BÁO CÁO KẾT THÚC MÔN — MÔN DEEP LEARNING

Đề tài: Sinh chú thích ảnh về cuộc sống quanh ta bằng mô hình học sâu

Họ tên SV: Chau Tiểu Long

MSSV: 21094341

Lớp: DHKHMT17B

Nhóm: 1

# Xây dựng mô hình

Bài toán: Sinh chú thích ảnh (Image Captioning) – cho một ảnh đầu vào, mô hình cần sinh ra một câu mô tả ngữ cảnh phù hợp bằng tiếng Anh.

Ý tưởng tổng thể: Kết hợp mô hình mã hóa (Encoder) và giải mã (Decoder) trong kiến trúc học sâu:

* Bộ mã hóa sử dụng mạng Vision Transformer (ViT) để trích xuất đặc trưng ảnh thay cho CNN truyền thống, tận dụng khả năng học quan hệ không gian toàn cục và biểu diễn mạnh mẽ của ViT.
* Bộ giải mã sử dụng GPT-2, một mô hình ngôn ngữ mạnh mẽ thuộc họ Transformer Decoder, để sinh câu chú thích dựa trên đặc trưng hình ảnh đã trích xuất.
* Mô hình được huấn luyện trên tập dữ liệu Flickr8k, bao gồm các ảnh kèm chú thích tiếng Anh.

Mô hình cơ sở sử dụng kiến trúc Encoder-Decoder, trong đó:

* Encoder: là Vision Transformer (ViT) từ thư viện transformers của Hugging Face.
* Decoder: là mô hình sinh ngôn ngữ GPT-2 đã được tinh chỉnh (finetuned).
* Việc huấn luyện sử dụng công cụ Seq2SeqTrainer do Hugging Face cung cấp, kết hợp cùng PyTorch để xử lý dữ liệu.

Lý do chọn mô hình cơ sở:

* Vision Transformer (ViT) cho hiệu quả tốt hơn CNN trong việc trích xuất thông tin hình ảnh nhờ khả năng nắm bắt các mối quan hệ không gian toàn cục.
* GPT-2 có khả năng sinh ngôn ngữ tự nhiên mượt mà, phù hợp cho việc tạo ra các chú thích ảnh có ngữ nghĩa và ngữ pháp tốt.
* Việc sử dụng mô hình và thư viện có sẵn từ Hugging Face giúp rút ngắn thời gian triển khai, dễ dàng tái sử dụng và tùy biến mô hình theo dữ liệu thực tế.

# Dữ liệu và tiền xử lý

· **Dữ liệu sử dụng:**

**Tên tập dữ liệu:** Flickr8k Dataset

**Đặc điểm:**

* Mỗi ảnh có từ 1 đến 5 chú thích tiếng Anh mô tả nội dung ảnh.
* Kích thước ảnh đa dạng, định dạng ảnh bao gồm JPG và PNG.
* Chú thích được lưu trong một tệp văn bản (captions.txt) dạng CSV gồm các cột: tên ảnh và câu chú thích.

**Chia dữ liệu:** Dữ liệu được chia thành hai tập:

**Training set:** 80%

**Validation set:** 20%

· **Tiền xử lý ảnh:**

* Ảnh được tải bằng thư viện PIL và chuyển sang chế độ RGB.
* Ảnh được biến đổi bằng torchvision.transforms, gồm các bước:
* Resize ảnh về kích thước chuẩn **(224, 224)**.
* Chuyển ảnh thành **Tensor**.
* Chuẩn hóa ảnh với **mean = (0.5, 0.5, 0.5)** và **std = (0.5, 0.5, 0.5)** (giá trị này đã được thay thế cho ImageNet để phù hợp hơn với mô hình cụ thể).
* Ảnh sau khi biến đổi được truyền vào ViTFeatureExtractor để trích xuất đặc trưng dưới dạng tensor (pixel\_values).

· **Tiền xử lý chú thích:**

* Chú thích được lấy từ cột caption của DataFrame.
* Sử dụng AutoTokenizer (với GPT-2) để:

**Tokenize** câu chú thích.

**Pad** câu đến chiều dài tối đa (max\_length = 50).

Gán các token đệm (pad\_token\_id) thành -100, để bỏ qua chúng khi tính toán hàm mất mát trong quá trình huấn luyện.

Cuối cùng, mô hình nhận đầu vào là cặp {pixel\_values, labels}.

# Thiết kế và huấn luyện mô hình

Mô hình được xây dựng theo kiến trúc **Encoder-Decoder** sử dụng thư viện transformers của Hugging Face:

* **Encoder:** Vision Transformer (ViT) – được sử dụng để trích xuất đặc trưng từ ảnh đầu vào. ViT chia ảnh thành các patch nhỏ, rồi xử lý chúng như một chuỗi đầu vào của Transformer.
* **Decoder:** GPT-2 – được dùng để sinh ra câu mô tả bằng tiếng Anh. GPT-2 là mô hình ngôn ngữ mạnh mẽ, có khả năng sinh văn bản mạch lạc và ngữ nghĩa tốt.

Hai thành phần trên được tích hợp bằng lớp VisionEncoderDecoderModel.from\_encoder\_decoder\_pretrained().

Cấu hình chính:

decoder\_start\_token\_id = tokenizer.bos\_token\_id

eos\_token\_id = tokenizer.sep\_token\_id

max\_length = 128

num\_beams = 4 (beam search)

no\_repeat\_ngram\_size = 3

length\_penalty = 2.0

Tham số huấn luyện:

Mô hình được huấn luyện bằng Seq2SeqTrainer với các cấu hình cụ thể:

**output\_dir:** 'VIT\_large\_gpt2' – thư mục lưu mô hình.

**Số epoch:** 3

**Kích thước batch:**

per\_device\_train\_batch\_size = 8

per\_device\_eval\_batch\_size = 8

**Tổng số mẫu huấn luyện:** 32,364

**Số bước tối ưu hóa (optimization steps):** 12,138

**Gradient Accumulation steps:** 1

**Learning rate:** 5e-5

**Warmup steps:** 1024

**Logging mỗi:** 1024 bước

**Lưu mô hình mỗi:** 2048 bước

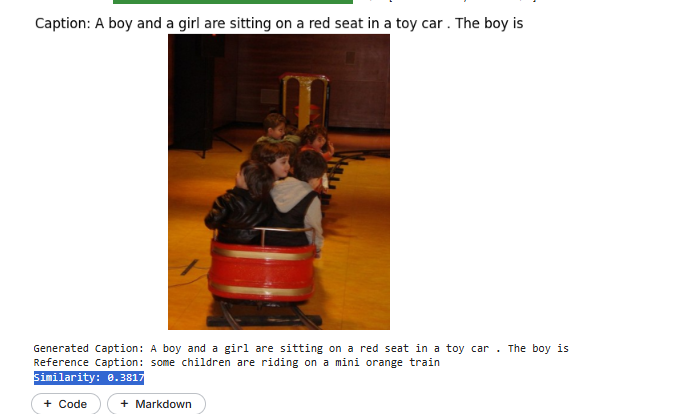
**Evaluation strategy:** "epoch" (đánh giá mỗi epoch)

**save\_total\_limit:** 1 (chỉ lưu lại 1 mô hình tốt nhất)

# 4. Đánh giá mô hình

* Phương pháp: Huấn luyện mô hình trên tập train, đánh giá bằng tập test.
* Chỉ số đánh giá:

Similarity: 0.3817



Câu có ý nghĩa gần giống.