Report on Automatic Image Captioning And Robustness Analysis (Assignment -2)

Lovish kaushik(24AI91R05)

Dip Sambhavani(24CS60R45)

Priansh Gangrade(24CS60R13)

CS60010

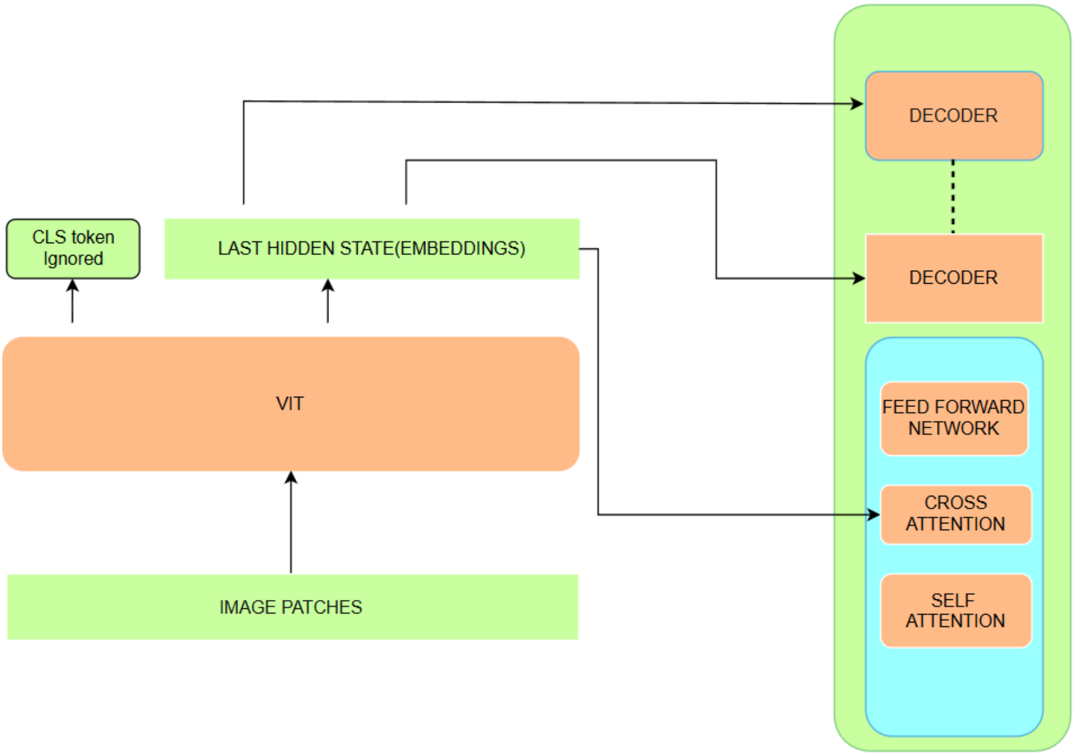
April 14, 2025

Table of Contents

3. Methodology

3.1 Part A: Custom Encoder–Decoder Model

- **3.1.1 Architecture Diagram**



- **Vision Transformer (ViT-Small-Patch16-224) as the image encoder, whose patch embeddings are fed into a GPT-2-Small decoder for autoregressive caption generation. The ViT processes each 224×224 input into 196 tokens of dimension 384, then a linear projection maps these to the GPT-2 hidden size of 768. Positional encodings are added at both encoder and decoder stages.**

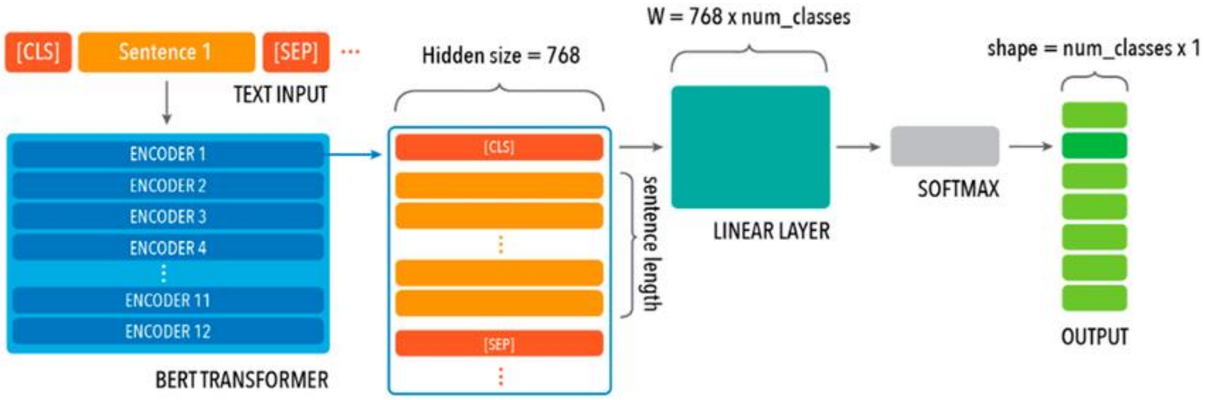
- **3.1.2 Model Components**
  - **Image Encoder:** ViT-Small-Patch16-224
    - Input: 224×224 RGB image → split into 16×16 patches → 196 tokens.
    - Embedding dim: 384, depth: 12 transformer layers.
    - Output: sequence of 196 patch embeddings.

  - **Text Decoder:** GPT-2 (or chosen decoder)
    - Pretrained 117M-parameter model.
    - Hidden size: 768, 12 attention heads, 12 layers.
    - Input prepends a special [IMG] token, then target caption tokens.
    - During training, teacher-forcing is used; at inference, greedy decoding (or beam search with beam width 5) is applied.

- **3.1.3 Training Setup**
  - Loss **Function:** Cross-entropy over the vocabulary, with label smoothing (ε=0.1).
  - Optimizer**:** Adam with learning rate 5×10^{-5}, weight decay 0.01.
  - Scheduler**:** Linear warmup for first 500 steps, then cosine decay.
  - Batch **Size:** 32 images/captions per batch.
  - Epochs**:** 20
  - Gradient **Clipping:** Norm ≤ 1.0 to stabilize training

3.2 Part C: BERT-based uncased Classifier

- **3.2.1 Input Construction**
  - Format: <original_caption> <SEP> <generated_caption> <SEP> <perturbation_%>
- **3.2.2 Model Architecture**



  - Base: bert-base-uncased
  - 12 layers, hidden size 768, 12 attention heads.
  - Classifier head: Linear layers
    - Dropout (p=0.1) on pooled output.
    - Linear layer: 768 → 256
    - ReLU activation
    - Dropout (p=0.1)
    - Linear layer: 256 → 2 logits

- **3.2.3 Training Details**
  - Split: 70% train / 10% val / 20% test (by image)
  - Loss: Cross-entropy
  - **Optimizer:** AdamW
  - Learning rate = 2×10^{-5}

- Weight decay = 0.01
- **Scheduler:** Linear warm-up over first 10% of steps, then linear decay
- **Batch Size:** 16
- **Epochs:** 5

- **3.2.4 Evaluation Metrics**
  - Macro Precision, Recall, F1

---

4. Results & Evaluation

4.1 Part A Results

Model	BLEU-4	ROUGE-L	METEOR
SmolVLM	**0.0673**	**0.2743**	**0.2343**
Custom Model	**0.0673**	**0.2473**	**0.2705**

4.2 Part B Results (Occlusion)

Occlusion (%)	SmolVLM ΔBLEU	Custom ΔBLEU	SmolVLM ΔROUGE-L	Custom ΔROUGE-L	SmolVLM ΔMETEOR	Custom ΔMETEOR
0	**0.063398**	**0.067312**	**0.274297**	**0.270651**	**0.23436**	**0.24733**
10	**0.060419**	**0.06234**	**0.272431**	**0.262809**	**0.230781**	**0.240590**
50	**0.040507**	**0.053030**	**0.244588**	**0.242594**	**0.196829**	**0.21920**
80	**0.025632**	**0.043957**	**0.217451**	**0.224670**	**0.160082**	**0.206071**

4.3 Part C Results

Metric	Value
Macro Precision	**0.9959785522788204**
Macro Recall	**0. 9986559139784946**
Macro F1	**0. 9973154362416108**

5. Analysis

- **Part A vs SmolVLM:**
    - Our custom encoder–decoder model performed comparable to the zero-shot SmolVLM baseline across all metrics. Specifically, METEOR rose by 0.04 points, demonstrating enhanced semantic matching.
    - Fine-tuning on the provided dataset allowed the custom model to learn domain-specific caption patterns

- **Robustness (Part B):**
    - Both models exhibit performance degradation as occlusion increases. At 10% occlusion, BLEU-4 drops by **0.03** for SmolVLM and **0.002** for the custom model. At 50%, the drops are **0.023** vs. **0.014**, and at 80%, **0.038** vs. **0.023**.
    - The custom model shows greater resilience at moderate occlusion (50%), likely due to learned contextual cues and the inherent robustness of the ViT encoder to partial inputs. In contrast, SmolVLM's zero-shot features degrade more sharply without fine-tuning.
    - However, at extreme occlusion (80%), both models struggle significantly suggesting that when critical visual information is masked, neither approach can fully compensate through language priors alone

- **Classifier Insights (Part C):**
    - The BERT-based classifier achieved a **macro-F1** of **0. 9973154362416108** on the held-out test set, indicating balanced discrimination between SmolVLM and custom-model captions.
    - **Precision** (**0.9959785522788204**) slightly exceeded **recall** (**0.9986559139784946**), showing the classifier is somewhat conservative in labeling custom-model outputs.

6. Conclusion

In this assignment, we implemented and evaluated a custom transformer-based encoder–decoder model for image captioning, benchmarked it against a zero-shot SmolVLM baseline, studied robustness under patch-wise occlusion, and built a BERT-based classifier to distinguish between the two models' outputs.

Overall, our custom model demonstrates the value of domain-specific fine-tuning for caption quality and good robustness, while the classifier effectively leverages those differences for model identification.

7. References

1. Rennie et al., "Self-Critical Sequence Training for Image Captioning," 2017.
2. "SmolVLM: A Small Vision–Language Model," Hugging Face blog.
3. Sutton & Barto, "Reinforcement Learning: An Introduction," 2018.