Image Caption Generator

Arun Madhusudhanan[1], **Tejaswini Dilip Deore**[1]

[1] Northeastern University, Boston
{madhusudhanan.a, deore.t} @northeastern.edu

Abstract

Describing an image automatically by generating meaningful captions is a fundamental problem in Artificial Intelligence. Recent advancements in Large Language Models have significantly improved the performance of this task. Previously the most common approach involved using a convolutional neural network (CNN) as encoder and a Recurrent neural network(RNN) as decoder for this task. In this project, we present an Image Caption Generator utilizing two distinct architectures: a convolutional neural network(CNN) encoder with long short-term memory(LSTM) decoder, and a transformer-based model using Vision Transformer (ViT) and Generative Pre-trained Transformer 2 (GPT-2). We compare performance of these models qualitatively and quantitatively. Noth architectures were evaluated using metrices like BLEU score, ROUGE score, METEOR score, and CIDEr score on Flickr8k dataset. Our findings indicate that both ViT-GPT2 model and CNN-LSTM model were able to generate meaningful and descriptive captions for images.

Introduction

Generating captions automatically requires understanding of both visual content of the image and natural language required to describe it. This ability of a model to generate captions describing the image content in natural language has many applications, like to empower visually impaired users by providing them better understanding of their surroundings, to enhance user experience on social media by streamlining the content, to give a better sense of surroundings for robot vision and navigation. However, this task requires understanding both the visual content of the image as well as the language required to describe it. In this project, we developed an image caption generator using state-of-the-art deep learning architectures. This project explores two distinct approaches: a hybrid model combining Convolutional Neural Networks (ResNet50 in this case) with Long Short-Term Memory(LSTM) network, and a transformer-based model with Vision Transformer(ViT) and Generative Pre-trained Transformer 2 (GPT-2).

Our primary objective is to evaluate and compare performance of these two architectures. Different metrices-BLEU(Papineni et al. 2002) score, ROUGE(Lin 2004) score, METEOR(Banerjee and Lavie 2005) score, and CIDEr(Vedantam, Lawrence Zitnick, and Parikh 2015) score provide a good estimate of how well the model is performing. Additionally, we conducted qualitative analysis by comparing model generated captions with reference captions from the dataset Flickr8k.

Background

Image captioning is a multi-modal task that requires integrating computer vision and natural language processing techniques. Early approaches to this problem often relied on template-based methods that combined object and scene recognition with rule-based language generation. However, these traditional methods were limited in their ability to generate diverse and fluent captions.

The advent of deep learning has led to significant advancements in image captioning. Vinyals et al.(Vinyals et al. 2015) proposed an encoder-decoder framework that used a CNN to encode the input image and an LSTM to generate the corresponding caption. The LSTM is a type of recurrent neural network (RNN) that is well-suited for modeling sequential data, such as natural language. It maintains an internal hidden state that allows it to selectively remember and forget relevant information as it processes the sequence. In our model, the LSTM takes the image features concatenated with the previously generated word as input, and predicts the next word in the caption sequence.

More recently, transformer-based models have emerged as a promising direction for image captioning. Huang et al. (Huang et al. 2019) introduced an image captioning model that used a vision transformer (ViT)(Dosovitskiy et al. 2020) to encode the visual input and a transformer-based language model to generate the captions.

Related Work

While previous studies have investigated the CNN-LSTM and transformer-based approaches independently, our work offers a direct comparison between the two architectures for the task of image captioning.

Vinyals et al. (Vinyals et al. 2015) pioneered the encoder-decoder framework for image captioning, demonstrating the effectiveness of using a CNN to encode the visual input and an LSTM to generate the captions. Subsequent works, such

as those by Xu et al. (Xu et al. 2015) and Rennie et al. (Rennie et al. 2017), built upon this foundation by introducing attention mechanisms and reinforcement learning techniques to further improve the captioning performance.

On the other hand, the transformer-based approaches, exemplified by the work of Huang et al. (Huang et al. 2019), have shown promising results by leveraging large-scale pretrained vision and language models. These methods capitalize on the powerful feature extraction and language modeling capabilities of transformers to generate more coherent and relevant captions.

Our study provides a comparative analysis of these two paradigms, shedding light on their relative strengths and limitations for the image captioning task. By evaluating the CNN-LSTM and ViT-GPT2 models on the same dataset and metrics, we aim to offer insights that can guide future research and development in this domain.

Project Description

Problem Formulation:

Image captioning is a task of generating a text describing contents of an image in natural language. Let I be the input image, which can be encoded as a set of visual features X = $\{x_1, x_2, ..., x_n\}$. The goal is to generate a sequence of words $S = \{w_1, w_2....., w_N\}$ that forms a grammatically correct sentence describing the context of an image I. We aim to find a model that could act as an approximation to function f shown in equation (1). The models considered are discussed in the next sections.

$$\{w_1, w_2....., w_N\} = f(\{x_1, x_2, ..., x_n\}) \tag{1}$$

Models

Image Caption Generator using CNN-LSTM: Our first approach towards image captioning utilized a convolutional neural network (CNN) for visual feature extraction coupled with a long short-term memory (LSTM) network for language modeling. Specifically, we employed the ResNet50 (He et al. 2016) architecture as the CNN component of the model.

The encoder- ResNet50 CNN (Figure: 1), takes a 224 $\times$ 224 pixel input image and outputs a 2048-dimensional feature vector encoding the visual content of the image. These visual features are then passed as input to the LSTM network, which is responsible for generating the corresponding caption word-by-word. Consider input image as I and its corresponding caption as $S = \{s_1, s_2 \ldots, s_T\}$, we can formulate this problem as shown in equations (2, 3, 4).
Where,

- CNN: convolutional neural network model
- W_{embd}: converts the input sentence to embedding based on vocabulary size before feeding to the LSTM network
- p_{t+1}: the probability of a specific word in vocabulary for captions

$$x_0 = CNN(I) \tag{2}$$

$$x_t = W_{emb} S_t, t \in 0, ...N-1 \tag{3}$$

$$p_{t+1} = LSTM(x_t), t \in 0, ...N-1 \tag{4}$$

During training, the LSTM is exposed to the ground truth caption sequences from the Flickr8K dataset, and it learns to maximize the likelihood of generating the correct words. The model is trained end-to-end using backpropagation through time, allowing the CNN and LSTM components to jointly optimize their parameters for the image captioning task.

At inference time, the model generates captions in an auto-regressive manner using either greedy search or beam search decoding.

- **Greedy search**: The model iteratively selects the word with the highest probability at each time step to generate the caption sequence. Given an input image I, the task is to generate a caption $S = \{s_1, s_2, \ldots, s_T\}$, where s_i represents the i-th word in the caption and T is the total number of words in the caption.

  At each time step t, greedy search selects the word s_t that maximizes the scoring function:

  $$s_t = \arg\max P(s_t | s_1, s_2, \ldots, s_{t-1}, I)$$

  where $P(s_t | s_1, s_2, \ldots, s_{t-1}, I)$ is the conditional probability of word s_t given the previously generated words and the input image I.

  The process continues iteratively until an end-of-sentence token or a maximum caption length is reached. This results in a sequence of words forming the generated caption. While simple and fast, this approach may lead to sub-optimal captions.

- **Beam search**: It is an approximation algorithm that explores the k-best hypotheses at each time step, where k is the beam width. This allows the model to consider multiple promising candidate sequences at once before committing to one final caption.

  Given an input image I, the task is to generate a caption $S = \{s_1, s_2, \ldots, s_T\}$, where s_i represents the i-th word in the caption and T is the total number of words in the caption.

  At each time step t, beam search selects the top-k candidate sequences, denoted as B_t^k, by expanding from the previously generated sequences B_{t-1}^k. The expansion step can be formulated as follows:

  $$B_t^k = Top{-}k \left(\bigcup_{b \in B_{t-1}^k} \{b \cdot s_t\} \right)$$

  where $\cdot$ represents concatenation, s_t denotes the set of possible next words at time step t, and k is the beam width.

  The conditional probability of a candidate sequence b at time step t is calculated as:

  $$P(b|I) = \prod_{i=1}^{t} P(s_i | s_1, s_2, \ldots, s_{i-1}, I)$$

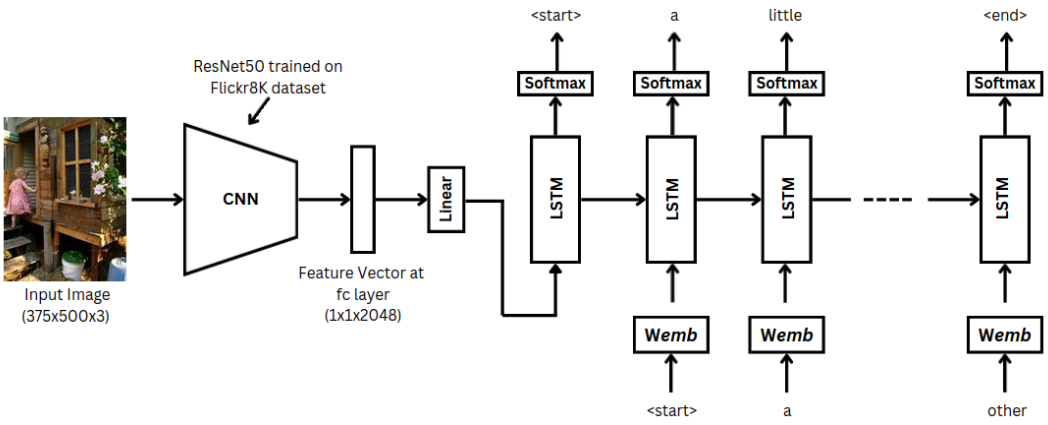

Figure 1: Image Caption Generator : CNN-LSTM model. This figure illustrates the architecture of an image caption generator model, which combines a ResNet50 convolutional neural network (CNN) for image feature extraction with a long short-term memory (LSTM) recurrent neural network (RNN) for generating captions. The CNN encodes the input image into fixed-size visual representations, while the LSTM decoder generates word-by-word captions based on the encoded image features.

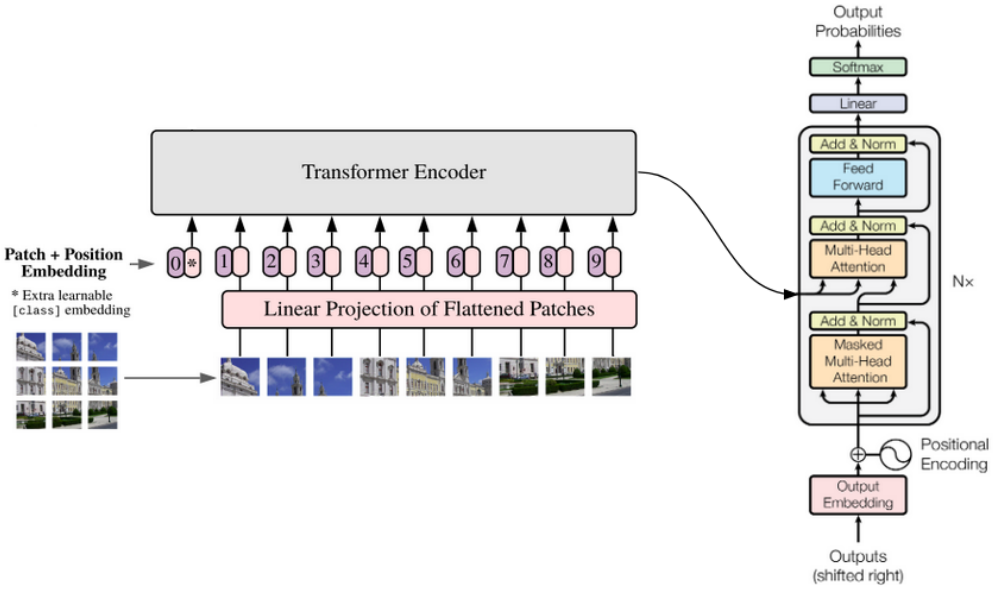

Figure 2: Image Caption Generator : ViT + GPT2 model. This figure depicts an image captioning model that merges a Vision Transformer (ViT) for image feature extraction with a Generative Pre-trained Transformer-2 (GPT2) model for generating captions. The ViT processes the input image to create fixed-size visual representations, while the GPT2 model generates captions based on these representations. By combining ViT's image understanding with GPT2's text generation capabilities, the model produces accurate and descriptive captions for diverse images.

The process continues recursively until it reaches a predefined maximum sequence length or when all beams end with an end-of-sentence token.

Once the maximum sequence length is reached or all beams end with an end-of-sentence token, the candidate sequence with the highest overall probability is selected as the final caption.

Beam search generally leads to better captioning performance compared to greedy search, at the cost of increased computational complexity.

In our experiments, we evaluated the CNN-LSTM model using both greedy search and beam search with k=3. The beam search approach yielded improved results over greedy search across all evaluation metrics.

ViT-GPT2 based Image Caption Generator: Our second approach to image captioning used transformer-based architectures for both the visual and language modeling components. Specifically, we utilized a vision transformer

(ViT) (Dosovitskiy et al. 2020) to encode the input image and a GPT-2 (Radford et al. 2019) language model to generate the captions as shown in Figure 2.

The ViT model divides the input image into a grid of 16 $\times$ 16 pixel patches, and linearly embeds each patch into a 768-dimensional vector representation. These visual tokens are then processed through multiple transformer encoder blocks, which apply self-attention and feed-forward operations to capture the spatial relationships and semantic content of the image.

The final ViT representation, a 768-dimensional vector encoding the entire image, is then passed as input to the GPT-2 language model. GPT-2 is a large-scale transformer-based language model that has been pre-trained on a vast corpus of text data to generate coherent and fluent natural language. Consider input image as I and its corresponding caption as $S = \{S_0, \ldots, S_N\}$, we can formulate this problem as shown in equations (5, 6, 7).
Where,

- ViT: visual transformer network model
- W_{embd}: converts the input sentence to embedding based on vocabulary size before feeding to the GPT2 network
- p_{t+1}: the probability of a specific word in vocabulary for captions

$$x_0 = ViT(I) \tag{5}$$

$$x_t = W_{emb} S_t, t \in 0, ...N-1 \tag{6}$$

$$p_{t+1} = GPT2(x_t), t \in 0, ..., sN-1 \tag{7}$$

During inference, the ViT encodes the input image, and the fine-tuned GPT-2 model generates the caption word-by-word in an auto-regressive manner, similar to the LSTM-based approach.

The key advantage of the ViT-GPT2 model is its ability to leverage large-scale pre-trained visual and language models, which have been shown to capture rich and representations that can be generalized. By fine-tuning these pre-trained models on the image captioning task, we can make use of their powerful feature extraction and language modeling capabilities to generate more accurate and diverse captions.

Experiments

Dataset

We selected the Flickr8k dataset due to its suitability for training deep learning models for image captioning given that we had limited computational resources. This dataset contains of 8000 images and each image in is paired with 5 descriptive captions annotated by human annotators.

We used 75% of our dataset for training our model and 25% of dataset for testing.

Evaluation Metrics

We evaluated both the CNN-LSTM and ViT-GPT2 models on the Flickr8K dataset using standard image captioning metrics, including BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). These metrics measure different aspects of the generated captions, such as n-gram precision, recall, semantic similarity, and consensus-based image description evaluation. In general, higher scores across these metrics indicate better captioning performance.

BLEU (Bilingual Evaluation Understudy) Score: BLEU is a metric that measures the n-gram overlap between the generated caption and the reference captions. It is calculated based on the precision of n-grams in the candidate caption against the n-grams in the reference captions. We report the BLEU-1 and BLEU-2 score, which measures the overlap of unigrams and bigrams.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score: ROUGE is a set of metrics used to evaluate automatic summarization and machine translation systems. We use the ROUGE-1 variant, which calculates the F-score based on the count of unigrams and ROUGE-L variant, which calculates the longest common subsequence-based F-score between the generated caption and reference captions. It captures sentence-level structure similarity.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) Score: METEOR is a metric that goes beyond just measuring unigram precision and recall. It calculates alignment between the generated caption and references based on exact, stem, synonym, and paraphrase matches between words and phrases. It aims to incorporate more linguistic intelligence into the evaluation.

CIDEr (Consensus-based Image Description Evaluation) Score: CIDEr computes the similarity between a generated caption and a set of reference captions by representing them as vectors of n-grams. The cosine similarity between these two vectors gives a score for the generated caption. CIDEr measures how well the generated captions match consensus in the human annotations.

Training Details:

CNN-LSTM Model: The CNN encodes the image into fixed-size visual representations, while the LSTM decoder generates word-by-word captions based on the input from the encoder.

1. **Encoder Training:**

   We performed transfer learning by using pre-trained ResNet50 weights. We added a linear layer to the end of the pre-trained ResNet50 and trained it on the Flickr8k dataset.

2. **Decoder Training:**

   We trained the LSTM decoder from scratch on the Flickr8k dataset. The training procedure can be broken down into the following steps: Given an image I and its true captions $S = \{w_1, w_2, \ldots, w_N\}$.

   (a) Extract the visual features $X = CNN(I)$ using the CNN encoder.

   (b) Initialize the LSTM hidden state with X.

   (c) At each time step t:

- Feed the previous ground truth word w_{t-1} and previous hidden state h_{t-1} to the LSTM.
- Compute the probability distribution over the vocabulary: $p(w_t|w_{1:t-1}, X) = LSTM(w_{t-1}, h_{t-1}, X)$.
- Compute the cross-entropy loss: $loss(t) = -\log p(w_t|w_{1:t-1}, X)$.

(d) Backpropagate the total loss $\sum_{t=1}^{N} loss(t)$ to update the LSTM parameters.

The training parameters for CNN-LSTM model used are listed in the Table 1.

Setting	Value
Epoch	30
Batch Size	32
Learning Rate	0.0003
Optimizer	Adam
Loss Function	Cross Entropy Loss

Table 1: Training parameters for CNN-LSTM model

ViT-GPT 2 Model: We performed fine-tuning of the Vision Transformer (ViT) by initializing it with pre-trained weights and then training it on the Flickr8k dataset. During this process, the pre-trained ViT model was adapted to the specific task of image feature extraction for caption generation. Similarly, we fine-tuned the Generative Pre-trained Transformer-2 (GPT-2) model on the Flickr8k dataset. By initializing the GPT-2 model with pre-trained weights and then training it further on the caption data, we enabled it to generate captions based on the encoded image features extracted by ViT. This fine-tuning process allowed the GPT-2 model to adapt its language generation capabilities to the specific context of image captioning. The pre-trained models were downloaded from Huggingface library. Due to limited availability of computational resources, we could fine tuned the model only for 3 epochs.

The training parameters used for ViT-GPT model are listed in the Table 2

Setting	Value
Epoch	3
Batch Size	4
Learning Rate	Default (Huggingface)
Optimizer	Adam
Loss Function	Cross Entropy Loss

Table 2: Training parameters for ViT-GPT2

Results

The models were trained using the chosen hyper parameters mentioned in Table 1 and 2. The loss curves were used to check if the model performance were improving while training. The loss curves obtained for CNN-LSTM model and ViT-GPT2 model are shown in figure 3 and 4. In the case of the CNN-LSTM model, both the training loss and validation loss consistently decreased over epochs, indicating improved performance. However, for the ViT-GPT2 model, while the training loss decreased steadily, the validation loss appeared to plateau, suggesting that the model might not be generalizing well to unseen data. It's possible that training the ViT-GPT2 model for more epochs could lead to a decrease in validation loss, indicating better generalization.

We performed comprehensive analysis on both models-CNN-LSTM and ViT-GPT2, and our analysis included both qualitative and quantitative comparison to gauge the characteristics of generated captions.

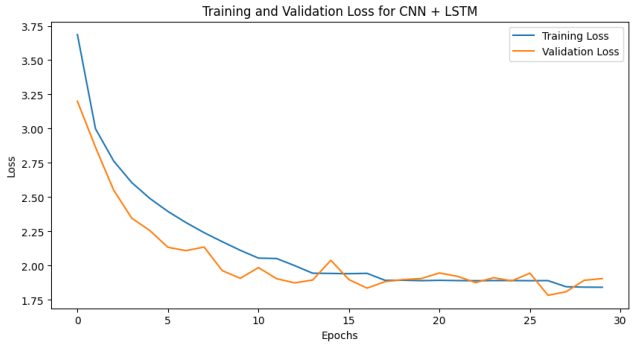

Figure 3: Training and validation loss for CNN-LSTM model

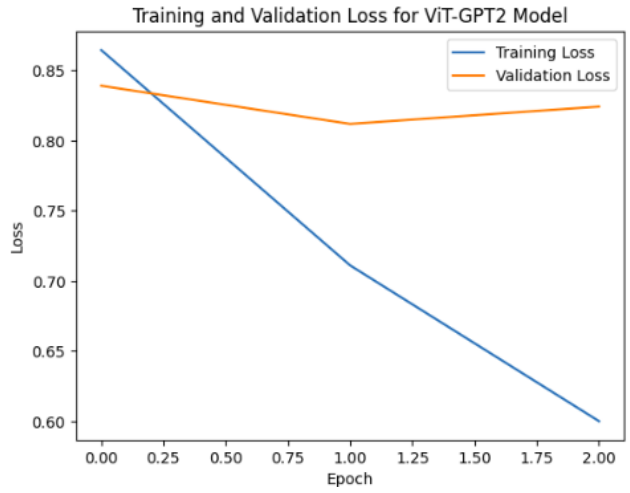

Figure 4: Training and validation loss for ViT-GPT2 model

Quantitative Analysis: We evaluated performance of our models based on the standard evaluation metrics explained in Evaluation Metrics section. Table 3 summarizes the quantitative results. The results for CNN-LSTM model were evaluated seperately using greedy search and beam search

From the results shown in Table 3, we can see that CNN-LSTM beam search performs better than both the CNN-LSTM greedy search and the ViT-GPT2 model across most evaluation metrics except CIDEr. The expectation was that the ViT-GPT2 model would outperform the CNN-LSTM

Model	BLEU-1	BLEU-2	ROUGE-1	ROUGE-L	METEOR	CIDEr
CNN + LSTM (Greedy Search)	36.397	8.203	28.998	26.02	23.575	7.165
CNN + LSTM (Beam Search with k=3)	**62.522**	**17.209**	**41.475**	**40.345**	**31.563**	6.037
Fine tuned ViT + GPT 2.0	41.109	13.76	36.11	34.05	31.11	**9.807**

Table 3: Quantitative Analysis: Performance measurement of models with evaluation metrics scores. The values highlighted in bold indicates the highest score obtained for its respective metric. CNN-LSTM Model with beam search appears to be outperforming other approaches except for CIDer.

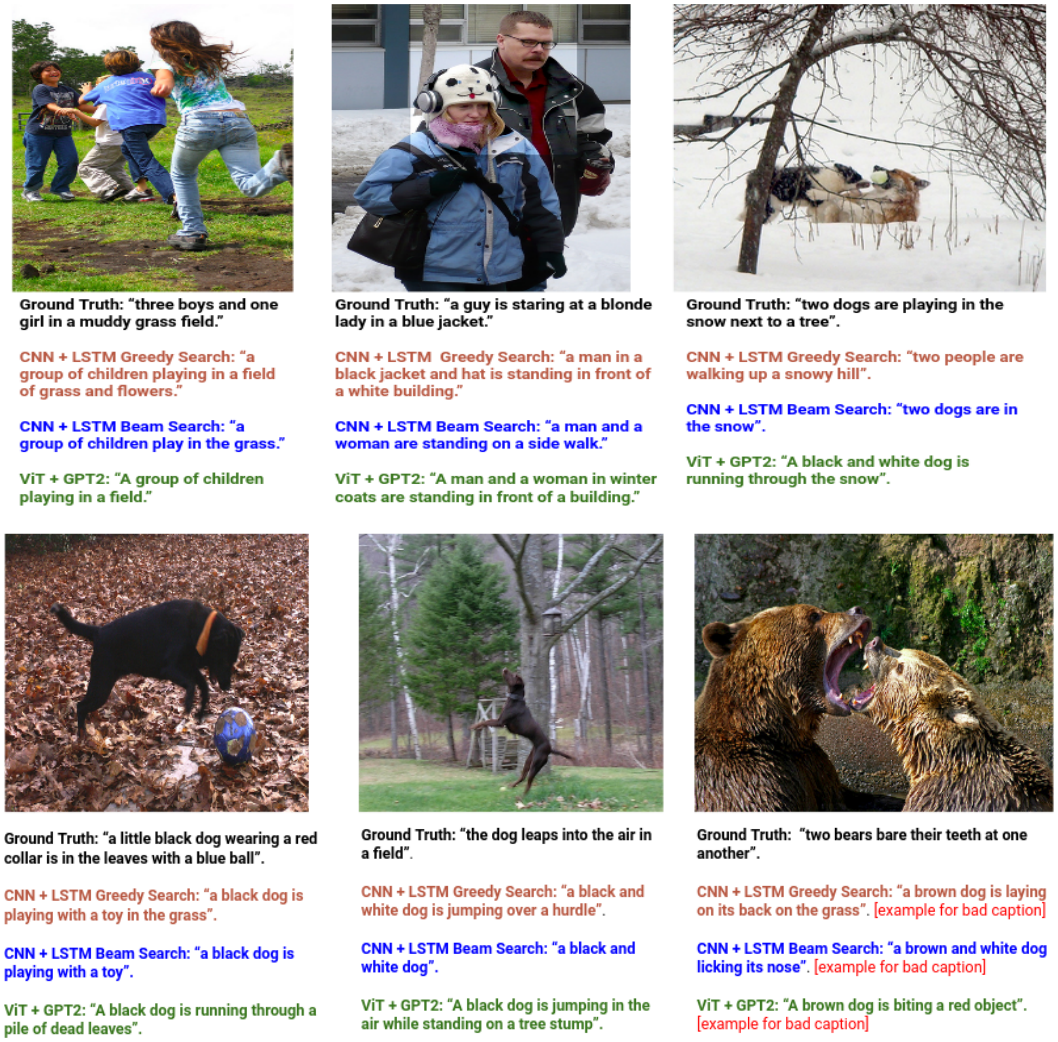

Figure 5: Qualitative Results: The figure displays a few images from the dataset. The caption in black color is true caption, one in red color is caption generated using CNN-LSTM(Greedy Search), one in blue color is caption generated using CNN-LSTM(Beam Search) and one in green color is generated by ViT-GPT2 model.

Model. This observation can be primarily attributed to two reasons.

1. Due to limited computational resources, we were only able to fine-tune the ViT-GPT2 model for a relatively small number of epochs - 3 epochs. Despite this constraint, the ViT-GPT2 model outperformed greedy search in all metrics, giving comparable results to CNN-LSTM beam search and outperforming it in terms of the CIDEr score.

2. The CIDEr score evaluates the similarity between the

generated caption and the reference caption and is specifically developed to evaluate the image captioning process. In contrast, other metrics such as BLEU, METEOR, and ROUGE are used in tasks such as machine translation, text summarization, and document understanding. This suggests that the ViT-GPT2 model generates captions that are more semantically similar to the reference captions. With further training and fine-tuning, we expect the ViT-GPT2 model to outperform the CNN-LSTM model.

Qualitative Analysis: In addition to quantitative analysis, we conducted qualitative analysis on the test dataset by comparing the generated captions from each model with the true captions. Some of the results are depicted in Figure 5.

The captions generated by both models were generally describing the contents although not always an exact match to the true captions. However, they conveyed meaningful content. For improving the performance of the CNN-LSTM model, we might consider adjusting hyperparameters such as the learning rate, dropout rate, or the size of the hidden layers. Additionally, fine-tuning the model architecture, such as incorporating attention mechanisms or experimenting with different types of recurrent units, could also be explored. To improve performance for the ViT-GPT2 model, we might be able to fine-tune the model on domain-specific data or perform additional pre-training on a larger and more diverse dataset. The captions generated using CNN-LSTM model with beam search and ViT-GPT2 model is found to be better than that predicted by CNN-LSTM model with greedy search.

Another observation was that CNN-LSTM beam search model appears to be generating smaller captions. This is a know issue with beam search and we could conisder fine tuning beverity penalty to make sure that model generates captions with sufficient lengths.

Despite only fine-tuning the GPT-ViT model for 3 epochs, its predictions were comparable to those of the CNN-LSTM models. This underscores the effectiveness of leveraging pre-trained models on extensive datasets and their ability to generalize to other datasets.

Each model also produced some inaccurate captions for certain images, as illustrated in Figure 5. This discrepancy might be due to the skewness in Flickr 8k dataset. There are more images images containing commonly occurring objects and scenes in the Flickr8k dataset, such as dogs, people, and outdoor scenes. Consequently, the models' performance is better on images with such objects and not as strong on others. One potential solution is to utilize larger datasets such as Flickr30K or MSCOCO dataset.

Acknowledgments

The successful completion of this project is attributed to the exceptional guidance and support of Professor Chris Amato. We also want to extend our gratitude to the Teaching Assistants whose invaluable assistance and support facilitated our progress throughout the course and project.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4634–4643.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. arXiv:1411.4555.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.