TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN VÕ CÔNG HUY - 52000765**

**TẠO TIÊU ĐỀ CHO ẢNH**

**BÁO CÁO CUỐI KỲ**

**THỊ GIÁC MÁY TÍNH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN VÕ CÔNG HUY - 52000765**

**TẠO TIÊU ĐỀ CHO ẢNH**

**BÁO CÁO CUỐI KỲ**

**THỊ GIÁC MÁY TÍNH**

Người hướng dẫn

**TS. Phạm Văn Huy**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Để hoàn thành dự án này, lời đầu tiên em xin được gửi lời cảm ơn sâu sắc đến thầy TS. Phạm Văn Huy trong suốt thời gian đồng hành trong môn học Thị giác máy tính, thầy đã rất trau chuốt trong từng buổi học để giúp em hiểu rõ về từng yếu tố nhỏ trong môn học, người đã giúp em từng bước hoàn thành dự án. Thầy đã hướng dẫn em từ những bước đầu tiên, gợi ý cho em những ý tưởng, lên kế hoạch thực hiện dự án, đến những bước cuối cùng hoàn thiện dự án. Sự tận tình của thầy đã giúp dự án đạt được kết quả như ngày hôm nay.

*TP. Hồ Chí Minh, ngày 8 tháng 3 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*N..V.C Huy*

*Nguyễn Võ Công Huy*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS Phạm Văn Huy. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 8 tháng 3 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*N.V.C Huy*

*Nguyễn Võ Công Huy*

**TẠO TIÊU ĐỀ CHO ẢNH**

**TÓM TẮT**

Đồ án sử dụng mô hình kết CNN (Convolutional Neural Networks) - LSTM (Long short term memory) để tạo các tiêu đề cho ảnh. Trong đó, việc trích xuất các đặc trưng của ảnh sẽ sử dụng Xception - là một mô hình CNN được huấn luyện dựa trên bộ dữ liệu imagenet và mô hình LSTM sẽ được sử dụng để tạo ra tiêu đề cho ảnh.

Ngoài ra để tăng thêm tính sinh động cho đồ án thì đồ án còn được xây dựng kết hợp thêm module web với (Flask - Python Framework và ReactJS - thư viện javascript front-end) để tải ảnh lên và hiển thị tiêu đề của mô hình trên website.

**GENERATE THE CAPTION FOR THE IMAGE**

**ABSTRACT**

The project uses the Convolutional Neural Networks (LSTM) model to create caption for the image. In it, extracting image features will use Xception - a CNN model that is trained based on the imagenet dataset and the LSTM model that will be used to generate the caption for the image.

In addition to making the project more vibrant, the project has been built with a combination of web modules (Flask - Python Framework and ReactJS - the javascript front-end library) to upload images and display the caption of uploaded image on the website.

**MỤC LỤC**

[DANH MỤC HÌNH VẼ vii](#_Toc160813662)

[DANH MỤC BẢNG BIỂU viii](#_Toc160813663)

[DANH MỤC CÁC CHỮ VIẾT TẮT ix](#_Toc160813664)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc160813665)

[1.1 Lý do chọn đề tài 1](#_Toc160813666)

[1.2 Mục tiêu thực hiện đề tài 1](#_Toc160813667)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 1](#_Toc160813668)

[2.1 Module AI 1](#_Toc160813669)

[2.1.1 Mạng nơ-ron tích chập 1](#_Toc160813670)

[2.1.2 Long short term memory – LSTM 2](#_Toc160813671)

[2.2 Module website 3](#_Toc160813672)

[2.2.1 Front-end 3](#_Toc160813673)

[2.2.2 Back-end 3](#_Toc160813674)

[CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT 4](#_Toc160813675)

[3.1 Mô hình kết hợp giữa CNN và LSTM 4](#_Toc160813676)

[CHƯƠNG 4. THỰC NGHIỆM 5](#_Toc160813677)

[4.1 Dữ liệu thực nghiệm 5](#_Toc160813678)

[4.2 Cài đặt thực nghiệm 5](#_Toc160813679)

[4.2.1 Môi trường 5](#_Toc160813680)

[4.2.2 Đánh giá mô hình 5](#_Toc160813681)

[CHƯƠNG 5. KẾT LUẬN 6](#_Toc160813682)

[5.1 Kết luận 6](#_Toc160813683)

[5.2 Hướng phát triển 6](#_Toc160813684)

[TÀI LIỆU THAM KHẢO 7](#_Toc160813685)

# DANH MỤC HÌNH VẼ

[Hình 4.1: Đánh giá kết quả bằng BLEU score 6](#_Toc160813686)

# DANH MỤC BẢNG BIỂU

# DANH MỤC CÁC CHỮ VIẾT TẮT

# MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Chúng ta nhìn vào một hình ảnh và bộ não có thể dễ dàng biết hình ảnh là gì, nhưng làm sao để một máy tính có thể biết được những gì hình ảnh đang thể hiện ? Các nhà nghiên cứu thị giác máy tính đã làm việc rất nhiều và họ đã từng cho rằng nó là không thể cho đến hiện tại! Với sự tiến bộ trong kỹ thuật học sâu (deep learning), sự sẵn có của các tập dữ liệu khổng lồ và sức mạnh máy tính cả về phần cứng lẫn phần mềm, chúng ta có thể xây dựng các mô hình có thể tạo ra phụ đề cho một hình ảnh (image captioning). Đây là những gì mà em sẽ thực hiện trong dự án này, với mạng nơ-ron tích chập – CNN và LSTM.

## Mục tiêu thực hiện đề tài

Mục tiêu của đồ án này là xây dựng một mô hình có khả năng mô tả nội dung của một bức ảnh đầu vào của người dùng được gửi lên từ trình duyệt đến server. Server sẽ đưa ảnh qua mô hình để xử lý và trả về nội dung mô tả của ảnh và gửi về hiển thị thông qua trình duyệt cho người dùng. Điều này không chỉ mang lại giá trị nghiên cứu mà còn hiện thực được một ứng dụng mà người dùng có thể thao tác trực tiếp từ trình duyệt của họ.

# CƠ SỞ LÝ THUYẾT

## Module AI

### Mạng nơ-ron tích chập

1. Tổng quan:

Một mạng nơ ron tích chập (CNN) là một kiến trúc mạng nơ ron học sâu thường được sử dụng trong Thị giác máy tính. Thị giác máy tính là một lĩnh vực của trí tuệ nhân tạo cho phép máy tính hiểu và giải thích về hình ảnh hoặc trực quan hóa dữ liệu.

Về học máy, CNN được ứng dụng cho các tập dữ liệu khác nhau như hình ảnh, âm thanh và văn bản. Các loại mạng nơ ron khác nhau được sử dụng cho các mục đích khác nhau, ví dụ như để dự đoán đoạn văn chúng ta sử dụng LSTM, tương tự như vậy cho việc phân loại hình ảnh chúng ta mạng CNN.

1. Ưu và nhược điểm:

Ưu điểm:

* Hoạt động tốt trong việc phát hiện đặc trưng trong hình ảnh, video và âm thanh.
* Không cần phải trích xuất đặc trưng ảnh một cách thủ công.
* Có khả xử lý lượng lớn dữ liệu và đạt độ chính xác cao.

Nhược điểm:

* Tốn kém tài nguyên khi tính toán, đòi hỏi rất nhiều bộ nhớ.
* Có khả năng bị overfitting nếu dữ liệu không đủ lớn.
* Yêu cầu lượng lớn dữ liệu đã được dán nhãn.

### Long short term memory – LSTM

1. Tổng quan:

Mạng trí nhớ ngắn hạn định hướng dài hạn còn được viết tắt là LSTM làm một kiến trúc đặc biệt của RNN có khả năng học được sự phục thuộc trong dài hạn (long-term dependencies) được giới thiệu bởi Hochreiter & Schmidhuber (1997). Kiến trúc này đã được phổ biến và sử dụng rộng rãi cho tới ngày nay. LSTM đã tỏ ra khắc phục được rất nhiều những hạn chế của RNN trước đây về triệt tiêu đạo hàm. Tuy nhiên cấu trúc của chúng có phần phức tạp hơn mặc dù vẫn dữ được tư tưởng chính của RNN là sự sao chép các kiến trúc theo dạng chuỗi.

1. Ưu và nhược điểm:

Ưu điểm:

* Tránh được vấn đề vanishing gradient.

Nhược điểm:

* Tăng độ khó cho quá trình huấn luyện.

## Module website

### Front-end

#### ReactJS

1. Tổng quan:

React là một thư viện UI phát triển tại Facebook để hỗ trợ việc xây dựng những thành phần (components) UI có tính tương tác cao, có trạng thái và có thể sử dụng lại được. React được sử dụng tại Facebook trong production, và www.instagram.com được viết hoàn toàn trên React.

Một trong những điểm hấp dẫn của React là thư viện này không chỉ hoạt động trên phía client, mà còn được render trên server và có thể kết nối với nhau. React so sánh sự thay đổi giữa các giá trị của lần render này với lần render trước và cập nhật ít thay đổi nhất trên DOM.

1. Ưu và nhược điểm:

Ưu điểm:

* Hiệu năng cao đối với các ứng dụng có dữ liệu thay đổi liên tục, dễ dàng cho bảo trì và sửa lỗi.
* Giúp viết các đoạn code JS dễ dàng hơn.
* Render được ở phía Server.

Nhược điểm:

* Khó tiếp cận đối với người mới.

### Back-end

#### Flask

1. Tổng quan:

Flask là một micro web framework được viết bằng Python, không yêu cầu tool hay thư viện cụ thể nào. “Micro” không có nghĩa là thiếu chức năng mà “micro” theo triết lý thiết kế là cung cấp một lõi chức năng “súc tích” nhất cho ứng dụng web nhưng người dùng có thể mở rộng bất cứ lúc nào. Flask luôn hỗ trợ các thành phần tiện ích mở rộng cho ứng dụng như tích hợp cơ sở dữ liệu, xác thực biểu mẫu, xử lý upload, các công nghệ xác thực, template, email, RESTful..., chỉ khác là khi nào bạn muốn thì bạn mới đưa vào thôi. Người dùng có thể tập trung xây dựng web application ngay từ đầu trong một khoảng thời gian rất ngắn và có thể phát triển quy mô của ứng dụng tùy theo yêu cầu.

1. Ưu và nhược điểm:

Ưu điểm:

* Tùy biến linh hoạt.
* Dễ cài đặt, phát triển.
* Nhiều tài liệu tham khảo.
* Cộng đồng lớn.

Nhược điểm:

* Có kiến thức nền tảng vững chắc.
* Không phù hợp cho người mới.

# MÔ HÌNH ĐỀ XUẤT

## Mô hình kết hợp giữa CNN và LSTM

Một trong những mô hình nơ rơn thú vị và thực tế hữu ích nhất đến từ sự pha trộn các loại mạng khác nhau với nhau thành các mô hình lai.Ví dụ: Hãy xem xét nhiệm vụ tạo tiêu đề cho hình ảnh. Trong trường hợp này, chúng ta có một hình ảnh đầu vào và một chuỗi đầu ra là tiêu đề cho hình ảnh vào.

Chúng ta có thể mô hình hóa điều này như một nhiệm vụ dự đoán chuỗi một-với-nhiều không ? Tất nhiên là có, nhưng làm thế nào để LSTM hoặc bất kỳ mô hình dự đoán chuỗi nào khác hiểu được hình ảnh đầu vào. Chúng tôi không thể nhập trực tiếp vào bộ căng hình ảnh RGB vì chúng không được trang bị để làm việc với các đầu vào như vậy. Đầu vào có cấu trúc không gian, giống như hình ảnh, không thể được mô hình dễ dàng với Vanilla LSTM tiêu chuẩn.

Chúng ta có thể trích xuất một số tính năng từ hình ảnh đầu vào không? Câu trả lời là có, đây chính xác là những gì chúng ta cần làm để sử dụng kiến trúc LSTM cho mục đích của chúng ta. Chúng ta có thể sử dụng kiến trúc CNN sâu để trích xuất các tính năng từ hình ảnh mà sau đó được đưa vào kiến trúc LSTM để phát ra tiêu đề.

Đây được gọi là mô hình LSTM của CNN, được thiết kế đặc biệt cho các vấn đề dự đoán chuỗi với các đầu vào chẳng hạn như hình ảnh hoặc video. Kiến trúc này liên quan đến việc sử dụng các lớp mạng nơ ron tích chập (CNN) để trích xuất các đặc trưng trên dữ liệu đầu vào kết hợp với LSTMs để thực hiện dự đoán chuỗi trên các vector đặc trưng.

Nói tóm lại, CNN và LSTM là một lớp mô hình sâu cả về không gian và thời gian và nằm ở ranh giới của Thị giác máy tính và Xử lý Ngôn ngữ Tự nhiên. Những mô hình này có tiềm năng rất lớn và ngày càng được sử dụng cho nhiều nhiệm vụ phức tạp như phân loại văn bản, chuyển đổi video, v.v. Đây là một kiến trúc chung của mô hình LSTM của CNN.

# THỰC NGHIỆM

## Dữ liệu thực nghiệm

Tên bộ dữ liệu: Flickr\_8K.

Nguồn gốc: Kaggle dataset.

Kích thước: 1 GB.

Định dạng mẫu: jpg.

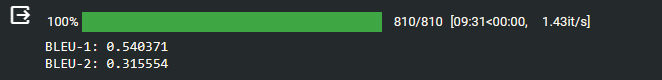
## Cài đặt thực nghiệm

### Môi trường

Sử dụng Google Collab.

### Đánh giá mô hình

Mô hình được đánh giá bằng cách sử dụng độ đo BLEU score trên tệp kiểm thử.



Hình 4.1: Đánh giá kết quả bằng BLEU score

# KẾT LUẬN

## Kết luận

Trong đề tài lần này, em đã đặt mục tiêu là mô hình có thể trả lời chính xác cho hình ảnh đầu vào từ phía người dùng. Nghiên cứu nhầm tận dụng lượng lớn dữ liệu hiện có và sức mạnh của máy tính về phần cứng lẫn phần mềm để thực hiện.

## Hướng phát triển

* Nâng cấp mô hình mạnh hơn để có thể đưa ra câu trả lời chính xác hơn và nhanh hơn.
* Tích hợp thành một module nhỏ cho một đồ án lớn hơn có chủ đề liên quan đến lấy tiêu đề từ ảnh.

# TÀI LIỆU THAM KHẢO

Tiếng Việt

phamdinhkhanh (2019), Lý thuyết về mạng LSTM phần 2, https://phamdinhkhanh.github.io/2019/04/22/Ly\_thuyet\_ve\_mang\_LSTM.htm

Tiếng Anh

Zeynep Tufekci (2022), https://medium.com/@zeyneptufekci.etu/cnn-lstm-for-image-captioning-with-progressive-loading-52a7