Image Captioning: Custom Model Benchmarking and Robustness Analysis with Ablation Study

Sachish Singla[a], **Swaminathan S K**[a] **and Tharun Selvam K**[a]

[a]*Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur , Email: {sachishs.15, swami2004, tharunselvam}@kgpian.iitkgp.ac.in*

Abstract—This report details the development and evaluation of a custom image captioning model based on the CLIP Vision Transformer and GPT-2 architecture, incorporating contrastive learning. An ablation study comparing frozen weights versus progressive unfreezing training strategies informed the final model configuration. The chosen model's performance is benchmarked against a zero-shot SmolVLM baseline on standard metrics (BLEU, ROUGE-L, METEOR, BERTScore). Furthermore, the robustness of both models is analyzed under varying levels of patch-wise image occlusion (10%, 50%, 80%). Finally, a BERT-based classifier is trained to distinguish between captions generated by the SmolVLM and the custom model, considering the occlusion level as input context. Results demonstrate that our custom caption generator combining CLIP encoder and GPT decoder outperforms SmolVLM even under occlusion settings.

Keywords—*Image Captioning, CLIP, GPT-2, Transformer, Robustness Analysis, Occlusion, BERT, Classification, Contrastive Learning, Ablation Study, Fine-tuning*

1. Introduction

This project focuses on Automatic Image Captioning, the task of generating textual descriptions for given images. The primary goals were:

1. To implement and train a custom encoder-decoder model for image captioning, leveraging a pre-trained vision encoder (CLIP ViT) and a language decoder (GPT-2), enhanced with contrastive learning. Different training strategies involving weight freezing were explored.
2. To benchmark the chosen custom model against an off-the-shelf Small Vision-Language Model (SmolVLM) in a zero-shot setting.
3. To analyze the robustness of both models by evaluating their performance degradation on images subjected to patch-wise occlusion at 10%, 50%, and 80% levels.
4. To build and evaluate a BERT-based classifier capable of identifying the source model (SmolVLM vs. Custom) given the original caption, generated caption, and occlusion percentage.

Evaluation was performed using standard metrics: BLEU, ROUGE-L, METEOR, and BERTScore for caption quality, and Precision, Recall, and F1-score for the classifier.

2. Pipeline

2.1. Part A: Custom Image Captioning Model

The final custom image captioning model follows an encoder-decoder architecture combined with a contrastive learning objective, selected based on comparative experiments.

2.1.1. Image Encoder

The vision encoder utilizes the pre-trained Vision Transformer (ViT) component from the CLIP model ('openai/clip-vit-base-patch32'). The encoder weights were kept frozen during training for the final reported model. The hidden state corresponding to the '[CLS]' token from the final layer is extracted as the primary image representation (dimension 768).

2.1.2. Decoder

A pre-trained GPT-2 model ('gpt2') serves as the text decoder. Cross-attention layers were enabled in the GPT-2 configuration to allow conditioning on the image features. In the final model, to enable structured input for the caption generation task, two special tokens — <|IMG|> and <|CAPTION|> — were introduced. These tokens mark the beginning of the image-conditioned input and the start of the caption text, respectively. Due to the addition of these tokens, the embedding layer was partially fine-tuned alongside the final two transformer blocks and the cross-attention layers, while the remaining parameters were kept frozen.

2.1.3. Encoder-Decoder Connection

An MLP connects the encoder's CLS token hidden state to the decoder's cross-attention dimension. The MLP architecture is: Linear(768 → 1536) → GELU Activation → Linear(1536 → 768). This non-linear projection adapts the image features for the decoder.

2.1.4. Contrastive Learning

To improve the alignment between image and text representations, a contrastive learning objective was added. This was done to leverage CLIP's training objective which involved corresponding images and text.

- **Image Projection Head**: The 768-dim CLS hidden state (before the MLP connector) is passed through a projection head (Linear(768→256) → ReLU → Linear(256→256)) to obtain a 256-dim image embedding.
- **Text Projection Head**: The input caption is processed by the GPT-2 decoder's transformer layers. The hidden state of the last token is extracted (768-dim) and passed through a similar projection head (Linear(768→256) → ReLU → Linear(256→256)) to obtain a 256-dim text embedding.
- **Contrastive Loss**: A symmetric contrastive loss (InfoNCE) is calculated between the normalized projected image and text embeddings using a temperature parameter.

2.1.5. Training Objective

The final loss function during training is a weighted sum of the standard language modeling (cross-entropy) loss from the GPT-2 decoder and the contrastive loss: $L_{total} = L_{LM} + w \times L_{Contrastive}$, where w was set to 1 based on experimentation.

2.2. Part B: Occlusion Robustness Analysis

The occlusion robustness analysis investigates how different image captioning models perform when parts of the input images are systematically obscured. This evaluation helps understand model resilience to partial visual information, which is critical for real-world applications where images may be partially occluded, corrupted, or incomplete.

2.2.1. Occlusion Methodology

Patch-wise occlusion was implemented by dividing each image into 16×16 pixel patches and randomly masking a specified percentage of these patches. Four occlusion levels were tested: 0% (original image), 10% (minor occlusion), 50% (moderate occlusion), and 80% (severe occlusion). The masked patches were set to zero values (black pixels), simulating the complete absence of visual information in those regions.

2.2.2. Models Evaluated

Two models were systematically compared:

- **Model A (SmolVLM)**: A pre-trained vision-language model designed for instruction-following image captioning tasks.
- **Model B (Custom Model)**: Our CLIP-ViT encoder + GPT2 decoder with contrastive learning as described in Section 2.1.

2.2.3. Evaluation Metrics

Model performance was evaluated using multiple complementary metrics:

- **BLEU**: Measures n-gram precision between generated and reference captions
- **ROUGE-L**: Assesses the longest common subsequence between generated and reference captions
- **METEOR**: Evaluates semantic similarity by considering synonyms and stemmed matches

For each model and metric, we calculated both absolute scores and relative performance degradation compared to the 0% occlusion baseline.

2.2.4. Comparative Analysis Framework

The analysis framework quantifies each model's robustness by measuring how rapidly performance deteriorates with increasing occlusion. This comparative approach identifies which architectural elements contribute to occlusion resilience in image captioning systems.

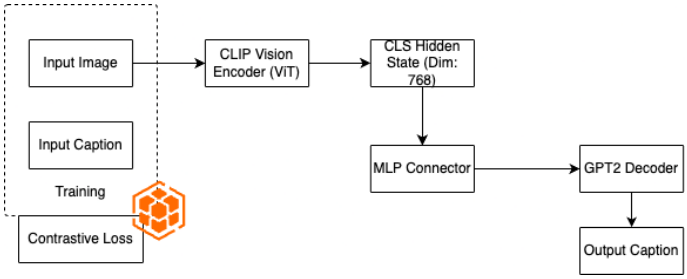


Figure 1. Architecture of the Final Custom Image Captioning Model.

2.3. Part C: Model Identification Classifier

A classifier was built to determine whether a given caption was generated by SmolVLM ('Model A') or the Custom Model ('Model B'), considering the occlusion context.

2.3.1. Input Representation

The input provided to the BERT classifier is a single text sequence. This sequence is created by joining the original ground truth caption, the model-generated caption, and the numerical occlusion percentage (treated as text). These three components are separated by the BERT special separator token, `[SEP]`. For example: "Original text `[SEP]` Generated text `[SEP]` 50".

2.3.2. Classifier Architecture

A pre-trained BERT model ('google-bert/bert-base-uncased') is used as the text encoder. The pooled output (corresponding to the '[CLS]' token) from BERT is fed into a classification head consisting of: Dropout (0.1) → Linear(768 → 256) → ReLU → Dropout (0.1) → Linear(256 → 2). The final layer outputs logits for the two classes (Model A vs. Model B).

2.3.3. Training Data and Split

The training data was constructed using the outputs from Part B (original captions, generated captions for both models at 0%, 10%, 50%, 80% occlusion, and the corresponding model label/occlusion level). This dataset was split into training (70%), validation (10%), and test (20%) sets, ensuring no overlap of original images between splits.

3. Results and Analysis

Evaluation was performed on the provided test set.

3.1. Part A: Baseline Comparison (0% Occlusion)

Table 1 compares the performance of the **final chosen Custom Model** (CLIP+GPT2+Contrastive, trained with the selected freezing strategy) against the zero-shot SmolVLM baseline on the original test images.

Table 1. Part A: Performance on Test Set (0% Occlusion)

Model	BLEU-4	ROUGE-L	METEOR
SmolVLM (Zero-Shot)	0.034	0.206	0.200
Custom Model (Final)	0.043	0.277	0.215

Analysis: From Table 1 we get that the custom model with CLIP encoder + GPT-2 decoder with contrastive loss consistently outperforms SmolVLM's inference performance on all the metrics. This will be further enchanced when testing with occlusion in the next section.

3.2. Ablation Study: Progressive Unfreezing vs. Frozen Weights

To determine the optimal training strategy for the custom CLIP-ViT + GPT-2 model, we compared two approaches:

1. **Frozen Weights (Final Custom Model):** Keeping the CLIP ViT encoder frozen and fine-tuning only the last two layers of the GPT-2 decoder (along with cross-attention and connector layers). This run corresponds to `experiment_image_captioning_run` (Blue lines in plots).
2. **Progressive Unfreezing:** A strategy where layers of the ViT encoder and/or more layers of the GPT-2 decoder are gradually unfrozen during training . This run corresponds to `experiment_progressive_unfreezing_run` (Red lines in plots).

Both runs utilized the same contrastive learning objective.

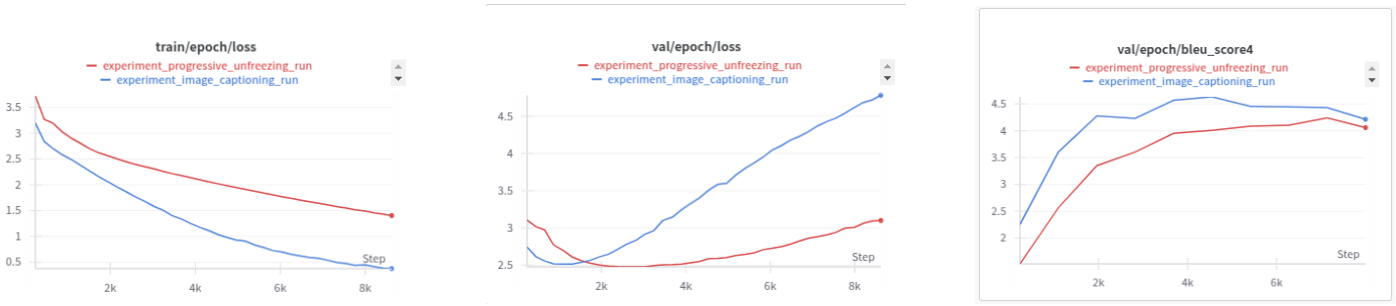
Analysis of Training Dynamics (Fig. 2):

- **Loss:** The progressive unfreezing strategy (Red) exhibited slower initial decrease but reached a potentially lower final training loss. The validation loss for progressive unfreezing showed a pattern of increasing after an initial dip, suggesting overfitting at later epochs. The frozen weight strategy (Blue) showed more stable validation loss with a faster convergence but also eventually overfitted.

Qualitative Analysis (Fig. 3): Comparing the generated captions for the example image at different training stages provides further insight.

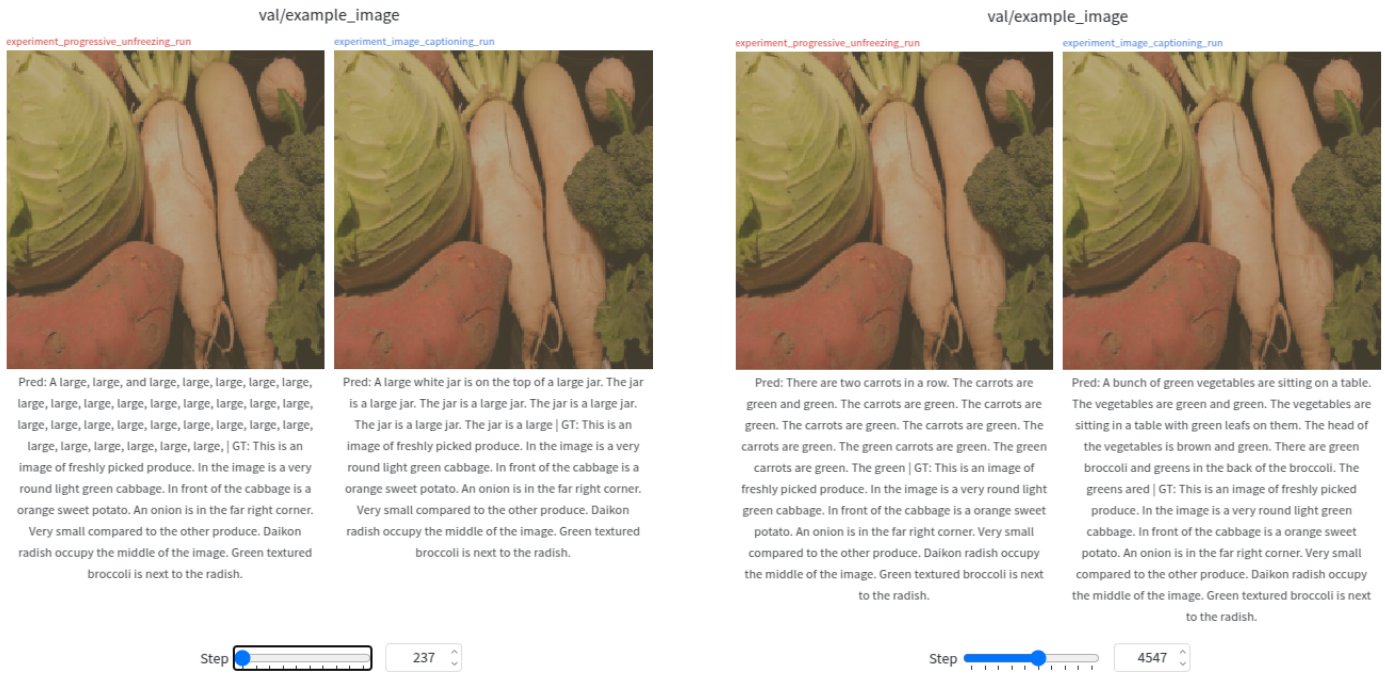
- **Early Stage (Fig. 3a):** At an early step (e.g., 237), the progressive unfreezing strategy produced a very repetitive caption. This is probably due to catastrophic forgetting encountered by the encoder and decoder models as soon as we unfreeze them. The frozen weights strategy generated slightly more structured but still basic caption, utilizing the pretrained knowledge.
- **Later Stage (Fig. 3b):** Towards the end of training (e.g., step 4547), the progressive unfreezing strategy yielded a still repetitive caption indicating that the model may never finetune over the small dataset we have, if we unfreeze completely. In contrast, the frozen weights strategy (our final model) produced a very coherent and relevant caption better aligned with the ground truth.

Ablation Conclusion: Based on the better validation BLEU scores, more stable validation loss, better qualitative caption generation, avoidance of overfitting observed in the frozen run, the strategy of keeping the majority of CLIP ViT frozen and fine-tuning only the later layers of the GPT-2 decoder was selected as the final custom model configuration for subsequent analysis in Parts A and B. The



(a) Training Loss **(b)** Validation Loss **(c)** Validation BLEU-4

Figure 2. Ablation Study: Training/Validation Loss and Validation BLEU-4 score comparison between Frozen Weights (Blue) and Progressive Unfreezing (Red).



(a) Early Training Step (e.g., Step 237) **(b)** Later Training Step (e.g., Step 4547)

Figure 3. Qualitative Example Inference Comparison at different training stages. Left side of each image: Progressive Unfreezing (Red); Right side: Frozen Weights (Blue).

contrastive loss was deemed beneficial over Cross-Entropy loss as it makes much more sense to bring together semantically similar captions and separate dissimilar ones. Moreover, the very idea of using CLIP-ViT is that the embedding space is grounded to both language and vision since it was jointly trained with the text encoder of CLIP.

3.3. Part B: Robustness Under Occlusion

The robustness of the final Custom Model and SmolVLM was assessed by evaluating performance on test images with 10%, 50%, and 80% random patch occlusion. Table 2 presents the absolute scores achieved under these conditions.

Table 2. Part B: Absolute Performance Scores Under Occlusion

Model	Occlusion	BLEU-4	ROUGE-L	METEOR
SmolVLM	10%	0.014	0.188	0.126
SmolVLM	50%	0.010	0.170	0.11
SmolVLM	80%	0.0036	0.137	0.088
Custom Model	10%	0.041	0.255	0.208
Custom Model	50%	0.031	0.230	0.187
Custom Model	80%	0.028	0.215	0.184

Analysis: From Table 2 we find that the occlusion going from 10 to 50 percent caused a significant drop in BLEU-4 of our model. However, the metrics still exceed all metrics of SmolVLM and the rest of the metrics seem pretty robust to occlusion.

3.4. Part C: Classifier Performance

The BERT-based classifier was trained to distinguish between captions generated by SmolVLM (Model A) and the Custom Model (Model B). Table 3 reports its performance on the held-out test set using macro-averaged metrics.

Table 3. Part C: Classifier Performance on Test Set

Metric (Macro Avg.)	Score
Precision	0.9973
Recall	0.9973
F1-Score	0.9974

Analysis: After just **3** training epochs over the dataset created from part B, we find that the BERT-based classifier can easily distinguish between the two models, as observed from the results in Table 3.

4. Conclusion

This project successfully implemented and evaluated a custom CLIP-GPT2 image captioning model with contrastive learning. An ablation study comparing frozen versus progressively unfrozen weights indicated that the frozen weight strategy yielded better performance/stabilityas compared to unfreezing. We hypothesize that this might be due to catastrophic forgetting by the pre-trained models and thus leading to bad convergence - however, a future direction would be to robustly test this hypothesis by experimenting over multiple seeds. This lead to the selection for the final model. The final custom model achieved performance exceeding the zero-shot SmolVLM baseline, particularly notable in the BLEU score. Robustness analysis in Part B revealed that both models degrade under occlusion, but the custom model showed greater resilience over all metrics and consistently outperformed SmolVLM. Finally, the Part C classifier demonstrated that the generated captions from the two models possess distinguishable features, allowing for successful classification with an F1-score of 0.9974. Future work could explore different encoder/decoder backbones, more advanced fusion techniques, or different robustness perturbations.