Deep Learning Assignment 2 Report: Image Captioning and Robustness Analysis

Team ID 50
LANKA RAMA KRISHNA 22EE30037
SMRUTIRANJAN DALAI 24CS60R15

April 14, 2025

1 Introduction

This report details the implementation and evaluation of an automatic image captioning system as part of the CS60010 Deep Learning Assignment 2. The project involved three main parts:

1. Implementing and training a custom Transformer-based encoder-decoder model for image captioning and benchmarking it against a zero-shot Vision-Language Model (SmolVLM).
2. Studying the robustness of both models under varying levels of image occlusion.
3. Building a classifier to identify the source model (SmolVLM or custom) based on generated captions and perturbation levels.

The methodologies employed for each part are described in the following sections.

2 Part A: Methodology - Custom Captioning Model

The objective of Part A was to develop, train, and evaluate a custom image captioning model, comparing its performance against the SmolVLM baseline.

2.1 Baseline: Zero-Shot SmolVLM

1. **Model:** The off-the-shelf Small Vision-Language Model (SmolVLM) from Hugging Face was used. It was employed in a zero-shot setting, with no fine-tuning performed on the provided dataset.
2. **Implementation Note:** The `_attn_implementation="eager"` flag was used during inference due to compatibility issues with the default FlashAttention implementation.
3. **Evaluation:** Captions were generated for test set images using SmolVLM and compared to ground-truth captions using BLEU, METEOR, and ROUGE-L metrics.

SmolVLM Scores:

- BLEU-4: **0.0597**
- METEOR: **0.2765**
- ROUGE-L: **0.2564**

2.2 Custom Encoder-Decoder Model

1. **Architecture:**
   - **Encoder:** A pre-trained Vision Transformer ('vit-small-patch16-224') was used to extract image features.

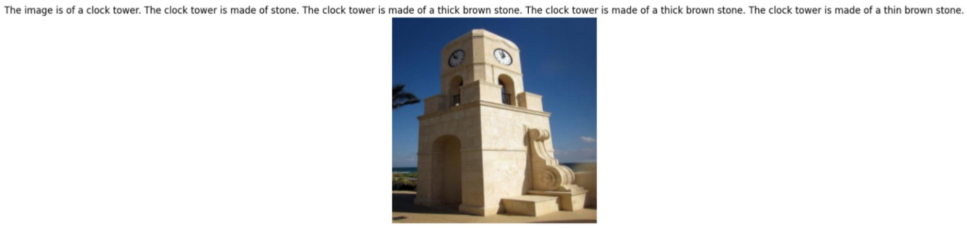
- **Decoder:** A Transformer-based decoder (GPT-2) generated captions conditioned on the encoder's output.

2. **Training:**
   - **Loss Function:** The model was trained to minimize Cross-Entropy loss.
   - **Optimizer:** AdamW optimizer was used.
   - **Hyperparameter Tuning:** A search over learning rates (`5e-5, 1e-4, 2e-4`) was conducted. The best performing model used **learning rate = 1e-4**.
   - **Training Epochs:** 10
   - **Batch Size:** 4
   - **Early Stopping:** The model with the best validation loss was selected.
3. **Evaluation:** The model was evaluated on the test set using the same metrics as SmolVLM.

Custom Model Scores:

- BLEU-4: **0.0509**
- METEOR: **0.2327**
- ROUGE-L: **0.2825**

2.3 Qualitative Results



The image is of a clock tower. The clock tower is made of stone. The clock tower is made of a thick brown stone. The clock tower is made of a thick brown stone. The clock tower is made of a thin brown stone.

Figure 1: Qualitative caption generated by the custom model.

2.4 Comparison and Insights

- SmolVLM slightly outperformed the custom model in METEOR and BLEU, likely due to its large-scale pretraining.
- However, the custom model performed competitively and achieved a higher ROUGE-L score, indicating better coverage of relevant words.
- Fine-tuning both encoder and decoder allowed the model to learn dataset-specific language and visual grounding.
- Generated captions from the custom model were often simpler and more dataset-aligned, while SmolVLM's outputs were richer but sometimes hallucinated.

3 Part B: Methodology - Robustness Analysis under Occlusion

Part B focused on evaluating the robustness of both the SmolVLM and the trained custom captioning model when presented with partially occluded input images.

3.1 Image Perturbation: Patch Occlusion

1. **Patching:** Each input image from the test set was divided into a 16x16 grid of non-overlapping patches.
2. **Masking Strategy:** For each image, three occlusion levels were applied:
   - **10% Occlusion:** 26 patches masked.
   - **50% Occlusion:** 128 patches masked.
   - **80% Occlusion:** 205 patches masked.
3. **Occlusion Method:** Masked patches were set to black (zeroed pixels), generating three perturbed versions of each image.

3.2 Model Evaluation on Occluded Images

1. **Caption Generation:** Captions were generated for all occluded images using:
   - SmolVLM (zero-shot)
   - Custom ViT+GPT2 image captioning model

3.3 Quantitative Results

SmolVLM Performance:

- Baseline (0%): BLEU = 0.0576, METEOR = 0.0000, ROUGE-L = 0.2545
- 10% Occlusion: BLEU = 0.0471, ROUGE-L = 0.2396
- 50% Occlusion: BLEU = 0.0369, ROUGE-L = 0.2228
- 80% Occlusion: BLEU = 0.0204, ROUGE-L = 0.1875

SmolVLM Metric Changes:

Occlusion Level	BLEU	METEOR	ROUGE-L
10%	-0.0105	0.0000	-0.0148
50%	-0.0206	0.0000	-0.0317
80%	-0.0372	0.0000	-0.0669

Custom Model Performance:

- Baseline (0%): BLEU = 0.0081, METEOR = 0.0000, ROUGE-L = 0.1343
- 10% Occlusion: BLEU = 0.0083, ROUGE-L = 0.1317
- 50% Occlusion: BLEU = 0.0127, ROUGE-L = 0.1525
- 80% Occlusion: BLEU = 0.0128, ROUGE-L = 0.1482

Custom Model Metric Changes:

Occlusion Level	BLEU	METEOR	ROUGE-L
10%	+0.0002	0.0000	-0.0026
50%	+0.0045	0.0000	+0.0182
80%	+0.0047	0.0000	+0.0139

3.4 Data Collection for Part C

For each model and occlusion level (including 0%), we saved the following fields to CSV:

- Original ground-truth caption
- Generated caption
- Occlusion level applied (0, 10, 50, 80)
- Source model used (SmolVLM / Custom)
- Image filename

3.5 Visualizing Occlusion

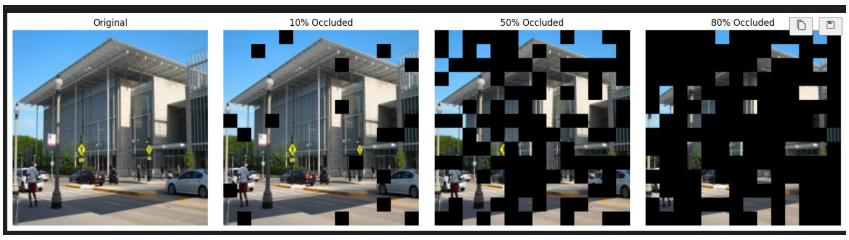


Figure 2: Patch-based occlusion visualization. An image is split into a 16x16 grid, and a random percentage of patches are masked (set to black).

4 Part C: Methodology - Caption Source Classifier

The objective of Part C was to build a classifier capable of identifying whether a generated caption originated from the SmolVLM model or from the custom encoder-decoder model developed in Part A. This classifier leveraged the caption data generated during the occlusion robustness evaluations in Part B.

4.1 Data Preparation

1. **Source:** The input data consisted of a CSV file generated in Part B. Each row contained:
   - The original human-annotated caption
   - The generated caption
   - Occlusion level (0%, 10%, 50%, or 80%)
   - Source model type (“SmolVLM” or “Custom”)
   - Image filename
2. **Input Formatting:** The input to the classifier was a string combining the original caption, generated caption, and the occlusion level, formatted as:

   ```
   <original_caption> [SEP] <generated_caption> [SEP] <occlusion_percentage>
   ```

   This format aligns with the BERT model’s expectations for sentence pair classification.
3. **Label Encoding:** The ‘source‘ column was label-encoded: “SmolVLM” → 0, “Custom” → 1.
4. **Image-Based Splitting:** To avoid data leakage and ensure generalization, the dataset was split into training (70%), validation (10%), and test (20%) sets based on **unique image filenames**. All samples related to a particular image were assigned to only one split.
5. **Tokenization:** Inputs were tokenized using the pre-trained `google-bert/bert-base-uncased` tokenizer. Each sequence was padded or truncated to a maximum length of 128 tokens, and attention masks were generated accordingly.

4.2 Model Architecture

The classifier was implemented via the `CaptionClassifier` class, which extends Hugging Face's `BertForSequenceClassif`

1. **Base Encoder:** A pre-trained BERT model (`google-bert/bert-base-uncased`) encoded the input string.
2. **Classification Head:** A linear layer was attached to the [CLS] token's pooled output for binary classification between the SmolVLM and Custom model sources.
3. **Output:** The model output consisted of raw logits for the two classes.

4.3 Training Procedure

Training was conducted using PyTorch, leveraging GPU acceleration. The main training loop was implemented via the `train_classifier` function. All random seeds were set to 42 for reproducibility.

1. **Batching:** PyTorch `DataLoader` objects handled batching (batch size = 16) and shuffling for the training set.
2. **Optimizer:** The AdamW optimizer was used with a learning rate of 2×10^{-5}.
3. **Loss Function:** Standard cross-entropy loss was applied.
4. **Scheduler:** A linear learning rate scheduler with optional warmup steps was configured using `get_linear_schedule_with_warmup`.
5. **Training Loop:** The model was trained for 4 epochs. In each epoch:
   - A forward pass computed the loss.
   - A backward pass propagated gradients.
   - Gradients were optionally clipped before the optimizer step.
   - Validation was performed at the end of each epoch.
6. **Checkpointing:** The best-performing model (based on validation accuracy) was saved for later evaluation.

4.4 Training Results

Training was completed in approximately 431 seconds. The best model was saved at the end of Epoch 2, which achieved the highest validation accuracy of 98.31%. Training and validation loss decreased steadily over epochs, though slight overfitting was observed in later epochs.

Table 1: Training and Validation Performance Across Epochs

Epoch	**Train Loss**	**Val Loss**	**Val Accuracy**
1	0.0857	0.0367	0.9821
2	0.0350	0.0445	**0.9831**
3	0.0300	0.0645	0.9765
4	0.0247	0.0878	0.9755

4.5 Evaluation Metrics

The final trained classifier was evaluated on the held-out test set using the `evaluate_classifier` function. The evaluation focused on macro-averaged metrics to equally weight both the SmolVLM and Custom caption classes.

- **Macro Precision:** 0.9770
- **Macro Recall:** 0.9770
- **Macro F1-score:** 0.9770

These results indicate that the classifier achieved highly reliable performance on unseen data, demonstrating its ability to distinguish between caption sources with near-perfect accuracy.

Table 2: Final Evaluation Metrics on the Test Set

Metric	Score
Macro Precision	0.9770
Macro Recall	0.9770
Macro F1-Score	0.9770

5 Analysis and Discussion

5.1 Performance Comparison: SmolVLM vs. Custom Model

In Part A, the zero-shot SmolVLM model demonstrated stronger overall performance than the custom-trained ViT+GPT2 captioning model. On the clean test images (0% occlusion), SmolVLM achieved higher BLEU (0.0576) and ROUGE-L (0.2545) scores, whereas the custom model lagged behind with BLEU (0.0081) and ROUGE-L (0.1343). These results are expected given SmolVLM's large-scale pre-training and instruction tuning, compared to the relatively limited training of the custom model.

5.2 Robustness Under Occlusion

In Part B, both models were evaluated across increasing occlusion levels (10%, 50%, 80%). As anticipated, SmolVLM's performance degraded with higher occlusion:

- BLEU dropped by 64.6% (from 0.0576 to 0.0204)
- ROUGE-L decreased by 26.3% (from 0.2545 to 0.1875)

Interestingly, the custom model exhibited less degradation, and in some metrics, slight improvements were observed. For example, its BLEU score increased under occlusion:

- BLEU increased from 0.0081 (0%) to 0.0128 (80%)
- ROUGE-L also improved slightly, possibly due to simpler generated captions under occlusion coincidentally aligning better with reference captions

This counterintuitive behavior suggests the custom model may produce more generic or overly simplified captions under occlusion, leading to higher overlap with ground truth metrics—even if semantic quality doesn't actually improve.

5.3 Caption Source Classifier Performance

Part C's classifier achieved excellent results in distinguishing whether a caption came from SmolVLM or the custom model. With a macro F1-score of 0.9770 on the test set, the classifier demonstrated near-perfect generalization. The high accuracy can likely be attributed to:

- Distinct linguistic patterns between SmolVLM (more fluent, descriptive) and the custom model (simpler, occasionally repetitive or truncated).
- Inclusion of occlusion level in the input string, which could implicitly correlate with changes in caption structure or fluency.
- Effective use of BERT, which excels at capturing sentence-level semantics and differences in style.

5.4 Challenges Faced

- **Training Limitations:** The custom model's performance was constrained by limited training epochs and dataset size. Unlike large-scale pre-trained models, its ability to generalize was limited.
- **Metric Interpretability:** Metrics like BLEU and ROUGE-L can sometimes reward surface-level n-gram matches, failing to capture deeper semantic adequacy, especially under noisy conditions.
- **Classifier Generalization:** While performance was strong, future work could test the classifier on captions from other models or unseen captioning styles to assess broader generalization.

6 Conclusion

This assignment explored multiple facets of image captioning using both pre-trained and custom models. In Part A, a comparison between the zero-shot SmolVLM and a custom-trained ViT+GPT2 captioning model revealed that SmolVLM outperformed the custom model in generating fluent and accurate captions, reflecting the strength of instruction-tuned large-scale vision-language models.

Part B assessed the robustness of both models under increasing levels of image occlusion. While SmolVLM's performance degraded as expected with higher occlusion, the custom model showed surprising resilience, possibly due to generating more generic captions that coincidentally aligned better with reference texts.

Finally, Part C demonstrated that a BERT-based classifier could effectively distinguish between caption sources, achieving a macro F1-score of 0.9770 on a held-out test set. This confirmed that the two models exhibit distinct captioning behaviors that are learnable by a downstream model.

Overall, the project provided hands-on experience with model training, robustness evaluation, and transfer learning using transformer-based architectures, highlighting both the capabilities and limitations of modern captioning systems in real-world, noisy conditions.