Training Compute-Optimal Large Language Models [Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Laurent Sifre] (https://arxiv.org/abs/2203.15556)
Автор обзора-рецензии: Козлова Ольга

1. Опишите суть работы в паре предложений, выделите ее основной вклад.

Авторы обнаружили, что принятые ранее правила скейлинга больших языковых моделей не работают. Для оптимального использования ресурсов при увеличении числа параметров модели в несколько раз, количество данных нужно увеличивать во столько же раз.

Основной вклад работы заключается в обновленных и уточненных правилах масштабирования больших моделей, которых сейчас становится все больше и больше. Это крайне важно, так как возможность ставить множество экспериментов в этой области представляется маловероятным ввиду высоких затрат временных и материальных ресурсов.

2. Когда написана работа? Опубликована ли она на какой-то конференции? Кто ее авторы, есть ли у них другие схожие работы? Подумайте как авторы пришли к идее статьи

Работа была опубликована 12 апреля 2022 года (блог-пост). Статья также была представлена на конференции NeurIPS 2022 в конце ноября (страница публикации).

Основные авторы: Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Laurent Sifre
Еще авторы: Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals

Большинство авторов занимается языковыми моделями уже продолжительное время, либо поддерживают large-scale архитектуры или внутренние инструменты DeepMind.

Авторы пришли к идеи статьи экспериментируя с большими моделями и заметив несостыковку с работой Kaplan et al. – подобранные там коэффициенты соотношения размера модели, объема тренировочных данных и используемых вычислительных ресурсов давали плохие результаты.

3. Какие из статей в списке ссылок оказали наибольшее влияние на данную работу? Можно ли выделить какие-то статьи, которые можно назвать базовыми для этой работы? Опишите в чем связь с этими работами.

Наибольшее влияние оказала статья Kaplan et al. (https://arxiv.org/abs/2001.08361). В ней было показано, что между количеством параметров в языковой модели и ее производительностью существует зависимость по степенному закону. После публикации в этой области обучаются все более крупные модели, ожидая улучшения производительности. В работе сделан примечательный вывод о том, что большие модели не должны обучаться до минимально возможного значения функции потерь, чтобы быть вычислительно оптимальными.

Хотя авторы Hoffmann et al. пришли к такому же выводу, по их оценкам, большие модели следует обучать на гораздо большем количестве данных, чем рекомендуют авторы. В частности, при увеличении вычислительного бюджета в 10 раз, по мнению Kaplan et al., размер модели должен увеличиться в 5,5 раз, в то время как количество данных должно увеличиться только в 1,8 раза. Вместо этого авторы Hoffmann et al. обнаружили, что размер модели и количество данных должны быть увеличены в равных пропорциях.

4. Кто цитирует данную статью? Есть ли у этой работы прямые продолжения, которые стоит прочесть тем, кто заинтересовался этой работой?

Всего 17 цитирований статьи. В основном работа используется как источник правил для скейлинга обучения.
Также встречаются работы, изучающие правила скейлинга для моделей в других областях, например в изображениях или графах, наблюдающие такие же зависимости.

5. Есть ли у работы прямые конкуренты?

Прямых конкурентов у статьи нет.

6. Опишите сильные, на ваш взгляд, стороны работы.

- **Вклад**: Закон масштабирования, изученный в данном эмпирическом исследовании, полезен для обучения больших языковых моделей (LLM). Авторы также утверждают, что существующие LLM сильно недоучены. Вывод, сделанный в этой статье, может показать интересное направление для сообщества, чтобы продолжать оптимизировать LLM: нам нужно обратить внимание на эффективное обучение данных вместо увеличения размера модели.
- **Оригинальность**: Хотя данная работа следует методологии предыдущего исследования, т.е. эмпирическому исследованию, в итоге в ней показан новый закон масштабирования.
- **Вариативность экспериментов**: Примечательно, что в ходе работы было проведено 3 эксперимента, каждый из которых подтвердил гипотезу.

7. Опишите слабые, на ваш взгляд, стороны работы.

- **Обоснованность:** Поскольку в данной работе используется эмпирический метод для исследования оптимального закона масштабирования, теоретическая основа не очень прочная. Можем ли мы как то судить о том, является ли модель недообученной, более обоснованным способом? Кроме того, учитывая, что три подхода к моделированию в данной работе опираются на эмпирические записи обучения, одной из проблем являются случайные факторы обучения.
- **Универсальность:** подобные эксперименты невозможно провести с небольшим объемом вычислительных ресурсов (в условиях небольшой лаборатории, например).

8. Предложите как можно было бы улучшить статью.

- Можно проверить как изменится соотношение в зависимости от данных – например, что будет при наличии повторяющихся данных, или что произойдет если данные будут плохого качества.
- Можно попробовать дать ответ на вопрос "что делать в условиях ограниченных ресурсов?" – например, малом количестве данных или ограниченности вычислительных ресурсов. Подобным вопросом задается Galactica (https://arxiv.org/abs/2211.09085)

9. Попробуйте на основе результатов статьи предложить исследование, не проведенное к текущему моменту, или идею применение в индустриальных приложениях.

Наиболее очевидное и необходимое исследование в данном случае, на мой взгляд, это теоретическое обоснование полученных результатов.