УДК 004.85

Ибрагимхалилов Руслан Теймурович,
ст. преподаватель каф. ИСТ
Поволжский государственный университет
телекоммуникаций и информатики
Ibragimkhalilov Ruslan Teymurovich,
Senior Lecturer, Department of Information Technologies,
Volga Region State University of
Telecommunications and Informatics

Воробьев Илья Романович, магистрант,
Поволжский государственный университет
елекоммуникаций и информатики
Vorobyov Ilya Romanovich , Master's Student,
Volga Region State University of
Telecommunications and Informatics

Морозов Дмитрий Денисович, магистрант,
Поволжский государственный университет
телекоммуникаций и информатики
Morozov Dmitry Denisovich , Master's Student,
Volga Region State University of
Telecommunications and Informatics

ТОНКАЯ НАСТРОЙКА БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: СОВРЕМЕННОЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ
FINE-TUNING OF LARGE LANGUAGE MODELS: CURRENT STATE AND DEVELOPMENT PROSPECTS

Аннотация. В статье рассматривается метод тонкой настройки (fine-tuning) как один из ключевых подходов в современной обработке естественного языка. Анализируется эволюция техники: от полного дообучения всех параметров модели до параметро-эффективных методов (PEFT). Особое внимание уделяется применению тонкой настройки для адаптации универсальных языковых моделей под узкоспециализированные задачи и предметные области. Сделан вывод о том, что развитие методов fine-tuning является магистральным направлением демократизации технологий искусственного интеллекта.

Abstract. The article discusses the fine-tuning method as one of the key approaches in modern natural language processing. The evolution of the technique is analyzed: from full fine-tuning of all model parameters to parameter-efficient methods (PEFT). Special attention is paid to the application of fine-tuning for adapting universal language models to highly specialized tasks and subject areas. It is concluded that the development of fine-tuning methods is the main direction for democratizing artificial intelligence technologies.

Ключевые слова: Тонкая настройка, большие языковые модели, обработка естественного языка, трансферное обучение, PEFT, LoRA, адаптация моделей.

Keywords: Fine-tuning, large language models, natural language processing, transfer learning, PEFT, LoRA, model adaptation.

Современный этап развития обработки естественного языка (NLP) характеризуется доминированием больших языковых моделей (LLM). Такие архитектуры, как GPT, BERT, LLaMA

и их аналоги, обучаются на колоссальных объемах текстовых данных, приобретая фундаментальные знания о языке, грамматике, фактологии и контексте [1, 2]. Однако универсальность этих моделей является одновременно и их ограничением: базовая модель способна продолжать текст или отвечать на общие вопросы, но неэффективна при решении строго определенных прикладных задач, таких как анализ тональности в отзывах клиентов, юридическое консультирование или медицинская диагностика. Решением этой проблемы является метод, известный как тонкая настройка. Данная работа ставит целью анализ современного состояния и перспективных направлений развития методов тонкой настройки языковых моделей.

Тонкая настройка (fine-tuning) представляет собой процесс адаптации предварительно обученной модели к конкретной целевой задаче. Суть метода можно описать метафорой: базовая модель – это выпускник университета, обладающий фундаментальными знаниями (она знает русский язык, умеет строить предложения и рассуждать). Тонкая настройка – это стажировка на конкретном рабочем месте, где выпускника доучивают выполнять специфические функции: писать судебные иски, ставить медицинские диагнозы по описанию симптомов или классифицировать новости по темам. В техническом плане это означает продолжение обучения модели, но уже на размеченном датасете, относящемся к предметной области, с корректировкой весов нейронной сети под конкретную downstream-задачу [3].

Долгое время классическим подходом был полный fine-tuning (Full Fine-Tuning), при котором обновляются все параметры модели. Для моделей с сотнями миллионов или миллиардами параметров это требует огромных вычислительных ресурсов и памяти. Например, полное дообучение модели BERT-base (110 млн параметров) все еще выполнимо, но аналогичная операция для GPT-3 (175 млрд параметров) становится непозволительной роскошью для большинства исследователей и компаний [4].

Ответом на это ограничение стало бурное развитие параметро-эффективных методов тонкой настройки (Parameter-Efficient Fine-Tuning, PEFT). Эти методы позволяют адаптировать гигантские модели, замораживая их исходные веса и добавляя лишь небольшое количество обучаемых параметров. Ключевая идея PEFT заключается в том, что для решения новой задачи не нужно изменять все "знания" модели, достаточно научить ее переключать внимание или немного корректировать выходные сигналы.

Наиболее популярным на сегодняшний день методом PEFT является LoRA (Low-Rank Adaptation). Вместо прямого обновления весовой матрицы модели, LoRA внедряет в архитектуру параллельные низкоранговые матрицы, которые и обучаются в процессе тонкой настройки [5]. Это позволяет сократить количество обучаемых параметров в десятки тысяч раз без потери качества. Другим значимым подходом является использование адаптеров (Adapters) – небольших обучаемых модулей, встраиваемых между слоями трансформера [6].

Актуальным трендом становится применение тонкой настройки для создания специализированных моделей в узких доменах. Медицина требует понимания сложной терминологии и логики постановки диагнозов, юриспруденция – оперирования статьями законов и прецедентами, финансы – анализа рисков и трендов. Базовые модели часто "зашумлены" общей информацией и недостаточно точны в этих областях. Тонкая настройка на корпусах медицинской литературы (как в проекте BioBERT [7]) или юридических документов позволяет модели достичь экспертного уровня понимания предметной области.

Еще одним перспективным направлением является тонкая настройка с использованием инструкций (Instruction Fine-Tuning). В этом случае модель учат не просто решать задачу, а следовать инструкциям на естественном языке. Это приводит к появлению моделей-помощников, способных отвечать на вопросы, писать код или резюмировать тексты в диалоговом режиме. Примером служат семейства моделей Flan-T5 или InstructGPT, которые являются результатом тонкой настройки базовых моделей на тысячах пар "инструкция-ответ" [8].

Развитие методов тонкой настройки неразрывно связано с решением проблемы катастрофического забывания (catastrophic forgetting), когда модель, дообучаясь на новых данных, утрачивает часть знаний, полученных на этапе предварительного обучения. Исследования в области регуляризации и методов многозадачного обучения (multi-task fine-tuning) направлены на сохранение баланса между приобретением новых навыков и сохранением старых [9].

Также нельзя обойти стороной вопросы этики и безопасности. Тонкая настройка на некачественных или предвзятых данных может усилить социальные бисы в модели или сделать ее уязвимой для генерации вредоносного контента. Разработка методов "выравнивания" (alignment), таких как RLHF (Reinforcement Learning from Human Feedback), позволяет откалибровать поведение модели в соответствии с человеческими предпочтениями и безопасностью [10].

Проведенный анализ позволяет сделать вывод, что тонкая настройка превратилась из простого инструмента дообучения в самостоятельное и важнейшее направление развития NLP. Основные тенденции здесь - уход от ресурсозатратного полного fine-tuning к параметро-эффективным методам (LoRA, адаптеры), глубокая специализация моделей под конкретные отрасли и активное использование инструкций для создания универсальных ассистентов. Дальнейшая эволюция методов fine-tuning будет направлена на повышение адаптивности моделей в режиме реального времени, обеспечение сохранности первоначальных знаний и усиление контроля за этичностью принимаемых решений. Это путь от создания "универсальных солдат" к формированию гибкого инструментария, доступного для решения широкого круга практических задач.

Список литературы:

1. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. - 2019. - P. 4171-4186.
2. Language Models are Few-Shot Learners / T.B. Brown [et al.] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). - 2020. - P. 1877-1901.
3. Howard J. Fine-tuned Language Models for Text Classification / J. Howard, S. Ruder // arXiv preprint. - 2018. - arXiv:1801.06146.
4. Strubell E. Energy and Policy Considerations for Deep Learning in NLP / E. Strubell, A. Ganesh, A. McCallum // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. - 2019. - P. 3645-3650.
5. LoRA: Low-Rank Adaptation of Large Language Models / E. Hu [et al.] // International Conference on Learning Representations (ICLR). - 2022.
6. Houlsby N. Parameter-Efficient Transfer Learning for NLP / N. Houlsby [et al.] // Proceedings of the 36th International Conference on Machine Learning (ICML). - 2019. - P. 2790-2799.
7. Lee J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining / J. Lee [et al.] // Bioinformatics. - 2020. - Vol. 36, № 4. - P. 1234-1240.
8. Chung W. Scaling Instruction-Finetuned Language Models / W. Chung [et al.] // arXiv preprint. - 2022. - arXiv:2210.11416.
9. Kirkpatrick J. Overcoming catastrophic forgetting in neural networks / J. Kirkpatrick [et al.] // Proceedings of the National Academy of Sciences. - 2017. - Vol. 114, № 13. - P. 3521-3526.
10. Ouyang L. Training language models to follow instructions with human feedback / L. Ouyang [et al.] // Advances in Neural Information Processing Systems (NeurIPS). – 2022.