

Генерация узкоспециализированных текстов с помощью дообученных больших языковых моделей в условиях ограниченных ресурсов

А. П. Степанов¹, Ю. А. Шичкина

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)

¹stepetal94@yandex.ru

Аннотация. Автоматическая генерация узкоспециализированных текстов до сих пор остается актуальной задачей. Несмотря на то, что большие языковые модели показывают хорошие общие результаты генерации, ряд текстов, например, медицинской тематики, не могут быть созданы нейросетью без дообучения. В общем случае этот процесс является ресурсозатратной операцией. В докладе рассмотрена методика, с помощью которой можно уменьшить количество потребляемой памяти графического процессора и выполнить тонкую настройку большой языковой модели, используя общедоступные бесплатные сервисы. В качестве набора для обучения выбраны медицинские заключения МРТ.

Ключевые слова: большая языковая модель; генерация узкоспециализированных текстов; эффективная по параметрам тонкая настройка

I. ВВЕДЕНИЕ

За последние несколько лет качество генерации текстов при помощи больших языковых моделей значительно улучшилось. Это связано с усовершенствованием их архитектуры, увеличением количества доступных датасетов для обучения, возрастанием мощности аппаратного обеспечения. Большие языковые модели стали способны решать многие специфические задачи без дополнительной тренировки. Тем не менее, спектр их возможностей по генерации ограничен обучающим набором данных. Если исходное качество созданного текста неудовлетворительно, то можно попробовать дообучить большую языковую модель. Среди основных видов дообучения выделяют [1]:

- Тонкую настройку при помощи инструкций (Instruction fine-tuning). В данном случае базовую языковую модель обучают на основе пар запрос-ответ. Чем больше объем и разнообразие инструкций, тем больше вероятность того, что языковая модель запомнит требуемый паттерн [2]. В качестве больших языковых моделей, обученных таким образом, можно отметить Stanford Alpaca, которая была получена на основе LLaMA 7b [3], а также представленную авторами статьи [4] языковую модель Med-PaLM 2. Главное достоинство подхода – возможность реализации полноценных диалоговых систем. Основной недостаток как в первом, так и во втором примерах – необходимость наличия исчерпывающего набора данных для обучения, что влечет за собой значительное увеличение

количества вычислительных ресурсов для его хранения и обработки.

- Полное дообучение (Full fine-tuning). Данный подход предполагает изменение всех весов модели. Это приводит к созданию ее новой версии и, как следствие, к забыванию (catastrophic forgetting) той информации, которая была получена на этапе изначальной тренировки. Этот подход требует не только значительных вычислительных ресурсов, но и сложной конфигурации среды, в которой будет производиться обучение.
- Эффективную по параметрам тонкую настройку языковой модели (Parameter-efficient fine-tuning, PEFT). Подобный метод позволяет значительно сократить используемое количество памяти за счет того, что часть весов модели не участвует в тренировке (frozen weights). Существуют разные подходы PEFT, но наиболее распространенным является метод на основе адаптеров LoRA (Low-Rank Adaptation) [5, 6] и его оптимизированная версия QLoRA (Quantized LoRA) [7], в которой размер каждого параметра сжимается в несколько раз. Стоит отметить, что с одной стороны, как показывают авторы работы [8], подобные подходы уступают полному дообучению, однако с другой – позволяют значительно сэкономить память графического процессора, время тренировки и бюджет проекта, что во многих случаях является критическим. Так, из работ [9 - 12] можно сделать вывод о том, что использование этого подхода позволяет получить результаты, близкие к SOTA решениям в разных областях науки и техники.

Сравнительная схема видов дообучения представлена на рис. 1.

Рис. 1. Схема видов дообучения

Стоит отметить, что, наборы данных, на которых дообучались большие языковые модели с открытым исходным кодом, такие, как, например, Llama [13], являются в основном англоязычными. Это приводит к тому, что изначальное качество генерации текстов на русском языке не удовлетворяет заявленным требованиям.

В докладе рассматривается методика эффективного по параметрам дообучения большой языковой модели и приводится пример результата обучения нейросети по генерации текста медицинского заключения МРТ на русском языке. В качестве среды для тренировки нейросети используются платформы Kaggle и Google Colab.

II. МАТЕРИАЛЫ И МЕТОДЫ

Методика дообучения сводится к представленным далее этапам.

A. Выбор модели обучения

Для правильного подбора базовой языковой модели для донастройки нужно оценить ресурсы. Типичное значение размера графической памяти, предоставляемое платформами для обучения без оплаты, составляет порядка 16 Гб. Этого хватает, чтобы выполнить тонкую донастройку большой языковой модели с 8 миллиардами параметров. Качество генерации при этом во многом зависит от содержательности обучающей выборки.

B. Подготовка набора данных

Этот этап является одним из самых трудоемких, т. к. именно на нем необходимо очистить «сырые» данные, а также привести их в формат, подходящий для обучения нейросети. Поскольку дообучается инструктивная языковая модель, обучающая выборка должна включать в себя json – файл, состоящий из массива элементов с полями input, instruction, output. Input – это строка с ожидаемым вводом, output – строка с тем, что планируется вывести, instruction – непосредственная команда для языковой модели. Например, «Сгенерируй текст согласно входным данным».

C. Выбор среды тренировки нейросети

В качестве сервисов можно использовать Google Colab или Kaggle. На бесплатной основе обе предлагают ~16Гб памяти графического процессора. Однако было замечено, что нагрузка на Google Colab выше, поэтому время использования GPU может варьироваться вплоть до блокировки на какой-то интервал. Kaggle предлагает фиксированное количество часов в неделю. Для загрузки обучающего набора данных в Google Colab можно использовать Google Disk, на Kaggle же придется импортировать требуемый набор данных.

D. Проведение дообучения и анализ результатов

Для дообучения с помощью метода PEFT применяют либо библиотеку transformers, либо более высокоуровневые библиотеки, основанные на ней. Одним из примеров последней является Unsloth, которая и использовалась при тестировании. Преимуществом является удобный программный интерфейс и поддержка сообщества. При использовании unsloth сначала необходимо загрузить модель в выбранном режиме квантования (например, 4 бита), следом загрузить набор данных и применить к нему шаблон диалога требуемой

модели, далее получить LoRA-адаптеры модели, потом создать объект тренера модели и выполнить дообучение. Результатом будут дообученная модель в качестве LoRA-адаптеров.

E. Сохранение весов дообученной модели

LoRA-адаптеры можно сохранить с целью дальнейшей интеграции в базовую модель, например, при помощи фреймворка ollama. Другой опцией сохранения является слияние дообученных весов в базовую модель и сохранение всей модели. В данном случае целесообразно применить экспорт в формат GGUF с квантованием для уменьшения объема. При сохранении можно сразу отправить результат в репозиторий на huggingface spaces.

III. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Проверка качества генерации дообученной языковой модели осуществлялась на примере создания описательной части медицинского заключения МРТ по краткой характеристике патологии.

В качестве базовой модели была выбрана Vikhr-Llama3.1-8B-Instruct-R-21-09-24. Данная модель натренирована на наборе данных Grandmaster-PRO-MAX и адаптирована под использование инструкций на русском языке [13]. Загрузка исходных файлов модели осуществлялась из репозитория проекта на Hugging Face Spaces.

Обучающий набор состоял из 936 документов в формате .doc. Процесс предобработки включал в себя выделение требуемых частей, обезличивание и формирование единого json – файла с инструкциями для языковой модели. Пример инструкции (для экономии места часть столбца «Output» опущена) приведен в табл. I.

ТАБЛИЦА I. ИНСТРУКЦИЯ ДЛЯ ОБУЧЕНИЯ ЯЗЫКОВОЙ МОДЕЛИ

Input	Instruction	Output
МР-признаки дисциркуляторной энцефалопатии.	Сгенерируй текст согласно входным данным	На серии МР-томограмм, взвешенных по T1 и T2 в аксиальной, сагиттальной и коронарной плоскостях, визуализированы суб- и супратенториальные структуры...

Дообучение языковой модели производилось в Google Colab с использованием библиотеки Unsloth и техники QLoRA. Было выбрано 4-х битное квантование параметров для экономии места. После обучения модель была экспортирована в формат GGUF и сохранена в репозиторий Hugging Face Spaces. Параметры обучения приведены в табл. II, а график функции потерь на рис. 2.

ТАБЛИЦА II. ПАРАМЕТРЫ ОБУЧЕНИЯ ЯЗЫКОВОЙ МОДЕЛИ

Количество примеров	Количество эпох	Количество шагов	Количество тренируемых параметров/общее количество параметров
936	3	351	167772160/8000000000

Таким образом, были рассмотрены подходы тонкой настройки больших языковых моделей. Приведена методика тренировки и представлены результаты эксперимента по генерации описательной части медицинского заключения МРТ. Было показано, что сгенерированный текст близок к ожидаемому. Направление будущих исследований включает в себя анализ качества при выборе другой базовой модели и оптимизацию процесса обучения. В заключении хочется отметить, что подход PEFT, несмотря на свое недавнее появление, уже показал отличные результаты на многих специализированных задачах, которые раньше требовали полного дообучения языковой модели. Учитывая потребность в экономии ресурсов, этот и альтернативные подходы тонкой настройки будут активно развиваться в будущем.

СПИСОК ЛИТЕРАТУРЫ

- [1] Parthasarathy, V.B., Zafar, A., Khan, A. and Shahid, A. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities [Electronic resource] // ArXiv. 2024 URL: <https://arxiv.org/abs/2408.13296> (accessed: 02.03.2025).
- [2] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Le, Q. V. Finetuned language models are zero-shot learners [Electronic resource] // ArXiv. 2021. URL: <https://arxiv.org/abs/2109.01652> (accessed 05.03.2025)
- [3] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T.B. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models [Electronic resource] URL: <https://crfm.stanford.edu/2023/03/13/alpaca.html> (accessed 07.03.2025)
- [4] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H. and Neal, D. Toward expert-level medical question answering with large language models // Nature Medicine. 2025. pp.1-8.
- [5] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. Lora: Low-rank adaptation of large language models // ICLR. 2025. Vol. 1. No. 2. P. 3.
- [6] Chunlei Xin, Yaojie Lu, Hongyu Lin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, and Le Sun. Beyond Full Fine-tuning: Harnessing the Power of LoRA for Multi-Task Instruction Tuning // In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2025. Pp. 2307–2317.
- [7] Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms // Advances in neural information processing systems. 2023. Vol. 36, Pp. 10088–10115.
- [8] Reece Shuttleworth, J.A., Torralba, A. and Sharma, P. Lora vs full fine-tuning: An illusion of equivalence [Electronic resource] // ArXiv. 2024. URL: <https://arxiv.org/abs/2410.21228> (accessed: 09.03.2025).
- [9] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W. and Yi, J. Parameter-efficient fine-tuning of large-scale pre-trained language models // Nature Machine Intelligence. 2023. Vol. 5. No. 3. Pp. 220–235.
- [10] Qiu J. et al., Large AI Models in Health Informatics: Applications, Challenges, and the Future // IEEE Journal of Biomedical and Health Informatics. 2023. Vol. 27. No. 12, Pp. 6074–6087.
- [11] Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M. and Zhou, B.. Enhancing chat language models by scaling high-quality instructional conversations [Electronic resource] // ArXiv. 2024. URL: <https://arxiv.org/abs/2305.14233> (accessed: 10.04.2025).
- [12] Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D. and Liu, X. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities // IEEE Communications Surveys & Tutorials. 2024.
- [13] Nikolich, Aleksandr, et al. Vikhr: Constructing a State-of-the-art Bilingual Open-Source Instruction-Following Large Language Model for Russia [Electronic resource] // ArXiv. 2024. URL: <https://arxiv.org/abs/2405.13929> (accessed: 14.04.2025).
- [14] Lin C. ROUGE: a package for automatic evaluation of summaries // Text Summarization Branches Out / Association for Computational Linguistics. Barcelona. 2024. Pp. 74–81.

Рис. 2. Функция потерь

Из табл. II можно сделать вывод о том, что донастройке подверглось только 2% от общего количества параметров.

Согласно рис. 2, после второй эпохи функция потерь изменяется незначительно => обучение можно ограничить только двумя эпохами.

Статистика дообучения представлена в табл. III.

ТАБЛИЦА III. СТАТИСТИКА ДООБУЧЕНИЯ ЯЗЫКОВОЙ МОДЕЛИ

Время обучения, мин	Максимальное использование GPU, Гб	Максимальное использование GPU, Гб
79.08	8.84	1.703

После проведения дообучения на запрос «МР-картина дисциркуляторной энцефалопатии» языковая модель отвечает: «На серии МР-томограмм, взвешенных по T1 и T2 в трёх плоскостях, визуализируются суб- и супратенториальные структуры. В белом веществе головного мозга отмечаются множественные участки изменения интенсивности сигнала, округлой формы, с четкими, ровными контурами, характеризующиеся гиперинтенсивным МР-сигналом на T2-ВИ и изогиперинтенсивным на T1-ВИ, диаметром до 3 мм, сосудистого характера. Также имеются зоны глиоза в области лобных долей, размерами до 5x4 мм (слева), и до 6x7 мм (справа). Боковые желудочки мозга симметричные, не расширены. III желудочек шириной до 3 мм. IV желудочек и базальные цистерны не изменены. Субарахноидальное пространство выражено неравномерно, умеренно расширено по конвекситальной поверхности лобных долей. Срединные структуры не смещены. Хиазмально-селлярная область без особенностей. Миндалины мозжечка расположены на уровне большого затылочного отверстия. Отмечается умеренное утолщение слизистой оболочки ячеек решетчатой кости и верхнечелюстных пазух.»

Оценка качества генерации осуществлена при помощи набора метрик ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [14], а именно ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum. Значения параметров приведены в табл. IV.

ТАБЛИЦА IV. ЗНАЧЕНИЯ МЕТРИК ROUGE

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0.699	0.574	0.626	0.626

Из таблицы можно сделать вывод о том, что дообученная языковая модель генерирует близкий, но не идентичный текст. Важно отметить, что при этом сохранена стилистика текста из обучающего набора.