Training Compute-Optimal Large Language Models

Рецензия

Козлова Ольга

Вклад работы

- Закон масштабирования, изученный в данном эмпирическом исследовании, полезен для обучения больших языковых моделей.

- 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

Публикации и авторы

- Блог-пост 12 апреля 2022
- NeurIPS 2022

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

Предшествующие работы

Kaplan et al. https://arxiv.org/abs/2001.08361

- Между количеством параметров в языковой модели и ее производительностью существует зависимость по степенному закону

- Большие модели не должны обучаться до минимально возможного значения функции потерь, чтобы быть вычислительно оптимальными

- При увеличении вычислительного бюджета в 10 раз, размер модели должен увеличиться в 5,5 раз, количество данных должно увеличиться только на 1,8 раза

Как появилась статья

- Обнаружили, что что-то не так и надо по-другому:
    - При увеличении вычислительного бюджета в 10 раз, размер модели должен увеличиться в 5,5 раз, количество данных должно увеличиться ~~только в 1,8 раза~~ также в 5,5 раз.

Продолжения / конкуренты

- Подобные зависимости также наблюдают и в других областях DL
- Прямых конкурентов нет

Сильные стороны

- **Вклад** – правильный закон скейлинга крайне важен для сообщества
- **Вариативность экспериментов** – гипотеза была подтверждена несколькими способами

Слабые стороны

- **Нет теоретических обоснований** – все выводы основаны на эмпирических наблюдениях

- **Воспроизводимость** – нет возможности проверить эксперименты без огромного объема вычислительных ресурсов

Возможные продолжения / улучшения

- Вывод теоретических обоснований
- Зависимость результатов экспериментов от данных (оригинальные эксперименты проведены на уникальных данных)
- Что делать в случае ограниченности данных / ресурсов?

Источники

- Статья
  - https://arxiv.org/abs/2203.15556
- Блог-пост
  - https://www.deepmind.com/blog/an-empirical-analysis-of-compute-optimal-large-language-model-training
- NeurIPS 2022
  - https://openreview.net/forum?id=iBBcRUlOAPR
- Письменная рецензия
  - https://docs.google.com/document/d/15Kk3Jx5-lebNneyFllCJp2C3_WlsbC6IzG7VijPukcw/edit?usp=sharing