Fine-Tuning Large Language Models

Summer

Introduction

Fine-tuning adapts pre-trained language models to specific tasks, leveraging their pre-trained knowledge for better performance compared to training from scratch. This document summarizes the considerations, methods, and challenges in fine-tuning large language models (LLMs).

Model Scale Considerations

- **Small Models**: Full fine-tuning is feasible on a single GPU.
- **Large Models**: Require multi-GPU setups and efficient fine-tuning methods due to memory constraints.

Data Requirements

- Factors influencing data needs:
  - Task complexity.
  - Domain similarity to pre-training data.
  - Model size.
- Emphasize quality over quantity: Clean, well-labeled, and domain-relevant data with balanced classes.

Fine-Tuning Techniques

Parameter-Efficient Fine-Tuning (PEFT)

- Tunes only a small subset of additional parameters, addressing memory and efficiency challenges.
- **Methods**:
  - Adapters: Linear projections inserted between model layers.
  - Prefix Tuning: Learns continuous task-specific vectors prepended to input.
  - Prompt Tuning: Learns soft prompts at input level.

  - Low-Rank Adaptation (LoRA): Updates low-rank matrices representing weight updates.
  - QLoRA: Combines LoRA with quantization techniques to reduce memory further.

Few-Shot Learning with SetFit

- Fine-tunes Sentence Transformers with as few as 8 samples per class.
- Combines embeddings from fine-tuned transformers with a lightweight classifier.

Challenges and Solutions

- **Challenges**:
  - High hardware requirements for large models.
  - Risk of catastrophic forgetting during fine-tuning.
- **Solutions**:
  - Gradient accumulation and checkpointing for memory efficiency.
  - Mixed precision training (e.g., FP16/BF16).
  - Flash Attention and optimized optimizers like Adafactor.

Model Merging

- Combines multiple fine-tuned models into a single model.
- Methods include task arithmetic, TIES (interference mitigation), and DARE (Drop and Rescale).

References

- Tunstall et al. (2022). *Efficient Few-Shot Learning Without Prompts.*
- Dettmers et al. (2024). *QLoRA: Efficient Finetuning of Quantized LLMs.*
- Houlsby et al. (2019). *Parameter-Efficient Transfer Learning for NLP.*