

Report

Generative AI model using GPT2 on Quora Question-Answer Dataset

Vaibhav Joshi

Date: 29 July 2024

Github link to the corresponding repo: [link](#).

Introduction

This report outlines the training of a GPT2 model (fine-tuning GPT2) on the [Quora Question-Answer Dataset](#).

Although two models, GPT-2 and T5 were trained, only the GPT-2 was selected as the final model, and so its details are mentioned in this text.

What is GenAI and GPT?

Generative Artificial Intelligence is an artificial intelligence capable of generating texts, images, videos, primarily using machine learning models.

GPT-2 is a Large Language Model which has the capability to answer questions about a topic from a text. GPT-2 uses attention mechanisms to focus into sections of more relevant texts only, and lesser on the less relevant texts.

Instructions for finding codes:

The corresponding code for the final GPT model can be found in `gpt_model.ipynb`. The T5 model's code can be found in `t5_model.ipynb`.

Literature Survey

- What is Lemmatization? <https://sciencedirect.com/topics/social-sciences/lemmatization>
 - "Lemmatization is the process of reducing words to their base or root form to facilitate analysis"
- Understanding Back Translation: <https://aclanthology.org/D18-1045.pdf>
 - "Back translation is the technique in which one sentence is translated into another language, and then translated back again to the original language, using some language translators."
- Paper of GPT-2 (Language models are unsupervised multitask learners): https://d4mucfpksyvv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
 - "Natural Language Processing tasks, such as question answering, reading comprehension, summarization, etc."

Methodology

Data Preprocessing, Cleaning

- Lemmatization was used to reduce words to their root words, or lemmas. This helps in improving the context of the questions/answers, and helps the model to understand better. Lemmatization helps to determine the context of the word.
- Tokenization was done on each question-answer pair, to feed to the model with tokens.
- The dataset contained many emojis, which were irrelevant for the model, so they were also removed.
- The dataset contained many URLs, which were also removed. I initially thought it would be correct to remove them, but then realized that for a domain specific-task, it may not be advisable to remove them, as they can potentially serve as an answer. This is a possible improvement in the model, to keep the URLs and then train.
- Stop-word removal can potentially remove the context of the sentence, and knowing the context is important for a QA model. Thus stop-word removal was not implemented.
- The dataset also has Hindi language sentences. These sentences were removed as well.

Model Selection and Evaluation

Specifically, two models were trained, one using T5, and the other using GPT. The T5 model took significantly much longer to train than the GPT model, and the GPT model gave better results than the T5 model. Due to this reason, the *GPT model was chosen for further hyperparameter-tuning and as the final model*.

The dataset was divided into two parts in ratio 8:2 for the train and validation sets respectively. The GPT model was trained for 2 epochs. Hyperparameter tuning: The max_length of the tokenized input was set to 1024, and the batch size was set to 2. The max_length input was key in the model performance; the higher the max_length of the tokenized_input, the model is able to get a bigger context of the question/answer, and thus its performance improves. It needs to be handled with care as it can lead to overfitting as well.

Results

After the final epoch of training the GPT2 model, the loss on the training set was 0.609, and the validation set loss was also around 0.603. Since the loss of the model was not high, this means that the model had learned significantly. And since the train and val set losses were similar, this means that the model was not overfitting, but may be learning properly.

The metrics ROUGE1, ROUGE2, and their corresponding precision, recall and F-scores were used to evaluate the model on 197 validation set examples. I could not validate the model on all

the examples because it was taking a lot of time to validate each row of the data. The average scores are shown below:

Rogue1 scores (order: precision, recall, f): 0.259 0.217 0.157

Rogue2 scores (order: precision, recall, f): 0.058 0.036 0.029

These scores must be considered in conjunction with human evaluations to get a comprehensive understanding of the model performance.

Some of the answers produced by the GPT model are shown below

Question: What industry needs lead generation and marketing?

Answer: Lead generation and marketing are two different field of business. Lead generation and marketing are two different fields, one is to connect with potential customer and build brand awareness. Lead generation and marketing are two different field that are both essential for business to grow and expand their reach. Here are some key point to consider: 1. Research and Optimize Your Website: Before starting your website, it's important to understand the content you're promoting. This will help you tailor your content to meet the need of your audience. 2. Create a Content Strategy: Develop a content strategy that outline your content strategy and how it will impact your business. This includes creating high-quality, engaging, and engaging content that resonates with your target audience. 3. Promote Your Products: Share your product or service to attract new customers. This can include product reviews, reviews, and other relevant information. 4. Engage in Social Media: Engaging in social medium can be an effective way to reach potential customers. Engaging in conversation about your product or service can help build relationship with your audience. 5. Use SEO Techniques: Use keyword research to identify relevant keywords that are relevant to your website's search engine results. This can include optimizing your website's

Question: how to use a proxy?

Answer: Proxy server are a type of web application that us the internet. They are used for various purposes, such as browsing, accessing websites, or accessing file on the Internet. Proxy server are used by various service provider to provide access to website and other data. They are used in various way to protect your privacy and security from unauthorized access. Some popular proxy server include Google Chrome, which provides a wide range of proxy server for various purposes. These proxy server allow user to browse the internet without any restriction or restrictions. However, some common practice involve using different proxies, such an IP address (IP), a well a setting up VPNs, which can be used to track and

access website from anywhere. Additionally, these proxy server allow user to access website directly through their mobile devices, which can help them stay connected with others online.

Question: What are the must-have iPhone apps?

Answer: There are many must-have iPhone apps available, including Google Play Store and Apple Music. Here are some of the most popular and widely used apps for Android: 1) Apps for music streaming: Spotify, YouTube Video Player (iTunes), and Apple Music playlists can be a great way to listen to your favorite song or create a playlist that is more relevant than others. 2)(Music streaming app): Spotify offer a wide range in it genre such as classical rock/rock, electronica, and pop. It's important to note these apps are not intended solely on a specific device, and they may require additional processing and maintenance to ensure they are user-friendly and user-friendly. 3.) App Store: Apple Store provides a wide selection of apps for Android devices, including iWatch TV, iWatch Pro Max HD, and iMovie. These apps are designed specifically for Android devices, and are designed with the goal of providing a seamless experience for both users' needs.

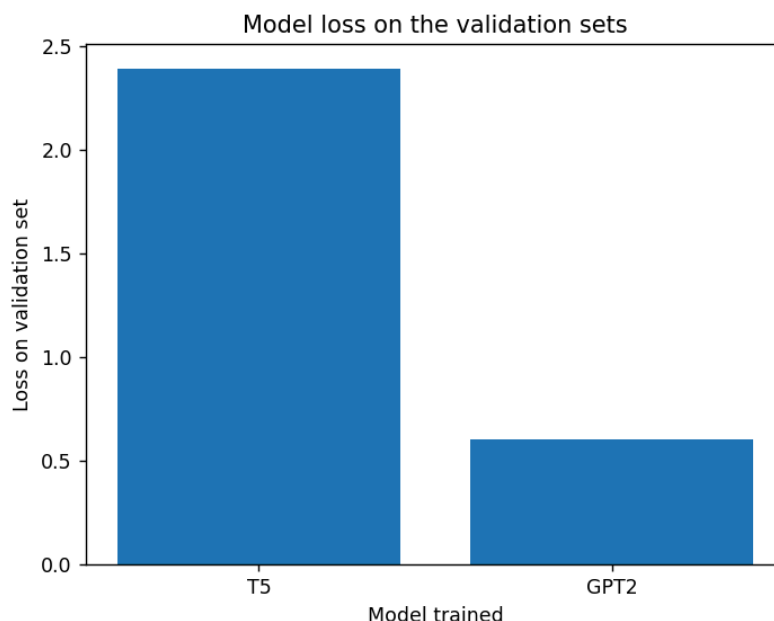
Question: What is the color code for 52M ohm?

Answer: I have a black and white laptop with a black screen.

The answers to the first three questions are meaningful. The answer to the last question is incorrect. The model did not paraphrase the answers it gave.

Visualization

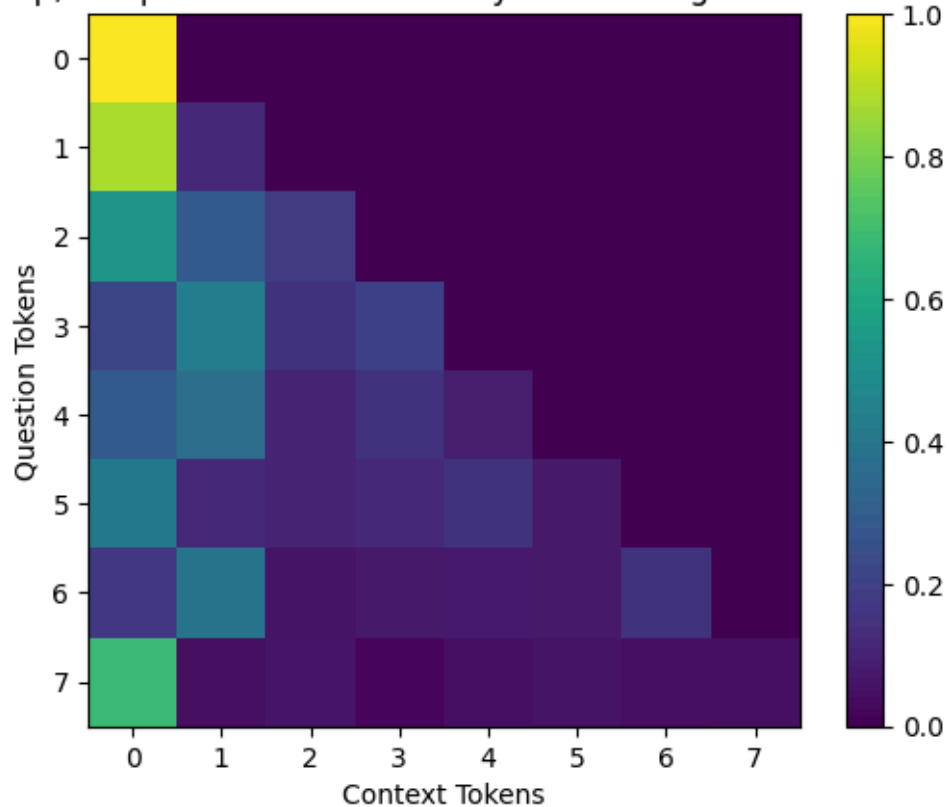
The model performances have been visualized below



A lower loss in GPT compared to T5 implies a better QA understanding of the GPT model compared to T5 model. This implies that the GPT model generalizes better than the T5 model.

Attention Heatmap for one question “what industry needs lead generation and marketing?”, helps to understand where all the model was putting more attention to: the tokens given more attention have more heat, and the tokens on which less attention was given are less hot, and are bluer:

Attention Heatmap, for question "what industry need lead generation and marketing?"



Note that the tokens “what”, “industry” and “marketing” are relatively hot, implying that the model gave more attention to them.

Insights and Recommendations

- Data Augmentation: Techniques such as back-translation can be incorporated to enhance the model performance
 - Back translation is the technique in which one sentence is translated into another language, and then translated back again to the original language, using some language translators. This helps in generating new different sentences for the same questions/answers, thus augmenting the data.
- It is to be importantly noted that the quora question-answers dataset contains answers from common people, and not necessarily from domain-experts of the question asked.

This means that, if there are 50 answers to one question (say), then not all the answers are the same, and this means that the model learns all the answers, whether they are correct or not. This hampers the performance of the model. As a resolution to this, the dataset must be manually curated (by domain-experts) to remove all the incorrect answers. This will improve the model performance.

- It was noted above that URLs may be important to keep in the data, so do not remove them from the data. This may hamper the performance of the model somewhat on some inputs, but could lead to slightly more contextual answers and may thus improve overall ROGUE score.
- Similarly it is not advised to implement stop-word removal, as it can potentially remove the context (meaning) of a sentence, in my opinion. As an example, the sentence "Today it may rain because the sky is cloudy", after stop word removal, will look something like this: "Today rain sky cloudy", I don't think this is useful for a generative AI model which needs to generate human readable texts. Stop-word removal may be useful for classification related tasks, not here.
- Use techniques such as classification models to classify the dataset into some fixed number of categories, and then train the gen-AI model along with these classification information.

Conclusion

A model was trained for the question-answer Quora dataset. The model may not be easy to train, because it contained a wide range of information, and, the correctness of the answers is a challenge to be addressed for next models, which can improve the performance of the model. Note that Chatgpt was utilized for this project.