Amazon QA Bot

Andy(Xiangyu) Cui, cui.xiangyu@northeastern.edu
Zichong Meng, meng.zic@northeastern.edu
Xichen Liu, liu.xic@northeastern.edu
Xueyan Feng, feng.xuey@northeastern.edu

Abstract

Our group explores the implementation and evaluation of Question Answering (QA) models using two natural language processing (NLP) models: GPT-2 and BERT. We will be focusing on the domain of Amazon product reviews. And we will test the performance and limitations of the two models in generating information based on product related textual information and questions.

The evaluation process ultilizes different metrics including loss, accuracy, and BLEU score. And we will use those metrics to measure the effectiveness of both models.Also, other assessments (for example human) were conducted to gain insights about the models' contextual understanding and their ability to handle varying styles and user intents present in the Amazon product reviews.

Our findings highlight the trade-offs between GPT-2 and BERT in the context of QA tasks on Amazon product reviews. We see that while GPT-2 does well in capturing broader contextual information and generating coherent responses, BERT works better in extracting precise details from the reviews.

The implications of this research extend to the broader field of NLP. Our research offers great insights into the practical application of GPT-2 and BERT in the real life e-commerce. And we hope the outcomes of this study will contribute to the ongoing discourse on the optimal use cases for different NLP architectures.

1. Introduction

Question Answering (QA) systems work by processing text queries and retrieving information to provide precise answers. In the field of NLP, Question Answering (QA) systems complete process by using neural networks to understand and interpret the text queries. The processing process suaully begins with the input text being tokenized and encoded into numerical representations. Then the encoded inputfed into the neural network. And the NLP model will processes this input, considering both the linguistic structure and the semantic context, and then determine the information to answer the query. Once the relevant information is identified, then it will synthesize and formulate a response.

In the scope of this paper, our team provides a detailed investigation into the training and performance evaluation of Question Answering (QA) models, focusing on utilizing two NLP Models: BERT and GPT2. We also emphasize the focus of this study to the analysis of Amazon product reviews question answerings.

GPT-2 is selected for its advanced language generation capabilities, offering the potential for creating responses that are like human assistant. Conversely, BERT is selected because of its very good at understanding of language context and extract information to form the answer because of its bidirectional nature.

Our research methodology incorporates a comprehensive analysis. We split a large corpus of Amazon product reviews into train and test datasetsis processed through and trained both GPT-2 and BERT models. We finetunes both of these two models and followed by evaluating the accuracy models' responses. To quantitatively assess the performance of GPT-2 and BERT post-training, we metrics such as accuracy, which measures the models' ability to provide correct answers, and/or BLEU scores (depends on the model), which evaluate the quality and coherence of the generated text, and the loss of the models. This methodological approach make sure a comprehensive evaluation which is very important to insights about the performance of the two models in processing and responding in the context of Amazon consumer reviews question answering.

This study also aims to identify the distinct advantages and constraints of each model when applied to the different scenarios posed by Amazon product reviews. These reviews, characterized by varied linguistic expressions, informal styles, and fluctuating lengths, serve as a challenging environment to test the versatility and resilience of QA models. We hope our comparison of BERT and GPT2 in this paper can contribute in enhancing the development, selectiong QA systems for the real world e-commerce applications.

2. Method

2.1 Dataset and data preprocessing

Our team uses the 'Toy Products on Amazon' dataset from Kaggle [1] for this Final project. This dataset includes many useful Amazon customer data including questions, answers, and reviews to many toy products that was sold on Amazon. However, out of many informations in the dataset, we focuses more on the product names product description, questions about this product, and the answer for the question in our research setting.

In order to use this dataset for language model training, specifically with BERT and GPT-2, we need to prepare and preprocess the dataset. To achieve this, we first thoroughly cleaned the dataset and extract the needed information from the dataset. This process involves removing missing information and correcting anomalies in the daya text, which also includes removing uncommon characters and duplicate answers. Also, for different model we reformatted the data differently to align with the requirementsto train the models since the nature of BERT and GPT2 are different. Upon completion of these steps, we compiled the data into a JSON file. This file is now ready to used for our model training.

2.2 Models

2.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) [2] is a encoder only transformer model developed

by Google AI Language. The most important feature of BERT is that this NLP model analyzes the context of a word in both from left to right and from right to left direction. And this important architecture of BERT makes it very good at natural language understanding because the bidirection approach allows for an depth context comprehension.

The BERT model is improved from the original transformer model but only utilized tha encoder part of the model and it's bidirectional. And it was used in many different settings in NLP and it achieved significant result in many NLP tasks. For example, sentiment analysis,information extraction, and language understanding.

2.2.2 GPT2

Generative Pre-trained Transformer 2 (GPT-2) [3] is a decoder only transformer model in the field of NLP, developed by OpenAI. GPT-2 is also improved and built upon from the original transformer model. But it only adopts the decoder part of the model. And because of the decoder only model architecture, GPT2 is excellent in text generation task. For our primary focus which is GPT-2 small, the model's architecture comprises 12 layers of transformer decoder blocks. For our primary focus for this project which is Question Answering (QA) for Amzon toy product tasks, GPT-2 will process input and use the contextual information gathered from the preceding text to generate predictions for the next word to generate an answer in a sequence.

2.3 Training

2.3.1 BERT

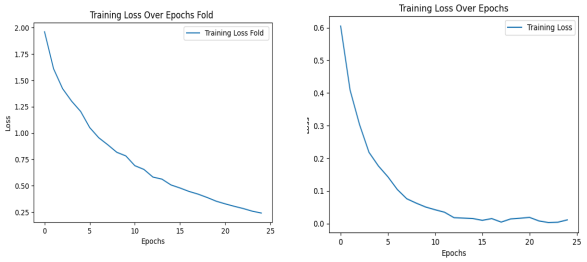
In the process of fine-tuning the BERT model on the Amazon product review dataset, we employed the following procedure. Initially, the model read in the preprocessed data from a json file and created a custom torch dataset that is consist of inputs with Product name + Product Description + Question + Answer.
We processed the data into a 70% training set and a 30% testing set. On the training set, we trained our model for 25 epochs using the Adam optimizer with initial parameters set to a learning rate (lr) of 2e-5, $\beta 1 = 0.9$, and $\beta 2 = 0.999$. We employed the Cross-Entropy Loss function as our loss function. This training process took approximately 20 minutes in total on a GTX 3060 Ti GPU.

2.3.2 GPT2

In the process of fine-tuning the Generative Pre-trained Transformer 2 (GPT-2) model on the Amazon product review dataset, a meticulous training procedure is employed. Initially, GPT-2-small, already pre-trained on a diverse range of text sources, is further adapted to the specific linguistic characteristics of consumer reviews. We first read in the preprocessed data from the json file and create a custom torch dataset that returns an input of Product+Product Description+Question+Answer. We processed the data into 70% training set and 30% testing set. We trained our model over 25 epochs on the training set using the Adam optimizer with initial parameters lr = 5e-5, $\beta 1 = 0.9$, and $\beta 2 = 0.999$. We use Crossentropy loss as our loss function.. And this training took approximately 12.5 minutes on a RTX3090 GPU.

3. Results

3.1 Experiment



(a) GPT 2 Average Train Loss

Figure 1: (a) Average train losses when training GPT2 model for 25 epoch with AdamW optimizer and LR = 5e-5 and BS = 8. (b) Average train losses when training BERT model for 25 epoch with AdamW optimizer and LR = 2e-5 and BS = 8

Model	Hyperparameters	Accuracy
GPT2	25 Epoch LR = 5e-5 BS = 8	0.7695 (BLEU)
BERT	25 Epoch LR = 2e-5 BS = 8	0.9264

Table 1: Average accuracy of GPT2 and BERT model on the test set.

3.2 K-Fold Cross-Validation

In this study, we performed a k-fold cross-validation which is employed to assess the performance of the proposed Question Answering (QA) model applied to Amazon product reviews. We implements the K-Fold Cross-Validation studies for BERT with 10 folds and GPT2 with 5 fold. With this method we can ensure that model can generalize across distinct subsets of the dataset. In our training for K-Fold Cross-Validation, for each fold, we reinitialized the model and configuration. And the dataset is splited into training and testing sets based on the which fold it is. And we fine tune the model like usual method. We also included a visual plot provide insights into the model's convergence patterns. This methodology is very useful for testing the robustness and reliability of our evaluation. And it offers valuable insights of the two model's performance across various data distributions.

Model	Hyperparameters	Fold number	Accuracy (BLEU)
GPT2	25 Epoch LR = 5e-5 BS = 8	Fold 0	0.7627
GPT2	25 Epoch LR = 5e-5 BS = 8	Fold 1	0.7622
GPT2	25 Epoch LR = 5e-5 BS = 8	Fold 2	0.7548
GPT2	25 Epoch LR = 5e-5 BS = 8	Fold 3	0.7486
GPT2	25 Epoch LR = 5e-5 BS = 8	Fold 4	0.7781

Table 2: K-Fold Cross-Validation average accuracy for GPT2 model.

Model	Hyperparameters	Fold number	Accuracy
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 0	0.9464
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 1	0.9246
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 2	0.9643
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 3	0.8770
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 4	0.9107

BERT	25 Epoch LR = 2e-5 BS = 8	Fold 6	0.9167
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 7	0.9405
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 8	0.9286
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 9	0.9405
BERT	25 Epoch LR = 2e-5 BS = 8	Fold 10	0.9107

Table 3: K-Fold Cross-Validation experiment for BERT model.

3.3 Ablation Study

We also conducted an ablation study to explore the impact of choosing different variations of hyperparameters on the models performance. For each fine-tuned model in the ablation study, we tests them on the same testing set of Amazon product question and answers. Note that here each configuration represents a unique combination of hyperparameter values, including batch size, learning rate and epochs and creates 10 different configurations. For better visualization, we visualized epoch losses, and we plot the average losses. And we calculate of the average BLEU score of the model on the testing dataset**.** This controlled ablation study provides a detailed exploration of the influence of key hyperparameters on the QA model's efficacy, offering insights into the optimal configuration for achieving superior performance in the context of Amazon product reviews.

Model	Hyperparameters	Accuracy (BLEU)
GPT2	25 Epoch LR = 5e-5 BS = 8	0.7695
GPT2	25 Epoch LR = 1e-5 BS = 4	0.7922
GPT2	25 Epoch LR = 1e-5 BS = 8	0.7912
GPT2	25 Epoch LR = 1e-4 BS = 4	0.7570
GPT2	25 Epoch LR = 1e-4 BS = 8	0.7592
GPT2	30 Epoch LR = 1e-5 BS = 4	0.7873
GPT2	30 Epoch LR = 1e-5 BS = 8	0.7923
GPT2	30 Epoch LR = 1e-4 BS = 4	0.7611
GPT2	30 Epoch LR = 1e-4 BS = 8	0.7435
GPT2	25 Epoch LR = 1e-6 BS = 2	0.7901

Table 4: Ablation studies with 10 different configurations and the average accuracy for GPT2 model.

Model	Hyperparameters	Accuracy (BLEU)
BERT	25 Epoch LR = 2e-5 BS = 8	0.9264

BERT	25 Epoch LR = 1e-5 BS = 4	0.9571
BERT	25 Epoch LR = 1e-5 BS = 8	0.9427
BERT	25 Epoch LR = 1e-4 BS = 4	0.9468
BERT	25 Epoch LR = 1e-4 BS = 8	0.9346
BERT	30 Epoch LR = 1e-5 BS = 4	0.9366
BERT	30 Epoch LR = 1e-5 BS = 8	0.9366
BERT	30 Epoch LR = 1e-4 BS = 4	0.9264
BERT	30 Epoch LR = 1e-4 BS = 8	0.9304
BERT	25 Epoch LR = 1e-6 BS = 2	0.9080

Table 5: Ablation studies to train BERT models with 10 different configurations.

4. Discussion

Our goal for this project is to train a customer QA bot that can answer questions about various amazon products. And as we compare the two models we used, we can see that BERT-base's accuracy is a little bit higher than GPT-2-small's performance on the testing set. So on the accuracy measurement level, BERT is probably a better option for this particular task. However, if we use human reading interpretation of selected testing cases, we can see that GPT2 is a generative model focused more on the generative part compared to BERT which focused on understanding part and extract information to present as output. Therefore, in real life, since we are trying to build a QA that can serve customers like a human QA assistant, we probably will consider GPT2 as well for its generative, improvising nature.

The K-Fold Cross Validation studies for both BERT and GPT-2 results show near the same average accuracy for all folds for each model.

The ablation studies for BERT shows that batch size of 4 and learning rate of 1e-5 works the best out of the 10 cases. However, the convergence of the model are not smooth as batch size of 8 and learning rate of 2e-5, so we do need to be careful when we select when hyperparameter to choose.

The ablation studies for GPT2 shows that a smaller learning rate than 5e-5 can improve the model performance by a few percent. However, smaller learning rates in the ablation studies results in a very large average training loss compared to LR=5e-5, therefore, we would need to be cautious about which configuration of hyperparameters to use.

Furthermore, even though our current approach reached a nice working result, we believe if we used a larger model for experiment we can probably achieve even more significant results. Unfortunately, we ran out of time for this idea.

5. Conclusion

In this project, we explore training BERT and GPT-2 models as a QA bot using Amazon Customer QA dataset and test their performance as a customer QA bot. In the

future, we would like to explore larger customer QA datasets, so that it can not only answer questions about amazon products but also explore how customer QA bot trained on larger datasets perform on different categories of products.

6. Github

https://github.com/AndyFCui/Amazon_QA_Robot/tree/main

References

[1]: Dataset: https://www.kaggle.com/datasets/PromptCloudHQ/toy-products-on-amazon/data

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional Transformers for language understanding," arXiv.org, https://arxiv.org/abs/1810.04805

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[4]: Question Answering Systems: Survey and Trends https://doi.org/10.1016/j.procs.2015.12.005

[5]: Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets https://doi.org/10.48550/arXiv.2008.02637

[6]: Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds https://doi.org/10.48550/arXiv.1911.02365