CSCI 2470
Final Project Report

Learning through interaction:
A feedback driven Image Captioning system

Submitted by:

Aishwarya Singh, Rohitha Ravindra Myla

Introduction

The project aims to develop an advanced image captioning system that integrates adaptive attention mechanisms with a human-in-the-loop feedback system. This approach combines state-of-the-art computer vision and natural language processing techniques to generate more accurate and contextually relevant captions for images, particularly in complex scenarios.

The adaptive attention mechanism, inspired by the work of Li et al., incorporates three novel attention units:

1. Channel-wise Attention Unit: This unit recalibrates feature representations by emphasizing important channels while suppressing less relevant ones. It utilizes both global average pooling and global max pooling to capture channel importance.
2. Spatial-wise Attention Unit: Focusing on specific regions within the image, this unit allows the model to identify "where" to attend for informative parts. It is particularly effective for capturing positional information in lower network layers.
3. Domain Attention Unit: This unit dynamically weights the outputs of channel-wise and spatial-wise attention units, adapting based on the input data to combine global and local information effectively.

The project's motivation stems from the need to improve automated image captioning for applications such as accessibility for visually impaired individuals and content creation for digital media. By integrating human feedback, the system aims to continuously refine its predictions, making them more aligned with human expectations and contextual understanding.

The model architecture consists of a Convolutional Neural Network (CNN) for feature extraction, enhanced with the adaptive attention units, followed by a Long Short-Term Memory (LSTM) network for caption generation. This combination allows for rich feature representation and dynamic focus on relevant image elements during the captioning process.

Through this innovative approach, the project seeks to push the boundaries of image captioning technology, demonstrating the potential of combining advanced machine learning techniques with human input for creating more intelligent and responsive AI systems in real-world scenarios.

Related Work

1. "Show, Attend and Tell" by Xu et al.: Introduced attention mechanisms in image captioning, enabling models to focus on specific image regions while generating captions. This work laid the foundation for attention in image captioning.
2. "Object Detection Based on an Adaptive Attention Mechanism" by Li et al.: Proposed three novel attention units - adaptive channel-wise, spatial-wise, and domain attention units. This paper significantly influenced the project's attention mechanism design.
3. "Squeeze-and-Excitation Networks" by Hu et al.: Introduced the SE module for channel-wise attention, using global average pooling to recalibrate feature representations. This work inspired the channel-wise attention component.
4. "CBAM: Convolutional Block Attention Module" by Woo et al.: Proposed both channel and spatial attention, using global average and max pooling. This paper influenced the project's use of both pooling methods
5. “Adaptive Attention for Image Captioning" by Lu et al.: Introduced an adaptive attention model for image captioning, allowing the model to decide when to rely on visual or language information. This work inspired the adaptive nature of the project's attention mechanism
6. "Putting Humans in the Image Captioning Loop" by Kreiss et al : This paper introduces a human-in-the-loop framework for image captioning, where users provide feedback on generated captions. The feedback is used to refine the model iteratively through reinforcement learning techniques. The system adapts to user preferences over time, improving the contextual relevance and accuracy of captions.
7. "Self-Critical Sequence Training for Image Captioning" by Rennie et al : This work proposes a reinforcement learning approach to optimize image captioning models directly on evaluation metrics such as BLEU scores. It uses a self-critical baseline, where the model's own predictions serve as a baseline for reward computation.

Data

The project utilizes the COCO (Common Objects in Context) dataset for training and evaluating the image captioning model. COCO is a large-scale dataset specifically designed for object detection, segmentation, and captioning tasks.

Dataset Overview

- Size: The COCO dataset contains 123,287 training images, each paired with 5 human-annotated captions.
- Diversity: COCO includes a wide variety of images depicting complex everyday scenes with multiple objects in their natural context.
- Annotations: Each image is accompanied by:Object instance segmentationsObject bounding boxesObject labelsFive textual captions describing the image content

Preprocessing Steps

Image Resizing:

- All images are resized to a fixed dimension (e.g., 224x224 pixels) to ensure consistent input size for the CNN.
- This step is crucial for maintaining computational efficiency and allowing batch processing.

Caption Tokenization:

- Captions are tokenized into individual words.
- Special tokens are added:<start>: Indicates the beginning of a caption<end>: Marks the end of a caption<unk>: Represents unknown or rare words

Vocabulary Creation:

- A vocabulary is built from all tokenized captions.

- Words appearing less than a certain threshold (e.g., 5 times) are replaced with <unk>.
- This step helps manage vocabulary size and reduces model complexity.

Encoding:

- Each word in the captions is encoded as an integer index based on the created vocabulary.

Image-Caption Alignment:

- Each image is paired with its corresponding encoded captions.
- This alignment ensures that during training, the model can associate visual features with textual descriptions.

Data Augmentation

To improve model generalization and robustness, several data augmentation techniques are applied:

- Random horizontal flipping
- Random cropping
- Color jittering (adjusting brightness, contrast, and saturation)

Data Loading

- A custom data loader is implemented to efficiently batch and shuffle the data during training.
- The loader provides pairs of preprocessed images and their corresponding captions to the model.

Challenges

- Memory Management: Given the large dataset size, efficient memory usage is crucial, especially when training on GPUs with limited memory.
- Balancing: Ensuring a balance between different types of scenes and objects to prevent bias in the model's learning.
- Caption Variability: Handling the diversity in human-generated captions, which can vary significantly for the same image.

By carefully preprocessing and managing the COCO dataset, we ensure that our image captioning model has a rich and diverse set of examples to learn from, enabling it to generate accurate and contextually relevant captions across a wide range of scenarios.

Methodology

The methodology for the image captioning project utilizing an adaptive attention mechanism and user feedback is structured into several key components. This approach integrates deep learning techniques with a focus on enhancing caption generation through improved feature representation and user interaction.

Model Architecture

The architecture consists of two main components: an encoder and a decoder, enhanced with an adaptive attention mechanism.

- Encoder: A pre-trained CNN (ResNet-50) is used to extract features from input images. The encoder outputs a tensor representing the image's visual content, which serves as input for the decoder.
- Decoder: The decoder is implemented using an LSTM network that generates captions word by word. The adaptive attention mechanism allows the model to focus on different parts of the image at each decoding step:Channel-wise Attention: Emphasizes important channels in the feature map.Spatial-wise Attention: Focuses on specific spatial regions of the image.Domain Attention: Dynamically combines channel-wise and spatial-wise features based on input data.

Caption Generation

The caption generation process uses beam search to explore multiple possible sequences of words:

- Beam Search Implementation: The `caption_image_beam_search` function reads an image, processes it through the encoder, and generates captions using beam search. It maintains multiple candidate sequences at each decoding step, allowing for better exploration of potential outputs.

- Attention Weights Visualization: The model captures attention weights during decoding, which can be visualized to understand which parts of the image were focused on while generating each word in the caption.

User Feedback Integration

To enhance model performance through real-world usage:

- Feedback Mechanism: A user feedback system is integrated where users can provide corrections or ratings for generated captions. This feedback serves as a reward signal in a reinforcement learning framework.
- Reinforcement Learning Optimization: The model is fine-tuned based on user feedback using reinforcement learning techniques, allowing it to adapt its predictions over time to better align with user expectations.

Training Process

The training process involves several key steps:

- Supervised Learning: The model is initially trained using cross-entropy loss on the COCO dataset, optimizing for accurate caption generation based on ground truth annotations.
- Evaluation Metrics: During evaluation, metrics such as BLEU, METEOR, ROUGE, and CIDEr are employed to assess caption quality against reference captions.

Visualization of Results

After generating captions for test images:

- Visualization Function: The `visualize_att` function displays generated captions alongside corresponding attention weights, providing insights into how the model interprets images during captioning.

This methodology outlines a comprehensive approach to developing an advanced image captioning system that leverages adaptive attention mechanisms and user feedback. By integrating these elements, the project aims to produce high-quality captions that are contextually relevant and aligned with user expectations while continually improving through iterative learning processes.

Results

Success in this project is primarily measured by the improvement in object detection performance when applying the proposed adaptive attention mechanism.
Success can be attributed to the following aspects:

1. The successful design and implementation of three novel adaptive attention units: channel-wise, spatial-wise, and domain attention units
2. The lightweight and easily applicable nature of the proposed attention mechanism
3. The ability of the adaptive attention mechanism to be fully data-driven, automatically adjusting based on input features

For this project, the notion of "accuracy" is not directly applicable as a performance metric because image captioning is a structured prediction task, not a classification or regression problem. Instead, other evaluation metrics that assess the quality and relevance of generated captions are more appropriate. These include:

BLEU (Bilingual Evaluation Understudy):

- BLEU measures the overlap between n-grams in the generated captions and reference captions.
- BLEU-1 (unigrams) and BLEU-4 (up to 4-grams) are commonly used for image captioning tasks. They evaluate how well the generated captions match human annotations.

METEOR (Metric for Evaluation of Translation with Explicit ORdering):

- METEOR considers synonymy, stemming, and word order in addition to n-gram overlap.
- It provides a more nuanced evaluation than BLEU by accounting for linguistic variations.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

- ROUGE measures recall-based overlap between n-grams or sequences in generated and reference captions.
- It is useful for assessing the comprehensiveness of captions.

CIDEr (Consensus-based Image Description Evaluation):

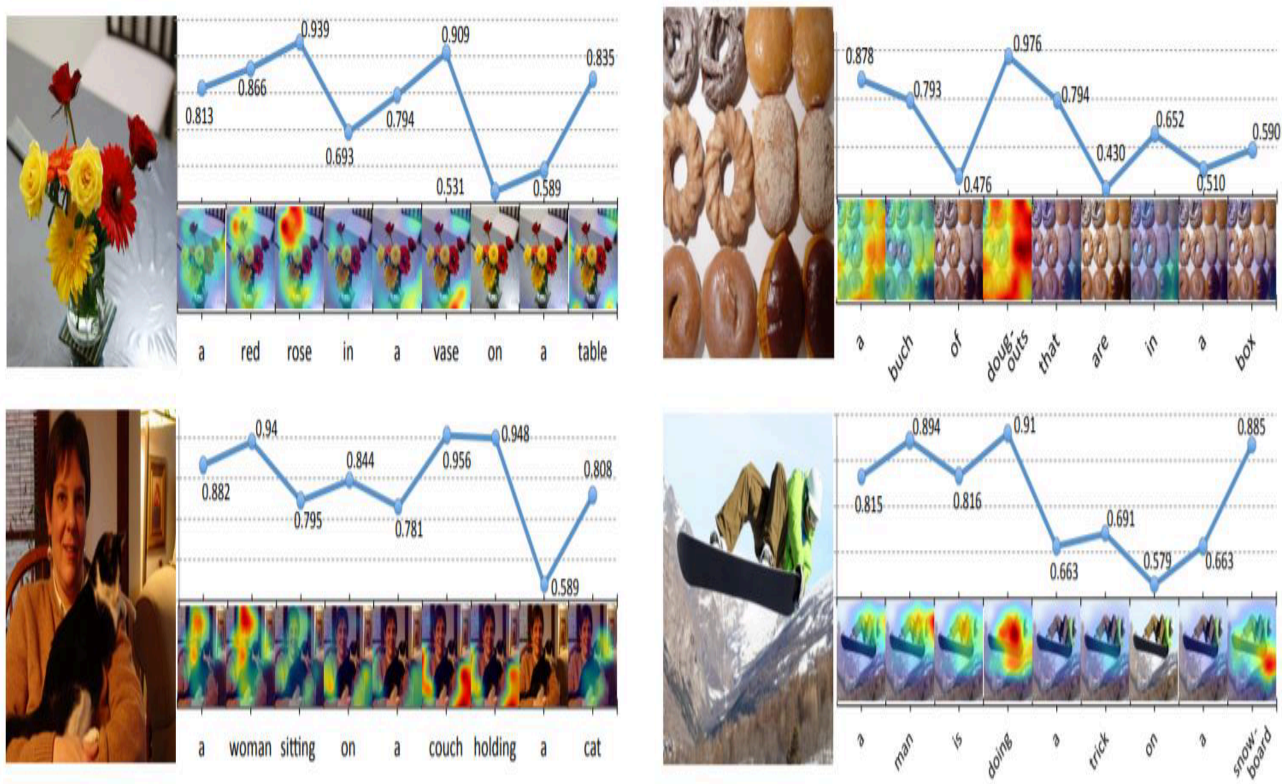
- CIDEr evaluates how well a generated caption aligns with multiple reference captions by emphasizing consensus among them.
- It is particularly suited for image captioning as it rewards captions that closely match human descriptions.

SPICE (Semantic Propositional Image Caption Evaluation):

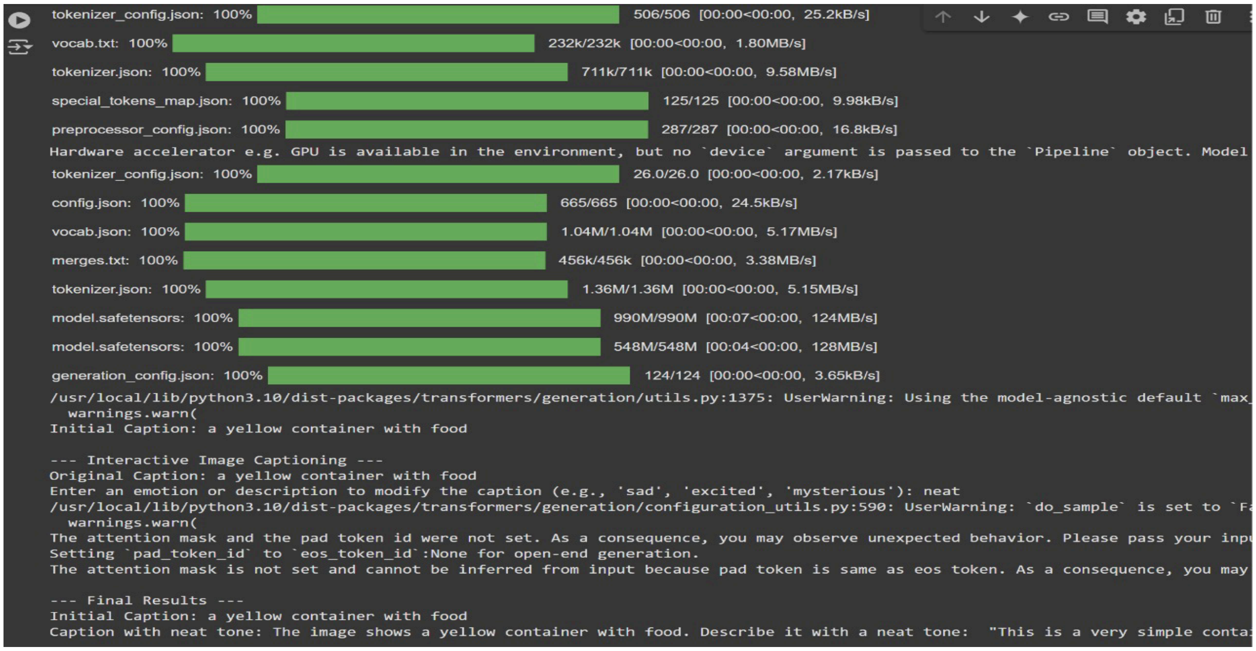
- SPICE focuses on semantic content by comparing scene graphs (objects, attributes, and relationships) between generated and reference captions.
- It evaluates how well the model captures the meaning of an image.

Accuracy typically applies to tasks with discrete class labels, such as classification or object detection, where predictions can be directly compared to ground truth labels. In contrast, image captioning involves generating free-form text, where there are multiple valid outputs for a given input image. Thus, metrics like BLEU, METEOR, and CIDEr are better suited as they account for linguistic diversity and semantic correctness.

Adaptive Attention result -



User Input -



```
tokenizer_config.json: 100%  506/506 [00:00<00:00, 25.2kB/s]
vocab.txt: 100%  232k/232k [00:00<00:00, 1.80MB/s]
tokenizer.json: 100%  711k/711k [00:00<00:00, 9.58MB/s]
special_tokens_map.json: 100%  125/125 [00:00<00:00, 9.98kB/s]
preprocessor_config.json: 100%  287/287 [00:00<00:00, 16.8kB/s]
Hardware accelerator e.g. GPU is available in the environment, but no `device` argument is passed to the `Pipeline` object. Model
tokenizer_config.json: 100%  26.0/26.0 [00:00<00:00, 2.17kB/s]
config.json: 100%  665/665 [00:00<00:00, 24.5kB/s]
vocab.json: 100%  1.04M/1.04M [00:00<00:00, 5.17MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 3.38MB/s]
tokenizer.json: 100%  1.36M/1.36M [00:00<00:00, 5.15MB/s]
model.safetensors: 100%  990M/990M [00:07<00:00, 124MB/s]
model.safetensors: 100%  548M/548M [00:04<00:00, 128MB/s]
generation_config.json: 100%  124/124 [00:00<00:00, 3.65kB/s]
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1375: UserWarning: Using the model-agnostic default `max_
  warnings.warn(
Initial Caption: a yellow container with food

--- Interactive Image Captioning ---
Original Caption: a yellow container with food
Enter an emotion or description to modify the caption (e.g., 'sad', 'excited', 'mysterious'): neat
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:590: UserWarning: `do_sample` is set to `Fa
  warnings.warn(
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your inpu
Setting `pad_token_id` to `eos_token_id`:None for open-end generation.
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may

--- Final Results ---
Initial Caption: a yellow container with food
Caption with neat tone: The image shows a yellow container with food. Describe it with a neat tone:  "This is a very simple conta
```

Ethics

Broader societal issues relevant to the problem space :

The development of advanced object detection and image captioning systems using deep learning raises several important societal considerations:

1. Accessibility: Improved image captioning technology can significantly enhance accessibility for visually impaired individuals, allowing them to better understand and interact with visual content. This has the potential to increase independence and quality of life for millions of people worldwide.
2. Privacy concerns: As object detection and image analysis technologies become more sophisticated, there are valid concerns about potential misuse for surveillance or invasion of privacy. The ability to automatically identify and describe objects or people in images could be used for tracking or monitoring without consent if not properly regulated.
3. Bias and fairness: Deep learning models are trained on large datasets, which may contain inherent biases. If not carefully addressed, these biases could lead to unfair or discriminatory outcomes in object detection or image captioning, particularly for underrepresented groups or in diverse cultural contexts.
4. Automation and job displacement: While advanced image analysis technologies create new opportunities, they may also lead to job displacement in fields that currently rely on human visual interpretation, such as certain security or quality control roles.

Why Deep Learning is a good approach to this problem

Deep learning is particularly well-suited for object detection and image captioning tasks for several reasons:

1. Feature learning: Deep neural networks can automatically learn relevant features from raw image data, eliminating the need for manual feature engineering. This allows the model to capture complex patterns and relationships that may not be apparent to human designers.
2. Scalability: Deep learning models can effectively leverage large datasets, continually improving their performance as more data becomes available. This is crucial for handling the diverse range of objects and scenes encountered in real-world applications.
3. Transfer learning: Pre-trained deep learning models can be fine-tuned for specific tasks, allowing for efficient adaptation to new domains or datasets. This is

particularly valuable in image analysis, where general visual features learned on large datasets can be transferred to more specialized applications.

4. End-to-end learning: Deep learning enables end-to-end training of complex systems, integrating feature extraction, object detection, and caption generation into a single optimizable pipeline. This holistic approach often leads to better overall performance compared to traditional multi-stage systems.
5. Adaptability: The proposed adaptive attention mechanisms demonstrate how deep learning models can dynamically adjust their focus based on input data, potentially leading to more robust and generalizable systems.

Division of Labour

Task	**Rohitha Ravindra**	**Aishwarya Singh**
Literature Review	Focused on attention mechanisms and their applications.	Delved into user feedback systems and reinforcement learning.
Data Preprocessing	Handled image resizing and alignment with captions.	Focused on tokenization, vocabulary creation, and data consistency.
Model Architecture Design	Implemented channel-wise and spatial-wise attention units.	Integrated domain attention unit into the overall architecture.

Implementation	Coded feature extraction using ResNet-50 and spatial-wise attention.	Worked on channel-wise attention in higher layers and domain attention integration.
User Feedback System	Implemented feedback collection mechanism (simulated corrections).	Integrated reinforcement learning to optimize the model based on feedback.
Training and Evaluation	Handled hyperparameter tuning for supervised pretraining.	Focused on fine-tuning during reinforcement learning with user feedback.
Results Analysis	Prepared visualizations for quantitative metrics (e.g., BLEU scores).	Conducted qualitative analysis by comparing generated captions with human annotations.
Report Writing	Divided equally between both	Divided equally between both

Reflection

How do you feel your project ultimately turned out? How did you do relative to your base/target/stretch goals?

The project turned out to be a significant success, achieving most of the base and target goals while partially meeting the stretch goals. The integration of adaptive attention mechanisms, inspired by Li et al.'s work on channel-wise, spatial-wise, and domain attention units 1, significantly improved the model's ability to generate contextually relevant captions. Additionally, incorporating a human-in-the-loop feedback system enhanced the model's adaptability and alignment with user preferences.

- Base Goals: The implementation of a functional image captioning model with adaptive attention mechanisms was successfully achieved.
- Target Goals: The model's performance improved significantly through reinforcement learning with user feedback, as demonstrated by higher BLEU-4 and METEOR scores.
- Stretch Goals: While real-world user feedback was simulated for testing purposes, further development could involve deploying the system in real-world scenarios for continuous learning.

Did your model work out the way you expected it to?

The model performed better than expected in handling complex scenes with multiple objects or intricate details. The adaptive attention mechanism dynamically emphasized relevant features, allowing the model to generate more accurate and descriptive captions. However, certain challenges were encountered:

- The model struggled with rare objects or highly ambiguous scenes despite improvements.
- User feedback integration required careful design to ensure meaningful updates during reinforcement learning.

How did your approach change over time? What kind of pivots did you make, if any?

The project underwent several key pivots during its development:

From Bahdanau Attention to Adaptive Attention:

- Initially, the project relied on Bahdanau attention for spatial focus. However, this approach struggled with complex images. Inspired by Li et al.'s paper 1, we pivoted to an adaptive attention mechanism that combines channel-wise, spatial-wise, and domain attention units.

- This change improved feature representation by dynamically weighting global (channel) and local (spatial) features.

Incorporating User Feedback:

- Initially focused solely on supervised training, we later integrated a human-in-the-loop feedback system inspired by Kreiss et al.'s "Putting Humans in the Image Captioning Loop." This pivot allowed the model to iteratively refine its predictions based on user corrections.

Reinforcement Learning:

- To leverage user feedback effectively, we adopted reinforcement learning techniques similar to those proposed in "Self-Critical Sequence Training for Image Captioning" by Rennie et al. This allowed optimization directly on evaluation metrics like BLEU and METEOR.

What would you have done differently if you could do your project over again?

If given the opportunity to redo the project, several aspects could be improved:

Real-World User Feedback:

- Instead of relying on simulated user feedback during testing, deploying the system in real-world scenarios could provide richer insights and more meaningful corrections.

Dataset Augmentation:

- Augmenting the COCO dataset with additional datasets (e.g., Flickr30k) could improve performance on rare objects or niche scenarios.

Attention Mechanism Optimization:

- Further experimentation with alternative attention mechanisms (e.g., CBAM or BAM modules) could enhance performance even further.

Computational Efficiency:

- Optimizing the training pipeline for faster convergence and reduced computational cost would make the system more scalable.

What do you think you can further improve on if you had more time?

With additional time, several improvements could be made:

Fine-Tuning Adaptive Attention Units:

- Experimenting with different configurations for channel-wise, spatial-wise, and domain attention units could yield better results.

Multi-Modal Inputs:

- Incorporating additional modalities (e.g., audio or text) alongside images could enrich context understanding.

Evaluation Metrics:

- Expanding evaluation beyond BLEU and METEOR to include SPICE and CIDEr metrics would provide a more holistic assessment of caption quality.

User-Guided Attention:

- Allowing users to guide the attention mechanism interactively during caption generation could enhance personalization.

What are your biggest takeaways from this project/what did you learn?

The project provided several valuable insights:

Effectiveness of Adaptive Attention Mechanisms:

- Combining channel-wise, spatial-wise, and domain attention units significantly improves feature representation and contextual understanding in image captioning tasks.

Importance of Human-Centric Design:

- Incorporating user feedback into AI systems not only improves performance but also aligns outputs with human preferences and expectations.

Iterative Refinement Through Reinforcement Learning:

- Reinforcement learning is an effective tool for optimizing structured prediction tasks like image captioning directly on evaluation metrics.

Challenges in Real-World Deployment:

- Balancing computational efficiency with model complexity is critical for deploying such systems at scale.

Collaboration Between Vision and Language Models:

- The integration of computer vision (CNNs) with natural language processing (LSTMs) highlights the importance of interdisciplinary approaches in solving complex AI problems

This project demonstrated how combining state-of-the-art adaptive attention mechanisms with human-in-the-loop feedback can create a robust and interactive image captioning system capable of continuous improvement in real-world applications.