МЕТОДИКА ОЦЕНКИ ЭНЕРГОПОТРЕБЛЕНИЯ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ НА ГРАНИЧНЫХ УСТРОЙСТВАХ И ЭНЕРГООРИЕНТИРОВАННОЕ ПРОЕКТИРОВАНИЕ

Насриддинов Бурхониддин Зухриддин угли

Преподаватель университета “PROFI”, г. Ташкент, Узбекистан

cyberXkiller98@gmail.com

Аннотация: В статье анализируются методика оценки энергопотребления моделей глубокого обучения на граничных устройствах с ограниченными вычислительными ресурсами и подходы к энерго-ориентированному проектированию. Цель исследования – обосновать ограниченность косвенных показателей, таких как число параметров и операций, предложить методику оценки, основанную на измерении энергопотребления на реальном устройстве, а также определить роль эффективных архитектур, дистилляции знаний и аппаратно-ориентированного поиска нейронных архитектур. Использованы методы системного и сравнительного анализа, а также моделирования. Предложена система показателей, совместно учитывающая энергию одного вывода, пиковый объём оперативной памяти, задержку и точность. Сделан вывод, что энергоэффективность должна закладываться как ограничение с самого начала проектирования, а не на завершающем этапе.

Ключевые слова: глубокое обучение, энергопотребление, граничные вычисления, TinyML, дистилляция знаний, поиск нейронных архитектур, методика оценки, микроконтроллер.

Введение. С широким распространением систем искусственного интеллекта их энергопотребление превратилось в самостоятельную научную проблему. Э. Струбелл и соавторы подсчитали, что обучение крупной языковой модели с поиском нейронной архитектуры требует энергии, эквивалентной

выбросу около 284 тонн углекислого газа [1]. Однако проблема энергопотребления характерна не только для крупных центров обработки данных. В сенсорных узлах, носимых устройствах и промышленных контроллерах, работающих от батареи или солнечной панели, каждая операция вывода (inference) сокращает срок автономной работы системы.

В Узбекистане развитие цифровой экономики и цифровизация промышленности, сельского хозяйства и коммунальной сферы определены приоритетными задачами государственных программ [10]. В этих направлениях растёт потребность в энергоэффективных моделях, обрабатывающих данные непосредственно на устройстве без передачи в облако.

В литературе сформировалось несколько направлений создания эффективных моделей. А. Ховард и соавторы предложили архитектуру MobileNets, основанную на свёртках, разделяемых по глубине [4]. Дж. Хинтон и соавторы обосновали метод дистилляции, переносящий знания большой модели в малую [5]. М. Тан и соавторы разработали поиск архитектуры, учитывающий задержку на реальном устройстве [9]. Вместе с тем во многих исследованиях эффективность по-прежнему оценивается числом параметров или операций, а эти показатели могут неверно отражать энергопотребление [2; 3]. Вопрос оценки энергопотребления по единой методике и включения его в процесс проектирования в качестве ограничения освещён недостаточно. Именно это определяет научную проблему исследования.

Цель исследования – предложить методику оценки энергопотребления моделей глубокого обучения на граничных устройствах и обосновать место методов энергоориентированного проектирования. Для достижения цели поставлены следующие задачи. Во-первых, проанализировать ограниченность косвенных показателей эффективности. Во-вторых, разработать методику оценки, основанную на измерениях на реальном устройстве. В-третьих, провести

сравнительный анализ эффективных архитектур, дистилляции знаний и поиска архитектур с точки зрения энергетического ограничения.

Материалы и методы. Методологическую основу исследования составляют несколько взаимодополняющих подходов. Метод системного анализа позволяет рассматривать модель на граничном устройстве как единую систему, состоящую из аппаратной платформы, программного обеспечения и нейронной сети. При этом учитывается, что энергопотребление зависит не только от архитектуры модели, но и от процессора, иерархии памяти и программной библиотеки [2].

Сравнительный метод применён для сопоставления эффективных архитектур, дистилляции знаний и аппаратно-ориентированного поиска архитектур по критериям энергопотребления, точности и требований к памяти [4; 5; 6; 7; 9]. При разработке методики оценки использованы принципы измерения энергии международного бенчмарка MLPerf Tiny [8]. Метод моделирования послужил для формирования выражения энергии одного вывода и математической постановки задачи оптимизации.

Результаты. *Ограниченность косвенных показателей.* Число параметров и операций удобно для быстрого сравнения моделей, но не отражает энергопотребление в полной мере. Т.-Дж. Ян и соавторы показали, что модель SqueezeNet, значительно более компактная по числу весов, потребляет больше энергии, чем AlexNet [3]. Причина в том, что энергия в основном расходуется на перемещение данных по иерархии памяти, объём которого зависит от типа слоя и размера промежуточных результатов (активаций) [2]. Следовательно, модель с меньшим числом параметров не всегда потребляет меньше энергии.

Методика оценки. В качестве основного показателя оценки предлагается энергия одного вывода на целевом устройстве. Она вычисляется за вычетом мощности устройства в режиме покоя:

$E_{inf} = \int_{t_0}^{t_1} (P(t) - P_{idle})\, dt$

где $P(t)$ – мощность, измеренная во время вывода; P_{idle} – мощность в режиме покоя; t_0 и t_1 – моменты начала и окончания вывода. Мощность измеряется внешним анализатором мощности или с помощью шунтирующего резистора, а результат принимается как среднее значение многократно повторённых измерений. Такой подход соответствует принципам энергетического режима бенчмарка MLPerf Tiny [8].

Помимо энергии, при оценке модели необходимо учитывать четыре дополнительных показателя: пиковое потребление оперативной памяти (SRAM), объём постоянной памяти (Flash), задержку и точность. Поскольку оперативная память микроконтроллеров обычно ограничена несколькими сотнями килобайт, пиковое потребление памяти часто становится решающим ограничением [7]. В результате задача проектирования ставится как задача условной оптимизации:

$max_m\ Acc(m)$ при $E_{inf}(m) \le E_{max}$, $M_{peak}(m) \le M_{SRAM}$, $T(m) \le T_{max}$

где m – модель, Acc – точность, M_{peak} – пиковое потребление памяти, M_{SRAM} – объём оперативной памяти устройства, T – задержка.

Методы энергоориентированного проектирования. Первое направление – эффективные архитектуры. Свёртки, разделяемые по глубине, в архитектуре MobileNets сокращают объём вычислений для ядер размером 3×3 в 8–9 раз по сравнению со стандартной свёрткой [4]. Семейство EfficientNet достигает высокой точности при малых ресурсах за счёт сбалансированного масштабирования глубины, ширины сети и размера входного изображения [6].

Второе направление – дистилляция знаний. В этом случае распределение вероятностей на выходе большой модели-«учителя» используется как дополнительный сигнал при обучении малой модели-«ученика» [5]. В результате малая модель достигает более высокой точности относительно своего размера, а потери точности в рамках энергетического ограничения сокращаются.

Третье направление – аппаратно-ориентированный поиск нейронных архитектур. В методе MnasNet в процесс поиска включена задержка, непосредственно измеренная на мобильном телефоне, в результате чего получена модель, в 1,8 раза более быстрая и на 0,5% более точная, чем MobileNetV2 [9]. Система MCUNet за счёт совместного проектирования поиска архитектуры и движка вывода достигла точности top-1 выше 70% на наборе ImageNet на коммерческом микроконтроллере [7].

Обсуждение. Сопоставление полученных результатов с подходами в литературе приводит к ряду выводов. Подход эффективных архитектур экономит энергию косвенно – за счёт сокращения числа операций [4; 6]. Подход поиска архитектур напрямую учитывает показатели на реальном устройстве, однако чаще оперирует задержкой, а не энергией [9]. Хотя задержка и энергия взаимосвязаны, они не тождественны: при изменении тактовой частоты процессора и типа памяти меняется и их соотношение [2].

Научная новизна исследования заключается в рассмотрении энергии не как показателя, проверяемого на завершающем этапе проектирования, а как ограничения, вводимого в задачу оптимизации с самого начала. Предложенная система показателей позволяет сравнивать методы эффективных архитектур, дистилляции и поиска архитектур по единому критерию.

Вместе с тем необходимо учитывать ряд ограничений. Измерение энергии на реальном устройстве требует дополнительного оборудования и времени, поэтому непосредственное измерение тысяч вариантов при поиске архитектуры непрактично. Для решения этой проблемы целесообразно использовать модель прогнозирования энергии, построенную на основе ограниченного числа измерений [3; 7].

Заключение. Проведённое исследование позволяет сделать ряд выводов.

Во-первых, число параметров и операций не является надёжным показателем энергопотребления. Энергоэффективность модели должна измеряться непосредственно на целевом устройстве.

Во-вторых, система показателей, включающая энергию одного вывода, пиковое потребление оперативной памяти, объём постоянной памяти, задержку и точность, позволяет всесторонне оценивать модели для граничных устройств.

В-третьих, эффективные архитектуры, дистилляция знаний и аппаратно-ориентированный поиск архитектур не исключают, а дополняют друг друга. Наибольший эффект даёт их совместное применение в условиях энергетического ограничения.

В-четвёртых, энергоэффективность должна закладываться как ограничение с самого начала проектирования. Дальнейшие исследования могут быть направлены на разработку модели прогнозирования энергии для задач мониторинга в местной промышленности и сельском хозяйстве и её проверку на реальных микроконтроллерных платформах.

Литература

1. Strubell E., Ganesh A., McCallum A. Energy and Policy Considerations for Deep Learning in NLP // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). 2019. P. 3645–3650.
2. 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.
3. 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.
4. Howard A. G., Zhu M., Chen B. et al. MobileNets: Efficient Convolutional Neural

Networks for Mobile Vision Applications. 2017. arXiv:1704.04861.

5. Hinton G., Vinyals O., Dean J. Distilling the Knowledge in a Neural Network. 2015. arXiv:1503.02531.
6. Tan M., Le Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR 97. 2019. P. 6105–6114.
7. Lin J., Chen W.-M., Lin Y., Cohn J., Gan C., Han S. MCUNet: Tiny Deep Learning on IoT Devices // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33.
8. Banbury C., Reddi V. J., Torelli P. et al. MLPerf Tiny Benchmark // NeurIPS Datasets and Benchmarks Track. 2021. arXiv:2106.07597.
9. Tan M., Chen B., Pang R. et al. MnasNet: Platform-Aware Neural Architecture Search for Mobile // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 2820–2828.
10. Указ Президента Республики Узбекистан «Об утверждении Стратегии «Цифровой Узбекистан – 2030» и мерах по её эффективной реализации» от 5 октября 2020 г. № УП-6079 // Национальная база данных законодательства Республики Узбекистан. URL: https://lex.uz.