**PROJECT REPORT**

**Image Caption Generator**

**Syed Areeb Ali (20K-1686)**

**Aliyan (20K-1098)**

**Muhammad Huzaifa (20K-0257)**

**Objective:**  
We will be doing image captioning on a large dataset using two different models and will compare and evaluate the accuracy of both the models.

**Problem Statement:**

The problem of image captioning is to generate accurate and coherent descriptions of images automatically. This involves understanding the content of the image and generating grammatically correct and contextually relevant captions, Traditional methods often fail to capture the context and details present in images. This project aims to address this challenge by leveraging deep learning models to automatically generate captions that accurately describe the content of images.

**Methodology:**  
We have done the following steps to build our image caption generator:

1. **Data Collection and Preprocessing**: The Flickr8k dataset of image 1GB and caption 2.2 MB, comprising images and corresponding captions, is used for training and evaluation. The dataset is preprocessed to remove irrelevant captions and ensure consistency.
2. **Model Architectures**:

**CNN with Transformer**: The image captioning model architecture combines a CNN with a Transformer-based encoder-decoder architecture. The CNN extracts high-level features from the input image, which are then passed to the Transformer encoder. The encoder processes the image features and generates a representation that is used by the Transformer decoder to generate captions word by word.

**CNN (VGG-16) with LSTM**: An alternative model architecture involves a CNN-LSTM network. The CNN (VGG-16) extracts features from the input image, which are then fed into an LSTM network. The LSTM processes the sequential features and generates captions word by word.

1. **Training**: Both model architectures are trained using the training dataset, with the objective of minimizing the loss function, which measures the discrepancy between the predicted captions and the ground truth captions. The training process involves iterative optimization of the model parameters using backpropagation and gradient descent.
2. **Evaluation**: The trained models are evaluated on the same dataset for validation to assess their performance in generating accurate and meaningful captions. Evaluation metrics such as loss and accuracy are computed to quantify the models' performance. The performance of the CNN with Transformer and CNN with LSTM architectures is compared to identify strengths and weaknesses.

**Result:**

So through the model of CNN with Transformer model we have found the average loss to be 0.1811 compared to the second model of CNN (VGG-16) with LSTM which gave the average loss of 0.4965 and we have concluded that CNN with Transformer model performs better as a image caption generator on large dataset.

**References:**

1. Vaswani, A., et al. (2017). "Attention is All You Need." Advances in Neural Information Processing Systems.
2. Xu, K., et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." International Conference on Machine Learning.
3. Brownlee, J. (n.d.). "How to Develop a Deep Learning Photo Caption Generator from Scratch." Machine Learning Mastery.
4. Keras Documentation. (n.d.). Keras API Reference.
5. TensorFlow Documentation. (n.d.). TensorFlow API Reference.