Low-rank self-play fine-tuning for small LLMs

Мун Павел

Научный руководитель: Грабовой А. В.
Научный консультант: Охотников Н. В.

Московский физико-технический институт

25 марта 2025 г.

Введение

Этапы обучения

- Предварительное обучение
- SFT - supervised fine-tuning
- RLHF - reinforcement learning from human feedback

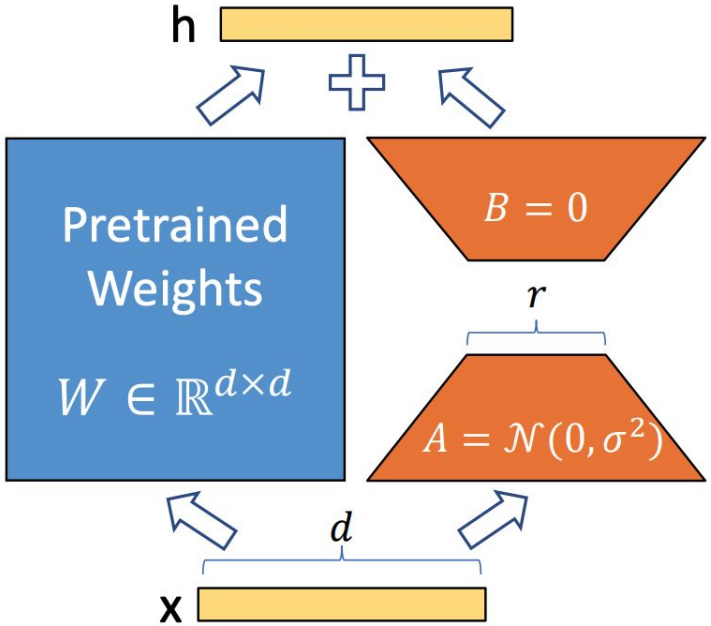
Постановка задачи

Проблема: сложность дообучения маленьких LLM в условиях ограниченных ресурсов

Предлагаемое решение: метод основанный на LoRA [1] и SPIN [2]

цель исследования: оценить оправданность метода в рамках ограниченных ресурсов для маленьких LLM.

Что такое LoRA



Что такое SPIN

- Цель противника: сгенерировать ответы, неотличимые от настоящих
- Цель игрока: уметь различать сгенерированные ответы от настоящих

Существующие решения

- Видеопамять: QLoRA [3]
- Человеческое участие: DPO [4]

Ожидаемый результат

- ответ на вопрос оправданности метода в условиях ограниченных ресурсов для маленьких LLM
- Сравнение с другими методами по метрикам качества, времени обучения, видеопамяти

Текущие результаты

- Верхняя оценка на ограничение количества параметров
- Прочитанные статьи: LoRA, SPIN
- Частично написанная статья
- Настроенный пайплайн обучения и обученная модель

План работы

- Новые датасеты
  - Winogrande
  - MMLU
  - Hellaswag
- Обучить больше архитектур
- Возможные улучшения метода
  - Дистиляция
  - Меньшие размеры float
  - И поиск иных подходов улучшения

Ссылки

- [1] *Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen*. LoRA: Low-Rank Adaptation of Large Language Models. https://openreview.net/forum?id=nZeVKeeFYf9
- [2] *Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, Quanquan Gu.* Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. https://openreview.net/forum?id=O4cHTxW9BS
- [3] *Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer*. QLoRA: Efficient Finetuning of Quantized LLMs. https://papers.nips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html
- [4] *Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, Chelsea Finn*. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html