Automatic Image Captioning with Model Benchmarking and Robustness Analysis

Preetham Battula - 22CS10015
Gavinikadi Aravind - 22CS10024
Kovvuru Kasyap - 22CS10039

Dr.Stone

Methodology

Part A: Implementing and Benchmarking a Custom Encoder-Decoder Model

Objective: Develop a transformer-based encoder-decoder model for image captioning, train it on the provided dataset, and benchmark its performance against SmolVLM's zero-shot capabilities.

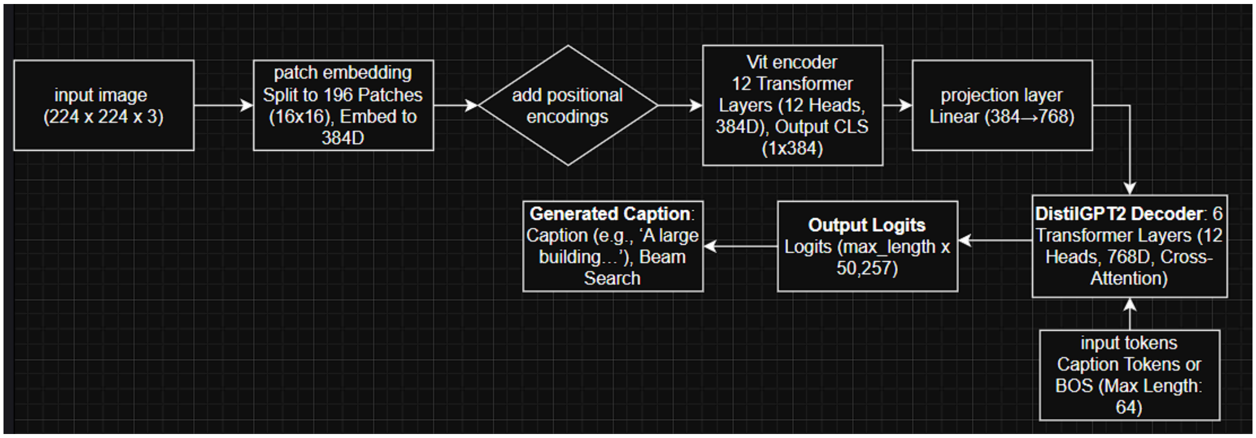
Approach:

- **Dataset**: The dataset (dataset.zip) includes RGB images and captions in train.csv and test.csv under /content/dataset/custom_captions_dataset/. The training set trains the custom model; the test set evaluates both models.
- **Zero-shot SmolVLM**: The zero_shot_captioning function uses HuggingFaceTB/SmolVLM-Instruct to generate test set captions without fine-tuning. Images are preprocessed with AutoProcessor, prompted with "Can you describe the image?", and captions (max length: 64) are generated in bfloat16 on a T4 GPU. BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR scores are computed via evaluate_captions.
- **Custom Model**: The ImageCaptionModel combines a Vision Transformer encoder (vit-small-patch16-224) and a DistilGPT2 decoder (distilgpt2) with cross-attention, designed for 15GB GPU memory.
- **Training**: The prepare_dataloader function creates a DataLoader (batch size: 8) with preprocessed images (224x224, normalized) and tokenized captions (max length: 64). The train_model function trains for 5 epochs using Adam (lr=5e-5) and cross-entropy loss, freezing the encoder and training the decoder and projection layer.
- **Evaluation**: The generate_captions function uses beam search (4 beams) to produce test captions. The evaluate_captions function computes BLEU, ROUGE, and METEOR scores, comparing the custom model to SmolVLM.

Custom Model: ImageCaptionModel:

- **Architecture**:
  - **Encoder (ViT)**: Processes a 224x224x3 image into 196 patches (16x16), embedded to 384D with positional encodings. Twelve transformer layers (12

heads, hidden size 384) output 197x384 vectors; the CLS token (1x384) represents the image.
  - **Projection**: A linear layer maps the 384D CLS embedding to 768D (DistilGPT2's dimension).
  - **Decoder (DistilGPT2)**: Six transformer layers (12 heads, hidden size 768) with cross-attention process tokenized captions and the projected embedding. Outputs logits (batch_size, max_length-1, 50,257).
  - **Forward**: Images yield CLS embeddings, projected to 768D, and fed to the decoder with input tokens, returning logits.
  - **Generation**: Beam search starts with BOS, generating captions up to 64 tokens.



Part C: Building a BERT-based Classifier for Model Identification

Objective: Develop a BERT-based classifier to distinguish between captions generated by SmolVLM and the custom ImageCaptionModel, using original captions, generated captions, and perturbation levels.

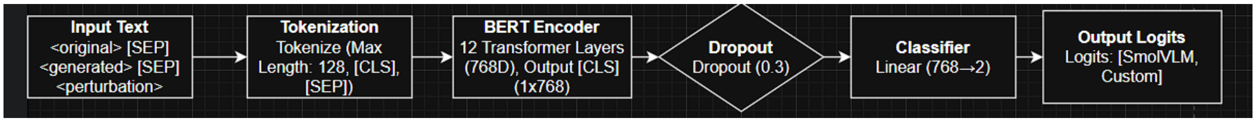
Approach:

- **Dataset**: The dataset (caption_classifier_dataset.csv, 5,568 samples) contains entries formatted as "<original_caption> [SEP] <generated_caption> [SEP] <perturbation_percentage>" with labels (0: SmolVLM, 1: custom). It is split by unique original captions into training (3,894 samples, 70%), validation (552 samples, 10%), and test (1,122 samples, 20%) sets to avoid image overlap.
- **Custom Model**: The CaptionClassifier class uses bert-base-uncased with a dropout and linear layer for binary classification, designed for GPU efficiency.
- **Training**: The CaptionClassifierDataset tokenizes inputs (max length: 128) using BERT's tokenizer. The train_classifier function trains with AdamW and cross-entropy loss, using a DataLoader with tuned batch sizes. Hyperparameter tuning tests learning rates (1e-5, 2e-5, 5e-5), batch sizes (8, 16, 32), and epochs (2, 3, 4) via hyperparameter_tuning, selecting the best based on validation accuracy.

- **Evaluation**: The evaluate_classifier function computes macro-averaged precision, recall, F1, and accuracy on validation and test sets, ensuring robust performance.

Custom Model: CaptionClassifier:

- **Architecture**:
  - **BERT Encoder**: Tokenizes input text into 128 tokens, processed by 12 transformer layers (12 heads, hidden size 768). Outputs a 768D pooled [CLS] embedding.
  - **Dropout**: Applies 0.3 dropout to prevent overfitting.
  - **Classifier**: A linear layer maps 768D to 2 logits (SmolVLM, custom).
  - **Forward**: Takes input IDs, attention mask, and token type IDs, returning logits (batch_size, 2).
- **Details**:
  - BERT: Pretrained weights are fine-tuned.
  - Training: Uses best hyperparameters (e.g., lr=1e-5, batch_size=16, epochs=2). Model saved as caption_classifier_model.pt.
  - Input: Combines captions and perturbation with [SEP] tokens.



Results

PART - A

Training results of custom model

Epoch no	1	2	3	4	5
Avg loss cross-entropy loss	`2.7940`	`2.4627`	`2.2908`	`2.1575`	`2.0348`

Testing results of custom model

bleu	**rouge-1**	**rouge-2**	**rouge-l**	**meteor**
`0.0465`	`0.360`	`0.103`	`0.274`	`0.244`

Testing results of SmolVLM model

bleu	**rouge-1**	**rouge-2**	**rouge-l**	**meteor**
`0.040`	`0.388`	`0.105`	`0.251`	`0.243`

PART - B

model	Occlusion%	Δ bleu	Δ rouge-1	Δ rouge-2	Δ rouge-l	Δ meteor
custom	10	-0.0037	-0.0095	-0.0064	-0.0023	-0.0080
custom	50	-0.0065	-0.0176	-0.0134	-0.0058	-0.0104
custom	80	-0.0141	-0.0339	-0.0285	-0.0213	-0.0243
SmolVLM	10	+0.0000	+0.0033	+0.0012	+0.0017	+0.0003
SmolVLM	50	-0.0042	-0.0131	-0.0088	-0.0048	-0.0115
SmolVLM	80	-0.0255	-0.0871	-0.0515	-0.0463	-0.0601

PART-C

LR,BS	Epochs	Accuracy	Precision	Recall	F1 Score	accuracy
LR=1e-05 BS=16	2	0.9837	0.9838	0.9837	0.9837	0.9837
LR=2e-05 BS=16	2	0.9837	0.9838	0.9837	0.9837	0.9837
LR=5e-05 BS=16	2	0.9837	0.9838	0.9837	0.9837	0.9837
LR=2e-05 BS=8	2	0.9837	0.9838	0.9837	0.9837	0.9837
LR=2e-05 BS=32	2	0.9837	0.9838	0.9837	0.9837	0.9837
LR=2e-05 BS=16	3	0.9837	0.9842	0.9837	0.9837	0.9837
LR=2e-05 BS=16	4	0.9837	0.9842	0.9837	0.9837	0.9837

Best hyperparameters found

Learning Rate: 1e-05,Batch Size: 16,Epochs: 2,Validation Accuracy: 0.9837

Validation set - Accuracy: 0.9837,**Precision:** 0.9838,**Recall:** 0.9837,**F1 Score :** 0.9837

Test set - Accuracy: 0.9837,**Precision:** 0.9838,**Recall:** 0.9837,**F1 Score :** 0.9837