LoRA: Low-Rank Adaptation of Large Language Models

Авторы статьи: Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen

Ссылка на статью: https://arxiv.org/abs/2106.09685

Автор исследования: Аюпов Шамиль

Контекст

Первый препринт вышел 17 июня 2021, затем статью обновили 16 октября. Статья недавняя, поэтому не представлена на конференциях, но авторы подались на ICLR 2022 (International Conference on Learning Representations). Сейчас (на момент написания этого текста) там прошел ревью-период и идет обсуждение авторов с ревьюерами.

Авторы

Все авторы (их 8) работают в Microsoft Research. Тут важно понимать, что у них есть веса GPT-3, так как Microsoft сотрудничает с OpenAI (вспоминаем Codex)
Два основных автора (помечены, как Equal Work):

Edward Hu, молодой исследователь (статьи пишет с 2019 года), который тем не менее пару раз уже поучаствовал в ICML и паре других конференций. У него 11 работ и 284 цитирований. Основная сфера его интересов – это NLP, хотя есть и пара работ про Adversarial атаки.

Yelong Shen, наоборот, уже опытный исследователь, самые известные его работы были опубликованы в 2012-2017 годы. У него ~77 работ и ~4k цитирования. Есть некоторая неоднозначность кол-ва его статей и цитирований (в разных источниках по-разному). Сфера интересов: information retrieval.

Интересно заметить, что самые цитируемые работы вышли в дотрансформерную эпоху (они как раз про то, как свертками и рекуррентными сетями вытаскивать полезную информацию), потом же цитируемость уменьшилась (видимо для этих задач повсеместно начали использовать трансформеры).

Вдохновители

В статье 65 ссылок, но главными источниками вдохновения я бы назвал следующие работы:

Li et al. "Measuring the Intrinsic Dimension of Objective Landscapes" (2018). Это статья, в которой рассказывается идея о том, что большие модели на самом деле лежат в пространствах маленькой размерности (intrinsic dimension).

Aghajanyan et al. "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning" (2020). Это эмпирически-теоретическая статья от Facebook, продолжение идеи выше, но больше про языковые модели. Эта статья была представлена на ACL 2021 как Outstanding Paper

LoRA использует эти идеи, представляя обновление как низкоранговую добавку.

Конкуренты

Отмечу, что идейных конкурентов много (много разных модификаций методов), но основными выделил бы следующие:

Houlsby et al. "Parameter-Efficient Transfer Learning for NLP" (2019). Это статья от Google Research, в которой представлены адаптерные слои для дообучения трансформеров.

Li et al. “Prefix-Tuning: Optimizing Continuous Prompts for Generation” (2021) – связанный с обучением для downstream задач отдельных наборов префиксных векторов.

Помимо этого есть и статьи, которые вышли примерно в это же время:

Mahabadi et al. “Compacter: Efficient Low-Rank Hypercomplex Adapter Layers” (2021) – продолжение идеи с адаптерами

Chen et al. “DSEE: Dually Sparsity-embedded Efficient Tuning of Pre-trained Language Models” (2021) – в ней пользуются идеей прунинга обновлений. Авторы уверяют, что в ней скорость инференса больше, чем в других подходах.

Продолжатели

Цитирований данных работ – 7.
Работ, которые продолжают анализировать именно LoRA и ее свойства пока нет, зато есть интересная работа **Mao** et al. “UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning” (2021) (Facebook AI), в которой авторы ищут ответ на вопрос: в разных задачах разные методы эффективного тюнинга дают разные по качеству результаты, можно ли эффективно и автоматически выбирать метод?

Дальнейшее исследование

Авторы сами выделяют несколько направлений, среди них:

- Метод ортогонален некоторым другим, можно ли их эффективно комбинировать?
- “Низкоранговость” метода может помочь ответить на вопрос: почему предобученные модели хорошо обобщаются на downstream задачи.

- Можно ли определить, когда предлагаемый метод будет работать хорошо, а когда нет? Другими словами когда использовать метод?

Кроме того, у меня самого возник вопрос: а как предложенные ранее методы влияют на скорость сходимости и как они сопоставляются друг с другом?

Приложение в индустрии

Данный метод может помочь расширить круг применений модели следующим образом: можно использовать одну модель и много параметров LoRA для разных задач, подменять их при применении. Учитывая сложность и цену эксплуатации огромных языковых моделей, это может увеличить количество разных задач, в которых применяется модель.
Хотя авторы и уверяют в том, что их подход не увеличивает время применения, в постановке выше это ложно – подмена параметров для разных задач требует времени.
Если же проблема времени применения для какой-то задачи критична, то модель под эту задачу можно склонировать и попробовать провести полный fine-tuning (для лучшего качества). В таком случае есть экспериментаторская ценность LoRA, метод помогает определить, когда вообще стоит тратить время и ресурсы на fine-tuning.