Comparison of fine-tuned DistilBert model for Conversational Question-Answering

Opeyemi Olanipekun (obolanipekun@ischool.berkeley.edu)

December 2022

Abstract

It is essential to enable machines to be effective at answering conversational questions. However, the understanding of conversational questions answering remains a major task and is often difficult for automated systems to understand. There has been an increased use of pre-trained models on various NLP tasks with great performance such as GPT[1], DialoGPT[2] , BERT[3] and RoBERTa[4]. Hence, in this paper, I propose to fine-tune a distilledBert model with a CoQA dataset and another with a SQuAD dataset and then evaluate both of them with a CoQA validation dataset to help better understand what kind of text dataset would enhance the performance of a model on a question answering dataset. In order to assess this, I conducted some experiments. My results support the idea that the dataset upon which a model is trained is important towards the performance of the model.

Introduction

My goal for this project was to build and compare the F1 score of two models which are trained on different datasets. Both models would be validated on a conversational question answering validation dataset. Conversation is a major part of human life where one question leads to the other. The conversational history nature of such questions and answers where one answer leads to another question makes it difficult for machines to comprehend and answer conversational questions. For humans, if we show a question from an existing conversation to a new person, it is highly likely that they would deviate from the original answer. This type of scenario makes it difficult for machines to learn conversational dataset. Hence, for humans to make an effective chat bot system, it would be important to train machines on conversational question answering dataset.

In this paper, I plan to fine-tune a DistilBert model on a conversational question answering (CoQA) dataset[5] and another DistillBert on a SQuAD 2.0 dataset[6] which is a reading comprehension dataset. Please note that in this paper, a model trained on the CoQA dataset would be referred to as DistilBert_CoQA while the one trained on the SQuAD 2.0 dataset would be referred to as DistilBert_SQuAD. Both models would be evaluated on a conversational validated dataset. CoQA is a difficult dataset because there is a connection between one question and the next. The CoQA dataset consists of a passage, questions and answers. There are two types of answers in this dataset. One is a free-form text while the other is a text span from the passage which serves as a rationale to the free-form answer. In this project, I treated the CoQA as a single turn reading comprehension in order to see if the model would catch the link a question or answer has with the previous one. So, in essence, I wanted to see if a model trained on a SQuAD 2.0 dataset would show a better F1 score on a CoQA validation dataset than a model trained on a CoQA dataset or vice versa. The result obtained in such analysis would help to better understand the best approach to training a model for a conversational question answering system.

Background

The CoQA dataset is a dataset that mimics the way humans converse on a daily basis where a question is asked and an answer is provided and depending on the answer a follow up question is then asked which then leads to another answer. This particular sequence continues. The CoQA dataset comprises 127,000 questions and answers which are obtained from 8,000 conversations and text passages from seven diverse domains[5]. Most of the earlier work on questions-answering has been on machine reading comprehension passage which is classified as single turn because it does not depend on the conversation history[7]. However, CoQA is classified as multi-turn where each question depends on the conversion history.

A number of models have been developed using the single-turn question answering machine reading comprehension such as BiDAF[8], DrQA[9] while others such as FlowQA[10] and SDNet[11] have been developed using the multi-turn conversational answering dataset. These models found the optimal answer

span using the passage and the conversation history. Pre-trained models such as BERT which have shown major performance improvements on language tasks can be explored on the CoQA dataset. Transformers have been found to be very successful in the conversation questions-answering task with the best model currently a RoBERTa based model with a F1 score of 90.7 outperforming human performance of 88.8.[7]

My task in this project would be modeled as a reading comprehension problem where I would analyze how a pretrained model (DistilBert) fine-tuned on a machine reading comprehension single-turn dataset (SQuAD 2.0) would perform on a CoQA dataset compared to the same model fine-tuned on a CoQA dataset and validated on a CoQA dataset. In this project, I focus on the extractive question-answering on finding a span in the passage which matches the question in the dataset. It would be interesting to see if the trained model would be able to extract the correct answer from a conversational dataset without dealing with the conversational history.

Methods

Data

The data used for this project are the CoQA and the SQuAD 2.0 dataset. The CoQA dataset was created by Reddy et al.[5] and was downloaded from their github link **(**https://stanfordnlp.github.io/coqa/) while the SQuAD 2.0 dataset was created by Rajpurkar et al[6] and downloaded from their github link (https://rajpurkar.github.io/SQuAD-explorer/). A quick analysis of both datasets showed that there are major differences. Nearly half of SQuAD 2.0 questions are dominated by "what" questions as seen in figure 1 (right) where I showed the first twenty most frequently used words used in starting the questions whereas the CoQA dataset has a distribution of other words such as "who", "how", "did", "where", "was" although questions starting with "what" still took a larger pie of the distribution (figure 1 left).

I also took a look at the length of the questions in both SQuAD 2.0 and CoQA. As you can see in figure 2 (left and right), the CoQA datasets questions are usually shorter compared to questions in the SQuAD 2.0 dataset. This might be due to the conversational nature of the datasets. We can also see that there are a number of questions in the CoQA datasets that only have a length of either 1, 2 or 3 words. Such questions would be difficult for machines to answer. However, it would be interesting to see if models trained on such datasets using extractive question answering can perform well on same (CoQA) validation datasets compared to a model trained on an actual machine question-answering dataset such as the SQuAD 2.0 datasets. It is also important to note that CoQA has about 11.1% questions with a "yes" answer and 8.7% questions with a "no" answer whereas SQuAD 2.0 does not have any question with a "yes" or "no" answer. This also makes it difficult for a machine to perform well on a conversional dataset.

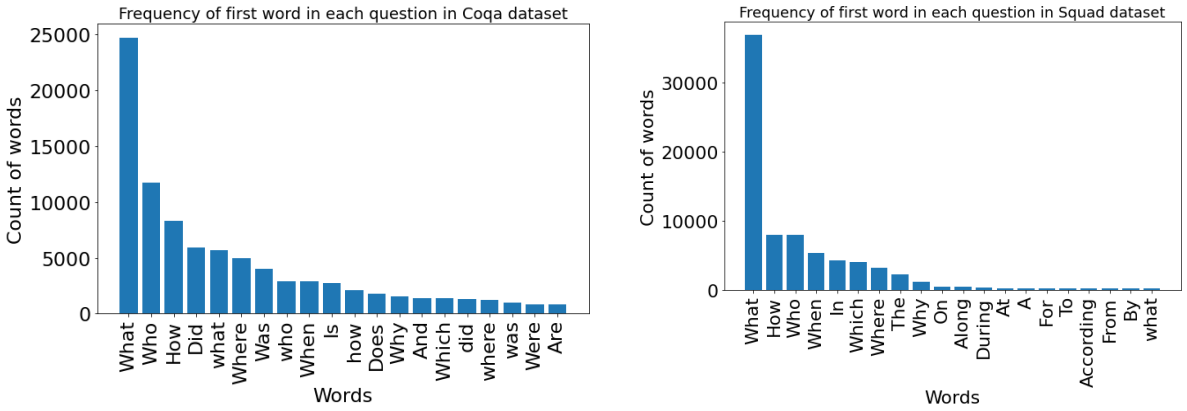

Figure 1. (left) showed the frequency of the first word of each question in the CoQA dataset. (right) showed the frequency of the first word of each question in the SQuAD 2.0 dataset.

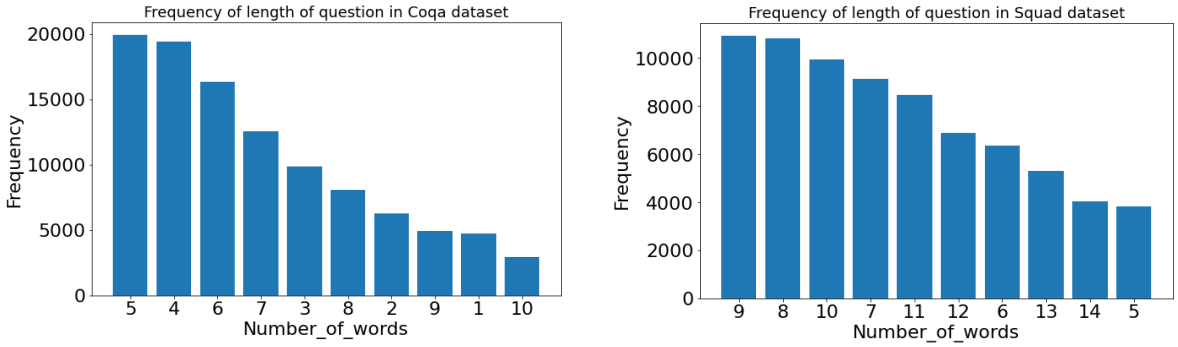

Figure 2. (left). showed the frequency of the length of questions in the CoQA dataset. (right) showed the frequency of the length of questions in the SQuAD 2.0 dataset.

Baseline

For my baseline, I use a distilBert base model (with a question and answering head) which is a smaller, faster, cheaper and lighter distilled version of Bert[12]. This model was downloaded from the hugging face repository and used for evaluation on the CoQA validation dataset without any fine-tuning of the model. The choice of this model as opposed to the actual BERT model released by the google team[3], was because of its size reduction of 40% compared to the size of BERT while retaining 97% of its language capabilities and being 60% faster[12]. For this project, I use F1 score to measure the model performance. The F1 score is the harmonic mean of the precision and the recall at word level between the predicted answer and the ground truth.

Modeling

For both the other 2 models fine-tuned on CoQA and SQuAD 2.0 datasets, I used the hugging face library as my source of the distilBert base model with question and answering head. Various experiments were performed to optimize the training of the models on their respective datasets. The models are optimized using adam optimizer with a weight decay to reduce chance of overfitting and also using an initial learning rate of 0.00005[7]. To obtain an optimized fine-tuned model, I experimented with freezing and not-freezing the distilBert layers while also adjusting the number of epochs and batch size. Throughout the experiment, I kept the initial learning rate and the optimizer constant due to time constraints and limited compute resources. The models were run on a paid version of the google colab (PRO+) with access to GPU and high RAM.

All of the models fine-tuned on the CoQA dataset were also evaluated on the CoQA dataset while the models fine-tuned on the SQuAD 2.0 dataset were evaluated on the SQuAD 2.0 validated dataset but the SQuAD 2.0 fine-tuned model which gave a better result on the SQuAD 2.0 validation dataset was then evaluated on the CoQA dataset to be compared with the performance of the model fine-tuned on the CoQA dataset and the baseline.

Results and Discussion

The result below showed the experimental set up and the result in terms of F1 score for all the models fine-tuned on CoQA and SQuAD including the baseline which was not fine-tuned. In my github repo (https://github.com/YemiOlani/Fall_2022_W266_Final_Project_Submission), I also showed the exact match accuracy of these models. As you can see below, the DistilBert_SQuAD model (model 8, 9, 10, 11, 12 and 13) were first evaluated on the SQuAD dataset and the one with the highest F1 score (model 10) with the number of epoch =1 and batch size = 16 with trainable parameters gave a better result on the SQuAD 2.0 validation dataset. This model was then evaluated on the CoQA validation dataset to compare with the best model of DistilBert_CoQA.

Model	Data Fine Tuned On	Data Validated On	Optimizer	Learning Rate	Epoch	Batch Size	Layers Frozen	F1 Score on validation dataset %
Model 1 (Baseline)	No fine tuning	CoQA	NA	NA	NA	NA	None	8.92%
model 2	CoQA	CoQA	AdamW	5e-5	1	16	0,1,2,3	47.85%
Model 3	CoQA	CoQA	AdamW	5e-5	1	16	0,1,2,3,4	42.42%
Model 4	CoQA	CoQA	AdamW	5e-5	1	16	None	49.15%
Model 5	CoQA	CoQA	AdamW	5e-5	3	16	None	48.82%
Model 6	CoQA	CoQA	AdamW	5e-5	5	16	None	46.23%
Model 7	CoQA	CoQA	AdamW	5e-5	1	32	None	48.14%
Model 8	SQuAD	SQuAD	AdamW	5e-5	1	16	0,1,2,3	77.68%
Model 9	SQuAD	SQuAD	AdamW	5e-5	1	16	0,1,2,3,4	67.86%
Model 10	SQuAD	SQuAD	AdamW	5e-5	1	16	None	82.59%
Model 10	SQuAD	CoQA	AdamW	5e-5	1	16	None	30.11%
Model 11	SQuAD	SQuAD	AdamW	5e-5	3	16	None	81.79%
Model 12	SQuAD	SQuAD	AdamW	5e-5	5	16	None	81.43%
Model 13	SQuAD	SQuAD	AdamW	5e-5	1	32	None	82.31%

Table 1: Experimental set up and results for CoQA and SQuAD DistiledBert fine-tuned Model.

It can be seen on table 1 above that both DistilBert_SQuAD and DistilBert_CoQA outperforms the baseline which was not fine-tuned. This confirms the importance of fine-tuning in the question-answering task. We can also see on table 1 that the DistilBert model fine-tuned on the CoQA dataset outperforms the model fine-tuned on the SQuAD dataset by about 19.04% F1 score when both are evaluated on the CoQA validation dataset. This might be due to the fact it was actually fine-tuned on the CoQA dataset itself. However, we can see that the DistilBert model did perform better on the SQuAD 2.0 dataset than on the CoQA when each of them were evaluated on their respective validation dataset. We can attribute this to the conversational nature of questions in the CoQA dataset as described previously. In addition to the fact that the next questions in the dataset builds on top of the previous question(s), some of the questions are just one word such as “where?”, “who?”. These words are not contained in the answer span or part of the context that contains the answer to the questions, hence making it difficult for the model to learn this in an extractive question-answering task. This further suggests that conversational dataset such as CoQA cannot be treated as a sole extractive question answering task.
We can also see on table 1 that unfreezing all the DistillBert layers is crucial to the performance of the model.

CoQA questions	True Answer	Predicted Answer By distillBert_CoQA	Predicted Answer By distillBert_SQuAD
What color was Cotton?	a little white kitten named cotton.	white kitten named cotton.	white

Where did she live?	in a barn near a farm house, there lived a little white kitten	once upon a time, in a barn near a farm house, there lived a little white kitten named cotton.	high up in a nice warm place above the barn
Who did she live with?	with her mommy and 5 other sisters.	she shared her hay bed with her mommy and 5 other sisters.	her mommy and 5 other sisters.
Whose paint was it?	the old farmer ' s orange paint,	cotton found a can of the old farmer ' s orange paint,	farmer ' s orange
What did Cotton's mother and siblings do when they saw her painted orange?	when her mommy and sisters found her they started laughing.	when her mommy and sisters found her they started laughing.	laughing.
What is the first phrase I learn?	" thank you "	thank you " is, of course, the first one. somehow, it seems inadequate.	
Why?	nicole has obviously told her mother that i am having heart surgery soon, so her mother has decided i need more nutrients.	so her mother has decided i need more nutrients.	
Who cast him?	farina, who had a long career as a police officer in chicago, got into acting through director michael mann, who used him as a consultant and cast him	farina, who had a long career as a police officer in chicago, got into acting through director michael mann, who used him as a consultant and cast him in his 1981 movie, " thief. "	michael mann,
Did Kendra tell him why?	she told him what happened	she told him what happened and he said that he was sure that everything would be okay.	
What does Kendra not want to miss?	she doesn ' t want to miss story time	kendra didn ' t want to miss the bus to school and hurried off to make it in time.	story time

Table 2: Typical comparison of predicted answers from DistillBert_CoQA and DistillBert_SQuAD with the ground truth answers.

I took a look at the first eighty questions, and their true and predicted answers (can be seen in my jupyter notebook report in my github repo), however I showed a few of these in Table 2 which shows a comparison of the predicted answers emitted by the DistilBert_CoQA and DistilBert_SQuAD with the ground truth answers. A thorough analysis of this comparison shows that DistilBert_SQuAD does better at predicting short answers rather than long extractive answers compared to DistilBert_CoQA. Since this project involves an extractive question-answering task, this might be one of the reasons that it has a lower F1 score than DistilBert_CoQA since the number of shared words between the predicted and the ground truth is the basis of the F1 score. DistilBert_SQuAD is good at answering questions that begin with "what". The reason this is so might be attributed to the fact that nearly half of the questions in the SQuAD 2.0 dataset is dominated by "what" questions (see figure 1-right). Since DistilBert_CoQA did better on extractive answers from the passage, it is expected to have a higher F1 score as seen in my result. DistilBert_CoQA also did better on questions with just one word such as "why?". This might be attributed to the fact that the model was actually trained on the CoQA datasets making the model to have learnt alot about this dataset unlike DistilBert_SQuAD which was never trained on the CoQA dataset.

Conclusion

Both DistilBert_CoQA and DistilBert_SQuAD showed improved performance over the baseline. DistilBert_CoQA gave a higher F1 score than DistilBert_SQuAD which can be attributed both to the fact that the model was actually trained on the CoQA dataset and also that most of its predicted answers are longer extractive answers obtained from the passage. DistilBert_SQuAD on the other hand, gave short correct answers. It is worth noting that in this project, the span text was used as the answers during training and evaluation. I did not use the conversational answer in the CoQA dataset for training and evaluating. Also, I did not consider the conversational history of the dataset during modeling. Despite this fact, the DistilBert_SQuAD model was good at predicting shorter correct answers. This attribute of DistillBert_SQuAD model shows that it would be a good model candidate upon which other architecture can be built to solve conversational challenges. A major take home in this project is that the dataset upon which a model is trained on can affect the performance of the model.

The next approach for this project would be to perform this task modeled as a conversational response generation problem rather than an extractive response generation and using the DistillBert_SQuAD as the foundation of the architecture.

References

1. Time Salimans Alec Radford, Karthik Narasimhan and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI
2. Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. arXiv:1911.00536v3
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
4. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints, page arXiv:1907.11692.
5. Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. arXiv e-prints, page arXiv:1808.07042.
6. Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Association for Computational Linguistics (ACL), pages 784–789, Melbourne, Australia.

7. Technical report on Conversational Question Answering Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang and Yunfeng Liu arXiv:1909.10772v1
8. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. arXiv e-prints, page arXiv:1611.01603.
9. Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. arXiv e-prints, page arXiv:1704.00051.
10. Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. FlowQA: Grasping Flow in History for Conversational Machine Comprehension. arXiv eprints, page arXiv:1810.06683.
11. Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering.
arXiv e-prints, page arXiv:1812.03593.
12. Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108v4