THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/9FkeVSkDSho*
- Link Github:
  - Proposal file: https://github.com/thanhluanpy/CS2205.APR2023/blob/main/ImageCaptionGenerator.pdf
  - Slides file: https://github.com/thanhluanpy/CS2205.APR2023/blob/main/ImageCaptionGenerator-Slide.pdf
  - Poster file: https://github.com/thanhluanpy/CS2205.APR2023/blob/main/ImageCaptionGenerator-Poster.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Trương Thanh Luân • MSSV: 220201016 	• Tự đánh giá (điểm tổng kết môn): 9/10 • Số buổi vắng: 0 • Link Github: https://github.com/thanhluanpy/CS2205.APR2023

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG MÔ HÌNH CNN & LSTM - GÁN NHÃN CHO HÌNH ẢNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

CNN & LSTM MODEL APPLICATION - IMAGE CAPTION GENERATOR

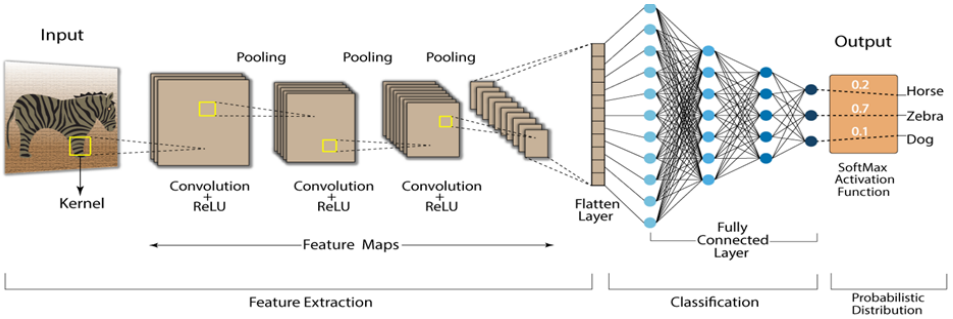
TÓM TẮT *(Tối đa 400 từ)*

Sử dụng ngôn ngữ tự nhiên và thị giác máy tính để nhận dạng ngữ cảnh của hình ảnh và mô tả chúng bằng ngôn ngữ tự nhiên [6]. Mô hình CNN và LSTM là sự kết hợp mạnh mẽ trong việc hiểu và phân loại hình ảnh một cách chính xác và có thể đưa ra mô tả trực quan phong phú, tạo ra các câu hoàn chỉnh bằng ngôn ngữ tự nhiên từ đầu vào là một hình ảnh [1]. CNN được sử dụng để trích xuất các đặc điểm hình ảnh và LSTM được sử dụng để dịch các đặc điểm hình ảnh thành câu [3]. Cấu trúc của mô hình tạo chú thích hình ảnh được thể hiện trong Hình 4. Mô hình hướng tới việc nhận diện và phân loại các đối tượng, từ con người đến động vật và vật thể, đồng thời nhận dạng hành động trong video [2]. Với sự kết hợp này, đã tạo ra một phương pháp gán nhãn hình ảnh hiệu quả, đáng tin cậy và tự động, mở ra tiềm năng ứng dụng rộng trong lĩnh vực xử lý hình ảnh và trí tuệ nhân tạo.

GIỚI THIỆU *(Tối đa 1 trang A4)*

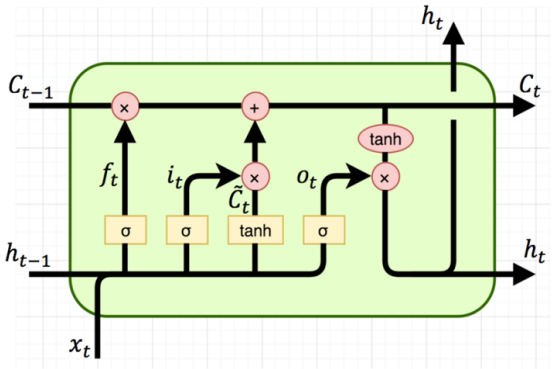
Để giúp những người khiếm thính hoặc các em nhỏ có thể biết được cảnh vật xung quanh hay hỗ trợ việc di chuyển (Image -> text -> voice), cũng như quản lý và tìm kiếm được hình ảnh dựa vào ghi chú [1, 3, 4]. Trong thời đại công nghệ phát triển với dữ liệu hình ảnh ngày càng phong phú, việc hiểu và phân loại hình ảnh đã trở thành một thách thức quan trọng trong lĩnh vực trí tuệ nhân tạo. Mô hình CNN (Convolutional Neural Network) và LSTM (Long Short-Term Memory) đã nổi lên như những công nghệ vượt trội trong việc xử lý và gán nhãn cho hình ảnh. Sự kết hợp của hai mô hình này đã tạo ra một cách tiếp cận đột phá, mở ra tiềm năng vô hạn và ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau.

Mô hình CNN là một kiến trúc mạng thần kinh nhân tạo được thiết kế đặc biệt cho việc phân tích hình ảnh. Nó tập trung vào trích xuất vector đặc trưng từ hình ảnh, giúp nhìn thấy và hiểu các đối tượng trong đó. Bằng cách sử dụng các lớp tích chập và lớp pooling, CNN có khả năng tự động học và trích xuất các đặc trưng quan trọng từ hình ảnh, tạo nên một biểu diễn trừu tượng và dễ dàng cho việc phân loại và gán nhãn [1].



Hình 1: mô hình CNN [7]

Trong khi đó, LSTM là một dạng đặc biệt của mạng nơ-ron hồi quy, nổi tiếng với khả năng mô hình hóa thông tin thời gian. LSTM có khả năng ghi nhớ và xử lý thông tin từ quá khứ và có thể phân tích chuỗi hình ảnh liên tiếp để dự đoán kết quả trong tương lai, chống hiện tượng "Vanishing Gradient" [1, 2], làm cho nó trở thành công cụ mạnh mẽ cho việc phân loại hành động và dự đoán xu hướng [1, 2].



Hình 2: mô hình LSTM [4]

Sự kết hợp CNN và LSTM mang lại những lợi ích đáng kể trong việc hiểu và phân loại hình ảnh. Khả năng trích xuất đặc trưng không gian của CNN kết hợp với khả năng mô hình hóa thông tin thời gian của LSTM tạo nên một cách tiếp cận toàn diện, đem lại độ chính xác và tự động hóa cao trong việc gán nhãn hình ảnh và phân loại hành động.

Với sự phát triển không ngừng của công nghệ, cùng với sự phát triển của tập dữ liệu và khả năng tính toán, mô hình này hứa hẹn sẽ đem lại những đóng góp quan trọng và ứng dụng đa dạng trong tương lai. Góp phần thay đổi cách chúng ta tương tác với hình ảnh, mang lại sự tiện ích và tiếng nói mới cho lĩnh vực này.

Dataset: Flickr8K

Input: tập ảnh bất kỳ cần gán nhãn.

Output: tập ảnh đã được gán nhãn.

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Hình 3: hình được gán nhãn [4]

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Đạt được độ chính xác cao: xây dựng mô hình có khả năng nhận diện và phân loại đối tượng, vật thể và hành động trong hình ảnh một cách chính xác và đáng tin cậy.
- Tự động hóa quy trình gán nhãn: tự động xử lý quá trình gán nhãn hình ảnh, giúp tiết kiệm thời gian và công sức so với việc gán nhãn thủ công. Điều này đảm bảo tính hiệu quả và khả năng tự động hóa trong việc xử lý hình ảnh.
- Mở rộng ứng dụng: ngoài việc gán nhãn đối tượng và vật thể, mô hình cũng có thể nhận dạng và phân loại các yếu tố khác trong hình ảnh như màu sắc, vị trí, biểu cảm và ngữ cảnh. Điều này mang lại tiềm năng ứng dụng rộng lớn trong các lĩnh vực như quảng cáo, thương mại điện tử, y tế và nhiều lĩnh vực khác.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

1. NỘI DUNG

Trích xuất vector đặc trưng ảnh bằng mô hình CNN: mô hình CNN được sử dụng để trích xuất các đặc trưng không gian từ hình ảnh. Các lớp tích chập của CNN giúp tìm ra các đặc trưng cấu trúc như cạnh, góc, hoặc hình dạng của các đối tượng trong hình ảnh.

Mô hình hóa thông tin thời gian bằng LSTM: sau khi trích xuất vector đặc trưng ảnh, các đặc trưng này được đưa vào LSTM để mô hình hóa thông tin thời gian. LSTM có khả năng hiểu và nhớ các quan hệ thời gian trong chuỗi hình ảnh, cho phép nắm bắt các thông tin liên quan đến sự xuất hiện và biến đổi của các đối tượng trong hình ảnh theo thời gian.

Kết hợp sức mạnh của CNN trong việc trích xuất đặc trưng không gian và LSTM trong việc

mô hình hóa thông tin thời gian, giúp quá trình gán nhãn hình ảnh chính xác.

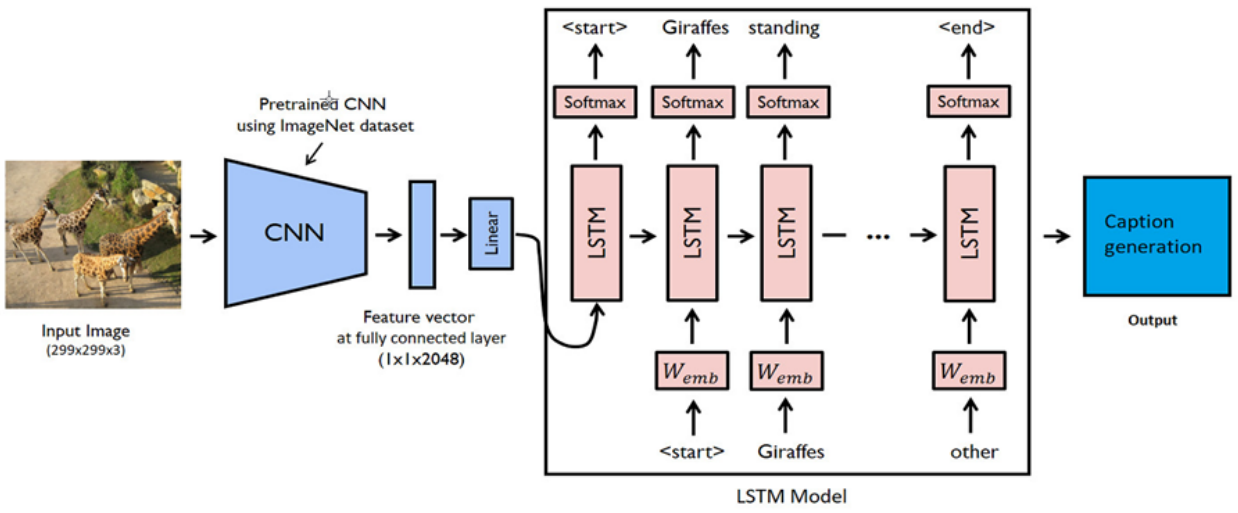
2. PHƯƠNG PHÁP:

Chuẩn bị dữ liệu: sử dụng dữ liệu Flickr8K Dataset và Flickr8K Text [5] để huấn luyện mô hình với 8000 ảnh và tiền xử lý (hình ảnh và văn bản) trước khi đưa vào mô hình [3]. Điều này bao gồm việc chia 6000 ảnh thành tập huấn luyện, 1000 ảnh cho tập xác nhận và 1000 ảnh tập kiểm tra, mỗi ảnh có 5 ghi chú bằng tiếng anh [4]. Thực hiện các biến đổi, chuẩn hóa dữ liệu (loại bỏ chữ số, kí tự đặc biệt, dấu câu và các từ không hữu ích) [3] và chuyển đổi định dạng hình ảnh thành dữ liệu phù hợp cho mô hình.

Xây dựng mô hình CNN & LSTM: các đặc trưng hình ảnh sẽ được trích xuất từ InceptionV3, một mô hình CNN được đào tạo trên bộ dữ liệu Imagenet và sau đó cung cấp các đặc trưng này vào mô hình LSTM, mô hình LSTM chịu trách nhiệm tạo chú thích hình ảnh [6].

Huấn luyện và đánh giá mô hình: được huấn luyện trên tập dữ liệu huấn luyện và dựa trên tập kiểm tra để đánh giá hiệu suất của mô hình. Quá trình huấn luyện bao gồm việc tối ưu hóa các tham số thông qua việc tính toán gradient và điều chỉnh các trọng số mạng Neural. Đánh giá mô hình dựa trên các chỉ số như độ chính xác, độ đo và ma trận nhầm lẫn (confusion matrix).

Gán nhãn cho hình ảnh mới: sau khi mô hình đã được huấn luyện và đánh giá, nó có thể được sử dụng để gán nhãn cho hình ảnh mới. Hình ảnh mới sẽ trải qua quá trình trích xuất đặc trưng không gian bằng CNN và mô hình hóa thông tin thời gian bằng LSTM để dự đoán nhãn tương ứng.



Hình 4: Mô hình tạo chú thích hình ảnh [6]

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- *Độ chính xác cao: kết quả dự đoán của mô hình đáng tin cậy và chính xác.*
- *Tự động hóa quá trình gán nhãn: giúp tự động hóa quá trình gán nhãn cho hình ảnh, giảm thiểu sự can thiệp của con người, tốc độ xử lý nhanh và tiết kiệm thời gian.*
- *Xử lý hình ảnh phức tạp và chuỗi hình ảnh: có khả năng xử lý hình ảnh phức tạp và chuỗi hình ảnh liên tiếp. Nhận diện các đối tượng và hành động trong chuỗi hình ảnh, cung cấp thông tin chi tiết và toàn diện.*
- *Tính linh hoạt và ứng dụng rộng rãi: cho phép áp dụng vào nhiều lĩnh vực khác nhau. Các ứng dụng bao gồm nhận dạng và phân loại hình ảnh, xử lý video, nhận dạng hành vi, ...*

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

Articles

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan: Show and Tell: A Neural Image Caption Generator. In arXiv:1411.4555, 2014. 1, 2

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell: Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In arXiv:1411.4389, 2014. 1, 2, 3, 4

[3] Ayush Kumar Poddar, Dr. Rajneesh Rani: Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language. In ICCECE, 2020.

Books

[4] Nguyễn Thanh Tuấn: Sách Deep Learning Cơ Bản. Tái bản lần 2 - Tháng 08, Năm 2020.

Websites

[5] Dataset: Flickr8K Dataset & Flickr8K Text. Url: https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b

[6] Data Flair Blog. Url: https://data-flair.training/blogs/python-based-project-image-caption-generator-cnn

[7] Basic CNN in Deep Learning. Url: https://www.analyticsvidhya.com/blog/2022/03/basics-of-cnn-in-deep-learning