TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN NGỌC HƯƠNG GIANG - 52100019**

**HUỲNH THỊ MỘNG TRINH - 52100132**

**IMAGE CAPTION**

**BÁO CÁO CUỐI KỲ**

**NHẬP MÔN  
THỊ GIÁC MÁY TÍNH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN NGỌC HƯƠNG GIANG - 52100019**

**HUỲNH THỊ MỘNG TRINH - 52100132**

**IMAGE CAPTION**

**BÁO CÁO CUỐI KỲ**

**NHẬP MÔN  
THỊ GIÁC MÁY TÍNH**

Người hướng dẫn

**TS. Phạm Văn Huy**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**LỜI CẢM ƠN**

Chúng em xin chân thành gửi lời cảm ơn đến quý thầy cô của khoa Công nghệ thông tin trường Đại học Tôn Đức Thắng nói chung và thầy Phạm Văn Huy nói riêng đã đồng hành cùng em trong môn học Nhập môn thị giác máy tính này.

Chúng em thật sự rất biết ơn sự giảng dạy tận tình của thầy nhờ đó mà chúng em có thể hoàn thành bài báo cáo một cách tốt nhất có thể.

Trong quá trình thực hiện báo cáo không tránh khỏi còn nhiều sai sót, chúng em kính mong nhận được những ý kiến góp ý từ quý thầy cô để bài báo cáo được hoàn thiện hơn.

Một lần nữa, chúng em xin chân thành cảm ơn quý thầy cô.

*TP. Hồ Chí Minh, ngày 11 tháng 3 năm 2025*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

Giang

Nguyễn Ngọc Hương Giang

Trinh

Huỳnh Thị Mộng Trinh

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng chúng 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 Báo cáo 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 chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung Báo cáo của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 11 tháng 3 năm 2025*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

Giang

Nguyễn Ngọc Hương Giang

Trinh

Huỳnh Thị Mộng Trinh

**IMAGE CAPTION**

**TÓM TẮT**

Đề tài này tập trung vào việc xây dựng hệ thống Image Captioning, cho phép tạo chú thích tự động cho hình ảnh bằng trí tuệ nhân tạo. Hệ thống sử dụng mô hình BLIP (Bootstrapping Language-Image Pre-training) để phân tích nội dung ảnh và sinh mô tả văn bản phù hợp. Người dùng có thể tải ảnh lên thông qua giao diện web, sau đó ảnh được xử lý trên server bằng mô hình AI, và kết quả sẽ hiển thị trực tiếp trên trang web.

Hệ thống được triển khai theo kiến trúc client-server, trong đó backend đảm nhận nhiệm vụ tiền xử lý ảnh, trích xuất đặc trưng và sinh chú thích, còn frontend hiển thị kết quả cho người dùng. Công nghệ sử dụng gồm PyTorch, Transformers, Flask/FastAPI cho backend và React/HTML, CSS, JavaScript cho giao diện web.

Ứng dụng của hệ thống rất đa dạng, từ tự động tạo mô tả sản phẩm trong thương mại điện tử, hỗ trợ người khiếm thị nhận diện nội dung hình ảnh, đến quản lý và phân loại hình ảnh thông minh. Với sự kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên, hệ thống không chỉ tối ưu hóa việc mô tả ảnh mà còn nâng cao trải nghiệm người dùng trong nhiều lĩnh vực.

**IMAGE CAPTION**

**ABSTRACT**

This project focuses on developing an Image Captioning system that enables automatic image annotation using artificial intelligence. The system employs the BLIP (Bootstrapping Language-Image Pre-training) model to analyze image content and generate appropriate textual descriptions. Users can upload images through a web interface, after which the images are processed on the server by the AI model, and the generated captions are displayed directly on the website.

The system is implemented using a client-server architecture, where the backend is responsible for image preprocessing, feature extraction, and caption generation, while the frontend displays the results to the user. The technologies used include PyTorch, Transformers, Flask/FastAPI for the backend and React/HTML, CSS, JavaScript for the web interface.

This system has a wide range of applications, from automatically generating product descriptions in e-commerce, assisting visually impaired individuals in understanding image content, to intelligent image management and classification. By integrating computer vision and natural language processing, the system not only optimizes image description processes but also enhances user experience across various domains.

**MỤC LỤC**

[DANH MỤC HÌNH VẼ vii](#_Toc14615)

[DANH MỤC BẢNG BIỂU viii](#_Toc11883)

[DANH MỤC CÁC CHỮ VIẾT TẮT ix](#_Toc15348)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc24575)

[1.1 Lý do chọn đề tài 1](#_Toc3678)

[1.2 Mục tiêu thực hiện đề tài 2](#_Toc29296)

[1.3 Phạm vi nghiên cứu 3](#_Toc31896)

[1.4 Ý nghĩa của đề tài 3](#_Toc28468)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 4](#_Toc1947)

[2.1 Tổng quan về Image Captioning 4](#_Toc17256)

[2.1.1 Giới thiệu về bài toán Image Captioning 4](#_Toc16358)

[2.1.2 Các phương pháp tiếp cận 4](#_Toc15848)

[2.2 Mô hình BLIP (Bootstrapped Language-Image Pretraining) 5](#_Toc8161)

[2.2.1 Lịch sử ra đời: 5](#_Toc681)

[2.2.2 Kiến trúc mô hình 6](#_Toc10796)

[2.2.3 Ưu nhược điểm 8](#_Toc17985)

[2.2.4 Ứng dụng: 9](#_Toc29587)

[CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN 10](#_Toc31385)

[3.1 Tổng quan về hệ thống 10](#_Toc15746)

[3.2 Thu thập và xử lý dữ liệu 10](#_Toc7687)

[3.3 Triển khai mô hình BLIP 11](#_Toc11560)

[3.4 Xây dựng giao diện người dùng 11](#_Toc10344)

[CHƯƠNG 4. HIỆN THỰC CODE 11](#_Toc19630)

[4.1 Xây dựng model 11](#_Toc11771)

[4.2 Xây dựng code call API 12](#_Toc12076)

[4.3 Xây dựng giao diện 14](#_Toc19967)

[CHƯƠNG 5. KẾT QUẢ VÀ ĐÁNH GIÁ 14](#_Toc5444)

[5.1 Kết quả đạt được 14](#_Toc29546)

[5.2 Hạn chế 15](#_Toc8071)

[CHƯƠNG 6. KẾT LUẬN 16](#_Toc24627)

[TÀI LIỆU THAM KHẢO 17](#_Toc17682)

# DANH MỤC HÌNH VẼ

[Hình 2.1 BLIP Model Architecture and Pre-training Objectives 7](#_Toc192632922)

[Hình 4.1 Giao diện tạo chú thích ảnh 15](#_Toc192632923)

[Hình 5.1 Kết quả mô tả ảnh 17](#_Toc192632924)

[Hình 5.2 Kết quả mô tả ảnh 17](#_Toc192632925)

# DANH MỤC BẢNG BIỂU

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| ViT | Vision Transformer |
| R-CNN | Region-based Convolutional Neural Network |
| CNN | Convolutional Neural Network |

# MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Trong thời đại công nghệ phát triển nhanh chóng, Image Captioning trở thành một trong những lĩnh vực quan trọng kết hợp giữa Thị giác máy tính (Computer Vision) và Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP). Việc nghiên cứu và phát triển hệ thống Image Captioning không chỉ mang ý nghĩa học thuật mà còn có nhiều ứng dụng thực tiễn, hỗ trợ cải thiện trải nghiệm người dùng trong nhiều lĩnh vực khác nhau.

Trước tiên, Image Captioning giúp máy tính hiểu và mô tả nội dung hình ảnh bằng ngôn ngữ tự nhiên, mở ra nhiều khả năng ứng dụng trong thực tế. Khác với các bài toán nhận diện đối tượng (Object Detection) chỉ dừng lại ở việc xác định danh tính của các thực thể trong ảnh, Image Captioning cung cấp ngữ cảnh rõ ràng hơn bằng cách tạo ra mô tả hoàn chỉnh và có ý nghĩa.

Bên cạnh đó, đề tài này có ứng dụng thực tiễn cao, đặc biệt trong các lĩnh vực như:

* Hỗ trợ người khiếm thị: Giúp họ tiếp cận nội dung hình ảnh thông qua mô tả văn bản hoặc giọng nói.
* Tìm kiếm hình ảnh thông minh: Cải thiện khả năng tìm kiếm dựa trên mô tả nội dung thay vì chỉ dựa vào từ khóa hoặc metadata.
* Tự động tạo mô tả sản phẩm trong thương mại điện tử: Giúp các nền tảng thương mại tạo mô tả sản phẩm một cách tự động, tiết kiệm thời gian và chi phí.
* Quản lý và phân loại hình ảnh: Hỗ trợ tổ chức dữ liệu hình ảnh hiệu quả hơn, đặc biệt trong các kho lưu trữ lớn.

Ngoài ra, nghiên cứu Image Captioning còn giúp nâng cao hiểu biết về các mô hình học sâu hiện đại như BLIP (Bootstrapping Language-Image Pre-training), Transformers, CNN-LSTM và các cơ chế Attention, từ đó giúp tiếp cận sâu hơn với các kỹ thuật tiên tiến trong AI.

Với những lý do trên, việc nghiên cứu và triển khai hệ thống Image Captioning không chỉ là một bài toán thú vị về mặt kỹ thuật mà còn có tiềm năng ứng dụng lớn trong thực tế, giúp cải thiện đáng kể khả năng tương tác giữa con người và máy tính thông qua hình ảnh và ngôn ngữ.

## Mục tiêu thực hiện đề tài

Mục tiêu của đề tài Image Captioning là xây dựng một hệ thống tự động tạo chú thích văn bản cho hình ảnh với độ chính xác cao và mô tả tự nhiên. Cụ thể, đề tài hướng đến các mục tiêu sau:

* Phát triển mô hình nhận diện và mô tả hình ảnh chính xác:
* Xây dựng hệ thống có khả năng phân tích hình ảnh và tạo ra mô tả phù hợp bằng ngôn ngữ tự nhiên.
* Cải thiện độ chính xác của mô hình trong việc nhận diện các thực thể trong ảnh, bất kể điều kiện ánh sáng, góc chụp hay sự tương đồng giữa các đối tượng.
* Đảm bảo tính tự nhiên và đa dạng ngữ nghĩa trong mô tả:
* Tránh tình trạng lặp lại hoặc mô tả đơn điệu bằng cách tạo ra các câu văn có sự phong phú và linh hoạt.
* Áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên để đảm bảo câu chú thích có cấu trúc rõ ràng và gần gũi với cách diễn đạt của con người.
* Tích hợp khả năng hiểu ngữ cảnh của hình ảnh:
* Không chỉ nhận diện các đối tượng trong ảnh mà còn hiểu được bối cảnh và mối quan hệ giữa chúng.
* Cải thiện khả năng mô tả hành động hoặc môi trường xung quanh để tăng tính chính xác và tự nhiên cho câu chú thích.
* Tăng cường khả năng tổng quát hóa của mô hình:
* Xây dựng mô hình có thể làm việc tốt với nhiều loại hình ảnh khác nhau, từ động vật hoang dã đến vật nuôi trong nhà hoặc các đối tượng phức tạp hơn.
* Ứng dụng các phương pháp học sâu như BLIP để nâng cao hiệu suất và độ tin cậy của hệ thống.
* Xây dựng hệ thống có khả năng triển khai thực tế:
* Thiết kế giao diện web trực quan, cho phép người dùng tải ảnh lên và nhận kết quả mô tả nhanh chóng.
* Đảm bảo hệ thống hoạt động ổn định, có thể mở rộng và tối ưu hiệu suất để phục vụ nhiều ứng dụng khác nhau như hỗ trợ người khiếm thị, tìm kiếm hình ảnh thông minh hoặc tự động tạo mô tả sản phẩm.

## Phạm vi nghiên cứu

Phạm vi nghiên cứu của đề tài tập trung vào việc phát triển một hệ thống Image Captioning dựa trên mô hình học sâu. Các nội dung chính bao gồm:

* Dữ liệu nghiên cứu: Sử dụng các bộ dữ liệu chuẩn cho bài toán Image Captioning như Flicker8k để huấn luyện và đánh giá mô hình.
* Mô hình nghiên cứu: Áp dụng các mô hình tiên tiến là BLIP kết hợp Attention để cải thiện hiệu suất.
* Công nghệ sử dụng:
  + Ngôn ngữ lập trình: Python.
  + Framework: TensorFlow, PyTorch.
  + Thư viện hỗ trợ: Hugging Face Transformers, NLTK.
  + Giao diện web: HTML, CSS, FastAPI/Flask để triển khai API.
* Hạn chế: Nghiên cứu tập trung vào mô tả hình ảnh đơn lẻ, chưa xem xét video hoặc chuỗi hình ảnh liên tục.

## Ý nghĩa của đề tài

Đề tài nghiên cứu về Image Captioning mang lại nhiều ý nghĩa quan trọng cả về mặt học thuật lẫn ứng dụng thực tế:

* Góp phần phát triển công nghệ AI: Đề tài giúp nâng cao hiểu biết về cách kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên, góp phần vào sự phát triển của trí tuệ nhân tạo.
* Ứng dụng rộng rãi trong thực tế:
  + Hỗ trợ người khiếm thị tiếp cận thông tin thông qua mô tả hình ảnh bằng văn bản hoặc âm thanh.
  + Cải thiện tìm kiếm hình ảnh thông minh dựa trên nội dung thay vì từ khóa.
  + Tự động hóa quá trình tạo mô tả sản phẩm trong thương mại điện tử.
  + Giúp quản lý dữ liệu hình ảnh hiệu quả hơn, đặc biệt trong kho lưu trữ lớn.
  + Thúc đẩy nghiên cứu trong lĩnh vực học sâu: Giúp nghiên cứu sinh, lập trình viên tiếp cận với các mô hình hiện đại như Transformer, CNN-LSTM, Attention, BLIP và cải thiện kỹ năng xây dựng mô hình AI.

# CƠ SỞ LÝ THUYẾT

## Tổng quan về Image Captioning

### Giới thiệu về bài toán Image Captioning

Image Captioning là một bài toán trong lĩnh vực thị giác máy tính kết hợp với xử lý ngôn ngữ tự nhiên, với mục tiêu tạo ra các mô tả bằng ngôn ngữ tự nhiên cho hình ảnh. Hệ thống Image Captioning giúp máy tính có thể hiểu và diễn giải nội dung của một bức ảnh, không chỉ nhận diện các đối tượng trong ảnh mà còn mô tả được ngữ cảnh và hành động liên quan.

### Các phương pháp tiếp cận

* Phương pháp truyền thống:
  + Dựa trên quy tắc và template: Các phương pháp này sử dụng các quy tắc tiền định để mô tả hình ảnh, chẳng hạn như gán nhãn cho từng đối tượng và kết hợp theo một mẫu câu cố định.
  + Hạn chế: Mô tả thiếu linh hoạt, không thể thích nghi với hình ảnh mới hoặc ngữ cảnh phức tạp.
* Phương pháp hiện đại (dựa trên học sâu - Deep Learning):
  + Sử dụng mô hình CNN-RNN: Kết hợp mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng hình ảnh và mạng nơ-ron hồi tiếp (RNN) hoặc LSTM để sinh mô tả văn bản.
  + Transformer và Attention Mechanism: Các mô hình tiên tiến như Transformer sử dụng cơ chế Attention để học cách tập trung vào các vùng quan trọng trong ảnh, giúp tạo ra mô tả chính xác và tự nhiên hơn.
  + BLIP (Bootstrapping Language-Image Pre-training): Một trong những mô hình tiên tiến nhất hiện nay, giúp cải thiện đáng kể hiệu suất trong bài toán Image Captioning.
* Lý do chọn BLIP
* Hiệu suất cao, được pre-trained trên tập dữ liệu lớn.
* Dễ sử dụng qua Hugging Face.
* Kết hợp thị giác và ngôn ngữ, tối ưu cho bài toán Image Captioning.

## Mô hình BLIP (Bootstrapped Language-Image Pretraining)

### Lịch sử ra đời:

BLIP (Bootstrapping Language-Image Pre-training) được giới thiệu trong bài báo BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, được công bố trên arXiv vào ngày 15 tháng 2 năm 2022.

BLIP được phát triển nhằm giải quyết những hạn chế của các mô hình tiền huấn luyện thị giác - ngôn ngữ (VLP) hiện có, đặc biệt là vấn đề về kiến trúc mô hình và chất lượng dữ liệu huấn luyện. Các tác giả nhận thấy rằng hầu hết các phương pháp trước đó đều gặp khó khăn khi áp dụng linh hoạt cho cả hai nhóm tác vụ: hiểu (understanding) và sinh (generation). Bên cạnh đó, dữ liệu huấn luyện thường được lấy từ các cặp hình ảnh - văn bản thu thập từ web, nhưng chứa nhiều nhiễu, ảnh hưởng đến chất lượng mô hình.

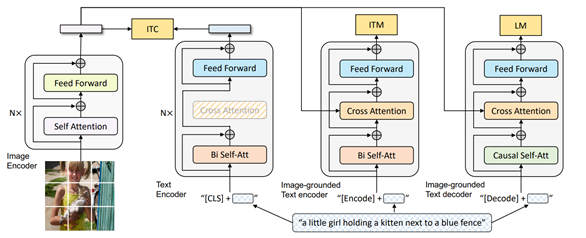
Để khắc phục những vấn đề này, nhóm nghiên cứu đã đề xuất BLIP với hai cải tiến chính:

Multimodal Mixture of Encoder-Decoder (MED): một kiến trúc linh hoạt có thể hoạt động như một bộ mã hóa hoặc bộ giải mã, giúp mô hình có thể thích ứng với nhiều nhiệm vụ khác nhau.

Captioning and Filtering (CapFilt): một phương pháp giúp lọc bỏ các mô tả kém chất lượng và tạo ra các mô tả hình ảnh chính xác hơn để cải thiện dữ liệu huấn luyện.

Những cải tiến này giúp BLIP đạt được hiệu suất vượt trội trên nhiều tác vụ thị giác - ngôn ngữ, đồng thời cải thiện đáng kể chất lượng dữ liệu huấn luyện so với các phương pháp trước đây như CLIP, ALBEF hay SimVLM..

### Kiến trúc mô hình



Hình 2.1 BLIP Model Architecture and Pre-training Objectives

(Nguồn: J. Li, D. Li, C. Xiong, and S. Hoi, *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. Salesforce Research, 2022. [Online]. Available: [https://github.com/salesforce/BLIP](https://github.com/salesforce/BLIP" \t "_new).)

Hình ảnh mô tả kiến trúc và quá trình huấn luyện của mô hình BLIP, bao gồm ba thành phần chính:

* Image Encoder (Bộ mã hóa hình ảnh)
* Nhận một hình ảnh đầu vào và chia nó thành các phần nhỏ (patch).
* Sử dụng mạng Transformer với Self-Attention để trích xuất đặc trưng của hình ảnh.
* Kết quả của bộ mã hóa hình ảnh được sử dụng trong các bước tiếp theo.
* Text Encoder (Bộ mã hóa văn bản)
* Nhận một câu mô tả văn bản (ví dụ: "a little girl holding a kitten next to a blue fence").
* Áp dụng Bi-directional Self-Attention để học thông tin của văn bản.
* Có thêm Cross Attention để kết hợp thông tin từ hình ảnh.
* Text Decoder (Bộ giải mã văn bản)
* Dùng để sinh văn bản mô tả hình ảnh.
* Sử dụng Causal Self-Attention thay vì Bi-directional Self-Attention.
* Dựa vào thông tin từ hình ảnh và văn bản để sinh ra mô tả mới.

BLIP (Bootstrapping Language-Image Pre-training) là một framework tiền huấn luyện đa phương tiện (Vision-Language Pre-training - VLP) được thiết kế để thực hiện cả hai nhiệm vụ hiểu và tạo văn bản dựa trên hình ảnh. Kiến trúc của BLIP bao gồm các thành phần chính sau:

Multimodal Mixture of Encoder-Decoder (MED):

* Unimodal Encoder: Mã hóa riêng biệt hình ảnh và văn bản. Hình ảnh được mã hóa bằng Visual Transformer (ViT), trong khi văn bản được mã hóa bằng một mô hình tương tự BERT.
* Image-grounded Text Encoder: Thêm các lớp cross-attention để tích hợp thông tin hình ảnh vào quá trình mã hóa văn bản. Mô hình này được huấn luyện với mục tiêu Image-Text Matching (ITM) để học cách kết hợp thông tin đa phương tiện.
* Image-grounded Text Decoder: Thay thế các lớp self-attention trong encoder bằng các lớp causal self-attention để tạo văn bản dựa trên hình ảnh. Mô hình này được huấn luyện với mục tiêu Language Modeling (LM) để tạo ra các mô tả văn bản từ hình ảnh.

Captioning and Filtering (CapFilt):

* Captioner: Một mô hình decoder được tinh chỉnh để tạo ra các mô tả văn bản từ hình ảnh. Nó tạo ra các captions tổng hợp cho các hình ảnh từ web.
* Filter: Một mô hình encoder được tinh chỉnh để lọc ra các cặp hình ảnh-văn bản nhiễu. Nó loại bỏ các captions không phù hợp với hình ảnh.

### Ưu nhược điểm

* Ưu điểm:

- Hiệu suất vượt trội: BLIP đạt được kết quả state-of-the-art trên nhiều nhiệm vụ đa phương thức, bao gồm truy xuất hình ảnh-văn bản, tạo chú thích hình ảnh, và trả lời câu hỏi trực quan.

- Khả năng chuyển đổi linh hoạt: BLIP có thể chuyển đổi linh hoạt giữa các nhiệm vụ hiểu và tạo văn bản, nhờ vào kiến trúc MED.

- Xử lý dữ liệu nhiễu hiệu quả: CapFilt giúp cải thiện chất lượng của tập dữ liệu tiền huấn luyện bằng cách tạo và lọc chú thích, giúp mô hình học từ dữ liệu nhiễu một cách hiệu quả hơn.

- Khả năng mở rộng: BLIP có thể mở rộng quy mô dữ liệu và mô hình, đạt được hiệu suất tốt hơn với tập dữ liệu lớn hơn và kiến trúc mô hình lớn hơn.

* Nhược điểm:

- Chi phí tính toán cao: Việc tiền huấn luyện và tạo chú thích tổng hợp đòi hỏi nhiều tài nguyên tính toán, đặc biệt là khi sử dụng kiến trúc lớn như ViT-L.

- Phụ thuộc vào chất lượng chú thích tổng hợp: Hiệu suất của BLIP phụ thuộc vào chất lượng của các chú thích tổng hợp được tạo ra bởi captioner. Nếu các

chú thích này không đủ đa dạng hoặc chính xác, hiệu suất của mô hình có thể bị ảnh hưởng.

- Khó khăn trong việc tối ưu hóa: Việc tối ưu hóa các tham số và quy trình tiền huấn luyện có thể phức tạp, đòi hỏi nhiều thử nghiệm và điều chỉnh.

### Ứng dụng:

Kiến trúc BLIP đã được ứng dụng rộng rãi trong các bài toán kết hợp giữa hình ảnh và văn bản, mang lại hiệu suất cao trong nhiều tác vụ đa phương thức. Dưới đây là một số ứng dụng chính của BLIP:

* Image Captioning (Tạo chú thích hình ảnh):
* BLIP có khả năng tạo ra các mô tả văn bản chính xác và chi tiết từ hình ảnh.
* Ứng dụng: Hỗ trợ người khiếm thị hiểu nội dung hình ảnh, tạo chú thích tự động cho các nền tảng chia sẻ hình ảnh như Instagram, Flickr.
* Image-Text Retrieval (Tìm kiếm hình ảnh bằng văn bản và ngược lại):
* BLIP có thể tìm kiếm hình ảnh phù hợp với mô tả văn bản hoặc tìm kiếm văn bản phù hợp với hình ảnh.
* Ứng dụng: Tìm kiếm hình ảnh trên các công cụ như Google Images, hoặc tìm kiếm sản phẩm dựa trên mô tả văn bản trong thương mại điện tử.
* Visual Question Answering (VQA - Trả lời câu hỏi dựa trên hình ảnh):
* BLIP có thể trả lời các câu hỏi liên quan đến nội dung của hình ảnh.
* Ứng dụng: Hỗ trợ giáo dục (ví dụ: trả lời câu hỏi về hình ảnh trong sách giáo khoa), hoặc trợ lý ảo thông minh.
* Visual Dialog (Đối thoại dựa trên hình ảnh):
* BLIP có thể tham gia vào các cuộc đối thoại liên quan đến hình ảnh, dựa trên cả hình ảnh và lịch sử đối thoại.
* Ứng dụng: Trợ lý ảo thông minh, chatbot hỗ trợ khách hàng dựa trên hình ảnh sản phẩm.
* Zero-shot Transfer Learning (Chuyển giao học không cần huấn luyện lại):
* BLIP có khả năng áp dụng trực tiếp vào các tác vụ mới mà không cần huấn luyện lại, nhờ vào khả năng tổng quát hóa mạnh mẽ.
* Ứng dụng: Áp dụng nhanh chóng vào các bài toán mới trong lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên.
* Video-Language Tasks (Các tác vụ kết hợp video và ngôn ngữ):
* BLIP có thể được áp dụng vào các tác vụ như tìm kiếm video bằng văn bản (text-to-video retrieval) hoặc trả lời câu hỏi dựa trên video (video question answering).
* Ứng dụng: Tìm kiếm video trên YouTube, phân tích nội dung video trong an ninh hoặc giáo dục.

# PHƯƠNG PHÁP THỰC HIỆN

## Tổng quan về hệ thống

Hệ thống nhận diện và tạo caption từ hình ảnh bao gồm hai thành phần chính: backend và frontend.

* Backend:
  + Sử dụng Flask để xây dựng API nhận dạng hình ảnh.
  + Mô hình BLIP từ Hugging Face được tích hợp và xử lý dữ liệu.
  + API nhận hình ảnh từ frontend, xử lý, sinh caption và trả về cho người dùng.
* Frontend:
  + Giao diện HTML/CSS/JavaScript cho phép người dùng tải hình ảnh lên.
  + Gửi request đến API Flask và nhận caption từ mô hình.
  + Hiển thị kết quả trǪ giao diện.
* Sơ đồ khối luồng dữ liệu:

Hình ảnh -> API -> Mô hình BLIP -> Caption -> Giao diện hiển thị.

## Thu thập và xử lý dữ liệu

Hệ thống cần nguồn dữ liệu hình ảnh chất lượng cao để huấn luyện và kiểm tra.

* Nguồn dữ liệu:
  + Hình ảnh tự chụp.
  + Dữ liệu từ COCO, Flickr30k, MS COCO Caption.
* Xử lý hình ảnh:
  + Chuyển đổi sang base64 để gửi qua API.
  + Resize đồng bộ, normalize giá trị pixel.
  + Biến đổi để đảm bảo phù hợp với mô hình BLIP.

## Triển khai mô hình BLIP

Mô hình BLIP (“Bootstrapped Language-Image Pretraining”) được tể tài từ Hugging Face để xử lý hình ảnh và sinh caption.

* Cài đặt và sử dụng:
* Tải và sử dụng AutoProcessor, BlipForConditionalGeneration.
* Thiết lập tham số max\_length, sử dụng CPU hoặc GPU.
* Tích hợp vào Flask API:
* API nhận base64 hình ảnh, xử lý và gửi qua mô hình BLIP.
* Trả về caption để frontend hiển thị.

## Xây dựng giao diện người dùng

Giao diện trực quan, tương tác mềm mại giúpa người dùng tải hình ảnh và xem kết quả.

Cách xây dựng:

* HTML/CSS/JavaScript cho giao diện.
* Tạo form tải hình, gửi request đến API.
* Hiển thị loading, xử lý lỗi khi API trả về.

# HIỆN THỰC CODE

## Xây dựng model

1. Tải và tiền xử lý dữ liệu

* Dataset flickr8k được tải từ Hugging Face (load\_dataset).
* Dữ liệu được chia thành tập train và validation (90% train, 10% val).
* Ảnh được hiển thị để kiểm tra với matplotlib.

1. Xây dựng dataset và dataloader

* ImageCaptioningDataset là một lớp Dataset tùy chỉnh để chuyển đổi ảnh và văn bản thành định dạng phù hợp cho mô hình BLIP bằng AutoProcessor.
* Dataloader được tạo với batch size = 2.

1. Tải mô hình BLIP

* BlipForConditionalGeneration từ Salesforce/blip-image-captioning-base được tải về.
* Mô hình được đưa lên cuda (nếu có).

1. Huấn luyện mô hình

* Dùng AdamW optimizer với learning rate 5e-5.
* Chạy qua 10 epochs, mỗi batch sẽ:
  + Trích xuất input\_ids và pixel\_values từ batch.
  + Tính loss và cập nhật trọng số bằng backpropagation.
  + Hiển thị loss sau mỗi 200 batch.

1. Lưu và nén mô hình: Mô hình sau khi huấn luyện được lưu vào thư mục /kaggle/working/Model\_BLIP\_Image\_Captioning và nén thành file zip.
2. Kiểm tra mô hình

* Chạy mô hình trên tập val để sinh caption.
* So sánh caption sinh ra với ground truth.
* Hiển thị ảnh và caption dự đoán.

1. Đánh giá mô hình bằng BLEU: Tính điểm BLEU-4 để đánh giá độ chính xác của caption dự đoán.

## Xây dựng code call API

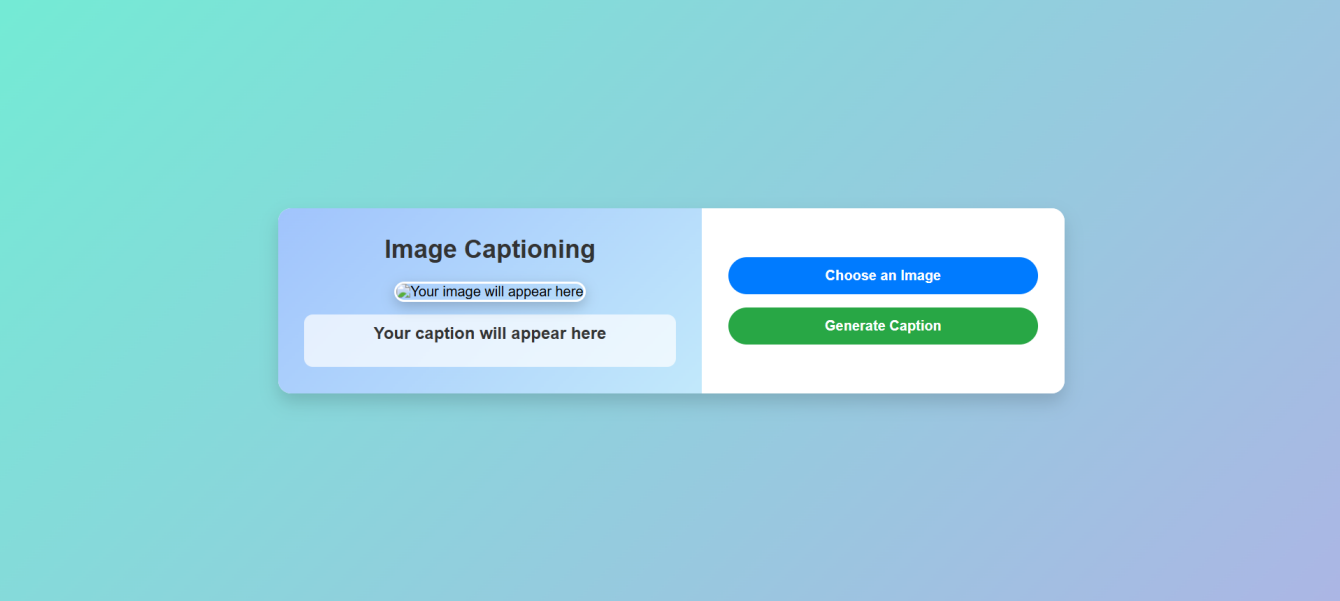
Mô tả thư viện:

* Flask: Framework để xây dựng API.
* request, jsonify: Hỗ trợ nhận dữ liệu từ client và trả về kết quả dưới dạng JSON.
* AutoProcessor, BlipForConditionalGeneration: Công cụ xử lý hình ảnh và mô hình BLIP để tạo caption.
* huggingface\_hub.login: Đăng nhập vào Hugging Face để tải model.
* torch: Thư viện cho các phép toán tensor, hỗ trợ GPU tăng tốc.
* Pandas: Có thể dùng để xử lý dữ liệu nếu cần.
* PIL (Pillow): Thư viện xử lý ảnh.
* base64, io: Hỗ trợ chuyển đổi ảnh từ định dạng base64 sang đối tượng Image.
* CORS (Cross-Origin Resource Sharing): Cho phép API hoạt động với các domain khác.

Cách hoạt động

* Nhận dữ liệu từ client:
  + API nhận một request POST với dữ liệu JSON chứa ảnh ở dạng base64.
  + data = request.get\_json() -> Lấy dữ liệu JSON từ request.
  + img\_data = data.get('image') -> Lấy chuỗi base64 từ JSON.
* Giải mã ảnh từ base64:
  + base64.b64decode(img\_data) -> Chuyển base64 thành bytes.
  + io.BytesIO(...) -> Chuyển bytes thành stream.
  + Image.open(...).convert("RGB") -> Chuyển stream thành ảnh PIL.
* Tiền xử lý ảnh:
  + processor(images=img, return\_tensors="pt") -> Chuyển đổi ảnh sang tensor phù hợp với BLIP.
  + to(device) -> Đưa dữ liệu lên GPU nếu có.
* Sinh caption bằng mô hình BLIP:
  + model.generate(pixel\_values=inputs["pixel\_values"], max\_length=50) -> Dự đoán caption.
  + processor.batch\_decode(...) -> Chuyển mã token thành câu hoàn chỉnh.
  + generated\_caption.replace(' - ', '-') -> Chuẩn hóa text.
* Trả kết quả về client: jsonify({"generated\_caption": generated\_caption}) -> Trả về caption dưới dạng JSON.

## Xây dựng giao diện



Hình 4.1 Giao diện tạo chú thích ảnh

- Giao diện ban đầu được xây dụng với 2 nút (1 nút chọn ảnh và 1 nút generate caption)

* Nút chọn tệp ảnh: Nút "Choose an Image" cho phép người dùng chọn một ảnh từ máy tính của họ.
* Nút tải ảnh và tạo caption: Nút "Upload and Generate Caption" gửi ảnh lên server để mô hình tạo caption.

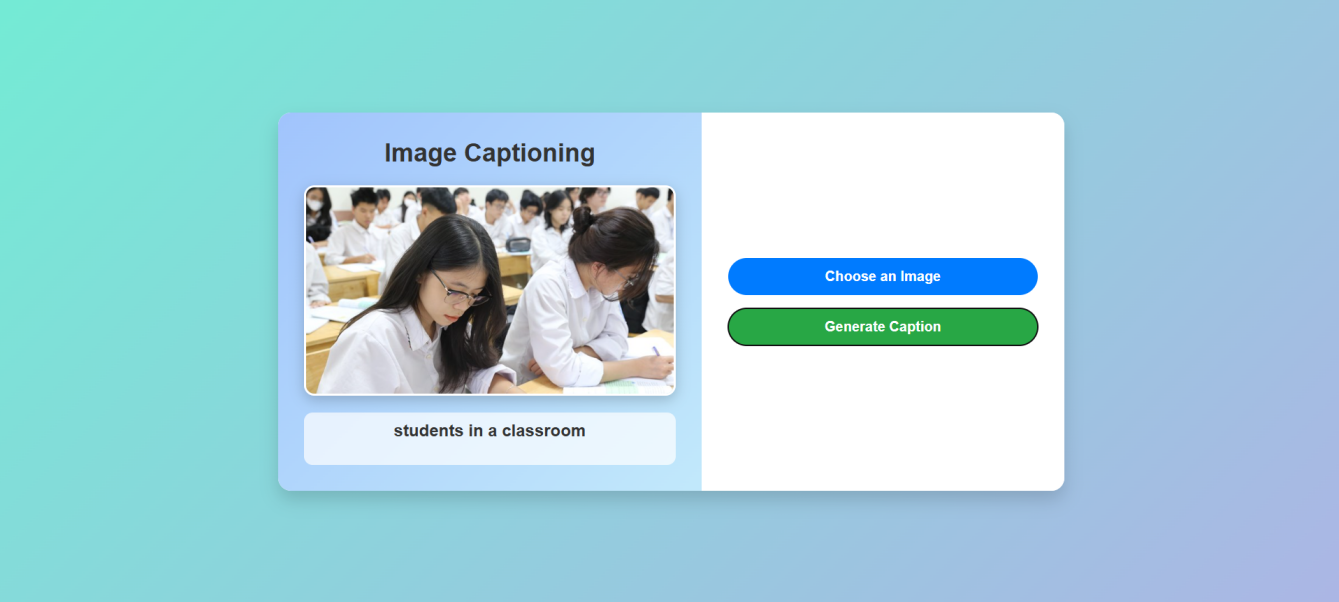
- Khu vực hiển thị ảnh: Sau khi người dùng chọn ảnh, ảnh được hiển thị ở khu vực trung tâm.

- Khu vực hiển thị caption: Phần caption được trả về và hiển thị dưới ảnh sau khi server xử lý ảnh.

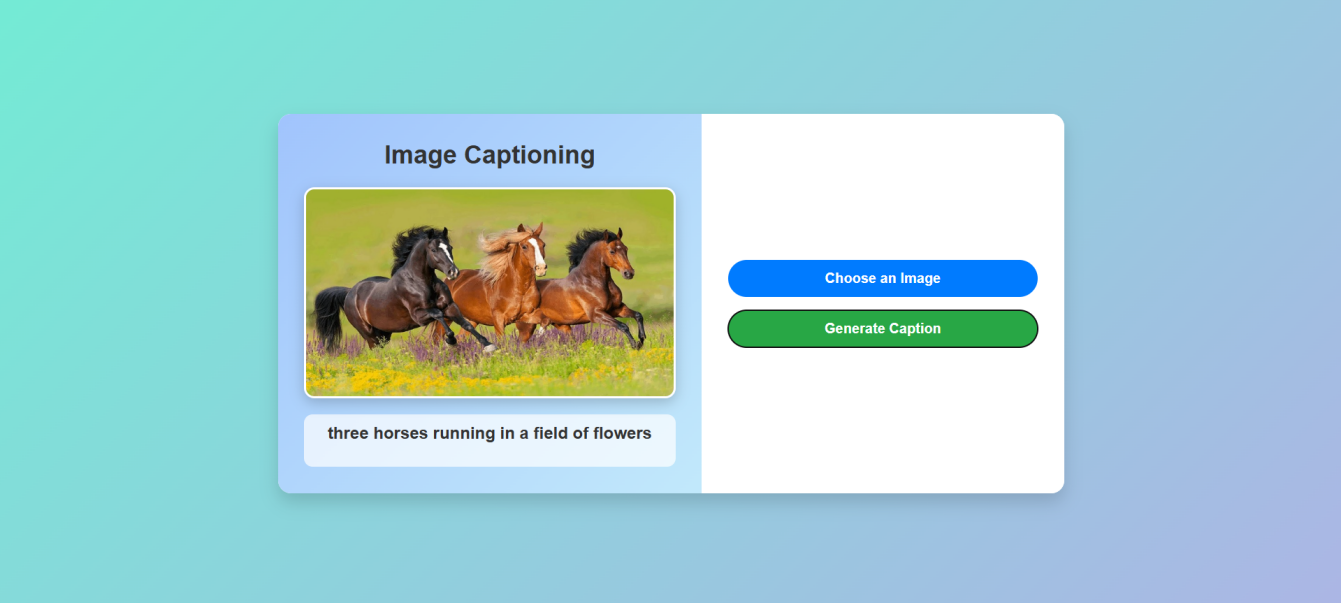
# KẾT QUẢ VÀ ĐÁNH GIÁ

## Kết quả đạt được

* Hệ thống hoạt động ổn định.
* Sinh caption chính xác, phù hợp với hình ảnh.
* Bảng minh họa kết quả caption sinh ra.
* Đánh giá bằng BLEU, METEOR, CIDEr.



Hình 5.1 Kết quả mô tả ảnh



Hình 5.2 Kết quả mô tả ảnh

## Hạn chế

* Phụ thuộc vào tài nguyên phần cứng.
* Chỉ hỗ trợ một số loại hình ảnh.

# KẾT LUẬN

Trong báo cáo này, chúng em đã triển khai thành công hệ thống Image Captioning bằng cách sử dụng mô hình BLIP (Bootstrapping Language-Image Pretraining). Hệ thống có khả năng tự động sinh mô tả cho hình ảnh đầu vào, giúp phân tích nội dung hình ảnh một cách thông minh.

Các bước thực hiện bao gồm:

* Xây dựng API Flask để nhận ảnh từ client, xử lý và trả về caption.
* Tích hợp mô hình BLIP từ Hugging Face để sinh mô tả hình ảnh.
* Tối ưu hóa hiệu suất bằng cách sử dụng GPU (nếu có) nhằm tăng tốc xử lý.
* Hỗ trợ CORS để cho phép frontend truy cập API dễ dàng.

Kết quả thực nghiệm cho thấy mô hình BLIP có thể tạo ra các caption chính xác và phù hợp với nội dung hình ảnh. Hệ thống này có tiềm năng ứng dụng rộng rãi trong nhiều lĩnh vực như hỗ trợ người khiếm thị, tìm kiếm ảnh thông minh, kiểm duyệt nội dung tự động và tạo mô tả ảnh trong thương mại điện tử.

# **TÀI LIỆU THAM KHẢO**

Tiếng Việt

1. [Thanh Dat Nguyen](https://viblo.asia/u/lil.thada) - 02/08/2023 - Giải quyết bài toán Vision-Language với BLIP-2 và InstructBLIP <https://viblo.asia/p/giai-quyet-bai-toan-vision-language-voi-blip-2-va-instructblip-W13VMeW7VY7>
2. [Nguyen Thanh Huyen](https://viblo.asia/u/huyennguyenthanh) - 02/07/2021- A Guide to Image Captioning (Part 1): Giới thiệu bài toán sinh mô tả cho ảnh <https://viblo.asia/p/a-guide-to-image-captioning-part-1-gioi-thieu-bai-toan-sinh-mo-ta-cho-anh-gAm5yr88Kdb>
3. [nttuan8](https://nttuan8.com/author/nttuan8/) - 16/06/2019 - Bài 15: Ứng dụng thêm mô tả cho ảnh (image captioning)<https://nttuan8.com/bai-15-ung-dung-them-mo-ta-cho-anh-image-captioning/>

Tiếng Anh

J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Salesforce Research*, 2022. [Online]. Available: [https://github.com/salesforce/BLIP](https://github.com/salesforce/BLIP" \t "_new).