УДК 004.032.26

Параметро-эффективное дообучение больших языковых моделей

И. В. Бабич, К. Н. Ефименко
Донецкий национальный технический университет
кафедра «Прикладная математика и искусственный интеллект»
E-mail: babichivanvictorovich@yandex.ru

Аннотация:
В статье приведён обзор теоретической основы parameter-efficient fine-tuning (PEFT) для больших языковых моделей (LLM). Рассмотрены ключевые методы низкоранговых адаптаций (LoRA) и их комбинация с низкобитной квантизацией (QLoRA). Дополнительно приведены рекомендации по выбору ранга адаптации и уровня квантизации на основе спектра Гессиана и эффективной размерности задачи. Успешность использования методов PEFT зависит от правильного выбора параметров, прежде всего ранга адаптации и глубины квантизации.

Введение

Современный этап цифровизации всех сфер жизнедеятельности отличается широчайшим использованием искусственного интеллекта [1-6]. Современные нейронные сети на основе архитектуры трансформеров, известные как большие языковые модели (Large Language Models, LLM), демонстрируют впечатляющие возможности в решении задач обработки естественного языка: генерации текста, автоматическом переводе, ответах на вопросы, извлечении информации и многих других. Высокое качество работы этих моделей достигается путём обучения огромного количества параметров (весов нейронной сети). Например, модель GPT-3 имеет порядка 175 миллиардов параметров.

Общая постановка проблемы

Большой масштаб порождает серьёзную проблему, известную как парадокс масштабирования: для адаптации уже обученной большой языковой модели под конкретную прикладную задачу (например, под медицинские или юридические тексты) необходимо выполнить её дополнительное обучение (дообучение, fine-tuning). При этом, из-за огромного размера LLM (десятки и сотни гигабайт), такое дообучение требует значительных вычислительных ресурсов – сотни графических процессоров (GPU) и большой объём оперативной памяти.

Таким образом, становится актуальной задача поиска параметро-эффективных методов дообучения (Parameter-Efficient Fine-Tuning, PEFT). Под параметро-эффективностью понимается возможность адаптации модели с минимальным изменением её исходных параметров. Формально задача параметро-эффективного дообучения ставится следующим образом:

Пусть имеется большая обученная нейросеть с параметрами (весами) θ_0. Необходимо найти такое минимальное изменение параметров $\Delta\theta$, что:

1. $\|\Delta\theta\| \ll \|\theta_0\|$, то есть изменение весов намного меньше их исходного значения.

2. Новая модель с параметрами $\theta_0+\Delta\theta$ минимизирует целевую функцию ошибки $L(\theta_0+\Delta\theta)$ на специфическом наборе данных.

Целью работы является адаптация модели к новой задаче, минимально затрагивая её исходные параметры и сократив вычислительные затраты на обучение.

Общая концепция и подходы PEFT

Основная идея параметро-эффективного дообучения состоит в том, чтобы ограничить пространство изменений параметров. Вместо того, чтобы изменять все миллиарды параметров исходной модели, мы разрешаем модели изменять лишь небольшую часть или выполнять изменения по определённым правилам. Благодаря этому существенно сокращается объём памяти и вычислительные ресурсы, необходимые для дообучения.

Рассмотрим наиболее распространённые подходы PEFT:

1. Низкоранговые адаптации (Low-Rank Adaptation, LoRA)

LoRA предлагает представлять изменения в весах нейросети в виде специальной матрицы низкого ранга [7]. Что такое низкий ранг? Это значит, что матрица изменений может быть записана как произведение двух матриц меньшего размера. Например, если исходная матрица имеет размер 1000×1000 (1 миллион элементов), то её низкоранговая версия может быть представлена двумя матрицами размером 1000×10 и 10×1000, что в сумме даёт лишь 20 тысяч элементов. Это радикально уменьшает объём параметров для

обучения, сохраняя при этом достаточную гибкость для адаптации.

Формально это записывают как:

$$\Delta W = B \times A,$$

где B и A – небольшие матрицы низкого ранга, которые и будут обучаться вместо исходных весов W.

Таким образом, исходные веса модели остаются нетронутыми, а изменения выражаются через небольшое количество новых параметров.

2. Квантизация весов и адаптация (QLoRA)

Квантизация (англ. quantization) – это способ представления чисел (параметров) с меньшей точностью. Например, вместо 32-битного вещественного числа можно использовать 8-битное или даже 4-битное представление. Это приводит к уменьшению памяти, необходимой для хранения модели. Однако квантизация вносит ошибки округления, которые снижают точность модели. Метод QLoRA совмещает низкоранговые адаптации (LoRA) и квантизацию так, чтобы минимизировать потери точности. Сначала модель сжимается за счёт квантизации (до 4 бит на параметр), а затем поверх неё накладываются небольшие адаптационные матрицы LoRA с более высокой точностью (например, 16 бит), что позволяет компенсировать потери качества, вызванные квантизацией [8].

3. Частичное дообучение (BitFit и IA^3)

Эти методы ещё сильнее ограничивают изменения параметров. Например, в методе BitFit изменяются только отдельные смещения (biases) нейронов, а веса остаются неизменными. В IA^3 изменения ограничены масштабированием активаций, то есть модель изменяет только коэффициенты, на которые умножаются промежуточные значения внутри слоёв сети. Оба подхода дают существенную экономию в ресурсах, хотя и менее гибкие, чем LoRA.

4. Дообучение через префиксы (Prompt/Prefix tuning)

Эти методы вообще не изменяют исходную модель. Вместо этого они добавляют к входу модели небольшие обучаемые последовательности (префиксы), которые подсказывают нейросети, как нужно адаптироваться к новой задаче. Таким образом, параметры самой нейросети вообще не меняются, меняется лишь специальный входной контекст.

Все указанные подходы объединяет одно – они решают задачу адаптации нейронных сетей минимальными средствами. Среди всех методов низкоранговые адаптации LoRA и их вариант с квантизацией QLoRA стали наиболее популярными благодаря оптимальному балансу гибкости и эффективности. В дальнейших разделах статьи мы более глубоко рассмотрим именно их математическое обоснование и теоретические свойства.

Низкоранговая адаптация (LoRA)

Низкоранговая адаптация (Low-Rank Adaptation, LoRA) является одним из наиболее эффективных подходов к параметро-эффективному дообучению больших языковых моделей. Главная идея LoRA заключается в том, что любые изменения в параметрах модели представляются в виде матриц низкого ранга. Рассмотрим, почему это важно.

Весовые матрицы в трансформерах обычно имеют очень большой размер. Например, одна весовая матрица слоя может содержать миллионы или даже сотни миллионов элементов. При стандартном дообучении все эти элементы приходится изменять. LoRA предлагает вместо полного изменения весов добавить к исходной матрице весов небольшую низкоранговую матрицу.

Важное теоретическое преимущество подхода LoRA заключается в его экспрессивности, то есть в способности модели, использующей низкоранговую адаптацию, воспроизводить решения, доступные при полном изменении всех параметров. Если ранг r выбран не меньше, чем так называемый эффективный ранг задачи (определяемый рангом матрицы Гессиана целевой функции ошибки в точке исходного минимума), то низкоранговое решение способно достичь практически того же минимального значения ошибки, что и полное дообучение [7].

Иными словами, LoRA сохраняет «выразительную силу» полной настройки параметров, но значительно экономнее в плане ресурсов, так как существенно ограничивает пространство поиска оптимальных параметров.

LoRA с квантизацией параметров (QLoRA)

Одним из усовершенствований низкоранговой адаптации является подход QLoRA (Quantized Low-Rank Adaptation), который сочетает низкоранговые адаптации и метод квантизации параметров.

Квантизация параметров – это метод представления весов нейронной сети с меньшим количеством бит. Например, вместо стандартного представления весов в формате 32 бита на каждый параметр используются 4-битные веса. Это приводит к восьмикратному сокращению объёма памяти, необходимого для хранения модели [9].

Однако такой метод квантизации вносит неизбежные погрешности округления, называемые квант-шумом. Квант-шум может

существенно ухудшить точность работы нейронной сети, особенно если используется очень низкая точность (например, 4 бита).

Идея QLoRA заключается в следующем:

– исходные веса большой модели сжимаются путём 4-битной квантизации (обычно используется специальный формат NF4, в котором веса представлены в 4-битах).

– поверх этой квантизованной модели накладываются дополнительные низкоранговые адаптеры LoRA в формате с более высокой точностью (обычно 16 бит). Именно эти адаптеры компенсируют потери точности, вызванные квантованием.

Таким образом, QLoRA получает двойную выгоду: сокращает память за счёт квантизации и уменьшает количество обучаемых параметров за счёт низкоранговой структуры. При этом теоретически доказано, что квант-шум остаётся ограниченным (не превышает некоторой небольшой величины), и использование адаптеров позволяет практически полностью восстановить точность исходной модели после квантования.

Теоретические свойства и ограничения PEFT

Методы параметро-эффективного дообучения, такие как LoRA и QLoRA, обладают рядом важных теоретических свойств, которые делают их привлекательными как с практической, так и с теоретической точки зрения (рис. 1).

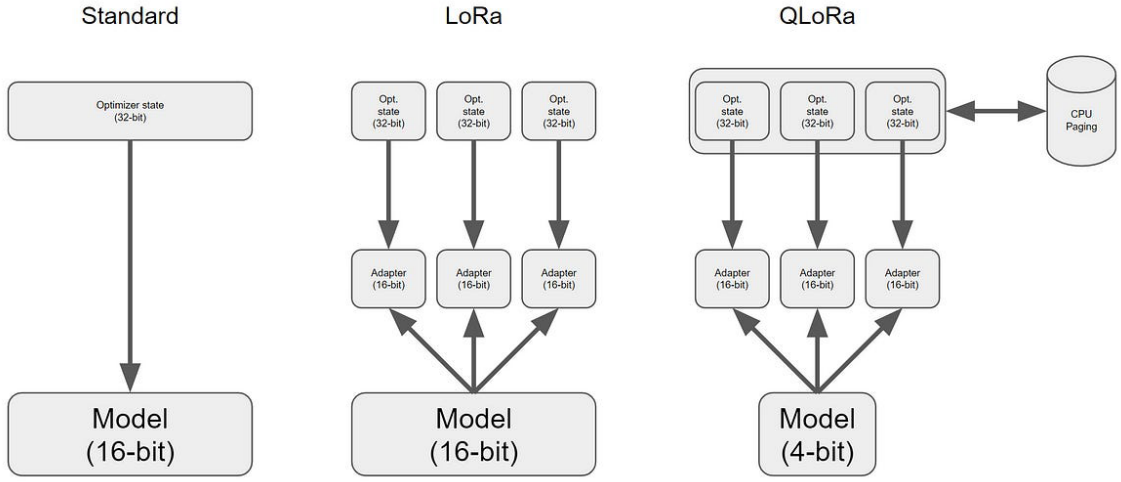

Рисунок 1 – Сравнение схем дообучения LLM: Standard, LoRA и QLoRA

Одним из ключевых преимуществ является то, что использование низкоранговых адаптаций ограничивает сложность модели с точки зрения теории обучения. Концептуально это связано с понятием VC-размера (Vapnik–Chervonenkis dimension), которое характеризует способность модели подстраиваться под любые данные. Чем выше VC-размер, тем больше риск переобучения, то есть модель начинает «запоминать» данные, теряя способность хорошо работать на новых примерах [10].

LoRA и QLoRA снижают VC-размер исходной большой модели за счёт ограничения числа адаптируемых параметров и пространства, в котором происходит обучение. Благодаря этому улучшаются теоретические границы обобщающей способности модели – она становится более устойчивой к переобучению и способна показывать хорошие результаты даже на относительно небольших наборах данных.

Другой важный эффект низкоранговых адаптаций связан с поведением функции ошибки модели. Исследования показывают, что низкоранговые изменения параметров модели чаще всего приводят к относительно «гладкому» изменению функции ошибки. Другими словами, если исходная модель уже находится в локальном минимуме ошибки, то дообучение с использованием небольших низкоранговых адаптеров чаще всего ведёт к минимальному изменению ошибки. Это говорит о том, что модели, адаптируемые через LoRA, находятся в более «плоских» зонах ландшафта ошибок, что также связано с лучшей способностью к обобщению и стабильностью решений.

Однако у подходов PEFT есть и существенные ограничения. Основное ограничение связано с выбором оптимального значения ранга. Слишком маленький ранг не даст модели достаточно гибкости для адаптации, а слишком большой – снизит преимущество метода в плане ресурсоёмкости. Также PEFT-методы не всегда совместимы с более сложными алгоритмами обучения второго порядка

(например, использующими матрицу Гессиана), так как оптимизация в сильно ограниченном пространстве параметров может плохо взаимодействовать с такими методами.

Параметро-эффективные подходы требуют тщательного выбора ограничений и учёта особенностей конкретной задачи, но при правильной реализации способны обеспечить существенные преимущества.

Выбор ранга и уровней квантизации

При практическом использовании низкоранговой адаптации (LoRA) и её комбинации с квантизацией (QLoRA) важнейшим вопросом становится правильный выбор двух основных параметров:

1. Ранга адаптации – насколько большим будет ранг низкоранговой матрицы, отражающей изменения параметров;

2. Уровня квантизации – количества бит, которое используется для хранения каждого параметра модели.

Эти параметры непосредственно влияют на компромисс между качеством и требуемыми ресурсами модели.

При выборе ранга основной идеей является баланс между гибкостью модели (способностью эффективно подстраиваться под задачу) и количеством ресурсов, которые потребуются для её обучения и хранения.

С теоретической точки зрения оптимальный выбор ранга напрямую связан с понятием эффективной размерности задачи. Эффективная размерность определяется тем, насколько разнообразны данные и насколько сложна целевая задача. Формально это связывается с спектром матрицы Гессиана, которая представляет собой вторые производные целевой функции ошибки модели по параметрам. Спектр матрицы Гессиана показывает, в каких направлениях параметров изменения наиболее существенно влияют на ошибку модели, а в каких – почти не влияют [9].

На практике это означает следующее: если спектр Гессиана быстро убывает (то есть имеется небольшое число «сильных» направлений, по которым изменения веса значительно снижают ошибку), то эффективная размерность задачи мала. Тогда для адаптации достаточно выбрать небольшой ранг. Если же спектр Гессиана убывает медленно, и эффективная размерность задачи велика (сложная задача, разнообразные данные), то необходим больший ранг.

Таким образом, на практике рекомендуют начинать адаптацию с относительно небольшого ранга (например, 8 или 16). Если качество модели оказывается неудовлетворительным, постепенно увеличивайте ранг до момента, когда качество начнёт заметно улучшаться. Обычно это значение будет оптимальным компромиссом. Для большинства практических задач средние значения ранга в пределах 32-64 обеспечивают хороший баланс между эффективностью и качеством.

При использовании QLoRA модель дополнительно сжимается за счёт низкобитного представления параметров. Ключевой параметр – число бит, выделяемых на параметр, влияет на компромисс между экономией памяти и точностью работы модели.

32-битное представление – стандартный формат хранения параметров, не вносит дополнительных ошибок, но требует много памяти.

8-битная квантизация – является распространённым и безопасным выбором, поскольку почти не ухудшает качество моделей и экономит в четыре раза больше памяти.

4-битная квантизация (NF4) – обеспечивает ещё большую экономию (до восьми раз по сравнению с 32-битным форматом), однако здесь уже необходимо учитывать наличие заметного квантового шума и компенсировать его увеличением ранга адаптации.

Для типовых приложений, где ресурсы ограничены, но требуется высокое качество, хорошим стартом является 8-битная квантизация. Если стоит задача максимально снизить потребление ресурсов (например, адаптация модели на мобильных устройствах), то переходят к 4-битному формату. При этом важно увеличить ранг адаптации (например, с 16-32 до 64-128), чтобы минимизировать падение точности, вызванное квантованием.

Выводы

Рассмотренные методы параметро-эффективного дообучения больших языковых моделей, в частности низкоранговая адаптация (LoRA) и её комбинация с квантизацией (QLoRA), обеспечивают эффективный компромисс между сложностью и гибкостью адаптации моделей. Их главная особенность заключается в ограничении пространства параметров, что позволяет значительно сократить объём вычислительных ресурсов и памяти, необходимых для дообучения.

С теоретической точки зрения ключевыми преимуществами LoRA и QLoRA являются высокая экспрессивность, благодаря которой модель сохраняет способность достигать оптимального минимума ошибки, и улучшенные свойства обобщения за счёт снижения VC-размера модели. Также важно отметить, что низкоранговые адаптации способствуют гладкости ландшафта ошибки, что улучшает устойчивость модели к изменениям данных и

уменьшает вероятность переобучения.

Тем не менее, успешность использования методов PEFT зависит от правильного выбора параметров, прежде всего ранга адаптации и глубины квантизации. Поэтому дальнейшие теоретические исследования должны быть направлены на более точное понимание связи между рангом адаптации, уровнем квант-шумов и способностью модели сохранять высокие показатели качества при минимальных вычислительных затратах.

Литература

1. Федяев, О. И. Интеллектуальная система принятия решений в отделении медицинского учреждения на основе нейросетевых, продукционных и статистических моделей / О. И. Федяев, В. С. Бакаленко // Статистика и Экономика, 2019. – № 16(3). – С. 70-77. – URL: https://doi.org/10.21686/2500-3925-2019-3-70-77

2. Дворяткина, С. Н. Интеграция фрактальных и нейросетевых технологий в педагогическом контроле и оценке знаний обучаемых / С. Н. Дворяткина // Вестник РУДН. Серия : Психология и педагогика, 2016. - Том. 14. - № 4. – С. 451-465.

3. Федяев, О. И. Прогнозирование остаточных знаний студентов по отдельным дисциплинам с помощью нейронных сетей / О. И. Федяев // Известия ЮФУ. Технические науки. – 2016. – С. 122-136.

4. AI в обучении: на что способны технологии уже сейчас? // EduTech. - 2022. - № 4 [49]. – 60 с.

5. Струнин, Д. А. Искусственный интеллект в сфере образования / Д. А. Струнин. — Текст : непосредственный // Молодой ученый. — 2023. — № 6 (453). — С. 15-16. — URL: https://moluch.ru/archive/453/99921/

6. Федяев, О. И. Автоматическая регистрация присутствия студентов на учебном занятии с помощью компьютерного зрения / О. И. Федяев, И. А. Коломойцева // XXI Национальная конференция по искусственному интеллекту с международным участием КИИ-2023 (Смоленск, 16-20 октября 2023 г.). Труды конференции. В 2-х томах. Т.1. – Смоленск: Принт-Экспресс, 2023. – С. 294-303.

7. Портал huggingface.co [Электронный ресурс] / Интернет-ресурс. – Режим доступа: https://huggingface.co/docs/text-generation-inference/conceptual/lora. – Загл. с экрана.

8. Портал huggingface.co [Электронный ресурс] / Интернет-ресурс. – Режим доступа: https://huggingface.co/blog/4bit-transformers-bitsandbytes. – Загл. с экрана.

9. Портал geeksforgeeks.org [Электронный ресурс] / Интернет-ресурс. – Режим доступа: https://www.geeksforgeeks.org/what-is-qlora-quantized-low-rank-adapter/. – Загл. С экрана.

10. Портал ru.wikipedia.org [Электронный ресурс] / Интернет-ресурс. – Режим доступа: https://ru.wikipedia.org. – Загл. с экрана.

Бабич И.В., Ефименко К.Н. Параметро-эффективное дообучение больших языковых моделей. *В статье приведён обзор теоретической основы parameter-efficient fine-tuning (PEFT) для больших языковых моделей (LLM). Рассмотрены ключевые методы низкоранговых адаптаций (LoRA) и их комбинация с низкобитной квантизацией (QLoRA). Дополнительно приведены рекомендации по выбору ранга адаптации и уровня квантизации на основе спектра Гессиана и эффективной размерности задачи. Успешность использования методов PEFT зависит от правильного выбора параметров, прежде всего ранга адаптации и глубины квантизации.*

Ключевые слова: *параметро-эффективное дообучение, PEFT, LoRA, QLoRA, квантизация, спектр Гессиана.*

Babich I., Efimenko K. Parameter-Efficient Fine-Tuning of Large Language Models. *This paper provides an overview of the theoretical foundations of parameter-efficient fine-tuning (PEFT) for large language models (LLM). It examines the key methods of low-rank adaptation (LoRA) and their combination with low-bit quantization (QLoRA). In addition, it offers recommendations for selecting the adaptation rank and quantization level based on the Hessian spectrum and the intrinsic dimension of the task. The success of using PEFT methods depends on the correct choice of parameters, primarily the adaptation rank and the depth of quantization.*

Keywords: *parameter-efficient fine-tuning, PEFT, LoRA, QLoRA, quantization, Hessian spectrum.*

Статья поступила в редакцию 12.03.2025
Рекомендована к публикации профессором Павлышом В. Н.