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

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

Задача

Дообучение небольших языковых моделей в условиях ограниченных ресурсов

Предлагается

метод, основанный на внедрении адаптеров LoRA и стратегии self-play

Цель

Исследовать оправданность применимости предложенного метода

Стратегия self-play

Цель противника: сгенерировать данные, неотличимые от настоящих

Цель игрока: уметь различать сгенерированные данные от настоящих

Задача дообучения на этапе SFT

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

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

Модель в виде условной плотности:

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

Задача минимизации функционала:

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

Двухшаговый SPIN

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

Обучение игрока:

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

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

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

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

Обновление противника:

$$p_{\boldsymbol{\theta}_{t+1}} = \underset{p}{\operatorname{argmax}} \mathbb{E}[f_{t+1}(\mathbf{x}, \mathbf{y})] - \lambda \mathbb{E}_{\mathbf{x} \sim q(\cdot)} \mathrm{KL}\left(p(\cdot|\mathbf{x}) || p_{\boldsymbol{\theta}_t}(\cdot|\mathbf{x})\right)$$

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

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

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

Одношаговый SPIN

Обозначим за $\boldsymbol{\Omega} \subset \boldsymbol{\Theta}$ - подпространство весов адаптеров LoRA, малоранговых разложений обучаемых матриц

Предложенный метод:

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

$$\boldsymbol{\Delta\theta}_{t+1} = \underset{\boldsymbol{\Delta\theta}\in\boldsymbol{\Omega}}{\operatorname{argmin}} L_{SPIN}(\boldsymbol{\theta}_0 + \boldsymbol{\Delta\theta}, \boldsymbol{\theta}_0 + \boldsymbol{\Delta\theta}_t)$$

Вычислительный эксперимент

Гипотеза

Значения метрики у модели, обученной предложенным методом будет выше, чем у модели на этапе SFT и не будет сильно отставать от модели обученной методом SPIN

Цель

Обучение моделей и сравнение их по метрикам, видеопамяти и времени обучения

Данные

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

Сравнение итераций SPIN

model	trainable params	BLEU (SFT)	BLEU (LoRA + SPIN)
qwen2.5 (lora_r = 8)	1M	0.06454 (93.4%)	0.06554 (93.2%)
qwen2.5 (lora_r = 16)	2.2M	0.06420 (92.9%)	0.06895 (98.0%)
qwen2.5 (without lora)	494M	0.06912 (100%)	0.07035 (100%)

Обучение моделей Qwen2.5-0.5B-Instruct. lora_r - параметр промежуточной размерности адаптеров. Модели с адаптерами обучались на T4 (16GB), а модель без параметров на A100 (40GB)

1. Снижение затрачиваемой видеопамяти в 2 раза
2. Разница метрики модели без адаптеров с моделью с адаптерами 2
3. Время обучения моделей с адаптерами 8-9 часов, а обучение модели без адаптеров на A100 2.5 часа

Выносится на защиту

1. Предложен метод дообучения в услових ограниченных ресурсов
2. Исследована оправданность применимости метода. Метод смог уменьшить используемую видеопамять в 2 раза при снижении значения метрики на 2%

Литература

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