

ОБЗОР СОВРЕМЕННЫХ МЕТОДОВ ЭНЕРГОЭФФЕКТИВНОГО МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВСТРАИВАЕМЫХ СИСТЕМ ОБРАБОТКИ АКУСТИЧЕСКИХ СИГНАЛОВ

С.А. Лемза¹, А.Ю. Бутов¹

¹ФГБОУ ВО «Воронежский государственный лесотехнический университет
имени Г.Ф. Морозова»

Аннотация. Представлен критический анализ современных методов энергоэффективного машинного обучения для встраиваемых систем обработки акустико-вибрационных сигналов. Рассмотрены аппаратно-алгоритмические архитектуры на базе 32-разрядных микроконтроллеров ARM Cortex-M (STM32U5, STM32H7, nRF52840), методы квантования и прореживания нейронных сетей, программные средства TensorFlow Lite for Microcontrollers и CMSIS-NN. Проанализированы практические реализации, демонстрирующие энергопотребление 50-150 мДж на цикл вычисления при точности классификации >95% на наборах данных ESC-10/ESC-50. Выявлены ограничения текущих подходов и определены перспективные направления исследований в области нейроморфных вычислений и распределённого обучения на периферии.

Ключевые слова: миниатюрное машинное обучение, встраиваемые системы, энергоэффективность, квантование нейронных сетей, обработка акустических сигналов, STM32U5, TensorFlow Lite, CMSIS-NN

REVIEW OF MODERN METHODS OF ENERGY-EFFICIENT MACHINE LEARNING FOR EMBEDDED ACOUSTIC SIGNAL PROCESSING SYSTEMS

S.A. Lemza¹, A.Yu. Butov¹

¹Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. A critical analysis of modern methods of energy-efficient machine learning for embedded acoustic-vibration signal processing systems is presented. Hardware-algorithmic architectures based on 32-bit ARM Cortex-M microcontrollers (STM32U5, STM32H7, nRF52840), neural network quantization and pruning methods, and software tools TensorFlow Lite for Microcontrollers

and CMSIS-NN are examined. Practical implementations demonstrating energy consumption of 50-150 mJ per inference cycle with classification accuracy >95% on ESC-10/ESC-50 datasets are analyzed. Limitations of current approaches are identified and promising research directions in the field of neuromorphic computing and edge distributed learning are defined.

Keywords: tiny machine learning, embedded systems, energy efficiency, neural network quantization, acoustic signal processing, STM32U5, TensorFlow Lite, CMSIS-NN

Введение

Лавинообразный рост числа устройств интернета вещей заставил индустрию пересмотреть саму архитектуру машинного обучения. Облака больше не справляются – вычисления спускаются на периферию. По прогнозам ABI Research, к 2030 году нас ожидает 2,3 миллиарда устройств со встроенным искусственным интеллектом, а рынок софта для периферийных вычислений уже в 2025 году превысит 40 миллиардов долларов. Встраиваемые системы распознавания акустических сигналов – будь то промышленная диагностика, медицинский мониторинг или голосовые интерфейсы – оказались перед жёсткой дилеммой. С одной стороны, глубокие нейросети требуют огромных вычислительных ресурсов. С другой – автономные устройства живут в жёстких энергетических рамках [1].

Парадигма TinyML (микро-машинное обучение) работает с моделями размером менее мегабайта на микроконтроллерах, у которых оперативной памяти – меньше двух мегабайт. Это заставляет радикально пересматривать классические архитектуры нейросетей. Данный обзор систематизирует современные аппаратные и алгоритмические решения, которые позволяют снизить энергопотребление на порядки, сохранив при этом приемлемую точность распознавания.

Аппаратная база: где рождается энергоэффективность

Эталонной платформой для приложений TinyML сегодня можно считать микроконтроллер STM32U5 (STMicroelectronics). В его основе – ядро ARM Cortex-M33 с тактовой частотой 160 МГц (240 DMIPS, 651 балл CoreMark). Иерархическая кэш-память (16 КБ для команд, 8 КБ для данных) и до 4 МБ флеш-памяти с коррекцией ошибок и ускорителем чтения (нулевые циклы ожидания). ОЗУ – до 2,5 МБ. Ядро умеет работать с цифровой обработкой сигналов, плюс опционально доступен сопроцессор с плавающей запятой – обе опции критичны для эффективного выполнения свёрток.

Но главное – энергопотребление. В активном режиме с динамическим масштабированием напряжения – 16 мкА на мегагерц. В режиме остановки с сохранением 16 КБ памяти – 1,7 мкА. В режиме ожидания – 300 нА. В режиме отключения – 110 нА. Технологический переход на 40-нанометровый процесс дал двойной выигрыш в энергоэффективности по сравнению с предыдущим поколением STM32L4, а производительность выросла втрое. Для сравнения: STM32H7 (480 МГц Cortex-M7, 1 МБ памяти) и Nordic nRF52840 (64 МГц Cortex-M4F со встроенным Bluetooth 5.0) предлагают другие компромиссы между скоростью, энергопотреблением и удобством беспроводной интеграции.

Программная кухня: как выжать максимум из железа

Фреймворк CMSIS-NN от ARM предоставляет оптимизированные ядра для свёрточных, полносвязных слоёв и функций активации. Он активно использует векторные инструкции цифровой обработки сигналов. Техника «im2col» (преобразование входных данных в матричную форму) сокращает количество обращений к памяти за счёт предварительного развёртывания входной информации. А формат «channels-last» (каналы в конце) обеспечивает последовательный доступ при поканальном квантовании. Эксперименты показывают: CMSIS-NN ускоряет выполнение в 4,6 раза и повышает пропускную способность, а энергоэффективность растёт в 4,9 раза по сравнению с базовыми реализациями на том же оборудовании.

Квантование и прореживание: уменьшаем модель, не теряя смысл

8-битное целочисленное квантование (симметричное или асимметричное масштабирование) сокращает размер модели в 4 раза, снижает вычислительную сложность, а точность падает незначительно – менее чем на 2% для большинства архитектур. Смешанное 4-битное квантование даёт ещё больше: экономия памяти 50%, сокращение вычислений на 75%, потеря точности – менее 5%. Это подтверждается экспериментами на наборах ImageNet и ESC-50.

Прореживание (pruning) нейронных сетей дополняет квантование. Структурное прореживание удаляет целые фильтры или каналы – вычисления остаются регулярными, что удобно для векторных процессоров. Неструктурное прореживание обнуляет отдельные веса, но требует разреженного представления данных. Комбинированные методы обучения с учётом квантования (quantization-aware training) моделируют эффекты квантования прямо на этапе обучения, минимизируя ошибку аппроксимации. Квантование после обучения

(post-training quantization) просто калибруется на репрезентативном наборе данных без дообучения [2]. Отдельная техника – дистилляция знаний (knowledge distillation) – переносит знания из большой модели-учителя в компактную модель-ученика. Так можно сохранить более 90% точности при 10-кратном сжатии.

Методологические основы программных средств и энергетических вычислений

TensorFlow Lite для микроконтроллеров предоставляет исполняющую среду размером ~20КБ, поддерживающую подмножество из ~100 операций TensorFlow через систему делегатов. Статическое выделение памяти исключает динамические запросы кучи, критичные для детерминированной задержки в реальном времени. Интеграция с CMSIS-NN осуществляется через делегирование ядер, перенаправляя сверточные операции на оптимизированные примитивы. Конвейер преобразования из TensorFlow в оптимизированный формат с квантованием обеспечивает размер модели 50-500КБ для типичных сверточных архитектур акустической классификации. Облачная платформа Edge Impulse для сквозного цикла машинного обучения объединила более 400000 проектов от 250000 разработчиков (по состоянию на октябрь 2025), автоматизируя выделение признаков (мел-частотные коэффициенты, спектрограммы), обучение моделей и развёртывание на целевые платформы через генерацию комплектов разработки. PyTorch Mobile и Apache TVM обеспечивают альтернативные пути от исследовательских сред к встраиваемому развёртыванию, однако уступают TensorFlow Lite для микроконтроллеров в размере и зрелости инструментария для сверхмаломощного сегмента.

Энергопотребление цикла вычислений определяется интегралом мгновенной мощности за время выполнения операции:

$$E_{\text{total}} = \int_0^{T_{\text{inf}}} (P_{\text{dyn}}(t) + P_{\text{stat}}(t)) dt$$

где P_{dyn} – мгновенная динамическая мощность, P_{stat} – статическая мощность утечки.

Основной вклад вносит динамическая составляющая:

$$P_{\text{dyn}} = \alpha \cdot C \cdot V^2 \cdot f$$

где α – коэффициент переключения, а V – напряжение питания.

Замеры на STM32U5 дают наглядную картину. Свёрточная сеть с примерно 100 тысячами параметров (5–10 свёрточных слоёв) на частоте 80–160 МГц

потребляет от 50 до 150 миллиджоулей за один проход вычислений. Цифры внушительные, но дальше начинается искусство оптимизации.

Динамическое масштабирование напряжения и частоты работает по простому принципу: снижаешь напряжение – энергопотребление падает квадратично, а производительность – линейно. Для приложений, которые не требуют мгновенного отклика, это идеальный баланс. А если добавить циклическое пробуждение – когда устройство просыпается только на время обработки акустического события, а остальное время дремлет в энергосберегающих режимах – цифры становятся ещё красноречивее. Обнаружение голосовой активности с порогом -40 дБFS позволяет системе 95% времени находиться в режиме остановки. Среднее энергопотребление в таком режиме падает ниже 1 милливатта.

Здесь выясняется неожиданный факт: вычисления сами по себе не так страшны, как доступ к памяти. Чтение 32-битного слова обходится в:

- из статической памяти (SRAM) – примерно 1 пикоджоуль;
- из флэш-памяти – уже 10 пикоджоулей;
- из динамической памяти (DRAM) – около 100 пикоджоулей.

Разница на два порядка объясняет, почему вычисления «на кристалле» (когда все данные уже в кэше или SRAM) оказываются настолько эффективнее, чем постоянные подгрузки из внешней памяти. Вывод прост: чтобы сделать нейросеть по-настоящему энергоэффективной, нужно минимизировать дальние прогулки данных по памяти.

Критический анализ современного состояния области

В 2022 году на страницах журнала Pattern Recognition появилась архитектура ACDNet, которая задала новую планку для классификации звуков на периферийных устройствах. Цифры впечатляют: точность 96,65% на наборе ESC-10, 87,10% на ESC-50 и 84,45% на UrbanSound8K. При этом модель весит всего 0,50 мегабайта, а вычислительная сложность составляет 14,82 миллиона операций с плавающей запятой. Что же такого особенного придумали авторы? Во-первых, разделимые по глубине свёртки – они резко сокращают число параметров. Во-вторых, механизмы временного внимания, которые выхватывают из потока данных самые информативные сегменты. И в-третьих – обработка необработанного сигнала без классического кратковременного преобразования Фурье. Это настоящий прорыв, потому что STFT обычно пожирает ресурсы.

Когда ACDNet развернули на микроконтроллере STM32H7, получили задержку всего 47 миллисекунд, а энергопотребление на один проход вычислений составило около 80 миллиджоулей.

Есть и другой подход – импульсные нейронные сети (SNN). Они работают по событийно-управляемому принципу: вычисления происходят только тогда, когда поступает входной импульс. В режиме тишины сеть практически не потребляет энергии. Это открывает интересные перспективы для акустических систем с низким рабочим циклом, но пока импульсные сети уступают традиционным по точности на сложных наборах данных.

Динамика изменения мембранного потенциала $v(t)$ описывается уравнением:

$$\tau_m \frac{dv(t)}{dt} = -(v(t) - v_{\text{rest}}) + R_m I(t)$$

где $v(t)$ – мембранный потенциал, τ_m – постоянная времени мембраны, v_{rest} – потенциал покоя, $I(t)$ – входной ток (сигнал).

Активация (спайк) происходит только при превышении порогового значения, что теоретически обеспечивает порядковое снижение энергопотребления по сравнению с классическими архитектурами. Нейроморфная система Intel Hala Point (20 петаопераций в секунду, >15 триллионов 8-битных операций на ватт) демонстрирует 100-кратное преимущество в энергоэффективности над графическими процессорами для разреженных, событийно-управляемых задач, однако отсутствие зрелых алгоритмов обучения импульсных сетей ограничивает практическое применение.

Текущие ограничения включают недостаточность вычислительных ресурсов для моделей на основе трансформеров, так как BERT и GPT требуют более 100 МБ памяти, а также отсутствие методологий совместного проектирования аппаратного и программного обеспечения для оптимизации специфичных приложений. К важным факторам относятся проблема смещения предметной области при развёртывании моделей, обученных на лабораторных данных, в реальные акустические условия с шумом и реверберацией, и фрагментация инструментальных средств, требующая специальной экспертизы в оптимизации под конкретные семейства микроконтроллеров. Энергетические затраты периферийного оборудования, при которых микрофоны потребляют 100–500 мкА, а аналого-цифровые преобразователи – 50–200 мкА, часто доминируют в бюджете системы, что требует комплексного управления энергопотреблением.

Заключение и перспективные направления исследований

Современные аппаратно-алгоритмические решения обеспечивают практическую реализацию глубоких нейронных сетей на микроконтроллерах класса ARM Cortex-M при энергопотреблении <1 мВт среднего уровня. Комбинация квантования, прореживания, архитектурного поиска и аппаратно-оптимизированных библиотек позволяет достигать $>90\%$ точности распознавания акустических образов при размере модели <1 МБ. Перспективные направления включают нейроморфные архитектуры с биологически-вдохновленными механизмами обучения для импульсных сетей, распределённое обучение на периферии без централизации приватных данных, автоматический поиск нейронных архитектур с многокритериальной оптимизацией компромиссов точность-задержка-энергопотребление, совместное проектирование аппаратного и программного обеспечения с настраиваемыми ускорителями для критичных операций, трансферное обучение и адаптацию с малым количеством примеров для минимизации требований к данным в специализированных областях, а также интеграцию механизмов внимания и адаптивных вычислений для динамического масштабирования сложности вычислений в зависимости от характеристик входного сигнала [3]. Развитие инструментальных средств автоматизированной оптимизации (квантование на уровне компилятора, автоматическая настройка) критично для демократизации миниатюрного машинного обучения и снижения барьера входа для прикладных разработчиков.

Список литературы

1. Исследование способов защиты информации от утечки по акустическому и виброакустическому каналам / В. К. Зольников, С. А. Сазонова, А. И. Заревич, С. С. Башун // Моделирование систем и процессов. – 2025. – Т. 18, № 2. – С. 27-40. – DOI 10.12737/2219-0767-2025-18-2-27-40. – EDN GBWORY.
2. Применение нейронных сетей для оптимизации энергопотребления СБИС / А. В. Ачкасов, А. С. Ягодкин, Ф. В. Макаренко, Н. Ю. Заленская // Моделирование систем и процессов. – 2025. – Т. 18, № 1. – С. 7-16. – DOI 10.12737/2219-0767-2025-7-16. – EDN MQXMNM.
3. Сазонова, С. А. Исследование и разработка моделей проектирования микросхем цифровой обработки информации / С. А. Сазонова, К. В. Зольников, В. Ф. Асминин // Моделирование систем и процессов. – 2025. – Т. 18, № 3. – С. 89-99. – DOI 10.12737/2219-0767-2025-18-3-89-99. – EDN LGIGCI.

References

1. Research on methods of information protection from leakage through acoustic and vibroacoustic channels / V. K. Zolnikov, S. A. Sazonova, A. I. Zarevich, S. S. Bashun // Modeling of Systems and Processes. – 2025. – Vol. 18, No. 2. – P. 27-40. – DOI 10.12737/2219-0767-2025-18-2-27-40. – EDN GBWORY.
2. Application of neural networks for optimization of VLSI energy consumption / A. V. Achkasov, A. S. Yagodkin, F. V. Makarenko, N. Yu. Zalenskaya // Modeling of Systems and Processes. – 2025. – Vol. 18, No. 1. – P. 7-16. – DOI 10.12737/2219-0767-2025-7-16. – EDN MQXMHM.
3. Sazonova, S. A. Research and development of design models for digital information processing microchips / S. A. Sazonova, K. V. Zolnikov, V. F. Asminin // Modeling of Systems and Processes. – 2025. – Vol. 18, No. 3. – P. 89-99. – DOI 10.12737/2219-0767-2025-18-3-89-99. – EDN LGIGCI.