

СОЗДАНИЕ IoT-ПРИЛОЖЕНИЙ С ПОМОЩЬЮ tinyML И МАШИННОГО ОБУЧЕНИЯ

РАДЖЕН БХАТТ (RAJEN BHATT)
ТИНА ШЮАНЬ (TINA SHYUAN)

С помощью крошечных датчиков «Интернет вещей» обеспечивает непрерывный мониторинг окружающей среды и различного оборудования. Достижения в области сенсорных технологий, микроконтроллеров и протоколов передачи данных сделали возможным массовое производство IoT-платформ с самыми разными вариантами подключения и по доступным ценам. Благодаря низкой стоимости IoT-оборудования датчики стали широко внедряться в общественных местах, жилых помещениях, а также в производственном оборудовании.

В рамках IoT датчики в режиме 24/7 отслеживают физические свойства, связанные с окружающей средой, и генерируют огромное количество данных. Например, акселерометры и гироскопы, установленные на вращающихся механизмах, постоянно регистрируют характер вибрации и угловую скорость ротора. Датчики качества воздуха непрерывно контролируют содержание газообразных загрязняющих веществ как внутри помещений, так и на открытом воздухе. Микрофоны в радионянях позволяют непрерывно слышать ребенка. Датчики в «умных» часах постоянно измеряют жизненно важные показатели здоровья. Анало-

гично широкий ряд других датчиков, таких как магнитометры, датчики давления, температуры, влажности, освещенности и т. д., измеряет физические условия везде, где бы они ни были установлены.

Алгоритмы машинного обучения (ML) позволяют обнаруживать в полученных данных интересные закономерности, которые невозможно распознать при ручном анализе и проверке. Конвергенция IoT-устройств и алгоритмов машинного обучения помогает создавать широкий спектр интеллектуальных приложений и расширять возможности пользователей, что стало возможным благодаря снижению энергопотребле-

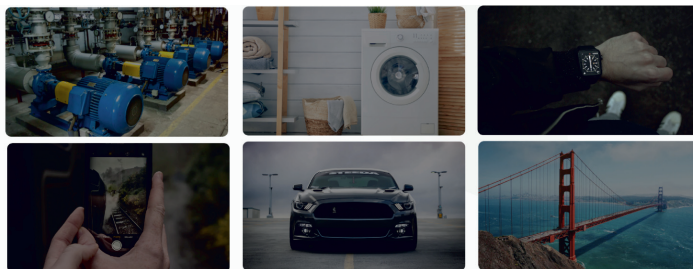
ния устройств, небольшой задержке и простоте алгоритмов, то есть с помощью технологий tinyML. Эта конвергенция коренным образом перестраивает многие отраслевые вертикали (рис. 1), включая, помимо прочего, носимые устройства, «умные» дома, «умные» предприятия («Индустрия 4.0»), автомобилестроение, машинное зрение и другие интеллектуальные потребительские электронные устройства.

tinyML С АВТОМАТИЗИРОВАННЫМ МАШИНЫМ ОБУЧЕНИЕМ

Алгоритмы машинного обучения, реализованные на крошечных микроконтроллерах в устройствах «Интернета вещей», представляют особый интерес из-за множества преимуществ:

- Конфиденциальность и безопасность данных: результаты интеллектуальной обработки данных генерируются непосредственно на встроенных микроконтроллерах. При этом информационные потоки не передаются в облако, а остаются на устройстве в локальной среде, где они конфиденциальны и защищены.
- Экономия энергии: алгоритмы tinyML потребляют гораздо меньше энергии благодаря отсутствию передачи больших данных.
- Низкая задержка и высокая доступность: поскольку алгоритмы обработки данных выполняются локально, задержка

Отрасли, которые может кардинально изменить tinyML



tinyML: низкое энергопотребление, малая задержка, простота, высокая эффективность

Хранение данных и простая аналитика

Визуализация данных

Системы на основе правил

IoT-узлы: датчики + микроконтроллеры + ввод-вывод + подключение

РИС. 1. ▶ tinyML расширяет возможности традиционных устройств «Интернета вещей». Все изображения предоставлены Qeexo

составляет порядка миллисекунд и не зависит от скорости и доступности сети.

Автоматизированное машинное обучение с использованием данных от датчиков включает этапы, показанные на рис. 2. Перед выполнением этих шагов должна быть проведена настройка датчиков и сбор качественных данных, пригодных для машинного обучения. Автоматизированная платформа машинного обучения, такая как Qeexo AutoML, создает простые и эффективные модели машинного обучения для микроконтроллеров класса Arm Cortex-M0-M4 и других совместимых устройств, одновременно управляя всем рабочим процессом.

tinyML НА БАЗЕ ARM Cortex-M0+

Распространение технологий «Интернета вещей» и потребность в крупномасштабном применении датчиков еще больше раздвигают границы архитектур микроконтроллеров и вычислений, необходимых для машинного обучения. Например, благодаря низкому энергопотреблению микроконтроллеры Arm Cortex-M0+, работающие на частоте 48 МГц, широко используются на платах датчиков, предназначенных для устройств «Интернета вещей». Этот микроконтроллер потребляет всего 7 мА на контакт ввода/вывода по сравнению с Cortex-M4, действующим на частоте 64 МГц и потребляющим 15 мА.

Низкое энергопотребление Cortex-M0+ достигается за счет уменьшения объема памяти и упрощения логики. Микроконтроллеры M0+ могут выполнять только 32-битные математические операции с фиксированной запятой, не поддерживают арифметику с насыщением и не имеют возможностей цифрового сигнального процессора. Основанная на этом микроконтроллере платформа Arduino Nano 33 IoT — одна из популярных IoT-платформ — имеет всего 256 кбайт флэш-памяти и 32 кбайт SRAM. В отличие от нее популярный сенсорный модуль с архитектурой Cortex-M4 — Arduino Nano 33 BLE Sense — может выполнять 32-битные операции с плавающей запятой, оснащен возможностями цифрового сигнального процессора и поддерживает арифметику с насыщением, а также имеет в четыре раза больше флэш-памяти и в восемь раз больше SRAM.

Развертывание алгоритмов машинного обучения на M0+ на порядок сложнее по сравнению с развертыванием на M4 из-за трех основных проблем:

- Вычисления с фиксированной запятой: типичное машинное обучение с использованием данных от датчиков включает цифровую обработку сигналов, выделение признаков и формирование логического решения. Для разработки эффективных моделей машинного обучения решающее значение имеет извле-

чение из поступающих с датчиков данных статистических и частотных характеристик (например, быстрое преобразование Фурье). Поток данных от датчиков, представляющие реальные физические явления, по своей природе нестационарны. В целом, чем качественнее информация, извлекаемая из нестационарных сигналов датчиков, тем больше возможностей для разработки эффективных моделей машинного обучения. Выполнение математических операций в представлении с фиксированной запятой при сохранении точности и результативности коммерческого уровня является сложной задачей. Конвейер машинного обучения с фиксированной запятой начинается с данных от датчика и проходит весь путь до вывода модели для последующей генерации выходных данных классификации/регрессии.

- Малый объем памяти: 256 кбайт флэш-памяти и 32 кбайт SRAM накладывают жесткие ограничения на размеры моделей машинного обучения и оперативной памяти, которую могут использовать модели. Реальные задачи машинного обучения часто имеют сложные границы принятия решений/классификации, отличающиеся большим количеством параметров. Использование древовидных ансамблевых моделей для решения таких сложных задач может привести к появлению углубленных деревьев



РИС. 2. ◀ Алгоритм Qeexo AutoML

РИС. 3. ►
Конвейер вывода
Qeexo AutoML M0+



и множества бустеров, влияя как на размер модели, так и на рабочую память. Уменьшение размера модели часто происходит за счет снижения эффективности модели — как правило, это не самый желательный результат.

- Низкая скорость процессора: низкая задержка всегда была ключевым показателем при выборе модели для коммерческого применения. Разница тактовой частоты в 16 МГц, которой мы жертвуем, выбирая архитектуру M0+ с 48 МГц, по сравнению с архитектурой M4 с 64 МГц имеет большое значение, когда дело доходит до измерения задержки на уровне миллисекунд.

Фреймворк AutoML M0+

Qeexo AutoML, разработанный для решения этих проблем, предоставляет конвейер машинного обучения с фиксированной запятой, тщательно оптимизированный для архитектуры Arm Cortex-M0+. Конвейер предусматривает обработку данных датчиков, выделение признаков и вывод с использованием древовидных ансамблевых алгоритмов, таких как Gradient Boosting Machine (GBM), Random Forest (RF) и eXtreme Gradient Boosting (XGBoost). Qeexo AutoML кодирует параметры ансамблевой модели с помощью эффективной структуры данных и использует интерпретации, что приводит к чрезвычайно быстрому расчету

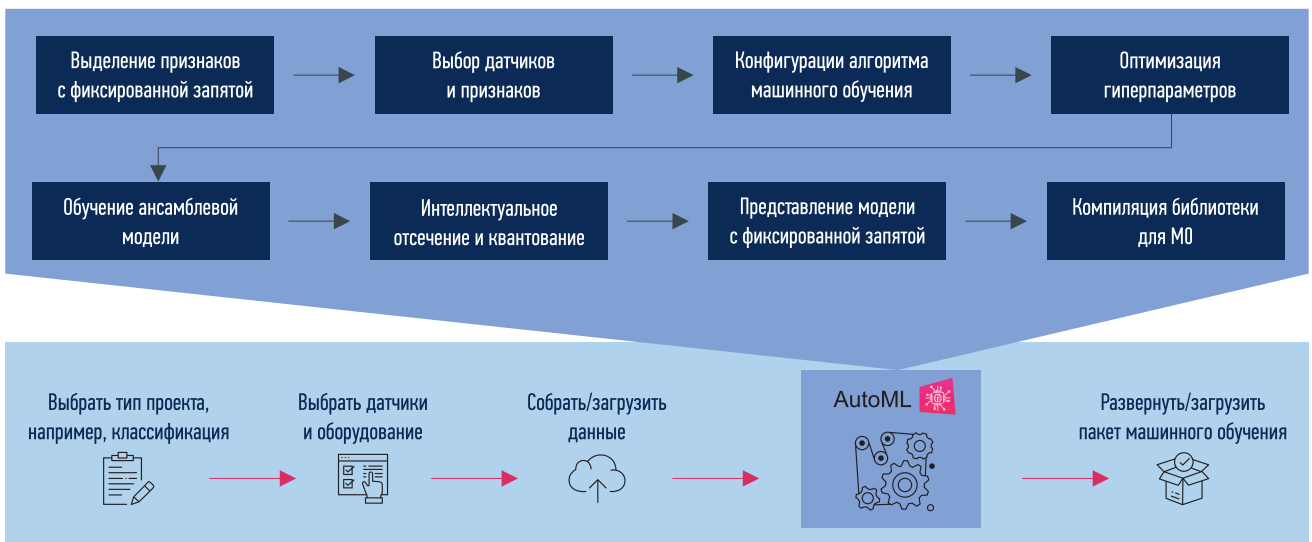
логического вывода. На рис. 3 показан конвейер машинного обучения с фиксированной запятой, разработанный Qeexo для Arm Cortex-M0+.

Qeexo AutoML выполняет запатентованное сжатие и квантование моделей для дальнейшего уменьшения объема памяти разработанных ансамблевых моделей без ущерба для эффективности классификации. На рис. 4 описан процесс обучения Qeexo AutoML для встроенной платформы Cortex-M0+.

Интеллектуальное отсеечение

Интеллектуальное отсеечение позволяет сжимать модели без потери эффективности. Проще говоря, Qeexo AutoML сначала строит пол-

РИС. 4. ▼
Конвейер обучения
Qeexo AutoML M0+



норазмерную ансамблевую модель в соответствии с рекомендациями оптимизатора гиперпараметров, а затем интеллектуально выбирает только самые мощные бустеры.

Такой подход, заключающийся в создании модели большего размера, а затем в ее интеллектуальном усечении, гораздо эффективнее, чем изначальное создание модели меньшего размера. Первоначальная модель большего размера предоставляет возможность выбрать эффективные бустеры (или деревья), что в конечном итоге приводит к повышению эффективности модели.

Как видно на рис. 5, сжатая ансамблевая модель составляет примерно 1/10 размера полной модели, но при этом обладает более высокой эффективностью перекрестной проверки (по оси X показано количество деревьев (или бустеров) в ансамблевой модели, а по оси Y — эффективность перекрестной проверки). Обратите внимание, что метод интеллектуального отсечения Qeexo AutoML выбирает только 20 самых мощных бустеров, что приводит к 90%-ному сжатию модели.

КВАНТОВАНИЕ АНСАМБЛЕВОЙ МОДЕЛИ

После обучения Qeexo AutoML выполняет квантование ансамблевых алгоритмов. Квантование после обучения — стандартная функция для моделей на основе нейронных сетей, которая по умолчанию поддерживается в таких фреймворках, как TensorFlow Lite. Однако запатентованный метод Qeexo квантования ансамблевых моделей может еще больше уменьшить размер модели, одновременно улучшив задержку на уровне микроконтроллера, практически не ухудшая эффективность модели. Конвейер Qeexo AutoML M0+ генерирует ансамблевые модели с фиксированной запятой, представленные с 32-битной точностью. Дополнительные опции для 16- и 8-битного квантования могут еще сильнее уменьшить размер модели — на 1/2 и 1/4 соответственно — при ускорении в 2–3 раза.

ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ TINYML

Какие существуют варианты использования tinyML? Возможности применения безграничны, здесь мы выделим некоторые из них:

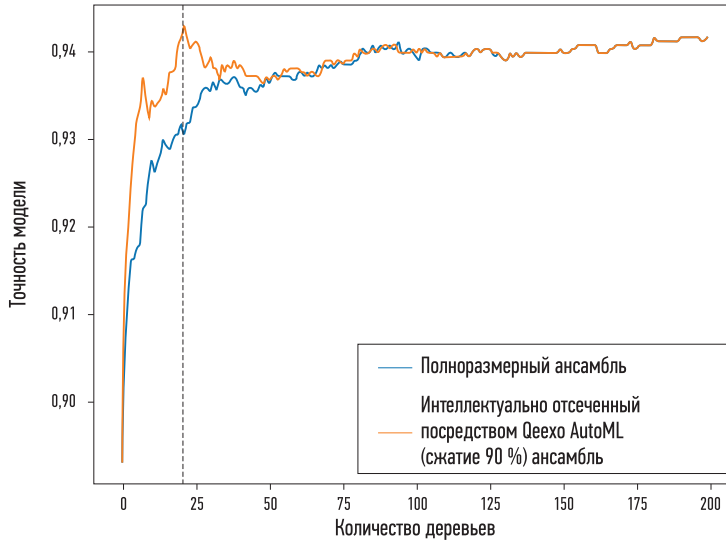


РИС. 5. ◀ Отсечение интеллектуальной модели Qeexo AutoML

1. Мы хотим создать «умную» стену с поддержкой искусственного интеллекта, с помощью нажатий на которую пользователи могут управлять освещением (включать и выключать, изменять интенсивность света). Мы можем определить жесты, связанные с включением, выключением и регулировкой интенсивности света, а затем собрать и маркировать данные о жестах с помощью модуля акселерометра и гироскопа, прикрепленного к задней части стены. Получив эти данные, Qeexo AutoML использует алгоритмы искусственного интеллекта для построения модели, способной распознавать жесты «стук» и «потирание». На видео по ссылке [1] можно увидеть прототип «умной» стены, разработанный Qeexo AutoML.

2. Используя машинное обучение и «Интернет вещей», мы хотим гарантировать, что с грузами обращаются с особой осторожностью в соответствии с рекомендациями по доставке. В приведенном по ссылке [2] видео показано, как транспортная тара с поддержкой искусственного интеллекта способна определить, как обращались с грузом начиная от отправителя и заканчивая пунктом назначения.

3. Конвергенция искусственного интеллекта и «Интернета вещей» может помочь в создании «умных» кухонных столешниц. На видео [3] показаны встроенные модели Qeexo AutoML для обнаружения различных кухонных приборов.

4. Мониторинг оборудования — один из наиболее перспективных

вариантов использования tinyML. В приведенном по ссылке [4] видео продемонстрировано распознавание различных типов неисправностей оборудования.

5. Обнаружение аномалий — это еще один сценарий, который значительно выигрывает от использования машинного обучения. Часто в промышленных условиях бывает трудно собрать данные о различных неисправностях, в то время как относительно легко мониторить оборудование в исправном рабочем состоянии. Просто наблюдая за исправным оборудованием, алгоритмы Qeexo AutoML могут разрабатывать системы искусственного интеллекта для обнаружения аномалий.

6. Распознавание активности с помощью датчиков, встроенных в носимые устройства, — еще один вариант применения, полезный в нашей повседневной жизни. В приведенном по ссылке [5] видео демонстрируется решение для распознавания активности, созданное с помощью Qeexo AutoML за считанные минуты. ●

ЛИТЕРАТУРА

- [Qeexo AutoML] Interactive Wall. <https://youtu.be/mc0aHoDb1hI>
- [Qeexo AutoML] Intelligent Shipping. <https://youtu.be/1S6irWy8G20>
- [Qeexo AutoML] Environment-Aware Countertop. <https://youtu.be/r7Mruf2vHA>
- [Qeexo AutoML] Predictive Maintenance & Anomaly Detection. <https://youtu.be/Rvd2GXDb800>
- Qeexo AutoML. Activity Tracking for Wearables Application. <https://youtu.be/wDdAPsMywEY>