Даньковский Евгений Васильевич,
Магистрант 1-го курса института Информационных систем и инженерно-компьютерных технологий, специальность «Прикладная информатика» (профиль «Системы искусственного интеллекта»), Инженер-электронщик,
АНО ВО «Российский новый университет» ООО «СЗМ»
Dankovsky Evgeny Vasilyevich,
First-year Master's student at the Institute of Information Systems and Engineering and Computer Technologies, majoring in Applied Informatics (major in Artificial ntelligence Systems), Electronics Engineer,
ANO VO "Russian New University" OOO "SZM"

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЗАДАЧАХ ОПТИМИЗАЦИИ ВСТРАИВАЕМЫХ ЭЛЕКТРОННЫХ СИСТЕМ
USING ARTIFICIAL INTELLIGENCE METHODS IN OPTIMIZATION PROBLEMS OF EMBEDDED ELECTRONIC SYSTEMS

Аннотация. В статье рассмотрены методы оптимизации нейронных сетей для встраиваемых электронных систем. Описаны квантование, сокращение параметров, дистилляция знаний и облегчённые архитектуры нейронных сетей. Показано, что данные методы позволяют снизить вычислительную нагрузку и энергопотребление без существенной потери точности моделей.

Abstract. This article examines methods for optimizing neural networks for embedded electronic systems. It describes quantization, parameter pruning, knowledge distillation, and lightweight neural network architectures. The study demonstrates that these methods can significantly reduce computational load and power consumption without substantial loss of model accuracy.

Ключевые слова: Искусственный интеллект, встраиваемые системы, TinyML, Edge AI, микроконтроллеры, обработка данных.

Keywords: Artificial intelligence, embedded systems, TinyML, Edge AI, microcontrollers, data processing.

Введение.

Сегодня технологии искусственного интеллекта активно используются в самых разных сферах жизни – от промышленности и медицины до бытовой электроники [1]. Особое значение приобретают интеллектуальные системы, функционирующие непосредственно на электронных устройствах с ограниченными вычислительными ресурсами. К таким системам относятся устройства интернета вещей, сенсорные платформы, носимая электроника, автономные робототехнические комплексы и интеллектуальные системы мониторинга.

С каждым годом увеличивается количество умных устройств, которые постоянно собирают и обрабатывают данные. Из-за этого возникает необходимость использовать алгоритмы машинного обучения прямо на самих устройствах, без постоянного обращения к облачным сервисам. Использование облачных вычислений не всегда является оптимальным решением, поскольку передача данных сопровождается задержками, увеличением энергопотребления и рисками, связанными с безопасностью информации. В связи с этим особую актуальность приобретает концепция Edge AI, предполагающая выполнение вычислений непосредственно на локальных устройствах.

Однако внедрение искусственного интеллекта во встраиваемые устройства связано с рядом сложностей. Большинство нейронных сетей требуют значительных вычислительных

ресурсов, объёмов памяти и высокой производительности процессоров. Встраиваемые устройства, напротив, обладают ограниченными аппаратными возможностями и работают в условиях низкого энергопотребления. Это делает необходимым использование специальных методов оптимизации моделей искусственного интеллекта.

Целью данной работы является анализ современных методов оптимизации нейронных сетей для применения в интеллектуальных встраиваемых электронных системах.

Актуальность проблемы и практическая значимость

Современная электроэнергетика находится в процессе цифровой трансформации, связанной с внедрением интеллектуальных систем мониторинга, автоматизированного управления и прогнозирования режимов работы энергосетей. Увеличение количества датчиков, интеллектуальных счетчиков и устройств Интернета вещей приводит к постоянному росту объема обрабатываемых данных. Передача всей информации в облачные центры обработки данных вызывает задержки, увеличивает нагрузку на каналы связи и снижает надежность функционирования критически важных энергетических объектов.

Одним из перспективных решений является использование искусственного интеллекта непосредственно на периферийных устройствах (Edge AI). Однако большинство современных нейронных сетей требуют значительных вычислительных ресурсов и объемов памяти, которыми не обладают встроенные контроллеры подстанций, интеллектуальные счетчики и устройства релейной защиты. Поэтому возникает задача оптимизации нейронных сетей для их эффективного применения в электроэнергетике.

Искусственный интеллект во встраиваемых электронных системах.

Встраиваемые электронные системы – это устройства, которые предназначены для выполнения конкретных задач и обычно работают внутри другой техники или оборудования [2]. В отличие от обычных компьютеров, такие устройства имеют небольшие размеры, ограниченный объём памяти и невысокую вычислительную мощность.

В настоящее время технологии искусственного интеллекта активно внедряются во встраиваемые системы различного назначения. Сегодня подобные системы используются в умных камерах, датчиках, системах видеонаблюдения, промышленной автоматизации, медицинском оборудовании и устройствах интернета вещей.

Использование нейронных сетей позволяет существенно повысить функциональные возможности электронных устройств [3]. Например, интеллектуальные камеры способны выполнять распознавание объектов в реальном времени, а системы промышленного мониторинга – прогнозировать возникновение неисправностей оборудования на основе анализа телеметрических данных.

Одним из перспективных направлений является технология TinyML, предполагающая выполнение алгоритмов машинного обучения на микроконтроллерах с минимальным объёмом памяти и низким энергопотреблением [4]. Применение TinyML открывает возможности создания автономных интеллектуальных устройств, работающих без постоянного подключения к облачным сервисам.

Несмотря на значительные преимущества, внедрение искусственного интеллекта во встраиваемые системы сопровождается рядом технических трудностей. Основной проблемой является высокая вычислительная сложность современных нейронных сетей. Большинство моделей глубокого обучения содержат миллионы параметров и требуют значительных вычислительных мощностей.

Ограничения встраиваемых электронных устройств.

Встраиваемые системы существенно отличаются от традиционных вычислительных платформ ограниченностью аппаратных ресурсов [5]. Это создаёт серьёзные сложности при реализации алгоритмов искусственного интеллекта.

Одним из ключевых ограничений является небольшой объём оперативной памяти. Большинство микроконтроллеров обладают объёмом памяти от нескольких сотен килобайт до нескольких мегабайт, тогда как современные нейронные сети могут занимать сотни мегабайт.

Другим важным фактором является ограниченная вычислительная производительность. Встраиваемые процессоры обладают значительно меньшей мощностью по сравнению с графическими ускорителями и центральными процессорами серверного уровня. Это приводит к увеличению времени выполнения операций и затрудняет реализацию сложных моделей глубокого обучения.

Отдельной проблемой является энергопотребление. Многие интеллектуальные устройства функционируют автономно и питаются от аккумуляторов или встроенных источников энергии. Высокая вычислительная нагрузка приводит к быстрому расходу энергии и снижению времени автономной работы устройства.

Кроме того, при передаче данных в облачные сервисы возникают задержки обработки информации. В задачах реального времени, например в системах автономного управления или промышленной автоматизации, даже небольшие задержки могут привести к снижению эффективности работы системы.

Методы оптимизации нейронных сетей.

Одним из наиболее распространённых способов оптимизации является квантование нейронных сетей [6]. Суть этого метода заключается в уменьшении объёма данных, используемых нейронной сетью при вычислениях. Например, вместо использования 32-битных чисел с плавающей точкой могут применяться 8-битные целые числа.

Использование квантования как одного из методов искусственного интелекта и нейронных сетей позволяет значительно сократить объём памяти, необходимый для хранения модели, а также повысить скорость выполнения вычислений. При этом снижение точности модели обычно остаётся незначительным.

Другим эффективным методом является сокращение параметров нейронной сети, также известное как pruning. Данный подход предполагает удаление наименее значимых нейронов и связей между ними. В результате уменьшается размер модели и сокращается количество выполняемых операций.

Также широкое распространение получил метод дистилляции знаний [6]. В рамках данного подхода большая и сложная модель используется для обучения более компактной модели. Компактная нейронная сеть перенимает основные закономерности работы исходной модели, сохраняя приемлемую точность при значительно меньших вычислительных затратах.

Ещё одним важным направлением является использование облегчённых архитектур нейронных сетей [7]. К таким архитектурам относятся MobileNet, SqueezeNet и EfficientNet. Данные модели изначально проектируются с учётом ограниченных вычислительных ресурсов и предназначены для работы на мобильных и встраиваемых устройствах.

Для повышения эффективности работы интеллектуальных систем также применяются аппаратные ускорители искусственного интеллекта. В современных микроконтроллерах и одноплатных компьютерах всё чаще используются специализированные нейронные процессоры, обеспечивающие ускорение операций машинного обучения.

Сравнение MobileNet, SqueezeNet и EfficientNet

Для встроенных систем широко используются специальные архитектуры нейронных сетей, разработанные с учетом ограниченных вычислительных ресурсов.

Таблица 1.

Архитектура	Размер модели	Особенности	Применение
MobileNet	Небольшой	Использует глубинно-разделимые свертки	Мобильные устройства, Edge AI
SqueezeNet	Очень маленький	В 50 раз меньше AlexNet	Микроконтроллеры и TinyML
EfficientNet	Средний	Высокая точность при небольшом числе параметров	Энергетика, компьютерное зрение

MobileNet использует глубинно-разделимые свертки (Depthwise Separable Convolution), что позволяет уменьшить количество операций примерно в 8-9 раз по сравнению с традиционными сверточными сетями.

SqueezeNet применяет специальные Fire-модули и обеспечивает точность, близкую к AlexNet, при размере модели менее 5 МБ.

EfficientNet использует масштабирование глубины, ширины и разрешения сети одновременно, что позволяет получить лучшее соотношение между точностью и вычислительной сложностью.

Для задач электроэнергетики наиболее перспективными являются MobileNet и EfficientNet, поскольку они обеспечивают высокую точность обнаружения неисправностей оборудования при относительно низких требованиях к аппаратным ресурсам.

Преимущества локального выполнения искусственного интеллекта.

Использование локального искусственного интеллекта во встраиваемых системах обладает рядом существенных преимуществ. Одним из главных достоинств является снижение задержек обработки информации. Поскольку вычисления выполняются непосредственно на устройстве, отсутствует необходимость передачи данных на удалённые серверы.

Дополнительным преимуществом является повышение уровня безопасности и конфиденциальности информации. Данные пользователя могут обрабатываться локально без передачи через внешние сети связи, что снижает вероятность утечки информации.

Кроме того, устройства с локальным искусственным интеллектом меньше зависят от качества интернет-соединения. Устройства способны функционировать автономно даже при отсутствии доступа к сети.

Следует отметить и экономическую эффективность локального выполнения вычислений. Использование периферийных вычислений позволяет снизить нагрузку на облачную инфраструктуру и уменьшить расходы на передачу данных.

Развитие технологий Edge AI способствует созданию нового поколения интеллектуальных электронных устройств, способных самостоятельно анализировать информацию и принимать решения в реальном времени [8].

Заключение.

Были рассмотрены основные способы уменьшения вычислительной нагрузки моделей искусственного интеллекта, такие как квантование, уменьшение количества параметров нейронной сети, дистилляция знаний и использование более лёгких архитектур моделей. Использование этих методов позволяет повысить эффективность работы интеллектуальных систем без серьёзной потери точности.

Развитие искусственного интеллекта и увеличение количества устройств интернета вещей делают всё более востребованным применение интеллектуальных алгоритмов во встраиваемых электронных системах. Однако ограниченные вычислительные возможности таких устройств требуют использования специальных методов оптимизации нейронных сетей.

Перспективными направлениями остаются технологии TinyML и Edge AI, которые позволяют запускать алгоритмы машинного обучения непосредственно на устройствах с ограниченными ресурсами. Важную роль также играет разработка специализированных аппаратных ускорителей, способных повысить скорость и эффективность работы искусственного интеллекта во встраиваемых системах.

Список литературы:

1. Бахтеев О.Ю. Машинное обучение и анализ данных. – Москва: ДМК Пресс, 2021. – 312 с.
2. Васильев А.Н. Встраиваемые системы и микроконтроллеры. – Санкт-Петербург: Питер, 2020. – 368 с.
3. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge: MIT Press, 2016. – 800 p.
4. Han S., Mao H., Dally W. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // International Conference on Learning Representations. – 2016.
5. Warden P., Situnayake D. TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. – Sebastopol: O'Reilly Media, 2020. – 490 p.
6. Chollet F. Deep Learning with Python. – New York: Manning Publications, 2021. – 504 p.
7. Li H., Ota K., Dong M. Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing // IEEE Network. – 2018. – Vol. 32. – №1. – P. 96–101.
8. Howard A. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications // Computer Vision and Pattern Recognition. – 2017.