International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 7239-7242

Image Captioning - A Deep Learning Approach

Lakshminarasimhan Srinivasan[1], Dinesh Sreekanthan[2], Amutha A.L[3]

[1,2] Student, Computer Science and Engineering, SRM Institute of Science and Technology
[3]Assistant Professor (O.G), Computer Science and Engineering, SRM Institute of Science and Technology

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 description sentences with proper and correct structure. In this study, the authors propose 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. We showcase the efficiency of our proposed model using the Flickr8K and Flickr30K datasets and show that their model gives superior results compared with the state-of-the-art models utilising the Bleu metric. The Bleu metric is an algorithm for evaluating the performance of a machine translation system by grading the quality of text translated from one natural language to another. The performance of the proposed model is evaluated using standard evaluation matrices, which outperform previous benchmark models.

INTRODUCTION

Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Recently, deep learning methods have achieved state-ofthe-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Instead of requiring complex data preparation or a pipeline of specifically designed models, a single end-to-end model can be defined to predict a caption, given a photo. In order to evaluate our model, we measure its performance on the Flickr8K dataset using the BLEU standard metric. These results show that our proposed model performs better than standard models regarding image captioning in performance evaluation.

The limitations of neural networks are determined mostly by the amount of memory available on the GPUs used to train the network as well as the duration of training time it is allowed. Our network takes around seven days to train on GTX 1050 4GB and GTX 760 GPUs. According to our results, our results can be improved by utilising faster and larger GPUs and more exhaustive datasets.

RELATED WORK

The image captioning problem and its proposed solutions have existed since the advent of the Internet and its widespread adoption as a medium to share images. Numerous algorithms and techniques have been put forward by researchers from different perspectives. Krizhevsky et al. [1] implemented a neural network using non-saturating neurons and a very efficient a unique method GPU implementation of the convolution function. By employing a regularization method called dropout, they succeeded in reducing overfitting. Their neural network consisted of maxpooling layers and a final 1000-way softmax. Deng et al. [2] introduced a new database which they called ImageNet, an extensive collection of images built using the core of the WordNet structure. ImageNet organized the different classes of images in a densely populated semantic hierarchy. Karpathy and FeiFei [3] made use of datasets of images and their sentence descriptions to learn about the inner correspondences visual data and language. Their work described a Multimodal Recurrent Neural Network architecture that utilises the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images. Yang et al. [4] proposed a system for the automatic generation of a natural language description of an image, which will help immensely in furthering image understanding. The proposed multimodel neural network method, consisting of object detection and localization modules, is very similar to the human visual system which is able to learns how to describe the content of images automatically. In order to address the problem of LSTM units being complex and inherently sequential across time, Aneja et al. [5] proposed a convolutional network model for machine translation and conditional image generation. Pan et. al [6] experimented extensively with multiple network architectures on large datasets consisting of varying content styles, and proposed a unique model showing noteworthy improvement on captioning accuracy over the previously proposed models. Vinyals et al. [7] presented a generative model consisting of a deep recurrent architecture that leverages machine translation and computer vision, used to generate natural descriptions of an image by ensuring highest probability of the generated sentence to

accurately describe the target image. Xu et al. [8] introduced an attention based model that learned to describe the image regions automatically. The model was trained using standard backpropagation techniques by maximizing a variable lower bound. The model was able to automatically learn identify object boundaries while at the same time generate an accurate descriptive sentence.

DATASET AND EVALUATION METRICS

For task of image captioning there are several annotated images dataset are available. Most common of them are Pascal VOC dataset, Flickr 8K and MSCOCO Dataset. Flickr 8K Image captioning dataset [9] is used in the proposed model. Flickr 8K is a dataset consisting of 8,092 images from the Flickr.com website. This dataset contains collection of day-to-day activity with their related captions. First each object in image is labeled and after that description is added based on objects in an image. We split 8,000 images from this corpus into three disjoint sets. The training data (DTrain) has 6000 images whereas the development and test dataset consist of 1000 images each.

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 can be 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:

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

Our model to caption images are built on multimodal 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 1.

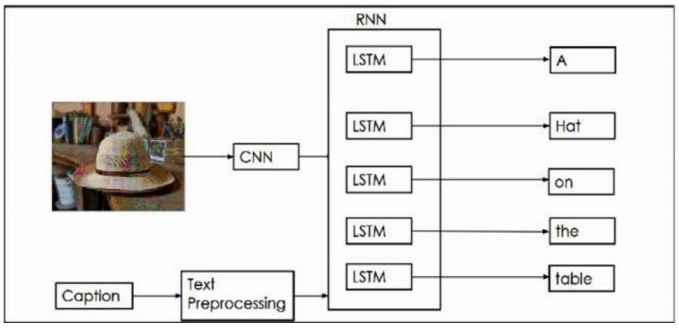

Figure 1. Architecture

The model consists of 3 phases:

A. Image Feature Extraction

The features of the images from the Flickr 8K dataset is extracted using the VGG 16 model due to the performance of the model in object identification. The VGG is a convolutional neural network which consists of 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.

TRAINING PHASE

During training phase we provide 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 we add 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 we have tried to minimize this loss function.

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

IMPLEMENTATION

The implementation of the model was done using the Python SciPy 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. Tensorflow library is installed as a backend for the Keras framework for creating and training deep neural networks. TensorFlow 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 Nvidia Geforce 1050 graphics processing unit which has 640 Cuda cores. In order to define structure of our network TensorFlow uses graph definition. Once graph is defined it can be executed on any supported devices. The photo features are pre-computed using the pretrained model and saved. These features are then loaded and them into our 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. The architecture of the model is shown in Figure 2.

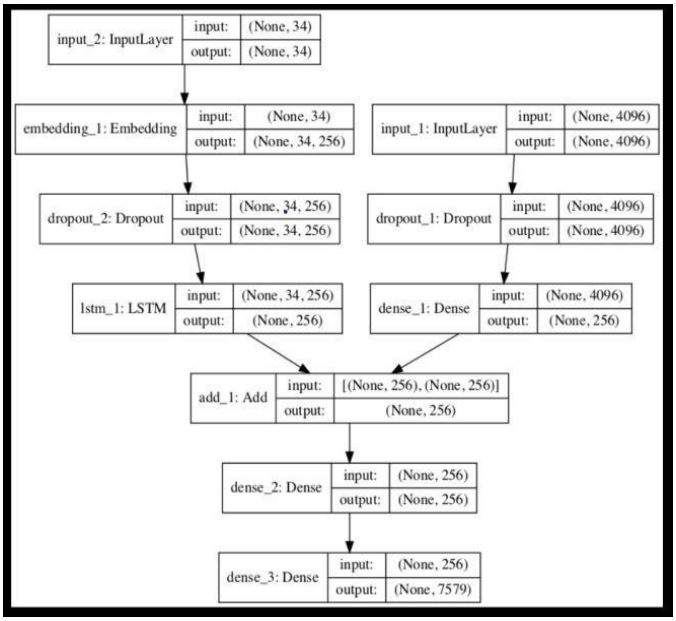

Figure 2. Image Captioning Model

RESULTS AND COMPARISION

The image captioning model was implemented and we were able to generate moderately comparable captions with compared to human generated captions. The VGG net model first assigns probabilities to all the objects that are possibly present in the image, as shown in Figure 3. The model 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. The generated sentences are shown in Fig 4. Generated sentence are black dog runs into the ocean next to a rock, while actual human generated sentences are black dog runs into the ocean next to a pile of seaweed., black dog runs into the ocean, a black dog runs into the ball, a black dog runs to a ball. This results in a BLEU score of 57 for this image.

Figure 3. Input Image

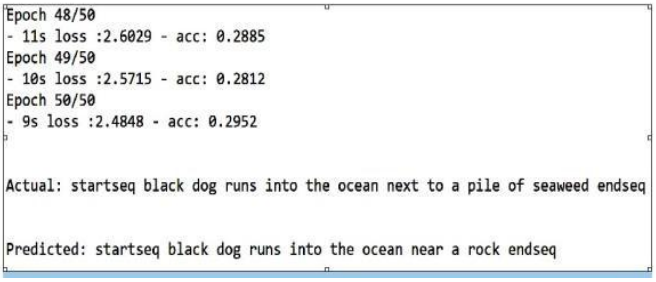

Figure 3. Output

Similarly in Fig. 5, 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 our generated sentence are very similar compared to human generated sentence.

Figure 5. Input Image

CONCLUSION

In this paper, the authors have implemented a deep learning approach for the captioning of images. The sequential API of Keras was used with Tensorflow as a backend to implement the deep learning architecture to achieve a effective BLEU score of

0.683 for our model. The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. In the future, the authors are working on alternating Pre-Trained Photo Models to improve the feature extraction of the model. Also, the authors are planning to improve achieve better performance by using word vectors on a much larger corpus of data such as news articles and other online sources of data. The configuration of the model was tuned, but other alternate configurations can be trained to see for improvement in the performance of the image captioning model.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, [Online] Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutional-neural-networks.pdf

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

[3] Andrej Karpathy, Li Fei-Fei, Deep VisualSemantic Alignments for Generating Image Descriptions, [Online] Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[4] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.pdf

[5] Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning, [Online] Available: https://arxiv.org/pdf/1711.09151.pdf

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

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator, [Online] Available: https://arxiv.org/pdf/1411.4555.pdf

[8] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, [Online] Available: https://arxiv.org/pdf/1502.03044.pdf [9] M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899

[9] BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA