АППАРАТНОЕ УСКОРЕНИЕ НЕЙРОННЫХ СЕТЕЙ

Русакович А. В.

Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь

Перцев Д.Ю., к.т.н., доцент

В данном материале рассмотрены основные методы ускорения обучения и выполнения нейронных сетей, а также история их развития.

В последнее десятилетие все более актуальной становится проблема обучения и выполнения нейронных сетей, особенно в условиях ограниченных аппаратных ресурсов.

С появлением первых нейронных сетей, таких как многослойный перцептрон [1], проблема ограниченности ресурсов была сдерживающим фактором для увеличения сложности модели и, следовательно, их точности и производительности. В 1980-е годы для обучения и выполнения первых нейронных сетей чаще всего применялся центральный процессор (CPU), подходящий для общих вычислений. Однако в пользу общности жертвовалось производительностью специфичных вычислений, например, векторных. Ситуация поменялась с появлением графических процессоров (GPU), аппаратная архитектура которых нацелена на работу с 3D объектами и одновременной работой с множеством данных. Известным примером является AlexNet [2], являющийся одной из первых сверточных нейронных сетей, пригодной к применению на GPU и позволившей выиграть соревнование по распознанию изображений ImageNet. Это дало толчок к развитию как самих GPU, так и к поиску других аппаратных ускорителей.

Успех сверточных нейронных сетей, накопление большого объема данных создавало запрос на улучшение аппаратных возможностей. Большинство нейронных сетей оперируют сверточными слоями, полносвязными слоями, а также различными функциями активации между ними. Поскольку и сверточные слои, и полносвязные слои можно представить как операцию матричного или тензорного умножения, одно из направлений ускорения была работа именно над оптимизацией этой операции. Так, в 2018 году NVIDIA презентовала новое поколение графических ускорителей серии RTX 2xxx, особенностью которой являлось наличие так называемых тензорных ядер [3], выполняющих операцию сплавленного сложения и умножения в матричном виде.

Поскольку основной специализацией графического процессора является графика, а не вычисления, большего ускорения можно достичь за счет дальнейшей специализации. Это привело к созданию и использованию тензорных вычислительных модулей (TPU) [4], используемых для обучения, и нейронных процессоров (NPU) для выполнения. Оба типа устройств отличаются характерным высоким соотношением производительности к единице потребляемой энергии, но при этом в своем использовании ограничены только обучением или выполнением модели нейронной сети.

Одно TPU ядро состоит из множества матричных вычислительных блоков, число которых зависит от версии, а также одного векторного и скалярного блоков. Ядра оперируют числами с пониженной точностью, в частности 8 или 16 бит, что позволяет упростить оборудование. Вместе с использованием квантования параметров и вычислений это позволяет значительно ускорить вычисления с небольшой потерей точности. Также, TPU ядра могут масштабироваться и объединяться в кластер для повышения эффективности.

В то же время NPU используются для выполнения нейронных сетей на маленьких устройствах, таких как смартфоны или Edge устройства. Для NPU также характерно использование вычислений с пониженной точностью в 8, 16 бит или даже под-байтовой точностью, а также использование SIMD/векторных блоков.

Таким образом наблюдается явное движение в сторону использования специализированных устройств на базе аппаратных решений для ускорения. В будущем можно ожидать дальнейшую специализацию решений только для определенного класса задач или архитектур. Например, устройств с аппаратным ускорением свертки или обработки звука.

Список использованных источников:

1.Хайкин, С. Нейронные сети. Полный курс: учебное пособие./ С. Хайкин — Диалектика, 2019. — 1104 p.

2.Классификация ImageNet с помощью глубоких сверточных нейронных сетей [Электронный ресурс] — Режим доступа: https://papers.nips.cc/paper_files/paper/2012-Abstract.html — Дата доступа: 22.03.2023.

3.Тензорные ядра NVIDIA [Электронный ресурс] — Режим доступа: https://www.nvidia.com/en-us/data-center/tensor-cores — Дата доступа: 20.03.2023.

4.Тензорные вычислительные модули, используемые для обучения [Электронный ресурс] — Режим доступа: https://cloud.google.com/tpu/docs/system-architecture-tpu-vm — Дата доступа: 21.03.2023.