УДК 004.912

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

***П.Ю. Мун[1], Н.В. Охотников[1], А.В. Грабовой[1]***

[1]Московский физико-технический институт (национальный исследовательский университет)

Большие языковые модели (LLM) демонстрируют исключительные возможности в широком спектре областей, которые могут требовать специализированных знаний, поэтому задача дообучения или выравнивания модели является важной. Обычно процесс обучения LLM состоит из нескольких этапов: предварительного обучения, и этапов дообучения: SFT, обучение на предварительно размеченном наборе данных, и RLHF, во время которого необходим эксперт-человек, оценивающий ответы модели. Предварительное обучение требует огромных вычислительных ресурсов, поэтому часто используют публичные предобученные модели (Llama3, Qwen2.5) и настраивают под целевую задачу. Но также проблема ограниченности ресурсов встречается и на этапах дообучения SFT и RLHF. Ограниченность ресурсов проявляется в недостатке видеопамяти для хранения и обновления параметров модели, необходимости в размеченных данных для повышения качества и времени обучения.

В работе исследуются методы повышения эффективности обучения моделей в условиях ограниченных ресурсов. В частности предлагается метод дообучения LLM, который значительно снижает потребление видеопамяти, а также убирает необходимость в прямом человеческом участии, как на этапе RLHF.

Целью работы является исследование оправданности применения предложенного метода к маленьким LLM в условиях ограниченных ресурсов.

Рассмотрим постановку задачи. Дан обучающий размеченный датасет $S = \{x_i, y_i\}_{i\in\{1..N\}}$, где $N$ - размер датасета, $x_i$ и $y_i$ - символьные последовательности. Обозначим за $\Theta$ - пространство всевозможных параметров трансформерной модели, $p_\theta$ - модель , а $\theta \in \Theta$ - ее параметры.

В задаче рассматривается обучение небольших языковых моделей, поэтому ставится ограничение на количество параметров модели, то есть $\|\Theta\| \leqslant K$, где K - константа.

Предложенный метод основан на двух идеях:

- Во-первых, внедрение адаптеров LoRA [1] в слои трансформерной модели. Метод предполагает сравнительно маленькую внутреннюю размерность пространства параметров и снижает ее с помощью встраивания адаптеров, малоранговых разложений матриц. Таким образом, вместо рассмотрения всего пространства $\Theta$, рассматривается лишь некоторое подпространство $\Omega$, $dim\Omega \ll dim\Theta$
- Во-вторых, механизме self-play для обучения адаптеров. Механизм состоит из последовательных игр текущей версии модели со своей предыдущей версией. Предыдущая версия генерирует ответы $y'$ по промптам $x$ датасета $S$ на этапе SFT, а модель пытается различить настоящий ответ $y$ от сгенерированного $y'$. Общий метод исследуется в статье [2], а в данной работе он применяется исключительно к адаптерам LoRA.

Таким образом, подбор модели осуществляется с помощью решения следующей задачи минимизации:

$$\Delta\theta_{t+1} = \underset{\Delta\theta\in\Omega}{\operatorname{argmin}}\, E\left[\ell\left(\lambda \log \frac{p_{\theta_0+\Delta\theta}(y\,|\,x)}{p_{\theta_0+\Delta\theta_t}(y\,|\,x)} - \lambda \log \frac{p_{\theta_0+\Delta\theta}(y'\,|\,x)}{p_{\theta_0+\Delta\theta_t}(y'\,|\,x)}\right)\right]$$

где t - итерация метода, $l = log(1 + exp(-t))$, $\lambda$ - коэффициент регуляризация,

**Литература**


1. *Edward J. Hu [et al.]. «LoRA: Low-Rank Adaptation of Large Language Models». B: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022. url:* https://openreview.net/forum?id=nZeVKeeFYf9
2. *Zixiang Chen [et al.]. «Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models». B: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024. url: https://openreview.net/forum?id=O4cHTxW9BS.158*