

ССЫЛКИ:

1. <https://ailynx.ru/news/llm/peft-parameter-efficient-fine-tuning-methods-for-llms/>
2. <https://colab.research.google.com/drive/1B9RsKLMa8SwTxLsxRT8g9OedK10zfBEP?usp=sharing&authuser=1>
3. <https://huggingface.co/docs/peft/index>
4. <https://arxiv.org/abs/2303.15647>
5. <https://arxiv.org/abs/2601.09185>
6. <https://arxiv.org/abs/2601.06788>
7. <https://arxiv.org/abs/2106.09685>
8. <https://arxiv.org/abs/2303.10512> (<https://github.com/QingruZhang/AdaLoRA>)
9. <https://arxiv.org/abs/2305.14314>

Глава 2. Методы параметро-эффективного дообучения языковых моделей

2.1. Проблема дообучения больших языковых моделей в условиях ограниченных ресурсов

Современные большие языковые модели (Large Language Models, LLM), такие как GPT, LLaMA, BERT и их производные, содержат от сотен миллионов до сотен миллиардов параметров. Полное дообучение (full fine-tuning) таких моделей требует значительных вычислительных ресурсов: большого объёма видеопамяти (VRAM), длительного времени обучения и хранения дополнительных состояний оптимизатора.

В условиях ограниченных вычислительных ресурсов, например, при использовании одной GPU или при развертывании моделей в прикладных сервисах, полное дообучение становится практически невозможным. Основные ограничения включают:

- нехватку видеопамяти для хранения параметров и градиентов;
- высокую стоимость хранения и передачи моделей;
- увеличение времени обучения и инференса;
- сложности повторного воспроизведения экспериментов.

Эти ограничения особенно актуальны для прикладных задач обработки естественного языка, где требуется адаптация моделей под конкретные домены или инструкции при фиксированном аппаратном бюджете.

2.2. Подходы к адаптации моделей: от полного fine-tuning к PEFT

Исторически основным способом адаптации языковых моделей являлось полное дообучение всех параметров. Однако было показано, что для многих задач обновление всех весов избыточно: эффективная адаптация возможна при изменении лишь небольшой подпространственной компоненты параметров модели.

Это наблюдение привело к развитию **Parameter-Efficient Fine-Tuning (PEFT)** - класса методов, направленных на снижение числа обучаемых параметров без существенной потери качества. К ключевым преимуществам PEFT относятся:

- снижение требований к памяти и вычислениям;

- ускорение обучения;
- возможность хранения и распространения только адаптационных параметров;
- повышение воспроизводимости экспериментов.

К основным направлениям PEFT относятся:

- fine-tuning смещений (BitFit);
- адаптерные слои (Adapters, Compacter);
- методы на основе низкоранговой параметризации (LoRA и его расширения);
- методы регуляризации градиентов (GaLore);
- комбинации PEFT с квантизацией.

2.3. Parameter-Efficient Fine-Tuning языковых моделей

Полное дообучение (full fine-tuning) больших языковых моделей предполагает обновление всех параметров нейронной сети, что требует значительных вычислительных ресурсов и большого объёма размеченных данных. В условиях ограниченной доступности вычислительных мощностей и памяти GPU данный подход становится трудно применимым, особенно для моделей с десятками и сотнями миллиардов параметров.

В связи с этим получили развитие методы **parameter-efficient fine-tuning (PEFT)**, направленные на адаптацию предварительно обученных языковых моделей за счёт обучения лишь небольшой части параметров при сохранении основной части весов замороженной. Основная идея PEFT заключается в том, что для адаптации модели под новую задачу нет необходимости изменять все параметры, поскольку большая часть знаний уже закодирована в предобученных весах.

К методам PEFT относятся адаптерные слои (Adapter Layers), методы на основе подсказок (Prompt Tuning, Prefix Tuning), а также методы, основанные на ре-параметризации обновлений весов. Последняя группа методов получила наибольшее распространение благодаря благоприятному компромиссу между качеством, числом обучаемых параметров и вычислительными затратами <https://arxiv.org/abs/2303.15647>

2.4. Reparametrization-based методы дообучения

Reparametrization-based методы основаны на предположении, что обновления весов нейронной сети в процессе дообучения имеют **низкую внутреннюю размерность**. Вместо прямого обновления матриц весов высокой размерности данные методы предлагают параметризовать изменения весов через низкоранговые преобразования.

Одним из первых подходов данного класса является метод **Intrinsic SAID**, предложенный Aghajanyan et al. (2020). Авторы исследуют внутреннюю размерность пространства fine-tuning и показывают, что для достижения сопоставимого качества обновления весов могут быть выполнены в низкоразмерном подпространстве. Для этого используется Fastfood-преобразование, позволяющее эффективно расширять параметры из пространства малой размерности в пространство весов модели.

Несмотря на теоретическую значимость, Intrinsic SAID обладает рядом практических ограничений. В частности, метод требует хранения вспомогательных

структур размерности $O(D)$, где D — число параметров модели, а также обновляет все веса модели, что делает его малопригодным для дообучения современных крупных языковых моделей.

Данные ограничения послужили мотивацией для разработки более практических и масштабируемых reparametrization-based методов, таких как LoRA и его модификации <https://arxiv.org/abs/2303.15647>

2.5. Метод Low-Rank Adaptation (LoRA) и его расширения

Метод **Low-Rank Adaptation (LoRA)** (<https://arxiv.org/abs/2106.09685> 2021г), является одним из наиболее популярных и практически применимых методов PEFT. В рамках данного подхода обновление матрицы весов представляется в виде произведения двух матриц малого ранга. При этом предобученные веса модели остаются замороженными, а обучаемыми являются только низкоранговые матрицы.

LoRA позволяет существенно сократить число обучаемых параметров и требования к памяти, при этом демонстрируя качество, сопоставимое с полным дообучением. На практике LoRA чаще всего применяется к матрицам проекций в механизме внимания трансформера, однако исследования показывают, что распространение LoRA на большее число слоёв может дополнительно повысить качество.

В последующих работах были предложены расширения LoRA, направленные на повышение выразительности и эффективности метода. Так, метод **KronA** использует кронекерово произведение для параметризации обновлений весов, обеспечивая более выгодное соотношение между числом параметров и рангом обновлений. Однако данный подход был протестирован преимущественно на моделях малого размера.

Метод **DoRA** предлагает разделить обновление весов на компоненты направления и масштаба, что позволяет устраниТЬ ограничения стандартного LoRA, связанные с жёсткой связью между этими характеристиками. Эксперименты показывают, что DoRA превосходит LoRA, особенно в режимах с малым рангом адаптеров и ограниченным объёмом данных.

Методы **GLoRA** и **AdaLoRA** направлены на дальнейшее повышение эффективности LoRA за счёт добавления обучаемых параметров масштабирования и адаптивного выбора ранга соответственно. Несмотря на улучшения качества, данные методы увеличивают сложность обучения и могут приводить к дополнительным затратам памяти, что ограничивает их применение в условиях жёстких ресурсных ограничений. <https://arxiv.org/abs/2303.15647>

OrthoGeoLoRA - метод PEFT, основанный на геометрических принципах.

Вместо $\Delta W = BA^T$ он использует разложение, подобное SVD: $\Delta W = B\Sigma A^T$. При этом матрицы А и В ограничены пространством Штифеля (их столбцы ортонормированы), а Σ - обучаемая диагональная матрица сингулярных значений

Ключевая инновация: Ограничение ортогональности реализуется через геометрическую перепараметризацию. Оптимизатор (например, Adam) работает с обычными (неограниченными) параметрами в Евклидовом пространстве, которые на каждом шаге дифференцируемо отображаются (например, через QR-разложение) в ортонормированные матрицы А и В. Это позволяет использовать стандартные

оптимизаторы, не нарушая их внутренние состояния (например, моментум), и делает метод "drop-in" заменой для LoRA.

Метод протестирован на задаче иерархического поиска концепций по Европейскому тезаурусу социальных наук (ELSST) - реалистичном сценарии для веб-инфраструктур социальных наук.

Создан синтетический датасет, где языковая модель генерировала разнообразные неявные описания для каждого концепта ELSST. Качество данных проверено экспертами.

OrthoGeoLoRA превзошел стандартный LoRA и его продвинутые варианты (AdaLoRA, DoRA, LoHa, LoKr) по всем метрикам (MRR, Recall@k, NDCG@k) при одинаковом бюджете параметров (ранге r).

OrthoGeoLoRA предотвращает коллапс ранга (сингулярные значения остаются распределенными), обеспечивает более быструю сходимость и стабильность оптимизации, а также эффективнее использует увеличение ранга по сравнению с LoRA.

Метод сохраняет параметрическую и вычислительную эффективность LoRA, но обеспечивает лучшие структурированные адаптации моделей. Это позволяет ресурсоограниченным организациям (университетам, НКО) эффективно настраивать модели для общественно значимых задач, таких как улучшение поиска научных данных по социальным вопросам (неравенство, миграция, здравоохранение).

OrthoGeoLoRA предлагает принципиальное геометрическое исправление внутренней параметризации LoRA, что приводит к более стабильной оптимизации, предотвращению коллапса ранга и повышению качества поиска структурированных концепций без увеличения числа обучаемых параметров <https://arxiv.org/abs/2601.09185>

2.6. Информационно-теоретический анализ PEFT и искусственная запутанность

Традиционный анализ методов дообучения языковых моделей в основном опирается на эмпирические метрики качества и вычислительные характеристики. Однако в последние годы появились работы, предлагающие рассматривать процессы fine-tuning с точки зрения информационной теории и геометрии параметрического пространства нейронных сетей. В частности, в работе Artificial Entanglement in the Fine-Tuning of Large Language Models предложен новый подход к анализу методов параметро-эффективного дообучения, основанный на понятии искусственной запутанности.

Искусственная запутанность представляет собой меру корреляций между компонентами параметров нейронной сети, рассматриваемых как высокоразмерный тензор. Данный подход заимствует инструменты из квантовой теории информации, однако используется исключительно как математический аппарат для анализа сложности внутренних представлений модели. В рамках данного подхода параметры и их обновления могут быть представлены в виде матричных продуктовых состояний (Matrix Product States, MPS), что позволяет количественно оценивать степень коррелированности параметрических обновлений.

С точки зрения данного анализа, полное дообучение (full fine-tuning) и parameter-efficient fine-tuning формируют принципиально различные внутренние структуры обновлений весов. Полное дообучение характеризуется более сложными и высоко-

коррелированными обновлениями, тогда как методы PEFT, такие как LoRA, ограничивают пространство допустимых изменений за счёт низкоранговой параметризации. В частности, показано, что обновления весов, полученные с помощью LoRA, демонстрируют профиль искусственной запутанности, подчиняющийся объемному закону с характерной «долиной запутанности», отражающей структурные ограничения метода.

Важным наблюдением является чувствительность внутреннего профиля искусственной запутанности к гиперпараметрам LoRA, таким как ранг адаптеров и коэффициент масштабирования. При определённых настройках LoRA внутренняя корреляционная структура существенно отличается от структуры, формируемой при полном дообучении, что указывает на принципиальные различия в параметрических представлениях, формируемых этими методами.

Несмотря на указанные различия во внутренней структуре обновлений, внешнее поведение модели, в частности выходы механизма внимания трансформера, демонстрируют существенно более простые корреляционные свойства. Показано, что матрицы внимания и их выходные представления обладают низкой искусственной запутанностью, подчиняющейся так называемому плоскому закону с логарифмической поправкой. Более того, данные свойства оказываются устойчивыми к выбору метода дообучения и его гиперпараметров.

Данное явление авторы интерпретируют как «свойство без волос» механизма внимания, согласно которому сложные и различающиеся внутренние параметрические структуры проецируются в схожие низко-коррелированные выходные представления. Иными словами, механизм внимания действует как оператор когерентного усреднения, сглаживающий различия между методами полного и параметро-эффективного дообучения на уровне итоговых представлений.

С практической точки зрения это свойство даёт объяснение высокой эффективности методов PEFT, в частности LoRA, несмотря на их ограниченную выразительность по сравнению с полным fine-tuning. Ограничение сложности внутренних корреляций служит своеобразной формой регуляризации, которая позволяет модели достигать сопоставимого качества генерации, особенно в условиях ограниченного объёма обучающих данных и вычислительных ресурсов.

Таким образом, информационно-теоретический анализ PEFT предоставляет дополнительную интерпретационную рамку для понимания компромисса между качеством и ресурсами при дообучении языковых моделей. В контексте данной работы представленные идеи используются для качественного объяснения экспериментальных результатов, полученных при исследовании влияния параметров LoRA, типа квантизации и объёма данных на эффективность дообучения моделей. <https://arxiv.org/abs/2601.06788>

2.7. Совместное применение LoRA и квантизации

Несмотря на снижение числа обучаемых параметров при использовании LoRA, замороженные веса языковой модели, как правило, хранятся в формате полной или полуточной арифметики (FP32 или FP16), что приводит к значительным требованиям к памяти GPU. В связи с этим в последние годы активно исследуются методы квантизации backbone-моделей, направленные на снижение объёма памяти и ускорение инференса.

Одним из наиболее значимых подходов является QLoRA, предложенный Dettmers et al. (2023). В данном методе веса предобученной модели квантуются до 4 бит с использованием NormalFloat-квантизации, а обучение осуществляется только для LoRA-адаптеров. Дополнительно применяется двойная квантизация и механизм динамического перемещения состояний оптимизатора между CPU и GPU.

Экспериментальные результаты показывают, что QLoRA позволяет дообучать модели с десятками миллиардов параметров на одной GPU при сохранении качества, сопоставимого с FP16 fine-tuning. Однако ряд исследований указывает на возможные деградации качества при пост-тренировочной квантизации, что подчёркивает важность квантизационно-осознанных методов обучения.

Таким образом, совместное применение LoRA и квантизации представляет собой перспективное направление для адаптации языковых моделей в условиях ограниченных вычислительных ресурсов. Вместе с тем влияние параметров LoRA, типа квантизации и объёма обучающих данных на качество генерации остаётся недостаточно систематизированным, что определяет актуальность дальнейших исследований в данной области <https://arxiv.org/abs/2303.15647>

2.8. Совмещение PEFT и квантизации для снижения требований к памяти

Несмотря на значительное сокращение числа обучаемых параметров, методы PEFT, такие как LoRA, предполагают хранение базовой модели в полной точности (FP16 или FP32). Это приводит к тому, что основная часть VRAM всё ещё расходуется на замороженные веса модели.

Для решения этой проблемы были предложены методы **квантизации весов базовой модели**, позволяющие хранить параметры в пониженной точности (8-бит, 4-бит и ниже). Наиболее значимым вкладом в этом направлении является метод **QLoRA**, в котором:

- базовая модель хранится в 4-битном формате NormalFloat (NF4);
- используется двойная квантизация для уменьшения накладных расходов;
- состояния оптимизатора при необходимости выгружаются в CPU-память.

QLoRA позволяет дообучать модели с десятками миллиардов параметров на одной GPU, сохраняя качество, сопоставимое с full fine-tuning. Однако было показано, что пост-тренировочная квантизация может приводить к деградации качества, что стимулировало развитие методов квантизационно-осознанного дообучения (LoftQ, IR-QLoRA, LQ-LoRA).

Таким образом, сочетание PEFT и квантизации является ключевым направлением для практического применения LLM в условиях ограниченных вычислительных ресурсов. <https://arxiv.org/abs/2303.15647>

2.9. Ограничения и открытые проблемы PEFT

Несмотря на значительные успехи методов parameter-efficient fine-tuning, ряд ограничений и нерешённых проблем остаётся актуальным.

Например, при недостаточном значении ранга выражительная способность низкоранговых обновлений может оказаться ограниченной, что приводит к ухудшению качества модели, особенно для сложных задач и доменных адаптаций

- фиксированный ранг LoRA не оптimalен;

- разные слои требуют разной размерности адаптации;
- деградация качества при малом ранге <https://arxiv.org/abs/2303.10512>.

Дополнительные сложности возникают при совместном применении PEFT и агрессивной квантизации. Хотя методы, такие как QLoRA, позволяют существенно снизить требования к памяти, они могут быть чувствительны к выбору типа данных (dtype), схемы квантизации и гиперпараметров обучения. Некорректные настройки способны приводить к нестабильности обучения и деградации качества генерации <https://arxiv.org/abs/2305.14314>

Ещё одной открытой проблемой является компромисс между качеством и производительностью модели. Снижение точности представления весов и ограничение параметрического пространства, с одной стороны, уменьшают вычислительные затраты, но с другой — могут увеличивать задержки инференса или снижать устойчивость модели. Оптимальный баланс между latency, потреблением ресурсов и качеством остаётся предметом активных исследований <https://arxiv.org/abs/2303.15647>.

Кроме того, влияние объёма обучающих данных на эффективность PEFT-методов изучено недостаточно систематически. В частности, остаётся открытым вопрос, в каких режимах (малые или средние объёмы данных) методы PEFT обеспечивают наибольший выигрыш по сравнению с полным fine-tuning.

2.10. Выводы по главе

В данной главе были рассмотрены основные подходы к параметро-эффективному дообучению языковых моделей, а также проанализированы их теоретические и практические особенности. Показано, что методы PEFT позволяют существенно снизить вычислительные затраты на обучение и адаптацию больших языковых моделей при сохранении конкурентоспособного качества.

Особое внимание было удалено методу LoRA и его расширениям, а также совместному применению PEFT и квантизации. Анализ существующих работ показывает, что сочетание LoRA и QLoRA представляет собой разумный и практически применимый базовый подход для дообучения языковых моделей в условиях ограниченных вычислительных ресурсов.

Рассмотренные модификации LoRA, такие как DoRA и AdaLoRA, направлены на повышение выразительности и эффективности низкоранговых адаптаций, однако сопровождаются увеличением сложности обучения и дополнительных требований к памяти. Это обосновывает необходимость их сравнительного анализа в экспериментальной части работы.

Выявленные ограничения и открытые вопросы, связанные с выбором ранга адаптеров, параметров квантизации и объёма данных, формируют основу для дальнейшего экспериментального исследования. В следующей главе будет описана постановка задачи, используемые данные и методология экспериментов, направленных на количественную оценку влияния указанных факторов на качество и эффективность дообучения языковых моделей.

