**MINOR -2**

**SYNOPSIS**

**ON**

**IMAGE CAPTION GENERATOR**

**Submitted By**

**NAME Roll NO**.

Niharika Agrawal R177219121

Prabhraj Singh R177219130

Pradumn Nathawat R177219131



**School of Computer Science**

***Under the guidance of***

**Bikram Pratim Bhuyan**

Designation

Department of Computer Science,

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**Dehradun-248007**

**Aug-Nov, 2021**

**1) Abstract**

Whenever an image appears in front of us our brain is capable of annotating or labeling it. But, what about computers? How can a machine process an image and label it with a highly relevant and accurate caption? It seemed quite impossible a few years back, but now with the enhancement of Computer Vision and Deep learning algorithms, availability of relevant datasets, and AI models, it becomes easier to build a relevant caption generator for an image. Abstract

Even Caption generation is becoming a growing business in the world, and many data annotation firms are earning billions from this. In this guide, we are going to build one such annotation tool which is capable of generating very relevant captions for the image with the help of datasets.

We can use techniques of Deep learning including LSTM (a type of Recurrent Neural Network) and Convolutional Neural Networks (CNN) is required for the same.

**2) Introduction**

Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain. Image captioning has a huge amount of application. NVIDIA is using image captioning technologies to create an application to help people who have low or no eyesight.

Initially, it was considered impossible that a computer could describe an image. With the advancement of Deep Learning Techniques and large volumes of data available, we can now build models that can generate captions describing an image.

Deep learning is one of the most rapidly advancing and researched fields of study that is making its way into all of our daily lives. It is a subfield of machine learning concerned with algorithms and inspired by the structure and function of the brain. The goal of image captioning is to convert a given input image into a natural language description.

**3) Motivation**

Generating captions for images is a vital task relevant to the area of both Computer Vision and Natural Language Processing. Mimicking the human ability of providing descriptions for images by a machine is itself a remarkable step along the line of Artificial Intelligence. The main challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language (like English). Traditionally, computer systems have been using pre-defined templates for generating text descriptions for images. However, 1 this approach does not provide sufficient variety required for generating lexically rich text descriptions. This shortcoming has been suppressed with the increased efficiency of neural networks. Many states of art models use neural networks for generating captions by taking images as input and predicting the next lexical unit in the output sentence.

**4) Related work**

Karpathy et.al [1] developed a model that generated text descriptions for images based on labels in the form of a set of sentences and images. They use multi-modal embeddings to align images and text based on a ranking model they proposed. Their model was evaluated on both full frame and region level experiments and it was found that their Multimodal Recurrent Neural Net architecture outperformed retrieval baselines.

Kiros et.al [2] proposed a neural network-based approach for generating text descriptions from image and for image retrieval from text. They used a Multimodal log bilinear model that was biased by the features of the input image.

Vinyals et.al [3] have proposed a model which uses CNN to generate image features which are passed to an LSTM network. Using word embeddings and feature vectors as intermediaries in determining gate values, they used beam search to obtain multinomial distribution to generate image captions.

NVIDIA is using image captioning technologies to create an application to help people who have low or no eyesight.

**5) Methodology**

In this project, we will first start off by using a dataset for our neural network model. For this we will be first cleaning (removing punctuations, converting text into lower case and removing words that contain number) our dataset. In our dataset we have 5 captions for a single image. We will be creating a description dictionary that maps images with a list of 5 captions.

We will be using transfer learning, where we don’t have to do anything on our own, we will be using the pre-trained model that has already been trained on large datasets and extract the features from these models and use them for our tasks. The extracted features for all images will be mapped with image names and along with their respective feature array. Then we will dump the features dictionary into a ‘.p’ extension file.

With this we will be loading the images into our model and train it. We would have to tokenize the vocabulary as the computer cannot understand English words and save it into another ‘.p’file.

Since our dataset contains large amount of dataset, we will be needed to yield batches of the dataset which would be done by the generator method.

Then we will be defining the model (eg: CNN-RNN model) for the training and testing purposes.

**6) Plan of Work**



**REFERENCE**

[1] Karpathy, A., and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306 (2014).

[2] Kiros, R., Salakhutdinov, R., and Zemel, R. Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (2014), pp. 595–603.

[3] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555 (2014).

[4] <https://data-flair.training/blogs/python-based-project-image-caption-generator-cnn/>

[5] <https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350>

[6] https://www.cse.iitk.ac.in/users/cs671/2015/\_submissions/vinsam/project/proposal.pdf