DL ASSIGNMENT 2 REPORT

PART A

Methodology:
We have implemented a custom image captioning model using a Vision Transformer (ViT) as the encoder and a GPT-2 language model as the decoder.

Architecture Overview:

1. Encoder: ViT (vit-small-patch16-224) pre-trained on ImageNet.
2. Decoder: GPt-2 with cross-attention enabled.
3. Connecter Layer: A linear layer to project ViT outputs to match GPT-2 input dimensions.
4. Token: A special <|img|> token is prepended to captions to denote the image input during training.

Training Strategy:

1. ViT is frozen to reduce overfitting and training cost.
2. GPT-2s last two transformer blocks are fine-tuned.
3. The connecter layer and final GPT-2 blocks are trained using CrossEntropyLoss, masking out the padding tokens.

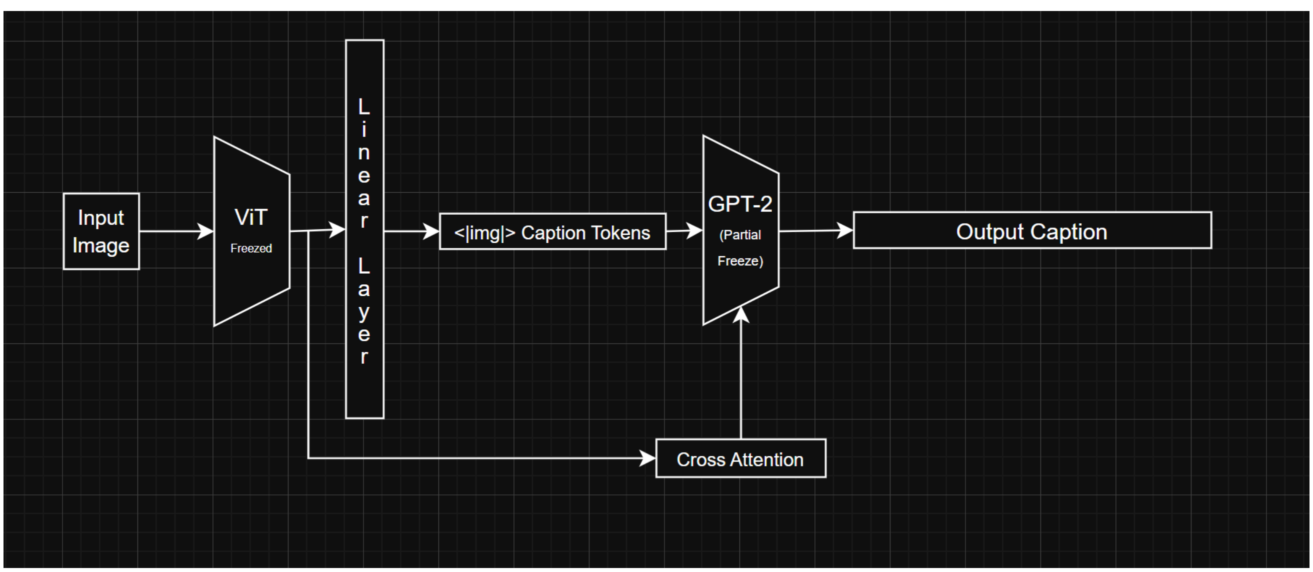

Figure 1: Custom Captioning Model Architecture

PART C

Methodology:
We have developed a classifier to identify the model that generated the caption.

Architecture:

1. Text Encoder: BERT (bert-base-uncased).
2. Classifier Head: Fully connected layer for classification.

Training Strategy:

1. The input text is tokenised using BERT.
2. CLS token embeddings are fed into a classification head.
3. CrossEntropyLoss is used for training.

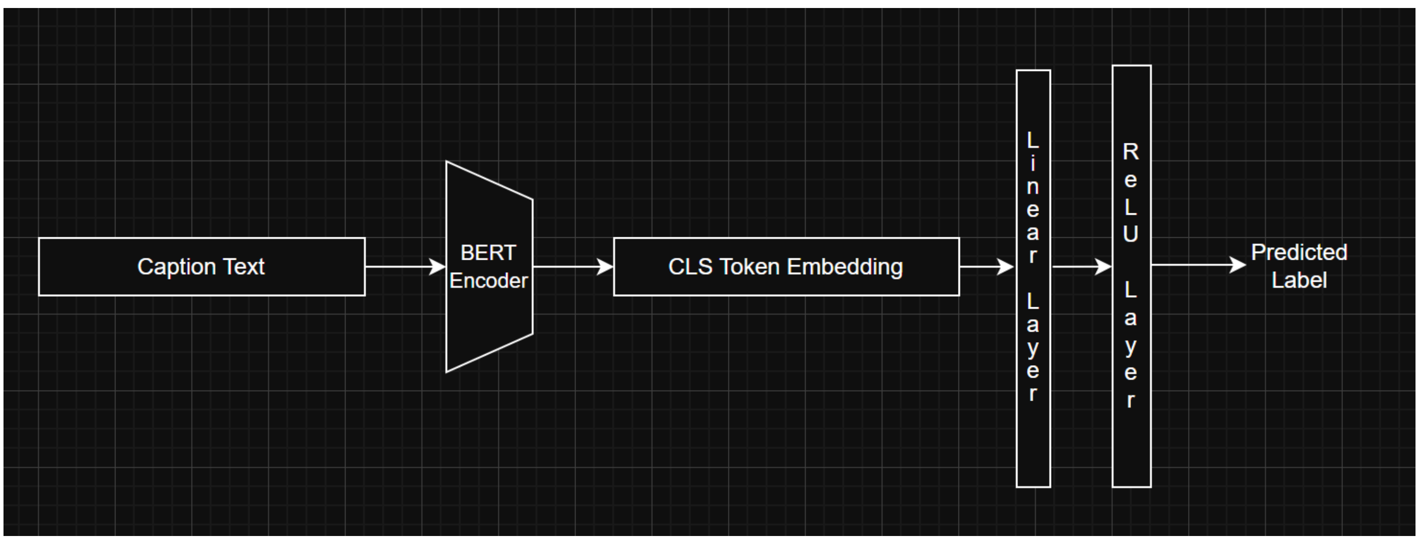

Figure 2: BERT-base based Classification Model

Results

Parts A & B

Model	Occlusion (%)	BLEU	ROUGE-L	METEOR	BERT (F1)
Custom	0	0.05919313	0.27218635	0.21563981	0.5337085
Custom	10	0.05032046	0.25813304	0.20328732	0.5367446
Custom	50	0.03073763	0.222820406	0.16705072	0.53461355
Custom	80	0.02834237	0.230663754	0.17424710	0.5356932
SmolVLM	0	0.05450976	0.239604271	0.27500039	0.5149172
SmolVLM	10	0.05176395	0.236930110	0.27178732	0.5165361
SmolVLM	50	0.03402938	0.211227373	0.24163033	0.50544107
SmolVLM	80	0.00914989	0.173578031	0.19219580	0.47384304

Part C

Metric	Value
Macro Precision	0.973185928410052
Recall	0.973103641913249
F1	0.973081319413875

Conclusion

1. The custom ViT-GPT2 model demonstrates a strong semantic alignment as reflected by consistent BERTScore F1 values.
2. SmolVLM performs better in fluency-oriented metrics like METEOR and ROUGE at lower occlusion; however, performance degrades as occlusion increases
3. The BERT-based caption generator classifier achieves a high F1 score of 97%, showing that different captioning models leave distinct semantic and linguistic patterns detectable through learned embeddings.