Deep Learning Assignment 2: Automatic Image Captioning

Team ID: *13* (Backpropagators)

Methodology

Part A: Implementing a Custom Encoder-Decoder Model

1. **Custom Encoder-Decoder Model:**

- **Architecture:**

  A custom image captioning model was built using the VisionEncoderDecoderModel class from transformers.

  - **Encoder:**

    A pre-trained Vision Transformer *(ViT-Small-Patch16-224, WinKawaks/vit-small-patch16-224)* was used as the image encoder.

  - **Decoder:**

    A pre-trained GPT-2 model *(gpt2)* was used as the text decoder.

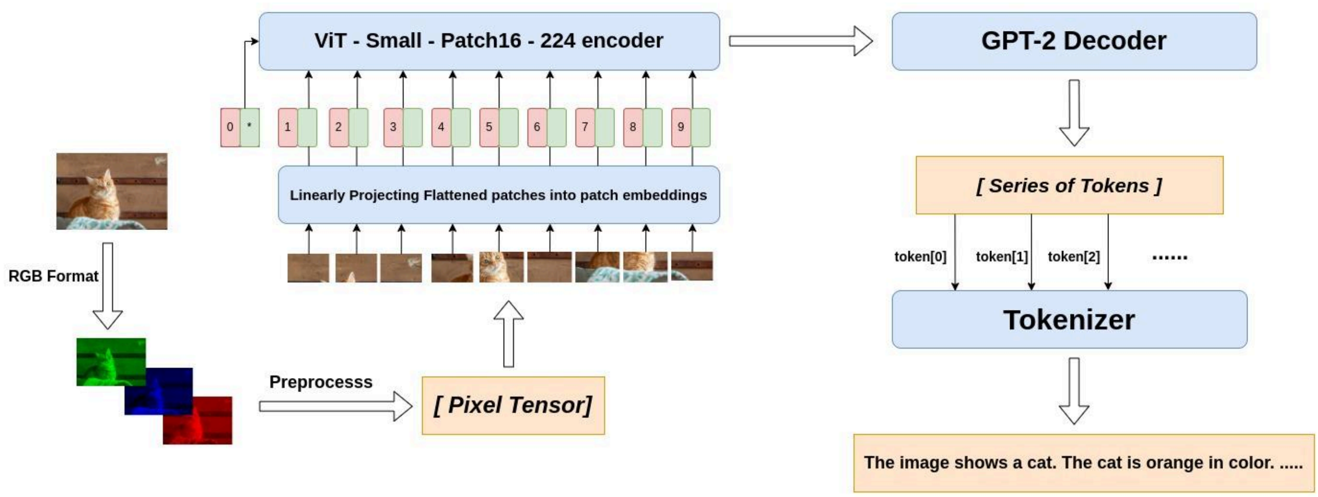
  - **Initialization:**

    The decoder's tokenizer *(AutoTokenizer for gpt2)* was used to set the model's configuration for padding and BOS tokens.

- **Training:**

  The model was trained on the provided dataset using the AdamW optimizer (learning rate = **5e-5**). Custom early stopping was implemented (patience=3 epochs based on validation loss), saving the best model state.

Also benchmarked it with smolVLM on our dataset.



Part B: Studying Performance Change Under Image Occlusion

- **<u>Occlusion & Evaluation:</u>**

  An occlusion function masked image patches (10%, 50%, 80%). Both SmolVLM and the custom model were evaluated and the generated captions, original captions, image IDs, and occlusion percentages were saved into final_raw_results.csv for use in Part C.

Part C: Building a BERT-based Classifier

1. **<u>Classifier Model and Training:</u>**

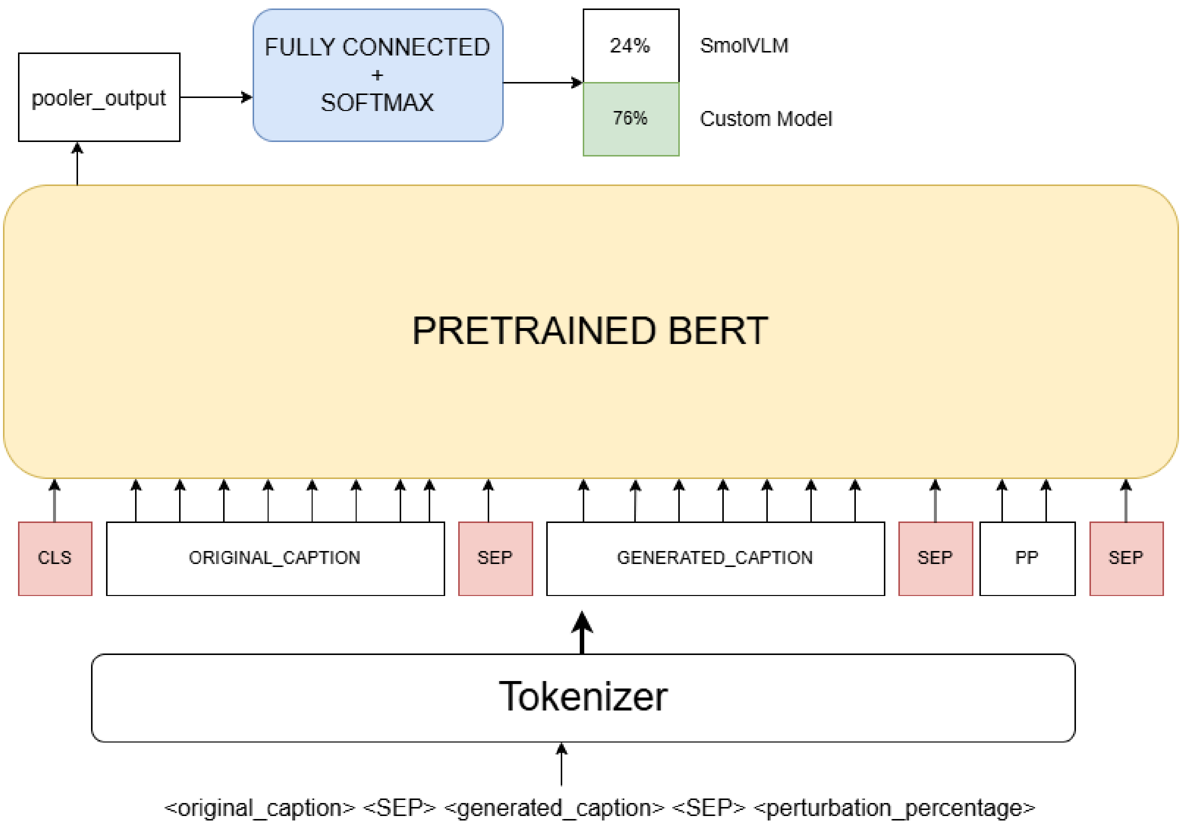
   - **<u>Architecture:</u>**

     A classifier was built on top of a pre-trained BERT-base-uncased model *(google-bert/bert-base-uncased)*.

     - The formatted input text is tokenized using the BERT tokenizer and sent to the BERT model.
     - The pooler_output (representation of the [CLS] token) is obtained.
     - Dropout is applied to the pooled output.
     - The result is passed through a final linear layer to output logits for the 2 classes (SmolVLM vs. Custom).

   - **<u>Training:</u>**

     The training was conducted over **5** epochs using a batch size of **16**, a learning rate of **2e-5,** and an AdamW optimizer configured with an epsilon value of **1e-8**. The loss function used is **cross-entropy loss**.



Evaluation Results Summary

Part A : Zero-Shot Evaluation Metrics

Model	BLEU	ROUGE-L	METEOR
SmolVLM	0.0575	0.2319	**0.2687**
Custom	**0.0697**	**0.2915**	0.2392

The **Custom model** consistently outperformed **SmolVLM** in BLEU and ROUGE-L during zero-shot evaluation, indicating better alignment with reference captions.

Part B: Performance on Occluded Images

Occlusion Level: 10%

Model	BLEU	ROUGE-L	METEOR
SmolVLM	0.0525	0.2253	**0.2601**
Custom	**0.0631**	**0.2837**	0.2333

Occlusion Level: 50%

Model	BLEU	ROUGE-L	METEOR
SmolVLM	0.0395	0.1763	0.1932
Custom	**0.0435**	**0.2525**	**0.2075**

Occlusion Level: 80%

Model	BLEU	ROUGE-L	METEOR
SmolVLM	0.0142	0.1045	0.1068
Custom	**0.0339**	**0.2408**	**0.1980**

Both models degraded in performance under image occlusion. However, the **Custom model** showed greater **resilience**, especially at 50% and 80% occlusion.

Performance Change (Occluded - Baseline)

Occlusion	Model	ΔBLEU	ΔROUGE-L	ΔMETEOR
10%	SmolVLM	-0.0050	-0.0066	-0.0086
	Custom	-0.0066	-0.0077	-0.0059
50%	SmolVLM	-0.0180	-0.0556	-0.0755
	Custom	-0.0262	-0.0390	-0.0317
80%	SmolVLM	-0.0433	-0.1275	-0.1619
	Custom	-0.0358	-0.0506	-0.0412

- The Custom model is significantly more robust to visual occlusion than SmolVLM, especially in preserving semantic and structural quality at higher occlusion levels.

Part C: BERT-based Source Classifier

Metric	Score
Macro Precision	0.9861
Macro Recall	0.9857
Macro F1	0.9857
Accuracy	0.9857

- The **BERT classifier** achieved nearly perfect precision and recall, demonstrating that the captions generated by each model have distinguishable patterns that can be learned effectively.