SAIDL Assignment- Q1

**Introduction:**

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but the human can largely understand them without their detailed captions.

Image captioning is a deep learning task which involves generating captions or textual descriptions for images. It requires an algorithm to understand and model the relationships between visual and textual elements, and to generate a sequence of output words.

The generation of captions from images has various practical benefits, ranging from aiding the visually impaired, to enabling the automatic and cost-saving labelling of the millions of images uploaded to the Internet every day.

**Review of Literature:**  
Recent advancements in image captioning have led to variety of approaches to perform the task. Some of them are simple encoder-decoder models, GAN based approaches, training a RL agent to generate captions given a reward, etc.

Previous encoder-decoder image captioning methods use LSTM that are unidirectional and relatively shallow in depth. Wang et al. [144] proposed a deep bidirectional LSTM-based method for image captioning. This method is capable of generating contextually and semantically rich image captions.

Another method introduced by Fang et al. to generate more comprehensive captions uses an image model to get the features, a language model, and a multimodal similarity model to train the model on an image captioning dataset.

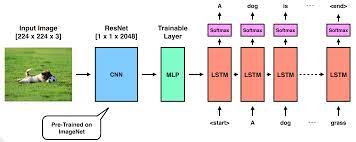
**Basic requirements:**

Existing deep learning-based image captioning methods use variants of image encoders to extract image features. The features are then fed into the neural network-based language decoders to generate captions.

The dataset used will be MSCOCO dataset. The MSCOCO data comprises 82783 training images, 40504 validation images and 40775 test images. Each image is accompanied by at least five captions of varying length. In the training set, there are 414113 captions in total, for an average of 5.002 captions per image.

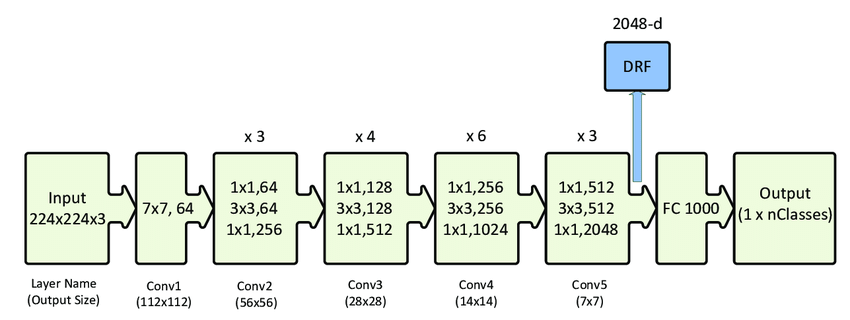
The hyperparameters for the model will be number of lstm layers, learning rate,dropout,etc.

**Model architecture:**



The model will have two parts:

1. A CNN architecture: This will be basically used to extract features from the input image.
2. LSTM: The LSTM takes the features as inputs and predicts words one after the other.



We will first pass the image through the ResNet 50 model, the last layer being 2048 × 1 output layer . The output from this will be the image representations. We then train a linear transformation that maps it to the LSTM network. The LSTM network will contain 2 layers and will be bidirectional in nature.

We then feed the representations as the first input of the LSTM network and each hidden state of the LSTM emits a prediction for the next word in the sentence.

Loss function used in our model is cross entropy loss.

**Metrics:**

Commonly used metrics in image captioning are:

* BLEU- Individual text segments are compared with a set of reference texts and scores are computed for each of them which are then averaged over the entire text. It is the commonly used metric for this task.
* ROUGE- It compares word sequences, word pairs, and n-grams with a set of reference summaries created by humans.
* CIDEr – it is an automatic consensus metric for evaluating image descriptions.

For comparison purposes we run a control i.e randomly generate captions for a given image from the vocabulary.