IMAGE CAPTIONING

This assignment aims to describe the content of an image by using CNNs and RNNs to build an Image Caption Generator. The model is implemented using Tensorflow and Keras. The dataset used is Flickr 8K, consisting of 8,000 images each paired with five different captions to provide clear descriptions. The implementation uses Python in a Google Colab notebook, where every step is documented.

Architecture

The model architecture consists of a CNN which extracts the features and encodes the input image and a Recurrent Neural Network (RNN) based on Long Short Term Memory (LSTM) layers. The most significant difference with other models is that the image embedding is provided as the first input to the RNN network and only once.

The following steps are performed:

1. Load dataset from local path or google drive
2. In order to extract the features from the images, a pre-trained CNN model, named Inception V3 is loaded.
3. Create a dictionary with a picture filename as the key and an array of captions as the value
4. Create a dictionary with the image filename as the key and the image feature extracted using the pre-trained model as the value.
5. Generate train and test set- This approach divides image_filenames, to avoid same image with different caption in train and test dataset.
6. Tokenize train labels- Generate a vocabulary and transform the train captions to a vector with their indices in the vocabulary.
7. Checkpoint- Create a TensorFlow checkpoint on a local path to save the encoder and decoder state while training
8. Test Stage- Evaluate random images