**19MIS1018- B DEVI PRASAD**

**SWE1904\_ CAPSTONE PROJECT**

**Description**

**Image and Video Captioning Using DeepLearning.**

**PROBLEM STATEMENT:**

Based on the content of an Image, this application will generate a caption for any natural image. Such an application might help blind people to see the world full with images. This model automatically generates natural language captions which then can be utilized for indexing and searching of images, tagging in social media, helping the visually impaired etc. The uses of such an application is immense. In this paper, we will build such an application using Convolution Neural Networks (CNN) for feature extraction and Long Short Term Memory (LSTM) for generating the captions.

**ABSTRACT:**

When we see an image, we can quickly recognize what is going on in the image, what objects are present and what they are doing. With the progress in Artificial Intelligence (AI), we are trying to do the same automatically by our computers. The need for such a system is increasing especially due to the advent of autonomous vehicles / semi-autonomous vehicles which involves reading and understanding millions of images. Automatically generating captions for any given image requires the use of Natural Language Processing (NLP) techniques and Neural Networks to classify the images. The ability for a computer to generate captions to an image has various business and individual benefits.

**EXISTING METHOD**

In most of the currently available systems, the model used for extracting features from the image is Inception deep learning model. But, Google in 2017 proposed a new model known as Xception which is a modified version of the Inception v3 model (also known as extreme inception). Xception, when compared to inception v3 is proven to provide a slightly better classification model for an image. It is trained on ImageNet dataset and it performs slightly better for small datasets but it outperforms Inception v3 by a significant margin when trained on large datasets.

Important features in an image can have high variations in area which makes it difficult to select the correct kernel size for the convolution layer. Going deeper to solve the problem can be computationally expensive as well. In 2014, a new deep convolutional neural network was proposed known as Inception to tackle the above mentioned problems. Rather than just going deeper, this architecture also went towards being more wider (in the form of levels) to increase efficiency. In doing so, it also reduces the need to go deeper as deeper models are prone to overfitting.

* Dataset : Flickr8k dataset.

**ARCHITECTURE**

