doi: 10.36724/2409-5419-2025-17-2-4-10

АНАЛИЗ ЭФФЕКТИВНОСТИ ПОСТОБУЧАЮЩЕГО КВАНТОВАНИЯ ДЛЯ ОПТИМИЗАЦИИ НЕЙРОННЫХ СЕТЕЙ

ТАТАРНИКОВА
Татьяна Михайловна[1]

РАСКОПИНА
Анастасия Сергеевна[2]

АННОТАЦИЯ

Введение: Технологии развиваются стремительно в том числе это касается нейронных сетей. После повеления глубокого обучения, с каждым годом модели становились все сложнее и глубже, что привело к тому что стало не хватать аппаратных мощностей. В статье рассматриваются современные методы оптимизации нейронных сетей с акцентом на постобучающее квантование как наиболее практичный подход для развертывания моделей в условиях ограниченных вычислительных ресурсов. **Методы:** Представлен обзор ключевых методов, включая прунинг, квантование и дистилляцию знаний, проведено сравнение их эффективности и применимости. Особое внимание уделено преимуществам и ограничениям PTQ, таким как сокращение размера модели, ускорение инференса и совместимость с промышленными фреймворками. В экспериментальной части представлены результаты квантования моделей MobileNetV2, BERT-base, YOLOv5s, EfficientNet-B0 и DistilBERT, анализируется влияние квантования на точность, скорость и компактность моделей. **Результаты** показали, что постобучающее квантование отлично справляется со своей задачей. Данный метод сумел сократить размер модели в 3 раза, ускорить инферес до 40 % и потерять в точность не больше 1.5%. Полученные результаты могут стать основой для дальнейших исследований оптимизации нейронных сетей , комбинирования метода квантования с другими методами и создания новых гибридных методов, которые будут брать все преимущества постобучающего квантования и нивелировать недостатки. Ведь постобучающее квантование особенно эффективен для мобильных и IoT-устройств, где критичны требования к энергопотреблению и памяти. А его и его использование для задач компьютерного зрения и обработки естественного языка – уже сейчас показывает применимость и перспективы.

Сведения об авторах:

[1] д.т.н., профессор, директор Института инфорамционных технологий и программирования ГУАП, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия, tm-tatarn@yandex.ru

[2] аспирант, ассистент Кафедры прикладной информатики ГУАП, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия, raskopina.anastasia@yandex.ru

КЛЮЧЕВЫЕ СЛОВА: *нейронные сети; постобучающее квантование; оптимизация моделей; прунинг; квантование; дистилляция знаний; ускорение инференса; сжатие модели.*

Для цитирования: *Татарникова Т.М., Раскопина А.С.* Анализ эффективности постобучающего квантования для оптимизации нейронных сетей // Наукоемкие технологии в космических исследованиях Земли. 2025. Т. 17. № 2. С. 4-10. doi: 10.36724/2409-5419-2025-17-2-4-10

Введение

В последнее время методы глубокого обучения стали развиваться все более и более стремительно. Это связанно с тем, что эффективность в задачах прогнозирования, классификации, компьютерного зрения и детекции объектов иногда превосходит даже человека. Отличные результаты искусственный интеллект также показывает и в задачах рекомендательных систем и обработки естественного языка. Данные системы плотно интегрированы в нашу ежедневную рутину и значительно облегчают нам жизнь. Но с развитием методов глубокого обучения начал расти и размер самих моделей, так как они становились все сложнее и глубже, что начало требовать значительные затраты на вычислительные ресурсы и требования к аппаратным мощностям. Данные тенденции стали проблемой во многих задачах, ведь часто приходится применят нейросети в условиях ограниченных ресурсов – например, на мобильных устройствах, встраиваемых системах и в приложениях Интернета вещей (IoT) [1].

В связи с данной проблемой в современном реалиях искусственного интеллекта остро встал вопрос оптимизации нейронных сетей. Существует множество методов оптимизации нейронных сетей, но не многие могут похвастаться той оптимизацией, которой позволит не понижать значительно точность модели. Большинство из них просто ускоряют весь процесс, но на столько занижают результативность модели, что применять их не имеет никакого смысла. Но существуют те методы, которые стараются держать баланс между точностью и эффективностью модели:

• Прунинг. Данный метод удаляет малозначимые веса или нейроны [2].

• Квантование. Данный метод снижает разрядность параметров.

• Дистилляция знаний. Данный метод передает знания от более крупной модели к более компактной [3].

Каждый и этих методов и алгоритмов имеет ряд своих преимуществ и ограничений. Выбор конкретного метода или их комбинация зависит от задачи, данных, требований к результатам и точности, а также от вычислительных ресурсов, выделенных под эту задачу.

Среди данных методов в последнее время уделяется большое внимание квантованию. Как было сказано ранее, данный метод снижает разрядность параметров. Это значит, например, что параметры, представленные в формате с плавающей запятой (FP32) будут преобразованы в более компактные форматы, что позволит ускорить сами модели и значительно сократить занимаемый ими объем памяти [4].

Квантование не обязательно применять во время обучения, его можно также применить и после. Квантование в процессе обучения называют quantization-aware training в сокращении QAT, а квантование после обучения называют post-training quantization в сокращении PTQ. В действительности возможность использовать постобучающее квантование является одним из самых привлекательных методов для исследователей. Ведь используя данный метод не требуется полностью переобучать модель и есть возможность применять его к уже готовым нейросетям, что действительно удобно для промышленного применения.

Тем не менее, выбор между различными методами оптимизации требует тщательного анализа. У каждого метода есть неоспоримые преимущества, но нужно учитывать, где именно эти преимущества помогут добиться наилучших результатов.

В частности, прунинг может обеспечивать высокую степень сжатия, но требует последующего переподгонки модели. Дистилляция знаний эффективна для переноса характеристик больших моделей, но связана с необходимостью обучения новой модели-ученика.

В данной статье будут рассматриваться различные методы оптимизации нейронных сетей. Но особое внимание будет уделено квантованию. Постобучающее квантование в этом контексте выступает как компромиссный подход, обеспечивающий значительное снижение вычислительных затрат при минимальных требованиях к дополнительным ресурсам.

Экспериментальная часть будет посвящена реализации метода квантования, интерпретации полученных результатов и анализу эффективности данного метода на реальных задачах.

Обзор методов оптимизации нейронных сетей

Рассмотрим ранее упомянутые методы, как прунинг,квантования и дистилляция знаний, более подробно.

Прунинг представляет собой процесс удаления избыточных параметров нейросети, таких как веса, нейроны или фильтры, которые оказывают незначительное влияние на её выходной сигнал. Идея этого метода заключается в том, что большая часть параметров модели в действительности не является критически важной для её корректной работы. Удаляя такие элементы, можно достичь существенного уменьшения объема модели и числа необходимых вычислений. Существует несколько подходов к реализации прунинга, различающихся степенью структурированности: от точечного удаления отдельных весов до вырезания целых структурных блоков, таких как каналы или фильтры. При этом важно учитывать, что неосторожный прунинг может привести к снижению точности модели, и поэтому часто применяется повторное обучение после удаления параметров для восстановления её первоначального качества. Несмотря на свою эффективность, прунинг требует тщательной настройки критериев важности параметров и может быть достаточно ресурсоёмким процессом [9].

Другим широко применяемым направлением оптимизации является квантование нейронных сетей. Оно основано на преобразовании весов и активаций из формата с высокой точностью, как правило 32-битных чисел с плавающей точкой, в форматы с меньшей разрядностью – например, в 16-битные или 8-битные целые числа. Это позволяет существенно уменьшить объем памяти, необходимый для хранения модели, и ускорить выполнение операций, особенно при использовании специализированных аппаратных решений.

Квантование может проводиться как после обучения модели, без дополнительной её адаптации, так и в процессе обучения, с учетом ограничений разрядности. Последний подход позволяет добиться более высокой точности, поскольку сеть постепенно приспосабливается к условиям низкой точности вычислений.

Кроме того, в рамках квантования могут использоваться различные схемы распределения значений, включая симметричные и асимметричные, в зависимости от характера весов и активаций. Благодаря своей простоте и совместимости с современными фреймворками и библиотеками, квантование стало одним из наиболее востребованных методов оптимизации нейросетей, особенно в промышленной практике и при развертывании моделей на устройствах с ограниченными ресурсами [10, 11].

Не менее интересным и перспективным направлением является дистилляция знаний, которая основывается на передаче информации от сложной и высокоточной модели-учителя к более компактной модели-ученику. При этом ученик обучается не на истинных метках обучающей выборки, а на выходных сигналах учителя, что позволяет добиться более глубокого понимания структуры данных. Такой подход способствует улучшению обобщающих способностей модели-ученика и позволяет существенно сократить её размер без серьёзной потери в точности. Наибольшую эффективность дистилляция демонстрирует в задачах классификации и обработки естественного языка, где важно сохранить контекстную и вероятностную структуру выходных данных. Однако реализация данного подхода требует наличия обученной модели-учителя и достаточного объема обучающих данных для передачи знаний, что может быть неприменимо в условиях ограниченного времени или ресурсов [12-15].

Постобучающее квантование нейронных сетей

Как говорилось ранее, постобучающее квантование позволяет не переобучать модель, а использовать ее на готовой обученной нейросети, что значительно может ускорить решение определенных задач.

Ключевыми этапами постуобучающего квантования являются:

• Анализ распределения весов и активаций, то есть сбор статистики, необходимой для правильного определения диапазонов квантования.

• Определение масштабных коэффициентов и нулевых смещений, которые обеспечивают переход от действительных чисел к целочисленным представлениям.

• Преобразование весов модели и параметров слоёв, то есть пересчёт данных в соответствии с выбранными уровнями квантования [16, 17].

Существует несколько разновидностей постобучающего квантования, отличающихся степенью вмешательства в модель и способом преобразования данных – динамическое и статическое.

Динамическое квантование предполагает, что веса модели квантуются заранее, а активации преобразуются в процессе выполнения. Этот подход часто применяется к моделям обработки естественного языка [18].

Статическое квантование, наоборот, требует предварительного преобразования как весов, так и активаций на основе калибровочного набора данных, что позволяет достичь более высокой производительности. Этот подход упрощает реализацию и снижает вероятность потери точности [19].

Конкретный выбор динамического или статического метода определяется требованиями к скорости инференса, допустимой потерей точности и характеристиками аппаратной платформы.

Постобучающее квантование обладает рядом значительных преимуществ. Во-первых, оно позволяет сократить размер модели: переход от FP32 к INT8 уменьшает объем занимаемой памяти примерно в 4 раза. Во-вторых, ускоряется инференс, так как целочисленные операции выполняются быстрее операций с плавающей точкой, особенно на специализированных ускорителях.

В-третьих, отсутствует необходимость в ресурсоемком переобучении, что делает PTQ крайне привлекательным в условиях ограниченного доступа к вычислительным ресурсам или исходному обучающему набору.

В-четвёртых, большинство современных фреймворков машинного обучения, включая TensorFlow Lite, ONNX Runtime и PyTorch, предоставляют встроенные инструменты для применения квантования.

Однако постобучающее квантование не является универсальным решением. В некоторых случаях оно может вызывать снижение точности, особенно при применении к сложным архитектурам или чувствительным задачам, где важны даже минимальные отличия в выходных значениях. Тем не менее, в подавляющем большинстве прикладных сценариев, особенно там, где модели используются для классификации или других задач с не слишком высокой чувствительностью к ошибкам, PTQ демонстрирует высокую эффективность и становится предпочтительным методом оптимизации.

Сравнительный анализ показывает, что постобучающее квантование выигрывает за счёт минимальных затрат времени и вычислительных ресурсов на подготовку, не требует дополнительных этапов обучения и может применяться к большинству существующих моделей без значительной модификации их структуры. В то же время прунинг и дистилляция остаются актуальными в ситуациях, когда приоритетом является максимальная точность при минимальном размере модели, и когда есть возможность провести повторное обучение [20].

Экспериментальная часть

В рамках экспериментальной части исследования были выбраны пять моделей: MobileNetV2, BERT-base, YOLOv5s, EfficientNet-B0 и DistilBERT. Все модели были обучены в двух вариантах: в исходной (полной) форме без применения квантования и в варианте после постобучающего квантования для оценки влияния квантования на точность, размер модели и скорость инференса. Эксперименты проводились на платформе на базе GPU с последующим переносом моделей на CPU для замеров производительности.

В процессе работы был использован язык программирования Python версии 3.10 и следующие основные библиотеки: PyTorch для обучения и квантования моделей, torchvision для работы с компьютерным зрением, а также Hugging Face Transformers и Datasets для работы с текстовыми задачами и загрузки датасетов. Для проведения всех этапов квантования применялись стандартные средства библиотеки PyTorch.

На начальном этапе работы с каждой моделью были загружены предобученные веса, после чего производилось дополнительное обучение на соответствующих задачах. Для MobileNetV2 использовался подмножество датасета ImageNet, модель обучалась с использованием оптимизатора Adam с начальными параметрами скорости обучения 0.001 и размером батча 128.

Для BERT-base был выбран датасет SST-2 для задачи классификации тональности текста. Модель дообучалась с использованием оптимизатора AdamW при скорости обучения 2e-5 в течение трёх эпох. Сначала оценивалась точность исходной модели без квантования. Затем модель подвергалась динамическому квантованию, затрагивающему только слои полносвязных нейронных сетей, с последующим сравнением результатов.

YOLOv5s была использована для задачи детекции объектов на подмножестве COCO. Модель дообучалась в течение 300 эпох с использованием оптимизатора SGD с моментумом 0.9 и начальной скоростью обучения 0.01. Модель в полной форме тестировалась на валидационном наборе, метрикой служила средняя точность (mean Average Precision, mAP). После тестирования YOLOv5s подвергалась статическому квантованию с калибровкой на изображениях и проводилась повторная оценка.

EfficientNet-B0 обучалась на датасете CIFAR-100 с применением базовых аугментаций, таких как случайная обрезка и горизонтальное отражение. Модель оптимизировалась с помощью Adam со скоростью обучения 0.001

DistilBERT применялся для задачи классификации новостей на датасете AG News. Fine-tuning модели происходил в течение четырех эпох при скорости обучения 3e-5 с использованием оптимизатора Adam. После обучения производилась оценка точности модели без квантования. Затем модель была подвергнута динамическому квантованию, с фокусом на ускорение инференса без значительного ухудшения качества классификации.

После обучения каждой модели переходили к этапу квантования. Для реализации квантования использовались встроенные инструменты библиотеки PyTorch. В зависимости от типа модели применялись разные схемы квантования.

Для моделей MobileNetV2, YOLOv5s и EfficientNet-B0 использовалось статическое постобучающее квантование. Для этого на этапе подготовки модели производилась замена исходных слоёв на квантованные аналоги с помощью методов из модуля torch.quantization. В процессе калибровки подавалось на модель ограниченное количество реальных обучающих данных, что позволяло определить оптимальные интервалы квантования для весов и активаций. После калибровки модель конвертировалась в квантованный формат и сохранялась для последующего тестирования.

Для моделей BERT-base и DistilBERT применялось динамическое квантование. Оно осуществлялось путём применения функции динамического квантования, которая автоматически заменяла подходящие слои на их квантованные версии во время инференса. Динамическое квантование не требовало дополнительной калибровки на обучающих данных и выполнялось значительно быстрее, чем статическое.

На всех этапах процесса квантования контролировалась правильность преобразований, проверяя соответствие архитектуры модели, а также тестировала работоспособность модели на валидационном наборе данных перед замером итоговых метрик. После квантования модели были сохранены в отдельных файлах для дальнейшего сравнения по размерам, скорости инференса и точности.

В таблице 1 можно увидеть характеристики моделей до квантования.

Таблица 1

Характеристики моделей до квантования

Модель	Точность (%)	Размер модели (МБ)	Время инференса (мс)	FLOPs	Использование RAM (МБ)
MobileNetV2	71.8	13.4	18	0.30	128
EfficientNet-B0	76.3	20.0	25	0.39	158
YOLOv5s	39.2 (mAP)	14.2	45	16.0	242
BERT-base	82.6	420	160	23.2	650
DistilBERT	79.1	250	95	11.2	380

В таблице 2 можно увидеть характеристики моделей после квантования.

Таблица 2

Характеристики моделей после квантования

Модель	Точность (%)	Размер модели (МБ)	Время инференса (мс)	FLOPs	Использование RAM (МБ)
MobileNetV2	71.2	3.4	12	0.20	84
EfficientNet-B0	75.1	5.2	17	0.26	92
YOLOv5s	38.5 (mAP)	3.7	28	10.0	150
BERT-base	81.8	105	112	15.7	520
DistilBERT	78.7	62	66	8.4	295

После обучения всех моделей в их исходных (FP32) форматах, я провела оценку их производительности по ключевым метрикам: точность, размер модели, время инференса, вычислительная сложность (FLOPs) и потребление оперативной памяти. Затем к этим же моделям было применено квантование, с последующим повторным измерением указанных характеристик. Это позволило наглядно сравнить эффективность до и после применения методов оптимизации (рис. 1).

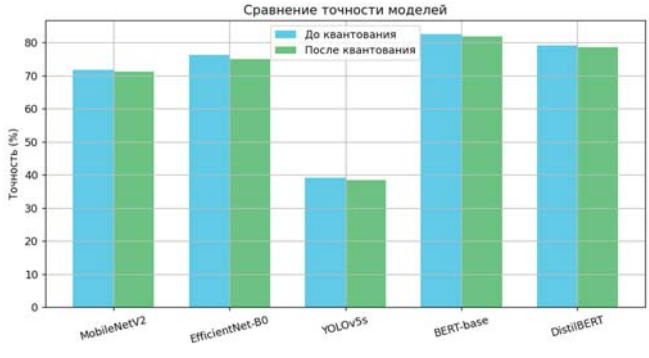

Рис. 1. Сравнение точности моделей

Из графика можно заметить, что квантование моделей оказывает минимальное влияние на точность, хотя и приводит к небольшому снижению. Для модели MobileNetV2 точность уменьшилась с 71.8% до 71.2%, что представляет собой малозаметное изменение. Однако для модели YOLOv5s снижение точности более выражено (с 39.2% до 38.5%). Модели с высокими показателями точности, такие как BERT-base (снижение с 82.6% до 81.8%) и DistilBERT (снижение с 79.1% до 78.7%), демонстрируют, что даже после квантования они сохраняют высокую эффективность.

После квантования размер всех моделей значительно сокращается (рис. 2).

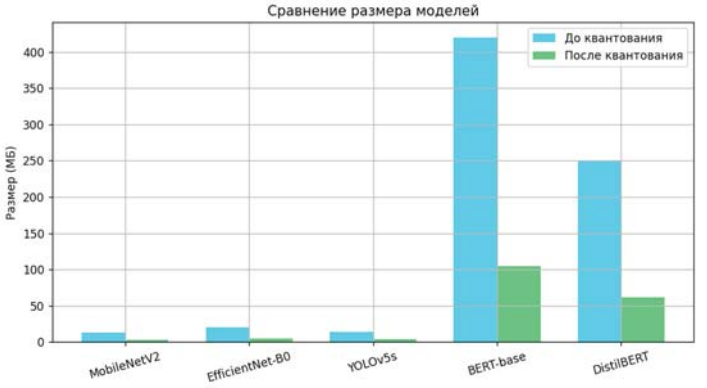

Рис. 2. Сравнение размеров моделей

Размер модели MobileNetV2 уменьшается с 13,4 МБ до 3,4 МБ, что делает её гораздо более компактной и подходящей для использования в условиях ограниченных вычислительных ресурсов. Для модели YOLOv5s размер также уменьшается с 14,2 МБ до 3,7 МБ, что значительно повышает её мобильность и применимость в реальных системах с ограничениями по памяти. Однако для крупных моделей, таких как BERT-base и DistilBERT, квантование даёт меньший эффект: размер BERT-base уменьшается с 420 МБ до 105 МБ, а DistilBERT – с 250 МБ до 62 МБ.

Время инференса всех моделей сокращается после квантования, что делает их более эффективными для реального применения, где важна скорость обработки (рис. 3).

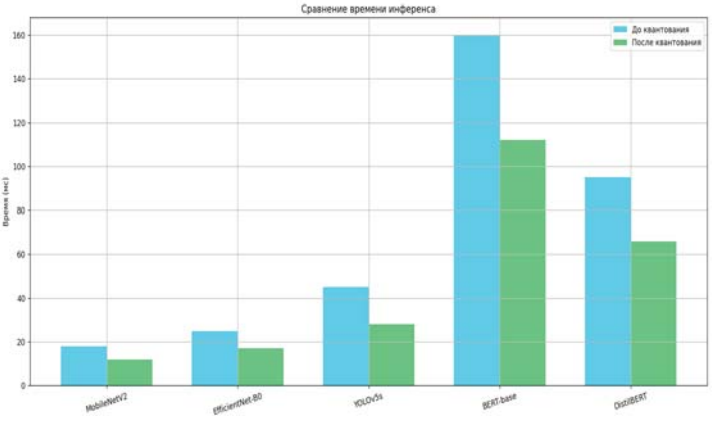

Рис. 3. Сравнение времени инференса

Для MobileNetV2 время инференса уменьшается с 18 мс до 12 мс, что улучшает скорость работы модели в реальных условиях. Для модели YOLOv5s время инференса также снижается с 45 мс до 28 мс, что существенно ускоряет её работу при обработке изображений в реальном времени. Модели BERT-base и DistilBERT также показывают снижение времени инференса, что особенно важно для задач обработки естественного языка, где необходима высокая скорость принятия решений. Эти результаты подчеркивают, что квантование не только уменьшает размер моделей, но и ускоряет их выполнение, что делает их более подходящими для применения в системах с ограниченными вычислительными ресурсами или с требованиями к скорости.

Заключение

В ходе проведённого исследования была продемонстрирована эффективность постобучающего квантования (PTQ) как метода оптимизации нейросетевых моделей для применения в условиях ограниченных вычислительных ресурсов. Экспериментальные результаты показали, что применение квантования позволяет существенно снизить размер модели – в среднем в 3-4 раза и ускорить время инференса на 30-40% без критичного ухудшения точности. Для всех рассмотренных моделей, включая MobileNetV2, EfficientNet-B0, YOLOv5s, BERT-base и DistilBERT, наблюдалось лишь незначительное снижение качества предсказаний (не более 1-1,5%), что подтверждает высокую практическую ценность данного подхода.

Особенно примечательным является тот факт, что крупные модели, такие как BERT-base, демонстрируют значительное снижение как объёма (с 420 МБ до 105 МБ), так и времени инференса (с 160 мс до 112 мс), сохраняя при этом приемлемый уровень точности. Это указывает на высокую адаптивность PTQ даже для трансформерных архитектур, традиционно считающихся ресурсоёмкими.

Полученные результаты подтверждают актуальность использования PTQ в задачах, требующих развёртывания нейросетей на устройствах с ограниченными вычислительными возможностями, включая мобильные приложения, IoT-системы и встраиваемые решения.

Постобучающее квантование доказало свою эффективность как баланс между простотой, производительностью и точностью. С развитием специализированных аппаратных ускорителей (например, TPU, NPU) и оптимизированных фреймворков (TensorFlow Lite, ONNX Runtime) PTQ станет ещё более востребованным инструментом разработчиков ИИ и исследователей.

Литература

1. *Козлова И.А., Дмитриева Т.И.* Постобучающее квантование сверточных нейронных сетей: особенности и перспективы // Информационные технологии и вычислительные системы. 2021. №6. С. 51-57.

2. *Жукова А.А., Пожидаев А.А.* Анализ методов квантования весов нейронной сети // Вестник Рязанского государственного радиотехнического университета. 2023. №4. С. 37-45.

3. *Быков В.Н., Чистяков К.М.* Использование структурной обрезки для ускорения вывода моделей нейронных сетей // Труды СПИИРАН. 2022. Т. 21, №2. С. 102-113.

4. *Мельников А.А., Соловьёв А.С.* Квантование нейросетей для развёртывания на мобильных устройствах // Современные информационные технологии и ИТ-образование. 2020. №16(1). С. 94-100.

5. *Горячев А.А., Иванова А.С.* Сравнительный анализ методов оптимизации нейросетей в задачах классификации изображений // Молодёжный научный вестник. 2022. №4. С. 12–18.

6. *Алексеев П., Квятковская И.Ю.* Применение нейросетей для распознавания условных графических обозначений радиоэлементов // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2021. №2. С. 47-56.

7. *Меньшиков Д.В., Преображенский А.П.* Сравнительный анализ результатов решения задачи определения тональности текста с использованием сверточных и рекуррентных нейронных сетей // Моделирование, оптимизация и информационные технологии. 2021. Т. 9, №4. С. 1-12.

8. *Алисултанова Э.Д., Моисеенко Н.А., Тасуев У.Р., Юсупова Р.В.* Система оптимизации работы клиентских служб на основе нейронных сетей // ТЭК России. 2021. №12(1). С. 15-20.

9. *Воронцов И.* Прорыв в обучении бинарных нейронных сетей: новый метод квантования обеспечивает их стабильность и высокое качество // Компьютерная оптика. 2024. Опубликовано: 27.12.2024. URL: https://zanauku.mipt.ru/2024/12/27/proryv-v-obuchenii-binarnyh-nejronnyh-setej-novyj-metod-kvantovaniya-obespechivaet-ih-stabilnost-i-vysokoe-kachestvo/

10. *Алексанян Г.К.* Экспериментальная проверка модели информационно-измерительной системы для мониторинга регионального вентиляционно-перфузионного соотношения легких человека // ТЭК России. 2021. №12(1). С. 9-14.

11. *Иванов С.В., Петров Д.А.* Методы сжатия нейронных сетей для встраиваемых систем // Журнал вычислительных технологий. 2022. Т. 27, №3. С. 45-53.

12. *Смирнова Е.Л., Кузнецов М.Н.* Энергоэффективные архитектуры нейронных сетей: обзор // Автоматика и телемеханика. 2023. Т. 84, №5. С. 789-798.

13. *Фёдоров А.Б., Никитин П.С.* Современные методы квантования для моделей глубинного обучения // Искусственный интеллект и принятие решений. 2024. №2. С. 34-42.

14. *Соколова Т.В., Лебедев К.И.* Стратегии прореживания сверточных нейросетей в задачах распознавания изображений // Распознавание образов и анализ изображений. 2021. Т. 31, №4. С. 567-574.

15. *Морозов Л.А., Егорова Н.С.* Развёртывание квантованных нейросетей на ПЛИС // Микропроцессоры и системы. 2023. Т. 89. С. 103-110.

16. *Кузьмина О.П., Тарасов В.В.* Сравнительное исследование методов постобучающего квантования // Нейрокомпьютинг. 2022. Т. 503. С. 123-130.

17. *Николаев Ю.Д., Сидоров А.Л.* Влияние квантования на производительность моделей глубинного обучения в задачах обработки текстов // Компьютерная лингвистика и интеллектуальные технологии. 2023. №1. С. 89-97.

18. *Орлова М.В., Денисов Р.К.* Обучение с учетом квантования для повышения устойчивости моделей // Журнал исследований искусственного интеллекта. 2024. Т. 67. С. 210-219.

19. *Павлов Д.С., Зайцев А.Н.* Методы уменьшения размера модели без существенной потери точности // Машинное обучение и анализ данных. 2021. Т. 7, №2. С. 145-152.

20. *Климов В.Е., Сорокин А.И.* Оценка влияния квантования на интерпретируемость моделей // Исследования когнитивных систем. 2022. Т. 70. С. 56-63.

ANALYZING THE EFFECTIVENESS OF POST-LEARNING QUANTIZATION FOR OPTIMIZING NEURAL NETWORKS

TATIANA M. TATARNIKOVA[1]
Saint Petersburg, Russia

ANASTASIA S. RASKOPINA[2]
Saint Petersburg, Russia

KEYWORDS: neural networks, post-learning quantization, model optimization, pruning, quantization, knowledge distillation, acceleration of inference

ABSTRACT

Introduction: Technologies are developing rapidly, including neural networks. After the advent of deep learning, the models became more complex and deeper every year, which led to a shortage of hardware. The article discusses modern methods for optimizing neural networks with an emphasis on post-learning quantization as the most practical approach for deploying models in conditions of limited computing resources. **Methods:** An overview of key methods, including pruning, quantization, and distillation of knowledge, is presented, and their effectiveness and applicability are compared. Special attention is paid to the advantages and limitations of PTQ, such as model size reduction, faster inference, and compatibility with industrial frameworks. The experimental part presents the results of quantization of MobileNetV2, BERT-base, YOLOv5s, EfficientNet-B0, and DistilBERT models, and analyzes the effect of quantization on the accuracy, speed, and compactness of the models. **The results** showed that post-learning quantization does an excellent job. This method was able to reduce the size of the model by 3 times, accelerate the inference by up to 40% and lose no more than 1.5% accuracy. The results obtained can become the basis for further research on neural network optimization, combining the quantization method with other methods, and creating new hybrid methods that will take all the advantages of post-learning quantization and offset the disadvantages. After all, post-learning quantization is especially effective for mobile and IoT devices, where energy consumption and memory requirements are critical. And its use for computer vision and natural language processing tasks is already showing applicability and prospects.

REFERENCES

[1] I.A. Kozlova, T.I. Dmitrieva, “Post-training quantization of convolutional neural networks: features and prospects”, *Information Technologies and Computing Systems*, 2021, No. 6, pp. 51-57. (In Russian)

[2] A.A. Zhukova, A.A. Pozhidaev, “Analysis of methods for quantizing neural network weights”, *Bulletin of the Ryazan State Radio Engineering University*, 2023, No. 4, pp. 37-45. (In Russian)

[3] V.N. Bykov, K.M. Chistyakov, “Using structural pruning to accelerate inference of neural network models”, *Proceedings of SPIIRAS*, 2022, Vol. 21, No. 2, pp. 102-113. (In Russian)

[4] A.A. Melnikov, A.S. Solovyev, “Quantization of neural networks for deployment on mobile devices”, *Modern Information Technologies and IT Education*, 2020, No. 16(1), pp. 94-100. (In Russian)

[5] A.A. Goryachev, A.S. Ivanova, “Comparative analysis of neural network optimization methods in image classification tasks”, *Youth Scientific Bulletin*, 2022, No. 4, pp. 12-18. (In Russian)

[6] P. Alekseev, I.Yu. Kvyatkovskaya, “Application of neural networks for recognizing schematic electrical symbols”, *Bulletin of the Astrakhan State Technical University. Series: Control, Computing Technology and Informatics*, 2021, No. 2, pp. 47-56. DOI: 10.24143/2072-9502-2021-2-47-56. (In Russian)

[7] D.V. Menshikov, A.P. Preobrazhensky, “Comparative analysis of results obtained in solving the task of text sentiment analysis using convolutional and recurrent neural networks”, *Modeling, Optimization and Information Technologies*, 2021, Vol. 9, No. 4, pp. 1-12. DOI: 10.26102/2310-6018/2021.35.4.012. (In Russian)

[8] E.D. Alisultanova, N.A. Moiseenko, U.R. Tasuev, R.V. Yusupova, “System for optimizing the work of client services based on neural networks”, *Fuel and Energy Complex of Russia*, 2021, No. 12(1), pp. 15-20. (In Russian)

[9] I. Vorontsov, “Breakthrough in training binary neural networks: a new quantization method ensures their stability and high quality”, *Computer Optics*, 2024. Published: 27.12.2024. Available at: https://zanauku.mipt.ru/2024/12/27/proryv-v-obuchenii-binarnyh-nejronnyh-setej-novyj-metod-kvantovaniya-obespechivaet-ih-stabilnost-i-vysokoe-kachestvo/. (In Russian)

[10] G.K. Aleksanyan, “Experimental testing of an information-measuring system model for monitoring regional ventilation-perfusion ratio of human lungs”, *Fuel and Energy Complex of Russia*, 2021, No. 12(1), pp. 9-14. (In Russian)

[11] S.V. Ivanov, D.A. Petrov, “Methods of neural network compression for deployment in embedded systems”, *Journal of Computational Technologies*, 2022, Vol. 27, No. 3, pp. 45-53. (In Russian)

[12] E.L. Smirnova, M.N. Kuznetsov, “Energy-efficient neural network architectures: a review”, *Automation and Remote Control*, 2023, Vol. 84, No. 5, pp. 789-798. (In Russian)

[13] A.B. Fedorov, P.S. Nikitin, “Advances in quantization techniques for deep learning models”, *Artificial Intelligence and Decision Making*, 2024, No. 2, pp. 34-42. (In Russian)

[14] T.V. Sokolova, K.I. Lebedev, “Pruning strategies for convolutional neural networks in image recognition tasks”, *Pattern Recognition and Image Analysis*, 2021, Vol. 31, No. 4, pp. 567-574. (In Russian)

[15] L.A. Morozov, N.S. Egorova, “Deployment of quantized neural networks on FPGA platforms”, *Microprocessors and Microsystems*, 2023, Vol. 89, pp. 103-110. (In Russian)

[16] O.P. Kuzmina, V.V. Tarasov, “Comparative study of post-training quantization methods”, *Neurocomputing*, 2022, Vol. 503, pp. 123-130. (In Russian)

[17] Y.D. Nikolaev, A.L. Sidorov, “Impact of quantization on the performance of deep learning models in NLP tasks”, *Computational Linguistics and Intellectual Technologies*, 2023, No. 1, pp. 89-97. (In Russian)

[18] M.V. Orlova, R.K. Denisov, “Quantization-aware training for improving model robustness”, *Journal of Artificial Intelligence Research*, 2024, Vol. 67, pp. 210-219. (In Russian)

[19] D.S. Pavlov, A.N. Zaitsev, “Techniques for reducing model size without significant accuracy loss”, *Machine Learning and Data Analysis*, 2021, Vol. 7, No. 2, pp. 145-152. (In Russian)

[20] V.E. Klimov, A.I. Sorokin, “Evaluation of quantization effects on model interpretability”, *Cognitive Systems Research*, 2022, Vol. 70, pp. 56-63. (In Russian)

INFORMATION ABOUT AUTHORS:

Tatyana M. Tatarnikova, *PhD, Professor, St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, tm-tatarn@yandex.ru*

Anastasia S. Raskopina, *postgraduate student, St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, raskopina.anastasia@yandex.ru*

For citation: T.M. Tatarnikova, A.S. Raskopina, “Analyzing the effectiveness of post-learning quantization for optimizing neural networks,” *H&ES Reserch*. 2025. Vol. 17. No. 2, pp. 4-10. doi: 10.36724/2409-5419-2025-17-2-4-10 (In Rus)