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: Lê Hoàng Phúc

MSSV: 21114851

Lớp: DHKHMT17CTT

Nhóm: 2

# 1. Xây dựng mô hình

 **Bài toán**: Mục tiêu là tạo ra chú thích cho ảnh (Image Captioning). Mỗi ảnh đầu vào sẽ được mô hình tạo ra một câu mô tả ngữ cảnh bằng tiếng Anh.

 **Ý tưởng tổng thể**: Sử dụng **Convolutional Neural Network (CNN)** để trích xuất đặc trưng từ ảnh và **Recurrent Neural Network (RNN)** (ở đây là **LSTM** - Long Short-Term Memory) để xử lý chuỗi từ và tạo ra câu mô tả cho ảnh. Sự kết hợp này tận dụng khả năng của CNN trong việc phân tích đặc trưng hình ảnh và RNN trong việc xử lý dữ liệu chuỗi, như câu văn.

 **Mô hình cơ sở**: Mô hình **InceptionV3**, đã được huấn luyện trên **ImageNet**, được sử dụng để trích xuất đặc trưng từ ảnh. InceptionV3 là một trong những mô hình CNN mạnh mẽ nhất và có khả năng nhận diện và phân loại ảnh rất tốt nhờ vào kiến trúc phức tạp của nó.

 **Lý do chọn mô hình cơ sở**: **InceptionV3** có thể trích xuất các đặc trưng ảnh có tính biểu cảm cao mà không cần phải huấn luyện lại từ đầu. Việc sử dụng mô hình đã huấn luyện sẵn giúp tiết kiệm thời gian huấn luyện và cải thiện độ chính xác của mô hình. Đồng thời, InceptionV3 đã được chứng minh là có khả năng phân loại hình ảnh rất tốt.

# 2. Dữ liệu và tiền xử lý

 **Dữ liệu sử dụng**: **Flickr8k** dataset bao gồm 8,000 bức ảnh và mỗi ảnh đi kèm với 5 chú thích mô tả bằng tiếng Anh. Dữ liệu này giúp mô hình học cách liên kết đặc trưng ảnh với ngữ nghĩa của các từ trong câu.

 **Định dạng ảnh**: Ảnh có định dạng **JPG/PNG** với kích thước đa dạng. Do đó, tiền xử lý ảnh rất quan trọng để đảm bảo các ảnh có thể được sử dụng làm đầu vào cho mô hình.

 **Chú thích**: Chú thích ảnh trong **Flickr8k** ban đầu có thể chứa các ký tự đặc biệt, chữ hoa và số. Vì vậy, chúng được xử lý bằng cách chuyển tất cả sang chữ thường, loại bỏ dấu câu và số, giúp dữ liệu chuẩn hóa và dễ dàng cho mô hình học. Ngoài ra, thêm các từ bắt đầu (**start**) và kết thúc (**end**) vào chú thích giúp mô hình hiểu được điểm bắt đầu và kết thúc của chuỗi văn bản.

 **Tiền xử lý ảnh**: Các ảnh được chuẩn hóa (pixel values được chuẩn hóa về khoảng [0, 1]) và thay đổi kích thước thành **299x299 pixels** để phù hợp với yêu cầu của mô hình **InceptionV3**. Kích thước này đảm bảo rằng ảnh có thể được đưa qua mô hình mà không gặp phải sự cố về kích thước đầu vào.

 **Tiền xử lý chú thích**: Chú thích được chuyển thành chuỗi văn bản đơn giản với các từ "start" và "end" ở đầu và cuối câu, tương tự như cách thức huấn luyện trong các bài toán sinh văn bản.

# 3. Thiết kế và huấn luyện mô hình

* **Mô hình**:
  + **Encoder**: Sử dụng đặc trưng ảnh được trích xuất từ **InceptionV3** làm đầu vào cho phần encoder. Encoder này giúp mô hình hiểu và lưu trữ thông tin về nội dung của ảnh.
  + **Decoder**: Sử dụng **LSTM** làm decoder để dự đoán các từ tiếp theo trong câu chú thích. Mô hình LSTM giúp mô hình xử lý các chuỗi dữ liệu có tính chất thời gian như văn bản, từ đó tạo ra các câu mạch lạc và có ngữ nghĩa.
  + **Kết hợp**: Đặc trưng ảnh từ encoder và chuỗi từ từ LSTM decoder sẽ được kết hợp để sinh ra câu mô tả hoàn chỉnh.
* **Tham số huấn luyện**:
  + **Batch size**: Sử dụng kích thước batch là **270** cho tập huấn luyện và **150** cho tập validation để giảm thiểu thời gian huấn luyện và tối ưu bộ nhớ.
  + **Epochs**: Huấn luyện trong **25 epochs** để mô hình có thể học đủ và không bị overfitting. Việc sử dụng **EarlyStopping** giúp dừng huấn luyện nếu mô hình không cải thiện hiệu suất sau một số epoch nhất định. **LearningRateScheduler** được dùng để điều chỉnh learning rate giúp mô hình học hiệu quả hơn.
* **Thư viện sử dụng**:
  + **TensorFlow/Keras**: Các thư viện này được sử dụng để xây dựng và huấn luyện mô hình học sâu.
  + **NLTK**: Thư viện này hỗ trợ trong việc xử lý văn bản như tokenization, chuẩn hóa từ ngữ.
  + **scikit-learn**: Dùng cho các bước tiền xử lý dữ liệu và các tác vụ học máy khác như split dataset, vector hóa.

# 4. Đánh giá mô hình

* **Phương pháp**: Mô hình được huấn luyện trên tập huấn luyện và đánh giá trên tập validation. Sau đó, hiệu suất của mô hình được đánh giá thông qua chỉ số BLEU (một chỉ số đo độ chính xác của mô hình trong việc tạo ra chú thích ảnh).
* **Chỉ số đánh giá**:

BLEU score: Average BLEU score on 100 test samples: 0.0536

Quan sát thực tế: Mô hình cho ra những chú thích gần đúng với đối tượng trong ảnh, nhưng ngữ cảnh cần phải cải thiện độ chính xác thêm.

