Critical analysis of transformer models on COVID-QA Dataset

Kaushik Ravindran
A59012879
Computer Science And Engineering
University Of California, San Diego
kravindran@ucsd.edu

Yash Khandelwal
A59012226
Computer Science And Engineering
University Of California, San Diego
ykhandelwal@ucsd.edu

Abstract

COVID-19 has affected people all over the world in a significant manner. Updating and learning about any kind of information on this is the need of the hour for most of us. The same can be achieved by creating a mechanism to answer people's queries. This can be accomplished by creating a QA model. We made use of the COVID-QA dataset to fine-tune on several models that were pre-trained on the SQuAD dataset. The results achieved on various metrics improve over what was achieved at that particular time. We achieved the maximum F1 score of 68.896 and an exact match of 43.564 with the ELECTRA base model. We also calculated the amount of CO_2 emissions taking place to train such models.

1 Introduction

Impact of COVID-19 is not unknown to the world. This is a new but infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). So extracting any kind of information on this disease is one of the primary concerns. Extracting such information is a challenging and non trivial problem. There is a demand to serve the purpose of extracting information to any kinds of queries. Question answering (QA) is a branch within natural language processing (NLP) and information retrieval fields, which is concerned with building systems that automatically answers questions posed by humans in natural language. A Question Answering (QA) system that would answer any query raised would solve this need. Studies in Artificial Intelligence (AI) have matured the task of field specific Question-Answering (QA). To build any end-to-end deep learning based QA system, dataset with a large volume is needed. This kind of QA would help researchers, developers, editors/associate editors of a Journal and also to answer queries for the common-man.

Lot of misinformation and false facts are being spread about COVID-19 which can negatively affect the society in multiple ways. There is a huge chance that people would believe such falsified information. Providing accurate and verified information from medical experts is therefore, a challenge which needs to be solved. Scrapping data from approved sources and building a system that can credibly answer natural language questions from humans requires a lot of processing and thus, a trivial solution does not exist and thus our work becomes necessary.

This paper would make use of existing models and fine-tune on COVID-QA dataset. The models selected for this tasks include models that are the state-of-the-art models used for QA tasks. This paper also analyzes the results that were generated by the models. The code used in this project can be found on github[1]. The fine-tuned models are available on this page.[2]

2 Related Work

The paper by (Möller et al., 2020) first presented dataset for QA on COVID named as COVID-QA. They made use of the COVID-19 scientific articles. They offered biomedical experts annotated 2019 question-answer pairs related to COVID-19 by considering 147 scientific articles.

They selected RoBERTa architecture (Liu et al., 2019a) and fine-tuned it on the SQuAD dataset as a baseline model. They continued to train their baseline model using their COVID-QA annotations in 5-fold cross validation manner.

The paper made use of two metrics to evaluate their model namely Exact Match and F1 scores. Compared to baseline score of 21.84 in terms of Exact Match, COVID QA model performed better with a score of 25.90. Similarly in terms of F1 scores, COVID QA model outperformed the baseline with a score of 59.53 as compared to 49.43 of

[1] github
[2] huggingface.co/armageddon

Parameter	BERT	RoBERTa	DistilBERT	ALBERT	ELECTRA
Sequence Length	512	512	512	512	512
No. of Trainable Parameters	Base 110M Large 340M	Base 125M Large 355M	Base 65M	Base 11M XXL 223M	Base 110M
Architecture	Encoder	Encoder	Encoder	Encoder	Encoder
No. of Blocks	Base 12 Large 24	Base 12 Large 24	Base 6	Base 12 XXL 12	Base 12
No. of Attention Heads	Base 12 Large 16	Base 12 Large 16	Base 12	Base 12 XXL 64	Base 4

Table 1: Table describing various parameters of the models in this paper

baseline.

From the results it was found that additional training on this domain-specific data leads to significant gains in performance. The overall scores were pretty low compared to SQuAD. It was assumed that low scores relate to more complex question/answer pairs on much longer documents and the lack of multiple annotations per question.The paper lacks in experimentation and just trains on one model. The assumption that it obtains better results when fine-tuned is a valid observation and should be considered in any experimentation.The experimention could have been performed on various models and checked if they resulted in any better metrics.

Another limitation is that they had just considered one model and did not try to obtain the results from other existing SOTA models.The work done by (Saikh et al., 2021) is similar to the (Möller et al., 2020) but they have considered a number of models to train their dataset. They make use of models like BERT, BioBERT and ClinicalBERT. They achieved a max F1 score of 37.19 and an exact match of 28.65. Results shows that ClinicalBERT produces better results than standard BERT and SOTA QA model. Thus it is necessary to have a look at how other models would perform on this dataset. We would look at the various models discussed in the paper.

BERT(Vaswani et al., 2017) is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pretrained BERT model can be fine-tuned with just one additional output layer. BERT is conceptually simple and empirically powerful.

RoBERTa(Liu et al., 2019b) builds on BERT's language masking strategy, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples. RoBERTa, modifies key hyperparameters in BERT,which allows RoBERTa to improve on the masked language modeling objective compared with BERT.

DistilBERT(Sanh et al., 2019) is a small, fast, cheap and light Transformer model based on the BERT architecture. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40%.

A Lite Bert (Sanh et al., 2019)(ALBERT), a deep-learning natural language processing (NLP) model, which uses 89% fewer parameters than the state-of-the-art BERT model, with little loss of accuracy. The model can also be scaled-up to achieve new state-of-the-art performance on NLP benchmarks.

ELECTRA (Clark et al., 2020)(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) is a new pre-training approach which aims to match or exceed the downstream performance of an MLM pre-trained model while using significantly less compute resources for the pre-training stage.

Table 1 succinctly summarizes details of the model architectures such as sequence length, heads, blocks etc.

3 Method

3.1 Dataset

COVID-QA is a QA dataset which was first released in the paper (Möller et al., 2020). It is a SQuAD (Rajpurkar et al., 2016) style annotated dataset which consists of 2019 question answer pairs related to the pandemic. This data was pulled from 147 scientific articles, and was annotated by experts in the biomedical field.

Since we have used the default hyper-parameters used by the authors of the respective models during

fine-tuning, the dataset is split only for train (90%) and test (10%).

3.2 Models

We have used various pre-trained SOTA transformer models which were discussed during the span of the course. We fine-tuned these models on the COVID-QA dataset. The pre-trained models that we have shortlisted for this task span a variety of transformer architectures which gained popularity over the years after BERT (Vaswani et al., 2017) was first proposed. The models are as below:

- BERT (base, large) - pre-trained on SQuAD 2.0 (Vaswani et al., 2017)
- RoBERTa (base, large) - pre-trained on SQuAD 2.0 (Liu et al., 2019b)
- DistilBERT (base) - pre-trained on SQuAD 2.0 (Sanh et al., 2019)
- ALBERT (base, XXL) - pre-trained on SQuAD 2.0 (Sanh et al., 2019)
- ELECTRA (base) - pre-trained on SQuAD 2.0 (Clark et al., 2020)

Despite our original plan to test the performance of long range attention models such as longformer and bigbird, these evaluation were cut short because of hardware restrictions.

3.3 Evaluation

We used the standard evaluation metrics for QA NLP task as listed below:

- **Exact Match:** This metric assigns a value of 1 if the predicted string exactly matches the answer string. In all other cases, it is 0.
- **F1 Score:** This is a more flexible metric which accounts for shared word across the predicted and the actual answer. Let n_s denote the number of shared words, n_p be the number of words in ground truth, and n_y be the number of words in the ground truth. Then:

$$precision = \frac{n_s}{n_p} \tag{1}$$

$$recall = \frac{n_s}{n_y} \tag{2}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

3.4 Pre-processing

The transformer models takes as input a sequence of tokens which were generated using the tokenizer defined for the model. The question and context are fed in a single sequence using special tokens for separation. The input to the model also consists of token type ids, which denote whether a token belongs to the question or the context in the input. Some models also need attention mask ids which can be used to control which tokens the model should attend to. Consider the example sequence below:

[CLS] This is the question ? [SEP] This is the context . [SEP]

For the above sentence, the token type ids are :

0 0 0 0 0 0 0 1 1 1 1 1 1

The output of the model is are a list of pair of logits for each token in the input sequence. Each pair of logit has a value for that token being the answer start and end. Finally, we look at all the logits generated by the model and select two indices within the context i and j such that $i < j$ and $logit(i, start) + logit(j, end)$ is maximum.

Usually the context of the question is quite long, and easily exceeds the maximum sequence length for the input of the transformer model. To address this issue, the context is split into several parts, and using these multiple features are generated. In each new generated feature, the question is present, but only the truncated part of the context is there. During splits, an overlap of tokens is added which is called stride. Setting this correctly is crucial as this can affect the performance of the model. We found empirically that the best performance was achieved with a stride length of 150, which is what we used for tokenization across all the models.

4 Experiments

4.1 Setup

All the models used in these set of experiments were pre-trained on SQuAD v2 dataset. The fine-tuning on the COVID-QA for each model was done on Google v3-8 TPU hardware which has 8 cores and 16GB RAM per core. The training was done for 3 epochs with a batch size of 8 per core. The code was written using PyTorch (Paszke et al., 2019). The *transformers* module from HuggingFace (Wolf et al., 2020) was used to download, fine-tune, and manage the various models. These

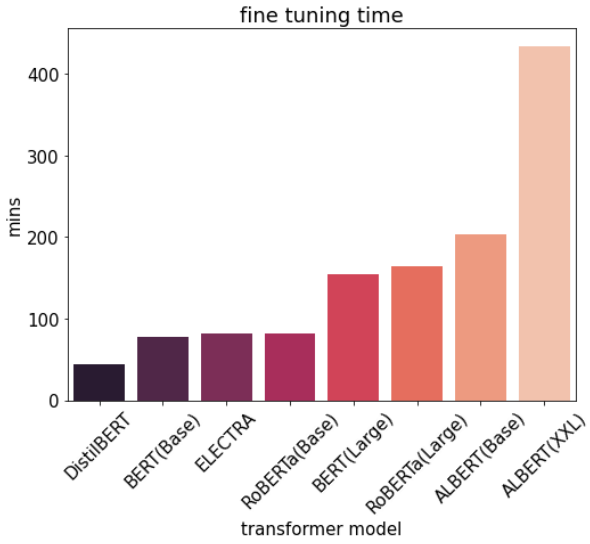

Figure 1: Time taken to fine-tune transformer models for 3 epochs

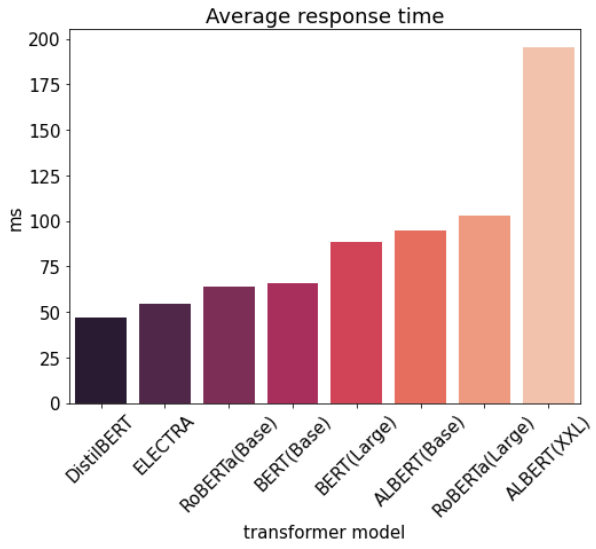

Figure 2: Average response time for each model

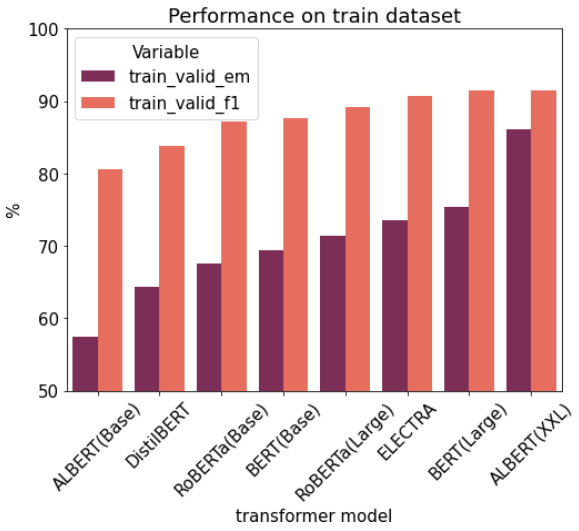

Figure 3: Performance of models on train dataset

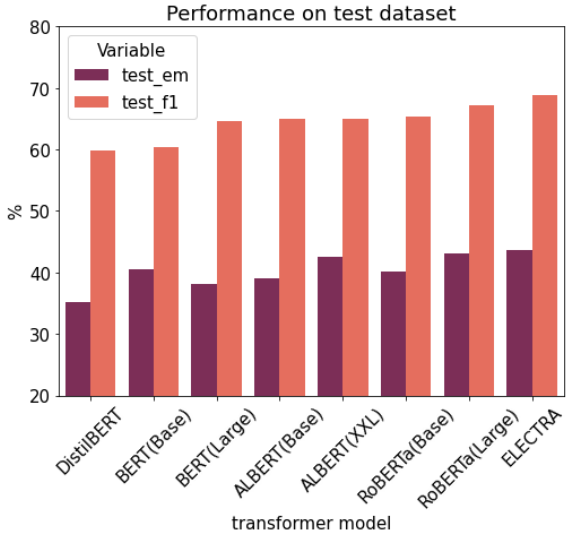

Figure 4: Performance of models on test dataset

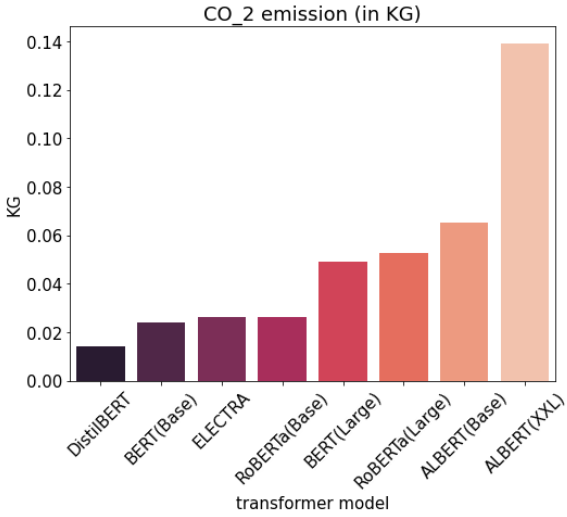

Figure 5: Environmental impact of training the models

models were finally uploaded to their respective model pages on Hugging-Face.

For each experiment, we noted the F1 and exact match scores on both the train and test dataset. We additionally noted down the training time, loss, and estimated carbon impact in CO_2 eq (in KG). The average response time of the models was also noted down.

4.2 Results

The results have been graphically visualized. For the time taken to fine-tune the models (see Figure 1), DistilBERT takes the least amount of time since it is a light version of BERT with fewer parameters, whereas ALBERT(XXL) takes the maximum amount of time. The large models take more time than their base counterparts. Similar kind of result is also observed for the average response time of the model (see Figure 2). We can see that both ALBERT(Base) and ALBERT(XXL) are huge models, requiring more computation than even the Large versions of BERT and RoBERTa.

ALBERT(XXL) quite outperforms the other transformer models when evaluating on the train dataset (see Figure 3), however falls quite behind on the unseen test dataset (see Figure 4). This could be hinting towards overfitting in these humongous models. We find that ELECTRA has the best overall performance. It takes less time to fine-tune, and has one of the lowest response times. It generalizes the best. This model achieves a maximum F1 score of 68.896 and an exact match of 43.564, outperforming the best performance presented in (Möller et al., 2020) by almost 40%.

DistilBERT is a lightweight model, but that also affects its performance, as it performs poorly on both train and test dataset.

Regarding the environmental impact (see Figure 5, we can clearly see that the amount of carbon emission is directly proportional to the number of trainable parameters in the model, with DistilBERT having the least impact, and ALBERT(XXL) having the maximum impact.

5 Conclusion

In this paper, we fine-tuned several SOTA transformer NLP models on the COVID-QA dataset. We use the latest transformer libraries from Hugging-Face, and the best in class v3 TPU hardware. We analyzed the performance of each of the models on the train and test datasets, and reported a critical analysis regarding the various metrics of the model.

In our analysis, we found that ELECTRA is the best overall, it takes less time to fine-tune, and has a small response time. It shows good performance on the train dataset and best performance on the test dataset. We also saw that the the computationally cheapest model is DistilBERT, however, it compromises on performance.

The humongous models such as BERT(Large), RoBERTa(Large), ALBERT(XXL) take a lot of computation power, and perform excellently on train dataset, but poorly on test dataset indicating overfitting and poor generalization capabilities.

Since carbon emission and power consumption of training these huge models is a concern, we also noted down the estimated carbon emission impact of the models.

COVID-19 has greatly impacted the word, and the problems can only be exacerbated with misinformation. A reliable source of truth that could potentially answer the questions of people may go well towards alleviating this issue. This work can be extended by using a lightweight document search mechanism such as tf-idf or Elastisearch to quickly shortlist a set of potential documents which were saved using a web crawler run on trusted sources such as WHO. These set of crudely shortlisted documents can then be used as context when a question is asked.

Acknowledgments

We are really grateful to Professor Ndapa Nakashole, Computer Science Department, University of California, San Diego and the Yutong Shao (TA) for their continuous support, encouragement and willingness to help us throughout this project.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Tanik Saikh, Sovan Kumar Sahoo, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Covidread: A large-scale question answering dataset on COVID-19. *CoRR*, abs/2110.09321.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.