УДК 004.056
doi:10.31799/1684-8853-2026-5-11-20
EDN: AZRPYT

Научные статьи
Articles

Математическая модель и экспериментальная оценка устойчивости квантованных DS-CNN для TinyML-распознавания речевых команд

М. А. Князев[а], *аспирант, orcid.org/0009-0007-3931-7442, maxiknyaz@mail.ru*
М. М. Гусев[б], *магистрант, orcid.org/0009-0004-7023-4328*
[а]*МИРЭА — Российский технологический университет, Вернадского пр., 78, Москва, 119454, РФ*
[б]*Национальный исследовательский университет «Высшая школа экономики», Таллинская ул., 34, Москва, 123458, РФ*

Введение: *квантование позволяет размещать нейронные сети распознавания ключевых слов на микроконтроллерах, но дискретизация весов и активаций изменяет классификационный запас и чувствительность модели к возмущениям.* ***Цель:*** *разработать математическую модель, связывающую ошибку квантования, классификационный запас, локальную устойчивость и ограничения развертывания, и применить ее для сравнения режимов численного представления.* ***Методы:*** *на бинарной подзадаче "yes/no" набора Google Speech Commands v2 сравнивались сверточные сети с разделением по глубине: исходная сеть с 32-разрядными числами с плавающей точкой и сети с восьми- и четырехразрядным целочисленным квантованием после обучения либо с учетом квантования при обучении. Использовались адаптивные атаки проецируемого градиентного спуска в пространстве признаков, акустические искажения с контролем отношения сигнал/шум, пороговый допуск, парето-анализ и интегральный показатель.* ***Результаты:*** *восьмиразрядное квантование после обучения сохранило близкую к исходной точность на чистых данных (95,67 против 95,83 %) и при атаке (90,70 против 91,59 %). Обучение с учетом восьмиразрядного квантования обеспечило соответственно 95,87 и 90,78 % при коэффициенте сохранения точности 0,947. Четырехразрядное квантование после обучения снизило исходную точность до 63,31 %, а обучение с учетом квантования восстановило ее до 88,55 %. Максимальный интегральный показатель 0,5862 получен для восьмиразрядной сети, обученной с учетом квантования, при медиане задержки 41,53 мс на Arduino Nano 33 BLE Sense. При отношении сигнал/шум −5 дБ целевая аугментация этой сети повысила точность с 50,85 до 67,75 %, сократив долю ошибок на 34,4 %.* ***Практическая значимость:*** *предложенная модель позволяет количественно выбирать режим квантования с учетом измеренных показателей качества и задержки при отдельном контроле памяти всей прошивки.*

Ключевые слова *— TinyML, распознавание ключевых слов, DS-CNN, обучение с учетом квантования, квантование после обучения, INT8, INT4, состязательные атаки, PGD, акустические искажения, KWS.*

Для цитирования: Князев М. А., Гусев М. М. Математическая модель и экспериментальная оценка устойчивости квантованных DS-CNN для TinyML-распознавания речевых команд. *Информационно-управляющие системы*, 2026, № 5, с. 11–20. doi:10.31799/1684-8853-2026-5-11-20, EDN: AZRPYT

For citation: Knyazev M. A., Gusev M. M. Mathematical model and experimental robustness assessment of quantized DS-CNNs for TinyML keyword spotting. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 5, pp. 11–20 (In Russian). doi:10.31799/1684-8853-2026-5-11-20, EDN: AZRPYT

Введение

Локальное распознавание ключевых слов (Keyword Spotting, KWS) используется в голосовых интерфейсах носимых устройств, умных колонок и других компонентов персонального интернета вещей. Перенос обработки на устройство уменьшает сетевую задержку и объем передаваемых речевых данных, но требует одновременного соблюдения ограничений постоянной и оперативной памяти, времени вывода и энергопотребления. В машинном обучении на микроконтроллерах (Tiny Machine Learning, TinyML) эти требования обычно выполняются сочетанием компактной архитектуры и целочисленного представления модели [1–10].

Квантование влияет не только на размер модели. Округление весов и активаций распространяется через слои, изменяет логиты, классификационный запас и положение локальной границы решений. Кроме того, дискретность вычислений меняет градиент, используемый при построении состязательных примеров [11–14]. Поэтому вывод об устойчивости нельзя делать только по разрядности: требуется модель, одновременно учитывающая ошибку квантования, локальную чувствительность, тип атаки и аппаратный бюджет.

Метод быстрого знакового градиента (Fast Gradient Sign Method, FGSM) и проецируемый градиентный спуск (Projected Gradient Descent, PGD) остаются базовыми средствами оценки

устойчивости моделей белого ящика [15, 16]. Для сетей с недифференцируемыми или дискретными преобразованиями применяются адаптивные способы обратного прохода и независимые атаки, позволяющие отделить реальную устойчивость от эффекта скрытого градиента [12, 17–19]. Для аудиосистем к этому добавляется различие между воздействием в пространстве признаков и искажением исходного сигнала: одинаковое обозначение бюджета возмущения в этих пространствах не имеет единого физического смысла [20–24].

Методы распознавания речевых команд на устройствах с ограниченными ресурсами развиваются в направлении компактных сверточных, остаточных и импульсных архитектур [1–5, 9, 10]. TFLite Micro обеспечивает исполнение целочисленных моделей на встроенных платформах, а стандартные схемы INT8 (восьмиразрядного целочисленного квантования) используют квантованные веса и активации с накоплением произведений в расширенном целочисленном формате [7, 8]. PTQ (Post-Training Quantization) применяет квантование после обучения, тогда как QAT (Quantization-Aware Training) вводит имитацию квантования в обучающий цикл и позволяет параметрам адаптироваться к ошибкам дискретизации [8, 11].

Влияние квантования на состязательную устойчивость неоднозначно. При небольшом изменении логитов INT8-модель может сохранять границу решений FP32 (32-разрядной модели с плавающей точкой), но грубая дискретизация способна исчерпать классификационный запас. Одновременно ступенчатые активации могут затруднять обычную градиентную атаку, не меняя фактической уязвимости [12–14, 17]. По этой причине современные протоколы сочетают многошаговые атаки, адаптивный обратный проход и независимые алгоритмы проверки [18, 19, 25]. Российские исследования также рассматривают состязательные воздействия, перенос атак и механизмы защиты моделей машинного обучения [26–32].

В речевых системах цифровые атаки в пространстве признаков и акустические искажения отвечают разным моделям угроз [20, 21]. Белый шум, изменение высоты тона и темпа характеризуют эксплуатационную устойчивость, если параметры преобразования не оптимизируются по функции потерь. Запросные и переносимые атаки, напротив, моделируют нарушителя с неполным знанием целевой сети [23, 24]. Такое разделение используется далее при интерпретации результатов.

Цель работы заключается в том, чтобы разработать математическую модель оценки состязательной устойчивости квантованных TinyML-аудиосистем и применить ее для сравнения пяти конфигураций сверточной сети с разделением по глубине (Depthwise Separable Convolutional Neural Network, DS-CNN): FP32, INT8-PTQ, INT8-QAT, INT4-PTQ и INT4-QAT (INT4 — четырехразрядное целочисленное квантование). Научная новизна заключается в совместном использовании оценки классификационного запаса, эмпирических показателей устойчивости, порогового допуска и парето-анализа. Практический результат состоит в количественном сопоставлении точности, устойчивости и размера модели с учетом измеренной задержки на микроконтроллере.

Данные и методика оценки устойчивости

Данные, признаки и архитектура

Эксперимент проводился на бинарной подзадаче распознавания команд “yes” и “no” набора Google Speech Commands v2 [6]. Обучающая, валидационная и тестовая части включали соответственно 6358, 803 и 824 записей. Частота дискретизации составляла 16 кГц, длительность входа — 1 с. Из аудиосигнала строилась логарифмическая мел-спектрограмма с 40 фильтрами, окном 32 мс и шагом 16 мс. Использовалось разбиение без общих дикторов; нормализация рассчитывалась по обучающей части. Признаки поступали на вход компактной DS-CNN с 4514 параметрами.

Для исключения различий исходных FP32-моделей выполнена согласованная серия из трех запусков. Сеть обучалась 50 эпох оптимизатором Adam с начальной скоростью обучения 10^{-3}. В каждом из трех запусков FP32 и варианты с QAT инициализировались одним выбранным состоянием FP32 и дообучались по 10 эпох Adam с шагом 10^{-4} и одинаковым порядком примеров. PTQ применялось к итоговому FP32; состояния выбирались по точности валидации. При обучении и оценке устойчивости в PyTorch 2.8.0 INT8/INT4 обозначают сетки W8A8/W4A4 (8/4 бита для весов и активаций соответственно); свертки и накопление выполнялись в FP32. Для аппаратного измерения задержки использованы преобразованные модели TensorFlow Lite. INT8 и INT4 выполнялись целочисленными ядрами, а контрольная модель FP32 была выполнена с вещественной арифметикой.

Воздействия и показатели

Основной сценарий состязательного воздействия реализован в пространстве признаков. Использовался ансамбль PGD со случайной инициализацией, а именно 40/100 шагов с кросс-

энтропией и 100 шагов с разностью неверного и верного логитов; шаги ε/20, ε/10 и ε/4 соответственно, по три старта. Основное сравнение проводилось при ε = 0,03 в норме $L\infty$ без ограничения признаков интервалом [0, 1]. Прямой проход сохранял округление, обратный использовал сквозную оценку градиента (Straight-Through Estimator, STE) с единичной производной внутри диапазона квантования и нулевой вне его. Учитывалось объединение исходных ошибок и ошибок всех итераций. На общей стратифицированной подвыборке из 128 записей первого запуска (по 64 каждого класса) дополнительно проверялись обычный градиент, перенос с FP32 и стохастическая аппроксимация с одновременным возмущением (Simultaneous Perturbation Stochastic Approximation, SPSA) [12, 17–19]. Диагностические PGD-40/100 и перенос PGD-100 с FP32 использовали разность логитов, шаг ε/10 и три старта. SPSA выполнялась за 128 шагов по 32 двусторонних направления с радиусом ε/4 и шагом Adam ε/20: 8323 запроса на пример; учитывались ошибки всех пробных точек. Дополнительный блок в пространстве сигнала включал аддитивный белый гауссовский шум на сетке отношения сигнал/шум (Signal-to-Noise Ratio, SNR) от 30 до −5 дБ.

Для каждой конфигурации определялись точность на исходных данных Acc_{clean}, точность после атаки Acc_{adv}, абсолютное снижение точности ΔAcc и коэффициент сохранения R_{ret}. Ресурсная составляющая содержит полный размер контейнера параметров S, в том числе заголовок и задержку L. Размер контейнера не включает код прошивки и среды выполнения и не равен занятому объему Flash. На Arduino измерялось время полного вызова dscnn_run при пакете 1 и фиксированном тензоре признаков 1 × 40 × 61. После пяти прогревочных вызовов выполнялись 50 измеряемых; для каждого сохранялись отсчеты DWT_CYCCNT до и после вызова. При частоте 64 МГц задержка в миллисекундах равна ΔC/64 000; медиану см. в табл. 2. В интервал включены распаковка весов INT4, свертки, пулинг, классификатор и argmax. Загрузка модели при инициализации, извлечение признаков и обмен через Serial исключены; прерывания оставлены включенными. Сценарные бюджеты составляли $S_{\max} = 1024$ КиБ и $L_{\max} = 100$ мс.

Схема анализа

Сначала сравнивались исходная и состязательная точности. Затем вычислялись нормированные размер и задержка, интегральная полезность и пороговый допуск. Парето-анализ выполнялся по четырем критериям: максимизация Acc_{clean} и Acc_{adv} при минимизации S и L.

Математическая модель

Квантование и классификационный запас

Пусть $x \in \mathbb{R}^{T}$ — нормированный аудиосигнал, $\varphi(x)$ — его частотно-временное представление, а $F_b(x) \in \mathbb{R}^{K}$ — логиты сети в режиме b. Предсказание задается как

$$\hat{y}_b(x) = \arg\max_{k \in \{1, \ldots, K\}} [F_b(x)]_k. \quad (1)$$

Целочисленное кодирование и восстановление вещественного значения определяются раздельно:

$$q_b(r; s, z_0) = clip\left(\left[\frac{r}{s}\right] + z_0, q_{\min}, q_{\max}\right); \quad (2)$$

$$Q_b(r; s, z_0) = s\left(q_b(r; s, z_0) - z_0\right),$$
$$q_{\min} = -q_{\max}, q_{\max} = 2^{b-1} - 1. \quad (3)$$

Послойное распространение квантования имеет вид

$$h_0 = Q^a_{b,0}(\varphi(x)), W_l^{(b)} = Q^w_{b,l}(W_l); \quad (4)$$

$$h_l = Q^a_{b,l}\left[\sigma_l\left(W_l^{(b)} h_{l-1} + b_l^{(b)}\right)\right], F_b(x) = h_L. \quad (5)$$

Для размеченного объекта введем классификационный запас

$$m_b(x, y) = [F_b(x)]_y - \max_{k \neq y} [F_b(x)]_k \quad (6)$$

и ошибку квантования по логитам

$$e_b(x) = \left\| F_b(x) - F_{32}(x) \right\|_\infty. \quad (7)$$

В режиме FP32 квантователи тождественны. Для оценки e_b используется вещественная копия тех же итоговых весов; для QAT это дает возможность отделить квантование от изменения весов при дообучении. В опыте веса и вход использовали симметричные диапазоны кодов (3): [−127; 127] / [−7; 7], выходы ReLU (Rectified Linear Unit — функции выпрямления) и усреднения являются беззнаковыми [0; 255] / [0; 15]. Масштабы весов задавались по максимуму модуля в выходном канале, а масштабы активаций — по максимуму модуля на 512 обучающих примерах и затем фиксировались. В реализации (5) смещения и конечные логиты остаются FP32; активация и квантователь последнего слоя тождественны.

Из определения нормы следует $[F_b]_y \geq [F_{32}]_y - e_b$ и $[F_b]_k \leq [F_{32}]_k + e_b$. Поэтому

$$m_b(x, y) \geq m_{32}(x, y) - 2e_b(x). \quad (8)$$

Соотношение (8) задает механизм влияния разрядности на качество: пока $2e_b(x) < m_{32}(x, y)$, квантование не исчерпывает запас FP32-модели

для данного объекта; при нарушении этого условия класс может измениться уже без внешнего воздействия.

Локальная устойчивость

Обозначение $F_b(z)$ относится к классификатору без повторного извлечения признаков. Радиус $r_b(x)$ определяется в норме L_∞ пространства признаков. Пусть разности логитов $g_{b,yk}(z) = [F_b(z)]_y - [F_b(z)]_k$ локально липшицевы с константой не более $2L_b$:

$$\left|g_{b,yk}(z+\Delta) - g_{b,yk}(z)\right| \le 2L_b \left\|\Delta\right\|. \quad (9)$$

Тогда достаточное условие сохранения класса при $\|\Delta\| \le \varepsilon_z$ и нижняя оценка локального радиуса имеют вид

$$m_b(x, y) > 2L_b\varepsilon_z,\ r_b(x) \ge \frac{m_b(x, y)}{2L_b}. \quad (10)$$

Совместно с (8) получаем

$$r_b(x) \ge \frac{\max\{m_{32}(x, y) - 2e_b(x), 0\}}{2L_b}. \quad (11)$$

Оценки (10), (11) являются условными. При $L_b > 0$ неравенство (9) должно выполняться во всем шаре L_∞, содержащем оцениваемый радиус. Точное округление при пересечении порогов этого не гарантирует. Измеряемый максимум конечных отношений $\frac{|\Delta g|}{2\|\Delta z\|_\infty}$ по PGD-возмущению и случайным направлениям является выборочной оценкой чувствительности, а не ее верхней границей.

Если предварительная обработка локально удовлетворяет

$$\left\|\varphi(x+\delta) - \varphi(x)\right\| \le L_\varphi \left\|\delta\right\|, \quad (12)$$

то для бюджета ε_x в пространстве сигнала достаточно условия

$$m_b(x,y) > 2L_bL_\varphi\varepsilon_x. \quad (13)$$

Таким образом, эффект QAT может проявляться через увеличение запаса m_b, уменьшение локальной чувствительности L_b или их сочетание. Напротив, снижение качества INT4 может соответствовать ситуации, в которой ошибка e_b исчерпывает запас для значительной доли объектов.

Метрики устойчивости

PGD в пространстве признаков задается рекурсией

$$z^{(t+1)} = \Pi_{B_\infty(z,\varepsilon_z)}\left(z^{(t)} + \alpha\,\mathrm{sign}\nabla_z L(F_b(z^{(t)}), y)\right). \quad (14)$$

При атаке в пространстве признаков предсказание в (16) вычисляется непосредственно по атакованному тензору. Для набора из N объектов точности и их производные определяются как

$$Acc_{clean}(b) = \frac{1}{N}\sum_{i=1}^{N} 1[\hat{y}_b(x_i) = y_i]; \quad (15)$$

$$Acc_{adv}(b, A) = \frac{1}{N}\sum_{i=1}^{N} 1[\hat{y}_b(x_i^A) = y_i]; \quad (16)$$

$$\Delta Acc = Acc_{clean} - Acc_{adv},\ R_{ret} = \frac{Acc_{adv}}{Acc_{clean}}. \quad (17)$$

Для оценки аддитивного шума используется энергетическое отношение сигнал/шум

$$SNR_{dB}(x,\delta) = 10\log_{10}\frac{\|x\|_2^2}{\|\delta\|_2^2} = 20\log_{10}\frac{\|x\|_2}{\|\delta\|_2}. \quad (18)$$

Допуск и интегральный показатель

Допустимое множество моделей задается системой порогов

$$F = \left\{ b : \begin{matrix} Acc_{clean}(b) \ge A_{\min},\ R_{ret}(b) \ge R_{\min} \\ S(b) \le S_{\max},\ L(b) \le L_{\max} \end{matrix} \right\}. \quad (19)$$

В расчетах приняты $A_{\min} = 0{,}95$, $R_{\min} = 0{,}90$, $S_{\max} = 1024$ КиБ и $L_{\max} = 100$ мс. Для сравнительного ранжирования определен исходный показатель

$$U_0(b) = 0{,}30Acc_{clean}(b) + 0{,}40Acc_{adv}(b) - 0{,}15\frac{S(b)}{S_{\max}} - 0{,}15\frac{L(b)}{L_{\max}}. \quad (20)$$

Пороговый показатель принимает значение $U(b) = U_0(b)$ для $b \in F$ и $U(b) = -\infty$ в противном случае. Такое разделение не позволяет высокой суммарной оценке компенсировать нарушение обязательного ограничения.

Результаты

Точность и устойчивость в пространстве признаков

Результаты ансамбля PGD при $\varepsilon = 0{,}03$ приведены в табл. 1. Для точности и R_{ret} указаны среднее и стандартное отклонения трех запусков на одних 824 записях; ΔAcc и S представлены средними. INT8-PTQ сохранил близкую к FP32 исходную точность (95,67 против 95,83 %) и изменил точность при атаке на 0,89 п.п. Коэффициенты сохранения также близки: 0,948 и 0,956.

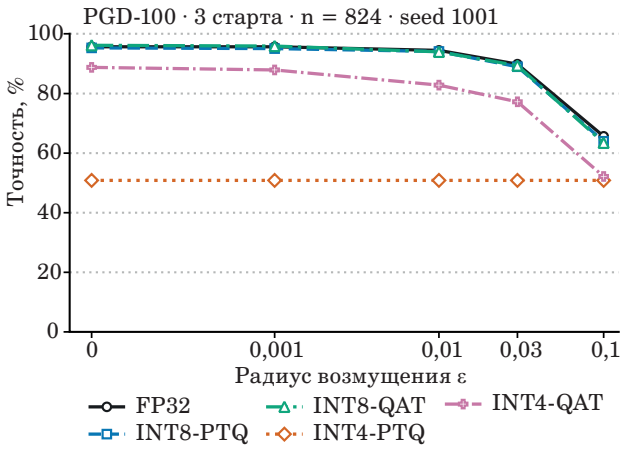
INT8-QAT показал точность при PGD 90,78 %, при этом снижение относительно исходных данных

■ ***Таблица 1.*** Точность DS-CNN при различных режимах квантования
■ ***Table 1.*** DS-CNN accuracy for different quantization modes

Модель	Acc_{clean}, %	Acc_{adv}, %	ΔAcc, п.п.	R_{ret}	Размер, КиБ
FP32	95,83 ± 0,25	91,59 ± 1,54	4,25	0,956 ± 0,014	21,20
INT8-PTQ	95,67 ± 0,37	90,70 ± 1,73	4,98	0,948 ± 0,015	14,91
INT8-QAT	95,87 ± 0,21	90,78 ± 1,58	5,10	0,947 ± 0,019	14,91
INT4-PTQ	63,31 ± 21,58	58,21 ± 12,86	5,10	0,942 ± 0,099	12,90
INT4-QAT	88,55 ± 2,79	75,24 ± 4,43	13,31	0,849 ± 0,023	12,90

составило 5,10 п.п. Парная разность Acc_{adv} между INT8-QAT и INT8-PTQ равна 0,08 ± 0,19 п.п. На общей подвыборке из 128 примеров INT8-QAT дает 99,22 % при обычной PGD-40 через округление и 91,41 % при STE-PGD-100; перенос с FP32 и SPSA дают 92,97 и 96,88 %. Прямые проходы с STE и точным округлением совпали. Для INT4-QAT STE-PGD-100 дает 82,03 %, для SPSA – 73,44 %, а объединение ошибок основного ансамбля и диагностических атак на той же подвыборке – 71,88 % (11 дополнительных ошибок относительно основного ансамбля). Следовательно, обычный градиент завышает устойчивость, а конечная STE-PGD не исчерпывает найденную уязвимость INT4; результаты подвыборки учитываются отдельно от табл. 1.

При INT4-PTQ исходная точность снизилась до 63,31 %. Большое значение R_{ret} = 0,942 в этом случае отражает сохранение уже низкой точности и не является признаком практической устойчивости. QAT восстановил исходную точность INT4 до 88,55 % и точность при PGD до 75,24 %. Прирост относительно INT4-PTQ составил соответственно 25,24 и 17,03 п.п., что демонстрирует способность обучения частично компенсировать ошибку грубой дискретизации.



■ Зависимость точности при состязательном воздействии от бюджета PGD-100
■ Accuracy under adversarial perturbations versus the PGD-100 budget

Зависимость от бюджета атаки показана на рисунке для первого запуска обучения (PGD-100 по разности логитов, шаг ε/10, три старта). INT8-конфигурации следуют близким траекториям, тогда как INT4-PTQ остается около 51 % вследствие потери качества еще на исходных данных. Кривая отдельной атаки не заменяет ансамбль табл. 1.

Расчет показателей ресурсов

Запуск выполнялся на Arduino Nano 33 BLE Sense с nRF52840, Cortex-M4F, 64 МГц, 1024 КиБ Flash и 256 КиБ оперативной памяти (RAM) (https://docs.arduino.cc/hardware/nano-33-ble-sense/; дата обращения: 15.09.2026). Использованы TensorFlow Lite for Microcontrollers (TFLM), CMSIS-NN (Common Microcontroller Software Interface Standard — Neural Network) 3.1.0 из CMSIS 5.9.0 и Arduino Mbed OS core 4.1.5 [7] (https://github.com/ARM-software/CMSIS-NN; дата обращения: 15.09.2026). Для INT8 применялись целочисленные ядра arm_convolve_s8, arm_depthwise_conv_s8 и arm_avgpool_s8 с накоплением в int32 и выходом int8. В INT4 веса диапазона [–7; 7] хранились по два в байте и распаковывались до int8 перед вызовом ядер; после сверток активации ограничивались 16 уровнями [0; 15] при хранении в int8 с теми же масштабами, что в PyTorch. FP32-конфигурация использует вещественный прямой проход.

Перед измерением на 824 исходных тестовых записях для INT8-PTQ, INT8-QAT, INT4-PTQ и INT4-QAT получено по 824/824 совпадения классов между платой и PyTorch. Максимальные абсолютные расхождения логитов составили $2{,}1 \cdot 10^{-3}$ для INT8 и $7{,}4 \cdot 10^{-3}$ для INT4. Эта проверка относится к исходным записям; оценки при PGD в табл. 1 получены в PyTorch. В табл. 2 приведены измеренные медианы задержки, нормированные показатели и значения U_0, рассчитанные по формуле (20) без промежуточного округления. Например, для INT8-QAT

$$U_0 = 0{,}30 \cdot 0{,}958738 + 0{,}40 \cdot 0{,}907767 - \\ - 0{,}15 \cdot \frac{14{,}9095}{1024} - 0{,}15 \cdot \frac{41{,}53}{100} \approx 0{,}5862. \quad (21)$$

■ ***Таблица 2.*** Задержка на Arduino и интегральная полезность конфигураций

■ ***Table 2.*** Arduino latency and composite utility of the configurations

Конфигурация	L, мс	S/S_{max}	L/L_{max}	U_0
FP32	72,43	0,0207	0,7243	0,542092
INT8-PTQ	41,87	0,0146	0,4187	0,584809
INT8-QAT	41,53	0,0146	0,4153	0,586249
INT4-PTQ	39,21	0,0126	0,3921	0,362070
INT4-QAT	38,96	0,0126	0,3896	0,506296

■ ***Таблица 3.*** Точность при акустических искажениях

■ ***Table 3.*** Accuracy under acoustic distortions

Условие	Контроль, %	Аугментация, %	Разность, п.п.
Исходные данные	95,95 ± 0,19	95,35 ± 0,14	–0,61 ± 0,21
SNR –5 дБ	50,85 ± 0,00	67,75 ± 3,65	16,90 ± 3,65
SNR 0 дБ	50,85 ± 0,00	79,18 ± 2,84	28,33 ± 2,84
SNR 10 дБ	51,17 ± 0,56	84,99 ± 5,80	33,82 ± 5,36
SNR 30 дБ	89,66 ± 3,26	93,50 ± 1,10	3,84 ± 2,22

Для S и точности использованы средние трех запусков обучения; L — медиана 50 запусков каждой конфигурации. Максимум U_0 получен для INT8-QAT: 0,586249 против 0,584809 у INT8-PTQ. Медиана INT8-QAT ниже FP32 на 42,66 %. По неокругленным значениям все пять конфигураций входят в парето-фронт; в каждой паре PTQ/QAT средние размеры контейнеров различаются лишь на 1,67 байта из-за заголовка, что не имеет самостоятельной практической значимости.

Для всех пяти конфигураций архитектурная сложность прохода составляет 2 618 944 операции умножения с накоплением (MAC) без учета активаций, квантования и обмена с памятью. Различия задержек характеризуют используемые реализации; уменьшение разрядности хранения весов не означает пропорционального уменьшения числа MAC.

Пороговый анализ дополняет интегральное ранжирование. Все пять медиан задержки ниже $L_{max} = 100$ мс. FP32, INT8-PTQ и INT8-QAT выполняют также пороги качества и размера и входят в F; INT4-конфигурации не достигают $A_{min} = 0,95$. Порог $S_{max} = 1024$ КиБ совпадает с полной емкостью Flash и используется для сравнения контейнеров, а не как измеренный свободный бюджет прошивки. Принадлежность F по этим показателям не подтверждает размещение всей прошивки в памяти или соблюдение предела задержки в каждом запуске.

Устойчивость к акустическим искажениям

Результаты в пространстве сигнала приведены в табл. 3. В отдельном парном опыте INT8-QAT контроль и аугментированная ветвь из общего FP32-состояния обучались ровно 10 эпох с одинаковыми остальными условиями. Сравнивались конечные состояния. Аугментация была представлена белым гауссовским шумом с вероятностью 0,5 и равномерным SNR от –5 до 10 дБ. Тестовые записи отделены по дикторам; шум совпадал по типу, но формировался независимо. При каждом SNR для каждой тестовой записи формировались три независимые реализации шума, одинаковые для контрольной и аугментированной модели. Точности трех реализаций усреднялись внутри каждого запуска, затем вычислялись среднее и стандартное отклонения трех запусков.

При –5 дБ точность выросла с 50,85 до 67,75 %, т. е. на 16,90 ± 3,65 п.п. Доля ошибок уменьшилась с 49,15 до 32,25 %, или на 34,4 %. Контрольная модель INT8-QAT при –5 дБ всегда выбирала команду “yes”, и 50,85 % соответствует доле этого класса (419/824). На исходных данных точность снизилась на 0,61 п.п. Этот результат показывает эффективность аугментации для исследованного белого шума; перенос на другие акустические условия не проверялся.

Ограничения исследования

Экспериментальная часть ограничена бинарной подзадачей “yes/no”, одной архитектурой DS-CNN и заданным ансамблем PGD-40/100. Перенос и SPSA дополнительно проверены на 128 примерах одного запуска. Поэтому численные границы не следует непосредственно переносить на полную 35-классовую задачу Speech Commands, другие архитектуры или иные нормы возмущения. Полученные значения характеризуют сравниваемые конфигурации в заданном протоколе, а не сертифицированный радиус устойчивости.

Сравнение включает в себя размеры контейнеров, число операций и медианы задержки на плате. Пиковая RAM и энергопотребление не измерялись. В целочисленных вариантах один тензор активаций $32 \times 20 \times 31$ занимает 19 840 байт при хранении в int8; это расчетный объем одного тензора, а не пиковая RAM с учетом арены TFLM, рабочих буферов ядер, распаковки весов и стека. Совпадение классов после конвертации на исходных данных не устанавливает совпадение устойчивости на атакованных входах.

Акустические результаты относятся к синтетическим искажениям сигнала; они описывают эксплуатационную устойчивость, но не полный тракт воспроизведения и записи в помещении.

Обсуждение

Математическая модель описывает различие между конфигурациями через два взаимосвязанных фактора: ошибку квантования e_b и локальную чувствительность L_b. Режим INT8-PTQ почти не изменил агрегированную исходную точность и приблизил кривую FP32 при PGD. Это согласуется с ситуацией, в которой ошибка квантования для большинства объектов меньше половины их исходного классификационного запаса. Следовательно, PTQ в данном режиме выступает прежде всего как средство уменьшения объема хранения, а не как самостоятельный механизм защиты.

Для проверки модели рассчитаны логиты на 824 записях трех запусков; объединенные наблюдения используются описательно. Для INT8-PTQ, INT8-QAT, INT4-PTQ и INT4-QAT медианы ошибки квантования по логитам относительно вещественной копии тех же весов равны 0,189; 0,195; 5,467 и 4,796; доли объектов с положительной правой частью (8) — 94,62; 94,54; 26,74 и 38,63 % соответственно. Среди первоначально правильно классифицированных объектов INT8-QAT медиана запаса равна 5,412 при сохранении класса под PGD и 0,830 при найденной атаке. Чувствительность оценивалась по сохраненному PGD-возмущению и трем случайным направлениям: по одному при радиусах 0,001; 0,01; 0,03 (генератор 31415). Медиана выборочной чувствительности равна 24,054 против 16,981 для FP32. Нарушений (8) при абсолютном допуске $2 \cdot 10^{-5}$ и относительном 10^{-6} к максимальному модулю логита вещественной копии не обнаружено.

Интегральный показатель и пороговый допуск решают разные задачи. U_0 задает ранжирование компромиссов, а множество F исключает конфигурации, нарушающие хотя бы одно обязательное требование. Поэтому большой R_{ret} INT4-PTQ не компенсирует низкую исходную точность. INT8-режимы позволяют уменьшить размер контейнера и медиану задержки при близком исходном качестве; в принятом сценарии максимум U_0 достигается у INT8-QAT.

Акустический блок дополняет оценку атак в пространстве признаков. Возмущение мелспектрограммы характеризует локальную геометрию классификатора, тогда как белый шум оценивает устойчивость всего тракта формирования признаков. Рост точности на 16,90 п.п. после целевой аугментации при –5 дБ подтверждает практическую ценность обучения на условиях, близких к границе эксплуатации. Такой подход совместим с TinyML, поскольку не увеличивает вычислительную сложность уже развернутой модели.

Заключение

Разработана математическая модель, связывающая ошибку квантования по логитам, классификационный запас и локальную чувствительность квантованной DS-CNN. Получена нижняя оценка $m_b \geq m_{32} - 2e_b$ и условная оценка локального радиуса при выполнении (9). Модель дополнена эмпирическими метриками, пороговым допуском и интегральным показателем развертывания.

Экспериментальное сравнение показало три характерных режима. INT8-PTQ сохранил близкую к FP32 исходную точность и близкое поведение при PGD. INT8-QAT обеспечил 95,87 % на исходных данных, 90,78 % при PGD и $R_{ret} = 0{,}947$. INT4-PTQ снизил исходную точность до 63,31 %, тогда как INT4-QAT частично восстановил ее до 88,55 %.

FP32 и обе INT8-конфигурации входят в F при принятых порогах качества, размера контейнера и медианы задержки. Максимальный $U_0 = 0{,}5862$ получен для INT8-QAT при $L = 41{,}53$ мс. INT8 сохраняет близкое к FP32 качество при меньших размере контейнера и медиане задержки; INT4-QAT применим при допустимости снижения исходной точности. Целевая аугментация при низком SNR дополняет квантование: при –5 дБ она повысила точность INT8-QAT на 16,90 п.п. и сократила долю ошибок на 34,4 %. Совместное применение математической модели, пороговых требований и парето-анализа позволяет выбирать режим квантования с учетом измеренных показателей качества и задержки при отдельном контроле памяти всей прошивки.

Литература

1. **Вольф Д. А., Туровский Я. А.** Распознавание речевых команд с применением сингулярного спектрального анализа в системах с ограниченными ресурсами. *Вестник Санкт-Петербургского государственного университета технологии и дизайна. Серия 1. Естественные и технические науки*, 2021, № 1, с. 3–11. doi:10.46418/2079-8199_2021_1_1, EDN: PNZXAY

2. **Левонович Н. И.** О методе распознавания голосовых команд с применением особого преобразования спектральных плотностей. *Моделирование и анализ данных*, 2022, т. 12, № 3, с. 49–57. https://psyjournals.ru/journals/mda/archive/2022_n3/Levonovich (дата обращения: 13.10.2026). doi:10.17759/mda.2022120304

3. **Сбоев А. Г., Рыбка Р. Б., Серенко А. В., Селиванов А. Г., Софонов А. О.** Классификация звука импульсной нейронной сетью, удовлетворяющей ограничениям нейропроцессора Алтай и при этом использующей полный динамический диапазон весов. *Вестник Военного инновационного технополиса «ЭРА»*, 2022, т. 3, № 3, с. 314–321. doi:10.56304/s2782375x22030093, EDN: HFEHPA
4. **Аксёнов А. А., Рюмина Е. В., Рюмин Д. А., Ивaнько Д. В., Карпов А. А.** Нейросетевой метод визуального распознавания голосовых команд водителя с использованием механизма внимания. *Научно-технический вестник информационных технологий, механики и оптики*, 2023, т. 23, № 4, с. 767–775. doi:10.17586/2226-1494-2023-23-4-767-775, EDN: ADSTLP
5. **Жолондковский В. В., Бочаров Ю. И., Бутузов В. А.** Нейросетевое устройство для распознавания речевых команд на базе аппаратных ускорителей семейства NeuroMatrix. *Безопасность информационных технологий*, 2024, т. 31, № 2, с. 134–141. doi:10.26583/bit.2024.2.09, EDN: MDJSMM
6. **Warden P.** Speech commands: A dataset for limited-vocabulary speech recognition, 2018. arXiv:1804.03209. https://arxiv.org/abs/1804.03209 (дата обращения: 13.10.2026). doi:10.48550/arXiv.1804.03209
7. **David R., Duke J., Jain A., Janapa Reddi V., Jeffries N., Li J., Kreeger N., Nappier I., Natraj M., Regev S., Rhodes R., Wang T., Warden P.** TensorFlow lite micro: Embedded machine learning on TinyML systems. *Proceedings of Machine Learning and Systems*, 2021, vol. 3, pp. 800–811.
8. **Jacob B., Kligys S., Chen B., Zhu M., Tang M., Howard A., Adam H., Kalenichenko D.** Quantization and training of neural networks for efficient integer-arithmetic-only inference. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713. doi:10.1109/cvpr.2018.00286
9. **Kim B., Chang S., Lee J., Sung D.** Broadcasted residual learning for efficient keyword spotting. *Interspeech 2021*, 2021, pp. 4538–4542. doi:10.21437/interspeech.2021-383
10. **López-Espejo I., Tan Z.-H., Hansen J. H. L., Jensen J.** Deep spoken keyword spotting: An overview. *IEEE Access*, 2022, vol. 10, pp. 4169–4199. doi:10.1109/access.2021.3139508
11. **Nahshan Y., Chmiel B., Baskin C., Zheltonozhskii E., Banner R., Bronstein A. M., Mendelson A.** Loss aware post-training quantization. *Machine Learning*, 2021, vol. 110, no. 11–12, pp. 3245–3262. doi:10.1007/s10994-021-06053-z
12. **Gupta K., Ajanthan T.** Improved gradient-based adversarial attacks for quantized networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 6, pp. 6810–6818. doi:10.1609/aaai.v36i6.20637
13. **Xiao Y., Liu A., Zhang T., Qin H., Guo J., Liu X.** RobustMQ: benchmarking robustness of quantized models. *Visual Intelligence*, 2023, vol. 1, art. 30. doi:10.1007/s44267-023-00031-w
14. **Ayaz F., Zakariyya I., Cano J., Keoh S. L., Singer J., Pau D., Kharbouche-Harrari M.** Improving robustness against adversarial attacks with deeply quantized neural networks. *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8. doi:10.1109/ijcnn54540.2023.10191429
15. **Goodfellow I. J., Shlens J., Szegedy C.** Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015. arXiv:1412.6572. https://arxiv.org/abs/1412.6572 (дата обращения: 13.10.2026).
16. **Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A.** Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. arXiv:1706.06083. https://arxiv.org/abs/1706.06083 (дата обращения: 13.10.2026).
17. **Athalye A., Carlini N., Wagner D.** Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research*, 2018, vol. 80, pp. 274–283.
18. **Croce F., Hein M.** Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research*, 2020, vol. 119, pp. 2206–2216.
19. **Carlini N., Athalye A., Papernot N., Brendel W., Rauber J., Tsipras D., Goodfellow I. J., Madry A., Kurakin A.** On evaluating adversarial robustness, 2019. arXiv:1902.06705. https://arxiv.org/abs/1902.06705 (дата обращения: 13.10.2026). doi:10.48550/arXiv.1902.06705
20. **Abdullah H., Warren K., Bindschaedler V., Papernot N., Traynor P.** SoK: The faults in our ASRs: An overview of attacks against Automatic Speech Recognition and speaker identification systems. *IEEE Symposium on Security and Privacy*, 2021, pp. 730–747. doi:10.1109/sp40001.2021.00014
21. **Bhanushali A. R., Mun H., Yun J.** Adversarial attacks on Automatic Speech Recognition (ASR): A survey. *IEEE Access*, 2024, vol. 12, pp. 88279–88302. doi:10.1109/access.2024.3416965
22. **López-Espejo I., Tan Z.-H., Jensen J.** A novel loss function and training strategy for noise-robust keyword spotting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, pp. 2254–2266. doi:10.1109/taslp.2021.3092567
23. **Wang Q., Zheng B., Li Q., Shen C., Ba Z.** Towards query-efficient adversarial attacks against Automatic Speech Recognition systems. *IEEE Transactions on Information Forensics and Security*, 2021, vol. 16, pp. 896–908. doi:10.1109/tifs.2020.3026543
24. **Tong C., Zheng X., Li J., Ma X., Gao L., Xiang Y.** Query-efficient black-box adversarial attacks on Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol. 31, pp. 3981–3992. doi:10.1109/taslp.2023.3304476

25. **Croce F., Andriushchenko M., Sehwag V., Debenedetti E., Flammarion N., Chiang M., Mittal P., Hein M.** RobustBench: A standardized adversarial robustness benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021, vol. 1.
26. **Куликов Д. А., Платонов В. В.** Состязательные атаки на системы обнаружения вторжений, использующих LSTM-классификатор. *Проблемы информационной безопасности. Компьютерные системы*, 2021, № 2, с. 48–56. EDN: CLLJVV
27. **Парфёнов Д. И., Болодурина И. П., Гришина Л. С., Жигалов А. Ю., Толмачев С. В.** Исследование эффективности трансферного обучения при проведении состязательных атак. *Современные информационные технологии и ИТ-образование*, 2022, т. 18, № 4, с. 861–871. doi:10.25559/sitito.18.202204.861-871, EDN: DDZPSG
28. **Фомичева С. Г., Беззатеев С. В.** Механизмы защиты моделей машинного обучения от состязательных атак. *T-Comm: Телекоммуникации и транспорт*, 2023, т. 17, № 10, с. 28–42. doi:10.36724/2072-8735-2023-17-10-28-42
29. **Гетьман А. И., Горюнов М. Н., Мацкевич А. Г., Рыболовлев Д. А., Никольская А. Г.** Состязательные атаки против системы обнаружения вторжений, основанной на применении методов машинного обучения. *Проблемы информационной безопасности. Компьютерные системы*, 2023, № 4, с. 156–190. doi:10.48612/jisp/eatr-5pxb-akt8, EDN: XTMDJV
30. **Вершков Н. А., Бабенко М. Г., Кучукова Н. Н., Кучуков В. А., Кучеров Н. Н.** Поперечнослойное разделение искусственных нейронных сетей для классификации изображений. *Компьютерная оптика*, 2024, т. 48, № 2, с. 312–320. doi:10.18287/2412-6179-co-1278
31. **Баев А. В., Самонов А. В., Сафонов В. М., Краснов С. В., Малышев С. Р.** Методы защиты моделей нейронных сетей от состязательных атак уклонения и отравления. *Автоматизация процессов управления*, 2024, № 4 (78), с. 39–48. doi:10.35752/1991-2927_2024_4_78_39
32. **Котенко И. В., Саенко И. Б., Лаута О. С., Васильев Н. А., Садовников В. Е.** Метод противодействия состязательным атакам на системы классификации изображений. *Вопросы кибербезопасности*, 2025, № 2 (66), с. 114–123. doi:10.21681/2311-3456-2025-2-114-123, EDN: MKWZFT

UDC 004.056
doi:10.31799/1684-8853-2026-5-11-20
EDN: AZRPYT

Mathematical model and experimental robustness assessment of quantized DS-CNNs for TinyML keyword spotting

M. A. Knyazev[a], Post-Graduate Student, orcid.org/0009-0007-3931-7442, maxiknyaz@mail.ru
M. M. Gusev[b], Master Student, orcid.org/0009-0004-7023-4328
[a]MIREA — Russian Technological University, 78, Vernadsky Ave., 119454, Moscow, Russian Federation
[b]HSE University, 34, Tallinskaia St., 123458, Moscow, Russian Federation

Introduction: Quantization enables keyword-spotting networks to operate on microcontrollers, but alters classification margins and sensitivity to perturbations. **Purpose:** To develop a model linking quantization error, classification margin, local robustness, and deployment constraints, and compare numerical representations. **Methods:** Depthwise separable convolutional networks have been evaluated on the binary "yes/no" subset of Google Speech Commands v2: a 32-bit floating-point baseline and 8-bit and 4-bit integer networks obtained by post-training quantization or quantization-aware training. The evaluation has combined adaptive projected-gradient-descent attacks in feature space, acoustic distortions with controlled signal-to-noise ratios, threshold admission, Pareto analysis, and a composite score. **Results:** Eight-bit post-training quantization has retained clean accuracy close to the baseline (95.67% versus 95.83%) and similar accuracy under attack (90.70% versus 91.59%). Eight-bit quantization-aware training has achieved 95.87% clean accuracy and 90.78% accuracy under attack, with an accuracy retention coefficient of 0.947. Four-bit post-training quantization has reduced clean accuracy to 63.31%, whereas quantization-aware training has restored it to 88.55%. With latency measured on Arduino Nano 33 BLE Sense, the eight-bit network trained with quantization awareness has achieved the highest composite score, 0.5862, at a median latency of 41.53 milliseconds. At a signal-to-noise ratio of −5 decibels, targeted augmentation of this network has improved accuracy from 50.85% to 67.75% and reduced the error rate by 34.4%. **Practical relevance:** The proposed model supports quantitative selection of a quantization mode using measured quality and latency, subject to separate checks of full-firmware memory requirements.

Keywords – TinyML, keyword spotting, DS-CNN, post-training quantization, quantization-aware training, INT8, INT4, adversarial robustness, PGD, acoustic distortions, KWS.

For citation: Knyazev M. A., Gusev M. M. Mathematical model and experimental robustness assessment of quantized DS-CNNs for TinyML keyword spotting. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 5, pp. 11–20 (In Russian). doi:10.31799/1684-8853-2026-5-11-20, EDN: AZRPYT

References

1. Wolf D. A., Turovsky Ya. A. Recognition of speech commands based on singular spectral analysis in systems with limited resources. *Vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta tekhnologii i dizajna. Seriya 1. Estestvennye i tekhnicheskie nauki* [Vestnik of St. Petersburg State University of Technology and Design. Series 1. Natural and Technical Sciences], 2021, no. 1, pp. 3–11 (In Russian). doi:10.46418/2079-8199_2021_1_1, EDN: PNZXAY
2. Levonovich N. I. Voice commands recognition method that uses special spectral density transform. *Modelirovanie i analiz dannyh* [Modelling and Data Analysis], 2022, vol. 12, no. 3, pp. 49–57. Available at: https://psyjournals.ru/en/journals/mda/archive/2022_n3/Levonovich (accessed 13 October 2026) (In Russian). doi:10.17759/mda.2022120304
3. Sboev A. G., Rybka R. B., Serenko A. V., Selivanov A. G., Sofonov A. O. Sound classification by a spiking neural network satisfying Altai neuroprocessor constraints. *Vestnik Voenno-*

go innovatsionnogo tekhnopolisa "ERA", 2022, vol. 3, no. 3, pp. 314–321 (In Russian). doi:10.56304/s2782375x22030093, EDN: HFEHPA
4. Axyonov A. A., Ryumina E. V., Ryumin D. A., Ivanko D. V., Karpov A. A. Neural network-based method for visual recognition of driver's voice commands using attention mechanism. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 767–775 (In Russian). doi:10.17586/2226-1494-2023-23-4-767-775, EDN: ADSTLP
5. Zholondkovskiy V. V., Bocharov Y. I., Butuzov V. A. Neural network device for recognizing speech commands based on hardware accelerators of the NEUROMATRIX family. *IT Security (Russia)*, 2024, vol. 31, no. 2, pp. 134–141 (In Russian). doi:10.26583/bit.2024.2.09, EDN: MDJSMM
6. Warden P. Speech commands: A dataset for limited-vocabulary speech recognition, 2018. arXiv:1804.03209. https://arxiv.org/abs/1804.03209 (accessed 13 October 2026). doi:10.48550/arXiv.1804.03209
7. David R., Duke J., Jain A., Janapa Reddi V., Jeffries N., Li J., Kreeger N., Nappier I., Natraj M., Regev S., Rhodes R., Wang T., Warden P. TensorFlow lite micro: Embedded machine learning on TinyML systems. *Proceedings of Machine Learning and Systems*, 2021, vol. 3, pp. 800–811.
8. Jacob B., Kligys S., Chen B., Zhu M., Tang M., Howard A., Adam H., Kalenichenko D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713. doi:10.1109/cvpr.2018.00286
9. Kim B., Chang S., Lee J., Sung D. Broadcasted residual learning for efficient keyword spotting. *Interspeech 2021*, 2021, pp. 4538–4542. doi:10.21437/interspeech.2021-383
10. López-Espejo I., Tan Z.-H., Hansen J. H. L., Jensen J. Deep spoken keyword spotting: An overview. *IEEE Access*, 2022, vol. 10, pp. 4169–4199. doi:10.1109/access.2021.3139508
11. Nahshan Y., Chmiel B., Baskin C., Zheltonozhskii E., Banner R., Bronstein A. M., Mendelson A. Loss aware post-training quantization. *Machine Learning*, 2021, vol. 110, no. 11–12, pp. 3245–3262. doi:10.1007/s10994-021-06053-z
12. Gupta K., Ajanthan T. Improved gradient-based adversarial attacks for quantized networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 6, pp. 6810–6818. doi:10.1609/aaai.v36i6.20637
13. Xiao Y., Liu A., Zhang T., Qin H., Guo J., Liu X. RobustMQ: benchmarking robustness of quantized models. *Visual Intelligence*, 2023, vol. 1, art. 30. doi:10.1007/s44267-023-00031-w
14. Ayaz F., Zakariyya I., Cano J., Keoh S. L., Singer J., Pau D., Kharbouche-Harrari M. Improving robustness against adversarial attacks with deeply quantized neural networks. *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8. doi:10.1109/ijcnn54540.2023.10191429
15. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015. arXiv:1412.6572. https://arxiv.org/abs/1412.6572 (accessed 13 October 2026).
16. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. arXiv:1706.06083. https://arxiv.org/abs/1706.06083 (accessed 13 October 2026).
17. Athalye A., Carlini N., Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research*, 2018, vol. 80, pp. 274–283.
18. Croce F., Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research*, 2020, vol. 119, pp. 2206–2216.
19. Carlini N., Athalye A., Papernot N., Brendel W., Rauber J., Tsipras D., Goodfellow I. J., Madry A., Kurakin A. On evaluating adversarial robustness, 2019. arXiv:1902.06705. https://arxiv.org/abs/1902.06705 (accessed 13 October 2026). doi:10.48550/arXiv.1902.06705
20. Abdullah H., Warren K., Bindschaedler V., Papernot N., Traynor P. SoK: The faults in our ASRs: An overview of attacks against Automatic Speech Recognition and speaker identification systems. *IEEE Symposium on Security and Privacy*, 2021, pp. 730–747. doi:10.1109/sp40001.2021.00014
21. Bhanushali A. R., Mun H., Yun J. Adversarial attacks on Automatic Speech Recognition (ASR): A survey. *IEEE Access*, 2024, vol. 12, pp. 88279–88302. doi:10.1109/access.2024.3416965
22. López-Espejo I., Tan Z.-H., Jensen J. A novel loss function and training strategy for noise-robust keyword spotting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, pp. 2254–2266. doi:10.1109/taslp.2021.3092567
23. Wang Q., Zheng B., Li Q., Shen C., Ba Z. Towards query-efficient adversarial attacks against Automatic Speech Recognition systems. *IEEE Transactions on Information Forensics and Security*, 2021, vol. 16, pp. 896–908. doi:10.1109/tifs.2020.3026543
24. Tong C., Zheng X., Li J., Ma X., Gao L., Xiang Y. Query-efficient black-box adversarial attacks on Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol. 31, pp. 3981–3992. doi:10.1109/taslp.2023.3304476
25. Croce F., Andriushchenko M., Sehwag V., Debenedetti E., Flammarion N., Chiang M., Mittal P., Hein M. RobustBench: A standardized adversarial robustness benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021, vol. 1.
26. Kulikov D. A., Platonov V. V. Adversarial attacks on intrusion detection systems using LSTM classifier. *Problems of Information Security. Computer Systems*, 2021, no. 2, pp. 48–56 (In Russian). EDN: CLLJVV
27. Parfenov D. I., Bolodurina I. P., Grishina L. S., Zhigalov A. Yu., Tolmachev S. V. Investigation of transfer learning efficiency in adversarial attacks. *Modern Information Technologies and IT-Education*, 2022, vol. 18, no. 4, pp. 861–871 (In Russian). doi:10.25559/sitito.18.202204.861-871, EDN: DDZPSG
28. Fomicheva S. G., Bezzateev S. V. Defenses for machine learning models from adversarial attacks. *T-Comm*, 2023, vol. 17, no. 10, pp. 28–42 (In Russian). doi:10.36724/2072-8735-2023-17-10-28-42
29. Getman A. I., Goryunov M. N., Matskevich A. G., Rybolovlev D. A., Nikolskaya A. G. Adversarial attacks against a machine learning based intrusion detection system. *Problems of Information Security. Computer Systems*, 2023, no. 4, pp. 156–190 (In Russian). doi:10.48612/jisp/eatr-5pxb-akt8, EDN: XTMDJV
30. Vershkov N. A., Babenko M. G., Kuchukova N. N., Kuchukov V. A., Kucherov N. N. Transverse-layer partitioning of artificial neural networks for image classification. *Computer Optics*, 2024, vol. 48, no. 2, pp. 312–320 (In Russian). doi:10.18287/2412-6179-co-1278
31. Baev A. V., Samonov A. V., Safonov V. M., Krasnov S. V., Malyshev S. R. Methods for protecting neural network models from evasion and poisoning attacks. *Automation of Control Processes*, 2024, no. 4(78), pp. 39–48 (In Russian). doi:10.35752/1991-2927_2024_4_78_39
32. Kotenko I. V., Saenko I. B., Lauta O. S., Vasiliev N. A., Sadovnikov V. E. A method of countering adversarial attacks on image classification systems. *Voprosy kiberbezopasnosti*, 2025, no. 2(66), pp. 114–123 (In Russian). doi:10.21681/2311-3456-2025-2-114-123, EDN: MKWZFT