Image Caption Generator

Hadia Shafiq MSCSF21M505
Faiza Saleem MSCSF21M509
mscsf21m505@pucit.edu.pk
mscsf21m509@pucit.edu.pk

Abstract

In this project, we make a deep neural networks based image caption generation method. We used a pre-trained model and LSTM to identify the caption of the image. As the deep learning techniques are growing, huge datasets and computer power are helpful to build models that can generate captions for an image. In this project, we have used deep learning techniques like CNN and RNN. Image caption generator is a process which involves natural language processing and computer vision concepts to recognize the context of an image and present it in English. We discuss Keras library, TensorFlow, numpy and jupyter notebooks for the making of this project.We also discuss about flickr8k dataset and CNN used for image classification.

1 Introduction

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but the human can largely understand them without their detailed captions. Image captioning has various applications in various fields such as biomedicine, commerce, web searching and military etc. Social media like Instagram , Facebook etc can generate captions automatically from images. [1].

Some Techniques on deep learning-based image captioning have shown in Fig.

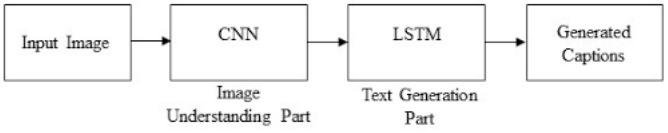

Figure 1: Image Caption Generator

2 Related Work

Some HandCrafted Techinques are used for image captioning model but there are too many advancements in deep neural networks that they have attracted a lot of attention. For Example: The recurrent neural network (RNN) [23] has attracted a lot of attention in the field of deep learning. It was originally widely used in the field of natural language processing and achieved good results in language modeling. Another method is Attention mechanism. the method stemmed from the study of human vision, is a complex cognitive ability that human beings have in cognitive neurology. When people receive information, they can consciously ignore some of the main information while ignoring other secondary information. This ability of self-selection is called attention.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In this project we used CNN and LSTM for generating captions for images

3 Methodology

The main aim of this project is to get a little bit of knowledge of deep learning techniques. We use two techniques mainly CNN and LSTM for image classification. So, to make our image caption generator model, we will be merging these architectures. It is also called a CNN-RNN model. CNN is used for extracting features from the image. We will use the pre-trained model Inception. LSTM will use the information from CNN to help generate a description of the image.

4 Experiments and Results

Flickr8k data set is a public benchmark dataset for image to sentence description. This data set consists of 8000 images with five captions for each image.The batch size we used for training is 30 the number of epoch we used are 10.

4.1 Qualitative evaluation

The image should be converted to suitable features so that they can be trained into a deep learning model. Feature extraction is a mandatory step to train any image in deep learning model. The features are extracted using pretrained model with inception V3

4.2 Quantitative evaluation

The main text file which contains all image captions is Flickr8k.token in our Flickr8ktext folder.We takes all descriptions and performs data cleaning. This is an important step when we work with textual data, according to our goal.We use the pre-trained model that have been already trained on large datasets and extract the features from these models and use them for our tasks.

5 Discussion

Feature Extractor – The feature extracted from the image has a size of 2048, with a dense layer, we will reduce the dimensions to 256 nodes. Sequence Processor – An embedding layer will handle the textual input, followed by the LSTM layer. Decoder – By merging the output from the above two layers, we will process by the dense layer to make the final prediction. The final layer will contain the number of nodes equal to our vocabulary size.

6 Conclusion

We have used Flickr8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. Although deep learning -based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for sometime.

References

[1] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.