International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Image Captioning Techniques

Ms. Divya [1], Tushar[2], Shivam Tyagi[3], Manan Arora[4], Rahul Singh[5]

[1] *Assistant Professor, CSE Department, HMRITM*
[2] *CSE Department, HMRITM*
[3] *CSE Department, HMRITM*
[4] *CSE Department, HMRITM*
[5] *CSE Department, HMRITM*

Abstract - In the past decade, the need of generating descriptive sentences automatically for images has generated a rising interest in natural language processing and computer vision research. Image captioning is an important task that requires semantic understanding of images and the ability to generate description sentences with correct structure. In this study, we propose a hybrid system that employs the use of multilayer Convolutional Neural Network (CNN) to generate a vocabulary which describes images and Long Short-Term Memory (LSTM) for accurately structuring meaningful sentences using the generated keywords. The convolutional neural network compares the target image with a large dataset of training images, after which it generates an accurate description using trained captions. We have showcased the efficiency of our proposed model using the Flickr8K and Flickr30K datasets and we can show that their model gives superior results when compared to the state-of-the-art models utilizing the Bleu metric. The Bleu metric is an algorithm that evaluates the performance of a machine translation system by grading the quality of text translated from one natural language to another. The performance of this proposed model has been evaluated using standard evaluation matrices, which are found to be outperforming previous benchmark models.

Key Words: image captioning, machine learning, Flickr8K, Keras, TensorFlow, nltk

1.INTRODUCTION

Caption generation is a very interesting AI (Artificial Intelligence) problem where a descriptive sentence is generated for a given image. It involves dual techniques ranging from computer vision to understand the content of the given image and a language model from the field of (NLP) natural language processing to turn the meaning of the image into words in an order that is meaningful. Image captioning has numerous applications like recommendations in editing softwares, 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-of-the-art results on examples of the above requirements. It's been demonstrated that the deep learning models are capable of achieving optimum results in caption generation problems. A single end-to-end model is able to predict a caption for an image without requiring complex data preparation or a pipeline of specifically designed models, For the evaluation of our model, we measure its performance on the Flickr8K dataset using the BLEU standard metric. The results show that our model is able to perform better than alternate standard models in regards with the image captioning in performance evaluation.

The limitation of neural networks is determined mostly by the amount of memory available on the GPU used to train the network as well as on the duration of training time provided. Our network takes around Five days to train on a GTX 1060 6GB and AMD RX570 GPUs. Our results, concludes that our results can be improved by making use of much faster and larger GPUs and more exhaustive datasets.

2. Related Works

Image captioning's solutions have already existed since the starting of the Internet and its adoption as a medium for sharing images. Various algorithms and techniques have been put forward by researchers. Krizhevsky et al. [1] implemented a neural network using non-saturating neurons and a very efficient and unique GPU implementation method of the convolution function. By utilizing a regularization method called dropout, they succeeded in reducing overfitting. Their neural network consisted of max pooling layers and a final 1000-way softmax. Deng et al. [2] introduced a new database in which they called ImageNet, which was an extensive collection of images built using the core of the WordNet structure. ImageNet organized the numerous classes of images in a very densely populated semantic hierarchy. Karpathy and FeiFei [3] used datasets of various images and their sentence descriptions for learning about the inner correspondences of visual data and language. Their work described a Multimodal Recurrent Neural Network architecture that utilised the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images. Yang et al. [4]

proposes a system that automatically generates a natural language description of an image, which can help greatly in improving the image understanding. The proposed multi model neural network method, that consists of object detection and localization modules, is very similar to the human visual system which is able to learn how to describe the content of images automatically. 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 very large datasets that consisted of varying content styles, and they proposed a unique model that showed noteworthy improvement in captioning accuracy over the previously proposed models. Vinyals et al. [7] presented a generative model that consisted of a deep recurrent architecture that leveraged machine translation and computer vision, used for generating natural descriptions of an image by ensuring highest probability of the generated sentence for accurately describing the target image. Xu et al. [8] introduced an attention-based model that learns to describe the image regions automatically. This model is trained using the standard backpropagation techniques by maximizing a variable lower bound. The model is able to automatically learn how to identify object boundaries while at the same time generating an accurately descriptive sentence.

3. Dataset and Evaluation metrics

For the task of image captioning there are numerous annotated images dataset available. Most common among them are the Pascal VOC dataset, Flickr 8K and MSCOCO Dataset. The Flickr 8K Image captioning dataset [9] is used in our proposed model. Flickr 8K is a dataset that consists of 8,092 images from the Flickr.com site. This dataset contains a collection of day-to-day activities with their related captions. First each object in an image is labelled and then a description is added based on objects in the image.

We have split 8,000 images from this dataset into three disjoint sets. Our training data (DTrain) has 6000 images while the development and test dataset consist of 1000 images each.

For evaluating the image-caption pairs, we have to evaluate their ability to associate previously unseen images and captions with each other. The evaluation of the model that generates natural language sentence can be done by the BLEU (Bilingual Evaluation Understudy) Score. This score describes how natural the sentence is compared to a human generated sentence. It is widely used to evaluate performance of Machine translation. Sentences are compared based on a 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 for captioning images is built on multimodal recurrent and convolutional neural networks. A Convolutional Neural Network is used for extracting the features from an image which is then along with the captions fed into a Recurrent Neural Network. The architecture of the image captioning model is shown in figure 1.

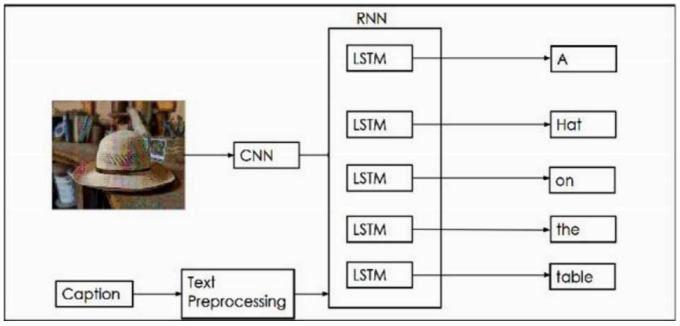


Figure 1. Architecture

The model consists of 3 phases:

A. Image Feature Extraction

The features of the images from the Flickr 8K dataset are extracted using the VGG 16 model due to the performance of the model in object identification. The VGG is a convolutional neural network that consists of 16 layer that have a pattern of 2 convolution layers followed by 1 dropout layer until the fully connected layer at the end. The dropout layers are present to reduce overfitting the training dataset, due to this model configuration learning very fast. These are then processed by a Dense layer to produce a 4096-vector element representation of the photo and is passed on to the LSTM layer.

B. Sequence processor

The function of a sequence processor is for the handling of the text input by acting as word embedding layer. The embedded layer consists of rules for extracting the required features of the text and it 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 this model combines the input from the Image extractor phase and the sequence processor phase using an additional operation which is then fed to a 256-neuron layer and then to a final output Dense layer which produces a softmax prediction of the next word in the caption over the entire vocabulary which was formed from the text data which was processed in the sequence processor phase. The structure of the network is to understand the flow of images and text is shown in the Figure 2.

TRAINING PHASE

During training phase, we provide a pair of input images and its appropriate captions to the image captioning model. The VGG model has been trained to identify all possible objects in a given image. While the LSTM part of the model is trained to predict each word in the sentence after it has seen the image as well as all the previous words. For every caption we also add two additional symbols that denote the starting and the end of the sequence. Whenever the stop word is encountered it stops generating sentence and marks it the end of string. Loss function for the model is calculated as given below, where I represent input image and S represents the generated caption. N is the length of the generated sentence. pt and St represents the probability and the predicted word at the time t respectively. During the process of training, we have to try to minimize this loss function.

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

4. IMPLEMENTATION

The implementation of this model was done by Python SciPy environment. Keras 2.0 is used to perform the deep learning model because of the bearing of the VGG net which is used for object identification. Tensorflow library is being installed as a backend for the Keras framework for creating and training deep neural networks. TensorFlow is a deep learning library produced by Google. It is providing a heterogeneous platform for the execution of algorithms i.e. it can also be run on low power devices like mobile as well as large-scale shared systems containing thousands of GPUs. The neural network is trained on the Nvidia Geforce 1050 graphics processing unit which has 640 Cuda cores. To explain the structure of our network TensorFlow is using graph definition. Once a graph is defined it can be performed on any supported devices. The photos that featured are pre-computed using the pre-trained model and is saved. These features are then being loaded into our model as the representation of a given photo in the dataset to reduce the redundancy of running each photo through a network each time we went to test a different language model configuration. The preloading of the image features is also done during the real-time implementation of the image captioning model. The architecture of the model is as shown in Figure 2.

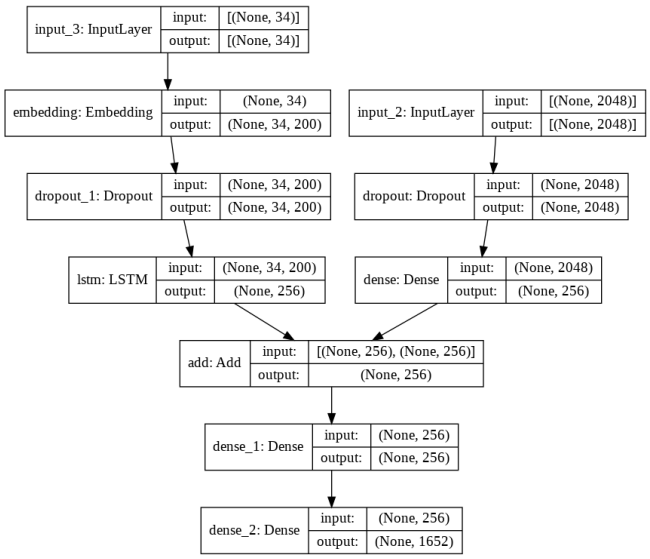
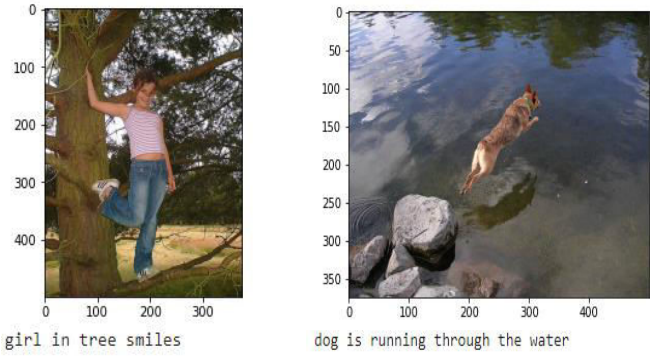


Figure 2. Image Captioning Model

5. CONCLUSIONS

The image captioning model is then implemented and we are able to produce moderately comparable captions compared to human generated captions. The VGG net model is first assigning probabilities to all the objects that are possibly presents in the image, as shown in Figure 3. Then model converts the image into a word vector. This word vector is then provided as input for LSTM cells which will be forming a sentence from this word vector. The generated sentences are shown in Figures below.

Input Images



Generated Images with Captioning

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-convolutionalneural-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.p df

[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