Dog Captioning

Marina Samprovalaki f3322310
Konstantinos Kivotos f3322304

May 2024

Introduction

Combining the Vision Transformer (ViT) [2] and Generative Pre-trained Transformer 2 (GPT2) [4] architectures has introduced fresh possibilities in computer vision tasks. Image captioning, crucial for visual comprehension, has traditionally employed distinct models for feature extraction and caption generation. However, the ViT-GPT2 architecture by nlpconnect, presents a unified framework for both tasks, offering potential enhancements in efficiency and performance. Our model undergoes training and evaluation utilizing a carefully curated dataset comprising dog images paired with corresponding captions. The dog dataset can be found here.

Model Architecture

We employed two key components in our model: an encoder transformer,called Vision Transformer (ViT) [2], and a decoder transformer based on GPT2 [4].

Vision Transformer (ViT)

The vision transformer, ViT [2], was pre-trained on a substantial dataset consisting of 14 million images sourced from the ImageNet [1]. This pre-training process equipped ViT with foundational visual understanding capabilities. For a comprehensive overview of ViT's pre-training methodology and specifics, please refer to the provided documentation here.

Decoder Transformer: GPT2

Also, the decoder transformer component of our model was built upon GPT2 [4]. GPT2 underwent extensive training on a vast corpus of English language data, enabling it to predict subsequent words in sentences with remarkable accuracy. Further details regarding the pre-training procedures and nuances of GPT2 can be explored here.

Dataset

Our dataset comprises 25.1 thousand entries of dog images accompanied by captions. We partitioned 70% for the training set, 10% for validation, and 20% for the test set. This division ensures not only model training but also independent evaluation, aiming for more representative performance metrics.

Set	Number of Rows
Training	9,022
Validation	2,507
Test	1,003

Table 1: Number of rows for each set

Data Acquisition and Preprocessing

In this section, we outline the steps taken to acquire and preprocess the dataset for our project. After accessing the dataset from the Hugging Face site, we utilized git clone to download it. However, we encountered a challenge during the project since the dataset contained images stored as raw bytes within a dictionary structure, necessitating the extraction of image paths for code compatibility. To address this issue, we loaded the dataset into a dataframe and stored the images locally in a repository. Subsequently, we integrated both the saved images and their corresponding paths into the dataframe. Lastly, we downloaded the ViT-GPT2 model, as well as the requisite feature extractor and tokenizer components.
It's crucial to emphasize that unlike traditional tokenizers, GPT2 utilizes beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens, but lacks dedicated decoder start and padding tokens. To ensure compatibility with the GPT2 architecture, we made necessary modifications to the tokenizer. This included updating the tokenizer to incorporate the required tokens for seamless integration with the model.

Training and Evaluation

In this section, we detail the training process and evaluation metrics we used for our model.

Training Process

For the training phase, we adopted the default configuration of 3 epochs and a batch size of 2 images. The selection of this batch size was driven by resource constraints, aiming to optimize computational efficiency while training the model effectively. During training, we implemented a function to compute

metrics, utilizing Rouge Score [3] for evaluation. Rouge assesses the similarity of n-grams between actual and predicted captions, providing valuable insights into model performance.

Evaluation Process

During evaluation, we utilized the generate function provided by Hugging Face to generate captions, followed by computation of BertScore [5] and Rouge [3] score. These scores provide quantitative assessments of the model's captioning performance. The model's performance yielding scores of 59.69 for BertScore [5] and 22.74 for Rouge [3].

A depiction of the initial 10 images from the validation set, accompanied by their corresponding gold captions and our generated captions, is illustrated in the figure below.

Figure 1: 10 first images along with their gold and predicted captions from the validation set.

Inference

During inference, we leveraged the test set created at the beginning, ensuring the model remained unbiased as it hadn't encountered these samples during training or validation. Employing the same generation function utilized for the validation data, we produced captions for each image in the test set. Following caption generation, we evaluated the model's predictions against the gold standard captions. The resulting evaluation metrics are shown in the table above:

Metric	Score
BertScore	0,597%
Rouge	0,231%

Table 2: Metrics for the Test Set

Test Set Image Example

For this image, the generated caption is below:

Figure 2: Random picture from the test set

```
[{'generated_text': 'a woman walks her dog in the park'}]
```

Figure 3: Predicted caption for the image above

As observed, the generated text accurately describes the image.

Limitations

While our study has provided valuable insights into image captioning using the ViT-GPT2 architecture, it is essential to acknowledge certain limitations and

avenues for future exploration. One notable limitation is the constrained scope of our experimentation. Due to time and resource constraints, we were unable to explore various aspects fully. For example, with more resources, we could have potentially utilized a larger dataset, allowing for more comprehensive model training and potentially improved performance. Additionally, better hyperparameter tuning could have been achieved with more computational resources, leading to optimized model configurations. Furthermore, our evaluation metrics could benefit from comparison with baseline models commonly used in the field. Conducting such comparisons would offer a more comprehensive understanding of our model's performance relative to existing approaches. In essence, while our study has yielded promising results, there remain numerous unexplored avenues for enhancing the efficacy and robustness of image captioning models.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[3] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.