TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

BÁO CÁO BÀI TẬP LỚN
XỬ LÝ NGÔN NGỮ TỰ NHIÊN

với đề tài: Image Captioning sử dụng mô hình CLIP theo mạng LSTM

Giảng viên hướng dẫn: ThS. Nguyễn Đình Quý

Sinh viên thực hiện: Nguyễn Hải Cường - 0174067 - 67CS1
Lã Minh Khánh - 4004267 - 67CS1
Trịnh Quỳnh Anh - 0279367 - 67CS1
Phạm Hồng Thái - 0127067 - 67CS1

Hà Nội, ngày 14 tháng 05 năm 2025

Mục lục

Lời nói đầu

Em xin chân thành cảm ơn!

Ngày 14 tháng 05 năm 2025

Từ các thành viên của nhóm

Chương I: Giới thiệu

1. Giới thiệu đề tài

Đề tài ”Tạo chú thích hình ảnh tự động (Image Captioning)” tập trung vào việc kết hợp mô hình học sâu để phân tích hình ảnh và tạo mô tả ngôn ngữ tự nhiên phù hợp. Phương pháp này ứng dụng mô hình CLIP (Contrastive Language-Image Pretraining) của OpenAI để trích xuất đặc trưng hình ảnh hiệu quả, đồng thời sử dụng mạng LSTM (Long Short-Term Memory) để sinh chuỗi chữ mô tả nội dung hình ảnh dựa trên các đặc trưng đó.

Trong dự án này, chúng ta sẽ khám phá cách mô hình CLIP mã hóa hình ảnh thành các embedding đa chiều, sau đó đưa vào mạng LSTM kết hợp với xử lý ngôn ngữ tự nhiên (NLP) thông qua thư viện NLTK để tối ưu hóa quá trình tạo chú thích. Mô hình sẽ học cách ánh xạ giữa đặc trưng hình ảnh và từ ngữ, đồng thời tạo ra câu mô tả chính xác và tự nhiên.

Ứng dụng của Image Captioning rất đa dạng, từ hỗ trợ người khiếm thị, tự động gắn thẻ hình ảnh, đến nâng cao trải nghiệm tìm kiếm hình ảnh. Dự án này không chỉ cung cấp cái nhìn sâu sắc về cách kết hợp thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (NLP), mà còn mở ra hướng phát triển các hệ thống AI đa phương thức (Multimodal AI) trong tương lai.

2. Mục tiêu, đối tượng và phạm vi

Mục tiêu

- Xây dựng mô hình Image Captioning tự động, kết hợp CLIP (ViT-B/32) để trích xuất đặc trưng hình ảnh và LSTM với cơ chế Attention để sinh chú thích.
- Tận dụng bộ dữ liệu Flickr8k (8,000 ảnh + chú thích) để huấn luyện và đánh giá mô hình.
- Ứng dụng NLTK để tiền xử lý văn bản (tokenization, stopwords removal) và đánh giá độ chính xác của caption (BLEU).

Đối tượng

- Bộ dữ liệu: Flickr8k (ảnh + chú thích tiếng Anh).
- Mô hình chính:
  - Encoder: CLIP (ViT-B/32) → trích xuất đặc trưng hình ảnh (global hoặc patch tokens).
  - Decoder: LSTM + Attention Mechanism → sinh caption dựa trên features từ CLIP.
- Công cụ: Python, PyTorch/TensorFlow, NLTK (xử lý ngôn ngữ).

Phạm vi

- Dữ liệu: Giới hạn ở Flickr8k (không mở rộng sang COCO hoặc Conceptual Captions).
- Mô hình:
  - Chỉ sử dụng CLIP ViT-B/32 (không so sánh với ResNet, Faster R-CNN).
  - Decoder dùng LSTM + Attention.
- Đánh giá: Tập trung vào độ chính xác (BLEU) và khả năng mô tả tự nhiên (qualitative analysis).

Chương II: Cơ sở lý thuyết

1. Image Captioning là gì?

Image Captioning (thuật toán gắn nhãn mô tả cho ảnh) là một thuật toán học máy được thiết kế để sinh ra một đoạn văn bản mô tả một hình ảnh. Thuật toán này được sử dụng rộng rãi trong các ứng dụng như tìm kiếm hình ảnh, dịch thuật, đánh giá hình ảnh, v.v.z

Chương III: Thuật toán