ISSN No.: 2455-6130
Volume 4, Issue 5, 2021, pp. 12-16

website: www.ijoetr.com

International Journal of Emerging Trends in Research

Image Caption Generator

Liya Ann Sunny[1], Sara Susan Joseph[2], Sonu Sara Geogy[3], Sreelakshmi K.S[4], Abin T Abraham[5]

[1] *Department of Computer Application, Saintgits College of Engineering, Kerala, India*

Abstract

For our project involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English phrase is needed. Here we use the other concepts of a CNN and LSTM model and build a working model. We can implemented our project using CNN (Convolutional Neural Networks) and LSTM (Long Short Term Memory). We can extracted the project from Xception which is a CNN model trained on the image net dataset. The features of LSTM is to generating the images. In our project Image Caption Generator we use the dataset is Flickr8K. There are use few other datasets can using in our project they are, Flickr30K and MSCOCO dataset compare to small Flickr8K dataset. The main advantage of our project is huge dataset can build better models.

Keywords: Convolutional Neural Network (CNN); Long Short Term Memory (LSTM)

1. Introduction

Image Caption Generator is now become great place to research on. This is the major area of computer vision and natural language processing concepts to recognize the context of an image. In our project the application is extensive and significant. It describing the content of images using natural processing is a fundamental and challenging task. The advancement of our project is computing power along with the availability of huge datasets and this is the main advantage. It can generate captions for an image. In computer vision tasks such as recognizing an object, action classification, image classification and scene recognition it in the form of a human-like sentence.

The goal of our project, is based on semantics of images should be captured here and expressed in the desired form of natural languages. In the real world it has a great impact, for instance by helping visually impaired people better understand the content of images on the web. The model, we will be merging CNN-RNN architectures. The feature extraction from images is done using CNN. We have used the pre-trained model Xception. The information received from CNN is then used by LSTM for generating a description of the image.

In our project we using these approaches are usually generic descriptions of the visual content and background information is ignored. The generic descriptions do not satisfy in emergent situations. The objective of our project is to develop a web based interface for users to get the description of the image and to make a classification system in order to differentiate images as per their description. It can also make the task which is complicated as they have to maintain and explore enormous amounts of data.

2. literature review

In literature review the various reference of the existing projects are taken into consideration which are similar to this current project.

1. Image Caption Generator using Big Data and Machine Learning

In this paper one of the most popular deep neural networks is the CNN is explained. There are multiple layers in CNN; such as convolution layer, non-linearity layer, pooling layer and fully-connected layer as well. It has an excellent performance in machine learning problems and one of the most common algorithms.

2. Comparison of Image Captioning methods

The great approaches is to generate descriptions with lack of specific information, such as named entities they are involved in the images. We will train our project using CNN-LSTM model so that it can generate a caption based on the image.

3. Image Caption Generator using Deep Learning

The problem of artificial intelligence that connects computer vision and natural language processing is describing the content of an image. The AI systematically analyse a deep neural networks based image caption generation method. The input is image, and the method as output in the form of sentence in English describing the content of the image.

4. A Neural Image Caption Generator

The deep neural network algorithm LSTM. LSTM is local both space as well as in time; the computational complexity is per time of step and also the weight pattern representation. The difference of other algorithm LSTM leads to many more successful runs, and learn must faster.

5. Image Captioning- A Deep Learning Approach

The content of an image using properly arranged English sentences is a tough challenging task, but it could is something very necessary for helping visually impaired people. These features are then fed into a RNN or a LSTM model to generate a description of the image in grammatically correct English sentences describing the surroundings.

3. System Architecture

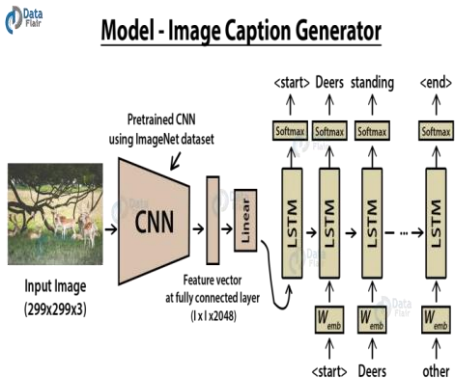

Fig 1: Proposed Model of Image Caption Generator

The proposed model of our project is as shown in the above figure 1. In this model, the input image is given and then CNN is used to create a dense feature vector as shown in this figure. This dense vector, is also called an embedding and this vector can be used as input into other algorithms, and its generate the suitable caption for given image as output.

In image caption generator, this dense vector becomes a representation of the image and use das the initial state of LSTM for generating meaningful captions, for the image. System Architecture of our system is shown below in Figure

This is our proposed system architecture will look like:

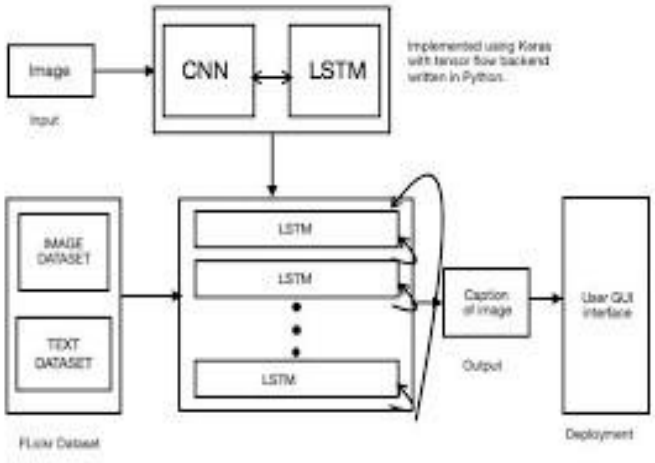

Fig 2: System Architecture of Image Caption Generator

4. Methodology

We have written data processing scripts to process raw input data (both images and captions) into proper format. A pre-trained CNN architecture as an encoder to extract and encode image features into a higher dimensional vector space. An LSTM- based Recurrent Neural Network as a decoder to convert encoded features to natural language descriptions. Algorithms:

a) Convolutional Neural Network

Convolutional Neural Networks (CNN) are specialized deep neural networks which processes the data has input shape like a 2D matrix. Image classification and identification can be easily done using CNN. It can determine whether an image is a bird, a plane, a dog or man etc.

The important features of an image can be extracted by scanning the image from left to right and top to bottom and finally the features are combined together to classify images. It deal with the images that have been translated, rotated, scaled and changes in perspective.

b) Long Short Term Memory

LSTM are type of RNN (Recurrent Neural Network) which is well suited for sequence prediction problems. In our project we can predict what the next words will be based on the previous text. It has shown itself effective from the traditional RNN by overcoming the limitations of LSTM.

c) Data Exploration

For our project, we have used the Flickr8K dataset. There are also other big datasets like Flickr30K and MSCOCO dataset but it can take weeks for systems having only CPU support just to train the network, so we used a small Flickr8K dataset. The huge dataset helps in developing a better model.

5. Evaluation

It has been implemented, it is unclear whether it is absolutely correct and can learn as expected. Therefore, the task is to train our model and validate that it will learn properly. After the model is correct, is a major component of this project and has yet to be finished. The sentence generation and performance evaluation is most important because it is the best way to present how well our caption generator can perform.

Evaluation Metrics

BLEU-1, BLEU-2, BLEU-3 and BLEU-4 (Bilingual Evaluation Understudy) is the most commonly reported metrics in evaluating the quality of text generation in natural language processing tasks. In our project we chose 4-gram BLEU score (BLEU-4) as our primary evaluation metric. The next steps is to also investigate other metrics (such as METEOR).

The rest of the metrics can be computed automatically assuming one has access to ground truth, i.e. human generated descriptions. The most commonly used metric so far in the image description literature has been the BLEU score, which is a form of precision of word n-grams between generated and reference sentences. We use this metric has some obvious drawbacks it has been shown to correlate well with human evaluations.

6. Results

Since our model is data driven and trained end-to-end, and given the abundance of datasets, we want to answer questions such as “how dataset size affects generalization”, “what kinds of transfer learning it would deal with weakly labeled examples” and “how it would deal with weakly labeled examples”. In this result, we performed experiments on five different datasets.

7. Conclusion

An Image Caption Generator has been developed using a CNN-RNN model. There are some key aspects about our project to note are that our model depends on the data so it cannot predict the words that are out of its vocabulary. A dataset consisting of 8000 images is used here. But for production-level models i.e. higher accuracy models, we need to train the model on larger than

100,000 images datasets so that better accuracy models can be developed. First of all we directly used pre-trained CNN network as part of our project.

References

[1] Image Caption Generator using Deep Neural Networks", J. Chen, W. Dong and M. Li, March 2018.

[2] "Understanding of a CNN", S.ALBAWI and T.A. MOHAMMED, 2017.

[3] "Study of LSTM", Neural Computation, S. Hochreiter, December 1997.

[4] "Image Caption Generator using Big Data and Machine Learning", C. Elamri and T. Planque, California, 2016.

[5] "An Survey on Automatic Image Caption Generation", Neuro computing, S. Bai and S.Aan, 13 April 2018.

[6] "Generating sequences with RNN", A. Graves.

[7] "Image Caption Generator- A deep Learning Approach", S.Liu and W. Deng, November 2015.