Project Design Phase-I
Solution Architecture

Date	10 November 2023
Team ID	592117
Project Name	Project – Image Caption Generator
Maximum Marks	4 Marks

Solution Architecture:

The architecture usually consists of two main components: a convolutional neural network (CNN) for image processing and a recurrent neural network (RNN) for generating captions.

Convolutional Neural Network (CNN):

Input Layer: Takes in the raw image pixels.
Convolutional Layers: Extract features from the image using convolutional operations. Popular architectures include VGG16, ResNet, or Inception.
Pooling Layers: Reduce the spatial dimensions of the features, helping to focus on important information.
Flattening Layer: Convert the 2D feature maps to a 1D vector.
Fully Connected Layers: Transform the features into a format suitable for input to the RNN.

Recurrent Neural Network (RNN):

Initial Hidden State: Obtained from the output of the CNN.
Word Embedding Layer: Converts words into vectors in a continuous vector space.
LSTM or GRU Layers: These recurrent layers capture the temporal dependencies in the sequence of words and help generate meaningful captions.
Fully Connected Layer: Maps the output of the RNN to the vocabulary size, producing a probability distribution over words.
Softmax Activation: Converts the output into probabilities, indicating the likelihood of each word.
Training:

The model is trained end-to-end using a dataset with image-caption pairs.
The loss function measures the difference between the predicted caption and the ground truth caption.
Backpropagation and optimization algorithms (e.g., Adam) update the model parameters.
Inference:

During inference, a new image is passed through the trained CNN to extract features.
The RNN then generates a caption one word at a time, considering the context and previously generated words.
The process continues until an end token is generated or a maximum caption length is reached.
Word Embeddings:

Pre-trained word embeddings (e.g., GloVe or Word2Vec) can be used in the word embedding layer to capture semantic relationships between words.
Attention Mechanism (Optional):

An attention mechanism can be added to the RNN to focus on different parts of the image when generating each word in the caption.
This architecture is a general overview, and various modifications and enhancements can be made based on specific requirements and advancements in the field.

Solution Architecture Diagram For (Image Caption Generator) :