Generate detailed captions of an image using Deep Learning

Khan Shayaan Shakeel
Computer Engineering
M. H. Saboo. Siddik COE
Mumbai, India
shayaankhan054@gmail.com

Masalawala Murtaza Shabbir
Computer Engineering
M. H. Saboo. Siddik COE
Mumbai, India
masalawala708@gmail.com

Qazi Faizan Ahmed
Computer Engineering
M. H. Saboo. Siddik COE
Mumbai, India
faizanqazi487@gmail.com

Abstract

This paper show the approach of automatically generating captions for the image. The paper shows how computer classifies different objects present in images and later combine them to produce captions. First our model detects all the objects and identify objects as words, then it generates the sequence from the detected words and later it re-rank the sentences to generate the final caption. The model is based on convolutional neural network and long short term memory network. The model is trained on Flickr8K dataset. Once the caption are generated the description is converted into speech with the help of either pyttsx3 or gtts package.

I. INTRODUCTION

For humans the process of describing an image is a very simple process but for computers we achieve this with the help of computer vision. The main goal of computer is to understand the scenario in an image. Not only it should understand the image but also it should be able to express it in human language. Image captioning is a process where the system must be capable enough to distinguish between the different objects and then later express it in terms of language which is understood by the humans. We create a system that links the objects in the image and creates a logical sequence. This logical sequence of description comes with the help of learning the data. With the help of dataset which consists of images with descriptions help us to train our model and predict the results.

"man in black shirt is playing guitar."

In the above diagram we can see that the caption generated is very accurate. The general idea is to divide the system into logically 2 modules where the first module is Image based model and the other is Language based model.
Image based model is build with the help of Convolutional Neural Network (CNN). This model is used to extract the features from the image. The CNN identifies different segments of an image and then assign weights to it which helps in the classification of the image. CNN is found to be very useful in image classification. But our main goal is to extract the features. CNN is generally used in layers where the output of the first layer is fed as the input to the second layer and so on. After a series of layers we get the vectorial representation of image which is fed as an input to the language base model.

Language base model is build with the help of Long Short Term Memory Network (LSTM). LSTM is a type of Recurrent Neural Network. LSTM is generally is used in sequence prediction problems. RNN can also be used for sequence prediction but the limitation with RNN is short term memory. As a result LSTM is found to be more efficient for predicting sequence.

Once the description is generated now the innovation part comes where we try to convert the generated caption into speech. For text to speech conversion there are many packages available but for our model we have used the pyttsx3 python package. Instead of pyttsx3 we can also use another python package viz. gtts. The main idea behind text to speech conversion is to create a social impact on visually impaired people. With the help of this project visually impaired people will come to know about what image consists of.

II. RELATED WORK

This section gives detailed information about the research work that has been done on Image Caption Generation. Recently the quality of image caption generation has improved considerably by using combinations of CNN to obtain vectorial representation of images and RNN to decode those representations into natural language sentences.

Yao at el have published a research paper that explains the process of Image and Video to Text conversion [11]. The entire process is based on Image Comprehension. The process is divided into three steps. In the first step visual features are extracted. In the second step the output of the first stage is given as input to second stage which converts it into textual description. In the final stage the description is transformed into semantically meaningful, human understandable captions. Users can not only obtain captions for images but for videos as well.

Li et al have published a paper that incorporates storytelling for videos [14]. The main aim is to produce coherent and concise stories for long videos. With the help of the Multimodal Embedding Research, they have designed a Residual Bidirectional RNN to use past and future

contextual knowledge. Multimodal embedding is also used for video clip phrases.

O. Vinayals et al have developed a model known as NIC which is a end-to-end neural network model that automatically generates caption for the input image [4]. The entire model is dependent on CNN which is used for features extraction and then later it is trained by a RNN to generate sentences. This system has proved to be producing accurate results for larger datasets. The model quantitative evaluations is done either by using BLEU or ranking metrics to assess the generated descriptions.

S. Shabir, S. Arafat et al have published that since there are many research is going on to find new ways for generating captions, they have given detailed overview over technical aspects and techniques of image captioning. The research paper is all about the most common process for image captioning to new ways that have been discovered. The research paper also talks about the all related points in detail. The paper has even proposed the fields where the potential efforts should be made in order to improve the results.

Hao Fang et al have published a system that divides the process of image caption generation into three major steps[10]. First the system reasons with the image sub-regions rather than the entire image. Next with the help of the CNN the features from the sub-regions are extracted and then fine-tuned on the training data. The training is done at Maximum Entropy (ME) from training data set descriptions. This training results in capturing of commonsense knowledge about the image through language statistics. The final stage is re-ranking of a set of high-likelihood sentences by a linear weighting. These weights are assigned on the basis of Minimum Error Rate Training (MERT). In addition to this they have used Deep Multimodal Similarity Model (DMSM) that maps the similarity between text and image. This in turn improves the selection of quality captions.

Kelvin Xu el at have proposed a system with two approaches[15]. The first one is soft deterministic attention mechanism that is trained on the basis of standard back-propagation methods and the second one is hard deterministic attention mechanism which is trained by maximizing an approximate variational lower bound or by REINFORCE. The paper showcase the how we can gain insights and interprets the results. It visualize where and what the attention is focused on in an image. The paper also show the usefulness of the caption generated by evaluating it against state of art performance.

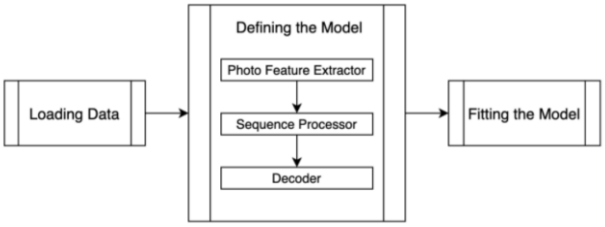
III. METHODOLOGY

System Design:

The entire module can be logically divided into two modules:

1. Image Based Model
2. Language Based model

In the Image Based Model the input image is converted into vectorial representations. For image based model convolutional neural networks are used in combinations. Image Based Model is also known as Encoder.
In the Language Based model the vectorial representations are converted into natural language. The vectorial representations are decoded with the help of the LSTM network. Language Based Model is also known as Decoder.



Both the model are integrated together to complete the entire process of image to text generation and later with the help of pyttsx3 or gtts we can convert the generated caption to speech which will help the visually impaired people.
The process is divided into following logical steps:

Step1: Firstly after pre-processing the data the data is loaded into the model for training

Step2: This step consists of three more steps:

Photo Feature Extractor: This module will extract the features from the image by using different combinations of convolutional neural networks.
Sequence Processor: The output from the Photo Feature Extractor is fed to the sequence processor. It uses Long Short Term Memory Network (LSTM) for managing text.
Decoder: It produces the most logically correct sequence by combining sequence processor and photo feature extractor.

Data Collection:

The dataset used for training and testing model is flickr_8k dataset. The dataset consist of two directories:

Flickr8k_dataset: It consists of 8092 photographs in JPEG format.
Flickr8k_text: It consists of number of files having descriptions for the image.

The dataset is divided into 6000 images for training, 1000 images for validation and 1000 images for testing.

Developing the model:

The entire model is divided into 2 modules:

1. Image Based Model – Convolutional Neural Network.
2. Language Based Model – Long Short Term Memory.

Convolutional Neural Network (CNN):

A Convolutional layer is also known as CNN or CoonvNet. It consist of three layers viz. convolutional layer, pooling layer, fully connected layer.

Convolutional layer:

All the load of computational work is handled by the convolutional layer. This layer performs dot product between two matrices, where one matrix represents the kernel i.e. learnable parameters and the other matrix represents restricted portion of the receptive field. The kernel is smaller thatn an image but is more in depth. The Kernel slides across the height and width of the image producing a two dimensional representation of the image.

Pooling Layer:

The pooling layer is used to derive summary statistics of nearby outputs at certain locations. This results in reducing the spatial size of representation which in turn reduces the computation and weights. Every slice of the representation has its own pooling operation.
Several pooling opeartons are there such as the average of the rectangular neighbourhood, L2 norm of the neighbourhood, weighted average based on the distance from the central pixel. But the most commonly used is the Max pooling, where the maximumof the neighbourhood is taken in consideration.

Fully Connected Layer:

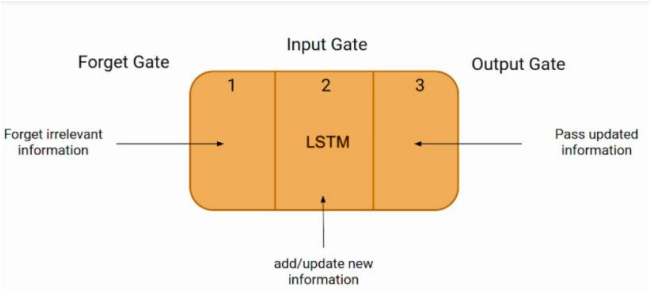
This layer helps to map the input and output of all the layers. The Neurons in this layer are fully connected with the all the neurons in the preceding and succeeding layer.

CNN is a Deep Learning Algorithm. CNN uses the concept of weights for image classification. CNN assign weights to different objects present in the image which helps in the classification of image. For vectorial representation of image layers of CNN are used together. The output of first layer is fed as input to the second layer and this process continues for all the subsequent layers.
After a series of convolutional network it is necessary to connect a fully connected layer. The output from the CNN layers are fed to the fully connected layers which result in a N dimensional vector which is in the encoded form.

Long Short Term Memory (LSTM):

LSTM is a type of Recurrent Neural network. RNN and LSTM are generally used for predicting orders. The idea behind using LSTM is that when we go into deep neural networks, if the gradients are very low or zero then training cannot take place which leads to poor prediction performance.

Long Short Term Memory is an advanced RNN algorithm which overcomes the limitations of traditional RNN. RNN remembers the past information and uses it for current operation but due to short term memory also known as vanishing gradient it cannot remember long term dependencies. LSTM overcomes the limitations of the traditional RNN and proves to be more efficient in long term sequence prediction.



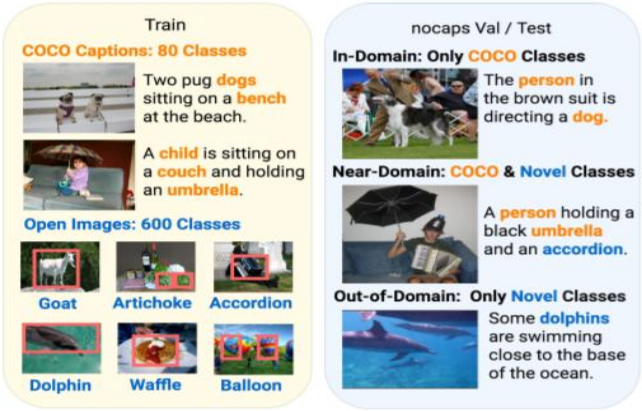
LSTM consists of three parts. The first parts tell whether the information coming from the previous timestamp is relevant or not. If it is found to be irrelevant the information is discarded. The second part tries to learn new information from the input provided to it. And finally the third part helps in updating the information from current timestamp to next timestamp. Depending upon the functionality of all the three parts they are known as the forget gate, the Input gate and the Output gate respectively.

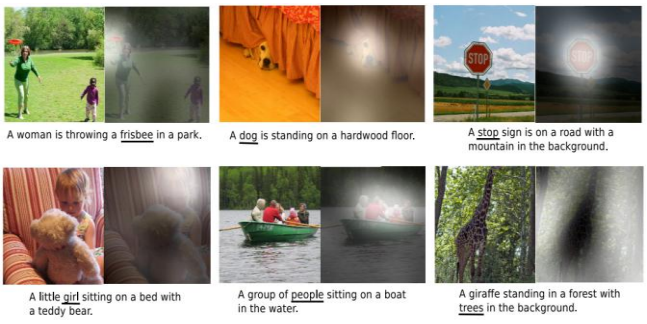
The major steps include:

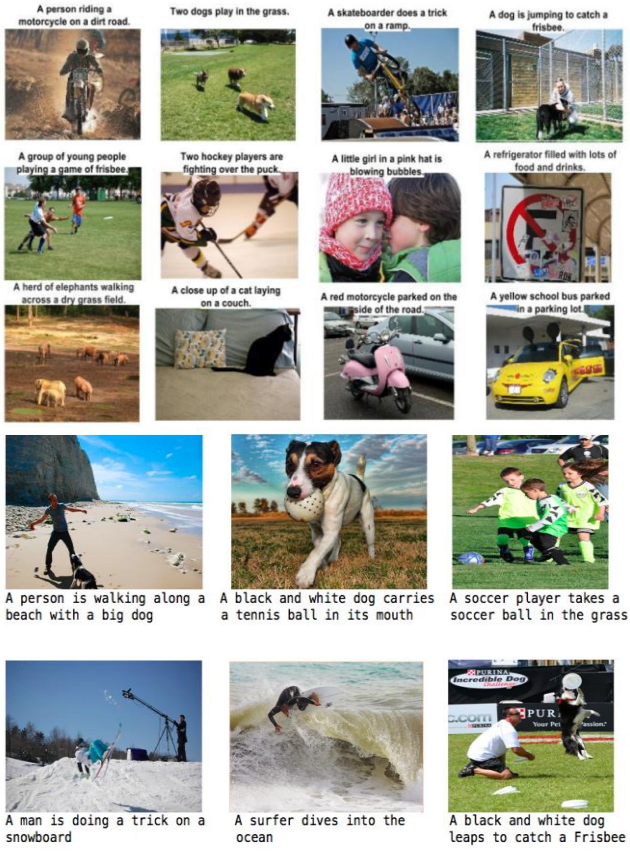
- Pre-processing the image: We use VGG16 model for extracting the features. VGG16 is called Visual Geometry Group. It comes preinstalled with the Keras library. We exclude the last layer of the classification model because we are interested in the features.
- Creating Vocabulary for the image: Machines cannot handle raw text. First the cleaning of the text is important which is done by splitting it into words, handling punctuations and removing words with numbers. Each unique word is mapped to a unique index value which could be understood by the machines.
- Training the model; The flickr_8k dataset consists of 6000 images in jpeg format for training.
- Tokenizing Vocabulary: The process of mapping the words with the unique index value is done by the Tokenizer class that comes with the Keras library.
- Data Generator: For this supervised model 6000 input images are provided and eacg image has 4096 lengh feature vector. This large data cannot be stored in memory so we use generator that yields batches.
- CNN-LSTM Model: With the help of CNN-LSTM we generate the captions.
- Testing the Model: After the model is trained we test the model against random images and evaluates the generated captions.

`

Examples:







IV. Future Scope

The model is currently trained with flickr_8k dataset. In future the CNN-LSTM model can be trained against the dataset containing much larger volume of images like 1000000 images which will improve the overall accuracy of the model. Instead of LSTM we can use another RNN algorithm known as Long Term Recurrent Convolutional Neural Network. LRCN combines a deep hierarchical visual feature extractor (such as NN) with a model that can learn to recognize temporal dynamics for task involving sequential data, linguistic, etc.

V. Conclusion

We have proposed a system that will generate logical captions for an image. The model can also be tested for its evaluation against BLEU and METEOR metric. We have developed a system that will be able to mimic human like behaviour for describing the image. In addition to that our model uses a very few hard coded assumptions. We hope that our research will encourage and help students for future work.

References

[1] V. Julakanti, "Image Caption Generator using CNN-LSTM Deep Neural Network", *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no., pp. 2968-2974, 2021.

[2] S.-H. Han and H.-J. Choi, "Domain-Specific Image Caption Generator with Semantic Ontology", *IEEE International Conference on Big Data and Smart Computing (BigComp), 2020.*

[3] M. Wang, L. Song, X. Yang, and C. Luo, " A parallel-fusion RNN-LSTM architecture for image caption generation ", *IEEE International Conference on Image Processing (ICIP), 2016.*

[4] S. Amirian, K. Rasheed, T. Taha and H. Arabnia, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap", *IEEE Access*, vol. 8, pp. 218386-218400, 2020.

[5] S. Shukla, S. Dubey, A. Pandey, V. Mishra, M. Awasthi and V. Bhardwaj, "Image Caption Generator Using Neural Networks", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 01-07, 2021.

[6] M. Panicker, V. Upadhayay, G. Sethi and V. Mathur, "Image Caption Generator", *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, no. 3, pp. 87-92, 2021.

[7] J. Karan Garg and Kavita Saxena, "Image to Caption Generator", *International Journal for Modern Trends in Science and Technology*, vol. 6, no. 12, pp. 181-185, 2020.

[8] F. Fang, H. Wang, and P. Tang, "Image Captioning with Word Level Attention", *25th IEEE International Conference on Image Processing (ICIP), 2018.*

[9] Y. Huang, J. Chen, W. Ouyang, W. Wan and Y. Xue, "Image Captioning With End-to-End Attribute Detection and Subsequent Attributes Prediction", *IEEE Transactions on Image Processing*, vol. 29, pp. 4013-4026, 2020.

[10] P. Mathur, A. Gill, A. Yadav, A. Mishra, and N. K. Bansode, " Camera2Caption: A real-time image caption generator ", *International Conference on Computational Intelligence in Data Science (ICCIDS), 2017*.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015*.

[12] Y. Zhou, Y. Sun, and V. Honavar, " Improving Image Captioning by Leveraging Knowledge Graphs", *IEEE Winter Conference on Applications of Computer Vision (WACV), 2019.*

[13] Y. Zhenyu and Z. Jiao, "Research on Image Caption Method Based on Mixed Image Features", *IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2019*.

[14] M. Tanti, A. Gatt, and K. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?", *Proceedings of the 10th International Conference on Natural Language Generation, 2017*.

[15] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares, "Image Captioning: Transforming Objects into Words.", *Proceedings of the 10th International Conference on Natural Language Generation, 2017*.

[16] A. Deshpande, J. Aneja, L. Wang, A.G. Schwing, D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.*

[17] R. Subash, "Automatic Image Captioning Using Convolution Neural Networks and LSTM", *Journal of Physics Conference, 2019.*

[18] B.Krishnakumar,K.Kousalya, S.Gokul,R.Karthikeyan, D.Kaviyarasu, "Image Caption Generation Using Deep Learning", *International Journal of Advanced Science and Technology, 2020.*

`