DOI 10.37882/2223-2966.2026.02-2.05

СРАВНИТЕЛЬНЫЙ АНАЛИЗ И ОЦЕНКИ МЕТОДОВ КВАНТИЗАЦИИ НЕЙРОННЫХ СЕТЕЙ КОМПАКТНЫХ АРХИТЕКТУР MOBILENET ДЛЯ РАЗВЕРТЫВАНИЯ НА МОБИЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ УСТРОЙСТВАХ ДЛЯ ВИЗУАЛЬНОГО МОНИТОРИНГА ЗАСЫПАНИЯ ВОДИТЕЛЕЙ В РАЗЛИЧНЫХ УСЛОВИЯХ ОСВЕЩЕНИЯ

COMPARATIVE ANALYSIS AND EVALUATION OF NEURAL NETWORK QUANTIZATION METHODS OF COMPACT MOBILENET ARCHITECTURES FOR DEPLOYMENT ON MOBILE COMPUTING DEVICES FOR VISUAL MONITORING OF DRIVERS' SLEEP IN VARIOUS LIGHTING CONDITIONS

A. Afanasyev

Summary. This paper presents a comparative analysis of the impact of quantization methods on the efficiency, performance, and accuracy of neural network models based on the compact MobileNetV1, MobileNetV2, and MobileNetV3 architectures. This analysis is based on the implementation of a driver sleep/wake binary classification task using a Samsung Galaxy A50 smartphone with the front camera. A comparative analysis of detection quality under different lighting conditions (daytime and nighttime) is conducted, and tradeoffs between accuracy, inference speed, and resource requirements are considered.

Keywords: MobileNet, smartphone, falling asleep, drowsiness, driver, visual monitoring, neural network, quantization.

Афанасьев Арсений Геннадьевич
Аспирант, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
afanasievag@bmstu.ru

Аннотация. В работе представлен сравнительный анализ воздействие методов квантизации на эффективность, производительность и точность нейросетевых моделей компактных архитектур MobileNetV1, MobileNetV2 и MobileNetV3 на примере реализации задачи бинарной классификации «засыпание/бодрствование» водителя с применением смартфона Samsung Galaxy A50 с использованием фронтальной камеры. Проводится сравнительный анализ качества детекции в разных условиях освещения (дневные и ночные), рассматриваются компромиссы между точностью, скоростью инференса (inference) и требованиями к ресурсам.

Ключевые слова: MobileNet, смартфон, засыпание, сонливость, водитель, визуальный мониторинг, нейросеть, квантизация.

Введение

Проблема сонливости, засыпания водителей является одной из ключевых причин дорожно-транспортных происшествий во всем мире. По данным различных исследований, утомление и микросон за рулем приводят к снижению скорости реакции, ухудшению восприятия дорожной обстановки и, как следствие, к повышению аварийности. Особенно опасны такие состояния в ночное время и при монотонных условиях движения. В связи с этим системы мониторинга состояния водителя (Driver Monitoring Systems, DMS) становятся важной частью современных транспортных систем [1]. Одним из подходов для создания этих систем является использование методов машинного обучения, которое успешно зарекомендовало себя в самых различных областях, таких как например, медицина [2], определение эмоционального содержания текста [3], поиск и обнаружение инфракрасных целей [4], прогнозирование загрузки виртуальных вычислительных систем [5-8], компьютерное зрение [9], распознавание лиц [10, 11], определение усталого состояния человека [12], и для других задач .

Ключевой задачей при реализации подобных систем на мобильных вычислительных устройствах является обеспечение высокой точности распознавания при ограниченных вычислительных ресурсах и энергопо-

треблении. Для этого могут применяться легковесные архитектуры сверточных нейронных сетей, таких как MobileNetV1, MobileNetV2 и MobileNetV3.[13] Дополнительным способом оптимизации является квантизация нейронных сетей [14], позволяющая уменьшить размер модели и ускорить вычисления за счет использования чисел пониженной разрядности.

Целью данной статьи является анализ методов квантизации для нейронных сетей семейства MobileNet для повышения эффективности решения задачи определения сонливости водителя с использованием фронтальной камеры смартфонов в различных условиях освещенности.

В качестве аппаратной платформы для проведения экспериментов был выбран смартфон Samsung Galaxy A50, оснащённый однокристальной системой Exynos 9610 (8 ядер: Cortex A73 и Cortex A53), графическим ускорителем Mali G72 MP3 и оперативной памятью объёмом 4 ГБ. Операционная система Android 11.

Емкость аккумулятора смартфона Samsung Galaxy A50 составляет 4000 мАч (миллиампер-часов). Эта литий-полимерная (Li-Pol) батарея обеспечивает непрерывную работу в режиме с фронтальной камерой около 20 часов.

Устройство не содержит выделенного нейропроцессора (NPU), поэтому инференс нейронных сетей выполняется и на CPU и GPU.

Фронтальная камера смартфона Samsung Galaxy A50 имеет разрешение 25 Мп и оптимизирована для работы как при дневном свете (для четкости), так и в условиях недостаточного освещения (для яркости и насыщенности).

Необходимо отметить, что квантизация наиболее эффективна, когда вычисления выполняются на CPU или NPU. Поэтому наличие GPU у смартфона Samsung Galaxy A50 не сильно влияет на эффективность вычислений при квантизации в случае int 8/int 4, так как GPU оптимизирован под вычисления в формате float16.

Данные ограничения делают Samsung Galaxy A50 показательной платформой для оценки эффективности квантизации моделей MobileNet с точки зрения скорости инференса и энергопотребления на массовых мобильных устройствах

Постановка задачи и требования к системе

Задача определения сонливости, засыпания водителя формулируется как задача бинарной сводной классификации зевков, состояний глаз, положения головы по видеопотоку с фронтальной камеры (рис. 1).

Пусть видеопоток, получаемый с фронтальной камеры смартфона, представлен в виде последовательности кадров:

$$X = \left\{ x_t \middle| x_t \in \mathbb{R}^{H \times W \times C}, t = 1,2,\ldots,T \right\},$$

где H, W — пространственное разрешение изображения,

C — количество цветовых каналов,

T — длина временного интервального окна анализа.

Задача определения засыпания водителя формулируется как задача бинарной классификации:

$$y = f(X, \theta), y \in \{0, 1\}$$

- $y = 1$ — соответствует состоянию засыпания,
- $y = 0$ — состояние бодрствования водителя,
- θ — параметры нейронной сети MobileNet.

Вероятность засыпания вычисляется с использованием сигмоидной функции активации:

$$P(y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

где z — выход последнего слоя нейронной сети.

Основными признаками засыпания являются:
- длительное закрытие глаз (метрика PERCLOS);
- частое моргание или, наоборот, редкое моргание;
- наклон головы;
- частота зевков.

Математически метрика PERCLOS (PERcentage of eye CLOSure) определяется следующим образом:

$$PERCLOS = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}(e_t < \tau),$$

где
- T — общее число кадров в анализируемом временном окне,
- e_t — нормированная степень раскрытия глаз на кадре t,
- τ — пороговое значение, соответствующее состоянию «глаза закрыты»,
- $\mathbb{I}(\cdot)$ — индикаторная функция, принимающая значение 1 при выполнении условия и 0 в противном случае.

Для непрерывного видеопотока метрика PERCLOS может быть представлена в эквивалентной временной форме:

$$PERCLOS = \frac{T_{closed}}{T_{window}},$$

где T_{closed} — суммарная длительность интервалов, в течение которых глаза водителя находятся в закрытом со-

Рис. 1. Система мониторинга засыпания водителя

стоянии, а T_{window} -длительность временного интервала наблюдения. При превышении заданного порогового значения система делает вывод о высокой вероятности засыпания водителя:

$$PERCLOS \geq P_{thr} \Rightarrow y = 1.$$

На практике значение порога P_{thr} выбирается в диапазоне 0,3–0,4 в зависимости от условий освещенности и требований к чувствительности системы.

Методы квантизации нейронных сетей

Квантизация нейронных сетей представляет собой процесс преобразования весов и значений активаций из чисел с плавающей точкой (обычно float32) в числа с меньшей разрядностью (int4, int8, int16, float16).

Основные преимущества квантизации:

- уменьшение размера модели;
- ускорение инференса;

- снижение энергопотребления;
- лучшая совместимость с мобильными ускорителями (Digital Signal Processor, DSP; Neural Processing Unit, NPU).

Основные методы квантизации

Пост-тренировочная квантизация (Post-Training Quantization, PTQ) — данный метод применяется после завершения обучения модели. Веса и/или значения активации преобразуются в int8 с использованием калибровочного набора данных. PTQ проста в реализации и не требует повторного обучения, однако может приводить к заметному снижению точности, особенно для сложных моделей и задач с высокой чувствительностью к шуму. Этот метод очень популярный для смартфонов. Потеря качества обычно небольшая (часто <1–2 %)

Квантизация с учетом обучения (Quantization Aware Training, QAT). При QAT эффекты квантизации моделируются непосредственно в процессе обучения. Это позволяет сети адаптироваться к пониженной разрядности и сохранить точность. Этот метод используется если важна максимальная точность. Точность почти как у float32. Недостатком является увеличение времени и сложности обучения.

Динамическая квантизация. При динамической квантизации веса хранятся в квантизированном виде, а значения активации квантизируются на лету во время инференса. Такой подход часто используется для рекурентных и полносвязных сетей, но реже применяется для сверточных моделей в задачах компьютерного зрения.

Адаптивная квантизация представляет собой расширение классических методов квантизации, при котором параметры квантизации (разрядность, масштабные коэффициенты или пороги) изменяются в зависимости от условий работы системы или характеристик входных данных. В отличие от статической квантизации, где используется единый набор параметров для всей модели, адаптивный подход позволяет учитывать неоднородность распределений активаций в различных слоях и изменчивость внешних факторов. В задачах мониторинга состояния водителя адаптивная квантизация может учитывать условия освещенности (день/ночь), уровень шума изображения и текущую нагрузку на вычислительные ресурсы. Например, при хорошей освещенности возможно применение более агрессивной int8-квантизации, тогда как в ночных условиях для чувствительных слоев целесообразно использовать повышенную разрядность (int16 или float16). Адаптивная квантизация позволяет достичь более выгодного компромисса между точностью и скоростью инференса, однако усложняет реализацию системы и требует дополнительных механизмов мониторинга условий работы. Для устройств без выделенного NPU, таких как Samsung Galaxy A50, данный подход представляет интерес прежде всего в исследовательском контексте, поскольку его практическая реализация на уровне мобильных фреймворков ограничена.

Архитектуры MobileNet и их особенности

MobileNetV1 является одной из первых архитектур, специально разработанных для мобильных и встраиваемых устройств. Основная идея заключается в использовании глубинно-разделимых сверток (Depthwise Separable Convolutions), которые разбивают стандартную свертку на две операции:

- Depthwise-свертка — применяется отдельно к каждому каналу входного тензора;
- Pointwise-свертка (1×1) — объединяет каналы.

Использование данной схемы позволяет существенно сократить вычислительные затраты. И это обеспечивает значительное ускорение инференса и снижение энергопотребления, что особенно важно для мобильных устройств. MobileNetV1 отличается простой структурой и предсказуемым поведением при оптимизации, однако может уступать более новым версиям по точности.

MobileNetV2 развивает идеи предыдущей версии и вводит два ключевых улучшения:

- инвертированные остаточные блоки (Inverted Residuals);
- линейные «бутылочные слои» (Linear Bottlenecks).

В отличие от классических остаточных соединений, MobileNetV2 сначала расширяет размерность признаков, а затем сжимает ее. Использование линейных активаций на выходе «бутылочного» слоя снижает потери информации при квантизации и сжатии. Это делает MobileNetV2 более устойчивой к агрессивным методам оптимизации.

MobileNetV3 является результатом автоматизированного поиска архитектур (NAS) и включает в себя дополнительные оптимизации:

- нелинейность h-SWish вместо ReLU;
- механизм внимания Squeeze-and-Excitation;
- более эффективное распределение вычислений между слоями.

Существуют две основные конфигурации: MobileNetV3-Large и MobileNetV3-Small. Для задач мониторинга водителя на смартфоне чаще используется версия Small, так как она обеспечивает лучший баланс между скоростью и точностью.

Анализ применения квантизации для MobileNetV1

MobileNetV1 благодаря своей простой структуре хорошо поддается пост-тренировочной квантизации.

Эксперименты показывают, что при использовании int8-квантизации размер модели уменьшается примерно в 4 раза, а скорость инференса на мобильном процессоре возрастает на 30–50 %.

Однако в задаче определения засыпания водителя, особенно в ночных условиях, MobileNetV1 демонстрирует заметное падение точности после PTQ. Это связано с тем, что слабоконтрастные изображения лица и шумы при плохом освещении требуют более точного представления признаков. В таких условиях предпочтительнее использовать QAT, что позволяет снизить потери точности до 1–2 % по сравнению с float32-моделью.

Анализ применения квантизации для MobileNetV2

MobileNetV2 считается более устойчивой к квантизации благодаря линейным «бутылочным» слоям. При пост-тренировочной int8-квантизации снижение точности в дневных условиях обычно не превышает 1%, а в ночных — 2–3 %. При использовании QAT метода MobileNetV2 показывает стабильные результаты как днем, так и ночью. Дополнительным преимуществом является более высокая точность по сравнению с MobileNetV1 при схожих вычислительных затратах. Для Samsung Galaxy A50, не оснащенного специализированным NPU высокого класса, MobileNetV2 в квантизированном виде является одним из наиболее сбалансированных вариантов.

Анализ применения квантизации для MobileNetV3

MobileNetV3 изначально проектировалась с учетом работы на мобильных устройствах и хорошо адаптирована к int8-инференсу. Использование h-swish и блоков внимания повышает выразительную способность сети, однако может усложнять квантизацию при отсутствии QAT.Пост-тренировочная квантизация MobileNetV3-Small дает хорошие результаты в дневных условиях, но в ночных сценариях возможно более заметное падение точности из-за чувствительности механизмов внимания к шуму. Применение QAT позволяет практически полностью нивелировать этот эффект. С точки зрения производительности MobileNetV3-Small в int8-формате демонстрирует наилучшее соотношение точности и скорости, что делает ее перспективным выбором для длительной работы системы мониторинга без значительного расхода аккумулятора смартфона.

Методика эксперимента

1. Сбор данных. Использовались видеозаписи с фронтальной камеры Samsung Galaxy A50 в реальных условиях. Дневные условия: яркий свет, естественное освещение. Ночные условия: искусственное освещение, слабая видимость. Данные были аннотированы вручную. Класс 0 — водитель бодрствует. Класс 1 — засыпание.
2. Предобработка. Детекция лица с помощью детектора MediaPipe Face Detection (BlazeFace)— самый быстрый по FPS с хорошей точностью детектор лица на смартфоне [15]. Нормализация размера изображения на 224×224. Аугментации (яркость, контраст, поворот) для устойчивости к условиям.
3. Обучение и оптимизация. Базовая тренировка моделей (float32). Настройка параметров квантизации через PTQ и QAT. Тестирование производительности на Samsung Galaxy A50 (табл. 1).

Таблица 1.

Точность моделей по условиям освещенности для формата float32

Модель	Дневные условия	Ночные условия
MobileNetV1	~87 %	~75 %
MobileNetV2	~90 %	~83 %
MobileNetV3-Small	~92 %	~85 %

Замечание. MobileNetV3-Large обычно даёт ещё лучший результат, но слишком тяжёлая для реального времени на Samsung Galaxy A50 без оптимизации.

4. Влияние PTQ. Значительное уменьшение размера моделей (до ~4). Небольшое падение точности (на 2–5 %). Ускорение инференса ~30–80 %.
5. Влияние QAT Уменьшение потерь точности (почти на уровне float32). Настройка веса моделей позволила адаптировать их к шуму ночных кадров. Требует больше времени тренировки и вычислительных ресурсов.
6. Сравнение по условиям освещения (табл. 2)

Таблица 2.

Сравнение по условиям освещения

Метод	Лучшая модель	Accuracy (днём)	Accuracy (ночь)
Float32	MobileNetV3	92 %	85 %
PTQ int8	MobileNetV3	90 %	82 %
QAT int8	MobileNetV3	91 %	84 %

7. Выявленные особенности использования на смартфоне:
- Температура мобильного устройства может снижать FPS;
- Фронтальная камера плохо справляется с шумом ночью;
- Требуется баланс между скоростью, точностью и энергопотреблением.
8. Практические рекомендации. Наличие целесообразности в адаптивной квантизацию для переключения между режимами работы нейронной сети в зависимости от освещения и временем суток.

Заключение

Методы квантизации позволяют эффективно адаптировать нейросети MobileNet к работе в реальном времени на мобильных устройствах подобных Samsung Galaxy A50, сохраняя приемлемую точность. Наблюдается чёткая зависимость между архитектурой модели, методом квантизации и условиями освещения. MobileNetV3 с QAT-оптимизацией наиболее универсален, особенно для сложных условий ночной трассы, тогда как PTQ даёт хорошую производительность с минимальными затратами на разработку.

Перспективы дальнейших исследований:

- Исследование адаптивной квантизации в реальном времени;
- Использование трансформеров с адаптивным порогом;
- Гибридные модели с отслеживанием зевков, состояния глаз и положения головы.

ЛИТЕРАТУРА

1. Driver_monitoring_system [Electronic resource] // En.wikipedia.org URL: https://en.wikipedia.org/wiki/Driver_monitoring_system (accessed: 04.01.2026).
2. Ся Т., Афанасьев Г.И., Афанасьев А.Г. Искусственный интеллект в стоматологии // Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2023. № 2-2. С.121–127.
3. Инь С., Афанасьев Г.И., Калистратов А.П. Метод применения нейронных сетей BERT-BiLSTM-Attention для определения эмоционального отношения автора к тексту//Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2023. № 7-2. С.55–58.
4. Ван Ч., Афанасьев Г.И., Афанасьев А.Г. Алгоритм обнаружения слабых инфракрасных целей на сложном фоне посредством нейросетевой модели YOLOv5 //Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2023. № 8-2. С. 54–59.
5. Зуев Д.А., Калистратов А.П., Семкин П.С., Афанасьев Г.И. Анализ производительности алгоритма HyperLogLog при потоковой обработке данных // Динамика сложных систем-XXI век. 2017. т.11. №4. С.51–55.
6. Калистратов А.П., Афанасьев Г.И., Подход к реализации моделирования производительности вычислительной системы //Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2018. № 6. С.58–62.
7. Калистратов А.П., Ревунков Г.И. Семкин П.С., Афанасьев Г.И., Влияние распределения системных ресурсов на производительность виртуальных машин // Динамика сложных систем-XXI век. 2017. т.11. №4. С.46–50.
8. Нестеров Ю.Г., Калистратов А.П., Афанасьев Г.И. Подход к применению машинного обучения в прогнозировании загрузки виртуальных вычислительных систем // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2019. № 11-2. С. 73–76.
9. Ма Л., Афанасьев Г.И., Афанасьев А.Г. Компьютерное зрение в интеллектуальных транспортных системах // Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2023. № 2-2. С.101–105.
10. Крутов Т.Ю., Афанасьев Г.И., Афанасьев А.Г. Сиамские нейросети для задачи распознавания лиц // Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2022. № 12-2. С.88–92.
11. Чжоу Х., Афанасьев Г.И., Афанасьев А.Г., Филатова А.Е. Система распознавания лиц на основе библиотеки алгоритмов компьютерного зрения OpenCV // Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2023. № 8-2. С.142–145.
12. Фэн Кэцзя, Афанасьев Г.И., Нестеров Ю.Г. Применение искусственного интеллекта в прогнозировании усталого поведения водителей при вождении автотранспортными средствами // Современная наука: Актуальные проблемы теории и практики. Серия Естественные и Технические Науки. 2022. № 12-2. С.190–195.
13. MobileNet [Electronic resource] // Keras.io URL: https://keras.io/api/applications/mobilenet (accessed: 04.01.2026).
14. A Survey On Neural Network Quantization A Survey On Neural Network Quantization [Electronic resource] /Jiawei Y., Zhongbo L., Zeqin F., Yongqiang X. // ACM Digital Library URL: https://dl.acm.org/doi/10.1145/3746709.3746773 (accessed: 04.01.2026).
15. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs [Electronic resource] /Valentin Bazarevsky V., Kartynnik Yu., Vakunov A., Raveendran K., Grundmann M. //Cornell University URL: https://arxiv.org/abs/1907.05047 (accessed: 04.01.2026).

Журнал «Современная наука: актуальные проблемы теории и практики»