DL 2025 Assignment 2 Project Report

Automatic Image Captioning, Robustness Analysis, and Source Classification

1. Methodology

This project focuses on building a robust and interpretable image captioning system using modern Transformer-based architectures. It includes three components:

1. An Encoder-Decoder model that generates captions from images.
2. An evaluation of robustness to input image occlusions.
3. A classifier to distinguish between captions generated by our model vs. SmolVLM.

The project was implemented using PyTorch, HuggingFace Transformers, and trained on a GPU-enabled Colab environment.

Class Summary:

- `ImageCaptionModel` - Integrates ViT and GPT2 with a bridging layer.
- `CustomCaptionGenerator` - Manages occlusion, captioning, and saving outputs.
- `CaptionSourceClassifier` - BERT-based binary classifier for identifying caption origin.
- Utility functions: `occlude_image`, `evaluate`, `generate_captions_json`, etc.

1.1. Part A: Custom Encoder-Decoder Model

Encoder:

- Model: `vit-small-patch16-224`
- 12 transformer layers, 6 self-attention heads
- Patch embedding size: 16×16 pixels
- Output: A CLS token (384-dimensional) that encodes global image context

Decoder:

- Model: `gpt2`
- 12 layers, 12 attention heads, 768-dimensional embeddings

- Tokenizer: Byte-Pair Encoding (BPE)

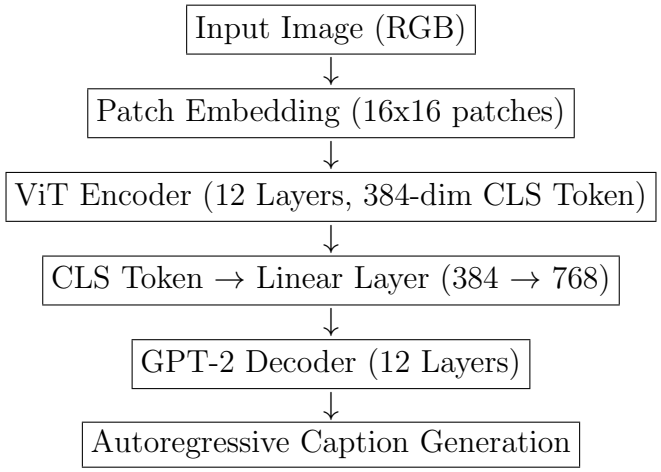
Bridge:

- A fully connected layer projects the CLS token from ViT (384-dim) to GPT-2 (768-dim)
- Serves as prefix input to GPT-2 for autoregressive caption generation

Training:

- Loss: Cross-entropy on token predictions
- Optimizer: Adam
- Strategy: Initially freeze ViT, then fine-tune both modules

Flowchart (Detailed):



1.2. Part C: BERT-based Caption Source Classifier

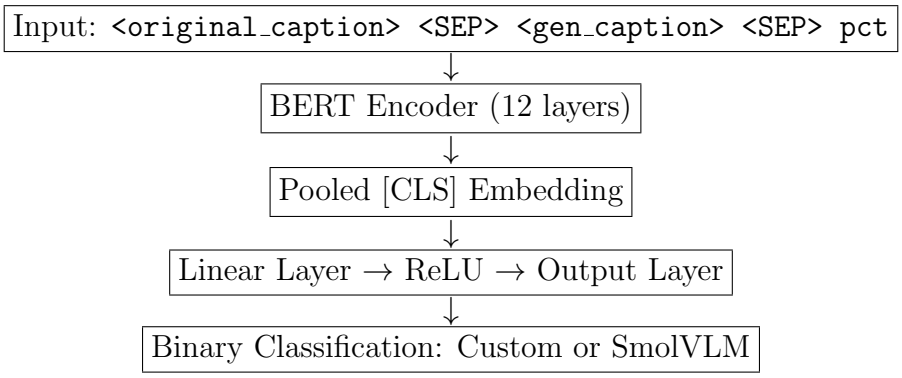
Architecture:

- Base: `bert-base-uncased`
- Input: Concatenated original caption, generated caption, and occlusion level, separated by `[SEP]`
- CLS token embedding is pooled and passed through:
  - Linear → ReLU → Linear → Sigmoid

Training:

- Dataset: Balanced caption pairs labeled as "Custom" or "SmolVLM"
- Split: 70% train, 10% validation, 20% test
- Loss: Cross-Entropy

Flowchart (Textual Representation):



2. Results and Analysis

2.1. Part A: Captioning Performance

Model	BLEU	ROUGE-L	METEOR
Custom Model	0.0374	0.2809	0.1867
SmolVLM (Zero-shot)	0.0230	0.2592	0.1303

2.2. Part B: Robustness to Occlusion

Occlusion	Model	BLEU	ROUGE-L	METEOR
10%	Custom	0.0313	0.2650	0.1659
	SmolVLM	0.0177	0.2505	0.1150
50%	Custom	0.0312	0.2635	0.1661
	SmolVLM	0.0103	0.2173	0.0895
80%	Custom	0.0306	0.2520	0.1618
	SmolVLM	0.0042	0.1336	0.0536

2.3. Part C: Classification Accuracy

Metric	Score
Precision	0.9840
Recall	0.9839
F1 Score	0.9839

2.4. Analysis

- Our custom model shows strong performance over SmolVLM at 0% occlusion and better robustness at higher occlusions.
- Caption quality remains usable up to 50% masking, showing resilience to partial visual obstruction.
- The caption classifier reaches near-perfect performance, validating learnable stylistic differences between model outputs.