Post-Training and Alignment of Large Language Models

Summer

Introduction

Post-training and alignment aim to adapt pretrained large language models (LLMs) for improved utility, preference alignment, and task-specific performance. This document provides an overview of algorithms, datasets, evaluation strategies, and emerging trends.

Why Post-Train?

- Enhance pretrained models with supervised fine-tuning (SFT) for task-specific adaptation.
- Align model outputs with human preferences through methods like preference optimization and reinforcement learning.
- Achieve better performance on domain-specific tasks.

Key Algorithms for Post-Training

Reinforcement Learning with Human Feedback (RLHF)

- Fine-tune models using supervised learning with prompt-response pairs.
- Train a reward model to evaluate response quality.
- Optimize outputs using reinforcement learning to maximize reward scores.
- Challenges:
  - Computational cost and instability.
  - Requires multiple LLMs (base model, SFT model, reward model).

Direct Preference Optimization (DPO)

- Replaces RLHF with direct optimization of binary preferences (good vs. bad responses).
- Faster and less resource-intensive than RLHF.
- Useful for verifiable outputs (e.g., math, code).

Other Preference Optimization Methods

- **Odds-Ratio Preference Optimization (ORPO)**: Maximizes the odds ratio of preferred responses.
- **Kahneman-Tversky Optimization (KTO)**: Models human preferences with binary labels, inspired by behavioral economics.

Datasets for Post-Training

- **Synthetic Data Generation**:
  - Seed prompts with strong LLMs (e.g., GPT-4) for task coverage.
  - Generate completions and filter for quality and diversity.
- **Data Distillation**:
  - Extract and refine data directly from existing LLMs.
  - Tools like Magpie and UltraFeedback enhance data collection.

Evaluation Strategies

- **Model-Free Evaluation**: Focus on specific tasks (e.g., math, code).
- **Model-Based Evaluation**: Use LLMs to judge other models' outputs (e.g., AlpacaEval).
- **Human Evaluation**: Gold standard but expensive and subjective.
- Address biases (e.g., positional, length biases) in model-based evaluation.

Case Study: Zephyr

- Base model: Mistral 7B.
- Methods: Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO).
- 100% synthetic data with DeepSpeed ZeRO-3 optimization.
- Results: Competitive performance with reduced training time and resources.

Emerging Trends

- Combining model-based and model-free evaluation for robust insights.
- Advances in preference optimization for more nuanced alignment.
- Increased reliance on synthetic datasets and lightweight tuning techniques.

References

- Ouyang et al. (2022). *Training language models to follow instructions with human feedback.*
- Rafailov et al. (2023). *Direct Preference Optimization.*
- Ethayarajh et al. (2024). *Kahneman-Tversky Optimization for preference alignment.*
- Cui et al. (2023). *UltraFeedback for LLM training.*