Caption Generator

Thenmozhi S
Associate Professor
Department of Computer Applications
PES University
Bangalore, India
thenmozhis@pes.edu

Vijaykumar R Pai
Department of Computer Applications
PES University
Bangalore, India
vijaykumarrpai@gmail.com

Abstract— **Humans have the ability to see visuals and comprehend the information associated with the visuals. The human brain automatically does this process. Can computers mimic the same? This question gives rise to this project "Caption Generator". Caption Generator is a machine learning application that identifies the action portrayed in the given image. The objective is to generate a caption that well describes the image. The machine has to be artificially trained to identify the captions as a meaning description of the given image. The application has to take the image as input and recognize the context of the image and describe them in a natural language like English. Suitable deep learning and artificial intelligence is used to achieve the objective.**

Keywords—Caption Generator, Convolutional Neural Network, Recurrent Neural Network, Long Short Term Memory (LSTM), VGG16 Model

I. Introduction

Humans can see an image and can tell what the image is about, but a machine cannot tell what the image is about. The process of describing the context of an image using simple English sentences is a very important task. Moreover, it could be of great assistance for visually impaired people as it would help them understand the context of the image available on the web or locally saved in their system. Also, it is very helpful in providing information about the images depicted in the scenarios such as images captured in smartphones. Moreover, a textual description of the images is not sufficient, but an image must also express how the objects are related to each other in terms of their attributes and activities involved in it.

Human beings usually describe a scene using natural languages that are concise and compact. However, computers describe the scene by taking an image that is a 2-D array. The objective is to map the images and captions to the same space and learning a mapping from the image to the sentences. The RNN method not only models the one-to-many (words) image captioning but also models many-to-one action generation and many-to-many image descriptions. The model has three components. The first component is a CNN that is used to understand the content of the image. The understanding of image answers the typical questions in Computer Vision such as "What are the objects?", "Where are the objects?" and "The features in the image?" For example, given an image of "Girl sitting on the grass-covered in paint," the CNN has to recognize the "Girl", "grass" and their relative locations in the image. The second component is an RNN that is used to generate a caption given the visual feature. For example, the RNN has to generate a sequence of probabilities of words given two words "Girl, grass". The third component is used to generate a caption by exploring the combination of the probabilities.

One of the biggest questions is how to make a computer understand the image is about and what the image is representing. Also, it would be a greater help to visually impaired people to help them better understand the content of the images on the web or locally saved in their system. It is very useful in providing more accurate and compact information of images.

Nowadays, the impact of social media is much more in anyone's life. Use of Caption Generator may be helpful in many situations such as providing assistance to help visually impaired people as it would help them understand the context of the images, Auto-subtitling.

The application work for captured images. It can even work for images that do not have any objects. In such cases, the application would not generate any caption depicting that the image is not having any object or features in it to predict the actions. The application is limited to accept one image at a time.

Machine learning can be defined as a computer system that is used to perform a specific task without using explicit instructions with the empirical study of algorithms and statistical models. With the improvement in Machine Learning techniques and accessibility of huge datasets and computation power, it is possible to create models that will generate captions for an image. These algorithms are used for building a model with the help of training data and then used to achieve the objective by testing the model on the testing data. The system is made to learn with the different images using CNN and LSTM algorithms. The model is to be tested for its accuracy with BLEU Score. The basic working of the application is that objects are identified from the images and features are extracted from it using pre-trained VGG16 model (CNN) and then fed to the LSTM model along with the captions to train. The trained model is then capable of generating captions for any images that are fed to it.

II. Related Work

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, (Journal of Artificial Intelligence Research submitted on 15th Jan 2016) have discussed about their project as a challenging problem as the model was not working properly with natural images that have recently received a huge amount of attraction from the computer vision and natural language processing communities. Also, they have classified the existing approaches based on how they conceptualized this problem. They have helped in reviewing the detailed description of existing models along with their advantages and disadvantages.

Jiuxiang Gu, Gang Wang, Jianfei Cai, Tsuhan (2017 IEEE International Conference on Computer Vision) explained about the effectiveness of their approach is validated on two datasets: Flickr30K and MS COCO. The extensive experimental results show that their method outperforms the vanilla recurrent neural network-based language models and is competitive with state-of-the-art methods. With 30000 images, the author was able to get 76 % accuracy.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga, (Journal on Deep Learning for Image Captioning submitted on 14th October 2018) said that although deep learning-based image captioning methods have achieved remarkable progress in recent years, a robust image captioning method that can generate high-quality captions for all images is yet to be achieved.

Kenneth P. Camilleri Marc Tanti, Albert Gatt, (Journal on Computer Vision and Pattern Recognition submitted on 7th August 2017) said that a recurrent neural network (RNN) is typically viewed as the primary ‘generation’ component. The authors suggest that the image features should be ‘injected’ into the RNN. They have viewed the RNN algorithm as only encoding the previously generated words. According to the authors, RNN algorithm should only be used to encode linguistic features and that only the final representation should be ‘merged’ with the image features at a later stage. Two architectures are being compared in this journal. As suggested RNNs are better viewed as encoders, rather than generators.

III. Theoretical framework

A. *Object Identification*

This is done by converting the image from RGB format to Greyscale to find the correlation between X coordinates. If there is a correlation between two X coordinates, then it means the object is the same. An image is converted to Greyscale format by converting the image in 2-D array. Having images in Greyscale format makes it easier to process without losing features.

B. *Feature Extraction From Images*

This is done by converting the image from RGB format to Greyscale to find the correlation between X coordinates. If there is a correlation between two X coordinates, then it means the object is the same. An image is converted to Greyscale format by converting the image in 2-D array. Having images in Greyscale format makes it easier to process without losing features.

C. *Caption Generation*

The extracted features dumped in pickle file are then passed through the VGG16 model along with a pre-processed bag of words for model generation. Simultaneously the CNN model is passed through the LSTM model for predicting the caption of the images and displaying the same.

IV. Design and modeling

A. *Overview*

The overall design of the system is depicted in the figure 1.

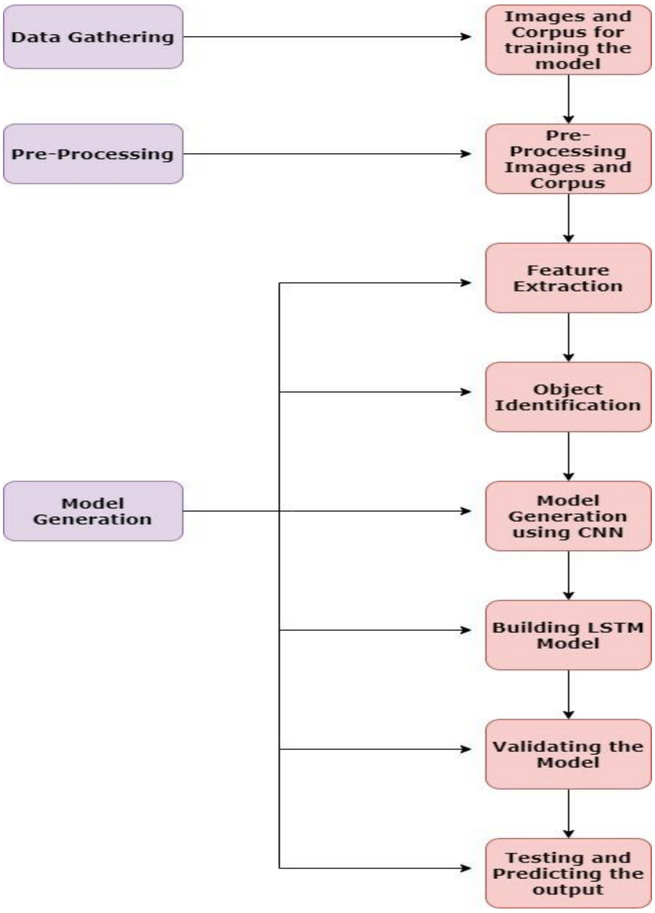

Data Gathering
Images and Corpus for training the model
Pre-Processing
Pre-Processing Images and Corpus
Feature Extraction
Object Identification
Model Generation
Model Generation using CNN
Building LSTM Model
Validating the Model
Testing and Predicting the output

Figure 1. Overview of the system

B. Data Gathering

For ML projects, the gathering of a valid dataset is the most basic and essential functional requirement. The dataset used for building the caption generator model is Flickr8k dataset. Out of 8092 images in the dataset, 6000 is used for training. The dataset also consists of around 60000 corpus for the images which are used as a textual description for training the model.

C. Pre-Processing

The corpus in the dataset is pre-processed for removal of stop words, articles, punctuation marks and digits from the textual descriptions. This is done to generate a bag of words thus helping in the mapping of words with the corpus. The file is saved in pickled format for training purpose.

D. Feature Extraction

This is possible with the help of VGG16 model, one of the CNN algorithm. Feature extraction is done to predict the action being portrayed in the image. Extracted features are then dumped in a pickle file format for model generation.

A sample output of Caption Generation is depicted in the figure 2.

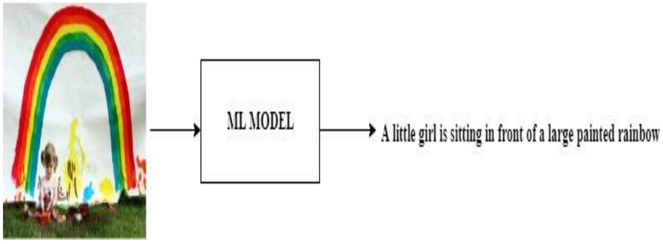

ML MODEL
A little girl is sitting in front of a large painted rainbow

Figure 2. Sample – Caption Generation

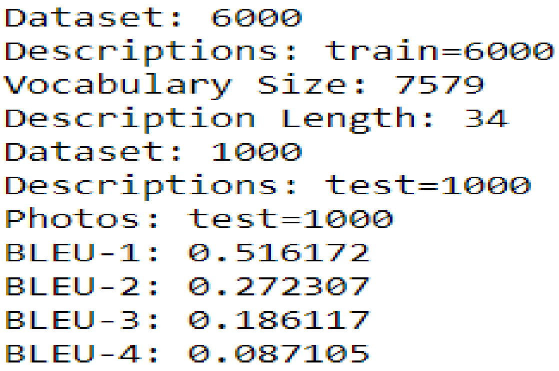
E. Caption Generation

Once the images and corpus are pre-processed, the images are passed through Pretrained CNN model (VGG16 Model) which has been used in this project. With the help of VGG16 Model, features are stored in a pickle format file called 'features.pkl' once the features are extracted from it. Then with the help of LSTM Algorithm, 'features.pkl' and the pre-processed corpus of the images stored as 'descriptions.txt' along with the token of words stored in 'tokenizer.pkl' are passed through the LSTM Model which starts generating the final ML Model with the help of Epochs for generating captions for the images. Among the generated epochs of models, the best model is selected for development of application based on BLUE Score that helps in the evaluation of the translated text.

V. Results and Discussion

A. Model Testing

Training was done by considering 6000 images consisting of different kinds of people, animals and scenarios with their unique features and characteristics. The model is tested with 2000 images of different types which is not there in the training dataset. The result is depicted as a BLEU Score in figure 3.



```
Dataset: 6000
Descriptions: train=6000
Vocabulary Size: 7579
Description Length: 34
Dataset: 1000
Descriptions: test=1000
Photos: test=1000
BLEU-1: 0.516172
BLEU-2: 0.272307
BLEU-3: 0.186117
BLEU-4: 0.087105
```

Figure 3. BLEU Score for Model Evaluation

BLEU Score (Bilingual Evaluation Understudy) helps in comparing quality of text which has been translated by the machine from one language to another. Here the language used is English. It is mainly used to know the prediction accuracy whether the caption generated for the image matches with the textual descriptions in the corpus.

50 % and above is considered as a very good score.

BLEU Score allows to get the different n grams weights specified for the calculation.

Very helpful in calculating different BLEU Score like individual and cumulative ngram scores such as single gram (one-gram) or word pairs (bigram).

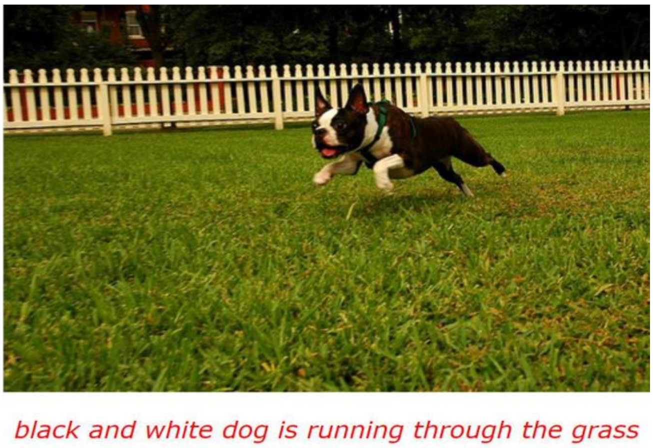


Figure 4. Correct Classification of Caption Generation

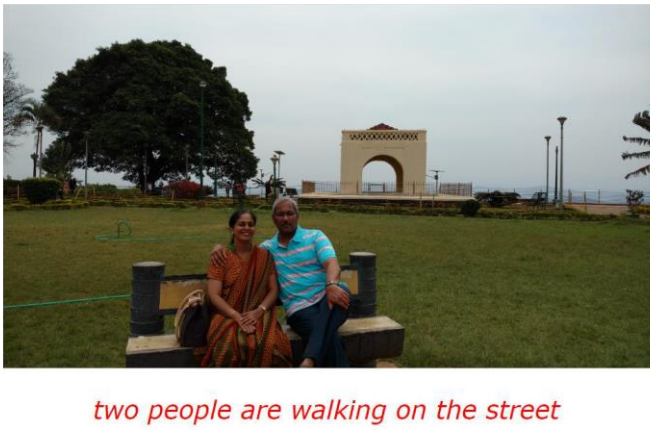


Figure 5. Wrong generation of caption

Figure 4 represents a precise description of the image and Figure 5 represents a wrongly generated caption. The reason behind the wrong generation of caption is because of lower BLEU Score. If the BLEU score is increased then possibly, a precise description will be generated for all the images.

B. *Scene Portrayed in the Image*

Image file is taken as an input, objects are identified from the images and simultaneously features are extracted from them with the help of CNN Algorithm and then fed to the LSTM Algorithm to generate caption for the image. If an image is blank, caption is not generated.

C. *Verification*

The model is tested on a real time and static images. The figure 4 represents only one sample caption generation. The model is tested with images captured from smartphones and digital cameras. Most of the time, the model produces correct result.

D. Future Work

The application should work for distorted and blurred images also. It should be able to generate captions for each frame in the video. A mobile application can be developed that will make the user more convenient to use. In the future, some hybrid algorithms can be used to achieve a higher accuracy that will help in better generation of captions for the images.

VI. Conclusion

The objective of the project is to identify the action portrayed in the given image. Almost 52% of the time the application gives the correct output. The generated caption describes the image in the same way as out human brain recognizes any visual. The caption generator is a useful application for visually impaired people as they might not know what kind of image is it and the kind of action taking place in it. Fine Tuning this application could ease their work. The application also helps in visual media for auto-subtitling. Besides, the whole application has been saved in the GitHub repository. So in the future, if the user requests for any changes, it can be easily done through git version control.

REFERENCES

[1] Jason Brownlee (2007), Deep Learning for Natural Language, Edition v1.1

[2] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, Retrieved from journal of Artificial Intelligence Research submitted on 15th Jan 2016 - https://arxiv.org/abs/1601.03896

[3] Jiuxiang Gu, Gang Wang, Jianfei Cai, Tsuhan Chen, An Empirical Study of Language CNN for Image Captioning, Retrieved from 2017 IEEE International Conference on Computer Vision - https://ieeexplore.ieee.org/

[4] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga, A Comprehensive survey of Deep Learning for Image Captioning, Retrieved from journal on Deep Learning for Image Captioning submitted on 14th October 2018 - https://arxiv.org/pdf/1810.04020.pdf

[5] Kenneth P. Camilleri Marc Tanti, Albert Gatt, What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?, Retrieved from Journal on Computer Vision and Pattern Recognition submitted on 7th August 2017 - https://arxiv.org/abs/1708.02043

[6] Flickr8k Dataset - https://www.kaggle.com/flickr8k

[7] Sumit Saha - A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way - https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[8] Rohit Thakur - Step by step VGG16 implementation in Keras for beginners - https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c

[9] Jason Brownee – Deep Learning for Natural Language Processing - https://machinelearningmastery.com/deep-learning-for-nlp/

[10] Shuang Bai – A survey on automatic image caption generation - https://www.sciencedirect.com/science/article/abs/pii/S0925231218306659?via%3Dihub#!

[11] Anonymous (2018) – VGG16 – Convolutional Network for Classification and Detection - https://neurohive.io/en/popular-networks/vgg16/