Image Captioning

Báo cáo cuối kỳ

21120071 - Nguyễn Thị Thanh Hoa
21120175 - Tô Ngọc Hân
21120184 - Lê Thị Minh Thư

Trường Đại học Khoa học Tự nhiên VNU-HCM
Khoa Công nghệ thông tin
Môn: Thị giác máy tính

Mục lục

Mục lục

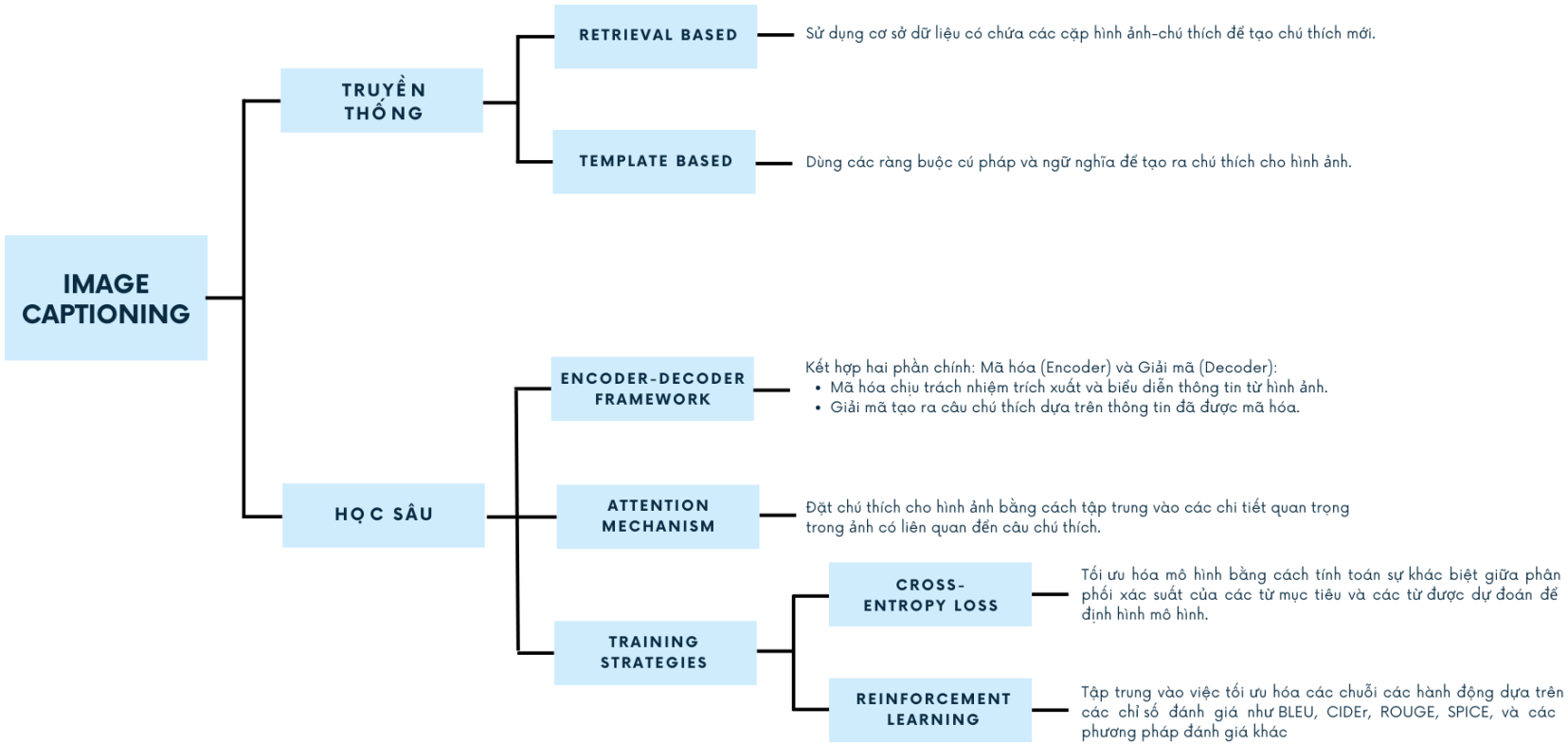
Image Captioning là gì?

- **Image Captioning** là quá trình tự động sinh ra mô tả văn bản cho các hình ảnh đầu vào. Mô tả này thường phản ánh các đặc điểm quan trọng của hình ảnh và có thể bao gồm thông tin về các đối tượng, hành động, và ngữ cảnh.
- **General system architecture**: Cấu trúc tổng quát của hệ thống Image Captioning bao gồm hai phần chính: mô hình trích xuất đặc trưng hình ảnh và mô hình sinh mô tả văn bản.
- **Ứng dụng**: Image Captioning được ứng dụng ở nhiều lĩnh vực như:
  + Tích hợp với các phần mềm, mô tả hình ảnh cho người khiếm thị.
  + Cải thiện khả năng tìm kiếm, phân loại số lượng lớn các hình ảnh dựa trên mô tả ảnh.
  + Cung cấp các mô tả sơ bộ về hình ảnh trong chẩn đoán Y khoa.
  + Tạo chú thích ảnh nhanh chóng trong ngành truyền thông và báo chí.

Image Caption trong đồ án

- Trong đồ án này, chúng tôi đề xuất mô hình CNN-LSTM để giải quyết bài toán Image Captioning, sử dụng cơ chế Encoder-Decoder.
- Để cải thiện mô hình, chúng tôi cũng đề xuất kết hợp mô hình CNN-LSTM với cơ chế chú ý (attention mechanism), giúp mô hình tập trung vào các phần quan trọng của hình ảnh trong quá trình sinh ra chuỗi mô tả.
- Mô hình của chúng tôi được thử nghiệm trên bộ dữ liệu Flickr8k. Và được đánh giá dựa trên các thông số: BLEU-4 và METEOR.

Các công trình nghiên cứu liên quan

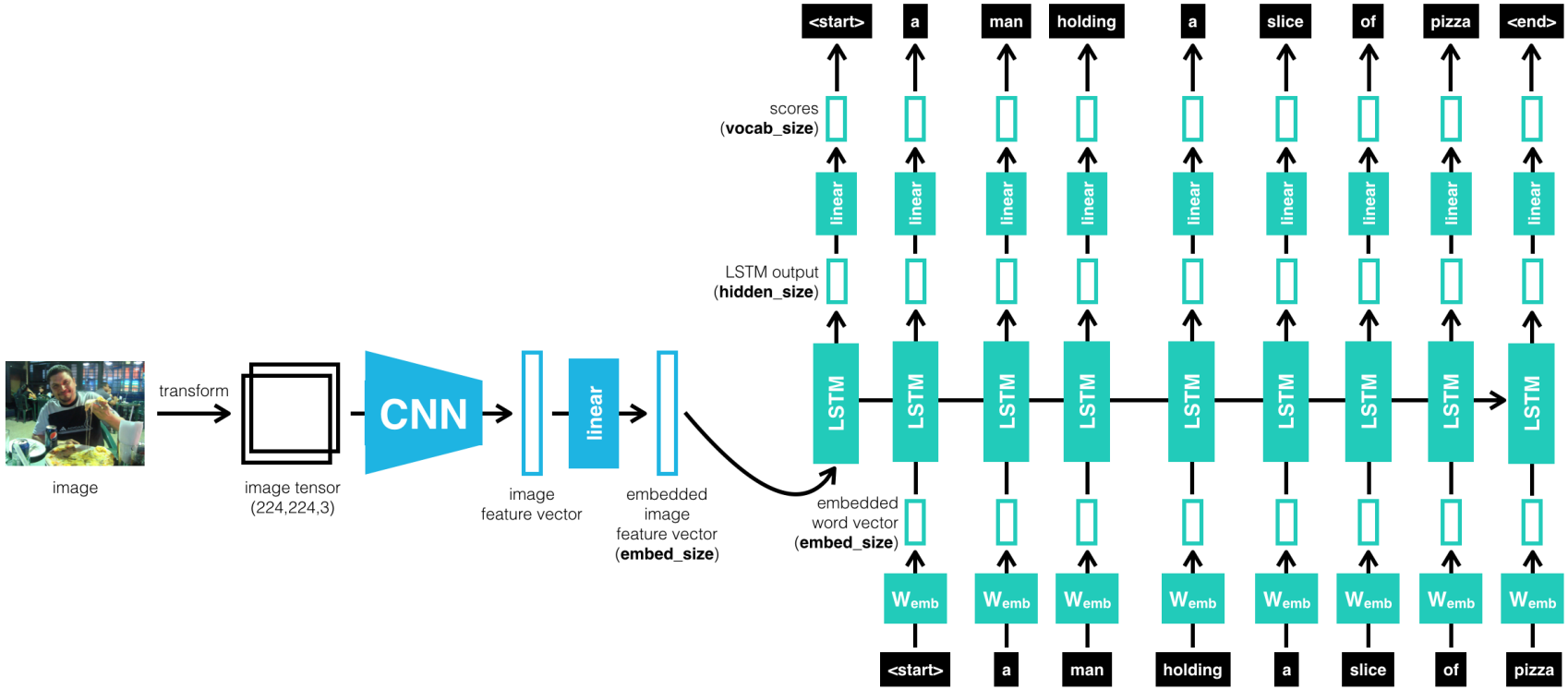


Hình: Các công trình nghiên cứu liên quan

Mục lục

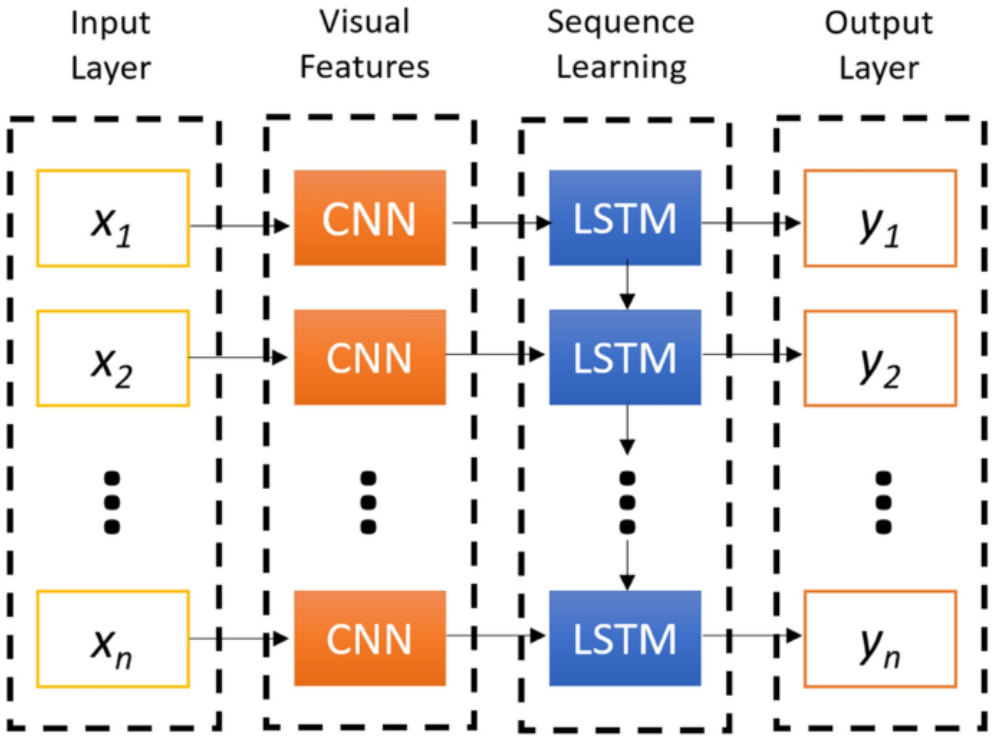
Mô hình CNN-LSTM

Đây là kiến trúc hoàn chỉnh của mô hình chú thích hình ảnh CNN-LSTM encoder-decoder:



Hình: CNN-LSTM Encoder-Decoder model

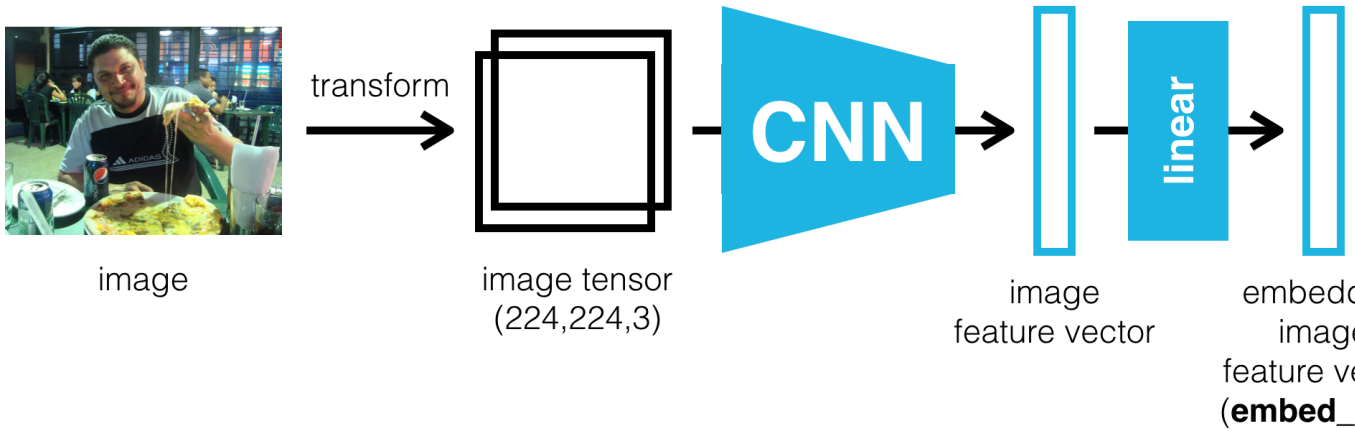
Mô hình CNN-LSTM



Hình: CNN-LSTM Encoder-Decoder model

Về cơ bản, bộ mã hóa CNN tìm các mẫu trong hình ảnh và mã hóa nó thành một vectơ được chuyển đến bộ giải mã LSTM để xuất ra một từ tại mỗi bước thời gian để mô tả hình ảnh tốt nhất. Khi đạt đến mã thông báo <end>, toàn bộ chú thích sẽ được tạo và đó là đầu ra của mô hình cho hình ảnh cụ thể đó.

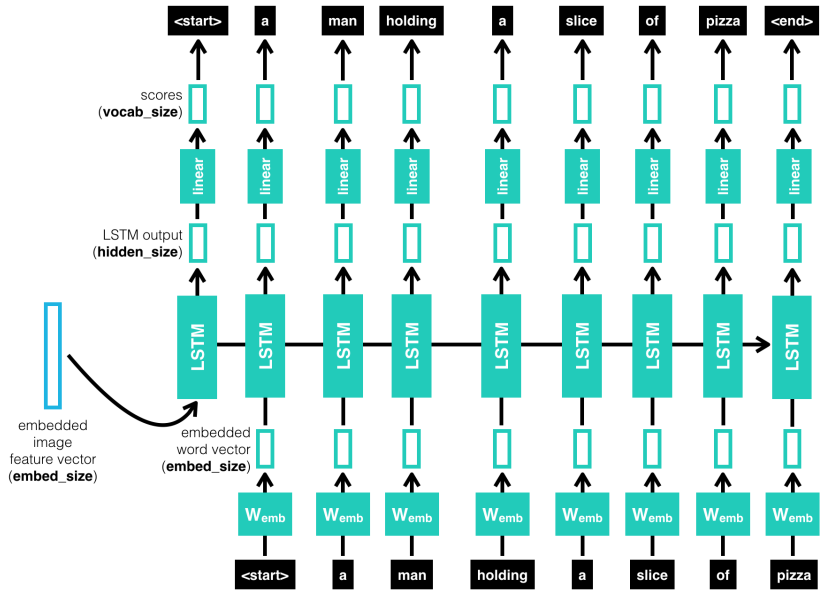
CNN Encoder



Hình: CNN Encoder

Bộ mã hóa dựa trên Mạng thần kinh tích chập (Convolutional Neural Network) mã hóa hình ảnh thành một biểu diễn nhỏ gọn đặc trưng (ở dạng nhúng - embedding). Bộ mã hóa CNN là ResNet (Residual Network). Những loại mạng này giúp giảm bớt các vấn đề về vanishing và exploding gradient. Nhóm sử dụng mô hình được đào tạo trước ResNet-50.

LSTM Decoder



Hình: LSTM Decoder

Bộ mã hóa CNN được theo sau bởi Mạng trí nhớ ngắn hạn định hướng dài hạn - Long Short–Term Memory (LSTM) tạo ra câu mô tả tương ứng. Vectơ đặc trưng được đưa vào "Bộ giải mã RNN"(được "mở"theo thời gian). Mỗi từ xuất hiện dưới dạng đầu ra ở trên cùng sẽ được đưa trở lại mạng dưới dạng đầu vào (ở dưới cùng) trong bước thời gian tiếp theo, cho đến khi toàn bộ chú thích được tạo. Mũi tên chỉ sang bên phải kết nối các hộp LSTM với nhau biểu thị thông tin trạng thái ẩn, đại diện cho "bộ nhớ"của mạng, cũng được phản hồi lại LSTM ở mỗi bước thời gian.

Mô hình CNN-LSTM

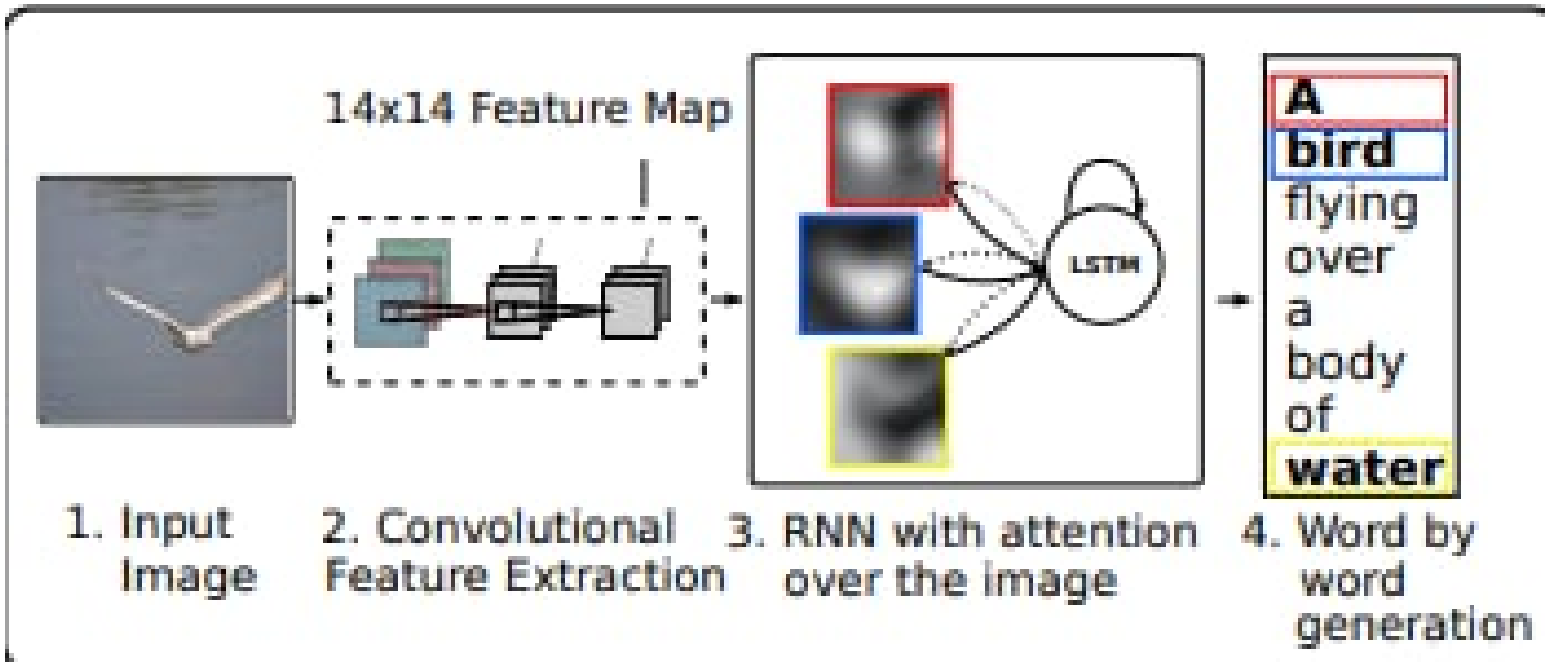
Hạn chế của mô hình CNN-LSTM: Toàn bộ thông tin của ảnh chỉ được đưa vào một lần duy nhất tại bước đầu tiên của LSTM.

Hướng có thể giải quyết: Đưa toàn bộ thông tin của ảnh vào từng bước.Tuy nhiên với những ảnh lớn, cứa nhiều thông tin và phức tạp thì xử lý như vậy không phù hợp.

Cơ chế Attention: giúp đưa thông tin của những vùng ảnh cần thiết vào từng bước sẽ phù hợp hơn so với hướng giải quyết trên.

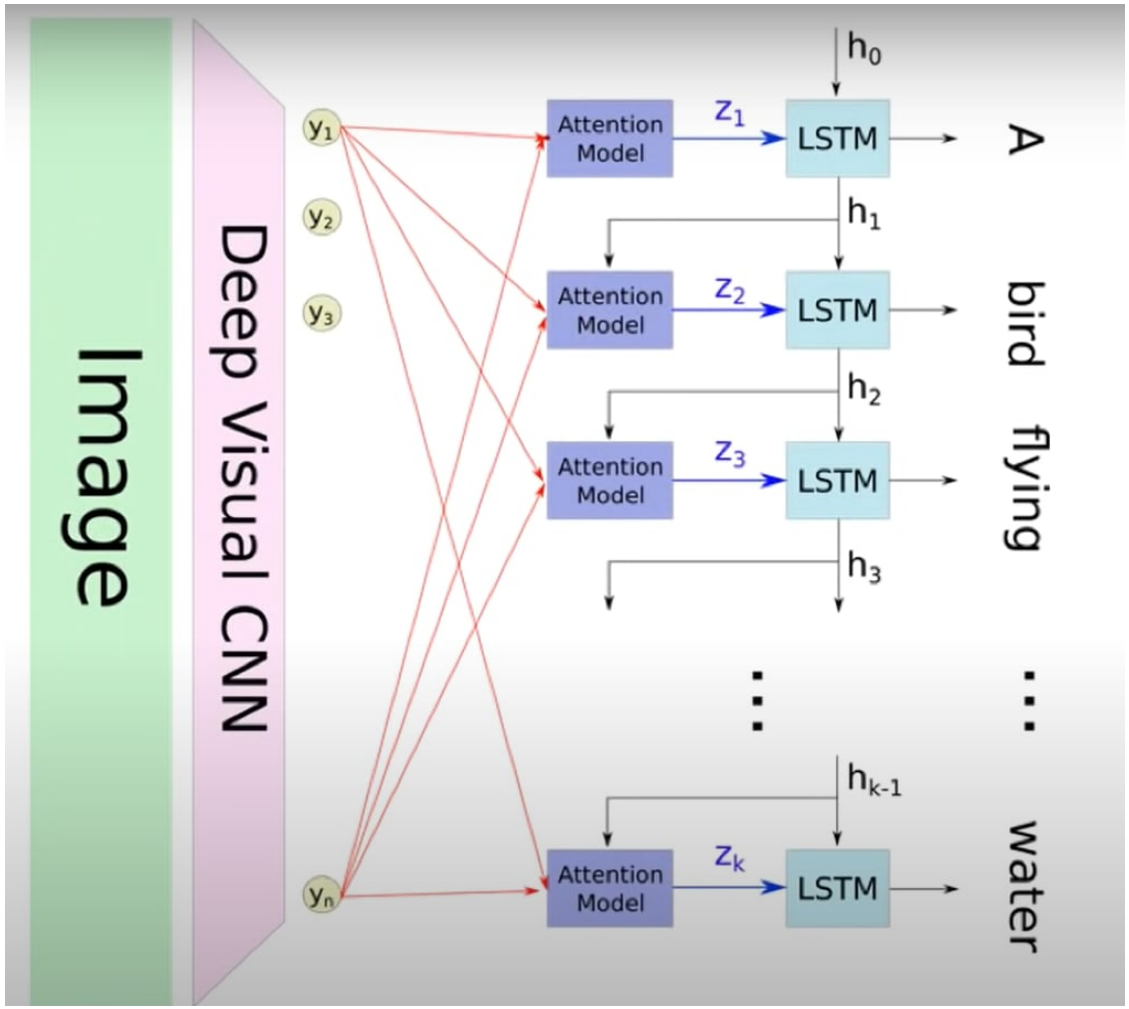
Mô hình CNN-LSTM-Attention

Đây là cấu trúc mô hình tự động mô tả nội dung ảnh có sử dụng cơ chế Attention tham khảo từ [Xu et al., 2016]



Hình: Minh họa cấu trúc mô hình sử dụng cơ chế Attention

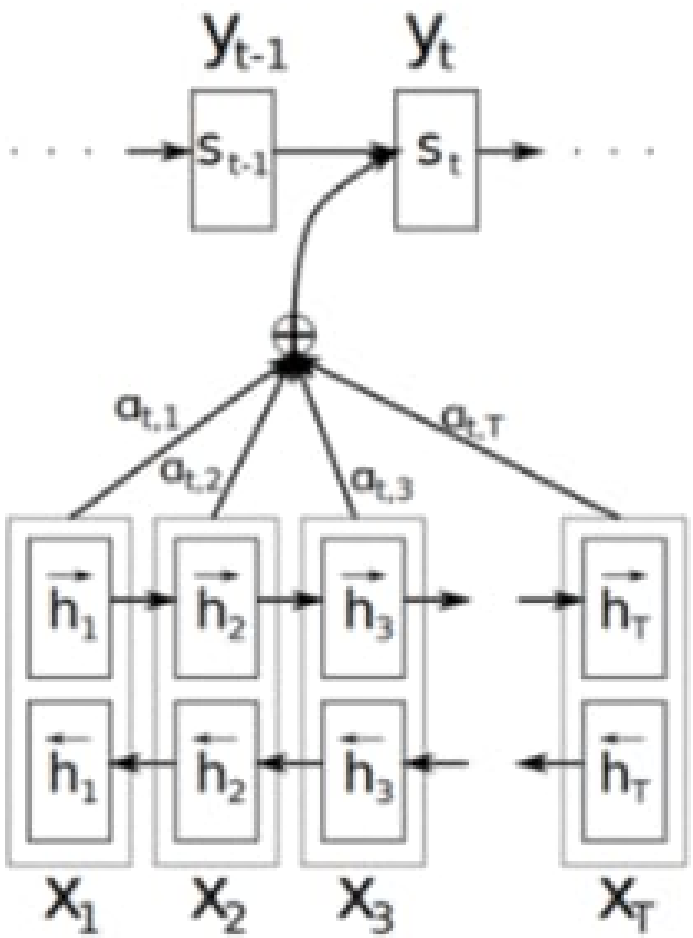
Mô hình CNN-LSTM-Attention



Hình: Minh họa mô hình CNN-LSTM-Attention với trạng thái ẩn tiếp theo h_t được tính toán dựa trên vec-tơ từ đầu vào tại bước đó , trạng thái ẩn trước đó h_{t1} và contex vector z_t. Cơ chế Attention nhận đầu vào là h_{t1} và y để tính toán context vector z_t.

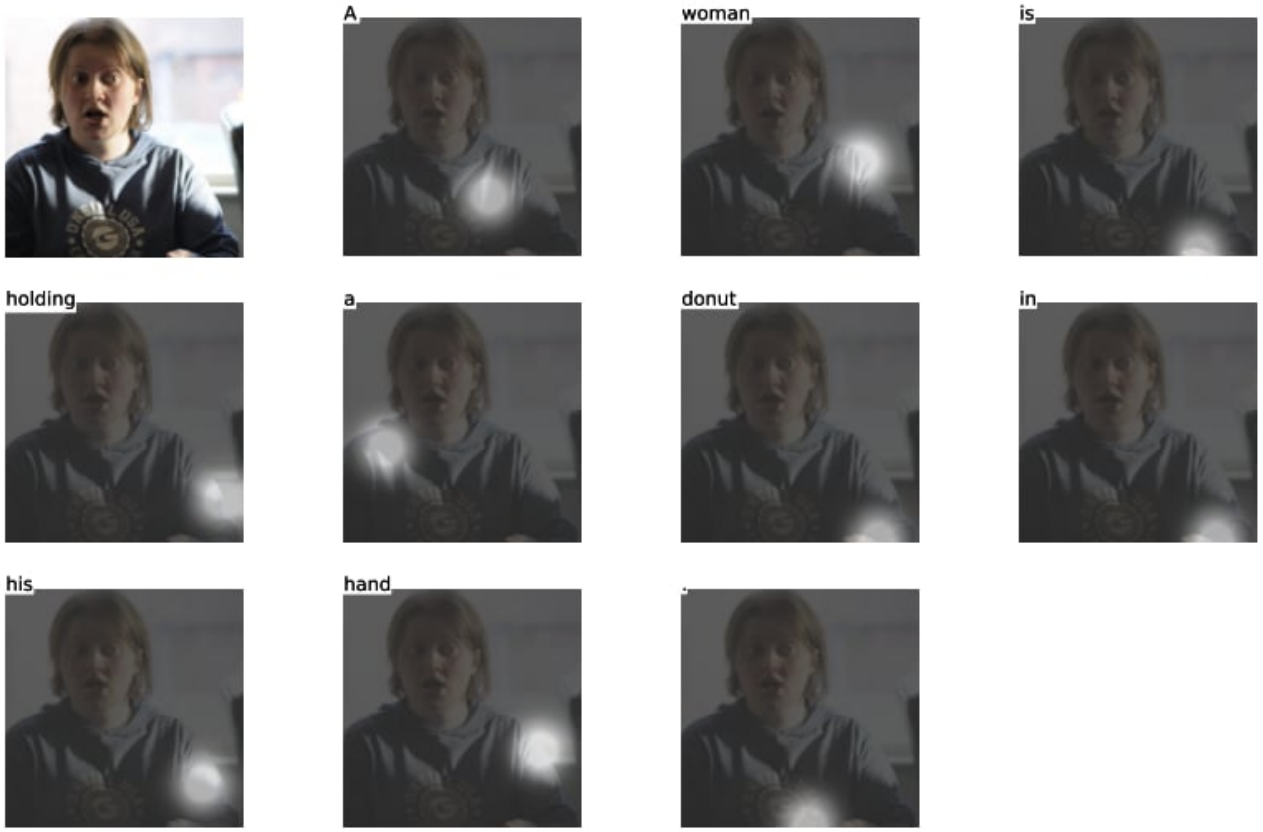
Cơ chế Attention

Mô hình sẽ tập trung vào những vùng có giá trị α cao. Các vùng này sẽ sáng hơn những vùng khác.



Hình: Minh họa cơ chế Attention

Cơ chế Attention



Hình: Minh họa các vùng ảnh mà mô hình tập trung vào

Mục lục

Flickr8k

Sử dụng bộ dữ liệu Flickr8k được tải trực tiếp từ Kaggle. Dữ liệu bao gồm một bộ 8000 ảnh và một file captions.txt. Image size của data này là (500,375,3). Mỗi ảnh sẽ có 5 captions làm nhãn. Cấu trúc file như sau:

```
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...

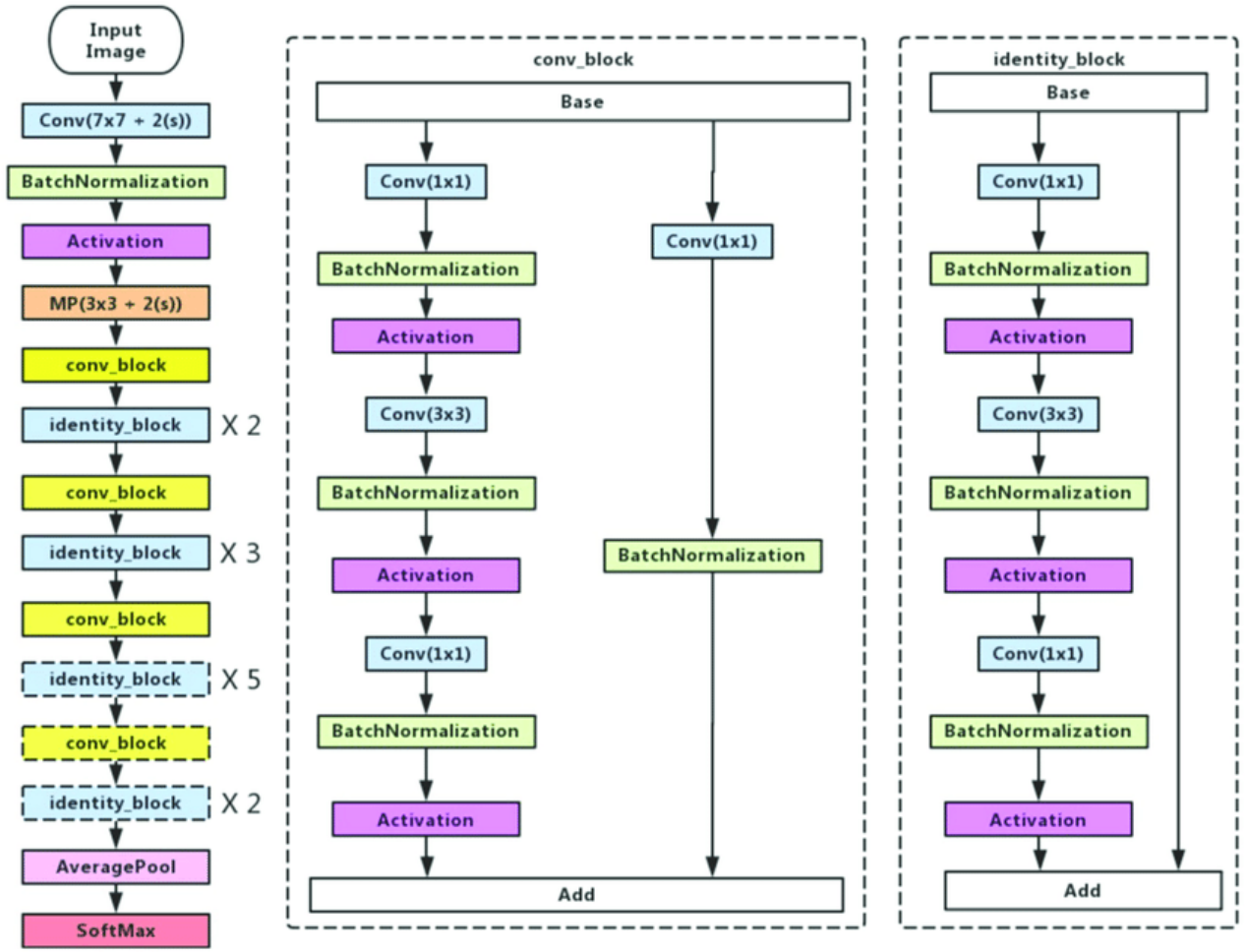
40455 rows × 2 columns

Hình: Cấu trúc file dữ liệu Flickr8k

Tập dữ liệu bao gồm hình ảnh của các cảnh quan, vật thể và chủ đề, vì thế nó rất phù hợp để giúp đánh giá hiệu suất của các mô hình chú thích hình ảnh trên nhiều tình huống trong thế giới thực. Bằng cách kết hợp Flickr8k vào thực nghiệm, nhóm nhắm đến việc tận dụng sự phong phú và đa dạng của nó để đánh giá kỹ lưỡng khả năng khái quát hóa của các kỹ thuật chú thích hình ảnh được sử dụng.

Kỹ thuật học chuyển tiếp

Mạng CNN được sử dụng là mạng Resnet-50 (gồm 16 tầng tích chập, 5 tầng max pooling và 3 tầng kết nối đầy đủ).



Hình: Minh họa kiến trúc mạng Resnet-50

Độ đo và thuật toán tối ưu

Độ đo: BLEU, METEOR.

- Sử dụng tỷ lệ trùng khớp n-gram (chuỗi ký tự liên tiếp n ký tự), BLEU xét giữa bản dịch máy và bản dịch của con người còn MENTEOR xét các yếu tố khác như ngữ nghĩa và ngữ pháp.
- Có miền giá trị [0, 1], điểm càng cao, chất lượng bản dịch càng tốt.
- BLEU có một số hạn chế, ví dụ như không tính đến ngữ nghĩa và ngữ pháp của bản dịch nên kết hợp METEOR để đánh giá chất lượng bản dịch máy toàn diện hơn .

Thuật toán tối ưu: Adam

- Kết hợp các ưu điểm của hai thuật toán tối ưu hóa khác là Momentum và RMSprop.
- Adam sử dụng hai biến trung gian m (trung bình động lượng của gradient), v (trung bình bình phương của gradient).
- Sau mỗi bước cập nhật tham số, Adam sẽ tính toán giá trị mới của m và v, sau đó sử dụng các giá trị này để điều chỉnh tốc độ học tập.

Kết quả tạo chú thích của 2 mô hình

Hai mô hình tạo ra một số chú thích có kết quả như sau
- Ví dụ 1:
  + Chú thích đúng: A person kayaking in the ocean .
  + Chú thích được mô hình không dùng Attention tạo ra: a man is standing on a beach looking at the water .
  + Chú thích được mô hình có dùng Attention tạo ra: a person is surfing on a boat .

Hình: Ảnh ví dụ 1

Kết quả tạo chú thích của 2 mô hình

- Ví dụ 2:
  + Chú thích đúng: A man is sitting on a bench , cooking some food .
  + Chú thích được mô hình không dùng Attention tạo ra: a boy in a red shirt is climbing a rock wall .
  + Chú thích được mô hình có dùng Attention tạo ra: a man in a red jacket is riding a bike on a wooden bench .

Hình: Ảnh ví dụ 2

So sánh hai mô hình dựa trên điểm đánh giá BLUE-4 và METEO

Kết quả đánh giá hai mô hình dựa trên độ đo BLUE-4 và METEOR được thể ở 2 bảng dưới đây

Bảng: Điểm BLUE-4

	Ví dụ 1	**Ví dụ 2**	**Ví dụ 3**	**Ví dụ 4**	**Ví dụ 5**
Không dùng Attention	0.037	0.086	0.092	0.086	0.021
Dùng Attention	0.056	0.132	0.089	0.499	0.023

Bảng: Điểm METEOR

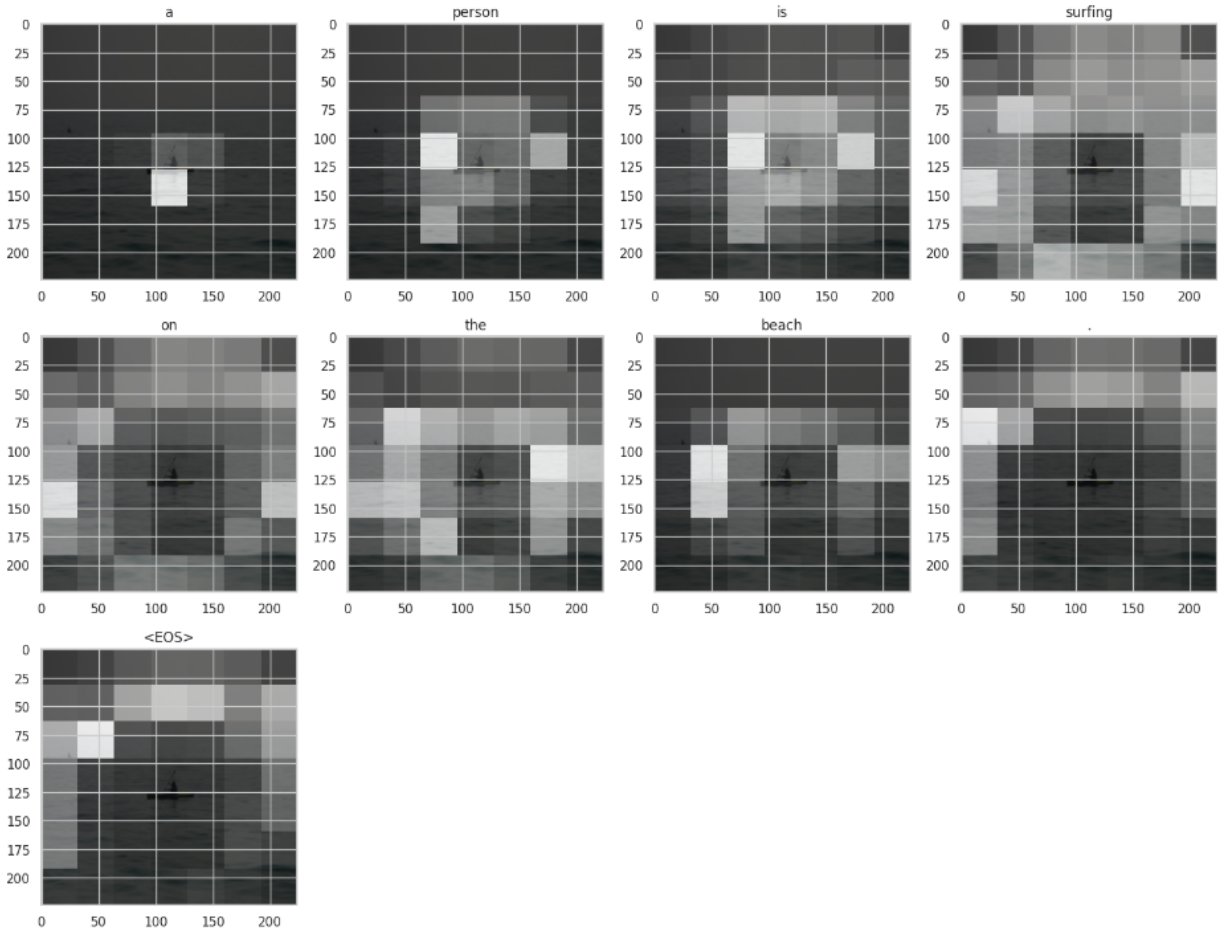
	Ví dụ 1	**Ví dụ 2**	**Ví dụ 3**	**Ví dụ 4**	**Ví dụ 5**
Không dùng Attention	0.168	0.422	0.463	0.216	0.187
Dùng Attention	0.221	0.603	0.326	0.763	0.106

So sánh hai mô hình dựa trên điểm đánh giá BLUE-4 và METEO

Nhận xét: Dựa trên kết quả đánh giá của độ đo BLUE-4 và METEOR sau khi thực nghiệm, có thể thấy mô hình CNN-LSTM-Attention có kết quả tốt hơn trên hầu hết tất cả các ví dụ thực nghiệm (ngoại trừ ví dụ 3). Mô hình CNN-LSTM-Attention được cung cấp thông tin những vùng ảnh cần tập trung tại từng bước trong quá trình phát sinh câu mô tả. Do đó, các từ chú thích được phát sinh dựa vào thông tin ảnh nhiều hơn và chính xác hơn.

Trực quan hóa sự chú ý (Attention) của mô hình

Hình ảnh dưới đây thể hiện sự trực quan hóa những vùng ảnh mô hình CNN-LSTM-Attention tập trung để phát sinh từ kế tiếp của mô hình mà nhóm đã xây dựng.



Hình: Trực quan hóa Attention của ảnh ví dụ 1

Mục lục

Tài liệu tham khảo I

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016).

Show, attend and tell: Neural image caption generation with visual attention.

arXiv:1502.03044.