КЛАССИФИКАЦИЯ И ОБЗОР МЕТОДОВ ПАРАМЕТРО-ЭФФЕКТИВНОЙ АДАПТАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ (LLM)

Нестеренков С. Н., Фурсанов С. А., Камышев С. В.
Кафедра программного обеспечения информационных технологий,
центр информатизации и инновационных разработок,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {s.nesterenkov, s.fursanov, s.kamyshev}@bsuir.by

В последние годы большие языковые модели (LLM, Large Language Models) стали ключевым инструментом в задачах обработки естественного языка. Их применение в специализированных предметных областях сталкивается с проблемами высокой вычислительной сложности и ограниченной доступности данных для дообучения. В данной работе представлен обзор существующих методов параметро-эффективной адаптации LLM, проводится их классификация и сравнительный анализ. Основное внимание уделено подходам, позволяющим адаптировать модели без изменения всех весов, что делает их более практичными для узкоспециализированных задач.

Введение

Большие языковые модели, такие как GPT и LLaMA, демонстрируют высокую эффективность в широком спектре задач генерации текста и семантического анализа [1,2]. Тем не менее, применение LLM в узкоспециализированных областях (например, медицина [3], юриспруденция, инженерные тексты, образование) требует адаптации к специфике терминологии и формулировок [4].

Это связано с тем, что полное дообучение (full fine-tuning) не только требует колоссальных вычислительных ресурсов и времени, но и несет в себе риск «катастрофического забывания» – явления, при котором модель, адаптируясь к новой узкой задаче, теряет часть общих знаний, полученных на этапе предварительного обучения на разнородных данных. Это делает модель менее гибкой и универсальной. Полное дообучение модели требует модификации всех параметров, что при сотнях миллиардов весов практически невозможно.

Появление методов параметро-эффективной адаптации (PEFT, Parameter-Efficient Fine-Tuning) позволяет решить эту проблему, изменяя лишь малую часть параметров модели [5]. Данные методы достигают высокой эффективности за счет того, что не перезаписывают исходные знания модели, а добавляют к ним легковесные, задачно-специфичные модули или производят низкоранговые корректировки весов. Такой подход не только радикально снижает вычислительные затраты и объем требуемых данных для дообучения, но и позволяет быстро переключаться между разными задачами, используя одну базовую модель. В следующем разделе представлена классификация и подробный обзор наиболее распространенных методов PEFT.

I. Обзор методов параметро-эффективной адаптации

Существующие подходы к PEFT можно разделить на несколько основных категорий:

- Adapter-tuning. Метод Adapter-tuning предполагает встраивание в архитектуру модели дополнительных небольших слоев – адаптеров, которые обучаются на специализированных данных, в то время как базовые веса трансформера остаются замороженными [5,6]. Такой подход позволяет использовать одну и ту же модель для различных задач, снижает вычислительные затраты и объем необходимых данных. Типичная архитектура адаптера представляет собой компактную последовательность слоев, часто по схеме «bottleneck», которая включает полносвязный слой, сжимающий эмбеддинги до меньшей размерности (down-project), нелинейную функцию активации (например, GeLU) и слой, восстанавливающий исходную размерность (up-project). Такие модули обычно встраиваются после блока внимания (Multi-Head Attention) и после блока прямой связи (Feed-Forward Network) в каждом слое трансформера. Ключевым преимуществом является их модульность: для новой задачи можно обучить и затем хранить лишь небольшой файл с весами адаптера, «подключая» его к замороженной базовой модели по мере необходимости.
- LoRA (Low-Rank Adaptation). Алгоритм LoRA основан на разложении весовых матриц на низкоранговые компоненты, которые подлежат обучению [7]. Основные веса модели остаются неизменными. LoRA позволяет изменять до 0,1–1 % параметров модели, что делает возможным использование LLM с десятками миллиардов весов даже на ограниченных ресурсах.

– Prompt-tuning и Prefix-tuning. Эти методы не модифицируют архитектуру модели, а используют обучаемые векторы-подсказки (prompts), подаваемые на вход модели [8]. Prompt-tuning оптимизирует фиксированные подсказки, а Prefix-tuning – обучаемые последовательности токенов, которые расширяют вход модели. Подходы особенно эффективны для задач генерации текста и диалоговых систем. Prefix-tuning работает на уровне скрытых состояний (hidden states) модели. Он добавляет последовательность непрерывных (не дискретных) векторов («префикс») к активациям каждого слоя трансформера, причем эти векторы обучаются методом обратного распространения ошибки. В отличие от него, Prompt-tuning является более простым методом, который работает только на входном уровне модели, оптимизируя непрерывные векторы-подсказки, конкатенируемые с эмбеддингами входных токенов. Со временем появились и более продвинутые вариации, такие как P-Tuning, который использует LSTM или MLP для генерации более эффективных промптов.

II. Классификация методов

С точки зрения количества изменяемых параметров, PEFT-методы можно классифицировать следующим образом:

- Минимальные изменения (0,1–1 % параметров) – LoRA, Prefix-tuning.
- Средние изменения (1–3 % параметров) – Adapter-tuning, P-tuning.
- Полное дообучение (100 % параметров) – классический fine-tuning, используется редко из-за вычислительных ограничений.

По назначению методы можно разделить на:

- Универсальные адаптеры – применимы к различным задачам без модификации основной модели.
- Задачно-специфические адаптеры – обучаются под конкретный домен, обеспечивая максимальное качество.

III. Сравнительный анализ

Выбор метода PEFT зависит от компромисса между эффективностью, затратами и универсальностью. Prompt-tuning наиболее экономен (менее 0.1% параметров), но эффективен только в крупных моделях (от 20 млрд параметров). LoRA (0.1-1% параметров) не замедляет инференс и идеальна для ограниченных ресурсов, хотя в основном адаптирует только механизмы внимания. Adapter-tuning (1-3% параметров) демонстрирует наибольшую стабильность в сложных многоэтапных задачах, но добавляет небольшую вычислительную нагрузку при inference. Prefix-tuning показывает выдающиеся результаты в генеративных сценариях, однако требует тщательной настройки гиперпараметров.

Важным практическим аспектом является возможность комбинирования методов. Например, адаптеры можно комбинировать с обучением промптов, а низкоранговые матрицы LoRA – с префикс-тюнингом. Такие гибридные подходы открывают новые возможности для балансировки между параметрической эффективностью и качеством адаптации.

Таким образом, LoRA оптимальна для задач классификации при ограниченных ресурсах, Adapter-tuning – для сложных задач, требующих максимального качества, Prefix-tuning – для генерации, а Prompt-tuning – для сверхбольших моделей. Перспективным направлением является разработка адаптивных методов, автоматически подбирающих тип и масштаб адаптации под конкретную задачу.

Заключение

Параметро-эффективные методы адаптации больших языковых моделей открывают новые возможности для внедрения LLM в специализированные предметные области. Дальнейшие исследования должны быть направлены на:

- Разработку гибридных методов, сочетающих преимущества различных подходов;
- Формирование критериев выбора метода в зависимости от задач и ресурсов;
- Создание экспериментальных прототипов адаптированных моделей для разных доменов.

IV. Список литературы

1. Vaswani A., Shazeer N., Parmar N. и др. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998-6008.
2. Touvron H., Martin L., Stone K. и др. LLaMA: Open and Efficient Foundation Language Models // arXiv preprint arXiv:2302.13971. – 2023.
3. Скалозуб К. А., Нестеренков С. Н., Ярмош А. Д. Градиентные методы оптимизации в компьютерном зрении для медицинской диагностики // Технологии передачи и обработки информации: материалы Междунар. науч.-техн. семинара, Минск, апр. 2025 г. / редкол. : В. Ю. Цветков [и др.]. – Минск : БГУИР, 2025. – С. 34-36.
4. Lee J., Yoon W., Kim S. и др. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. – 2020. – Vol. 36, № 4. – P. 1234-1240.
5. Houlsby N., Giurgiu A., Jastrzebski S. и др. Parameter-Efficient Transfer Learning for NLP // Proceedings of ICML. – 2019.
6. PPfeiffer J., Kamath A., Rücklé A. и др. AdapterFusion: Non-Destructive Task Composition for Transfer Learning // Proceedings of EACL. – 2021.
7. Hu E., Shen Y., Wallis P. и др. LoRA: Low-Rank Adaptation of Large Language Models // arXiv preprint arXiv:2106.09685. – 2021.
8. Li X. L., Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation // Proceedings of ACL. – 2021.