ViT - GPT2 Image Captioning Model

Jiadong HAO
haojd@umich.edu

Hanlong LIU
hanlongl@umich.edu

Chengcheng ZHANG
chengchz@umich.edu

1. Introduction

Image captioning has significant applications, such as helping visually impaired people understand visual content and automating image descriptions for social media platforms to improve accessibility. Traditional models like ”Show and Tell” [7], which uses Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), face limitations in capturing long-range dependencies and global context effectively.

To address the above problem, our project uses transformers, as their proved ability to model longer-distant relationships [6]. Specifically, we use Vision Transformer (ViT) [1] for visual feature extraction and GPT-2 [4] for caption generation. Our primary contributions are as follows.

1. Implement the proposed Vit-GPT2 image captioning model [3, 5, 2].
2. Train and tune two versions of our models (small and large) on datasets of four different scales.
3. Demonstrate that our best model outperforms the baseline model ”Show and Tell” in BLEU scores and BERT similarity.

2. Related work

Previous works in image captioning like “Show and Tell” have predominantly employed encoder-decoder architectures that combine Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [7]. While effective, these models exhibit limitations in modeling long-range dependencies and understanding global context, often leading to suboptimal performance.

Transformers have emerged as a promising solution to address these challenges. GPT-2 leverages extensive pretraining to generate contextually coherent and semantically rich captions [4]. ViT extends the capabilities of transformers to image processing, treating images as sequences of patches while only having minor modifications to the original transformer architecture [1].

3. Method

The ViT-GPT2 pipeline for image captioning is shown in Fig 1. The process begins by resizing input images to a fixed resolution (224 * 224) and dividing them into patches. These patches are flattened and passed through a pretrained ViT to extract visual feature embeddings.

Simultaneously, a pretrained GPT-2 text decoder takes in the tokenized “already-generated” caption and uses a self-attention mechanism to model dependencies within the text. Then, in the multi-head cross-attention component, the visual feature embeddings (last hidden states of ViTs) and the text embeddings generated by GPT-2’s self-attention will interact with each other, to predict next word while considering the image information.

4. Experiment

Initially, to validate our idea, we trained a small model, the ViT(base) + GPT2(Small), on a small dataset Flickr_8k, and found that our model had great unreleased power since the training loss and validation loss still showed a decreasing trend at the last epoch = 5. Hence, we built a much larger model, ViT(Large) + GPT2(Middle), and trained it on larger dataset like COCO_250k. The dataset and model briefs are shown in Table 1 and Table 2.

We evaluated our results using both quantitative and qualitative metrics. For quantitative evaluation, we use the BLEU (Bilingual Evaluation Understudy) score, which measures n-gram overlap between generated and reference captions. While effective, BLEU relies on exact word matching and may fail to capture the semantic similarity between sentences with different word choices. To address this, we introduced BERT similarity, which computes sentence-level semantic embeddings by averaging the last hidden states of BERT. The results shown in Table 3 demonstrates how our best model outperforms the ”Show and Tell”.

For qualitative evaluation, we generated captions for 50 images using our best model and evaluated the quality of the images. Some of the examples are shown in this link. We found 84% of them are very accurate, 14% are with minor errors, and 2% are invalid sentences.

5. Conclusion

This project demonstrates the effectiveness of ViT and GPT-2 for image captioning tasks, addressing the limitations of traditional CNN-RNN-based models. The most significant takeaway is the enhanced performance of our transformer-based models, evidenced by higher BLEU scores and improved BERT similarity metrics compared to the ”Show and Tell” baseline. Additionally, our approach proved to be computationally efficient, achieving superior results using fewer training epochs (5 epochs for us, 10 epochs for the baseline) and a fraction of the dataset size (40% of COCO for us and 100% COCO for the baseline).

Future work can focus on integrating more advanced visual feature extractors, such as hybrid ViT-CNN models, to further improve caption quality. Furthermore, applying reinforcement learning to optimize for human evaluation metrics, such as fluency and relevance, could enhance the practical applicability of the model.

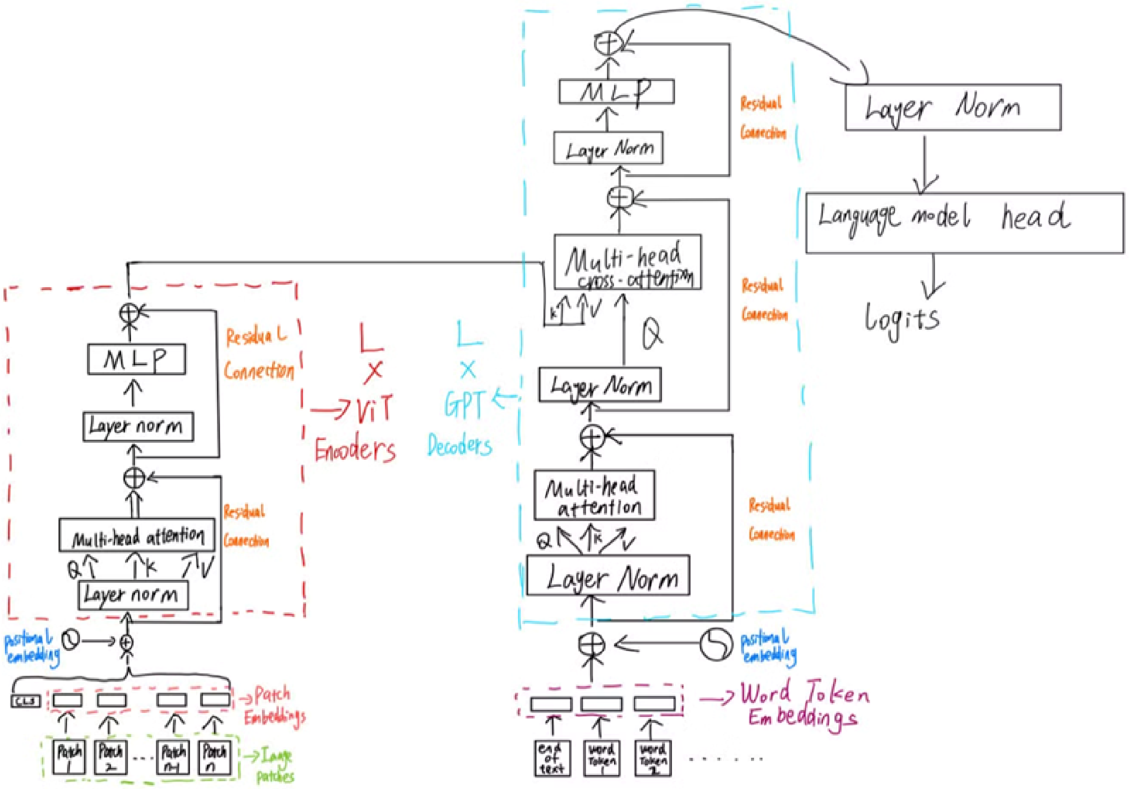

Figure 1: Vit-GPT2 Image Captioning Model.

Model	Trainable parameters	Embedding dim	Attention heads	Depth
ViT(Large) + GPT2(Middle)	758,933,504	1024	16	24
ViT(Base) + GPT2(Small)	238,603,776	768	12	12

Table 1: Model Hyperparameters.

Dataset	Training samples	Evaluation samples	Testing samples
COCO_250k	225,000	25,000	450
COCO_150k	135,000	15,000	450
Flickr_30k	27,000	3,000	450
Flickr_8k	7,200	800	450

Table 2: Dataset.

Dataset	Model	BLEU1	BLEU2	BLEU3	BLEU4	BERT similarity
COCO_250k	**ViT(Large) + GPT-2(Middle)**	**0.6480**	**0.5195**	**0.4333**	**0.3727**	**0.5550**
COCO_150k	ViT(Large) + GPT-2(Middle)	0.6401	0.5135	0.4281	0.3675	0.5506
COCO_150k	Vit(base) + GPT2(Small)	0.6196	0.4850	0.3980	0.3356	0.5433
Flickr_30k	Vit(base) + GPT2(Small)	0.5644	0.3979	0.2822	0.2097	0.2716
Flickr_8k	Vit(base) + GPT2(Small)	0.5311	0.3634	0.2541	0.1836	0.1729
COCO_Full	*Show and Tell (Baseline)*	*0.629*	*0.436*	*0.290*	*0.193*	*0.5223*

Table 3: Evaluation Results.

References

[1] Alexander Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[2] Burhanuddin Latsaheb. Image captioning with vit-gpt2. `https://www.kaggle.com/code/burhanuddinlatsaheb/image-captioning-vit-gpt2`, 2023. 1

[3] NLPConnect. Vit-gpt2 image captioning. `https://huggingface.co/nlpconnect/vit-gpt2-image-captioning`, 2023. 1

[4] Alec Radford, Jeff Wu, Rewon Child, et al. Language models are unsupervised multitask learners. *OpenAI*, 2019. 1

[5] Shreydan. Visiongpt2. `https://github.com/shreydan/VisionGPT2`, 2023. 1

[6] Kyuhong ShimWonyong Sung. A comparison of transformer, convolutional, and recurrent neural networks on phoneme recognition. *arXiv preprint arXiv:2210.00367*, 2022. 1

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1