Image Captioning and Generating text by GPT-2

Kayahan Kaya, Hashim Hashimov, Huseyn Garayev

Abstract

This report will discuss a project which is Image Captioning and generating text by GPT-2 model. As technology grows, advancements in Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP) are inevitable. Image Captioning became popular nowadays which assigns related caption based on the objects of the image. In this project, Image Captioning results are used to generate continuation sentence of the given image description with GPT-2.

1 Introduction

This project aims to work with Image Captioning algorithms and the GPT-2 model. Mainly, the algorithm should be able to generate a caption at first based on the input image. Then, the result caption will be input for GPT-2 which will create longer meaningful sentences. The main motivation behind this project was analyzing image captioning techniques and doing something interesting as GPT-2 results tend to be interesting and a continuation story of given image description.

Generally, based on our research, most approaches use LSTM architecture and Convolutional Neural Networks (CNN) for this kind of task. Section 2 (Related Work) will be discussed in detail.
Furthermore, in this project, LSTM and CNN (Inception v3 Pre-trained model) architecture had been used as well. Figure 1 shows an example result of image captioning part

Figure 1: Image Captioning Example

The current approach that had been used in this project generated accurate and good results.
In conclusion, we did image captioning and came to the decision that CNN and LSTM approach is efficient based on our research.
In the following sections, Related Work will be discussed in Section 2, Methodology in Section 3, Results and Evaluation in Section 4, Conclusion in Section 5.

2 Related Work

In this section, some related works similar to our topic will be summarized. Recently, Most of the researchers are focused on the task of image captioning. Most of the state-of-the-art approaches use deep learning techniques to solve this hard task. In this section, we will give a summary of articles that are related to our project topic. Our main goal is to get some intuition from previous work and creating our own way to solve this challenging task.
Novel researchers use deep-learning techniques to generate captions [2]. As an example paper [1] uses CNN for image information extraction and LSTM for caption generation, some researchers used image and caption datasets to train models however they are not accurate for unseen images in order to make image-specific caption.
Firstly, “Vision and Language: from Visual Perception to Content Creation by Tao Mei, Wei Zhang, Ting Yao” conducted research about Vision to Language and inverse. The vision-to-language part was similar and it shows different approaches. The typical architecture might be CNN encoder and LSTM decoder. However, there are several ways to extract information from the images:

- Directly taking outputs through layers
- Attention mechanism
- Extracting region level information
- and so on.

The point here is that results from image extraction will feed the LSTM decoder to generate accurate sentences. Paper used Reinforcement Learning to optimize LSTM decoders.
Additionally, the paper presents a Transformer-based encoder-decoder model which utilizes an attention mechanism.
Furthermore, Paper [2] introduces their own proposed model AICRL for automatic image captioning based on RestNet50 and LSTM with soft attention. It is also designed based on CNN and RNN [2]. The model is trained by using stochastic gradient descent as well.
From an attention mechanism perspective, as you can see, soft attention has been implemented as a new gate to the LSTM because there is no need to process the whole image [2].

In this paper [3], Authors propose a method of incorporating high-level concepts into the successful CNN-RNN approach, and show that it achieves a significant improvement on the state-of-the-art in both image captioning and visual question answering. They try to show that the same architecture can be used to incorporate external knowledge, which is critically essential for answering visual questions in general. Here is the summary of how their framework works;

Given an image, a CNN is first applied to produce the attribute-based representation V_{att}(5 most frequent words in image text). The internal textual representation is made up of image captions generated based on the image attributes. The hidden state of the caption-LSTM after it has generated the last word in each caption is used as its vector representation. These vectors are then aggregated as V_{cap} with average-pooling. The external knowledge is mined from the KB (Knowledge Base), and the responses are encoded by Doc2Vec, which produces a vector V_{know}. The three vectors V are combined into a single representation of scene content, which is input to the VQA(Vision to Answer) LSTM model that interprets the question and generates an answer.

In [4] proposed algorithm for generating sentences is not novel, but the approach is different. A pre-trained CNN (visual geometry group (VGG) net) model is used to extract objects from the image. After that, the text saliency approach is implemented for spotting the words, and the words are mapped to the embedding space using the Word2Vec algorithm. Eventually, the visual and textual features are fused in LSTm to generate the image caption. Experiments on the benchmark datasets such as (Flickr8k and Flickr30k) show that this model outperforms the baseline approaches. AUC scores are 0.806 and 0.798, which are outstanding scores. The captions generated from the image are very informative, and the text extraction adds specificity to the caption.

3 Methodology

3.1 Image Captioning

Image captioning, i.e., automatic generation of a natural language description from an image, has received much attention from both fields of Computer Vision (CV) and Natural Language Processing(NLP) (Vinyals et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2015). It is very useful for understanding images. Our project uses Image Captioning in order to get the correct caption for the image and then we will use this caption to generate longer and more sentences with GPT2 based on some dataset books.

In this project, encoder-decoder architecture had been used for captioning part and CNN for the vision part. Flickr8k data set containing 8000 images each associated with 5 captions have been used in this project. At the first stage of the process, the image is fed to the pre-trained CNN model (Inception V3), and the feature vectors are generated. Further, the vectors are gone through linear layers and passed to the encoder part of the LSTM model, while the word embeddings of the captions are fed to the decoder part. The output of each layer is passed through the softmax activation layer.

On the other hand, transfer learning is the idea of overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related ones.

Our current model consists of encoder-decoder architecture. The feature vectors of the images are obtained using a pre-trained CNN model (Inception v3) and fed to the encoder part, while the embeddings of words (GloVe word2vec) are fed to the decoder part.

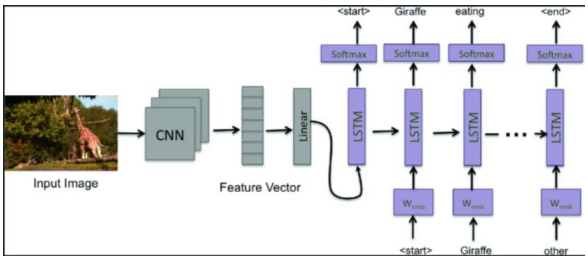

Figure 2: Encoder-Decoder Model

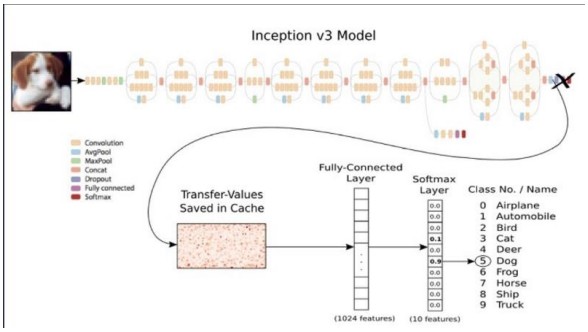

Figure 3: Transfer Learning using Inception V3 model

3.2 GPT-2 Architecture

GPT2 is a transformer-based architecture that is created by OpenAI. We used the model which is trained with 124m words.

A transformer model is a deep learning model which uses an attention mechanism to influence the weights of different parts of input data. It is similar to Recurrent Neural Networks (RNN) such that the transformer also handles data in a sequential way. Generally, the transformer is an encoder-decoder model.

In the GPT model, the BPE algorithm has 40000 tokens and it learned position embeddings with 512 positions.

Another point to mention is that 768-dimensional states and 12 attention heads exist in this model. The Feedforward network uses 3072 inner states too.

Moreover, in this project, the following books had been used as a dataset for GPT-2 model.

- "Miguel de Cervantes - Don Quixote"
- "Mark Twain - Tom Sawyer
- "Mark Twain - Huckleberry Finn"

As an example of the result, Figure 6 shows the possible results from GPT-2 based on Figure 5 caption. In Section 4, results will be discussed in detail with metrics.

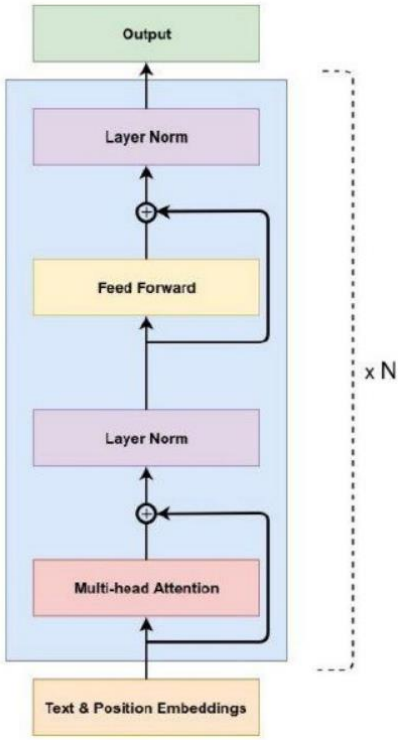

Figure 4: Transformer-Encoder Architecture

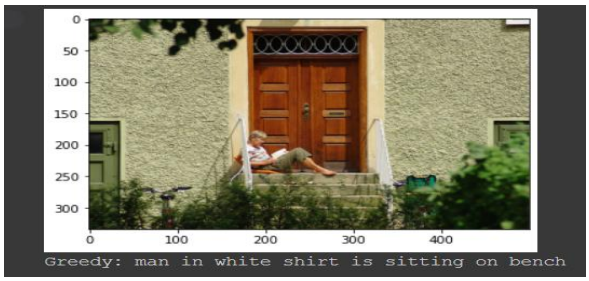

Figure 5: Image Captioning result example

```
man in red shirt is sitting on bench next to building.
"Let me
talk to you, señor," said he, "for I
can't help it."

"I
====================
man in red shirt is sitting on bench next to building
antechamber in the plaza. He is talking in a low voice.

"Go on, brother," he says, as he leans forward and takes in the
tone
====================
man in red shirt is sitting on bench next to building-place
man, who is a man, and he is a very sturdy person, and he is very
clumsy; and though he has a great deal of strength in his body,

====================
man in red shirt is sitting on bench next to building-mate
Chapman; and this is the gentleman, whom the landlord says
he has been observing the same thing all the day."

"I fancy so,
====================
man in red shirt is sitting on bench next to building
him a castle. "I am not," he says, "going to take a nap, but
it is my business to do so."
```

Figure 6: GPT-2 Result based on Figure 5

4 Results and Evaluation

To begin with, image captioning results had been evaluated with the help of the BLEU score. In this manner, it is important to mention about BLEU score. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. [5]
Also, the BLEU score ranges between 0-1, and its main advantage is the easiness of calculation. The comparison has done regardless of word order.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

Figure 7: BLEU score formula

As a result, Figure 8 shows the result of the image captioning part with the BLEU score respectively.


```
Greedy: dog is running through the grass

LABELS:
 black and gray dog
two black dogs fight
two black poodles fighting
two black poodles playfully tussle on grassy manicured lawn
two small black dogs facing each other with their mouths open
Bleu score 0.6389431042462724
```

Figure 8: Image Captioning Result with score

This caption feeds the GPT-2 model to get longer sentences based on indicated books.

```
two dogs are running through the grass,
and that the common people will not let them come out."

"I am afraid," said Sancho, "that the good people will not let

==================
two dogs are running through the grass."

"I am afraid so," said Sancho. "I have heard say that in
the village there is a fire-house where they burn the dead
==================
two dogs are running through the grass, and I am taking
out my
wine-skins."

"You are wrong, Sancho," said Don Quixote; "for my master
==================
```

Figure 9: GPT-2 Result

5 Conclusion

Although image captioning is challenging task, our model did quite well and produced good results. However, in our evaluation, Bleu score has some problems to evaluate it correctly.
On the other hand, GPT2 model is good but we could do it better to be more accurate. Generally, we will need more computing power for the future work. As a future work, we would like to apply attention mechanism as well.

References

[1] Tao Mei,Wei Zhang,Ting Yao. *Vision and Language: from Visual Perception to Content Creation*.

[2] Yan Chu, Xiao Yue,Lei Yu,Mikhailov Sergei,and Zhengkui Wang. *Automatic Image Captioning Based on ResNet50 and LSTM withSoft Attention*.

[3] Image Captioning and Visual Question Answering Based on Attributes and External Knowledge - Qi Wu, Chunhua Shen, Anton van den Hengel,Peng Wang, Anthony Dick

[4] Integration of textual cues for fine-grained image captioning using deep CNN and LSTM

[5] https://en.wikipedia.org/wiki/BLEU

[1] https://github.com/kayakayahan/nlp-imagecaptioning

[1] Code Repository