Comparative Analysis of RAG vs Fine-tuning Approaches Using SQuAD Dataset

Authors:

- Yadhukrishnan Pankajakshan, M.S. Computer Science
- Harshal Jorwekar, M.S. Data Science

Project Description

This project aims to conduct a systematic comparison between Retrieval-Augmented Generation (RAG) and Fine-tuning approaches for question-answering tasks. SQuAD presents an ideal testing ground as it contains 100,000+ questions posed by crowdworkers on Wikipedia articles, where answers are specific text segments from the corresponding passages. The study will specifically analyze the trade-offs between these two approaches in terms of accuracy, computational efficiency, and scalability.

Retrieval-Augmented Generation (RAG) and Fine-tuning represent two distinct approaches to enhancing Large Language Model (LLM) capabilities. RAG dynamically augments LLM responses by retrieving relevant information from external knowledge bases, enabling real-time access to up-to-date information without model retraining. In contrast, Fine-tuning involves adapting pre-trained models to specific tasks through additional training on specialized datasets, potentially offering better performance for focused applications but requiring periodic updates to maintain relevance.

Our primary research questions are:

1. Which approach provides better accuracy and consistency in question-answering tasks?
2. How do these approaches compare in terms of computational resources and scalability?
3. What are the trade-offs between model size, training time, and performance?

Summary of the Data

We will utilize the SQuAD v2.0 dataset, which contains 107,785 question-answer pairs spanning 536 articles. The dataset includes:

- Wikipedia articles as source documents
- Human-generated questions
- Answer spans within the documents
- Validation sets for accuracy measurement
- Questions that may be unanswerable (SQuAD 2.0 feature)

Data Distribution:

- Training set: 87,599 questions
- Development set: 10,570 questions
- Test set: 9,616 questions

The dataset is well-structured and doesn't require significant preprocessing, though we'll need to handle the answer span indexing appropriately. Initial analysis shows balanced distribution across different question types and consistent answer span lengths.

Methods

Our methodology comprises three main components:

1. RAG Implementation:
    - Vector embedding of Wikipedia articles using BERT-based embeddings
    - Implementation of a retrieval mechanism using FAISS
    - Integration with an open-source LLM (likely LLaMA-7B)
    - Prompt engineering for consistent evaluation
2. Fine-tuning Implementation:
    - Base model selection (BERT-large)
    - Parameter-efficient fine-tuning using LoRA
    - Implementation of span prediction mechanism
    - Optimization for question-answering task
3. Evaluation Framework:

```
def evaluate_model(predictions, ground_truth):
    metrics = {
        'exact_match': calculate_em_score(predictions, ground_truth),
        'f1_score': calculate_f1_score(predictions, ground_truth),
        'response_time': measure_response_time(),
        'resource_usage': track_resource_utilization()
    }
    return metrics
```

Preliminary Results

Initial experiments with a subset of SQuAD show that the baseline logistic regression model achieves an F1 score of 51.0%[1], while human performance stands at 86.8%[2] F1 score based on inter-annotator agreement.

References

1. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2383-2392.
2. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 784-789.