ВЛИЯНИЕ МЕТОДОВ КВАНТОВАНИЯ И ПРУНИНГА НА ЭНЕРГОЭФФЕКТИВНОСТЬ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ НА УСТРОЙСТВАХ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

Насриддинов Бурхониддин Зухриддин угли
Преподаватель университета “PROFI”, г. Ташкент, Узбекистан
cyberXkiller98@gmail.com

Аннотация: В статье проводится сравнительный анализ двух основных методов снижения энергопотребления при выполнении глубоких нейронных сетей на устройствах с ограниченными вычислительными ресурсами – квантования и прунинга. Цель исследования – обосновать механизм влияния этих методов на энергоэффективность с точки зрения энергии обращений к памяти и арифметических операций, а также предложить последовательность их совместного применения. Использованы методы системного и сравнительного анализа, а также обобщения экспериментальных результатов, опубликованных в литературе. Показано, что основным источником энергозатрат являются не арифметические операции, а чтение данных из памяти, поэтому наибольший эффект дают методы, сокращающие объём модели. Сделан вывод, что последовательное применение структурного прунинга и 8-битного целочисленного квантования является наиболее практичным подходом для устройств с ограниченными ресурсами.

Ключевые слова: глубокое обучение, энергоэффективность, квантование, прунинг, сжатие моделей, граничные вычисления, микроконтроллер, нейронная сеть.

Введение. В последние годы модели глубокого обучения широко применяются в распознавании изображений, обработке речи и технической диагностике. Однако большинство таких моделей разработано для

высокопроизводительных серверов, и их запуск на смартфонах, сенсорных узлах и микроконтроллерах затруднён из-за ограниченности памяти, вычислительной мощности и источника питания. Особенно в устройствах с батарейным питанием энергопотребление напрямую определяет срок работы системы. Поэтому оптимизация моделей по энергоэффективности стала актуальной научно-практической задачей.

В Узбекистане внедрение технологий искусственного интеллекта также определено одним из приоритетных направлений государственной политики [9]. Развитие систем сенсорного мониторинга в промышленности, сельском хозяйстве и коммунальной сфере повышает потребность в эффективных моделях, работающих локально, то есть непосредственно на граничных устройствах.

Проблема сжатия моделей широко исследована в зарубежной литературе. С. Хань и соавторы, объединив прунинг, квантование и кодирование Хаффмана, показали возможность уменьшения объёма модели в десятки раз [1]. Б. Джейкоб и соавторы предложили схему квантования, основанную исключительно на целочисленной арифметике [3]. В. Сзе и соавторы систематизировали подходы к эффективной обработке глубоких нейронных сетей на аппаратном и алгоритмическом уровнях [4]. Вместе с тем влияние этих методов на энергопотребление часто оценивается через косвенные показатели – число параметров или операций. Вопрос же обоснования последовательности применения методов на основе энергетической модели освещён недостаточно. Именно это определяет научную проблему исследования.

Цель исследования – обосновать механизм влияния методов квантования и прунинга на энергоэффективность и предложить последовательность их совместного применения на устройствах с ограниченными ресурсами. Для достижения цели поставлены следующие задачи. Во-первых, определить структуру энергопотребления при работе нейронной сети. Во-вторых, провести

сравнительный анализ видов квантования и прунинга. В-третьих, обосновать практическую последовательность их совместного применения.

Материалы и методы. Методологическую основу исследования составляют несколько взаимодополняющих подходов. Метод системного анализа позволяет рассматривать работу нейронной сети как совокупность взаимосвязанных элементов – чтения весов из памяти, хранения промежуточных результатов и выполнения арифметических операций.

Для оценки энергопотребления использованы данные об энергии элементарных операций, приведённые М. Хоровицем для технологии 45 нм [2]. Согласно этим данным, чтение 32-битного слова из внешней памяти DRAM требует около 640 пДж, 32-битное умножение с плавающей точкой – 3,7 пДж, а сложение 8-битных целых чисел – 0,03 пДж.

Сравнительный метод применён для сопоставления видов квантования (после обучения и с учётом квантования при обучении) и прунинга (неструктурного и структурного) по критериям энергопотребления, точности и аппаратной совместимости [1; 3; 6; 7; 8]. Метод обобщения результатов послужил для сведения опубликованных в литературе экспериментальных показателей в единую рамку анализа.

Результаты. *Структура энергопотребления.* Согласно данным Хоровица, чтение одного слова из памяти DRAM требует примерно в 170 раз больше энергии, чем умножение с плавающей точкой [2]. Отсюда следует важный вывод: на устройстве с ограниченными ресурсами основная часть энергии расходуется не на вычисления, а на перемещение весов и промежуточных результатов по иерархии памяти [4]. Следовательно, чем компактнее веса модели и чем полнее они помещаются во внутреннюю память SRAM устройства, тем больше экономия энергии.

Квантование. Квантование переводит веса и активации из 32-битного формата с плавающей точкой в формат меньшей разрядности, как правило, в 8-битный целочисленный. Это экономит энергию двумя путями. Во-первых, объём модели уменьшается в 4 раза, и энергия обращений к памяти соответственно сокращается. Во-вторых, целочисленные операции значительно дешевле операций с плавающей точкой [2; 3]. Квантование после обучения не требует дополнительного обучения, но на компактных архитектурах может сильнее снижать точность. Учёт квантования в процессе обучения уменьшает потери точности [3; 5; 7].

Прунинг. Прунинг удаляет веса или фильтры с низкой значимостью. С. Хань и соавторы, объединив прунинг и квантование, уменьшили объём модели AlexNet в 35 раз, а модели VGG-16 – в 49 раз без потери точности [1]. Однако разреженные матрицы, возникающие при неструктурном прунинге, на универсальных процессорах и микроконтроллерах без специальной поддержки не дают полного выигрыша в скорости или энергии. Структурный прунинг удаляет целые фильтры и формирует меньшую сеть, пригодную для обычных плотных вычислений [8]. Т.-Дж. Ян и соавторы, непосредственно учитывая энергетическую модель в процессе прунинга, снизили энергопотребление AlexNet в 3,7 раза, а GoogLeNet – в 1,6 раза при потере точности top-5 не более 1% [6].

Последовательность совместного применения. На основе анализа для устройств с ограниченными ресурсами предлагается следующая последовательность: 1) обучение базовой модели и измерение её энергопотребления на целевом устройстве; 2) энергоориентированный структурный прунинг – удаление фильтров начиная со слоёв с наибольшим энергопотреблением; 3) дообучение для восстановления точности; 4) 8-битное

целочисленное квантование с учётом квантования при обучении; 5) повторное измерение энергии, задержки и точности на целевом устройстве.

Важно, чтобы прунинг выполнялся до квантования: сначала упрощается структура сети, затем квантуются оставшиеся веса. При обратном порядке ошибка квантования может нарушить критерии прунинга.

Обсуждение. Сопоставление полученных результатов с подходами в литературе приводит к ряду выводов. Во многих исследованиях эффективность модели оценивается числом параметров или операций. Однако число операций не отражает энергопотребление в полной мере, поскольку слои с одинаковым числом операций могут резко различаться по числу обращений к памяти [4; 6]. Поэтому результат оптимизации следует оценивать по энергии, непосредственно измеренной на целевом устройстве.

Научная новизна исследования состоит в рассмотрении квантования и прунинга не как отдельных методов, а как единой последовательности, основанной на энергетической модели. Основным критерием при этом выступает не число параметров, а энергия обращений к памяти.

Вместе с тем необходимо учитывать ряд ограничений. Квантование до 4 бит и ниже теоретически даёт большую экономию, однако на многих микроконтроллерах такие форматы не поддерживаются на аппаратном уровне [7]. Кроме того, данные Хоровица относятся к определённому технологическому процессу, и точные значения на других устройствах отличаются. Тем не менее закономерность, согласно которой обращение к памяти значительно дороже вычислений, сохраняется как общая [2; 4].

Заключение. Проведённое исследование позволяет сделать ряд выводов.

Во-первых, на устройствах с ограниченными ресурсами основным источником энергопотребления моделей глубокого обучения являются

обращения к памяти. Поэтому методы, сокращающие объём модели, имеют решающее значение для экономии энергии.

Во-вторых, 8-битное целочисленное квантование уменьшает объём модели в 4 раза и поддерживается на аппаратном уровне большинством устройств. Это делает его наиболее практичным методом.

В-третьих, для устройств общего назначения структурный прунинг предпочтительнее неструктурного, поскольку он даёт реальную экономию без специального оборудования.

В-четвёртых, последовательность из структурного прунинга, дообучения и квантования должна применяться на основе энергетической модели. Дальнейшие исследования могут быть направлены на экспериментальную проверку этой последовательности на реальных микроконтроллерных платформах.

Литература

1. Han S., Mao H., Dally W. J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // International Conference on Learning Representations (ICLR). 2016. arXiv:1510.00149.
2. Horowitz M. Computing's Energy Problem (and What We Can Do About It) // IEEE International Solid-State Circuits Conference (ISSCC). Digest of Technical Papers. 2014. P. 10–14.
3. Jacob B., Kligys S., Chen B. et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2704–2713.
4. Sze V., Chen Y.-H., Yang T.-J., Emer J. S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey // Proceedings of the IEEE. 2017. Vol. 105, No. 12. P. 2295–2329.

5. Krishnamoorthi R. Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper. 2018. arXiv:1806.08342.
6. Yang T.-J., Chen Y.-H., Sze V. Designing Energy-Efficient Convolutional Neural Networks Using Energy-Aware Pruning // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 5687–5695.
7. Gholami A., Kim S., Dong Z. et al. A Survey of Quantization Methods for Efficient Neural Network Inference. 2021. arXiv:2103.13630.
8. Li H., Kadav A., Durdanovic I., Samet H., Graf H. P. Pruning Filters for Efficient ConvNets // International Conference on Learning Representations (ICLR). 2017. arXiv:1608.08710.
9. Постановление Президента Республики Узбекистан «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта» от 17 февраля 2021 г. № ПП-4996 // Национальная база данных законодательства Республики Узбекистан. URL: https://lex.uz.