## **1. Các cải tiến**

* Xây dựng lại khung code cho gọn gàng, dễ hiểu.
* Thử với các mô hinh pre-trained trích xuất đặc trưng khác như densenet201. Nhận xét là densenet201 tốt cho việc nghiên cứu, học tâp hơn là inceptionv3
* Tiếp tục tập trung vào lstm. Tìm hiểu, trainning mô hình khác(LSTM\_2)
* Có thêm val-loss
* Xây hàm Bleu tính toán độ chính xác(Độ khớp của mô hình trên tập test)
* Hiểu sâu hơn về code

## **2. Giới thiệu**

## Chào mừng các bạn đến với buổi thuyết trình về mô hình sinh mô tả hình ảnh!

## Hình ảnh là một phần quan trọng trong cuộc sống của chúng ta. Chúng ta sử dụng hình ảnh để giao tiếp, ghi nhớ và hiểu thế giới xung quanh mình. Tuy nhiên, không phải ai cũng có thể hiểu được tất cả các thông tin được truyền tải trong một hình ảnh.

## Mô hình sinh mô tả hình ảnh là một loại trí tuệ nhân tạo có thể tạo ra văn bản mô tả nội dung của một hình ảnh. Các mô hình này được đào tạo trên một bộ dữ liệu gồm các hình ảnh và văn bản mô tả tương ứng.

## **2. Ứng dụng**

+ Chú thích tự động cho hình ảnh trên mạng xã hội, website, website upload picture, ... để giúp cho các công cụ tìm kiếm hiểu hơn về hình ảnh.

+ Phân tích hình ảnh quảng cáo và sản phẩm để hiểu xu hướng và sở thích của khách hàng.

+ Trợ giúp người khiếm thị: Mô hình có khả năng tạo ra mô tả văn bản cho hình ảnh có thể giúp người khiếm thị hiểu được nội dung của hình ảnh. Ví dụ, mô hình có thể mô tả chi tiết về hình ảnh mà người khiếm thị không thể nhìn thấy.

+ Tìm kiếm hình ảnh: Mô hình có thể giúp tìm kiếm hình ảnh dựa trên mô tả văn bản. Người dùng có thể nhập mô tả văn bản, và mô hình sẽ tìm kiếm các hình ảnh phù hợp với mô tả đó.

+ Giám sát video và an ninh: Trong lĩnh vực giám sát video, mô hình có thể tạo ra mô tả cho các cảnh quan trọng trong video, giúp quản lý an ninh hiểu rõ hơn về những gì đang xảy ra.

+ Giáo dục: Mô hình có thể tạo ra mô tả cho các hình ảnh giảng dạy trong giáo dục trực tuyến. Điều này có thể giúp sinh viên hiểu rõ hơn về nội dung hình ảnh.

+ Ứng dụng di động và trải nghiệm người dùng tương tác: Mô hình có thể được tích hợp vào ứng dụng di động và ứng dụng thực tế ảo để tạo ra trải nghiệm tương tác thú vị. Ví dụ, ứng dụng có thể tạo ra mô tả cho các đối tượng xung quanh người dùng trong thực tế ảo.

+ Mô tả hình ảnh y tế: Trong lĩnh vực y tế, mô hình có thể tạo ra mô tả cho hình ảnh y tế như hình ảnh MRI hoặc X-quang, giúp bác sĩ và nhân viên y tế hiểu rõ hơn về tình trạng của bệnh nhân.

## **3. Bài toán học sâu**

Ý tưởng: Dùng đặc trưng của ảnh và dùng các từ phía trước để dự đoán từ tiếp theo của caption.

Example:

<Image> + startseq -> A. <Image> + startseq A -> girl .

<Image> + startseq A girl -> going .

<Image> + startseq A girl going -> into .

<Image> + startseq A girl going into -> a .

<Image> + startseq A girl going into a -> wooden .

<Image> + startseq A girl going into a wooden -> building

<Image> + startseq A girl going into a wooden building-> endseq.

## **4. Mô tả dữ liệu**

### **4.1 Mô tả tập Dataset**

Flickr 8k Dataset là một tập dữ liệu phổ biến được sử dụng trong các nhiệm vụ liên quan đến mô hình học máy xử lý hình ảnh và ngôn ngữ tự nhiên. Dữ liệu này chứa 8000 ảnh và mỗi ảnh đi kèm với năm mô tả (caption) khác nhau. Việc này giúp mô hình học máy có khả năng học cách hiểu và mô tả nội dung của ảnh theo nhiều cách khác nhau.

Dữ liệu được chia thành ba tập chính:

● Training Set (tập huấn luyện): Gồm 6000 ảnh và tương ứng với mỗi ảnh là năm mô tả khác nhau. Có 6800\*5 = 30000 câu mô tả.

● Validation Set, Test Set dùng chung bộ dữ liệu: Gồm 1200 ảnh được sử dụng để đánh giá, test mô hình trong quá trình huấn luyện.

### **4.2 Trích xuất đặc trưng hình ảnh**

Sử dụng các mô hình đã được train sẵn như: Inception, DenseNet. Cả Inception và DenseNet là hai kiến trúc mô hình học sâu phổ biến được thiết kế để trích xuất đặc trưng từ hình ảnh.

Ví dụ InceptionV3:

+ Input: Image ( 299x299x3 )

+ Output: Image ( 1x1x2048 )

( Bỏ đi layer cuối )

### **4.3 Caption preprocessing**

- Chuyển các ký tự trong cột 'caption' về chữ thường

- Loại bỏ các ký tự không phải là chữ cái trong cột 'caption'

- Loại bỏ các khoảng trắng dư thừa trong cột 'caption'

- Loại bỏ các từ có độ dài là 1 trong cột 'caption'

- Thêm "startseq" vào đầu và "endseq" vào cuối mỗi câu trong cột 'caption'

vd : “A girl going into a wooden building .”

=> :“startseq girl going into wooden building endseq”

**4.4 Caption Tokenization**

- Mã hóa văn bản thành các chuỗi số .

- Xây dựng từ điển từ vựng từ các câu mô tả (captions) .

- Xác định kích thước từ vựng .

- Xác định độ dài lớn nhất của câu mô tả trong tập dữ liệu.

### **4.5 Dữ liệu đầu vào**

Giả sử ảnh được trích xuất đặc trưng bằng InceptionV3:

Dữ liệu đầu vào gồm:

● Dữ liệu ảnh có kích thước (2048,)

● Một phần của caption. Ta padding một phần caption lên đến kích thước lớn nhất của caption là (34, ).

Dữ liệu đầu ra: Từ tiếp theo trong caption.

Ví dụ: hình ảnh 2654514044\_a70a6e2c21.jpg với mô tả đầu tiên “startseq brown dog running endseq”.

| Dữ liệu ảnh | Một phần của caption | Từ tiếp thep |
| --- | --- | --- |
| Ảnh đầu vào (2048, ) | startseq (34,) | brown |
| Ảnh đầu vào (2048, ) | startseq brown (34,) | dog |
| Ảnh đầu vào (2048, ) | startseq brown dog (34,) | running |
| Ảnh đầu vào (2048, ) | startseq brown dog running (34,) | endseq |

### **4.6 Word embedding**

Pre-trained GLOVE model được sử dụng cho quá trình word embedding. Ta có file glove.6B.200d.txt chứa tất cả các từ và mỗi từ gắn với vector đại diện với độ dài 200. Dựa vào file glove để thêm vector cho tập từ điển 1652 làm đầu vào cho embedding. Embedding không cần train.

## **5. Mô tả model**

Các mô hình có thể sử dụng như LSTM, Attension, Transformer.

Attension và Transformer là các mô hình học sâu mạnh mẽ nhưng cũng có một số hạn chế: Mô hình phức tạp có thể khó hiểu và triển khai. Yêu cầu nhiều dữ liệu đào tạo hơn so với các mô hình RNN truyền thống.

LSTM là một mô hình học sâu đơn giản, dể triển khai, dể train

Vì vậy nhóm chúng em quyết định tập trung vào LSTM

Mô hình được thiết kế theo kiến trúc Merging, tức là xử lý riêng biệt ảnh và văn bản đầu vào, sau đó kết hợp hai đối tượng lại ở bước cuối

Đối với công việc xử lý ảnh, chúng tôi sử dụng model Inception v3 có sẵn của Google. Model này đã được cắt bỏ lớp fully connected với mục đích dùng để trích xuất các đặc trưng của ảnh đầu vào. Vector đặc trưng được đưa qua lớp Dropout để tránh overfitting và tiếp tục qua lớp Dense để chuẩn hóa về chiều 256.

Đối với xử lý văn bản đầu vào, dữ liệu văn bản sẽ được chuyển thành dạng vector thông qua mô hình word embedding Glove.6b.200d. Dữ liệu sẽ được đưa qua lớp Dropout và tiếp tục được đưa vào lớp LSTM. Tại đây, lớp LSTM tính toán đầu ra và chuẩn hóa về chiều 256. Hình ảnh đã chuẩn hóa (có chiều 256) sẽ được cộng trực tiếp với vector đầu ra của lớp LSTM (có chiều 256) để tạo thành tổ hợp dữ liệu hình ảnh và từ. Tổ hợp này sau đó sẽ đưa qua lớp fully connected để dự đoán xác suất của từ tiếp theo trong câu văn.

## **6. Kết quả**

Mô hình LSTM\_1 đơn giản, nhưng có hiệu quả nhỉn hơn một chút nhỏ mô hình LSTM\_2