International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 12 | Dec - 2021 ISSN: 2582-3930

Image Captioning Using Deep Learning

Harshul Jain[1], Daksh Banswal[2], Namita Goyal[3], Vandana Choudhary[4]

[1,2]B. Tech. Student, Information Technology & Maharaja Agrasen Institute of Technology
[3,4]Assistant Professor, Information Technology & Maharaja Agrasen Institute of Technology

---

Abstract - With increasing advancements in artificial intelligence, image captioning has become a popular topic. Image captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep learning domain. As it has a wide range of uses. It uses two topics of artificial intelligence which are natural language processing, and computer vision. This project aims to provide a brief explanation about the image in mostly one sentence and requires the methodology of computer vision and natural language processing. Recent development in deep learning and the availability of image caption datasets such as COCO and Flikr. Our model is trained using Flickr8k data using Convolution Neural Network (CNN) as an encoder and Recurrent Neural Network (RNN) as decoder show that our model is generating relevant captions for the image.

Key Words: Image Captioning, Deep Learning, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN)

1. INTRODUCTION

Image captioning is a process in which the input image is determined and a description is provided in the form of caption using natural language processing. Image captioning is not only the application of computer vision but also the application of the integration of computer vision and natural language processing. It is the primary implementation of contextual understanding of an image and text generation based on that understanding. This is the reason why image caption generation has been a very popular challenge for machine learning algorithms. There have been many different advances in the field of computer vision and natural language processing which has improved the accuracy of image caption generating models in general. This process proves to be a very important technology and has several applications. We must first understand how important this problem is to real world scenarios. Let's see few applications where a solution to this problem can be very useful.

Self driving cars — automatic driving is next generation advancement and use of challenging machine learning algorithms and if we can properly caption the scene around the car, it can give a boost to the self driving system.

Aid to the blind — we can create a product for the blind which will guide them travelling on the roads without the support of anyone else. We can do this by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning

CCTV cameras are everywhere today, but along with viewing the world, if we can also generate relevant captions, then we can raise alarms as soon as there is some suspicious activity going on somewhere. This could probably help reduce some crime and/or accidents.

Automatic Captioning can; make Google Image Search as good as Google Search, as then every image could be first converted into a caption and then search can be performed based on the caption.

2. Background

Machine Learning is the study of computer algorithms that refers to learning using algorithms that parse data, learn from that data and then apply it to predict an efficient future decision. In other words, machine learning is perfect example of intersection of statistics and computer Science The main focus of machine learning is all about computers being able to think and act with less human intervention. In recent years, the smart technologies are quickly becoming mainstays of life, the demand of machine learning experts and data scientists have grown exponentially (almost 75% over the past four years) and are poised to keep growing due to large data flow in the market and need to store and analysis data that help computer programs optimize their functionality from vast quantities of data to make important business decisions.

Deep Learning is a subset of machine learning and is related to algorithms for artificial neural networks inspired by the structure and working of a living brain. These neural networks consist of multiple layers which

extract features from the dataset fed to them and analyze these features by giving them certain weights. Traditionally, layers that appear later in the neural net model extracts and analyze higher or more complex features. For example, in image recognition models, lower layers would either check the edges or borders and size of the images, whereas the higher layers would try to identify the subjects, object borders, categories of the objects present in the image or identify the letters and characters spotted inside the image

Convolutional Neural Network (CNN)

In the past few decades, Deep Learning has proved to be a very powerful tool because of its ability to handle large amounts of data. One of the most popular deep neural networks is Convolutional Neural Networks. Neural networks are a subset of machine learning. Each node connects to another and has an associated weight and threshold.

CNN's were first developed and used around the 80s. It used to recognize handwritten digits. It was mostly used in the postal sectors to read zip codes, pin codes, etc. The main thing to train and also requires a lot of computing resources. This is one of the major drawbacks for CNNs at that period and hence CNNs were limited to the postal sectors and failed to enter the world of machine learning.

In deep learning, a convolutional neural network is a class of deep neural networks. Mostly used to analyze visual imagery. Now when we think of neural networks we think about matrix multiplications but that is not the case with ConvNet. It uses a special technique called Convolution. In mathematics, convolution is a mathematical operation on two functions that produces a third function that expresses how the shape of one is modified by the other.

Limitations

It just recognizes patterns and details that are so minute that they go unnoticed by the human eye. But when it comes to understanding the contents of an image it fails.

Recurrent Neural Network (RNN)

A recurrent neural network is a type of artificial neural network which uses sequential data or time-series data. These deep learning algorithms are commonly used for ordinal or temporal problems. Such as language translation, natural language processing, speech recognition, and image captioning; they are incorporated into popular applications such as Siri, voice search, and Google Translate. Recurrent neural networks utilize training data to learn. They are distinguished by their memory as they take information from prior inputs to influence the current input and output. The output of the recurrent neural network depends on the prior elements within the sequence.

Types of recurrent neural network

One to one, one to many, many to one, many to many
Common activation functions
An activation function determines whether a neuron should be activated. The nonlinear functions typically convert the output of a given neuron to a value between 0 and 1 or -1 and 1. Some of the most commonly used functions are
Sigmoid, Tanh, Relu.
Variant RNN architectures
Bidirectional recurrent neural networks (BRNN)
Long short term memory (LSTM)
Gated recurrent units (GRU)

Long short term memory (LSTM)

Learning to store information for some time intervals. LSTM can learn to bridge minimal time lags in excess of 1000 discretetiem steps by enforcing constant error flow through "constant error carrousels" within special units. Recurrent network can in principle ise their feedback connections to store representations of recent input events in form of activations. This is potentially significant for many applications, including speech processing, non-markovian control and music composition. To construct an architecture that allows for constant error flow through special, self connected units without the disadvantages of the naïve approach, we extend the constant error carrousel CEC embodied by the self connected linear unit.

At a high-level LSTM works very much like an RNN cell. Here is the internal functioning of the LSTM network. The LSTM consists of three parts, as shown in the image below and each part performs an individual function.

The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp.

These three parts of an LSTM cell are known as gates. The first part is called **Forget gate, the** second part is known as **the Input gate** and the last one is **the Output gate**.

3. Experiment

Dataset Used

In our project, we have used the Flickr8k image dataset with different sizes to train the model for understanding how to discover the relation between images and words for generating captions. It contains 8091 images in JPG format; each image has 5 different captions.
The images are bifurcated as follows in the code:
• Training Set — 6001 images
• Dev Set — 1001 images
• Test Set — 1001 images

There is also some text files related to the images. One of the files is "Flickr8k.token.txt" which has each image along with its 5 captions. Every line contains the <image name>#i <caption>, where 0≤i≤4 i.e. the name of the image, caption number (0 to 4) and the actual caption.

1000268201_693b08cb0e.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way.
1000268201_693b08cb0e.jpg#1 A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2 A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3 A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4 A little girl in a pink dress going into a wooden cabin .

Model generation

Merge-model architecture is used in this project to create an image caption generator. In this model, the encoded features of an image are used along with the encoded text data to generate the next word in the caption. In this approach, RNN is used only to encode text data and is not dependent on the features of the image. After the captions have been encoded, those features are then merged with the image vector in another multimodal layer which comes after the RNN encoding layer. This architecture model provides the advantage of feeding preprocessed text data to the model instead of raw data. The model development has the following blocks:

- Image feature extractor
- Text Preprocessor
- Output predictor
- Fitting the Model
- Caption Generation

a) Image Feature Extractor

The feature extractor needs an image 224x224x3 size. The model uses ResNet50 pretrained on ImageNet dataset where the features of the image are extracted just before the last layer of classification. Another dense layer is added and converted to get a vector of length 2048.

b) Text Preprocessor

To define the vocabulary, 8253 unique words are tokenized from the training dataset. As computers do not understand English words, we have represented them with numbers and mapped each word of the vocabulary with a unique index value and we encoded each word into a fixed sized vector and represented each word as a

number. Also, we maintain a list for each caption that stores the next word at each sub-iteration. Further, one hot encoding is applied on the list that contains the next word. Further, both partial sequence and one hot encoded next word are converted into arrays.

c) Output Predictor

In the training of the model, we first applied the Sequential model, for the images, which contains a Dense layer that uses 'relu' as the element-wise activation function. Then we added a Repeat vector layer with argument '40', which would repeat the input 40 times. Then we applied another sequential model, in which we used the Embedding layer as the first layer of the model.

Output vector from both the image feature extractor and the text processor are of same length (128) and a decoder merges both the vectors using an addition operation. This is then fed into two dense layers. The first layer is of length 128 and the second layer makes a prediction of the most probable next word in the caption. This layer uses softmax activation function to predict the most probable next word in the vocabulary.

d) Fitting the Model

After building the model, the model is fit using the training dataset. The model is made to run for 210 epochs and the best model is chosen among the 210 epochs by computing loss function on Flickr8k development dataset. The model with the lowest loss function is chosen for generating captions.

e) Caption Geneartion

To test our trained model, we input an image to the model. Next the image is fed into the feature extractor to recognize what all objects and scenes are depicted in the image, after resizing it. The process of caption generation is done using the RNN trained model. Then for that image, sequentially, word-by-word the caption is generated by selecting the word with maximum weight for the image at that particular iteration. The indexed word is converted to word and then appended into final caption. When tag is detected or the size of the caption reaches 40, the final caption is generated and printed along with the input image.

5) Performance measure

Loss value would decrease, and accuracy value would increase as we move from lower to higher number of epochs. Also, more the number of epochs, more would be the smoothness of these curves.

Through the testing phase of our implemented model, we found out that the model can generate sensible descriptions of images in valid English sentences. The generated captions are helpful enough to tell about the objects or elements in the image. Such Image Captioning can be helpful to visually impaired people, for image search and autonomous driving system etc. The loss and accuracy of the system has helped us achieve such good results.

The accuracy was found around 89%.

Some sample images from within and outside the dataset have been tested and we have got the following results.

6)Output

a young girl is standing on a sidewalk . endofseq

a black dog climbing out of the pool . endofseq

a little girl is sitting in a colorful toy car . endofseq

a man with a helmet is standing on a street with a bicycle . endofseq

a hiker in a red jacket is carrying rocks and the arms . endofse

a black and white dog is playing with a rope next to a wooden rope . endof

4) Conclusion and future scope

Our described model is based on a CNN that encodes an image into a compact representation, followed by an RNN that generates corresponding sentences based on the learned image features. It worked quite well when tested on several images. The captions it generated for the images were quite accurate. But the source of input image also played an important role in feature extraction and hence caption generation. Certain images are not well recognized and we found out that there is, still some scope of improvement. The experiment can be tried with other models as well.

REFERENCES

1. A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In ACL, 2010
2. deeplearning.ai Coursera
3.http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43274.pdf
4.http://hankyujang.com/Papers/ImageCaptioningUsingDeepLearning.pdf
5.https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/
6.Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, 2015
7. A Comprehensive Survey of Deep Learning for Image Captioning by Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga