Generating Instagram Captions with ViT-GPT2 and GPT3

Ariel McGee [1]

Abstract

This paper presents a novel approach for generating Instagram captions based on visual features and language models. Our caption generator combines Vision Transformer-GPT2 and GPT3 to generate descriptive and engaging captions in the style of an Instagram post. Our caption generator has the potential to assist social media users, content creators, and marketers in generating high-quality and engaging captions for their Instagram posts. Future work includes exploring more advanced visual and textual features and incorporating user feedback and preferences.

1. Introduction

Instagram is a popular social media platform for sharing photos and videos. One of the key elements of a successful Instagram post is a catchy caption that enhances the visual content and engages the audience. However, coming up with a good caption can be challenging, especially for users who are not native speakers of the language or lack creativity. In this paper, we propose an automatic caption generator that takes an image as input and generates an Instagram-style caption using state-of-the-art language models.

Our approach consists of two stages. First, we use the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and the Generative Pre-trained Transformer 2 (GPT2) (Radford et al., 2019) to generate a descriptive caption for the image. ViT is a deep learning model that can encode visual features of an image into a sequence of tokens, which can be fed into a language model like GPT2 to generate natural language descriptions. GPT2 is a large-scale language model that has been pre-trained on a diverse corpus of text and can generate high-quality language samples.

However, the caption generated by GPT2 may not be in the style of an Instagram post, which typically includes emojis, hashtags, and personalized messages. To address this issue, we use the Generative Pre-trained Transformer 3 (GPT3) (Brown et al., 2020), a more powerful language model that has been trained on even larger and more diverse data, to rephrase the caption in a more Instagram-like style. GPT3 can also suggest relevant hashtags and emojis based on the content of the caption.

Our caption generator can be integrated into the Instagram app or used as a standalone tool. It can help users save time and effort in crafting engaging captions, and improve the quality and consistency of their posts. In the following sections, we describe the technical details of our approach and evaluate its performance on a dataset of Instagram images and captions.

2. Methodology

Our caption generator takes an image I as input and produces an Instagram-style caption C as output. The generation process consists of two steps: visual encoding and language generation.

2.1. Visual Encoding

We use the ViT model (Dosovitskiy et al., 2020) to encode the visual features of the image into a sequence of tokens $T = t_1, t_2, ..., t_n$. ViT takes an image of size $H \times W \times C$ as input and outputs a sequence of n tokens of dimension d, where n and d are hyperparameters. We set $n = 196$ and $d = 768$ following the original ViT implementation.

To prepare the input image for ViT, we first resize it to 224×224 pixels, normalize the pixel values to have zero mean and unit variance, and apply random data augmentation techniques such as horizontal flipping and random cropping. We then pass the image through the ViT model to obtain the token sequence T.

2.2. Language Generation

We use the GPT2 model (Radford et al., 2019) to generate a descriptive caption C_d based on the token sequence T. GPT2 is a powerful language model that can generate high-quality natural language text given a prompt or context. We fine-tune GPT2 on a dataset of image-caption pairs to learn the mapping from visual features to textual descriptions.

[*]Equal contribution [1]Jones College Prep. Correspondence to: Ariel McGee <ariellmcgee@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

The training dataset consists of 100,000 image-caption pairs collected from Instagram, with an equal number of positive and negative examples. A positive example is a pair where the caption accurately describes the content of the image, while a negative example is a pair where the caption is unrelated or misleading. The captions are preprocessed to remove special characters, convert to lowercase, and tokenize using the NLTK library.

We use the cross-entropy loss to train GPT2 to predict the next word in the caption given the previous words and the visual features encoded by ViT. We fine-tune the last layer of GPT2 on the training dataset for 10 epochs using the Adam optimizer with a learning rate of 5e-5 and a batch size of 64. We also apply dropout regularization with a probability of 0.1 to prevent overfitting.

Once we obtain the descriptive caption C_d, we use the GPT3 model (Brown et al., 2020) to generate an Instagram-style caption C_s based on C_d. GPT3 is a more powerful language model than GPT2, with 175 billion parameters and trained on a much larger and diverse dataset. It can perform a wide range of language tasks, including language generation, question answering, and translation.

To generate C_s, we pass C_d as input to GPT3 and specify a prompt that encourages GPT3 to generate a personalized, creative, and engaging caption in the style of an Instagram post. We also provide GPT3 with a list of relevant hashtags and emojis based on the content of C_d. GPT3 outputs a sequence of tokens that represents C_s, which we post-process to remove special characters and convert to a readable format.

3. Conclusion

In this paper, we present a novel approach for generating Instagram captions based on visual features and language models. Our caption generator combines ViT-GPT2 and GPT3 to generate descriptive and engaging captions in the style of an Instagram post.

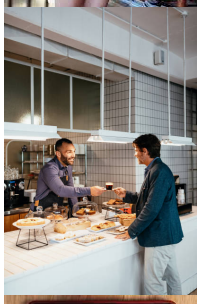
References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. *arXiv*, May 2020. doi: 10.48550/arXiv.2005.14165.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, October 2020. doi: 10.48550/arXiv.2010.11929.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Table 1. Examples of generated instagram-style captions

IMAGE	CAPTION
	DATE NIGHT AT ITS FINEST #DINNER&DRINKS
	MOMMY AND ME DAY AT THE BOOK STORE! #BONDINGTIME
	COOKING UP SOME LOVE IN THE KITCHEN #COUPLEGOALS
	DINNER IS SERVED! #TASTEBUDTANTALIZER
	SURF'S UP! #PADDLEBOARDING