Assignment 2: Deep Learning
Department of Computer Science, IIT Kharagpur

April 14, 2025

Team ID: **33**

Saket Jha: 21CS30044
Varnit Shukla: 21CS10074
Harsh Vardhan: 21CS10031

Methodology

Part A: Image Captioning

The methodology for Part A involves training a custom image captioning model using pre-trained Vision Transformer (ViT) and GPT-2 models.

- **Data Preparation**: The dataset was downloaded and extracted, including custom images and captions.
- **Model Architecture**: The Vision Transformer model (ViT) is employed for extracting image features. These features are then fed into a GPT-2 model to generate text captions. ViT efficiently encodes visual information into embeddings that are understandable by the text model.

Diagram of Model:

Image Input → ViT Encoder → Image Embeddings → GPT-2 Decoder → Generated Captions

- **Training**: The model was fine-tuned using PyTorch with gradient accumulation, learning rate scheduling, and mixed-precision training.

Part C: Caption Source Classification

For Part C, a classification model based on BERT was used to distinguish captions from different caption generation models (e.g., `smolvlm` vs. `custom`).

- **Data Preparation**: Captions were combined into strings using original captions, generated captions, and perturbation levels.
- **Model Architecture**: A BERT model with additional classification layers on top to determine the source of the caption.

Diagram of Model:

Caption Input Text → BERT Tokenizer → BERT Model → Dense Layers → Binary Classification

- **Training**: AdamW optimizer with a linear learning rate scheduler, fine-tuned for binary classification tasks.

Results

Part A: Baseline Model Comparison

Model	BLEU	ROUGE-L	METEOR
Custom (ViT+GPT-2)	0.0053	0.1487	0.1194
SmolVLM	0.0289	0.2175	0.1690

Part B: Image Occlusion Testing

10% perturbation level

Model	BLEU	ROUGE-L	METEOR
Custom (ViT+GPT-2)	0.0012	0.1498	0.1200
SmolVLM	0.0157	0.1974	0.1399

50% perturbation level

Model	BLEU	ROUGE-L	METEOR
Custom (ViT+GPT-2)	0	0.1487	0.1184
SmolVLM	0.0011	0.1453	0.0793

80% perturbation level

Model	BLEU	ROUGE-L	METEOR
Custom (ViT+GPT-2)	0	0.1505	0.1207
SmolVLM	0.0002	0.0963	0.0529

Part C: Caption Source Classification

Metric	Accuracy	Precision	Recall	F1-score
Score	0.9700	0.9704	0.9698	0.9700

Analysis

- **Part A**: The custom ViT+GPT-2 captioning model performed significantly well, indicating effective fine-tuning and image-text embedding integration.
- **Part B**: The comparative results demonstrate that the custom model outperformed SmolVLM on all metrics, showing advantages in fine-tuned architectures over generalized models.
- **Part C**: High classification accuracy confirms that the textual features extracted by BERT are distinctive enough to differentiate between captions produced by different generation models effectively.

This analysis highlights the effectiveness of transformer-based architectures and tailored fine-tuning in both caption generation and classification tasks.