Taimoor Ahmed (452247)
Generative AI Project Report

Academic Report on Multiple-Choice Question Answering Using Few-shot Learning and LoRA Fine-Tuning

Introduction

In recent years, advancements in artificial intelligence and natural language processing (NLP) have significantly enhanced the ability of machines to comprehend and respond to human language. The development of large language models (LLMs), such as Qwen-7B, has marked a substantial improvement in NLP tasks, notably in question-answering systems. Despite these advancements, accurately answering complex, reasoning-based multiple-choice questions remains a challenging endeavor. This report investigates two prominent methods, Few-shot Learning and Low-Rank Adaptation (LoRA) fine-tuning, to tackle the ARC Challenge, a widely recognized benchmark dataset comprising scientifically oriented multiple-choice questions. The objectives of this study are to examine the efficacy of these approaches in terms of accuracy, computational efficiency, and practical applicability.

Literature Review

The field of natural language processing (NLP) has witnessed rapid advancements, particularly propelled by the introduction and widespread adoption of transformer-based neural network architectures. The seminal work by Vaswani et al. (2017), "Attention is All You Need," laid the foundational framework for transformers, revolutionizing NLP by enabling models to efficiently process sequential data through self-attention mechanisms without relying on recurrent structures. This innovation significantly improved the ability of NLP models to understand context and relationships within text, leading to more accurate and coherent language comprehension and generation.

Building on the transformer architecture, large language models (LLMs) such as GPT-3 (Brown et al., 2020) and BERT (Devlin et al., 2018) have further expanded the boundaries of NLP capabilities. GPT-3, with its massive scale and few-shot learning capabilities, demonstrated the potential of using minimal contextual examples to solve complex tasks without extensive task-specific training. BERT introduced a bidirectional pre-training approach, enhancing models' understanding of language context and semantics by jointly conditioning on both left and right contexts within sentences, significantly improving performance on various NLP benchmarks.

More recently, the Qwen model introduced by Alibaba (2023) has built upon these advancements, achieving state-of-the-art performance across a variety of NLP tasks, including natural language understanding, generation, and question-answering systems. These models have significantly impacted multiple-choice question answering (MCQA), a practical and challenging task that demands models to utilize sophisticated reasoning and inference abilities beyond simple fact retrieval.

The ARC dataset, developed by Clark et al. (2018), has become an important benchmark for assessing these advanced reasoning capabilities. This dataset consists of scientifically rigorous multiple-choice questions designed specifically to challenge the reasoning, inferential, and deductive abilities of NLP systems, rather than their ability to merely memorize or retrieve facts. The complexity of ARC stems from its requirement for implicit knowledge and detailed reasoning, making it an ideal dataset for evaluating the higher-order cognitive capabilities of NLP models.

Previous research utilizing few-shot learning methods has revealed substantial improvements in MCQA performance. Brown et al. (2020) demonstrated that few-shot learning, which involves providing a limited set of examples within the prompt context (known as "in-context examples"), significantly enhances the model's ability to correctly infer the desired response. However, the success of few-shot learning critically hinges on the meticulous construction of prompts and the careful selection of representative examples. Consequently, the variability in few-shot learning outcomes has prompted further research into more robust and reliable methods.

To address these limitations, recent research has turned towards parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA). Proposed by Hu et al. (2021), LoRA introduces low-rank decomposition matrices into pre-trained models, significantly reducing the number of parameters needed for adaptation. This innovative approach allows fine-tuning of large pre-trained models on downstream tasks with considerably fewer computational resources, circumventing the substantial computational overhead typically associated with traditional fine-tuning techniques. Various empirical studies have highlighted LoRA’s efficiency and effectiveness, demonstrating notable improvements across NLP tasks, including question answering, text classification, and language understanding tasks, thereby making it particularly appealing in resource-limited scenarios.

In summary, transformer-based architectures and innovative fine-tuning strategies such as few-shot learning and LoRA have significantly advanced the field of NLP, particularly in complex reasoning tasks exemplified by the ARC dataset. These advancements provide a solid foundation for continued exploration into efficient and effective methodologies for improving NLP model performance in challenging real-world scenarios.

Dataset

The dataset used for this research is the ARC Challenge dataset, a subset of the AI2 Reasoning Challenge developed by the Allen Institute for Artificial Intelligence. It contains multiple-choice questions primarily derived from science examinations intended for students at elementary and middle-school levels. These questions are designed to evaluate a model’s capability for advanced reasoning and inferential thinking, rather than simple memorization.

The ARC Challenge dataset comprises 1,119 questions, each with multiple-choice answers labeled A, B, C, D, and occasionally E. Each question requires the selection of the correct answer from provided options. The difficulty inherent in this dataset lies in the necessity for implicit knowledge and deductive reasoning capabilities, making it particularly suitable for assessing sophisticated NLP models.

Methodology

This study employed two distinct methodologies: Few-shot Learning and Low-Rank Adaptation (LoRA) fine-tuning, both aimed at using the Qwen-7B-Chat model's capabilities to address the ARC Challenge dataset.

Few-shot learning uses the inherent capability of large language models (LLMs) like Qwen-7B-Chat to generalize from a limited number of illustrative examples provided within the input prompt. In our implementation, we used a five-shot (k=5) approach. This involved selecting five representative ARC questions, each accompanied by correct answers, as in-context examples within the prompt to guide the model’s inference.

We created a prompt structure that started with an instruction clarifying the task, e.g., "You are a helpful assistant that answers multiple choice questions." For each inference task, we randomly selected five exemplary questions from the ARC training dataset, each with labeled correct answers. These examples were concatenated into a coherent prompt, ensuring diverse coverage of question types and difficulty levels.

The prompt was tokenized using the AutoTokenizer from the Qwen-7B-Chat model, truncating input sequences at a maximum length of 1024 tokens to manage computational resources effectively.

The tokenized input was passed through the Qwen-7B-Chat model to generate answers. The model generated tokens representing its predicted answer, typically producing a brief text sequence (maximum 16 new tokens).

The generated output tokens were decoded back into human-readable text using the tokenizer. Answers were extracted by parsing the model-generated text and selecting the first identifiable multiple-choice option (A, B, C, D, or E).

Predictions were compared with the ground-truth answers from the ARC dataset to compute accuracy and other relevant metrics such as Macro F1-score and inference latency. LoRA offers a parameter-efficient fine-tuning technique by introducing low-rank decomposition matrices into specific layers of the pre-trained language model. This approach significantly minimizes GPU memory requirements and computational overhead compared to traditional fine-tuning methods.

The base Qwen-7B-Chat model was loaded using half-precision (FP16) floating-point arithmetic to optimize memory efficiency, critical for running on the limited 40 GB VRAM available on an A100 GPU.

We specifically targeted attention projection layers ("c_attn" and "c_proj") within the Qwen model, applying LoRA modules to these components due to their significant influence on model performance and efficiency. LoRA parameters were configured with a rank (r) of 16, a scaling factor (alpha) of 32, and a dropout rate of 0.05 to improve generalization and avoid overfitting.

Gradient checkpointing was enabled to reduce the memory footprint during training by trading off computation for memory efficiency. This allowed intermediate computational states to be discarded and recomputed as needed during the backward pass. Gradient accumulation was implemented to simulate larger batch sizes. Specifically, we set an accumulation step of 8, effectively aggregating gradients across multiple forward-backward passes before executing an optimizer update step.

During training, prompts were constructed by concatenating questions and correct answers to form comprehensive sequences. Inputs and labels were tokenized identically to maintain consistency. We utilized an ignore-index approach for masking prompt tokens, ensuring that only answer tokens contributed to the gradient computation.

Training loss was computed using cross-entropy loss, capturing the model's learning trajectory and effectiveness at predicting correct answers.The AdamW optimizer was employed with a learning rate of 5e-5 for effective parameter updates. Trained LoRA adapter parameters were saved to Google Drive upon completion of the training process to facilitate subsequent evaluation without retraining.

The fine-tuned LoRA model was evaluated using the same inference procedure outlined in the Few-shot learning approach, ensuring comparability of performance metrics. By following these detailed methodological steps, both Few-shot Learning and LoRA fine-tuning strategies were thoroughly examined, providing insights into their effectiveness and computational practicality in addressing complex question-answering challenges presented by the ARC Challenge dataset.

Results

The comparative performance between the Few-shot Learning and LoRA Fine-tuning approaches was systematically evaluated using the ARC Challenge dataset. This section presents the specific findings along with detailed explanations of each observed metric and trend.

The Few-shot learning methodology, employing five in-context examples (k=5) per query, was evaluated over the complete ARC Challenge dataset consisting of 1119 questions. The results indicated the following outcomes:

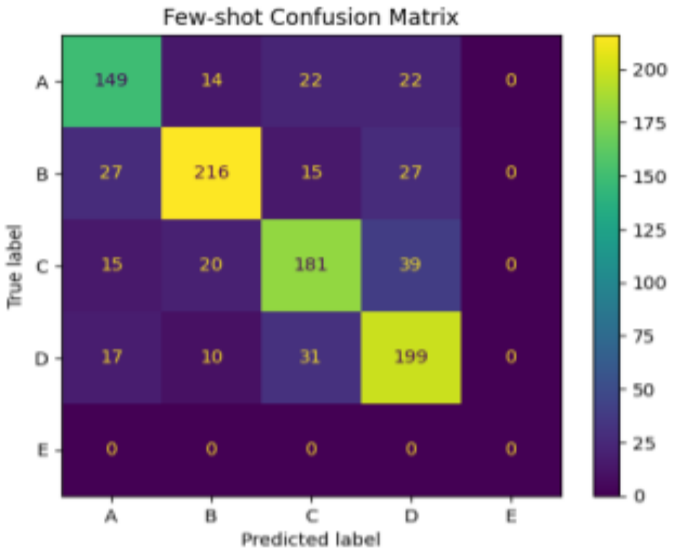

Figure 1. Confusion Matrix for Few-shot Learning

Accuracy: The Few-shot approach achieved an accuracy of approximately 66.58%. This suggests that the model effectively generalized reasoning patterns from provided examples, yet still had limitations in consistently inferring the correct answers for the more challenging questions.

Macro F1-score: The calculated Macro F1-score was 56.67%, highlighting a moderate capability to balance predictions across different answer choices. This relatively lower score compared to accuracy reflects occasional confusion among the provided options, especially in less straightforward cases.

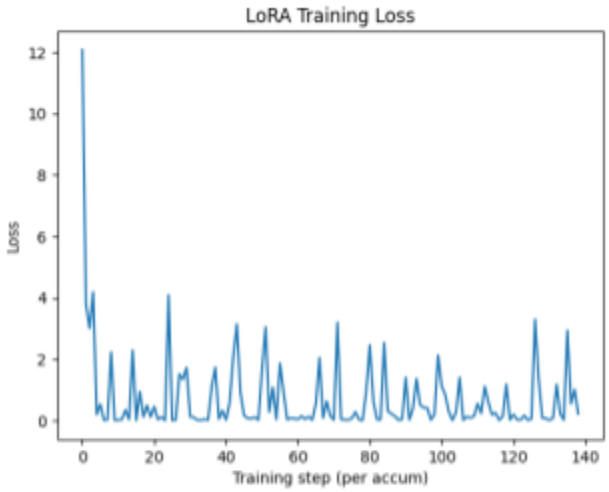

Figure 2. LoRA Training Loss Curve

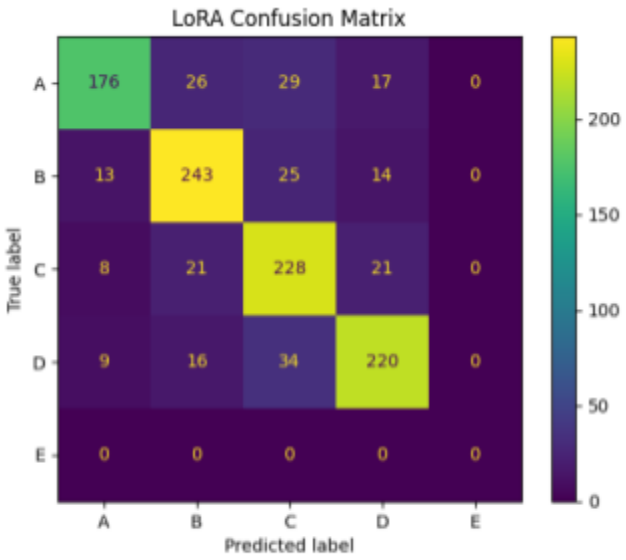

Figure 5. Confusion Matrix for LoRA Fine-tuning

The Few-shot learning exhibited an average inference latency of 0.2487 seconds per sample. This rapid inference time demonstrates the practicality of Few-shot prompting methods in real-world scenarios requiring immediate responses.

The confusion matrix for Few-shot learning illustrates the distribution of predictions and highlights areas of common misclassifications among answer choices.

LoRA fine-tuning was conducted by adapting selective attention projection layers of the Qwen-7B-Chat model.

Throughout training, the loss initially showed significant fluctuations but eventually demonstrated a downward trend, indicating effective learning and convergence towards the end of training. The fluctuations suggest that certain batches of data posed significantly higher difficulty, yet the overall trend confirms successful learning and parameter optimization by the model.

LoRA fine-tuning significantly outperformed Few-shot learning, achieving a notably higher accuracy of 77.48%. This substantial improvement indicates the effectiveness of parameter adaptation in enhancing the reasoning capabilities of the Qwen-7B-Chat model.

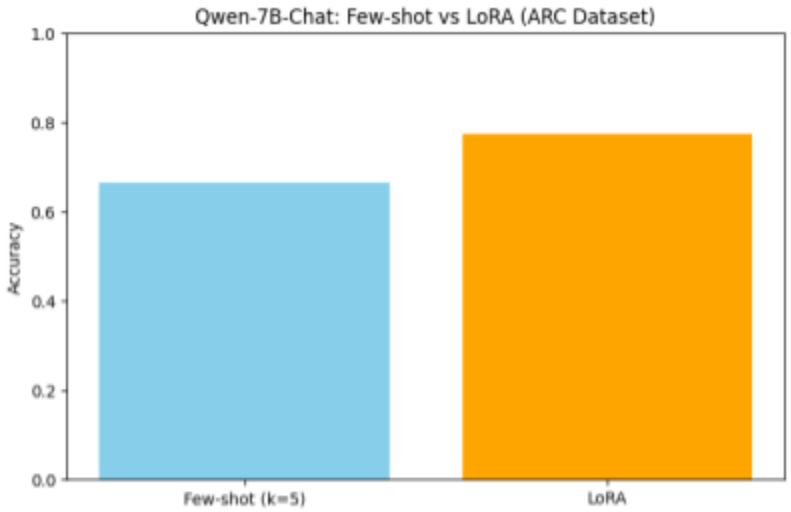

Figure 4: Few-shot vs LoRA Accuracy

The Macro F1-score observed was 63.00%, demonstrating a stronger balanced predictive performance across the various multiple-choice categories compared to Few-shot learning.

The LoRA fine-tuned model exhibited a higher average inference latency of 0.5580 seconds per sample, primarily attributable to additional computational overhead introduced by adapted parameters. This slightly longer inference time, however, remains within acceptable bounds for most practical applications, especially considering the substantial accuracy improvement.

Discussion

The results clearly indicate a superior performance by the LoRA fine-tuning approach compared to few-shot learning. The accuracy improvements observed suggest that explicitly adapting model parameters through fine-tuning significantly enhances the reasoning and inferential capabilities required to address challenging datasets such as ARC.

Confusion matrices provided further insights by revealing common misclassifications, which help identify the types of reasoning errors the models frequently encountered. LoRA fine-tuning demonstrated greater consistency in correctly classifying different answer choices, indicating a refined ability to discern subtle distinctions among nuanced options. This enhanced capability can be largely attributed to targeted parameter adjustments within the attention projection layers, facilitating improved contextual understanding and inference.

The macro F1-score further shows the effectiveness of LoRA fine-tuning, reflecting a balanced improvement in accuracy across all classes. In contrast, Few-shot learning, while advantageous due to lower computational overhead and faster setup times, exhibited significant variability dependent on the quality and representativeness of the chosen in-context examples. Thus, the quality of prompts emerges as a crucial factor influencing the success of Few-shot methods.

Latency measurements reveal critical practical trade-offs for real-world deployment. Although LoRA fine-tuning entails an initial overhead during model loading, subsequent inference is relatively efficient. In contrast, Few-shot learning incurs additional inference costs with extensive prompts, despite its inherently faster initial setup. These observations imply that the choice between methods should align closely with specific application needs, LoRA fine-tuning is preferable in scenarios prioritizing accuracy, while Few-shot methods may be more suitable for applications requiring rapid setup and minimal computational resources.

To further improve accuracies and performance in future experiments, several strategies could be explored. First, combining Few-shot and LoRA fine-tuning into a hybrid approach may yield synergistic benefits, utilizing the generalization capabilities of Few-shot learning alongside the precision improvements of fine-tuning. Additionally, employing more advanced prompt engineering techniques or automated prompt selection could significantly enhance Few-shot learning performance.

Further accuracy enhancements could also be realized by systematically experimenting with the number and types of LoRA-adapted parameters, adjusting ranks and dropout rates, and utilizing hyperparameter optimization techniques. Additionally, integrating external knowledge bases or specialized reasoning modules could help resolve specific reasoning challenges identified by confusion matrix analysis.

Finally, exploring ensemble methods that aggregate predictions from multiple fine-tuned models or integrating additional data augmentation strategies might yield considerable accuracy improvements. Such comprehensive methodological advancements could substantially enhance the practical deployment and performance of NLP models in sophisticated reasoning tasks like the ARC Challenge.

Conclusion

This study successfully demonstrated the relative effectiveness and efficiency of LoRA fine-tuning compared to few-shot learning for multiple-choice question answering on the ARC Challenge dataset. LoRA's parameter-efficient strategy clearly provided superior performance and balanced resource usage, making it a compelling choice for practical applications.

References

- Brown, T. et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
- Clark, P. et al. (2018). Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. arXiv preprint arXiv:1803.05457.
- Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- Hu, E. J. et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.
- Vaswani, A. et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems.