Рецензия на статью “Pixelated Butterfly: Simple and Efficient Sparse training for Neural Network Models“

Выполнил: Тряпицын Александр
Группа БПМИ192

Суть работы:
Авторы рассматривают метод статической спарсификации произвольной модели без потери качества и адаптируют его для современного вычислительного оборудования. В частности - авторы рассматривают паттерны спарсификации “sparse&low-rank matrices” (Candes et al. 2011, Udell et al. 2019, Chen et al. 2021), захватывающие локальную и глобальную информацию, и “butterfly matrices” (Parker et al. 1995, Dao et al. 2019), особый вид матриц позволяющий избежать комбинаторной проблемы поиска лучшего паттерна. Они отмечают, что этот метод не является предпочтительным для использования на современных GPU из-за особенностей блочного доступа к памяти у графического процессора. Авторы предлагают свое решение этой проблемы за счет блочного паттерна спарсификации и приводят сопутствующие эксперименты со всеми современными моделями. Результат - ускорение обучения и вычисления модели без ухудшения качества.

Контекст:
Статья была выпущена 30 ноября 2021 и представлена на конференции ICLR 2022 исследователями из 3 различных университетов (Stanford U., Peking U., U. at Buffalo) и 2 исследовательскими отделами компаний (SambaNova Systems, Adobe). При этом основной вклад внесли исследователи из Stanford University - Tri Dao и Beidi Chen.

Tri Dao имеет 259 цитирований в 2022 с индексом Хирша 12 и уже несколько лет занимается областью спарсификации вычислений и моделей машинного обучения.

Beidi Chen имеет 133 цитирования в 2022 с индексом Хирша 11 и занимется исследованием эффективности обучения и вычисления моделей машинного обучения на современных GPU.

Текущая статья является продолжением целый серии статей ключевых авторов, посвященных спарсификации моделей, и вдохновлена их предыдущими статьями Learning fast algorithms for linear transforms using butterfly factorizations (Tri Dao et al. 2019) и Scatterbrain: Unifying sparse and low-rank attention (Beidi Chen, Tri Dao et al. 2021).

Цитирования:
Данную статью относительно мало цитируют на фоне других статей авторов. Данная работа находится у них на 10-13 месте по цитирования. Тем не менее есть несколько интересных цитирований:

1. Towards Sparsification of Graph Neural Networks (Hongwu Peng et al. 2022)
2. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness (Tri Dao et al. 2022)

А также прямое продолжение - Monarch: Expressive Structured Matrices for Efficient and Accurate Training (Tri Dao, Beidi Chen et al. 2022)

Конкуренты:
Есть несколько основных статей, в которых исследуется спарсификации моделей машинного обучения другими подходами:

1. Прунинг (Lee et al. 2018, Evci et al. 2020)
2. Теория Лотерейных билетов (Frankle et al. 2018)
3. Хеширование (Chen et al. 2019, Kitaev et al. 2020)

Все они исследуются вопросы динамические паттерны спарсификации, из-за чего обучение модели может быть существенно дольше

Также есть работы, посвященные оптимизации конкретных видов комбинаций слоев на основе спарсификаций

1. Layer Agnostic Sparsity: Most existing work targets a single type of operation such as attention (Child et al. 2019, Zaheer et al. 2020)

Однако во многих реальных приложениях боттлнек по времени вычисления составляют MLP слои (Wu et al. 2020), к которым работа выше не применима

Сильные стороны:

1. Актуальность и значимость, так как сейчас инфренс и обучение моделей это одна из самых больших сложностей в крупных ml/data-oriented компаниях
2. Есть большое число экспериментов с sota моделями из разных областей
3. Сам метод спарсификации не усложняет обучение за счет статического паттерна и является обобщенным, применимым к любой архитектуре моделей

Слабые стороны:

1. В работе практически нет численных сравнений с другими методами спарсификации. Даже с методами, предложенными самими авторами ранее

Дальнейшие исследования:
Авторами рассмотрен по сути простой способ спарсификации и адаптирован к современному оборудованию. Можно пытаться адаптировать более сложные способы.

Применение в индустрии:
Удешевление обучения и инфренса моделей для больших компаний и массово (может быть реализован в клаудах)