**УДК 004.8** **DOI: 10.18413/2518-1092-2025-10-4-0-6**

**Хрупин Д.С., Шапцев В.А.**

# МЕТОД КВАНТОВАНИЯ НЕЙРОННЫХ СЕТЕЙ ОБНАРУЖЕНИЯ НА ВСТРАИВАЕМЫХ СИСТЕМАХ

Тюменский государственный университет,
ул. Володарского, 6, г. Тюмень, 625003, Россия

*e-mail: Khrupin24@mail.ru, vashaptsev@ya.ru*

**Аннотация**
Квантование моделей является ключевым методом развертывания высокопроизводительных нейросетевых детекторов объектов на устройствах с ограниченными ресурсами. Однако стандартные подходы к квантованию, такие как PTQ, QAT и даже методы смешанной точности, оптимизируют распределение числа битов по показателю чувствительности слоев, игнорируя семантическую специфику задачи. Это приводит к существенному снижению точности при различении семантически близких классов, что критично для многих практических приложений. В статье предложен новый подход к квантованию со смешанной точностью, который учитывает семантику задачи. Введена метрика семантической значимости компонентов сети, вносящих ключевой вклад в различение трудноразличимых классов. На её основе формируется гетерогенная конфигурация битности, которое обеспечивает высокую точность критически важных частей модели, допуская агрессивное сжатие остальных. Представлен план экспериментальной валидации подхода на задаче определения типа транспортного средства. Ожидается значительно лучший компромисс между точностью и ресурсоемкостью модифицированной нейросетевой модели по сравнению со стандартными техниками квантования.

**Ключевые слова**: квантование нейронных сетей; распознавание объектов; встраиваемые системы; глубокое обучение; сжатие моделей; адаптивная длина двоичных значений коэффициентов; смешанная точность; метрика семантической значимости

**Для цитирования:** Хрупин Д.С., Шапцев В.А. Метод квантования нейронных сетей обнаружения на встраиваемых системах // Научный результат. Информационные технологии. – Т.10, №4, 2025. – С. 70-76. DOI: 10.18413/2518-1092-2025-10-4-0-6

**Khrupin D.S.
Shaptsev V.A.**

# QUANTIZATION METHOD FOR DETECTION NEURAL NETWORKS ON EMBEDDED SYSTEMS

Tyumen State University, 6 Volodarsky St., Tyumen, 625003, Russia

*e-mail: Khrupin24@mail.ru, vashaptsev@ya.ru*

**Abstract**
Model quantization is a key method for deploying high-performance neural network object detectors on resource-constrained devices. However, standard quantization approaches, such as PTQ, QAT, and even mixed-precision methods, optimize the distribution of bits based on the sensitivity of layers, ignoring the semantic specificity of the task. This leads to a significant decrease in accuracy when distinguishing between semantically similar classes, which is critical for many practical applications. The article proposes a new approach to mixed-precision quantization that takes into account the semantics of the task. A metric of semantic significance of network components that make a key contribution to the discrimination of difficult-to-distinguish classes is introduced. Based on it, a heterogeneous bit configuration is formed, which ensures high accuracy of critically important parts of the model, allowing aggressive compression of the rest. A plan for experimental validation of the approach on the task of determining the type of vehicle is

presented. A significantly better compromise between accuracy and resource intensity of the modified neural network model is expected compared to standard quantization techniques.
**Keywords:** neural network quantization; object recognition; embedded systems; deep learning; model compression; adaptive binary coefficient length; mixed precision; semantic significance metric
**For citation:** Khrupin D.S., Shaptsev V.A. Quantization Method for Detection Neural Networks on Embedded Systems // Research result. Information technologies. – T.10, №4, 2025. – P. 70-76.
DOI: 10.18413/2518-1092-2025-10-4-0-6

## *ВЕДЕНИЕ*

Глубокие нейронные сети продемонстрировали высокую эффективность в задачах компьютерного зрения, включая обнаружение объектов [1]. Такие архитектуры, как YOLO [2], SSD [3] и EfficientDet [4], стали стандартом де-факто. Однако высокая вычислительная нагрузка и потребность значительного объема памяти ограничивают их применение на встраиваемых устройствах, где ресурсы ограничены [5]. Для преодоления этого барьера развиваются методы оптимизации нейронных сетей, среди которых – «квантование» [6, 7]: дискретизация (целочисленное представление) как входных значений, так и весов входов слоёв. Оно позволяет существенно сократить размер модели, снизить требования к пропускной способности памяти и ускорить вычисления посредством низкоточной целочисленной арифметики современных процессоров и специализированной электроники (NPU, TPU) [8].

Применяемые в задачах классификации изображений методы квантования приводят к существенной потере точности в более сложных случаях, таких как обнаружение объектов [9]. Это связано с несколькими факторами.

***Больший динамический диапазон активаций***. В сетях обнаружения наблюдается большой разброс значений входных признаков; особенно в слоях, отвечающих за предсказание координат и размеров ограничивающих рамок. [10]

***Чувствительность различных частей сети***. Компоненты сети имеют разную чувствительность к погрешности квантования. Так регрессионные оценки координат могут быть более чувствительны, чем данные классификации [11].

***Сложность задачи.*** Обнаружение требует одновременно и локализацию, и классификацию объектов. Это делает итоговую метрику (например, mean Average Precision, mAP), чувствительную к ошибкам этих компонентов. [12]

Ведущиеся исследования предлагают разные стратегии квантования. Post-Training Quantization (постобучающее преобразование модели) привлекательно простотой: не требует переобучения модели; но часто приводит к заметному падению точности, особенно при использовании коротких двоичных последоваетльностей (ниже INT8) [13, 14]. Для компенсации этой потери предложен Quantization-Aware Training (QAT). Он «имитирует» квантование во время обучения, позволяя сети адаптироваться к низкоточной арифметике [15]. QAT обеспечивает лучшую точность, но требует доступ к полному набору данных обучения, значительных вычислительных ресурсов и памяти в переобучении.

Кроме базовых методов Post-Training Quantization (PTQ) и QAT, исследованы: смешанный (mixed-precision), где разные слои квантуются с разной битностью [16,17]; метод адаптивного выбора параметров квантования [18]; аппаратно-ориентированное квантование, учитывающее специфику электронной платформы [19].

Также стоит отметить, что даже продвинутые методы смешанной точности, такие как HAWQ [19] или HAQ [16], хотя и представляют собой шаг вперед по сравнению с однородным квантованием, все же имеют фундаментальное ограничение. Они основывают свои стратегии распределения битности на низкоуровневых, не зависящих от семантики задачи метриках. Например, HAWQ использует информацию из матрицы вторых производных функции потерь для оценки чувствительности слоев, предполагая, что слои с большей кривизной поверхности потерь более чувствительны к ошибкам квантования. Такой подход, будучи математически обоснованным,

не делает различий между ошибками, которые ведут к путанице между классами "автобус" или "велосипед", и «седан» или "универсал". С точки зрения гессиана, это могут быть равнозначные по величине ошибки. Предлагаемый подход исходит из предположения, что для прикладной задачи эти ошибки имеют совершенно разную цену, и модель оптимизации должна это учитывать.

Ниже предложена тактика квантования в задачах определения типов транспортных средств (ТС), обеспечивающая лучший баланс точности обнаружения и требуемых ресурсов.

## *МАТЕРИАЛЫ И МЕТОДЫ (ОБЗОР)*

Наиболее распространенной схемой является аффинное равномерное квантование [20]. Рассмотрим базовый пример квантования по стратегии PTQ.

Для заданного тензора $\boldsymbol{r}$ с вещественными значениями (FP32), его квантованное (целочисленное, INTN) представление $\boldsymbol{q}$ получается масштабированием и сдвигом:

$$\boldsymbol{r} = \boldsymbol{s} \cdot (\boldsymbol{q} - \boldsymbol{z}). \quad (1)$$

Здесь $\boldsymbol{q} = \boldsymbol{round}(\frac{r}{s} + \boldsymbol{z})$ – операция округления, $\boldsymbol{s}$ – шаг квантования, $\boldsymbol{z}$ – целочисленная точка нуля (zero-point). Параметры $\boldsymbol{s}$ и $\boldsymbol{z}$ определяют диапазон квантования $[\boldsymbol{s}(\boldsymbol{q\_min} - \boldsymbol{z}), \boldsymbol{s}(\boldsymbol{q\_max} - \boldsymbol{z})]$, где $\boldsymbol{q}$_min и $\boldsymbol{q}$_max – минимальное и максимальное значения для выбранного типа INTN (например, 128 и 127 для INT8). Эти параметры могут вычисляться как один набор ($\boldsymbol{s}$, $\boldsymbol{z}$) для всего тензора (Per-tensor) или как отдельный набор ($\boldsymbol{s}$, $\boldsymbol{z}$) для каждого канала сверточного слоя или каждой строки/столбца весовой матрицы (Per-channel/ Per-axis).

В таблице представлены данные о 3-х методиках квантования. Каждый имеет свой компромисс «точность – простота – вычислимость».

*Таблица*

Характеристика основных стратегий квантования нейронных сетей

*Table*

Characteristics of the main neural network quantization strategies

| **Параметр** | **Post-Training Quantization** | **Quantization-Aware Training** | **Mixed-Precision Quantization** |
|---|---|---|---|
| Базовый принцип | Квантование предобученной FP32-модели. | Симуляция эффектов квантования во время обучения/дообучения. | Назначение разной битовой ширины разным слоям/компонентам сети. |
| Этап применения | После обучения. | Во время обучения или дообучения. | Битность - после обучения; финальная настройка - часто QAT. |
| Входные требования | Обученная FP32-модель; небольшой калибровочный датасет (для статич.). | Обучающий датасет; архитектура модели; тренировочный пайплайн. | Обученная FP32-модель; агоритм битности; калибр. /обуч. датасет. |
| Ожидаемая точность (vs FP32) | Умеренное (INT8), значительное (<INT8) падение . Статич. PTQ < Динам. PTQ / QAT. | Минимальное падение (INT8); лучшая устойчивость к <INT8 (из трех). | Лучшее соотношение точность/эффективность; зависит от стратегии и QAT. |
| Сложность реализации | Низкая (особенно статическое PTQ). | Высокая (модификация обучения, гиперпараметры). | Очень высокая (алгоритм поиска битности, возможно сложный QAT). |

| | | | |
|---|---|---|---|
| Затраты на применение метода | Низкие (1 проход калибровки для статич.). | Высокие (полное обучение/дообучение). | Высокие/очень высокие (анализ чувствительн./ поиск + опц. QAT). |
| Влияние на производительность вывода | Статич.: значит. ускорение, низкая задержка. Динамич.: меньшее ускорение из-за рантайм-вычис. S/Z. | Значительное ускорение, низкая задержка (параметры фиксированы). | Зависит от распределения битности и поддержки аппаратурой; потенциально макс. ускорение. |
| Гибкость битовой ширины | Обычно однородная (INT8/INT4 и т.д.). | Обычно однородная, можно адаптировать для смешанной. | Максимальная (основная идея метода). |
| Устойчивость к <INT8 | Низкая. | Лучшая из трех стандартных подходов. | Агрессивное квантование нечувствительных частей; устойч. опред. стратегией. |
| Аппаратные особенности | Статич.: хорошо с акселераторами. Динамич.: менее эффективно аппаратно. | Хорошо поддерживается акселераторами (вывод статичен). | Требует гибкой аппаратно-программную поддержку эффективного исполнения. |
| Типичный сценарий применения | Быстрое развертывание; небольшие потери точности; нет доступа к обучающим данным/пайплайну. | Требуется максим. точность при низкой битности; доступны ресурсы для обучения. | Достижение оптимального баланса точность-ресурсы; таргетирование специфичных аппаратных бюджетов. |
| Ключевые преимущества | Простота, скорость внедрения, не требует переобучения. | Высокая точность, лучшая адаптация сети к квантованию. | Наилучший компромисс точность-эффективность; гранулярный контроль. |
| Ключевые ограничения | Потеря точности (особенно <INT8); зависимость статич. PTQ от калибровки; оверхед динам. PTQ. | Сложность реализации. Высокие вычислительные затраты на обучение; полный датасет. | Сложность распределения битности; зависимость от аппаратной поддержки; высокие затраты на настройку. |
| Возможность комбинации | - | Для финальной настройки MPQ. | Комбинируется с QAT для лучшей точности. |

## *ОБСУЖДЕНИЕ ТАБЛИЦЫ*

Существующие методы квантования, включая подходы со смешанной точностью, как правило, распределяют битовую ширину, опираясь на общие характеристики слоев (глубина, диапазон значений) или на оценки их чувствительности к ошибке квантования, часто измеряемой как степень влияния на общую точность модели или на изменение весов/активаций. Хотя эти подходы эффективны для сжатия ресурсной потребности, они игнорируют семантику решаемой

задачи. [21] А специфика задачи может иметь дополнительные факторы эффективной работы нейросети.

В задаче детекции и классификации типов транспортных средств встречаются как легко различимые категории (легковой автомобиль и автобус), так и семантически близкие классы (седан и хэтчбек, фургон и микроавтобус, типы грузовиков). Успешное различение последних требует сохранения тонких, высокоуровневых признаков, извлекаемых нейронной сетью. Стандартные стратегии квантования, не учитывающие эту семантическую особенность, могут непропорционально ухудшить способность модели различать именно эти, иногда критически важные, но трудноразличимые классы.

Сформулируем гипотезу: тактика квантования, адаптивно назначающая большую битовую ширину тем компонентам нейронной сети, которые вносят наибольший вклад в различение семантически сложных или критически важных для задачи классов ТС, даёт наилучший компромисс между точностью классификации и производительностью вычислений. Такой подход учитывает семантику задачи в процессе регулирования параметров квантования. При этом реализуется идентификация и защита от агрессивного квантования тех компонентов сети (слои, каналы, отдельные фильтры), которые наиболее важны для различения близких классов.

## *ЭТАПЫ РЕАЛИЗАЦИИ ПРЕДЛАГАЕМОГО МЕТОДА*

1. ***Идентификация критических классов.*** Выявление пар или групп классов ТС с высоким уровнем взаимной путаницы по матрице ошибок исходной FP32-модели на валидационном наборе данных. Эти классы определяются как "критические". Альтернативно: критические классы могут быть заданы в условии задачи.
2. ***Оценка семантической значимости компонентов.*** Для каждого кандидата на квантование (компонент сети: сверточный слой или канал) вычисляется метрика семантической значимости. Значение метрики отражает важность компонента для верной классификации объектов критических классов (шаг 1).

Основная идея оценки заключается в измерении чувствительности модели к «повреждению» конкретного компонента. В качестве такого повреждения выступает его пробное агрессивное квантование. Если производительность модели на критических классах значительно падает после квантования компонента, значит, он имеет высокую семантическую значимость. В качестве меры производительности целесообразно использовать функцию потерь, так как она более чувствительна к изменениям уверенности модели, чем метрики точности.

Метрику семантической значимости $S_c$ для компонента с можно определить как прирост функции потерь на критическом наборе данных $D_{crit}$ при применении к этому компоненту сильного возмущения (агрессивного квантования):

$$S_c = \mathbb{E}_{(x,y)\in D_{crit}}[L(M'_c(x), y)] - \mathbb{E}_{(x,y)\in D_{crit}}[L(M(x), y)], \quad (2)$$

где:

– $S_c$ – искомая метрика семантической значимости для компонента $\boldsymbol{C}$.

– $M$ – исходная модель (FP32).

– $M'_c$ – модель, в которой к компоненту $\boldsymbol{C}$ применено агрессивное квантование (например, до INT4 или INT2). Остальные компоненты остаются в FP32.

– $D_{crit}$ – валидационный поднабор данных, содержащий только объекты критических, трудноразличимых классов.

– $L$ – функция потерь для задачи классификации.

– $x, y$ – входное изображение и истинная метка класса из набора $D_{crit}$.

– $\mathbb{E}$ – оператор математического ожидания, на практике – усреднение значений функции потерь во всех примерах из набора данных $D_{crit}$.

***Интерпретация метрики:***

Высокое значение $S_c$ указывает на то, что компонент $\boldsymbol{C}$ вносит существенный вклад в различение семантически близких классов. Его «повреждение» приводит к значительному росту ошибки на критических примерах. Такие компоненты являются главными кандидатами на сохранение высокой точности.

Значение $S_c$, близкое к нулю, означает, что данный компонент не является критически важным для данной специфической подзадачи. Следовательно, его можно квантовать более агрессивно, не опасаясь значительного ухудшения способности модели различать сложные классы.

Эта процедура выполняется итеративно для всех рассматриваемых компонентов сети (слоев или каналов), в результате чего формируется своего рода «карта семантической значимости». Эта карта и служит основанием для последующего адаптивного распределения битовой ширины.

3. ***Адаптивное распределение битовой ширины.*** По значениям метрики формируется гетерогенная тактика квантования: компонентам с высокой семантической значимостью назначается большая битовая ширина, менее значимые компоненты квантуются более агрессивно.

4. ***Финальная стадия.*** После определения гетерогенной конфигурации битности для всей модели применяется процедура QAT. На этом этапе в граф вычислений вставляются узлы "fake quantization", которые симулируют процесс квантования и округления для каждой группы компонентов в соответствии с назначенной ей битностью. Модель дообучается в течение нескольких эпох на полном обучающем наборе данных. Это позволяет весам модели тонко подстроиться под ошибки, вносимые квантованием, и компенсировать потенциальную потерю точности, особенно на границах диапазонов квантования.

## *ПЛАН ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ*

1. Формирование обучающего множества данных.
2. Выбор архитектуры нейросетевой модели обнаружения и её обучение.
3. Формирование множества сравниваемых моделей: исходное -{PTQ, QAT, MPQ}.
4. Программная реализация предложенного подхода.
5. Вычисления, визуализация и получение значений метрик. Интерпретация данных эксперимента.

## *ЗАКЛЮЧЕНИЕ*

Обзор методов квантования нейронных сетей и анализ их ограничений привёл к формулированию гипотетического подхода, который учитывает данные о сложности задачи обнаружения транспортного объекта в процессе гетерогенного распределения битовой ширины кванта (длины двоичного кода) в настраиваемых значениях коэффициентов нейросети. Запланирован вычислительный эксперимент по оценке эффективности этого подхода, по проверке гипотезы о высокой эффективности соответствующего метода.

**Список литературы**

**References**

1. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks // arXiv. URL: https://arxiv.org/abs/1506.01497 (дата обращения: 11.04.2025).

2. Redmon J., Farhadi A. YOLOv3: An Incremental Improvement // arXiv. URL: https://arxiv.org/abs/1804.02767 (дата обращения: 11.04.2025).

3. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C., Berg A.C. SSD: Single Shot MultiBox Detector // arXiv. URL: https://arxiv.org/abs/1512.02325 (дата обращения: 11.04.2025)

4. Tan M., Pang R., Le Q.V. EfficientDet: Scalable and Efficient Object Detection // arXiv. URL: https://arxiv.org/abs/1911.09070 (дата обращения: 11.04.2025).

5. Chen Y., Krishna T., Emer J.S., Sze V. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks // ResearchGate. URL:

https://www.researchgate.net/publication/292869497_Eyeriss_An_Energy-Efficient_Reconfigurable_Accelerator_for_Deep_Convolutional_Neural_Networks#references (дата обращения: 11.04.2025).

6. Jacob B., Kligys S., Chen B., Zhu M., Tang M., Howard A., Hartwig A., Kalenichenko D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // arXiv. URL: https://arxiv.org/abs/1712.05877 (дата обращения: 11.04.2025).

7. Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper // arXiv. URL: https://arxiv.org/abs/1806.08342 (дата обращения: 11.04.2025).

8. Jouppi, N. P., et al. In-Datacenter Performance Analysis of a Tensor Processing Unit // arXiv. URL: https://arxiv.org/abs/1704.04760 (дата обращения: 11.04.2025).

9. Ron Banner, Yury Nahshan, Elad Hoffer, Daniel Soudry. Post-training 4-bit quantization of convolution networks for rapid-deployment // arXiv. URL: https://arxiv.org/abs/1810.05723 (дата обращения: 11.04.2025).

10. Yelysei Bondarenko, Markus Nagel, Tijmen Blankevoort. Understanding and Overcoming the Challenges of Efficient Transformer Quantization // arXiv. URL: https://arxiv.org/abs/2109.12948 (дата обращения: 27.08.2025).

11. Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, Kurt Keutzer. ZeroQ: A Novel Zero Shot Quantization Framework // arXiv. URL: https://arxiv.org/abs/2001.00281 (дата обращения: 11.04.2025).

12. Rundong Li, Yan Wang. Fully Quantized Network for Object Detection // ResearchGate. URL: https://www.researchgate.net/publication/334729962_Fully_Quantized_Network_for_Object_Detection (дата обращения: 11.04.2025).

13. Markus Nagel, Mart van Baalen, Tijmen Blankevoort, Max Welling. Data-Free Quantization Through Weight Equalization and Bias Correction // arXiv. URL: https://arxiv.org/abs/1906.04721 (дата обращения: 11.04.2025).

14. Migacz, S. 8-bit inference with TensorRT. // GTC 2017.

15. Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, Yuheng Zou. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients // arXiv. URL: https://arxiv.org/abs/1606.06160 (дата обращения: 11.04.2025).

16. Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, Song Han. HAQ: Hardware-Aware Automated Quantization with Mixed Precision // arXiv. URL: https://arxiv.org/abs/1811.08886 (дата обращения: 11.04.2025).

17. Automatic Mixed Precision package – torch.amp — PyTorch 2.6 documentation // pytorch.org. URL: https://pytorch.org/docs/stable/amp.html (дата обращения: 10.04.2025).

18. Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, Dharmendra S. Modha. Learned Step Size Quantization // arXiv. URL: https://arxiv.org/abs/1902.08153 (дата обращения: 11.04.2025).

19. Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, Kurt Keutzer. HAWQ-V2: Hessian Aware Trace-Weighted Quantization of Neural Networks // arXiv. URL: https://arxiv.org/abs/1911.03852 (дата обращения: 10.04.2025).

20. Sambhav R. Jain, Albert Gural, Michael Wu, Chris H. Dick. Trained Quantization Thresholds for Accurate and Efficient Fixed-Point Inference of Deep Neural Networks // arXiv. URL: https://arxiv.org/abs/1903.08066 (дата обращения: 10.04.2025).

21. Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, Kurt Keutzer. A Survey of Quantization Methods for Efficient Neural Network Inference // arXiv. URL: https://arxiv.org/abs/2103.13630 (дата обращения: 11.04.2025).

**Хрупин Данила Станиславович,** аспирант кафедры информационных систем, Тюменский государственный университет, г. Тюмень, Россия
**Шапцев Валерий Алексеевич,** доктор технических наук, профессор кафедры информационных систем, Тюменский государственный университет, г. Тюмень, Россия

**Khrupin Danila Stanislavovich,** Postgraduate Student of the Information Systems Department, Tyumen State University, Tyumen, Russia
**Shaptsev Valeriy Alekseevich,** Doctor of Technical Sciences, Professor of the Information Systems Department, Tyumen State University, Tyumen, Russia