2.Xây dựng hệ thống

2.1 Mô hình Encoder-Decoder  
2.1.1 Khái niệm

Mô hình Encoder-Decoder là một trong những kiến trúc phổ biến trong học sâu, đặc biệt là trong các bài toán xử lý chuỗi (sequence-to-sequence). Mô hình này bao gồm hai phần chính:

Encoder: Mô hình encoder chịu trách nhiệm tiếp nhận đầu vào, thường là ảnh, và chuyển đổi chúng thành một dạng biểu diễn có kích thước nhỏ hơn nhưng vẫn đầy đủ thông tin. Trong bài toán này, encoder có thể là một mạng nơ-ron tích chập (CNN) đã được huấn luyện trên một tập dữ liệu lớn (ví dụ: ImageNet). Đầu ra của encoder là một vector đặc trưng của ảnh, đại diện cho các thông tin quan trọng từ bức ảnh đó.

Decoder: Decoder nhận thông tin từ encoder và sinh ra chuỗi đầu ra. Trong bài toán Image Captioning, chuỗi này là một câu văn mô tả nội dung của ảnh. Decoder thường được xây dựng bằng cách sử dụng các mô hình học sâu như LSTM (Long Short-Term Memory) hoặc Transformer để xử lý các chuỗi từ vựng.  
2.1.2 **Encoder: Trích xuất đặc trưng hình ảnh với CNN**

Hình ảnh là dữ liệu đầu vào của mô hình, và để mô hình có thể hiểu được nội dung hình ảnh, cần phải trích xuất các đặc trưng (features) quan trọng từ hình ảnh. Điều này được thực hiện thông qua một mạng CNN.

Trong mô hình của tôi, phần trích xuất đặc trưng hình ảnh sử dụng một mô hình CNN pre-trained, như **ResNet**, **Inception**, hoặc **VGG**. Mô hình CNN này được huấn luyện trên các tập dữ liệu lớn như ImageNet, giúp chúng học được các đặc trưng hình ảnh cơ bản như hình dạng, màu sắc, kết cấu, và các đối tượng trong ảnh.

Cụ thể, tôi sử dụng một mô hình trích xuất đặc trưng đã được huấn luyện trước để lấy các vector đặc trưng từ hình ảnh. Những vector này được đưa vào phần Decoder (LSTM) để giúp mô hình hiểu rõ hơn về nội dung hình ảnh.

2.1.3 **Decoder: Sinh chú thích hình ảnh với LSTM**

LSTM (Long Short-Term Memory) là một loại mạng nơ-ron hồi tiếp (recurrent neural network - RNN) được thiết kế để giải quyết các vấn đề liên quan đến việc học chuỗi dữ liệu dài, đặc biệt là vấn đề **vanishing gradient** mà các mô hình RNN truyền thống gặp phải.

Trong bài toán tạo chú thích hình ảnh, LSTM đóng vai trò quan trọng trong việc học cách tạo ra chuỗi từ ngữ mô tả từ các đặc trưng hình ảnh. Mô hình LSTM nhận vào các từ của caption dưới dạng các số nguyên đã được mã hóa (tokenized) và chuyển chúng thành chuỗi từ có nghĩa, có thể mô tả chính xác nội dung hình ảnh.

Mô hình sử dụng **LSTM** để học mối quan hệ giữa các từ trong caption và đặc trưng hình ảnh. LSTM nhận đầu vào là chuỗi từ (caption) và đầu ra là chuỗi các từ tiếp theo trong caption. Quá trình huấn luyện giúp mô hình học cách kết hợp các đặc trưng hình ảnh với các từ trong caption để tạo ra chú thích phù hợp.

**2.1.4 Quá trình huấn luyện mô hình**

Quá trình huấn luyện của mô hình được thực hiện qua các bước sau:

* **Tiền xử lý dữ liệu**: Dữ liệu đầu vào bao gồm hình ảnh và chú thích tương ứng. Chú thích được tiền xử lý, bao gồm việc thêm các token đặc biệt như <start> và <end> để đánh dấu sự bắt đầu và kết thúc của caption. Các từ trong caption sau đó được mã hóa thành các chỉ số bằng cách sử dụng một **tokenizer**.
* **Trích xuất đặc trưng hình ảnh**: Mỗi hình ảnh được đưa qua mô hình CNN pre-trained để trích xuất các đặc trưng hình ảnh (features). Các đặc trưng này sẽ đóng vai trò quan trọng trong việc giúp mô hình học cách sinh chú thích phù hợp.
* **Huấn luyện mô hình**: Mô hình được huấn luyện với một tập dữ liệu lớn, bao gồm các đặc trưng hình ảnh và các caption tương ứng. Mô hình sử dụng hàm mất mát **sparse categorical crossentropy** để tối ưu hóa các trọng số của mạng, giúp mô hình có thể tạo ra các caption chính xác hơn.
* **Kiểm tra và đánh giá mô hình**: Sau khi huấn luyện xong, mô hình được kiểm tra trên một tập dữ liệu kiểm tra riêng biệt để đánh giá khả năng tạo chú thích chính xác cho hình ảnh chưa được nhìn thấy trong quá trình huấn luyện.

2.2 Mục tiêu

Mục tiêu của mô hình **Encoder-Decoder kết hợp CNN và LSTM** là tạo ra một hệ thống tự động sinh chú thích cho hình ảnh (image captioning). Cụ thể, mô hình này được huấn luyện để đạt được các mục tiêu sau:

1. **Tự động mô tả hình ảnh**: Mô hình có khả năng tự động sinh ra một chú thích ngắn gọn và chính xác cho hình ảnh đầu vào, mô phỏng cách con người mô tả nội dung của một bức tranh hoặc ảnh chụp.
2. **Hiểu và phân tích nội dung hình ảnh**: Mô hình học cách phân tích và trích xuất các đặc trưng quan trọng của hình ảnh, như đối tượng, cảnh vật, hành động, và các yếu tố ngữ nghĩa trong bức ảnh.
3. **Tạo câu văn tự nhiên**: Mô hình sinh ra các câu chú thích có cấu trúc ngữ pháp và ngữ nghĩa hợp lý, phản ánh đúng nội dung hình ảnh. Đây là kết quả của việc kết hợp các đặc trưng hình ảnh với một mạng LSTM để tạo chuỗi từ ngữ có ý nghĩa.
4. **Ứng dụng trong các hệ thống trợ giúp người khiếm thị**: Cung cấp các giải pháp hỗ trợ người khiếm thị trong việc hiểu được nội dung của hình ảnh mà họ không thể nhìn thấy. Đây là một trong những ứng dụng thực tiễn quan trọng của bài toán tạo chú thích hình ảnh.

2.3 Vai trò

Mô hình Encoder-Decoder kết hợp CNN và LSTM đóng vai trò quan trọng trong việc phát triển các hệ thống thông minh trong các lĩnh vực như thị giác máy tính, xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo (AI). Cụ thể:

1. **CNN - Vai trò trích xuất đặc trưng hình ảnh**:
   * **Trích xuất đặc trưng**: CNN giúp trích xuất các đặc trưng quan trọng từ hình ảnh, chẳng hạn như các đối tượng, cảnh vật, và các chi tiết khác, chuyển chúng thành các vector đặc trưng có thể sử dụng trong mô hình học sâu.
   * **Xử lý thông tin hình ảnh**: CNN giúp mô hình hiểu được thông tin thị giác, chẳng hạn như nhận diện các đối tượng, mối quan hệ không gian giữa các đối tượng trong ảnh, giúp mô hình có thể tạo ra chú thích chính xác.
2. **LSTM - Vai trò sinh chuỗi văn bản**:
   * **Tạo chuỗi từ ngữ mô tả**: LSTM giúp mô hình học cách tạo ra các câu từ các từ đã được mã hóa, dựa trên các đặc trưng hình ảnh. LSTM đóng vai trò quan trọng trong việc xử lý chuỗi và học các mối quan hệ giữa các từ trong câu.
   * **Dự đoán từ tiếp theo**: LSTM học cách dự đoán từ tiếp theo trong chuỗi chú thích dựa trên các từ đã xuất hiện trước đó và các đặc trưng hình ảnh. Điều này giúp mô hình sinh ra những câu chú thích có cấu trúc và ngữ pháp hợp lý.
3. **Mô hình Encoder-Decoder - Vai trò kết nối giữa đặc trưng hình ảnh và văn bản**:
   * **Kết hợp thông tin hình ảnh và văn bản**: Mô hình Encoder-Decoder kết hợp đặc trưng hình ảnh từ CNN và chuỗi từ trong LSTM để sinh ra một câu chú thích cho hình ảnh. Encoder giúp chuyển đổi hình ảnh thành một biểu diễn số học, trong khi Decoder sử dụng biểu diễn này để tạo ra văn bản mô tả.
   * **Tạo ra chú thích ngữ nghĩa**: Mô hình này học cách kết hợp hai loại dữ liệu (hình ảnh và văn bản) để tạo ra chú thích phù hợp với nội dung hình ảnh, thể hiện sự kết hợp của các kỹ thuật thị giác máy tính và xử lý ngôn ngữ tự nhiên.

Các thư viện

1. json

* Chức năng: Thư viện json trong Python được sử dụng để đọc và ghi dữ liệu ở định dạng JSON. Trong đoạn mã trên, thư viện này được dùng để lưu trữ tokenizer (một đối tượng dùng để chuyển đổi văn bản thành các chỉ số số học) vào tệp JSON. Điều này giúp mô hình có thể dễ dàng tái sử dụng trong các lần chạy sau mà không cần phải huấn luyện lại tokenizer

2. os

* Chức năng: Thư viện os cung cấp các hàm giúp thao tác với hệ thống tệp và thư mục trong Python. Trong đoạn mã này, os.path được sử dụng để kiểm tra sự tồn tại của tệp và thư mục, xử lý các đường dẫn tệp, và tạo các thư mục mới nếu chúng chưa tồn tại

3. numpy

* Chức năng: numpy là thư viện phổ biến trong Python cho việc tính toán với các mảng n-dimensional. Trong đoạn mã trên, thư viện numpy được sử dụng để xử lý các mảng đặc trưng hình ảnh. Đặc trưng hình ảnh sau khi được trích xuất sẽ được chuyển đổi thành mảng numpy để phù hợp với mô hình học sâu. numpy giúp thực hiện các phép toán với mảng, như reshaping mảng hoặc xử lý các giá trị thiếu (missing values

4. tensorflow

* Chức năng: tensorflow là thư viện mã nguồn mở được sử dụng rộng rãi trong các ứng dụng học sâu (deep learning). Trong mã nguồn, tensorflow đóng vai trò quan trọng trong việc xây dựng mô hình học sâu, huấn luyện mô hình, và lưu trữ mô hình sau khi huấn luyện.

5. model.caption\_model

* Chức năng: Đây là một mô-đun có thể chứa định nghĩa mô hình captioning (tạo chú thích cho hình ảnh). Mặc dù trong đoạn mã này, caption\_model không được sử dụng trực tiếp, nhưng có thể nó chứa lớp mô hình được sử dụng để tạo chú thích cho hình ảnh trong dự án. Mô hình captioning thường bao gồm một mạng nơ-ron học sâu để dự đoán chú thích văn bản cho mỗi hình ảnh đầu vào.

6. feature\_extraction.feature\_extraction

* Chức năng: Thư viện này bao gồm các hàm dùng để tiền xử lý ảnh và trích xuất đặc trưng từ các bức ảnh. Cụ thể, preprocess\_image có thể thực hiện các bước như thay đổi kích thước ảnh, chuẩn hóa ảnh, hoặc chuyển đổi ảnh thành dạng có thể đưa vào mô hình học sâu. extract\_features là hàm trích xuất đặc trưng từ ảnh, giúp chuyển ảnh thành các vector đặc trưng, thường được sử dụng làm đầu vào cho các mô hình học sâu. Những đặc trưng này là các biểu diễn số học giúp mô hình hiểu được nội dung của ảnh.

7. feature\_extraction.load\_feature\_extraction\_model

* Chức năng: Thư viện này cung cấp các hàm để tải mô hình đã được huấn luyện từ trước, có thể là các mô hình học sâu như Inception, VGG, hoặc ResNet, dùng để trích xuất đặc trưng từ ảnh. Mô hình này thường đã được huấn luyện trên một lượng lớn dữ liệu và có khả năng trích xuất các đặc trưng có ý nghĩa từ hình ảnh đầu vào. Hàm load\_feature\_extraction\_model sẽ tải mô hình này để sử dụng cho việc trích xuất đặc trưng từ các bức ảnh trong tập dữ liệu.

8. sys

* Chức năng: Thư viện sys cung cấp các chức năng liên quan đến hệ thống và trình thông dịch Python. Sys được sử dụng để thay đổi sys.path, giúp thêm thư mục hiện tại vào danh sách tìm kiếm các mô-đun Python, từ đó giúp Python có thể tìm thấy các mô-đun mà người lập trình đã tạo ra trong cùng thư mục với mã nguồn. sys.stdout được thay đổi để in ra các ký tự Unicode, giúp hiển thị đúng các ký tự đặc biệt khi làm việc với văn bản tiếng Việt trong chương trình.

Cấu trúc tệp của project

A screenshot of a computer program

Description automatically generated

Dataset: thư mục chứa dữ liệu đầu vào

* Flicker8k\_Dataset/Images: Có chứa 8091 hình ảnh
* captions: Có chứa các tệp văn bản và chú thích của hình ảnh.

feature\_extraction

* feature\_extraction.py: File này thực hiện các bước tiền xử lý và trích xuất đặc trưng từ ảnh bằng mô hình InceptionV3 trong TensorFlow/Keras. Đây là một phần trong quy trình xử lý ảnh để sử dụng các đặc trưng hình ảnh cho các nhiệm vụ học máy như mô hình tạo chú thích ảnh (image captioning).
* load\_feature\_extraction\_model.py: File này chứa một hàm để tải mô hình InceptionV3 đã được huấn luyện trước trên tập dữ liệu ImageNet và cấu hình lại mô hình này để chỉ trích xuất đặc trưng từ ảnh, không bao gồm lớp phân loại cuối cùng.

model:

caption\_model.py: File này định nghĩa một mô hình học sâu dùng để tạo caption (mô tả văn bản) cho ảnh, sử dụng kiến trúc Encoder-Decoder với LSTM. Mô hình này kết hợp đặc trưng của ảnh và chuỗi văn bản để tạo ra một caption cho ảnh đầu vào.  
  
results:

* found\_images.log: Tệp này ghi lại các ảnh đã được tìm thấy hoặc sử dụng trong quá trình huấn luyện.
* missing\_images.log: Tệp này chứa thông tin về các ảnh bị thiếu hoặc không được tìm thấy.
* model\_weights.weights.h5: Tệp trọng số của mô hình đã huấn luyện.
* model.keras: Tệp mô hình đã huấn luyện chứa cấu trúc và trọng số.

templates:

* index.html: Đây là giao diện người dùng của ứng dụng Flask, nơi người dùng có thể tải ảnh lên và nhận caption cho ảnh đó.

Upload: Thư mục này chứa các ảnh mà người dùng tải lên để xử lý và tạo caption.

app.py: Đây là tệp chính của ứng dụng Flask, nơi bạn xử lý các route HTTP và các logic liên quan đến việc tải ảnh lên và tạo caption.

captions.txt: Tệp này lưu trữ các caption đã được tạo cho ảnh đã xử lý. Mỗi lần bạn tạo caption cho một ảnh, caption đó sẽ được ghi vào tệp này

tokenizer.json: Tệp JSON này chứa tokenizer (máy phân tích từ ngữ) để chuyển các từ thành các chỉ mục số mà mô hình có thể hiểu.

train.py: Tệp này dùng để huấn luyện mô hình học sâu (deep learning model) sử dụng dữ liệu ảnh và các caption tương ứng.

Tệp văn bản chứa tất cả chú thích hình ảnh là caption.txt trong folder dataset

Nội dung của tệp là tên hình ảnh và mô tả tương ứng được phân tách với nhau bằng các dòng mới và mỗi hình ảnh có 5 mô tả