Учёные раскрыли, как снизить затраты на обучение больших языковых моделей.

Специалисты из Microsoft – это американская многопрофильная компания, занимающаяся разработкой программного обеспечения и производством компьютерной техники. Она была основана в 1975 году Биллом Гейтсом и Полом Алленом и на сегодняшний день является одной из самых крупных и известных IT-компаний в мире.

 Среди продуктов Microsoft наиболее известными являются операционные системы Windows, пакеты офисных приложений Office, браузер Internet Explorer и поисковая система Bing. Кроме того, компания занимается разработкой программного обеспечения для серверов, баз данных, игровых консолей Xbox и многих других устройств.

 Microsoft также предоставляет услуги облачных вычислений и хранения данных через свою платформу Azure, а также занимается разработкой искусственного интеллекта и других инновационных технологий. Компания имеет филиалы по всему миру и сотрудничает с многими крупными корпорациями и организациями." data-html="true" data-original-title="Microsoft" >Microsoft и Бэйханского Университета разработали инновационную технику для тонкой настройки больших языковых моделей (LLM), которая значительно снижает затраты.

Новая методика «MoRA» представляет собой параметрически эффективную технику тонкой настройки (PEFT-методы (Parameter-Efficient Fine-Tuning) — это подходы к дообучению LLM-моделей, которые позволяют сократить количество обучаемых параметров, сохраняя при этом высокую производительность модели.

 Основная идея PEFT-методов заключается в адаптации небольшого числа параметров, а не всей модели, что делает процесс дообучения более эффективным и менее ресурсоёмким.

 Примеры PEFT-методов включают Low-Rank Adaptation (LoRA), адаптацию через мягкие промпты (prompt tuning) и другие техники, использующие небольшие модули или слои для дообучения." data-html="true" data-original-title="PEFT" >PEFT), устраняющую ограничения другого популярного метода – LoRA (low-rank adaptation). MoRA особенно полезна, когда необходимо обучить модель новым знаниям. С ростом популярности PEFT-методов в бизнес-среде, MoRA может стать важным инструментом для разработчиков LLM-приложений.

Преимущества и недостатки LoRA

Классическая тонкая настройка требует обновления всех параметров модели, что становится затратным и медленным процессом при наличии миллиардов параметров. PEFT-методы позволяют найти оптимальное подмножество параметров, необходимых для настройки модели под конкретную задачу.

LoRA (Low-Rank Adaptation) — метод адаптации LLM-моделей, при котором веса предобученной модели фиксируются, а к ним добавляются обучаемые матрицы низкого ранга. Метод позволяет значительно уменьшить количество параметров, необходимых для дообучения модели на новые задачи, сохраняя её производительность." data-html="true" data-original-title="LoRA" >LoRA стала популярной благодаря способности обновлять параметры через матрицы низкого ранга, что значительно снижает требования к памяти. Однако LoRA не всегда справляется с более сложными задачами, такими как математическое рассуждение и постоянное предварительное обучение.

Введение MoRA

LoRA (слева) использует матрицы низкого ранга, а MoRA (справа) использует одну квадратную матрицу для точной настройки с эффективным использованием параметров

Для устранения ограничений LoRA учёные представили MoRA, которая использует квадратные матрицы вместо низкоранговых. Главная идея MoRA заключается в использовании обучаемых параметров для достижения максимального ранга в пространстве исходных размеров модели. В отличие от LoRA, входные и выходные размеры адаптера MoRA не совпадают с исходной моделью, поэтому была разработана функция сжатия/декомпрессии, которая преобразует данные между двумя пространствами.

Результаты тестирования MoRA

Кривая потерь MoRA очень похожа на полную настройку для задач по запоминанию знаний

Тестирование моделей LoRA и MoRA одинакового размера показало, что MoRA значительно превосходит LoRA в задачах запоминания и приближается к производительности полностью настроенной модели. В задачах настройки инструкций и математического рассуждения MoRA показала результаты, сравнимые с LoRA, но в постоянном предобучении в биомедицинской и финансовой сферах MoRA превзошла LoRA.

PEFT для бизнеса

Тонкая настройка является важной задачей для корпоративных приложений LLM. Она позволяет компаниям использовать меньшие модели для задач, ранее требовавших дорогих передовых моделей. LoRA и её варианты являются золотым стандартом параметрически эффективной тонкой настройки. Существует множество инструментов и платформ для создания адаптеров LoRA, таких как S-LoRA, позволяющий запускать тысячи адаптеров на одном GPU.

Ученые выпустили реализацию MoRA с открытым исходным кодом, совместимую с LoRA. Это может оказаться важным инструментом для корпоративных приложений, которые хотят добавить новые знания в базовые модели.

На перекрестке науки и фантазии — наш канал