chương 1 Giới thiệu 1](#_Toc407486277)

[1.1 GIỚI THIỆU 1](#_Toc407486278)

[1.2 CÁC NGHIÊN CỨU LIÊN QUAN 1](#_Toc407486279)

DANH MỤC HÌNH

DANH MỤC BẢNG

CÁC TỪ VIẾT TẮT

# Giới thiệu

## Giới thiệu

Tạo ra các câu mô tả cho hình ảnh một cách tự động là một nhiệm vụ rất gần với scene understanding (hiểu ngữ cảnh) - một trong những nhiệm vụ chính của lĩnh vực Computer Vision. Một mô hình tạo ra các câu mô tả từ hình ảnh không những đủ mạnh mẽ để xác định được những đối tượng nào có bên trong hình ảnh mà còn phải nắm bắt và diễn tả được các mối quan hệ của các đối tượng bằng ngôn ngữ tự nhiên. Vì vậy mà việc tạo ra các câu mô tả trực quan cho hình ảnh từ lâu đã được xem là một vấn đề khó. Đây là một thách thức rất quan trọng cho các thuật toán của việc học máy. Một mô hình như vậy phải bắt chước được khả năng tuyệt vời của con người đó là nén một lượng lớn thông tin thành ngôn ngữ, mô tả một cách trực quan, nổi bật và đó là một thách thức lớn đối với lĩnh vực máy học (machine learning) và nghiên cứu trí tuệ nhân tạo.

Trong những năm gần đây, có rất nhiều nghiên cứu về bài toán mô tả hình ảnh được thực hiện hoá nhờ các mạng Neural Network và số lượng lớn dữ liệu cho việc huấn luyện. Các nghiên cứu gần đây đã đạt được nhiều thành công đáng kể bằng cách sử dụng kết hợp giữa mạng Convolution Neural Network để trích xuất các thông tin, đặc điểm của hình ảnh thành một vec-tơ và mạng Recurrent Neural Network để chuyển đổi các vec-tơ đó thành câu mô tả.

Một trong những khía cạnh đáng chú ý của hệ thống thị giác ở con người là sự “chú ý” (Rensink, 2000; Corbetta & Shulman, 2002). Thay vì nén toàn bộ hình ảnh, thể hiện chúng ở một dạng dữ liệu tĩnh, sự “chú ý” cho phép mô hình nắm bắt được các thông tin nổi bật khi cần thiết. Điều này rất quan trọng khi hình ảnh chứa nhiều vật thể rất lộn xộn. Trích xuất các thông tin nổi bật của hình ảnh từ lớp cuối trong mạng Convolution Neural Network là một phương pháp hiệu quả đã được sử dụng rộng rãi trước đây. Tuy nhiên, phương pháp này tiềm ẩn khả năng có thể làm mất các thông tin hữu ích cho mô hình. Sử dụng các thông tin thô ở cấp thấp có thể hạn chế được nhược điểm này. Nhưng để làm việc với các thông tin cấp thấp này bắt buộc phải có một cơ chế mạnh mẽ để điều hướng mô hình nhận diện được các thông tin quan trọng.

## CÁC NGHIÊN CỨU GẦN ĐÂY

Trong những năm gần đây, có rất nhiều phương pháp đã được đề xuất cho bài toán mô tả nội dung ảnh. Nhiều phương pháp sử dụng mạng Recurrent Neural Network, lấy ý tưởng từ thành công của mô hình tuần tự (sequence-to-sequence) trong bài toán dịch ngôn ngữ - machine translation (Cho et al., 2014; Bahdanauet al., 2014; Sutskever et al., 2014; Kalchbrenner & Blunsom, 2013). Cấu trúc mã hoá-giải mã (encoder-decoder) trong machine translation rất phù hợp với bài toán mô tả ảnh bởi vì có thể coi quá trình chuyển đổi một bức ảnh thành câu mô tả tương tự như quá trình dịch từ một ngôn này sang một ngôn ngữ khác trong machine translation.

Cách tiếp cận đầu tiên sử dụng Neural Network cho bài toán mô tả nội dung ảnh được viết trong bài báo khoa học Kiros et al. (2014a), tác giả đã sử dụng mô hình multimodal log-bilinear (một mô hình Language Model đơn giản mà dự đoán từ xuất hiện tiếp theo trong câu dựa trên các từ đã xuất hiện trước đó) kết hợp với các thông tin, đặc điểm đã được trích xuất từ ảnh để tạo ra câu mô tả. Trong bài báo khoa học Mao et al. (2014), tác giả cũng sử dụng phương pháp tương tự nhưng thay thế mạng Feedforward Neural Network bằng một mạng Recurrent Neural Network. Còn trong hai bài báo Vinyals et al. (2014) và Donahueet al. (2014), tác giả sử dụng mạng lưới Recurrent Neural Network với các đơn vị long short-term memory (LSTM - Hochreiter & Schmidhuber, 1997), không giống như các bài báo Kiros et al. (2014a) và Mao et al. (2014) mà mô hình của họ nhận hình ảnh làm dữ liệu đầu vào ở mỗi bước trong quá trình tạo ra câu mô tả, thì mô hình trong bài báo Vinyals et al. (2014) chỉ nhận hình ảnh làm dữ liệu đầu vào tại bước đầu tiên của quá trình. Cùng với dữ liệu hình ảnh, các bài báo Donahue et al. (2014) và Yao et al. (2015) cũng áp dụng kỹ thuật trên với dữ liệu video, cho phép mô hình của họ có thể tạo ra câu mô tả cho cả một video.

Hầu hết các mô hình được sử dụng trong các bài báo trên, dữ liệu thông tin của hình ảnh được biểu diễn dưới dạng một vec-tơ được lấy từ lớp cuối của một mạng Convolutional Network đã được huấn luyện trên một tập dữ liệu hình ảnh. Trong bài báo Fang et al. (2014) thì tác giả lại đề xuất một phướng pháp với ba bước bằng cách kết hợp với bài toán nhận diện vật thể (object detections), mô hình này ban đầu huấn luyện bộ nhận dạng (detectors) dựa trên một multi-instance learning framework. Một mô hình Language Model huấn luyện trên tập dữ liệu các câu mô tả được áp dụng vào dự liệu đầu ra của bộ nhận dạng. Không giống như các mô hình này, mô hình của chúng tôi đề xuất không sử dụng trực tiếp các bộ nhận dạng này mà thay vào đó học các thông tin tiềm ẩn trong hình ảnh ngay từ đầu. Điều này cho phép mô hình của chúng tôi có thể học được khái niệm trừu tượng trong hình ảnh.

**Tài liệu tham khảo**