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 (hay thuật toán sinh văn bản dựa theo ảnh) là một thuật toán giúp máy tính sinh ra một đoạn văn bản mô tả một ảnh. Đây là một bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing, NLP) và được ứng dụng rộng rãi trong các ứng dụng như dịch thuật, tìm kiếm hình ảnh, đánh giá hình ảnh, dịch thuật, v.v.

Thuật toán này thực hiện nhiệm vụ dự đoán chú thích cho một hình ảnh nhất định dựa theo các đặc điểm thuộc tính có trong bức ảnh. Các ứng dụng phổ biến trong thế giới thực của nó bao gồm hỗ trợ người khiếm thị có thể giúp họ điều hướng qua các tình huống khác nhau. Do đó, chú thích hình ảnh giúp cải thiện khả năng tiếp cận nội dung cho mọi người bằng cách mô tả hình ảnh cho họ.

Theo đó, thuật toán này gồm input-output như sau:

- Input: Một hình ảnh;
- Output: Một chuỗi các từ mô tả hình ảnh dựa theo các đặc điểm.

2. Ứng dụng

Image Captioning (tạo chú thích hình ảnh tự động) có nhiều ứng dụng thiết thực trong đời sống, công nghệ và nghiên cứu, bao gồm:

1. **Hỗ trợ người khiếm thị:** Mô tả hình ảnh tự động giúp người khiếm thị hiểu nội dung ảnh thông qua giọng nói (screen reader).

   Ví dụ: Ứng dụng như Seeing AI (Microsoft) sử dụng AI để mô tả cảnh vật, chữ viết, cảm xúc.

2. **Tìm kiếm hình ảnh thông minh:** Cải thiện kết quả tìm kiếm bằng cách hiểu nội dung ảnh thay vì chỉ dựa vào metadata hoặc tags.

   Ví dụ: Google Images sử dụng AI để phân tích và gán nhãn ảnh.

3. **Mạng xã hội & Nền tảng chia sẻ ảnh:** Tự động gợi ý chú thích (caption) cho ảnh đăng tải trên Facebook, Instagram.

   Ví dụ: Phát hiện nội dung không phù hợp (ảnh bạo lực, ảnh hưởng xấu).

4. **Y tế & Chẩn đoán hình ảnh:** Mô tả tự động kết quả X-quang, MRI, CT scan để hỗ trợ bác sĩ.

   Ví dụ: AI mô tả tổn thương trong ảnh y tế.

5. **Giáo dục & Học tập:** Tạo mô tả cho hình ảnh trong sách giáo khoa, tài liệu học tập.

   Ví dụ: Hỗ trợ học ngoại ngữ (ghép ảnh với từ vựng).

6. **An ninh & Giám sát:** Tự động mô tả sự kiện trong camera an ninh.

   Ví dụ: Phát hiện hành vi đáng ngờ qua phân tích hình ảnh.

7. **Thương mại điện tử:** Tự động gán nhãn sản phẩm từ ảnh (”Áo thun màu xanh, chất cotton”).

   Ví dụ: Pinterest sử dụng AI để đề xuất sản phẩm tương tự.

8. **Robot & Xe tự lái:** Giúp robot/xe tự lái ”hiểu” môi trường xung quanh qua mô tả bằng ngôn ngữ.

   Ví dụ: Tesla sử dụng AI để nhận diện vật thể và cảnh báo nguy hiểm.

3. Các mô hình phổ biến

Trong thế giới hiện tại, có nhiều mô hình Image Captioning được phát triển và áp dụng, trong đó có những mô hình được sử dụng phổ biến như:

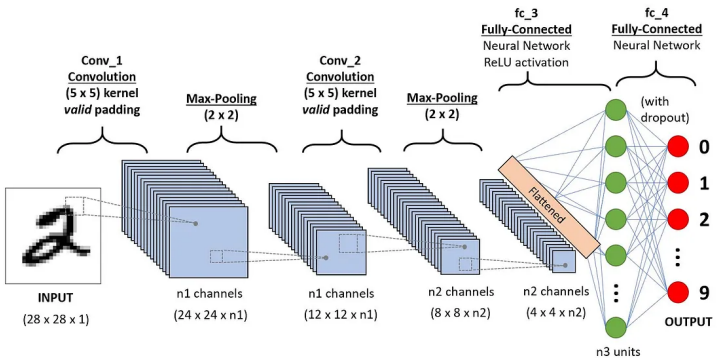
1. Mô hình dựa trên CNN và LSTM: đây là các mô hình dạng cổ điển và được sử dụng phổ biến nhất. Mô hình này sử dụng CNN để trích xuất đặc trưng từ ảnh và LSTM để sinh ra chú thích. Trong đó:
   - Show and Tell (2015)
     - Kiến trúc: CNN + LSTM
     - Ý tưởng: dùng CNN làm encoder, LSTM làm decoder.
     - Ưu điểm: đơn giản, hiệu quả với dữ liệu nhỏ.
   - Show, Attend and Tell (2015)
     - Bổ sung cơ chế Attention để tập trung vào vùng ảnh quan trọng khi sinh từng từ.
     - Ưu điểm: Cải thiện độ chính xác cho ảnh phức tạp.
   - NIC (Neural Image Captioning) (2015)
     - Kiến trúc: CNN + LSTM, ngoại trừ việc nó sử dụng ResNet (hoặc VGG16) làm encoder thay vì CNN thông thường.
     - Ứng dụng tốt cho Flickr8k, Flickr30k.
2. Mô hình dựa trên Transformer: đây là các mô hình mới và được phát triển gần đây. Mô hình này sử dụng kiến trúc Transformer để sinh ra chú thích. Trong đó:
   - Vision Transformer + Transformer Decoder (2020)
     - Kiến trúc: CNN + Transformer
     - Encoder: ViT chia ảnh thành các patch và mã hóa thành các token.
     - Decoder: Transformer sinh caption dựa trên token ảnh.
     - Ưu điểm: Hiệu suất cao với dữ liệu lớn.
   - OFA (One For All) (2021)
     - Đa nhiệm: Unified model cho Image Captioning, VQA, Text-to-Image.
     - Kiến trúc: Transformer, các patch được huân luyện trước có thể sử dụng trên đa dạng tác vụ.
   - BLIP (Bootstrap Language-Image Pre-training) (2022)
     - Kết hợp CLIP và GPT: Vừa hiểu ảnh, vừa sinh văn bản mạch lạc.
     - Ứng dụng: Tạo caption chất lượng cao, chỉnh sửa caption.
   - GIT (GenerativeImage2Text) (2022)
     - Tận dụng Vision Transformer (ViT) và Transformer Decoder.
     - Đặc điểm: Huấn luyện trên lượng dữ liệu khổng lồ (hàng tỷ ảnh)

3. Ngoài ra còn có các mô hình khác như VinVL (Sử dụng CNN và Transformer), hay CoCa (Contrastive Captioners).

4. Các kiến thức liên quan

4.1 CNN

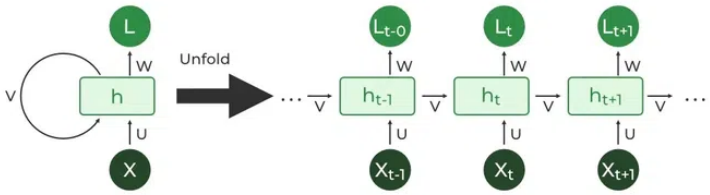
CNN (Convolutional Neural Network) là một loại mạng nơ-ron tích chập được sử dụng phổ biến trong xử lý ảnh và nhận dạng hình ảnh. Chúng được thiết kế để trích xuất đặc trưng từ ảnh bằng cách áp dụng các lớp tích chập và các lớp tổng hợp.



Hình 1: CNN

4.2 RNN

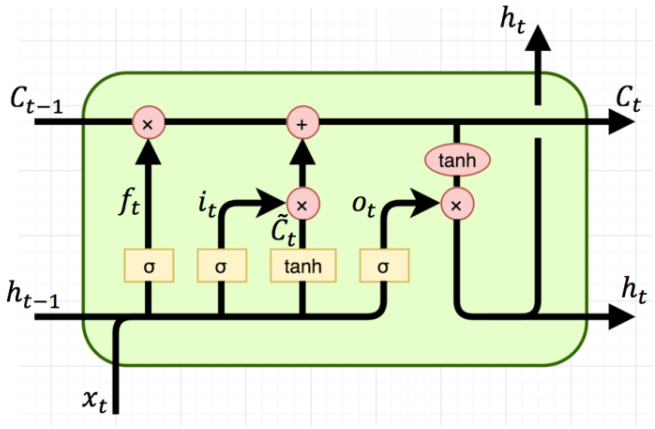
RNN (Recurrent Neural Network) là một loại mạng nơ-ron có trạng thái được sử dụng phổ biến trong xử lý ngôn ngữ tự nhiên và xử lý chuỗi thời gian. RNN được thiết kế để xử lý chuỗi dữ liệu có tính tuần tự, trong đó mỗi phần tử của chuỗi được xử lý dựa trên thông tin từ phần tử trước đó.



Hình 2: RNN

4.3 LSTM

LSTM (Long Short-Term Memory) là một loại mạng nơ-ron có trạng thái được sử dụng phổ biến trong xử lý ngôn ngữ tự nhiên và xử lý chuỗi thời gian. LSTM được thiết kế để xử lý chuỗi dữ liệu có tính tuần tự, trong đó mỗi phần tử của chuỗi được xử lý dựa trên thông tin từ phần tử trước đó.



Hình 3: RNN

Điểm khác biệt giữa LSTM và RNN là LSTM có thể lưu trữ thông tin dài hạn hơn so với RNN thông qua các cổng ghi và đọc. Điều này giúp LSTM có thể xử lý chuỗi dữ liệu có độ dài lớn hơn so với RNN.

4.4 Mô hình CLIP

CLIP (Contrastive Language-Image Pre-training) là một mô hình học sâu được thiết kế để học cách biểu diễn hình ảnh và văn bản bằng cách so sánh chúng với các cặp ảnh-đoạn văn bản. Mô hình này được huấn luyện trên một tập dữ liệu lớn và được sử dụng để học cách biểu diễn hình ảnh và văn bản trong không gian vector.

Chương III: Thuật toán