PHÁT SINH MÔ TẢ CHO ẢNH VỚI CƠ CHẾ ATTENTION TRÊN HÌNH ẢNH VÀ TĂNG CƯỜNG THÔNG TIN KHÁI NIỆM

Lương Quốc An, Võ Hồ Việt Khoa
Khoa Công nghệ Thông Tin,
Trường Đại học Khoa học Tự Nhiên, ĐHQG-HCM
1412020@student.hcmus.edu.vn, 1412255@student.hcmus.edu.vn

Tóm tắt

Bài toán phát sinh câu mô tả cho hình ảnh có rất nhiều ứng dụng từ hỗ trợ người khiếm thị cho đến tích hợp vào các hệ thống trợ lý ảo thông minh. Tuy nhiên, bài toán này vẫn đang là một thách thức. Để máy tính mô tả được một hình ảnh, máy tính cần phải có khả năng hiểu những thông tin từ hình ảnh cũng như khả năng xử lý ngôn ngữ tự nhiên. Đặc biệt là khả năng tập trung vào chi tiết của nội dung hình ảnh cũng như phải đảm bảo tính toàn vẹn và ngữ pháp của câu mô tả. Để giải quyết các vấn đề đó, nhóm sử dụng một bộ dữ liệu nổi tiếng là COCO, bao gồm gần 130000 ảnh, mỗi ảnh ứng với 5 câu mô tả, mô tả ảnh theo những cách khác nhau. Cùng với đó, bộ dữ liệu chứa các hình ảnh của 80 đối tượng, trong nhiều cảnh trí khác nhau, thực hiện nhiều hoạt động khác nhau sẽ giúp mô hình huấn luyện giải quyết được các vấn đề nêu trên. Trong công trình này, từ cơ sở những phương pháp đã tìm hiểu, nhóm nghiên cứu và đề xuất một phương pháp cải tiến cho mô hình phát sinh câu mô tả bằng cách kết hợp giữa thông tin từ hình ảnh và các nhãn ngữ nghĩa. Nhóm sinh viên cũng đề xuất một phương pháp beam search cải tiến để đảm bảo tính toàn vẹn của câu mô tả. Kết quả cho thấy một số cải thiện so với một số công trình trước đây, tuy nhiên vẫn còn nhiều hạn chế.

Từ khóa: mô tả ảnh, attention.

IMAGE CAPTION GENERATION WITH ATTENTION ON IMAGE AND CONCEPTS AUGMENTATION

Quoc-An Luong, Viet-Khoa Vo-Ho
Faculty of Information and Technology, University of Science, VNU-HCM
1412020@student.hcmus.edu.vn, 1412255@student.hcmus.edu.vn

Abstract

Generating image caption which fully describe the main contents of an image by a sentence has a wide range of application in our life such as supporting system for visual-impaired people or virtual assistant system. However, this problem is still a hard challenge in computer vision. In order to generate a good caption, the computer is required not only to understand visual information from images but also to be able to process natural language. Moreover, the caption should describe the detailed content of the image as well as ensure the grammar and completeness of the sentence. To resolve the those problems, we use a popular dataset of COCO, which comprises of roughly 130k images, each comes with 5 sentences describing it in different ways. Also, the dataset covers about 80 objects in many different scenes, doing different activities will provide a comprehensive information for training the model to tackle problems mentioned above. In our work, based on some previous researches, we propose a new model for generating image caption from combination of visual information and tags of concept. We also propose a modified beam search method to ensure the completeness of generated sentences. Our result shows some improvement compared to some previous works. Our generated captions are completed and contain more details about the object in the image but there are still some limits in our model compared to the current state of the art.

Key words: image captioning, attention.