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: Nguyễn Hùng Anh

MSSV: 21064051

Lớp: DHKHMT17B

Nhóm: 1

# 1. Xây dựng mô hình

* Bài toán: **Sinh chú thích ảnh (Image Captioning)** là bài toán trong lĩnh vực thị giác máy tính kết hợp xử lý ngôn ngữ tự nhiên, nhằm mục đích **tự động tạo một câu mô tả bằng tiếng Anh cho một bức ảnh đầu vào**. Đây là bài toán có tính ứng dụng cao trong thực tế như:
  + Tự động tạo mô tả cho ảnh trong thư viện
  + Hỗ trợ người khiếm thính hiểu nội dụng trong ảnh
  + Ứng dụng trong tìm kiếm ảnh theo nội dung
* Ý tưởng tổng thế:
  + Trích xuất đặc trưng ảnh (Image Feature Extraction):
    - Sử dụng một mạng kiến trúc mạng CNN là InceptionV3 để trích xuất đặc trưng (feature vector) biểu diễn nội dung của ảnh.
  + Sinh chuỗi mô tả (Caption Generation):
    - Dùng một mạng LSTM (Long Short-Term Memory) để sinh ra câu mô tả từ đặc trưng ảnh và các từ đã sinh trước đó.
    - Quá trình sinh caption bắt đầu từ token đặc biệt startseq và kết thúc với endseq.
* Mô hình cơ sở:
  + Mô hình sử dụng kiến trúc kết hợp giữa:
    - CNN encoder: Inception V3 (pre-trained trên ImageNet), đùng để trích xuất đặc trưng ảnh.
    - RNN decoder: Một lớp Embedding + LSTM + Dense để sinh mô tả từng từ một.
    - Word Embedding: Glove pre-trained embeddings (100D) để biểu diễn dưới dạng vector.
* Lý do chọn mô hình cơ sở:
  + InceptionV3 là một trong những CNN mạnh, nhẹ và được huấn luyện tốt trên ImageNet, cho phép trích xuất đặc trưng ảnh hiểu quả.
  + LSTM là có khả năng phụ thuộc dài hạn, rất phù hợp với bài toán như sinh mô tả từ ảnh.
  + Glove embeddings là giúp mô hình học ngữ nghĩa từ tốt hơn và tăng tốc độ hội tụ nhờ sử dụng vector từ đã huấn luyện.
  + Kiến trúc encoder-decoder là giải pháp cổ điện và hiểu quả cho các bài toán biến đầu vào thành chuỗi (Image -> Sentence), dễ triển khai và kiểm soát.

# 2. Dữ liệu và tiền xử lý

* Dữ liệu sử dụng: Trong bài toán này, bộ dữ liệu sử dụng là COCO2014 (Common Object in contex)
  + Đặc điểm ảnh:
    - Bộ dữ liệu chứa hàng ngàn hình ảnh với nội dung đa dạng (cảnh vật, người, động vật, đồ vật, …)
    - Ảnh có định dạng JPG
    - Kích thước ảnh không đồng nhất, cần chuẩn hóa khi đưa vào mô hình
  + Đặc điểm chú thích (Captions):
    - Mỗi ảnh đi kèm nhiều mô tả bằng tiếng Anh (ít nhất 5 chú thích cho mỗi ảnh)
    - Các câu mô tả có độ dài trung bình từ 5 đến 20 từ, diễn đạt nội dung ảnh một cách ngắn gọn.
* Tiền xử lý ảnh
  + Resize chuyển đội kích thước ảnh về đúng yêu cầu bởi CNN (299x299)
  + Chuấn hóa pixel được scale giá trí pixel về [-1,1] cho mô hình CNN Inceptionv3
  + Feature extraction thì thay vì dùng ảnh trực tiếp, trích xuất vector đặc trưng từ tầng cuối cùng (ví dụ: từ Global Average Pooling layer của InceptionV3).
  + Lưu trữ Lưu các đặc trưng trích xuất (feature vectors) vào file .pkl để tăng tốc độ huấn luyện
* Tiền xử lý chú thích
  + Làm sạch văn bản:
    - Chuyển toàn bộ chữ thành chữ thường.
    - Loại bỏ dấu câu, ký tự đặc biệt, số.
    - Xóa từ quá ngắn hoặc không hợp lệ.
  + Chèn token đặc biệt thì Thêm từ khóa startseq vào đầu mỗi câu và endseq vào cuối câu để đánh dấu bắt đầu và kết thúc chuỗi.
  + Tạo từ điển từ (Vocabulary)
    - Xây dựng tập từ vựng gồm các từ xuất hiện đủ tần suất (ví dụ: >5 lần).
    - Gán chỉ số (index) cho từng từ để mô hình có thể xử lý.
  + Mã hóa chú thích:
    - Chuyển các từ thành chỉ số dựa trên từ điển.
    - Sử dụng padding để chuẩn hóa độ dài các chuỗi.
  + Embedding Layer:
    - Gắn mỗi chỉ số từ với một vector embedding (ví dụ: sử dụng GloVe).

# 3. Thiết kế và huấn luyện mô hình

* Mô hình
* Tham số huấn luyện
* Thư viện sử dụng

# 4. Đánh giá mô hình

* Phương pháp: Huấn luyện mô hình trên tập train, đánh giá bằng tập test.
* Chỉ số đánh giá:
  + BLEU score
  + Quan sát thực tế: xem chú thích mô hình tạo ra có phù hợp không