TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Đức Tài - Phạm Hoàng Hải Đăng
Lê Hồng Thạch - Trần Đình Tấn Phát
Phan Nguyễn Gia Huy - Nguyễn Đỗ Bảo Kiên

ĐỒ ÁN LÍ THUYẾT GIỚI THIỆU TRÍ TUỆ NHÂN TẠO

IMAGE CAPTIONING

CHƯƠNG TRÌNH CHÍNH QUY

Tp. Hồ Chí Minh, tháng /2024

Ngày 17 tháng 12 năm 2024

Mục lục

ĐÁNH GIÁ THÀNH VIÊN:

- Nguyễn Đức Tài: nhiệt tình, chăm chỉ, hoạt bát.
- Phạm Hoàng Hải Đăng: chăm chỉ, chịu khó, .
- Lê Hồng Thạch: nhiệt tình, chịu khó.
- Trần Đình Tấn Phát: năng nổ, hoạt bát.
- Phan Nguyễn Gia Huy: cần cù, chăm chỉ.
- Nguyễn Đỗ Bảo Kiên: hòa đồng, chăm chỉ.

ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH CHO TỪNG YÊU CẦU:

- Source code: xong
- slide: xong
- Report:xong

1 Introduction: Giới thiệu bài toán.

1.1 Khái niệm:

Image-captioning là việc làm cho máy tính có thể tự sinh mô tả cho một bức ảnh được nhập vào. Giống như cách mà chúng ta nhìn vào một bức ảnh và hiểu bức ảnh đó có nội dung là gì, máy tính cũng có thể làm tương tự.
VD:

1.2 Ứng dụng của image-captioning:

- Sinh mô tả cho ảnh: Số lượng ảnh lớn, khổng lồ việc gán nhãn thủ công sẽ mất thời gian và tiêu tốn tiền bạc. Hãy để AI gãn nhãn dùm
- Các ứng dụng hỗ trợ người khiếm thính (tích hợp với chuyển text thành audio để mô tả ảnh cho người bị mù. . .)
- ETC. . .

Input:

Output:

```
A child in a pink dress is climbing up a set of stairs in an entry way .
A girl going into a wooden building .
```

1.3 Cách mà mô hình image captioning hoạt động:

Hãy nghĩ về chú thích hình ảnh như một sự hợp tác giữa hai thành phần thiết yếu của bộ não máy tính:

- Con mắt (Mạng thần kinh tích chập - CNN): Cũng giống như chúng ta có mắt để nhìn, máy tính có CNN để phân tích hình ảnh. Các mạng này giúp máy tính xác định các yếu tố quan trọng trong hình ảnh, chẳng hạn như tai hoặc đuôi của mèo. Những yếu tố chính này được dịch thành một tập hợp các số đặc biệt mà máy tính hiểu được. Những số đặc biệt này được gọi là "vector embeddings".
- Miệng (Mạng thần kinh tuần hoàn - RNN): "Miệng"của máy tính là RNN. Nó lấy những con số đặc biệt đó (vector embeddings) từ CNN và kết hợp chúng với sức mạnh của từ ngữ. Nó giống như thể chúng ta đang dạy máy tính thuật lại một câu chuyện về hình ảnh. RNN lấy từng từ một và bắt đầu tạo thành một câu.

2 Methodology: Giải thích phần xử lý dữ liệu, giới thiệu mô hình, giải thích mô hình. Phần này t chỉ làm cái sườn chính thôi nên ae tự tìm kiếm them thông tin để soạn rõ hơn từng bước!

2.1 Xử lý dữ liệu (Data Preprocessing)

2.1.1 Chuẩn bị dữ liệu chú thích ảnh (Captions):

Dữ liệu thường được sử dụng có thể là bộ Flickr 8k, Flickr 30k hoặc bộ dữ liệu MS COCO. Ở đây mình sử dụng dữ liệu Flickr 8k được tải trực tiếp từ kaggle. Cách tải các bạn có thể xem trong file notebook.

Dữ liệu bao gồm một bộ 8000 ảnh và một file captions.txt. Mỗi ảnh sẽ có 5 captions làm nhãn. Cấu trúc file như sau: Image size của data này là (500,375,3)

Ví dụ về 1 ảnh và 5 captions của nó:

```
n = 17
img = Image.open(image_path + df['image'][5*n])
```

```
df = pd.read_csv("/content/train/captions.txt")
df
```

	image	caption
0	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set o...
1	1000268201_693b08cb0e.jpg	A girl going into a wooden building .
2	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse .
3	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playh...
4	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a woo...
...	...	...
40450	997722733_0cb5439472.jpg	A man in a pink shirt climbs a rock face
40451	997722733_0cb5439472.jpg	A man is rock climbing high in the air .
40452	997722733_0cb5439472.jpg	A person in a red shirt climbing up a rock fac...
40453	997722733_0cb5439472.jpg	A rock climber in a red shirt .
40454	997722733_0cb5439472.jpg	A rock climber practices on a rock climbing wa...

40455 rows × 2 columns

```
plt.figure(figsize=(7, 7))
plt.imshow(img)
print(df['caption'][5*n:5*n+5])
```

- Từ tệp captions.txt, các chú thích được tải lên và ánh xạ vào các hình ảnh tương ứng thông qua mã hình ảnh (image ID).
- Mỗi hình ảnh có 5 chú thích mô tả, tổng cộng có 40,455 chú thích cho 8,091 hình ảnh.

2.1.2 Tiền xử lý chú thích:

- Chuyển toàn bộ văn bản sang chữ thường (lowercase).
- Loại bỏ các ký tự không phải chữ cái (non-alphabetical characters).
- Xóa các khoảng trắng thừa.
- Thêm các token đặc biệt startseq và endseq để đánh dấu điểm bắt đầu và kết thúc của mỗi chú thích.

Ví dụ:

2.1.3 Token hóa văn bản (Tokenization):

- Sử dụng Tokenizer để chuyển đổi các từ trong chú thích thành các chỉ số số học, tạo điều kiện cho mô hình học tập.
- Lưu tokenizer vào tệp để tái sử dụng trong các bước sau.

2.1.4 Kích thước từ vựng và độ dài tối đa:

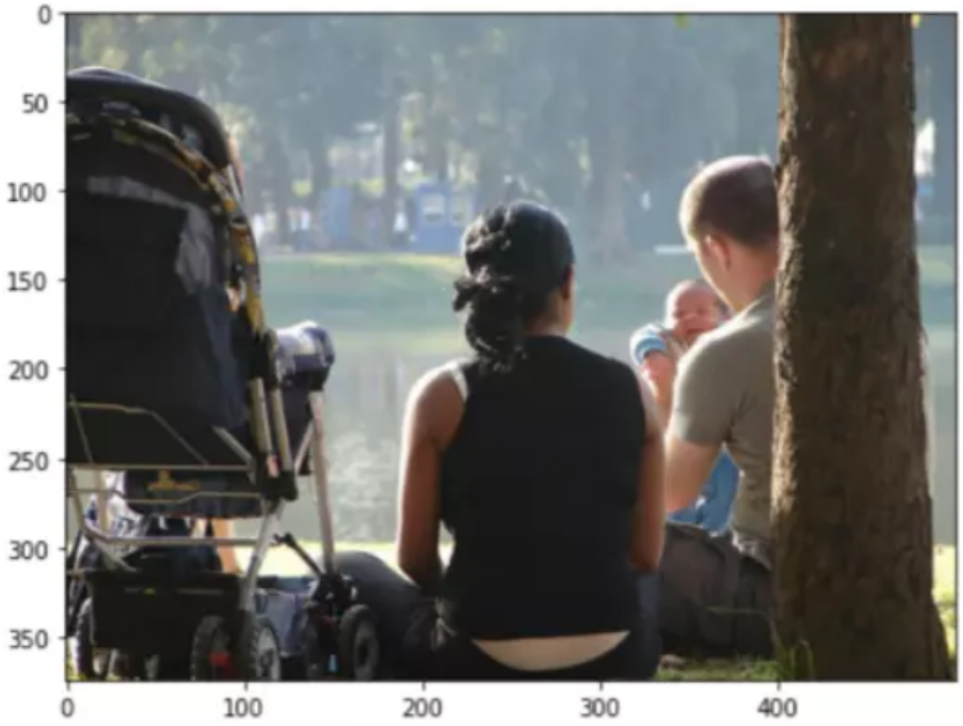
- Kích thước từ vựng: 8,768 từ.
- Độ dài chú thích tối đa: 34 từ.

2.2 Giới thiệu mô hình (Model Introduction) Mô hình bao gồm hai thành phần chính:

2.2.1 Trích xuất đặc trưng từ hình ảnh:

- Sử dụng mô hình VGG16 đã được huấn luyện trước trên tập dữ liệu ImageNet để trích xuất đặc trưng từ hình ảnh.
- Lớp Fully Connected (FC) cuối cùng của VGG16 bị loại bỏ để chỉ giữ lại các đặc trưng dạng không gian.

```
85    A couple and an infant , being held by the mal...
86    A couple sit on the grass with a baby and stro...
87    A couple with their newborn baby sitting under...
88    A man and woman care for an infant along the s...
89    Couple with a baby sit outdoors next to their ...
Name: caption, dtype: object
```



3

3.1 Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP):

Sử dụng mô hình LSTM (Long Short-Term Memory) để tạo ra các chú thích từ các đặc trưng hình ảnh.
LSTM là một loại mạng nơ-ron hồi tiếp (RNN) giúp xử lý dữ liệu chuỗi (như văn bản) và có khả năng ghi nhớ thông tin dài hạn, rất phù hợp cho bài toán tạo chú thích ảnh.

4 Giải thích mô hình(Model Explanation)

4.1 Kiến trúc mô hình

- **Đặc trưng hình ảnh(Image Feature)**
  Các đặc trưng từ VGG16 được nén lại thành một đầu vào vector.

  VGG16: Một mạng nơ-ron CNN (Convolutional Neural Network) được sử dụng để nén hình ảnh thành một đầu vào vector. VGG16 đã được huấn luyện trên ImageNet và có thể chuyển đổi hình ảnh thành các đặc trưng mạnh mẽ.

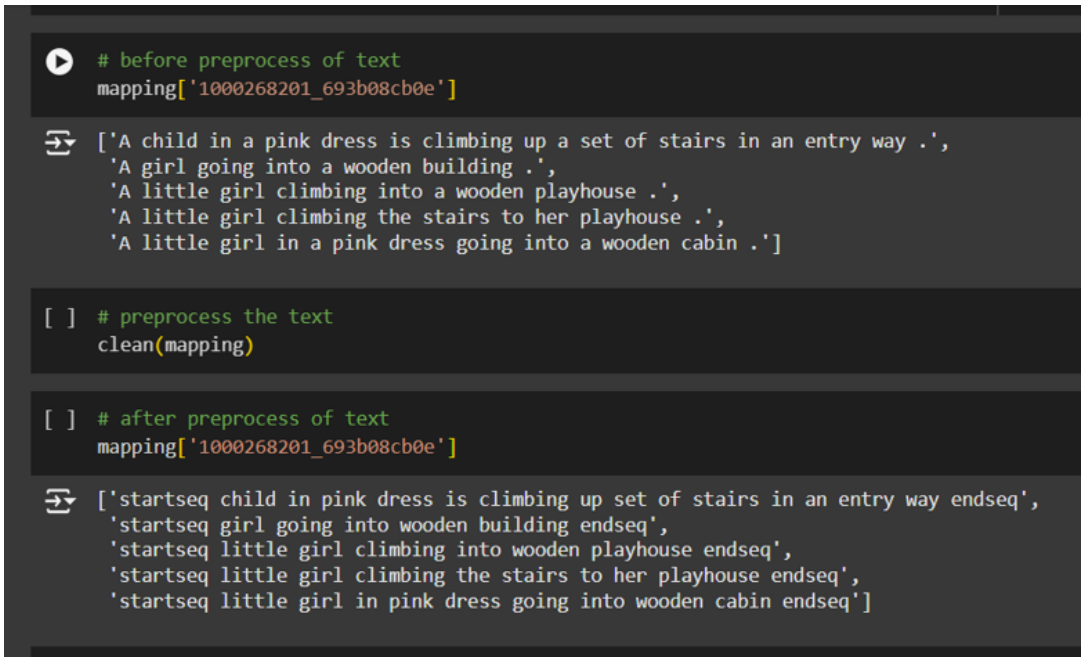
- **Chú thích dạng chuỗi(Caption Sequence)**
  Chú thích được token hóa và đưa vào LSTM.

  Token hóa: Chú thích được chia thành các từ (tokens) và được đưa vào LSTM (Long Short-Term Memory) để xử lý chuỗi các từ này. LSTM giúp duy trì thông tin lâu dài và giải quyết vấn đề của chuỗi dữ liệu.

- **Tích hợp đầu ra:**
  Đặc trưng hình ảnh và chuỗi chú thích được kết hợp thông qua lớp Dense.

  Đầu ra cuối cùng là một chuỗi từ dự đoán mô tả hình ảnh.

  Lớp Dense: Đặc trưng hình ảnh và chuỗi chú thích được kết hợp thông qua lớp Dense. Lớp Dense sẽ tích hợp các đầu vào và tạo ra một đầu ra cuối cùng.



Đầu ra cuối cùng: Đầu ra cuối cùng là một chuỗi từ dự đoán mô tả hình ảnh. Mô hình sẽ tạo ra một chuỗi từ dựa trên đặc trưng hình ảnh và chuỗi chú thích.

4.2 Quá trình huấn luyện:

- Tối ưu hóa hàm mất mát(loss function) sử dụng Categorical Crossentropy) Categorical Crossentropy: Đây là hàm mất mát phổ biến được sử dụng trong các bài toán phân loại. Nó đo lường sự khác biệt giữa chuỗi chú thích dự đoán và chuỗi chú thích thực tế.
- Áp dụng thuật toán Adam Optimizer để cập nhật trọng số. Adam Optimizer: Đây là một thuật toán cập nhật trọng số phổ biến trong mô hình học sâu. Adam kết hợp lợi ích của Gradient Descent với các cải tiến để tăng tốc độ và hiệu quả của quá trình huấn luyện.
- Mục tiêu là dự đoán từ tiếp theo trong chú thích dựa trên từ hiện tại và đặc trưng hình ảnh.

Các bước cụ thể trong quá trình huấn luyện:
1.Tách đặc trưng từ hình ảnh: Sử dụng mạng CNN như VGG16 để nén hình ảnh thành một đầu vào vector.
2.Token hóa chuỗi chú thích: Chú thích được chia thành các từ (tokens) và được đưa vào LSTM để xử lý chuỗi các từ này.
3.Kết hợp đầu ra: Đặc trưng hình ảnh và chuỗi chú thích được kết hợp thông qua lớp Dense.
4.Huấn luyện mô hình: Sử dụng hàm mất mát Categorical Crossentropy và thuật toán Adam Optimizer để cập nhật trọng số của mô hình.
5.Dự đoán từ tiếp theo: Mục tiêu của mô hình là dự đoán từ tiếp theo trong chú thích dựa trên từ hiện tại và đặc trưng hình ảnh.

BÁO CÁO LÀM VIỆC HÀNG TUẦN

- **(a) Ngày họp**: chủ nhật hàng tuần.
- **(b) Thành viên tham dự**: tất cả thành viên.
- **(c) Phân công**:
  - Thạch,Phát,Huy: viết latex(trong 1 tuần)
  - Đăng: powerpoint(trong 1 tuần)
  - Tài: source code(2 tuần)
  - Kiên: phần còn lại của công việc
- **(d) Tiến độ làm các tuần**:
  - tuần 1:
    - Nội dung phân công: tìm hiểu thông tin về project 2.

    * Người được phân công: cả nhóm.
    * Tiến độ làm: hoàn thành.
    * Công việc đã làm: tìm hiểu về image caption.
    * Công việc chưa làm: Phân chia công việc project.
- tuần 2:
    * Nội dung phân công: làm latex, powerpoint,source code, quay video.
    * Người được phân công:
        · Thạch, Huy, Phát: Latex.
        · Đăng: powerpoint.
        · Tài: source code.
        · Kiên, Đăng: quay video
    * Tiến độ làm: 50% công việc.
    * Công việc đã làm: Latex, powerpoint.
    * Công việc chưa làm: source code và quay video .
- Tuần 3:
    * Nội dung phân công: source code và quay video thuyết trình.
    * Người được phân công:
        · Tài làm source code
        · Kiên và Đăng quay video
    * Tiến độ làm: 100% công việc.
    * Công việc đã làm: Latex, powerpoint, source code và quay video.
    * Công việc chưa làm: hết .