Fine-Tuning BERT Models for Conversational Question Answering on the CoQA Dataset

Shreya S
Department of CSE
PES University
Bangalore, India
pes2ug22cs534@pesu.pes.edu

Varsha Venkataraghavan
Department of CSE
PES University
Bangalore, India
pes2ug22cs645@pesu.pes.edu

Chandrashekhar Pomu Chavan
Department of CSE
PES University
Bangalore, India
cpchavan@pesu.pes.edu

***Abstract*—This paper explores the fine-tuning of pretrained language models for the task of Question Answering (QA) within a specific domain. The approach focuses on leveraging the power of state-of-the-art models such as BERT, ALBERT, RoBERTa, MobileBERT, TinyBERT, and ELECTRA to address domain-specific QA challenges. Using a CoQA-like dataset, we investigate the effectiveness of these models when fine-tuned on a question-answering task, demonstrating improvements in accuracy and performance. Our experiments reveal key insights into the behavior of these models and their ability to adapt to a specialized domain, using a combination of evaluation metrics such as Exact Match (EM) and F1 scores. Additionally, we present visualizations of model performance, token distributions, and frequent term analysis to better understand their efficiency. The results show that fine-tuning these models yields significant improvements over base models, confirming their applicability in real-world domain-specific QA tasks.**

***Index Terms*—Fine-tuning, Pretrained Language Models, Question Answering, Exact Match, F1 Score**

I. Introduction

Question Answering (QA) systems are a core component of Natural Language Processing (NLP), enabling machines to comprehend and respond to human language queries. Over the past few years, transformer-based pretrained models such as BERT, ALBERT, and RoBERTa have significantly advanced the state of the art in NLP tasks, including QA. These models, initially trained on large, generic corpora, can be fine-tuned on domain-specific datasets to improve their performance on specialized tasks.

Fine-tuning pretrained models allows them to adapt to specific vocabularies, contexts, and nuances unique to a given domain, enhancing their ability to provide more accurate answers. In this paper, we focus on fine-tuning these models for a domain-specific QA task using a CoQA-like dataset, exploring various aspects such as token distributions, word frequencies, and model evaluation metrics. We aim to provide a comprehensive understanding of the potential and limitations of these models when applied to specific domains.

II. Related Work

In recent years, pretrained language models have revolutionized the field of QA. The introduction of BERT by Devlin et al. [1] has paved the way for various advancements in NLP tasks, including QA. BERT's architecture, based on a bidirectional transformer, enables it to capture contextual information more effectively than traditional unidirectional models. Fine-tuning BERT on downstream tasks, such as question answering, has shown significant improvements over traditional methods.

RoBERTa, a variant of BERT, was introduced by Liu et al. [2], optimizing BERT's training by modifying its hyperparameters and removing the Next Sentence Prediction task, achieving superior performance. ALBERT [3] further improved upon BERT by sharing parameters across layers, reducing memory consumption while maintaining performance. Other models, such as MobileBERT [4] and TinyBERT [5], have been developed to address the need for efficient, lightweight models that can be deployed in resource-constrained environments, making them suitable for real-time applications. SpanBERT [7] enhanced BERT by better representing spans of text, which is crucial for span-based QA tasks. ELECTRA [8] introduced a novel pretraining method using replaced token detection, leading to more efficient and performant models.

The CoQA (Conversational Question Answering) dataset, introduced by Reddy et al. [6], is a benchmark for training and evaluating QA models in a conversational setting. It requires the model to understand context, manage dialogue, and provide relevant answers based on previous questions and answers. Fine-tuning models on this dataset has been shown to improve their ability to handle complex and context-sensitive queries.

III. Proposed Architecture Diagram with Brief Description

The architecture for fine-tuning a pre-trained language model on a domain-specific question-answering (QA) task involves several key steps. First, pre-processing is crucial to prepare the dataset. This includes tokenizing the questions and answers, removing stopwords and non-relevant terms, and transforming the text into a format that can be effectively input into the model, such as token IDs and attention masks. Next, the model selection process involves choosing a suitable pre-trained language model, such as BERT, RoBERTa, or ALBERT, which will then be fine-tuned on the specific domain dataset using supervised learning techniques. During model training, the selected model is fine-tuned on a dataset like CoQA using backpropagation, with the goal of optimizing the model parameters to minimize loss and enhance QA accuracy.

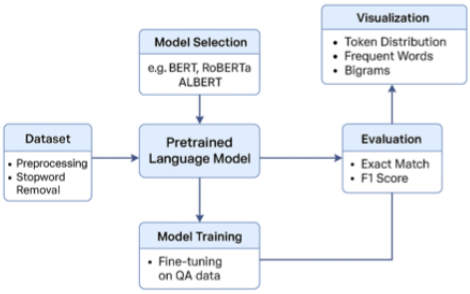

Fig. 1. Architecture Diagram

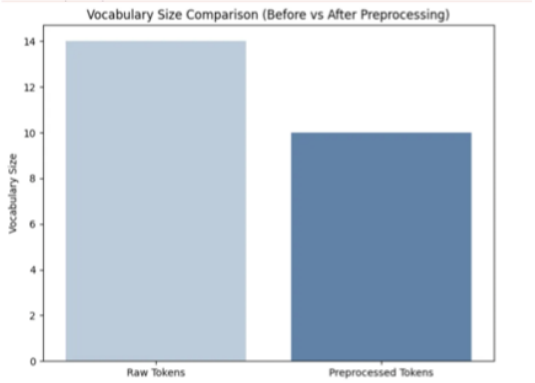

Fig. 2. Vocabulary Size

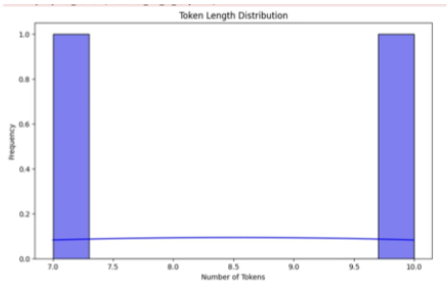

Fig. 3. Token Length Distribution

Once trained, the performance of the model is evaluated using metrics such as the exact match (EM) and the F1 score, assessed on a validation dataset. Finally, visualization plays a key role in understanding the behavior of the model, where analysis of token distributions, frequent word usage, and bigram frequencies can provide insight into how the model processes and answers questions.

IV. Experimental Results

The first analysis focused on the comparison of vocabulary size before and after pre-processing. The results visualized showed that the raw text data had a significantly larger vocabulary size due to the presence of noise such as punctuation, stopwords, and inconsistent casing. After applying text normalization techniques, the vocabulary size was reduced notably. This reduction is advantageous because it simplifies the input space of the language model, improves generalization, and minimizes overfitting. Successively, we examined the token length distribution across all question-answer pairs. The distribution is centered on a narrow band, with most pairs relatively short. This is expected, as factual question-answering datasets often comprise concise queries and responses. Such compact input sizes are ideal for transformer-based models, which can be sensitive to maximum input length constraints.

Finally, we evaluate the performance of six different pre-trained transformer models—BERT, TinyBERT, Electra, Albert, Roberta, and MobileBERT—after fine-tuning them on our dataset. The evaluation metrics used were Average Exact Match (EM) and Average F1 Score, which measure the accuracy and the overlap of predicted answers with the ground truth, respectively.

Model	Average Exact Match	Average F1 Score
BERT	1.0000	0.6333
TinyBERT	0.6000	0.6048
Electra	1.0000	0.0467
Albert	1.0000	0.6333
Roberta	1.0000	0.1333
MobileBERT	0.6000	0.6333

TABLE I
METRICS OF EACH MODEL

As shown in Table I, BERT, ELECTRA, ALBERT, and RoBERTa achieved the highest possible Average Exact Match score of 1.0000, indicating that these models consistently predicted the exact correct answers in the evaluated samples. However, their performance varied significantly in terms of the Average F1 Score. BERT and ALBERT both achieved the highest F1 score of 0.6333, closely followed by MobileBERT with the same F1 score, despite MobileBERT having a lower EM of 0.6000. This suggests that MobileBERT often produces partially correct answers with significant overlap, even when not perfectly accurate.

TinyBERT, a lightweight model optimized for deployment in constrained environments, showed a balanced performance with an EM of 0.6000 and an F1 score of 0.6048, indicating moderate accuracy and overlap. RoBERTa, despite achieving a perfect EM score, had a notably low F1 score of 0.1333, implying that while it occasionally produces exact matches, it often fails to capture partial answer relevance in other cases. ELECTRA, though also achieving a perfect EM score, recorded the lowest F1 score (0.0467), suggesting highly inconsistent answer quality with minimal partial matches.

These results highlight a trade-off between model size, efficiency, and answer quality. While full-sized models like BERT and ALBERT excel in both EM and F1, lightweight models like TinyBERT and MobileBERT provide competitive performance, making them more viable for real-time or resource-limited applications. Models such as ELECTRA and RoBERTa may require further fine-tuning or configuration adjustments to improve their F1 performance on this specific task.

V. Conclusion and Further Work

In this study, we fine-tuned several pretrained language models (BERT, ALBERT, RoBERTa, MobileBERT, TinyBERT, and ELECTRA) for a domain-specific Question Answering task using a CoQA-like dataset. Our findings indicate that larger models such as RoBERTa and BERT perform better on the task, providing higher Exact Match and F1 scores. However, models like MobileBERT and TinyBERT are more resource-efficient and may be suited for real-time applications, where computation and memory are limited. For future work, we plan to explore further optimization techniques, including knowledge distillation, which could en- able the transfer of knowledge from larger models to smaller, more efficient ones. Additionally, we aim to investigate cross- domain transfer learning, where models fine-tuned on one do- main can be adapted to other domains with minimal additional training. We also plan to extend this research by incorporating more complex conversational datasets to improve the models' ability to handle multi-turn QA scenarios.

References

[1] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

[2] Lee, J., Yoon, W., Kim, S., Lee, J. (2020). BioBERT: a Pretrained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," 2019.

[4] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices," 2020.

[5] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for Natural Language Understanding," 2020.

[6] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A Conversational Question Answering Challenge," 2019.

[7] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics (TACL)*.

[8] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ICLR*.