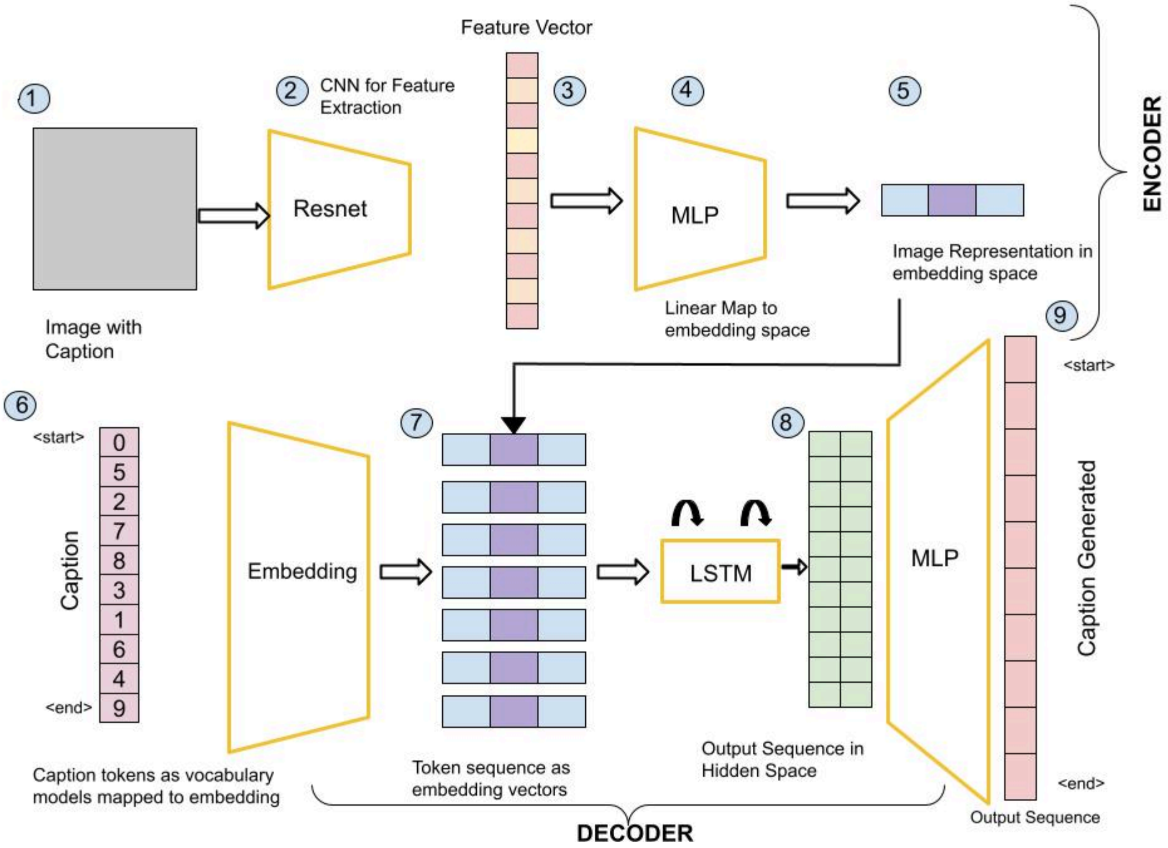
Image Captioning

PART A:
Architecture:

The below diagram is the overall architecture of the CNN+LSTM model.



Feature Vector
1
2
CNN for Feature Extraction
3
4
5
Resnet
MLP
ENCODER
Image with Caption
Linear Map to embedding space
Image Representation in embedding space
9
<start>
6
<start>
0
5
2
7
8
3
1
6
4
9
<end>
Caption
Embedding
7
8
LSTM
MLP
Caption Generated
<end>
Caption tokens as vocabulary models mapped to embedding
Token sequence as embedding vectors
Output Sequence in Hidden Space
Output Sequence
DECODER

Encoder (CNN-based):

1. **Preprocessing**: Resized,and normalised the images to ensure consistent input dimensions of size (3,224,224).
2. **Pretrained ResNet-50:** Utilised a frozen ResNet-50 model to extract high-level features.

3. **Embedding:** Projected the extracted features into a lower-dimensional space for visual representation.

Decoder (LSTM-based):

1. **Word Embeddings:** Converted caption words into dense vectors for semantic understanding.
2. **LSTM Decoder:** Employed a LSTM network with image features and word embeddings for sequential caption generation.
3. **Caption Generation:** Predicted the next word in sequence based on current context during training.
4. **Loss Calculation:** Compared the predicted word probabilities with ground truth using cross-entropy loss.
5. **Training:** Optimised the encoder and decoder parameters jointly via backpropagation and gradient descent.
6. **Evaluation:** Assessed the model performance using ROUGE_L score for caption similarity.

Model Parameters:

- Batch SIze = 128
- Number of epochs = 10
- Hidden layer size = 256
- Embedding layer size = 128
- Vocabulary size = 2553
- Learning Rate = 0.001

RESULTS:

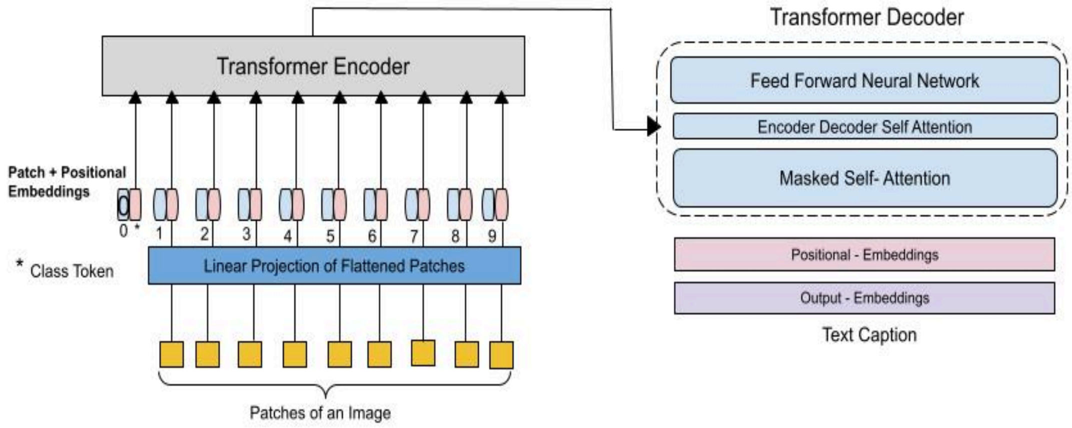
Rouge_L mid scores : (Based on longest common subsequence in the prediction)

- Precision indicates the proportion of words in the predicted captions that are relevant to the ground truth captions.
- Recall measures the proportion of words in the ground truth captions that are successfully captured by the predicted captions.
- Precision=0.26
- Recall=0.248
- Fmeasure=0.23961258318402084

PART B:

Diagram:

The below diagram is the overall architecture of the ViT + GPT-2 model.



1. Input Processing:

- **Preprocessing:**
    - Captions are tokenized using the default tokenizer of GPT-2, it involves splitting the caption text into individual tokens, representing words or subwords.
    - Special tokens indicating the beginning and end of sequences (`bos_token_id` and `eos_token_id`) are added to mark the start and end of captions.
    - Padding or truncation is applied to ensure a fixed length of the tokenized sequences, with a maximum length of 239 tokens.
    - Images are resized and normalized to a standard size of 224x224 pixels to ensure uniformity in image dimensions.

- **VIT Encoder**: The input images are passed through the Vision Transformer (VIT) encoder, which extracts high-level visual features from the images.. The output of the VIT encoder is a set of visual embeddings that capture rich visual information from the input images. These embeddings represent the features of the image in a high-dimensional space.

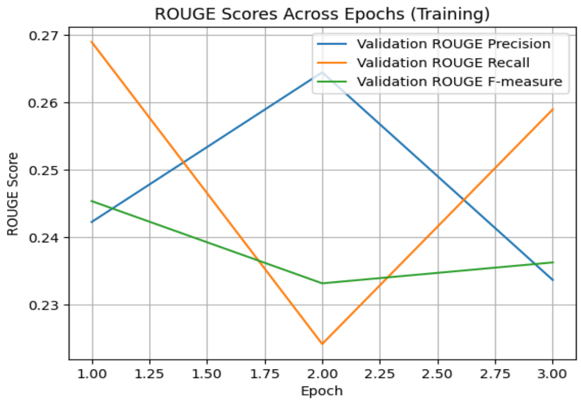
2. Caption Generation (GPT-2 Decoder):

- **Input Preparation:** The visual embeddings obtained from the VIT encoder serve as the input to the GPT-2 decoder along with special tokens indicating the start and end of the caption.
- **Language Modelling:** GPT-2 decoder is a language model trained on large text corpora. It generates captions by predicting the next word in the sequence based on the context provided by the input embeddings and previously generated words.
- **Decoding strategy:** Beam search was used to decode the generated caption.

3. Training Process:

- **End-to-End Training:** The entire model, consisting of the VIT encoder and GPT-2 decoder, is trained end-to-end using a dataset of images paired with their corresponding captions with cross entropy as the loss function.

Epochs Vs performance on validation dataset :



Model Hyperparameters :

- Number of training epochs = 3
- Learning rate = 5e-5
- Evaluation_steps = 715

- Logging_steps = 1024
- Save_steps = 2048
- Warmup_steps = 1024
- Per_device_train_batch_size = 8
- Per_device_evaluation_batch_size = 8

RESULTS :

1. **Precision (0.3407):**

   Approximately 34.07% of the words in the predicted captions match the words in the ground truth captions.
2. **Recall (0.3007):**

   About 30.07% of the words in the ground truth captions are covered by the predicted captions.
3. **F-measure (0.3025):**

   The obtained F-measure of approximately 0.3025 indicates the overall effectiveness of the model in generating captions.