Image Captioning by ViT/BERT, ViT/GPT

Preprint · March 2023

CITATIONS
0

READS
2,640

2 authors, including:

Wing Man Casca Kwok
Northeastern University
4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Image Captioning ViT/BERT, ViT/GPT

Wing Man Casca, Kwok
Northeastern University
kwok.wi@northeastern.edu

***Abstract*—The objective of the project is to design and develop an advanced artificial intelligence image captioning system that is capable of generating captions for images or video frames without human involvement.**

This system incorporated the latest natural language processing and computer vision state of the art, to correlate the underlying patterns and relationship between visual and textual data. In our work, the system is trained on the Flickr8k dataset, the images and captions are encoded and concatenated with a vision transformer, followed by decoding the extracted features using BERT and GPT-2.

I. Introduction

The latest advances in multi-modal deep learning have demonstrated remarkable success, as seen, for example, from DallE's ability to generate photorealistic synthetic images with solely textual inputs. The breakthrough also benefits traditional industries of all sorts such as medical fields, to illustrate, medical imaging could now achieve a higher segmentation accuracy when a CNN is also trained with medical textual records when making predictions.

Therefore, we have come up with the idea in this research work, to build an image-to-text system that extracts semantic information from visual content. Our aim is to develop a deep learning model that can generate image captions to improve content retrieval efficiency. The system could apply to automatic sports commenting, audio book creation, image recognition and categorization, etc.

In our work, we have adopted an encoder and decoder approach, utilizing Vision Transformer (ViT) as the encoder to generate image embeddings, and compare the outcomes by 2 different decoders - BERT and GPT2.

Transformer and LSTM are closely related. They both employ attention mechanism to correlated element-wise relationship, but one distinct advantage of transformer-based is its ability to process sequences in parallel. Unlike traditional recurrent neural networks (RNNs) which process input sequences sequentially, transformers do not have any sequence constraints, allowing them to process all elements in the sequence simultaneously. This parallelization of the computation can result in significant speed-ups, making it possible to process longer sequences that would be otherwise computationally infeasible for traditional RNNs.

Transformer has emerged as the current state of the art such as BERT and GPT in not just natural language processing but also dominating the literature in computer vision. Vision transformer (ViT) was designed based on the foundation of BERT and was published in 2021. It has gained popularity due to the attention mechanism that enables the network to capture spatial information and correlations between image patches. In computer vision literature, image embeddings are generated by dividing the image into patches such as 16 x 16 pixels, and the attention mechanism provides additional benefits by attending the weights between local and global patches, enable the convolutional network to extract not only visual features but also spatial information. For example, without the transformer, despite we are still able to extract features such as eyes, noses and mouths on a face, we would not be able to capture correlations, such as the eyes being above the nose in a human face.

II. Related work

A. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*

The Visual Attention paper proposed an encoder and decoder architecture, implemented the convolutional network VGG as the encoder, and the attention-based model LSTM as the decoder for generating captions. The model generates one word at each timestamp based on the image embeddings, the previous hidden state, and the previously generated caption.

B. *ImageBERT: Cross-Modal Pre-training with large scale weak-supervised image-text data*

In contrast to Visual Attention, which predicts the language sequence conditioned on the image embedding, ImageBERT utilizes Faster-RCNN, a regional proposal network commonly used for object detection tasks in computer vision, to generate the image context token. The image embeddings are also masked before being fed to the BERT decoder.

Similar to Visual Attention, ImageBERT concatenates image embeddings and sentence as a single input into a BERT decoder. As a contrast to other BERT based framework, some approaches like ViBERT and LXMERT applies a single modal transformer to visual content and sentence respectively, and combined the two modalities in later stage.

III. Method

The Flickr8k dataset is a publicly available collection of 8,000 images, each with five captions describing the content of the image. It was designed for use in image captioning research.

For the captions of each image, they are usually in length around 10 words, such that they would not be over verbose. All images in the Flickr8k dataset were collected from Flickr, which is a commonly adopted website for photo sharing.

Within the dataset, there were a wide range of topic selections, including sports, animal, landscapes etc. Here is an example of 15 random selection of the dataset (Figure 1).

Fig. 1. Flickr8k dataset

In our work, we have adopted an encoder and decoder approach, utilizing Vision Transformer (ViT) as the encoder to generate image embeddings, and compare the outcomes by 2 different decoders - BERT and GPT2. All together there are 16 layers in the encoder and 16 layers in the decoder.

They key part is the cross-attention head which takes the key and value from encoder, and query from decoder embeddings to achieve the attention between image patches of which represents location in an image. At each training step, the model predicts sequence of words by comparing the ground truth caption.

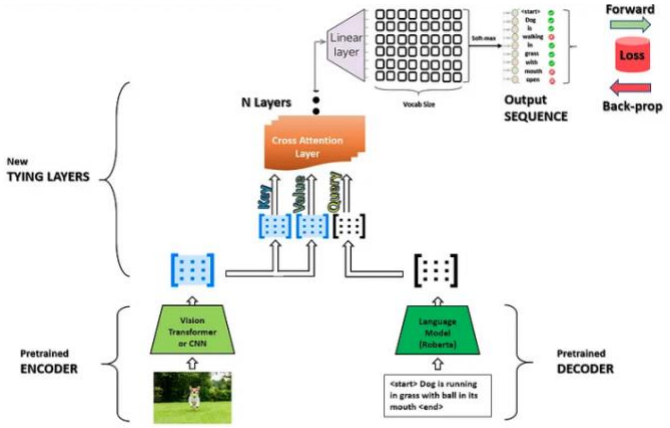

Fig. 2. Encoder decoder architecture(source: medium.com)

The network was fine-tuned from pre-trainings, with batch size 32, learning rate 5e-4, BERT with 7 epochs, GPT-2 with 5 epochs, image normalized by ImageNet mean and STD. Dataset is split into 85% training and 25% validation. Testing dataset comes from sources out of the dataset.

IV. RESULTS

Below shows the generated captions after training. These examples exhibit an impressive level of human-like fluency and naturalness with semantical meaningfulness.

The first 2 examples illustrate the validation result. The last 2 examples are testing images out of the dataset for inference.

Upon observing the results, we can see that the prompts generated by both BERT and GPT-2 align with the semantic meaning of the images. While BERT generates sequences that are within the defined sequence length, GPT-2 tends to produce longer texts due to its emphasis on prompt generation during training.

1) Validation Results 1

BERT : "a black dog and a black and white dog are playing on the street."

GPT-2: "A black dog and a white dog are running on the street"

Ground truth:

	image	caption
5	1001773457_577c3a7d70.jpg	A black dog and a spotted dog are fighting
6	1001773457_577c3a7d70.jpg	A black dog and a tri-colored dog playing with each other on the road .
7	1001773457_577c3a7d70.jpg	A black dog and a white dog with brown spots are staring at each other in the street .
8	1001773457_577c3a7d70.jpg	Two dogs of different breeds looking at each other on the road .
9	1001773457_577c3a7d70.jpg	Two dogs on pavement moving toward each other .

2) Validation Results 2

BERT: "a girl in a pink sweater is climbing up a set of stairs"

GPT-2: "A little girl in a pink dress is standing in front of a wooden fireplace. She is looking out into the woods"

	image	captio
0	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way
1	1000268201_693b08cb0e.jpg	A girl going into a wooden building
2	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse
3	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playhouse
4	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a wooden cabin

3) Inference Results 1

BERT: "a young woman wearing a white shirt and black sunglasses is smiling"

GPT-2: "A woman in a black shirt is standing in front of a white building. She is smiling. Another woman is standing next to her."

4) Inference Results 2

BERT: "a soccer player takes a shot of the player in the distance."

GPT-2: "A group of soccer players are playing a game on a field. One player has the ball. Another player in front of him. The goal is to kick it. The players on the left are trying to stop the player on the right."

BERT: "a soccer game in progress."

GPT-2: "A soccer player in a red uniform is chasing after a ball. The player in white is trying to catch it"

3) Model Performance

ROUGE-N measures the number of matching n-grams between the model-generated text and a human-produced reference. In this project, Rouge-1 is implemented to measure the ratio of the number of unigrams that the ground truth appears in model outputs.

Result 1: BERT, batch size = 64, sample = 6472, LR = 5e-4

Epoch	Training Loss	Validation Loss	Rouge1 Precision	Rouge1 Recall	Rouge1 Fmeasure
1	No log	2.672490	0.304400	0.314700	0.296700
2	3.145700	2.445663	0.345100	0.340900	0.330900
3	2.520800	2.273501	0.386200	0.383800	0.371900
4	2.177700	2.168446	0.385000	0.420000	0.387600
5	1.882800	2.101561	0.418700	0.421400	0.405600
6	1.555900	2.180405	0.434400	0.435200	0.419400
7	1.214900	2.398801	0.430600	0.426800	0.413300

Result 1: GPT-2, batch size = 64, sample = 6472, LR = 5e-4

Epoch	Training Loss	Validation Loss	Rouge1 Precision	Rouge1 Recall	Rouge1 Fmeasure
1	No log	2.782453	0.052800	0.508900	0.094700
2	2.990100	2.478712	0.069600	0.527500	0.121200
3	2.455500	2.348075	0.064200	0.576600	0.114200
4	1.909600	2.443358	0.069600	0.584500	0.122800
5	1.413500	2.765103	0.074400	0.598800	0.130700

Due to limited GPU availability, we have trained our models for a limited number of epochs. Each BERT training epoch took 1 hour, and GPT-2 required 2 hours.

In general, we observe that the rouge1 precision has kept on improving upon number of epochs for both BERT and GPT-2 decoders. We have noticed a significantly lower rouge1 precision from the GPT-2 results, in contrast from the generated captions, that the generated text matches with the image's semantic meanings.

GPT models are typically trained using a language modeling objective, where the model is trained to predict the next token in a sequence given the previous tokens. This training objective is useful for text generation tasks, but may not be as effective as MLM for generating high-quality text representations; it tends to generate long sequences, which could also impact the values of rouge-1 measurement.

V. Conculsions

In this project, we have achieved success in demonstrating the multi-modality competence between vision and language transformers through cross-attentions. We have explored the effectiveness of using BERT and GPT-2 as decoders in generating image captions, and have discovered that each model has its own unique strengths and weaknesses in this task.

Our results show that both BERT and GPT-2 are capable of generating captions that match the semantic meanings of the corresponding images. However, we have observed that GPT-2 tends to generate longer sequences, which may have an impact on the precision of the rouge-1 measurement.

Despite their differences, we believe that both models have the potential to be used in real-world applications of image captioning, with each model having its own set of advantages that can be leveraged depending on the specific task at hand. Our findings provide insights into the strengths and weaknesses of these models and can inform future research in this area.

VI. References

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[2] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

[3] ImageBERT: Cross-Modal Pre-training with large scale weak-supervised image-text data

[4] Flickr8k dataset.

[5] Two minutes NLP — Learn the ROUGE metric by examples
https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499

[6] Image Captioning Using Hugging Face Vision Encoder Decoder — Step 2 Step Guide (Part 2)
https://medium.com/@kalpeshmulye/image-captioning-using-hugging-face-vision-encoder-decoder-step-2-step-guide-part-2-95f64f6b73b9