Повышение качества и снижение стоимости обучения больших языковых моделей

Косарев Евгений Александрович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: evgenijkkk@yandex.ru

Научный руководитель *— Воронцов Константин Вячеславович*

Большие языковые модели с высоким качеством способны решать множество задач от определения спама до семантического анализа данных. Однако они все еще демонстрируют низкое качество на узкоспециализированных задачах, например в определении манипулятивности или поляризованности текста. Для решения этой проблемы предлагается экономящий вычислительные ресурсы метод, способный повышать качество языковой модели на разнообразных специализированных задачах.

Введем два режима обучения:

- **High-res** - обучающих данных больше 2 тысяч примеров.
- **Low-res** - обучающих данных меньше 2 тысяч, обычно несколько сотен примеров. В этом режиме собирается существенно меньше данных, что ускоряет и удешевляет подготовку обучающего множества.

Обучать большие языковые модели ресурсозатратно, так как требуется оптимизация большого числа параметров. Для экономии ресурсов, предлагается модифицировать метод P-tuning [1], назовем его P-tuning & bias. В исходном методе к входу в модель слева и справа добавляется несколько специальных обучаемых токенов, остальные параметры модели не обучаются. Модификация заключается в том, что специальные токены будут добавляться только слева, а к обучаемым параметрам добавятся $bias$ из всех линейных преобразований внутри модели вида $R(X) = AX + bias$.

Сравнивать ресурсозатратность и качество метода будем с обучением (Full Finetune) всех параметров модели. В качестве базового алгоритма выберем модель с тремя миллиардами параметров [2], назовем ее 3B. В данной работе были собраны 15 бенчмарков, задач текстовой аналитики на основе открытых текстовых коллекций и закрытых данных компании Яндекс, с их помощью будем оценивать качество модели.

	Качество lowres	Качество highres	Число GPU	Часы на highres
Full Finetune	0,5097	0,6883	64	8.23
P-tuning	0,5136	0,6787	**8**	**0.96**
P-tuning & bias	**0,5221**	0,6840	**8**	1.03

Таблица 1: Усредненные показатели качества и стоимости обучения по 15 задачам

Метод P-tuning & bias лучший в режиме lowres, схож с Full Finetune в highres, требует в 8 раз меньше видеокарт и обучается в 8 раз быстрее, чем Full Finetune.

Улучшим базовую модель 3B, единожды дообучив ее на 75 генеративных и классификационных задачах из открытого доступа в режиме Full Finetune (было проверено, что они не пересекаются с задачами из бенчмарков). Получим модель 3B(T), которую методом P-tuning & bias дообучим на решение произвольных задач, качество сравним на 15 бенчмарках:

Бечнмарк	3B **lowres**	3B(T) **lowres**	3B **highres**	3B(T) **highres**
Классификация	0,5466	**0,6006**	0,7079	0,7031
Генерация	0,5007	**0,6019**	0,6631	**0,6946**
Вместе	0,5221	**0,6013**	0,6840	**0,6985**

Таблица 2: Сравнение обощающих способностей 3B VS 3B(T)

Модель 3B(T) превосходит 3B в lowres и highres режимах в классификационных и генеративных задачах.

Литература

1. Liu X. et al. GPT understands, too //AI Open. – 2023.
2. Веса и описание Huggingface 3B модели: `https://huggingface.co/openlm-research/open_llama_3b/tree/main`