STAT 5243 Project 3

**Abstract**

Image captioning is a challenging task at the intersection of computer vision and natural language processing, requiring a model to generate meaningful textual descriptions from visual inputs. Building on experience gained from earlier assignments using LSTM-based approaches on the Flickr8k dataset, this project explores a Transformer-based model—specifically a Vision-Encoder-Decoder architecture integrating a ViT encoder and a GPT-2 decoder. By fine-tuning a pre-trained ViT-GPT2 model on Flickr8k, the model can generate reasonably descriptive captions after limited training time, demonstrating the potential of transformer-based architectures for image-to-text tasks. Further training and parameter tuning would likely improve performance.

**Introduction**

Image captioning aims to produce a coherent, contextually appropriate textual description of an image, capturing objects, actions, and the overall scene. Traditional approaches typically relied on Convolutional Neural Networks (CNNs) to extract visual features and Recurrent Neural Networks (RNNs)—often Long Short-Term Memory (LSTM) networks—to model the sequential nature of language. While LSTMs can effectively handle short-range dependencies, they often struggle with long-range contextual reasoning and dynamically shifting linguistic patterns.

Transformer-based architectures, leveraging self-attention rather than recurrence, have recently revolutionized both language and vision tasks. Pre-trained vision-language models, such as those integrating a Vision Transformer (ViT) encoder with a GPT-2 decoder, enable large-scale feature learning and can generate text grounded in visual content. Although these models may demonstrate some zero-shot capabilities, the training objectives used in pre-training are not always equivalent to the user’s final objectives (training ≠ obj ≠ user object). In practice, fine-tuning such models on a specific dataset often outperforms zero-shot inference (0-shot < fine-tune), because the task-specific fine-tuning process better aligns the model with the user’s actual needs.

This project is inspired by my earlier coursework that implemented LSTM-based image captioning and incorporated beam search to improve the descriptive quality of captions produced for the Flickr8k dataset. With a deeper understanding of Transformer architectures, a natural question emerges: can a fine-tuned Transformer-based, pre-trained vision-text model produce more accurate, contextually rich captions that also meet user-defined goals than an LSTM-based approach?

**Methods**

**Dataset**: Flickr8k dataset is used, and this contains approximately 8,000 images, each with five captions. The dataset is split into training, validation (dev), and test sets. Images are diverse, depicting daily life scenes, animals, and people.

**Data Preprocessing**:

**Images**: Images are processed to a format compatible with the Vision Transformer (ViT) feature extractor. Each image is resized and normalized as required by the model’s pre-trained weights.

**Captions**: Captions are tokenized and a fixed sequence length (max length = 40 tokens). Special tokens such as <START> and <EOS> are included to mark sequence start and end.

**Model Architecture**:

The model to be fine-tuned is a pre-trained Vision-Encoder-Decoder model consisting of:

Vision Encoder (ViT): A Vision Transformer that encodes images into a set of latent features.

Text Decoder (GPT-2): A GPT-2 language model decoder that takes the encoder’s output and generates captions. The decoder uses self-attention and cross-attention layers to relate visual features and generated text tokens.

**Training Procedure**:

Hyperparameters: AdamW optimizes with a learning rate of 1e-4, batch size of 16, and train for a small number of epochs.

Loss Function: The model is trained with a standard cross-entropy loss between predicted tokens and ground-truth captions.

**Code Availability**

The code for data loading, model fine-tuning, and inference is available in a public GitHub repository: <https://colab.research.google.com/drive/1EJWmZmPDRU0zcEtaDeBcAN324qy4o8f1?usp=sharing>

**Results**

Before training, or before section 6, please jump to the last section, i.e. section7, for pre-trained model to see the result.A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Before conducting any fine-tuning, the pre-trained Vision-Encoder-Decoder model generated captions that were surprisingly accurate for certain test images. In this example, given an image depicting two dogs playing in the snow, the zero-shot model outputted: "Two dogs are playing in the snow". This caption effectively captured both the action (“playing”) and the setting (“in the snow”), as well as the number of dogs involved, demonstrating the strength of the pre-trained model’s generalization capabilities.

However, after performing just a single epoch of fine-tuning on the Flickr8k dataset, the generated output for the same image changed to: "A black and brown dog is running". In this case, not only did the caption fail to reflect the presence of two dogs, it also introduced a grammatical error and oversimplified the scenario. The loss of accuracy and fluency suggests that, with minimal training steps, the model might have begun to drift away from its initial general-purpose language and vision grounding. In other words, limited fine-tuning may have led the model to a less optimal local minimum, thereby degrading the quality of its captions rather than improving them.

**Discussion**

Although the zero-shot performance of a pre-trained model can be remarkably strong, additional training sessions of transformer could be better. This extended training would help the model better align with the dataset, ultimately achieving higher accuracy and more natural language generation.

However, it’s worth noting, that LSTMs can also perform well under these conditions. Since each sentence is relatively short, the issue of vanishing memory may have a smaller impact, making LSTMs less prone to significant errors. In contrast, the Transformer’s strengths become more apparent in tasks that involve longer texts or more complex structures, such as translating an entire movie script, where its ability to model long-range dependencies and nuanced attention patterns truly shines.

The reason for not training five epochs this time was largely practical. Previously, the training process seemed processing, leading to the early termination of runs. After introducing tqdm, it became clear that training was indeed progressing, though slowly. After taking more than half an hour, one epoch is done. It’s likely that five full epochs of fine-tuning would yield more ideal results. Past experience with LSTM projects suggests that increased training cycles often produce significant improvements in caption quality.

**Conclusion**

In summary, the zero-shot performance of the pre-trained Vision-Encoder-Decoder model on the Flickr8k dataset demonstrated promising descriptive capabilities even without task-specific fine-tuning. However, a small amount of suboptimal training did not improve the captions and, to some sense, diminished the model’s initial quality. This outcome underscores the importance of sufficiently fine-tuning. Given more training epochs, proper parameter tuning, and a more patient optimization process, the Transformer-based model would likely surpass its initial performance and consistently produce more accurate, fluent, and contextually appropriate image captions.