

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

Павел Юрьевич Мун

Московский физико-технический институт

Курс: Автоматизация научных исследований

Эксперт: А. В. Грабовой

Консультант: Н. В. Охотников

2025

Задача и цель исследования

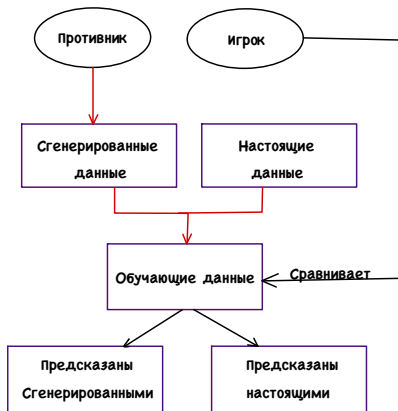
Задача

Дообучение небольших языковых моделей в условиях ограниченных ресурсов

Цель

Исследовать оправданность применимости предложенного метода

Стратегия self-play



1. Zixiang Chen и др. . «Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models». Openreview, 2024 URL: <https://openreview.net/forum?id=O4cHTxW9BS..>
2. Edward J Hu и др. LoRA: Low-Rank Adaptation of Large Language Models Openreview, 2022 URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
3. Rafael Rafailov и др. Direct Preference Optimization: Your Language Model is Secretly a Reward Model// NeurIPS, 2023

Дообучение на SFT

Множество Θ - пространство всевозможных параметров, p_{θ} - модель, а $\theta \in \Theta$ - ее параметры. $\mathbf{x} = [x_1, \dots, x_n] \sim q(\cdot)$, $\mathbf{y} = [y_1, \dots, y_m] \sim p_{data}(\cdot|\mathbf{x})$ - последовательности символов, интерпретируются как промпт и истинный ответ

В задаче ставится ограничение на количество параметров модели, то есть $\|\Theta\| \leq K$

Модель в виде условной плотности:

$$p_{\theta}(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^m p_{\theta}(y_j|\mathbf{x}, y_{<j})$$

Задача минимизации функционала:

$$L_{SFT}(\theta) = -\mathbb{E}_{\mathbf{x} \sim q(\cdot), \mathbf{y} \sim p_{data}(\cdot|\mathbf{x})} [\log p_{\theta}(\mathbf{y}|\mathbf{x})] \quad (1)$$

Двух шаговый SPIN

Обозначим за $\Omega \subset \Theta$ - подпространство весов адаптеров LoRA

Рассмотрим итерацию $t+1$, противником является p_{θ_t} , которая по промптам \mathbf{x} генерирует ответы \mathbf{y}' , а игроком $p_{\theta_{t+1}}$.

Обучение игрока:

$$f_{t+1} = \operatorname{argmin}_{f \in \mathcal{F}_t} \mathbb{E}[l(f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}'))]$$

$$\mathbf{x} \sim q(\cdot)$$

$$\mathbf{y} \sim p_{data}(\cdot | \mathbf{x})$$

$$l(t) = \log(1 + e^{-t})$$

Обновление противника:

$$p_{\theta_{t+1}} = \operatorname{argmax}_p \mathbb{E}[f_{t+1}(\mathbf{x}, \mathbf{y})] -$$

$$\lambda \mathbb{E}_{\mathbf{x} \sim q(\cdot)} \text{KL}(p(\cdot | \mathbf{x}) || p_{\theta_t}(\cdot | \mathbf{x}))$$

$$\mathbf{x} \sim q(\cdot)$$

$$\mathbf{y} \sim p_{data}(\cdot | \mathbf{x})$$

$$l(t) = \log(1 + e^{-t})$$

Одношаговый SPIN

Одношаговый алгоритм:

$$L_{SPIN} = \mathbb{E}_{\mathbf{x} \sim q(\cdot), \mathbf{y} \sim p_{data}(\cdot|\mathbf{x})} \left[\ell \left(\lambda \log \frac{p_{\theta}(\mathbf{y}|\mathbf{x})}{p_{\theta_t}(\mathbf{y}|\mathbf{x})} - \lambda \log \frac{p_{\theta}(\mathbf{y}'|\mathbf{x})}{p_{\theta_t}(\mathbf{y}'|\mathbf{x})} \right) \right]$$

$$\Delta \theta_{t+1} = \operatorname{argmin}_{\Delta \theta \in \Omega} L_{SPIN}(\theta_0 + \Delta \theta, \theta_0 + \Delta \theta_t)$$

Вычислительный эксперимент

Гипотеза

Значения лосса у модели, обученной предложенным методом будет ниже, чем у модели на этапе SFT

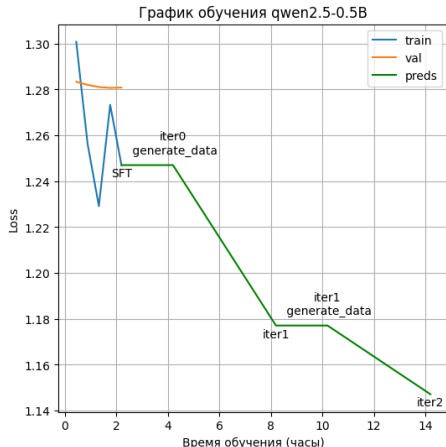
Цель

Обучение моделей предложенным методом и просто с адаптерами LoRA и сравнение по метрикам, лоссу и времени обучения

Данные

Рассматривается датасет ultrachat_200k, модели обучаются на 1% данных, примерно 2000 объектов

Сравнение итераций SPIN



- ▶ SFT - момент окончание этапа SFT
- ▶ $\text{iter}(k)\text{generate_data}$ - окончание генерации данных противником для $k + 1$ итерации SPIN
- ▶ $\text{iter}(k)$ - окончание k -ой итерации SPIN

- ▶ Предложен метод дообучения в условиях ограниченных ресурсов
- ▶ Обоснована применимость адаптеров LoRA к методу SPIN
- ▶ Получены некоторые практические результаты