A Deep Learning Approach to Image Captioning

Rohith Chityala,
chityalar@coventry.ac.uk
Student id: 10013258

Introduction:
Until a few years ago, it was almost seemed impossible that a machine can generate a description for an image. But now, due to the developments of Deep Learning and the data available, several models are being designed which can generate accurate captions for a given image. Deep Learning is a part of Machine learning which is designed in such a way that a computer can think like a human brain and take decisions accordingly. In recent years, Deep Learning has seen a noticeable growth and applications in solving real world problems. One such unique application is vision-language objective, an Image Captioning, which requires a combination of both Computer Vision and Natural Language Processing technologies.

Purpose:
Image captioning has a wide range of applications directly and indirectly. For example, image tagging for e-commerce, photo sharing services and online catalogs. In e-commerce, when a product image is uploaded, automatic tagging and captioning of the product specifications can be done (type, color of clothing in terms of clothes). One major application of Image captioning can be found in the social media platforms such as Instagram and Facebook. The models employed by them are becoming smarter every day and are giving more accurate results. Another indirect use of image captioning is to help visually impaired people by converting the text generated for an image into verbal description. This type of technology is known to be effective and is being widely used in the form of virtual assistants by companies like Amazon, Apple, and Google.

Methodology:
In this project, the Flickr 8k dataset containing of 8091 images in JPEG format is used which also contains a Flickr8k_text file that has different captions with the image names. The main goal is to develop a model which generates a caption for given input image, describing clearly about the actions in the image. Basically, it can be divided into two steps- image based model, also called as encoder, which assigns importance to various features present in the input image. This is achieved by a deep learning algorithm called Convolutional Neural Network(CNN). The second step is language based model which is the Long Short Term Memory(LSTM). LSTM is a type of Recurrent Neural Network mostly used in machine translation, speech recognition etc. The output feature vector generated from the CNN is then fed as input to the LSTM model which then generates natural language sentences. The main tasks include preprocessing the images, creating the vocabulary for the images, train the model, evaluate the model, and test it on new images to predict the captions. Here, model evaluation is done by a metric called Bilingual Evaluation Understudy Score(BLEU). The score 1.0 indicates a perfect match and 0.0 indicates a perfect mismatch. Finally, the model is tested on a completely new image and accurate captions are generated. All the above mentioned process is achieved by implementing libraries such as TensorFlow and Keras in Python.

Dataset: https://www.kaggle.com/adityajn105/flickr8k?select=Images.

Keywords:
Deep learning, Neural Networks, NLP, Computer Vision, CNN, LSTM, TensorFlow, Python.