IMAGE CAPTIONING USING CNN & LSTM

Akash Verma

Student, Department Of Information Technology , Ajay Kumar Garg Engineering College

Abstract--In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing and computer vision research. Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating descriptive sentences with proper and correct structure. In this study, the author has discussed a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The convolutional neural network compares the target image to a large dataset of training images, then generates an accurate description using the trained captions.

Keywords -- CNN , LSTM , MS-COCO , BLEU;

1. INTRODUCTION

Image Captioning is an Artificial Intelligence problem that refers to the process of generating textual description from an image. The task of image captioning is divided into two modules - one is an image based model which extracts the features and nuances out of our image and the other is the language based model which translates the features and objects given by our object based model to a natural sentence. However, it is expected that the sentence is meaningful, self-contained and grammatically and semantically correct. In other words, the caption shall describe the image correctly and not rely on any additional information. Image captioning has various applications such as for visually impaired persons , for social media ,usage in virtual assistants and several other applications. The Convolutional Neural Network is used to generate vocabulary describing the images and Long Term Short Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. I showcase the efficiency of my proposed model using the **MS-COCO 2014 dataset** and measure its performance using **BLEU** standard metric.

2. RELATED WORK

Ever since researchers started working on object recognition in images, it became clear that only giving the names of the objects recognized does not gives as good impression as a human described image. The image captioning problem and the given solutions have existed since the advent of the internet. Numerous algorithms have been put forward by various researchers from different perspectives. There have been many variations and combinations of different techniques since 2014. Krizhevsky had also implemented a neural network using non-saturating neurons, By employing a regularization model called dropout, they succeeded in reducing over fitting. Deng introduced a new database called Image Net. It was a collection of images that was built using the core of the word net structure. It organized the different classes of images in a densely populated semantic hierarchy. Yang proposed a system that automatically generated natural language description of an image. The model consisted of object detection and localization modules that is very similar to the human visual system that describes the image automatically. Pan experimented with the multiple network architectures on large datasets and proposed a model that showed improvement on captioning the accuracy over previously proposed models. Aneja proposed a CNN for machine translation and conditional image generation. Vinyals represented a model that consisted of deep recurrent architecture that focused on machine translation and computer vision . It used to generate the natural descriptions of an image by ensuring the highest probability of the generated sentence to accurately describe the image. Natural Language problems have already been investigated for more than five years now. Recent progress in Artificial Intelligence has greatly improved the performance of the models. However the results are still not satisfying. Machines cannot imitate human brains and the way communicate and so it remains an ongoing task, As long as the machines do not think , talk and behave like humans, natural language descriptions will remain a challenge to be solved. Due to the increasing amount of information on this topic , it is very difficult to keep on track with the researches and the results achieved in the image captioning field.

3. DATASET

I have used the **MS-COCO 2014 dataset** to train the model. The dataset contains over 82,000 images, each of which has at least 5 different caption annotations. The code below downloads and extracts the dataset automatically

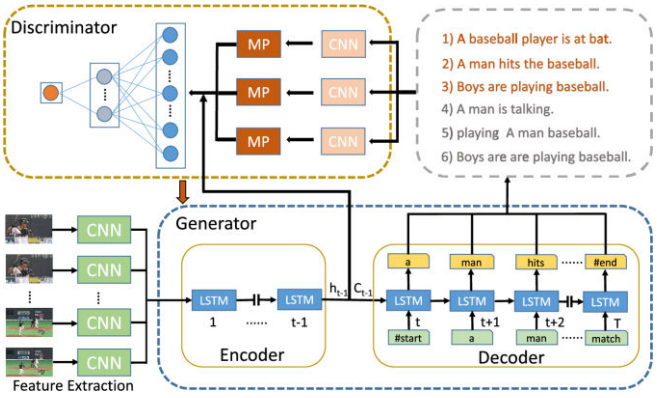
Train images
http://msvocds.blob.core.windows.net/coco2014/train2014.zip
Validation images
http://msvocds.blob.core.windows.net/coco2014/val2014.zip
Captions for both train and validation
http://msvocds.blob.core.windows.net/annotations-1-0-3/captions_train-val2014.zip
Takes 10 hours and 20 GB. I have downloaded necessary files for this.

4. EVALUATION METRICS

In order to evaluate the image-caption pairs, we need to evaluate their ability to associate previously unseen images and captions with each other. The evaluation of model that generates natural language sentence is done by the **BLEU (Bilingual Evaluation Understudy)** Score. It describes how natural sentence is compared to human generated sentence. It is widely used to evaluate performance of Machine translation. Sentences are compared based on modified n-gram precision method for generating BLEU score where precision is calculated using following equation

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{ngram' \in C'} Count(ngram')}$$

My model to caption images is built on recurrent and convolutional neural networks. A Convolutional Neural Network is used to extract the features from an image which is then along with the captions is fed into an Recurrent Neural Network. The architecture of the image captioning model is shown in figure 2.



The model consists of 3 phases:

A. Image Encoding

The features of the images from the MS-COCO 2014 dataset is extracted using the VGG 16 model due to the performance of the model in object identification. The VGG 16 is a convolutional neural network which consists of 16 layer which has a pattern of 2 convolution layers followed by 1 dropout layers until the fully connected layer at the end. The dropout layers are present to reduce overfitting the training dataset, as this model configuration learns very fast. These are processed by a Dense layer to produce a 4096 vector element representation of the photo and passed on to the LSTM layer.

B. Sequence Processor

The function of a sequence processor is for handling the text input by acting as a word embedding layer. The embedded layer consists of rules to extract the required features of the text and consists of a mask to ignore padded values. The network is then connected to a LSTM for the final phase of the image captioning.

C. Decoder

The final phase of the model combines the input from the Image extractor phase and the sequence processor phase using an additional operation, then fed to a 256 neuron layer and then to a final output Dense layer that produces a softmax prediction of the next word in the caption over the entire vocabulary which was formed from the text data that was processed in the sequence processor phase. The structure of the network to understand the flow of images and text is shown in the Figure 2.

5. TRAINING PHASE

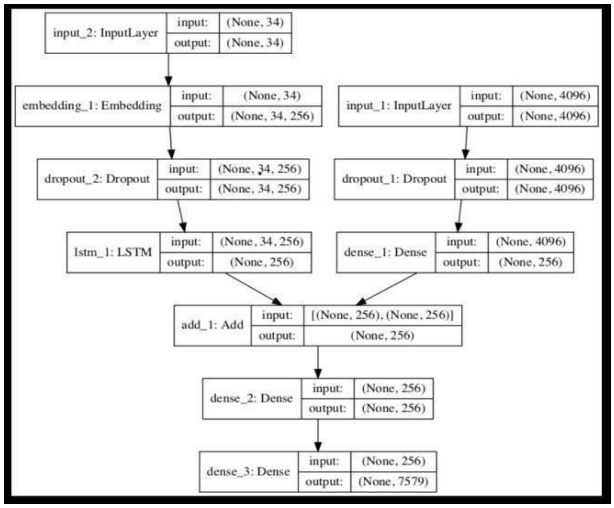
During training phase, I have provided a pair of input image and its appropriate captions to the image captioning model. The VGG model is trained to identify all possible objects in an image. While LSTM part of model is trained to predict every word in the sentence after it has seen image as well as all previous words. For each caption I have added two additional symbols to denote the starting and ending of the sequence. Whenever stop word is encountered it stops generating sentence and it marks end of string. Loss

function for model is calculated as, where I represents input image and S represents the generated caption. N is length of generated sentence. Pt and St represent probability and predicted word at the time t respectively. During the process of training I have tried to minimize this loss function.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

6. IMPLEMENTATION

The implementation of the model was done using the Python3, Google Colab environment. Keras 2.0 was used to implement the deep learning model because of the presence of the VGG net which was used for the object identification. Tensor Flow library is installed as a backend for the Keras framework for creating and training deep neural networks. Tensor Flow is a deep learning library developed by Google. It provides heterogeneous platform for execution of algorithms i.e. it can be run on low power devices like mobile as well as large scale distributed system containing thousands of GPUs. The neural network was trained on the Google Colab GPU. In order to define structure of the network, Tensor Flow uses graph definition. Once graph is defined, it can be executed on any supported devices. The photo features are pre-computed using the pre-trained model and saved. These features are then loaded into the model as the interpretation of a given photo in the dataset to reduce the redundancy of running each photo through the network every time we want to test a new language model configuration. The preloading of the image features is also done for real time implementation of the image captioning model.



7. EXAMPLES

The image captioning model was implemented and it was able to generate moderately comparable captions when compared to human generated captions. The VGG net model first assigns probabilities to all. The model then converts the image into word vector. This word vector is provided as input to LSTM cells which will then form sentence from this word vector.

Ex.1-
Generated sentence are “many car runs on the Road” , while actual human generated sentences “A Car Racing Match” or “F1 racing match” or “A game of car”. This results in a BLEU score of 57 for this image.

Ex.2-
Similarly in this example, generated sentence is “ A man wearing black shirt is standing in ice”, whereas the actual sentence is “ A man is drilling in ice’”.
While calculating BLEU score of all image in validation dataset we get average score of 60.1 , Which shows that the generated sentence is very similar compared to human generated sentence.

8. CONCLUSION AND FUTURE WORK

My end-to-end neural network system is capable of viewing an image and generating a reasonable description in English depending on the words in its dictionary generated on the basis of tokens in the captions of train images. The model has a convolutional neural network encoder and a LSTM decoder that helps in generation of sentences. The purpose of the model is to maximize the likelihood of the sentence given in the image.

Experimenting the model with MS-COCO 2014 dataset show decent results. I evaluated the accuracy of the model on the basis of BLEU score . The model was able to generate moderately comparable captions when compared to human generated captions. In Future more data can be collected and different evaluation metrics can be used to increase the accuracy of the model.

REFERENCES

1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E.Hinton, ImageNet Classification with Deep Convolutional Neural Networks.

2. Jia Deng, Wei Dong, Richard Socher Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A large-Scale Hierarchical Image Database.

3. Zhongliang Yang, Yu-Jin Zhang , Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization.

4. Jia-Yu Pan, Hyung-Jeong Yang , Pinar Duygulu, Automatic Image Captioninh, Conference: Multimedia and Expo ,2004. ICME ' 04. 2004 IEEE International Conference on,Volume:3. .