Efficient Fine-Tuning of Large Language Models for Emotion Detection: A Comparative Study

Homework 7

Applied NLP

G, Likith Vinayaka
LXG210034

Introduction:

The three methodologies employed in this project are related to fine-tuning large language models for emotion detection in text data. Emotion detection is treated as a multi-label classification problem, aiming to predict the presence or absence of 11 different emotions in a given text input. The methodologies are as follows:

1. LoRA (Low-Rank Adaptation):

Weights & Biases Project: gemma-lora

- A parameter-efficient fine-tuning technique that introduces a small number of trainable parameters to adapt a pre-trained language model.
- Fine-tunes the google/gemma-1.1-2b-it model.

2. IA3 (Instruction-Aware Architecture Adaptation):

Weights & Biases Project: gemma-ia3

- Another parameter-efficient fine-tuning method that specifically targets the attention mechanism in transformer models.
- Also fine-tunes the google/gemma-1.1-2b-it model.

3. QLoRA (Quantized LoRA):

Weights & Biases Project: gte-qwen-qlora

- A variation of LoRA that combines it with quantization techniques to further reduce the memory footprint.
- Fine-tunes the Alibaba-NLP/gte-Qwen1.5-7B-instruct model from the MTEB leaderboard.

Methodologies:

LoRA, IA3, and QLoRA are parameter-efficient fine-tuning techniques designed to adapt pre-trained language models for specific tasks. They differ in their approach to introducing trainable parameters and optimizing model performance while reducing computational and memory requirements.

Architectures:

1. google/gemma-1.1-2b-it:

This is a large language model developed by Google, with 1.1 billion parameters. It is a transformer-based model pre-trained on a large corpus of text data.

2. Alibaba-NLP/gte-Qwen1.5-7B-instruct:

This is a even larger language model developed by Alibaba, with 7 billion parameters. It is also a transformer-based model, but it is specifically designed for instruction-following tasks, making it potentially well-suited for emotion detection.

Data Preprocessing:

In all three methodologies, the data preprocessing steps involve loading the training data from a CSV file, splitting it into training, validation, and test sets, and tokenizing the text using pre-trained tokenizers from the Hugging Face Transformers library. Additionally, a subset of 1,000 samples is created from the training set for faster experimentation.

The text data is tokenized using the corresponding pre-trained tokenizers (AutoTokenizer) from the Hugging Face Transformers library. The tokenized data is then converted into the format required by the Hugging Face Datasets library.

Results and Observations/Metrics:

1. LoRA:
- Validation F1-macro: 0.7472527987876625
- Validation Accuracy: 0.17424629535002556
- Validation F1-micro: 0.8407736021430452

2. IA3:
- Validation F1-macro: 0.6425829295537843
- Validation Accuracy: 0.06710952137625617
- Validation F1-micro: 0.7856645143308404

3. QLoRA :
- Validation F1-macro: 0.306213 (at step 40)
- Validation Accuracy: 0.010901 (at step 40)
- Validation F1-micro: 0.394021 (at step 40)

Comparison between the three approaches:

LoRA demonstrated the highest performance across all evaluated metrics, followed by IA3. QLoRA, while promising, showed lower performance, possibly due to the different and larger base model used.

Among the three approaches, LoRA achieved the highest validation F1-macro score of 0.7472, indicating better overall performance in the multi-label emotion detection task. IA3 had a slightly lower F1-macro score of 0.6425, but it still outperformed QLoRA, which had an F1-macro of 0.306213 at step 40.

In terms of validation accuracy, LoRA again performed the best with 0.1742, followed by IA3 with 0.0671, and QLoRA with 0.010901 at step 40.

For validation F1-micro, LoRA achieved the highest score of 0.8407, followed by IA3 with 0.7856, and QLoRA with 0.394021 at step 40.

It's important to note that the QLoRA approach was applied to a different and larger model (Alibaba-NLP/gte-Qwen1.5-7B-instruct), which could partially explain its lower performance compared to LoRA and IA3, so the final performance might improve with further training.

Challenges Encountered:

- **Limited Computing Power:** The computational demands of the models, especially gte_qwen_1.5, made it challenging to train on the entire dataset.
- **Data Size:** Due to computational limitations, only a subset of the data could be used for training, affecting the model's performance.
- **Model Complexity:** The complex architectures of the models required extensive tuning and longer training times, making them computationally expensive.

Observations and Conclusion:

Based on the provided results, the LoRA approach appears to be the most effective method for fine-tuning the google/gemma-1.1-2b-it model for emotion detection. It achieved the highest scores across all evaluated metrics (F1-macro, accuracy, and F1-micro) on the validation set.

The IA3 approach also performed well, although slightly lower than LoRA. It demonstrates the effectiveness of targeting the attention mechanism for this task.

The QLoRA approach showed promising results but underperformed compared to LoRA and IA3. However, it's important to consider that the model used (Alibaba-NLP/gte-Qwen1.5-7B-instruct) was different and significantly larger than the one used in the other two methodologies.

Overall, the LoRA and IA3 techniques proved to be effective parameter-efficient fine-tuning methods for emotion detection using the google/gemma-1.1-2b-it model. The QLoRA approach, while promising, may require further fine-tuning and optimization to achieve better performance on the larger Alibaba-NLP/gte-Qwen1.5-7B-instruct model.

It's worth noting that these observations are based on the provided methodologies and results, and further experiments and evaluations on different datasets and settings may be necessary to draw more comprehensive conclusions.