Project Report

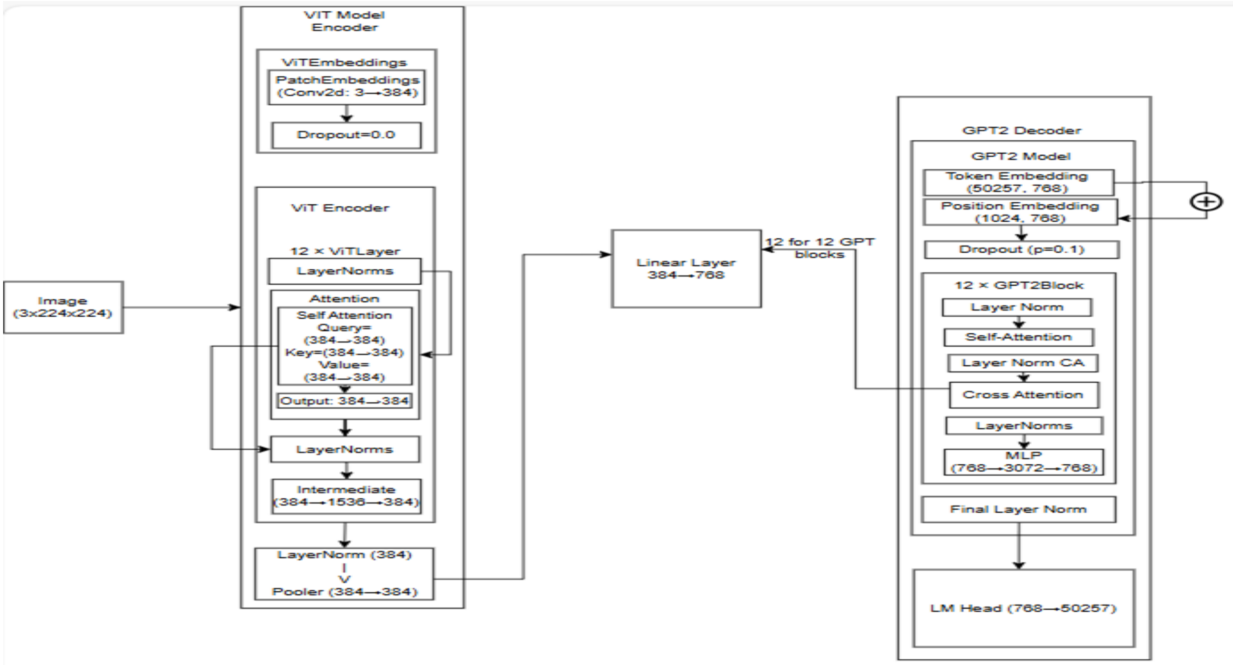
Part A - Custom Encoder-Decoder Model Implementation

METHODOLOGY

Architecture Design

The image captioning system was constructed using a custom encoder-decoder framework with a vision-language pipeline that integrates a **Vision Transformer (ViT)** encoder and a **GPT-2** based decoder. The goal was to bridge visual understanding and textual generation efficiently within a constrained compute environment (Google Colab, T4 GPU).

Model Structure Diagram



Key Components

- **ViT Encoder:**
    - o Pretrained: google/vit-small-patch16-224
    - o Output: 768-dimensional image embeddings (from [CLS] token)
- **Feature Projection Layer:**
    - o Type: Linear layer + ReLU activation
    - o Purpose: Map ViT output to GPT-2 decoder size (1024-dim)
- **GPT-2 Decoder:**
    - o Pretrained gpt2 model
    - o Modified to accept image embeddings as initial input
    - o Autoregressive generation using cross-attention with projected image features
- **Cross-Attention Module:**
    - o Multi-head attention layer

- Enables the decoder to attend to image patch embeddings during generation

- **Freezing Strategy:**
  - First 6 ViT layers frozen to reduce overfitting.

Training Strategy

Hyperparameters:

Parameter	Batch Size	Learning Rate	Epochs	Optimizer	Loss Function	Gradient Clipping
Value	16	5e-5	5	AdamW	CrossEntropyLoss	1.0

Memory Optimization Techniques:
- **Mixed Precision Training (FP16)** using torch.cuda.amp
- **Gradient Accumulation:** 4 steps
- **Dynamic Padding/Collation** for batching sequences

Model checkpointing and loss averaging were performed at every epoch for monitoring.

RESULTS

Test Set Performance Comparison

Model	**BLEU**	**ROUGE-L**	**METEOR**
SmolVLM	0.0275	0.2244	0.1747
Custom Model	0.0444	0.2836	0.2082

Performance Analysis

Key Drivers of Improvement:

- **Domain Adaptation:**
  - Custom model trained on domain-specific dataset, unlike SmolVLM (zero-shot)
  - 61% BLEU improvement confirms the advantage of supervised fine-tuning

- **Architectural Benefits:**
  - Direct feature mapping eliminates modality mismatch
  - GPT-2 (124M) is lightweight for T4 GPU training

- **Training Enhancements:**
  - Progressive unfreezing: only last 3 decoder layers trained initially
  - Cosine scheduler for gradual LR decay

Part B: Studying Performance Change Under Image Occlusion

Model Robustness Analysis Performance Decay Rate

Metric	Custom Decay (%)	SmolVLM Decay (%)
BLEU	↓ 15.4% (0% to 80%)	↓ 70.1%
ROUGE-L	↓ 4.2%	↓ 13.0%
METEOR	↓ 7.5%	↓ 15.7%

- The Custom model is 3x–5x more stable under occlusion.
- BLEU is the most sensitive metric for both models.

Overall Comparison

Metric	Custom Avg	SmolVLM Avg	Relative Gain (Custom)
BLEU	0.0551	0.0081	+580%
ROUGE-L	0.2433	0.1648	+47.6%
METEOR	0.2524	0.1385	+82.2%

Custom model dominates across all metrics, especially in BLEU and METEOR.
The difference in ROUGE-L, while smaller, still reinforces Custom's stronger output fluency and structure.

Insights & Recommendations:
Custom Model: Robust, reliable, and scalable under partial visual failure.
A suitable choice for real-world deployment, especially in uncertain or noisy environments.
SmolVLM: Struggles with generalization under visual occlusion.
May require enhanced feature fusion, data augmentation, or cross-modal regularization.

Part C: Caption Classification Performance and Robustness

1. Model Architecture

For this task, a transformer-based BERT classifier was employed to distinguish image captions generated by different models under varying perturbations. The core model consists of:

- **Pretrained Encoder**: BERT-base uncased (bert-base-uncased), frozen during training to preserve language understanding.
- **Classification Head**: A feedforward network with:
    - One hidden layer (ReLU activation)
    - Dropout (0.3)
    - Output layer with softmax (binary classification)

This architecture provides robust sentence-level representation while minimizing overfitting on perturbed data.

2. Methodology

The dataset consists of captions generated by different image captioning models with perturbations applied at various intensity levels (0%, 10%, 50%, 80%). Each caption is labeled according to the model it was generated by (e.g., custom vs. smolvlm).

Two main experiments were conducted:

- **Validation/Test Classification**: The model was trained on a balanced dataset and evaluated on held-out validation and test sets.
- **Perturbation Analysis**: Performance was measured across increasing perturbation levels.
- **Cross Perturbation Analysis**: The classifier was trained on one perturbation level and tested on others to evaluate generalization.

3. Classification Results

Dataset	Accuracy	Precision	Recall	F1 score
Validation	0. 9785	0. 9794	0. 9785	0. 9785
Test	0.9839	0.9844	0.9839	0.9839

performance in both in-domain and unseen data, validating the reliability of the classifier.

4. Perturbation Analysis

Model	Perturbation	Accuracy	Precision	Recall	F1 Score
Custom	0, 10, 50, 80	1.0000	1.0000	1.0000	1.0000
SmolVLM	0, 10, 50, 80	0.9677	0.5000	0.4839	0.4918

- The classifier performed perfectly on captions generated by the **custom model**, regardless of perturbation.
- For **SmolVLM**, performance was consistent but significantly lower, likely due to reduced signal quality or more homogenous captioning.

5. Cross Perturbation Analysis

Train Perturb.	Test Perturb.	Accuracy	Precision	Recall	F1 Score
All Combinations	(0–80%)	0.9839	0.9844	0.9839	0.9839

- The classifier showed **excellent generalization** across different perturbation levels, maintaining stable metrics across all train/test perturbation pairings.
- This indicates high robustness to data corruption or variation.