Trình tạo chú thích hình ảnh sử dụng CNN và LSTM

Swarnim Tripathi
swarnim0711@gmail.com Đại
học Galgotias, Ấn Độ

Ravi Sharma
ravi.sharma@galgotiasuniversity.edu.in
Đại học Galgotias, Ấn Độ

Tóm tắt - Đối với bài báo này, chúng tôi sử dụng CNN và LSTM để nhận biết chú thích của hình ảnh. Tạo chú thích hình ảnh là một hệ thống hiểu được các tiêu chuẩn xử lý ngôn ngữ tự nhiên và thị giác máy tính để nhận dạng mối liên hệ của hình ảnh bằng tiếng Anh. Trong bài báo nghiên cứu này, chúng tôi thận trọng theo đuổi một số khái niệm quan trọng về chú thích ảnh và các quy trình quen thuộc của nó.

Chúng tôi nói về thư viện Keras, sổ ghi chép numpy và jupyter để tạo bài báo này. Chúng tôi cũng nói về lickr_dataset và CNN được sử dụng để phân loại ảnh.

Từ khóa- CNN,LSTM,chú thích hình ảnh, học sâu.

GIỚI THIỆU

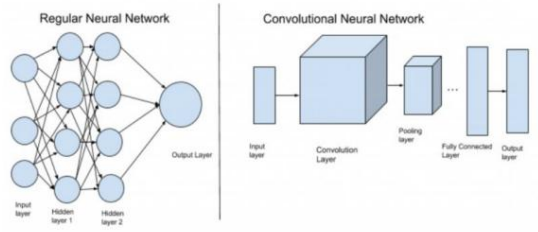
Mỗi ngày chúng ta thấy rất nhiều ảnh xung quanh, trên mạng xã hội và trên báo. Con người chỉ có thể tự nhận ra ảnh. Con người chúng ta có thể chọn ra những bức ảnh mà không cần chú thích được chỉ định nhưng mặt khác máy móc cần hình ảnh để được đào tạo trước rồi mới tự động tạo ra chú thích cho ảnh.

Chú thích hình ảnh có thể mang lại nhiều lợi ích, ví dụ như hỗ trợ người khiếm thị sử dụng chức năng chuyển văn bản thành giọng nói thông qua phản hồi thời gian thực về việc bao quát tình huống qua nguồn cấp dữ liệu camera, cải thiện giải trí y tế xã hội bằng cách sắp xếp lại chú thích cho ảnh trong nguồn cấp dữ liệu xã hội cùng với tin nhắn thành giọng nói.

Giúp trẻ em nhận biết các chất hơn nữa để có được kiến thức về ngôn ngữ. Chú thích cho mọi bức ảnh trên web toàn cầu có thể tạo ra các bức ảnh chân thực nhanh hơn và chi tiết hơn khi khám phá và lập chỉ mục. Chú thích hình ảnh có nhiều gói khác nhau trong nhiều lĩnh vực bao gồm y sinh học, thương mại, tìm kiếm trên internet và hải quân và nhiều phương tiện khác. Phương tiện truyền thông xã hội như Instagram, Facebook, v.v. có thể tạo chú thích thường xuyên từ hình ảnh. Mục tiêu chính của bài nghiên cứu này là có được một chút chuyên môn về các chiến lược học sâu. Chúng tôi sử dụng hai chiến lược đặc biệt là CNN và LSTM để phân loại hình ảnh.

KỸ THUẬT CHỈ DẪN HÌ NH ẢNH

CNN - Hệ thống nơ-ron tích chập là hệ thống nơ-ron quan trọng cụ thể có thể tạo ra thông tin có hình dạng thông tin, ví dụ, mạng lưới 2D và CNN có giá trị khi làm việc với hình ảnh. Nó kiểm tra hình ảnh từ góc trái sang góc phải và xuyên qua để trích xuất các điểm nổi bật đáng kể từ hình ảnh và hợp nhất phần tử để mô tả hình ảnh. Nó có thể xử lý các hình ảnh được diễn giải, xoay, thu nhỏ và sửa đổi. Hệ thống nơ-ron tích chập là một phép tính học sâu sắc tiếp nhận hình ảnh thông tin, phân bổ ý nghĩa cho các thành phần/phản đối khác nhau trong hình ảnh và nhận dạng chúng từ nhau.

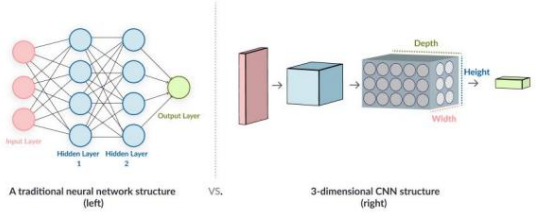


Hình 1 Kiến trúc CNN Việc xử lý trước cần thiết trong ConvNet không đáng kể khi so sánh với các phép tính thứ tự khác. Mặc dù các kênh được thiết kế thủ công theo các chiến lược thô sơ, nhưng với sự chuẩn bị đầy đủ, ConvNets phù hợp để học các kênh/điểm nổi bật này. Cấu trúc của hệ thống cong giống như thiết kế mạng nơ-ron bên trong tâm trí con người & được lấy cảm hứng từ cách tổ chức của vỏ não thị giác. Các nơ-ron đơn lẻ

phản ứng với các nâng cấp chỉ trong một khu vực giới hạn của trường nhìn thấy được gọi là trường mở. Sự phân loại các trường như vậy bao gồm tổng hợp các hình ảnh vùng.

CNN: Kiến trúc - Một mạng nơ-ron thô sơ thuần túy, bất kể vị trí nào tất cả các nơ-ron trong một lớp hợp nhất với tất cả các nơ-ron trong lớp tiếp theo đều không hiệu quả khi phân tích hình ảnh và video lớn. Đối với một hình ảnh có kích thước bình thường với nhiều thành phần hình ảnh được gọi là pixel & màu 3 tông (RGB tức là màu đỏ, màu xanh lá cây, màu xanh lam), phạm vi hạn chế sử dụng hệ thống nơ-ron được chấp nhận sẽ là hàng tấn, & điều đó có thể dẫn đến quá mức.

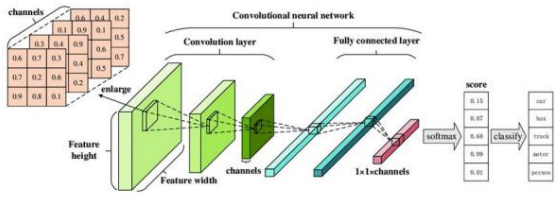
Để hạn chế số lượng hạn chế và nhận dạng hiệu quả của hệ thống thần kinh trên các phần quan trọng của hình ảnh, CNN sử dụng một sắp xếp 3D trong đó mỗi điều chỉnh của các tế bào thần kinh chia nhỏ một khu vực nhỏ hoặc "điểm nổi bật" của hình ảnh. Thay vì tất cả các tế bào thần kinh bỏ qua các lựa chọn của chúng đến lớp thần kinh tiếp theo, mỗi nhóm tế bào thần kinh dành nhiều thời gian để phân biệt một phần của hình ảnh, chẳng hạn như mũi, tai trái, miệng hoặc chân. Kết quả cuối cùng là một điểm phạm vi, minh họa cách hợp lý mà mỗi khả năng được bầu làm một phần của lớp.



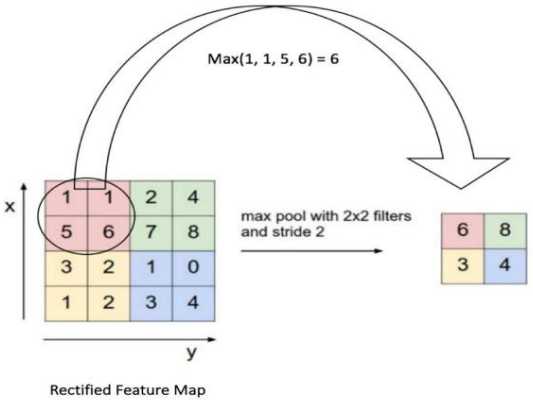
Hình 2 Hoạt động của CNN

CNN hoạt động như thế nào?

Như chúng ta đã thảo luận trước đây, một mạng nơ-ron được kết nối đầy đủ, trong đó đầu vào ở các lớp trước được kết nối với mọi đầu vào ở các lớp sau sẽ thuận tiện cho nhiệm vụ trong tầm tay, theo CNN, các nơ-ron trong một tế bào có thể được kết nối với một vùng tế bào cụ thể trước nó, thay vì tất cả các nơ-ron theo cách hoàn toàn giống nhau.



Điều này giúp giảm độ phức tạp của mạng nơ-ron và thu được ít năng lực tính toán hơn. Theo máy tính mới theo hình ảnh chuẩn với việc sử dụng các số ở mỗi pixel. Khi chúng ta thường so sánh hai hình ảnh chúng ta kiểm tra giá trị pixel của từng pixel. Kỹ thuật này chỉ giúp chúng ta so sánh hai hình ảnh giống hệt nhau nhưng khi chúng ta giữ các hình ảnh khác nhau để so sánh thì việc so sánh sẽ không thành công. Trong CNN, việc so sánh hình ảnh diễn ra từng phần.



Hình 3 Bản đồ đặc điểm của hình ảnh CNN

Lý do chính đằng sau việc sử dụng thuật toán CNN là đây là thuật toán duy nhất lấy hình ảnh làm đầu vào và dựa trên hình ảnh đầu vào để vẽ bản đồ đặc điểm, tức là phân loại từng pixel dựa trên sự giống nhau và khác nhau. CNN phân loại các pixel và tạo ra một ma trận, được gọi là bản đồ đặc điểm. Bản đồ đặc điểm là tập hợp các pixel tương tự được đặt trong một danh mục riêng biệt. Các ma trận này đóng vai trò quan trọng trong việc tìm ra bản chất của sự vật trong hình ảnh đầu vào.

Thêm thông tin về CNN -

Có tổng cộng 3 loại lớp trong mô hình CNN-

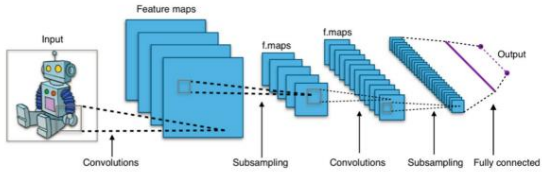
1. Tích chập
2. Gộp chung
3. Kết nối đầy đủ

Ở lớp đầu tiên, hình ảnh đầu vào được đọc qua CNN và trên nền tảng đó, một bản đồ đặc điểm được tạo ra.
Từ bản đồ đặc điểm đó, nó đóng vai trò là đầu vào cho các lớp sau, tức là cho lớp Pooling. Trong lớp pooling, bản đồ đặc điểm được chia thành các phần đơn giản hơn để kiểm tra cẩn thận bối cảnh của hình ảnh. Lớp này làm cho bản đồ đặc điểm dày đặc hơn để khám phá thông tin quan trọng nhất

về bức tranh.

Lớp thứ nhất và thứ hai tức là Convolutional và Pooling được thực hành rất nhiều lần, tùy thuộc vào hình ảnh để có được thông tin dày đặc về hình ảnh. Bản đồ đặc trưng dày đặc hơn được tạo ra nhờ hai lớp này. Và bản đồ đặc trưng dày đặc này được lớp cuối cùng tức là Fully Connected sử dụng.

Lớp này thực hiện phân loại. Nó sắp xếp các pixel theo mức độ giống nhau và khác nhau.
Phân loại được thực hiện ở mức độ đặc biệt để nắm được bản chất của bức tranh, giúp xác định các đối tượng, người, sự vật, v.v.



Hình 4 Các lớp hình ảnh được quét

Các lớp này giúp CNN định vị và tìm thấy các đặc điểm của hình ảnh một cách rõ ràng. Việc trích xuất các đặc điểm quan trọng có trong hình ảnh của các đầu vào có độ dài cố định được chuyển thành các đầu ra có kích thước cố định.

Các kỹ thuật CNN được sử dụng rất nhiều, ví dụ:

· Tầm nhìn máy tính— trong lĩnh vực khoa học y tế, phân tích hình ảnh chỉ được thực hiện thông qua CNN. Cấu trúc bên trong của cơ thể được kiểm tra dễ dàng với sự trợ giúp của công nghệ này.

Trong điện thoại di động, nó được sử dụng cho rất nhiều mục đích, ví dụ, để tìm tuổi của người đó, để mở khóa điện thoại bằng cách xem hình ảnh từ
máy ảnh.

Trong công nghiệp, nó được sử dụng rộng rãi để cấp bằng sáng chế hoặc bản quyền cho những hình ảnh cụ thể.

· Khám phá dược phẩm— được sử dụng rộng rãi để khám phá thuốc/dược phẩm, bằng cách phân tích các đặc tính hóa học và tìm ra loại thuốc tốt nhất để chữa một vấn đề cụ thể.

Nguồn gốc của LSTM:-

LSTM lần đầu tiên được tìm kiếm bởi hai nhà nghiên cứu người Đức - Sepp Hochreiter và Jurgen Schmidhuber, vào năm 1997. LSTM là viết tắt của long short-term memory (bộ nhớ dài hạn ngắn hạn). Trong lĩnh vực học sâu về mạng nơ-ron hồi quy, LSTM giữ một vị trí quan trọng. Yếu tố đặc biệt về LSTM là nó không chỉ lưu trữ dữ liệu đầu vào mà còn có thể cung cấp dự đoán về các tập dữ liệu tiếp theo thông qua chính nó. Mạng LSTM này lưu giữ dữ liệu đã lưu trữ trong một khoảng thời gian cụ thể và trên cơ sở đó dự đoán hoặc đưa ra các giá trị tương lai cho dữ liệu. Đây là mục đích chính khiến LSTM được sử dụng ở đây nhiều hơn so với RNN truyền thống.

Vấn đề với RNN (Recurrent Neural Mạng lưới):-

RNN là một phần của bộ quy tắc học sâu được thực hiện để xử lý một số tác vụ máy tính phức tạp hoặc phức hợp như phân loại mục và nhận dạng giọng nói. RNN được thực hiện để giải quyết một loạt các hoạt động phát sinh theo chuỗi, với thông tin của mọi tình huống hoàn toàn dựa trên số liệu thống kê từ các tình huống trước đó.

Một cách tinh tế, chúng tôi có ý định ưu tiên các RNN có bộ sưu tập dữ liệu mở rộng và khả năng cao hơn. RNN này có thể được sử dụng để thực hiện nhiều vấn đề thực tế như dự báo hàng tồn kho và tăng cường nhận dạng giọng nói. Tuy nhiên, RNN không được sử dụng để giải quyết các vấn đề thực tế và đó là do vấn đề Vanishing Gradient.

Vấn đề biến mất độ dốc -

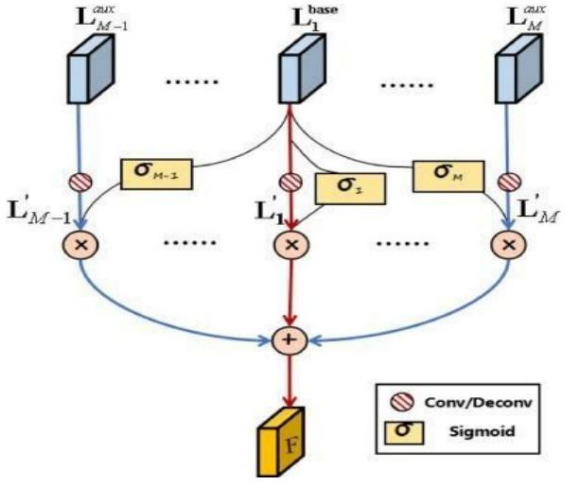
Vấn đề gradient biến mất này là nguyên nhân chính khiến cho hoạt động của RNN trở nên khó khăn. Nhìn chung, kỹ thuật của RNN được thực hiện sao cho nó lưu trữ dữ liệu trong một khoảng thời gian ngắn và lưu trữ một số mảng dữ liệu. RNN không thể nhớ tất cả các giá trị dữ liệu và trong một khoảng thời gian dài. RNN chỉ có thể lưu trữ một số dữ liệu trong một khoảng thời gian ngắn. Do đó, khả năng ghi nhớ của RNN chỉ có lợi cho các mảng dữ liệu ngắn hơn và trong khoảng thời gian ngắn.
Vấn đề gradient biến mất này trở nên rất nổi bật so với RNN truyền thống - để giải quyết một vấn đề cụ thể, nó thêm rất nhiều bước thời gian, dẫn đến mất dữ liệu khi chúng ta sử dụng backpropagation. Với rất nhiều bước thời gian, RNN phải lưu trữ các giá trị dữ liệu của mỗi bước thời gian, dẫn đến lưu trữ ngày càng nhiều giá trị dữ liệu và điều đó không khả thi trong trường hợp của RNN. Và bằng cách này, vấn đề gradient biến mất được hình thành.

Có thể làm gì để giải quyết tình trạng biến mất này?
Vấn đề gradient với RNN -

Để giải quyết vấn đề này, chúng ta sẽ sử dụng Long short-term memory (LSTM), là một tập hợp con của RNN. LSTM về cơ bản được xây dựng để khắc phục vấn đề Vanishing Gradients. Điều đặc biệt về LSTM là nó có thể bảo toàn các giá trị dữ liệu trong khoảng thời gian dài

của thời gian và do đó có thể giải quyết được vấn đề độ dốc biến mất.

LSTM được xây dựng theo cách mà chúng luôn chứa lỗi. Và do những lỗi này, LSTM tiếp tục nghiên cứu các giá trị dữ liệu qua nhiều bước thời gian. Do nghiên cứu các giá trị dữ liệu nhiều lần, nên việc nghiên cứu lan truyền ngược theo thời gian và các lớp trở nên dễ dàng.



Hình 5 Cổng trong LSTM

Lý tưởng nhất là theo sơ đồ trên, LSTM sử dụng một số cổng để lưu trữ dữ liệu và sau đó xử lý dữ liệu và gửi kết quả đến cổng cuối cùng. Khi chúng ta nói về RNN, chúng thường truyền dữ liệu đến cổng cuối cùng mà không cần bất kỳ xử lý nào.
Từ các cổng này trong LSTM, toàn bộ mạng có thể định hình dữ liệu theo nhiều dạng, bao gồm lưu trữ dữ liệu và xem xét dữ liệu từ các cổng. Các cổng trong LSTM có khả năng độc lập để đưa ra phán đoán liên quan đến các sự kiện và dữ liệu.

Hơn nữa, những cánh cổng này có khả năng tự đưa ra quyết định bằng cách mở hoặc đóng cổng.

Việc hiểu các cổng LSTM để lưu trữ dữ liệu trong một khoảng thời gian mang lại lợi ích cho LSTM so với RNN.

Kiến trúc của LSTM :-

Kiến trúc của LSTM rất đơn giản, bao gồm 3 cổng chính, lưu trữ dữ liệu trong thời gian dài hơn và giúp giải quyết những khó khăn mà RNN không giải quyết được.

3 cổng chính của LSTM bao gồm:

· Cổng quên – công việc chính của cổng quên là lọc dữ liệu, tức là xóa tất cả dữ liệu đó không cần thiết trong tương lai để giải quyết một nhiệm vụ cụ thể. Cổng này chịu trách nhiệm về hiệu suất chung của LSTM, nó tối ưu hóa dữ liệu.

· Cổng vào – LSTM bắt đầu từ cổng này, tức là cổng vào. Cổng này lấy dữ liệu đầu vào từ người dùng và cung cấp dữ liệu đầu vào cho các cổng khác.

· Cổng đầu ra – Cổng này có chức năng hiển thị kết quả mong muốn theo cách phù hợp.

Công dụng của Mạng bộ nhớ dài hạn ngắn hạn:

LSTM được sử dụng sâu sắc và chủ yếu cho nhiều nhiệm vụ học sâu bao gồm dự báo dữ liệu dựa trên dữ liệu trước đó. 2 minh họa đáng chú ý bao gồm dự đoán văn bản và dự đoán thị trường chứng khoán.

Dự đoán văn bản - LSTM được sử dụng rất nhiều trong việc dự đoán văn bản. Bộ nhớ dài hạn, sự hiểu biết về LSTM khiến nó đủ khả năng dự đoán các từ tiếp theo trong câu. Đây là kết quả của mạng LSTM trong việc tự dự đoán các từ tiếp theo. Trước tiên, LSTM lưu trữ dữ liệu, cảm giác của các từ, kiểu dáng của các từ, cách sử dụng các từ trong một tình huống cụ thể, v.v. và trên cơ sở đó dự đoán các từ tiếp theo. Dữ liệu được lưu trữ, tức là dữ liệu đầu vào được sử dụng thêm cho mục đích sử dụng trong tương lai.

Minh họa tốt nhất về dự đoán văn bản là Chatbot, được sử dụng rộng rãi trên các trang web thương mại điện tử và ứng dụng di động.

Dự đoán thị trường chứng khoán - Trong thị trường chứng khoán, LSTM cũng lưu trữ dữ liệu hoặc xu hướng mà thị trường hoạt động tại một thời điểm cụ thể, tại một thời điểm cụ thể và trên cơ sở đó dự đoán các biến thể và xu hướng tiếp theo của thị trường. Đây là một nhiệm vụ khó khăn để dự đoán biến thể trên thị trường chứng khoán vì các biến thể của thị trường rất khó để dự đoán và dự báo. Mô hình LSTM phải được đào tạo theo cách cung cấp các giá trị chính xác cho người dùng. Để làm được điều đó, rất nhiều dữ liệu phải được lưu trữ trong một thời gian dài, thậm chí có thể mất nhiều ngày.

Tìm hiểu thêm về LSTM -

LSTM về cơ bản là một phần của RNN, có khả năng lưu trữ nhiều giá trị dữ liệu hơn so với RNN. LSTM được sử dụng rộng rãi ngày nay trong mọi lĩnh vực. Sơ đồ đơn giản nhất của LSTM được hiển thị bên dưới. Nó bao gồm 3 cổng chính là, cổng Forget, cổng vào, cổng ra. Các cổng này có khả năng lưu trữ dữ liệu và đưa ra đầu ra mong muốn.
Bất cứ khi nào nói về mạng LSTM, ba cổng luôn xuất hiện.

Sơ đồ bên dưới cho thấy kiến trúc đơn giản nhất của LSTM:

Hình 6 Hoạt động của LSTM

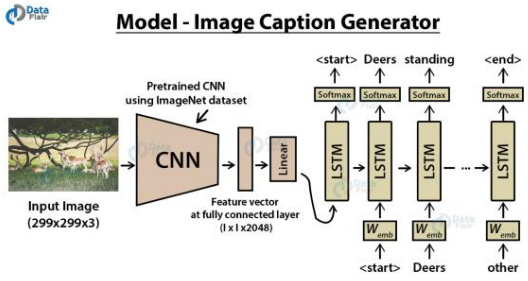
Mô hình tạo chú thích hình ảnh:-

Để chuẩn bị một mô hình tạo chú thích hình ảnh, chúng tôi sẽ tóm tắt hai kiến trúc khác nhau. Nó được gọi là CNN-LSTM

mô hình. Vì vậy, trong phần này chúng ta sẽ sử dụng hai kiến trúc này để lấy chú thích cho hình ảnh đầu vào.

· CNN - được sử dụng để trích xuất các đặc điểm quan trọng từ hình ảnh đầu vào. Để thực hiện điều này, chúng tôi đã lấy một mô hình được đào tạo trước để xem xét có tên là Xception.

· LSTM - được sử dụng để lưu trữ dữ liệu hoặc các tính năng từ mô hình CNN và xử lý thêm để hỗ trợ tạo chú thích hay cho hình ảnh.



Hình 7 Mô hình CNN-LSTM

Kiến trúc tập tin dự án :-

Để phục vụ mục đích nghiên cứu, chúng tôi đã tải xuống bộ dữ liệu bao gồm các tệp sau:

· Flickr8k_Datasets – Tệp này chứa tất cả các hình ảnh mà trước tiên chúng ta phải đào tạo mô hình của mình. Có 8091 hình ảnh.

· Flickr8k_texts – Thư mục này chứa các tệp văn bản và chú thích được định dạng sẵn cho hình ảnh.

Các tập tin sau đây được thiết lập để chúng tôi chạy hệ thống này nhằm kiểm tra hoạt động của mô hình CNN-LSTM.

· Model – Thư mục này sẽ chứa tất cả các mô hình đã được đào tạo, được đào tạo lần đầu tiên. Đây sẽ là quy trình một lần để đào tạo mô hình.

· Description.txt – Đây là tệp sẽ chứa tên hình ảnh và chú thích liên quan sau khi xử lý sơ bộ.

· Feature.p – Tệp này liên kết hình ảnh và các chú thích liên quan được trích xuất từ Xception, đây là mô hình CNN được đào tạo trước.

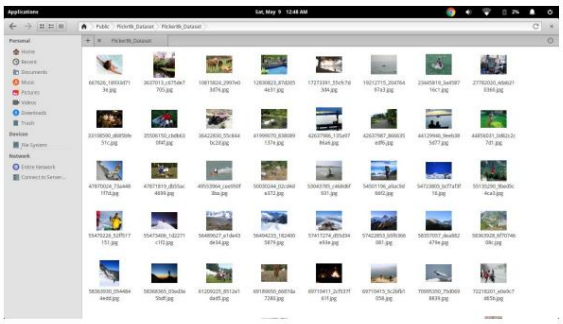
· Tokenizers.p – Tệp này chứa một biểu thức mà chúng tôi gọi là token được tổng , và những mã thông báo này là quát hóa với giá trị chỉ mục.

· Models.png – Biểu đồ biểu diễn phần mở rộng của mô hình CNN-LSTM.

· Testing_captions_generator.py – Đây là tệp Python được sử dụng để tạo chú thích cho hình ảnh.

· Training_captions_generator.ipynb – Về cơ bản đây là một Jupyter notebook, nói ngắn gọn là một ứng dụng dựa trên web. Chúng tôi sử dụng nó để đào tạo mô hình của mình và trên cơ sở đó tạo ra chú thích cho hình ảnh đầu vào của chúng tôi.

Hình 8 Cấu trúc tệp dự án



Hình 9 Flickr_Dataset

PHẦN KẾT LUẬN

Mô hình CNN-LSTM được xây dựng dựa trên ý tưởng tạo chú thích cho các hình ảnh đầu vào. Mô hình này có thể được sử dụng cho nhiều ứng dụng khác nhau. Trong bài này, chúng tôi đã nghiên cứu về mô hình CNN, mô hình RNN, mô hình LSTM và cuối cùng chúng tôi xác thực rằng mô hình đang tạo chú thích cho các hình ảnh đầu vào.

TÀI LIỆU THAM KHẢO

[1] Abhaya Agarwal và Alon Lavie. 2008. Meteor, m-bleu và m-ter: Các số liệu đánh giá cho mối tương quan cao với thứ hạng của con người về đầu ra dịch máy. Trong Biên bản báo cáo của Hội thảo thứ ba về dịch máy thống kê. Hiệp hội Ngôn ngữ học tính toán, 115–118.

[2] Ahmet Aker và Robert Gaizauskas. 2010. Tạo mô tả hình ảnh bằng cách sử dụng các mẫu quan hệ phụ thuộc. Trong Biên bản báo cáo của cuộc họp thường niên lần thứ 48 của Hiệp hội Ngôn ngữ học tính toán. Hiệp hội Ngôn ngữ học tính toán, 1250–1258.

[3] Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould. 2016. Spice: Đánh giá chú thích hình ảnh mệnh đề ngữ nghĩa. Tại Hội nghị Châu Âu về Thị giác Máy tính. Springer, 382–398.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. 2017. Sự chú ý từ dưới lên và từ trên xuống đối với chú thích hình ảnh và vqa. Bản in trước arXiv arXiv:1707.07998 (2017).

[5] Jyoti Aneja, Aditya Deshpande và Alexander G Schwing. 2018. Chú thích hình ảnh tích chập. Trong Biên bản Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu. 5561–5570.

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, và những người khác. 2016. Chú thích thành phần sâu: Mô tả các danh mục đối tượng mới mà không cần dữ liệu đào tạo ghép nối. Trong Biên bản báo cáo của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu.

[7] Dzmitry Bahdanau, Kyunghyun Cho và Yoshua Bengio. 2015. Dịch máy thần kinh bằng cách học chung để căn chỉnh và dịch. Trong Hội nghị quốc tế về biểu diễn học tập (ICLR).

[8] Shuang Bai và Shan An. 2018. Khảo sát về việc tạo chú thích hình ảnh tự động. Neurocomputing. Khảo sát máy tính của ACM, Tập. 0, Số 0, Điều 0. Ngày chấp nhận: Tháng 10 năm 2018. 0:30 Hossain và cộng sự.

[9] Satanjeev Banerjee và Alon Lavie. 2005. METEOR: Một phép đo tự động để đánh giá MT với mối tương quan được cải thiện với phán đoán của con người. Trong Biên bản hội thảo acl về các biện pháp đánh giá nội tại và bên ngoài cho máy bản dịch và/hoặc tóm tắt, Tập 29. 65–72.