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**

Chuyên Ngành Hệ Thông Tin

GVHD**: TS. NGUYỄN THIÊN BẢO**

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

GVHD**: TS. NGUYỄN THIÊN BẢO**

Sinh viên: **NGUYỄN ĐÌNH LỘC**

MSSV: 15110243

**PHẠM QUỐC BẢO**

MSSV: 15110161

TP. HỒ CHÍ MINH – 09/2018

**LỜI CẢM ƠN**

**TÓM TẮT**

**MỤC LỤC**

[**DANH MỤC HÌNH vii**](#_Toc526116647)

[**DANH MỤC BẢNG viii**](#_Toc526116648)

[**CÁC TỪ VIẾT TẮT ix**](#_Toc526116649)

[Chương 1 Giới thiệu 1](#_Toc526116650)

[**1.1 Giới thiệu 1**](#_Toc526116651)

[**1.2 CÁC NGHIÊN CỨU GẦN ĐÂY 2**](#_Toc526116652)

[**1.3 TẬP DỮ LIỆU FLICK8K 3**](#_Toc526116653)

[Chương 2 MÔ TẢ HÌNH ẢNH VỚI MÔ hình ATTention 5](#_Toc526116654)

[**2.1 CẤU TRÚC CỦA MÔ HÌNH 5**](#_Toc526116655)

[**2.1.1 BỘ MÃ HOÁ: CONVOLUTIONAL NEURAL NETWORK 5**](#_Toc526116656)

[**2.1.2 BỘ GIẢI MÃ: RNN VỚI CÁC ĐƠN VỊ LSTM 6**](#_Toc526116657)

[**2.1.3 MÔ HÌNH ATTENTION 7**](#_Toc526116658)

DANH MỤC HÌNH

DANH MỤC BẢNG

CÁC TỪ VIẾT TẮT

# Giới thiệu

## Giới thiệu

Bài toán tạo ra câu mô tả nội dung 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 thị giác máy tính (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 đó 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. 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 và mô tả một cách trực quan, nổi bật. 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 nhờ sự phát triển của lĩnh vực học sâu (Deep Learning) với các mạng thần kinh nơ-ron nhân tạo (Artificial Neural Network - ANN) và số lượng lớn dữ liệu sẵn có 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 (CNN) để trích xuất các thông tin, đặc điểm của hình ảnh thành một vec-tơ và mạng Recurrent Neural Network (RNN) để chuyển đổi các thông tin 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 xử lý toàn bộ thông tin về hình ảnh cùng một lúc, sự “chú ý” cho phép mô hình tập trung vào một phần của hình ảnh và xử lý các thông tin nổi bật đó khi cần thiết. Điều này rất quan trọng khi hình ảnh chứa nhiều vật thể lộn xộn và chồng chéo lên nhau. Trích xuất các thông tin của hình ảnh từ lớp cuối trong mạng CNN 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 một số thông tin hữu ích. Sử dụng các thông tin thô ở cấp độ 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

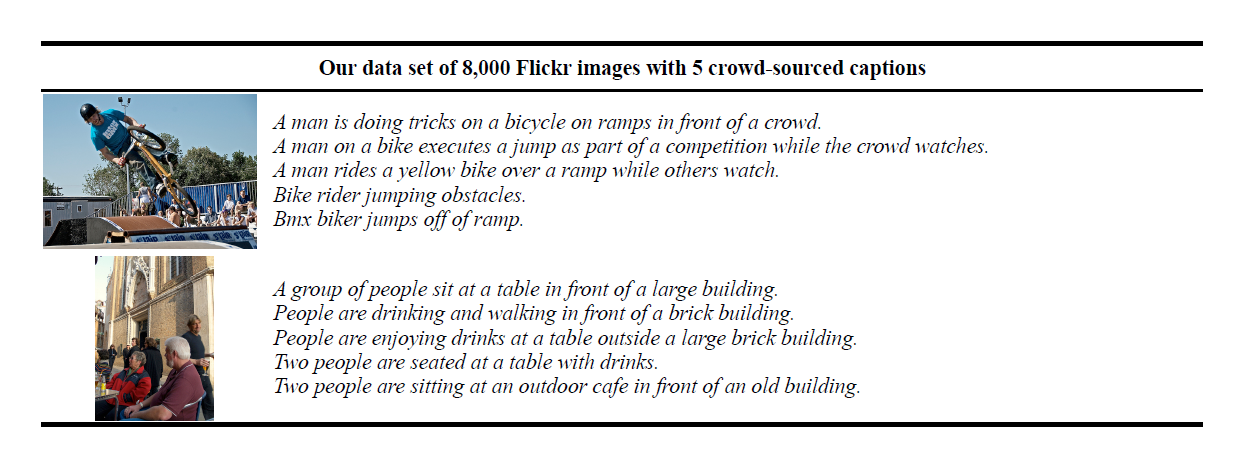
Trong những năm gần đây, có rất nhiều phương pháp đã được đề xuất cho bài toán mô tả nội dung ảnh. Nhiều phương pháp sử dụng RNN, lấy ý tưởng từ thành công của mô hình tuần tự (sequence-to-sequence) trong bài toán dịch ngôn ngữ - machine translation (Cho et al., 2014; Bahdanauet al., 2014; Sutskever 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 RNN .Còn trong hai bài báo Vinyals et al. (2014) và Donahueet al. (2014), tác giả sử dụng mạng lưới RNN 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 CNN 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.

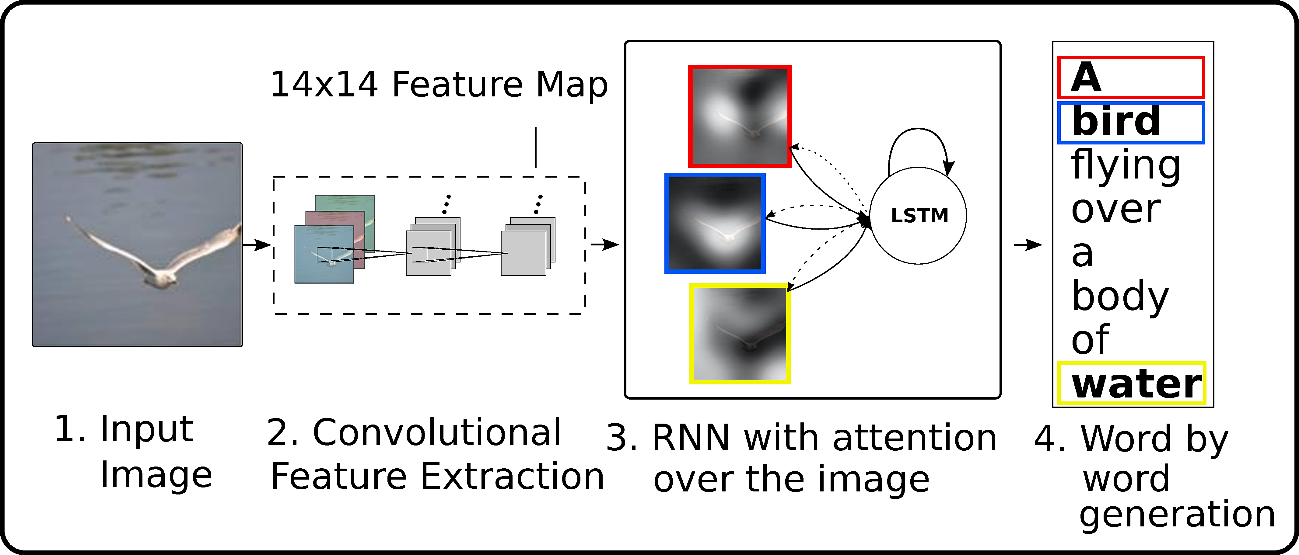
## TẬP DỮ LIỆU FLICK8K

Tập dữ liệu Flickr8k có 8092 hình ảnh được lấy từ trang web flickr.com. Nội dung của những tấm hình trong tập dữ liệu này là tập trung về các hành động của người và động vật (chủ yếu là chó). Những tấm hình đó không chứa bất kỳ những người nổi tiếng hay danh lam thắng cảnh nào. Ngoài ra trong tập dữ liệu Flickr8k còn có chứa những câu mô tả về những tấm hình riêng biệt. Để tránh lỗi cú pháp, những người gắn nhãn đã vượt qua các bài kiểm tra về ngữ pháp mới có thể ghi chú thích và mô tả các tình huống, sự kiện và đối tượng cho tập dữ liệu này. Những câu đó sẽ được gom lại thành một tập các câu mô tả riêng biệt cho từng tấm hình cụ thể, bởi vì một hình ảnh có thể có nhiều câu mô tả khác nhau. Những câu mô tả bức hình có độ dài trung bình là 12 từ. Dưới đây là hình ảnh mô tả tập dữ liệu Flickr8k:



# MÔ TẢ HÌNH ẢNH VỚI MÔ hình ATTention

## CẤU TRÚC CỦA MÔ HÌNH

Mô hình cho bài toán mô tả nội dung ảnh sẽ nhận vào một hình ảnh ở định dạng thô và tạo ra câu mô tả cho hình ảnh đó. Dựa trên ý tưởng của mô hình sequence-to-sequence với cơ chế attention trong bài toán máy dịch – tập trung vào các từ có ngữ cảnh cần thiết ở mỗi lần sinh từ tiếp theo, một mô hình image-to-sequence với cơ chế attention được sử dụng – tập trung vào các mảng hình cần thiết trong bức ảnh ở mỗi lần sinh từ mô tả tiếp theo. Với cấu trúc mã hoá – giải mã, một mạng CNN được sử dụng làm bộ mã hoá – mã hoá hình ảnh thành các thông tin cần thiết, một mạng RNN làm bộ giải mã – giải mã các thông tin về hình ảnh thành câu mô tả và để thực hiện cơ chế attention, một mô hình attention sẽ đảm nhận giao tiếp trung gian giữa CNN và RNN.

Hình 1. Mô hình mô tả nội dung ảnh với cơ chế attention.

### BỘ MÃ HOÁ: CONVOLUTIONAL NEURAL NETWORK

Mạng CNN đảm nhận công việc trích xuất các thông tin về hình ảnh thông qua các lớp của nó, thay vì lấy các thông đã được nén lại trong một véc-tơ ở lớp fully connected (các lớp với kết nối dày đặc của mạng ANN) khi mà các thông tin riêng lẻ của hình ảnh đã được nén lại khó cho việc thực hiện cơ chế attention, một lớp convolution được sử dụng. Các thông tin từ lớp convolution được trính xuất thành một bộ véc-tơ tương ứng với từng phần của hình ảnh.

Cụ thể bao gồm L véc-tơ được trính xuất, mỗi véc-tơ có số chiều là D.

14x14x512

L = 14x14

D = 512

Convolution

Hình 2. Bộ véc-tơ được trích xuất từ lớp convolution.

Trích xuất các thông tin từ lớp convolution cho phép bộ giải mã có thể lựa chọn tập trung vào từng phần cụ thể của hình ảnh.

### BỘ GIẢI MÃ: RNN VỚI CÁC ĐƠN VỊ LSTM

Mạng RNN với các đơn vị LSTM (Hochreiter & Schmidhuber, 1997) được sử dụng để tạo ra câu mô tả, mỗi bước mô hình sinh ra từ tiếp theo trong câu dựa vào véc-tơ ngữ cảnh (context vector), trạng thái trước đó của nơ-ron (previous hidden state) và từ được sinh ra trước đó.

Ở đây, , , , , lần lượt là cổng forget, cổng input, cổng output, memory và hidden state của đơn vị LSTM. Véc-tơ là véc-tơ ngữ cảnh, chưa các thông tin về hình ảnh ứng với từng mảng của hình, được tính toán dựa vào mô hình attention. là ma trận word embeding với là one-hot véc-tơ biểu thị từ được sinh ra trước đó. và là các tham số cần học của mạng. là hàm sigmoid activation.

Các giá trị khởi tạo memory và hidden state của LSTM được dự đoán thông qua hai mạng ANN (, ) bằng giá trị trung bình cộng của bộ véc-tơ trích xuất từ lớp convolution:

Trong mô hình này, mạng RNN sử dụng một “deep output layer” (Pascanu, et al., 2014) để dự đoán từ tiếp theo trong câu sử dụng hidden state, véc-tơ ngữ cảnh và từ dự đoán trước đó.

### MÔ HÌNH ATTENTION

Véc-tơ ngữ cảnh chứa các thông tin về từng phần của hình ảnh ở thời điểm t trong quá trình tạo câu mô tả. được tính đưa theo bộ véc-tơ trích xuất được từ lớp convolution tương ứng với từng vị trí khác nhau trên hình ảnh. Đối với từng vị trí , mô hình tính toán ra một trọng số dương tương ứng với tầm quan trọng, ảnh hưởng mà vị trí được “chú ý” trong quá trình tạo ra từ tiếp theo trong câu. Trọng số của từng véc-tơ được tính toán bằng mô hình attention dựa trên một mạng ANN với dữ liệu đầu vào là véc-tơ và giá trị hidden state trước đó .

Khi các trọng số (có tổng bằng 1) được tính, véc-tơ ngữ cảnh được xác định bằng kỳ vọng của bộ véc-tơ (Bahdanau, et al., September 2014):

**TÀI LIỆU THAM KHẢO**

Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, & Yoshua. (September 2014). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, . . . Yoshua. (2014). How to construct deep recurrent neural networks. *ICLR*.