УДК 004.08
doi:10.31799/1684-8853-2026-3-2-13
EDN: WKWZGY

Научные статьи
Articles

Метод сжатия глубоких нейронных сетей, основанный на геометрически контролируемом прореживании

Т. М. Татарникова[a], *доктор техн. наук, профессор, orcid.org/0000-0002-6419-0072, tm-tatarn@yandex.ru*
А. С. Раскопина[a], *аспирант, ассистент, orcid.org/ 0009-0002-0276-607X*
[a]*Санкт-Петербургский государственный университет аэрокосмического приборостроения, Б. Морская ул., 67, Санкт-Петербург, 190000, РФ*

Введение: *высокие вычислительные ресурсы, энергозатраты и время для решения задач с применением технологий глубокого обучения обусловили поиск решений сжатия моделей нейронных сетей без существенной потери качества результата.* ***Цель:*** *разработать метод сжатия глубоких нейронных сетей – сокращения вычислительной сложности и числа параметров сверточных нейронных сетей без существенной потери точности решения задачи классификации.* ***Результаты:*** *разработан новый метод геометрически контролируемого прореживания модели нейронной сети, основанный на жадном отборе кандидатов структурного прореживания с контролем изменения геометрии представлений. Предложена метрика контролирования сохранения геометрии представлений в виде матрицы межклассовых сходств, вычисляемой по центроидам классов в пространстве признаков. Введен параметр допустимого бюджета деформации геометрии представлений и предложен подход к его выбору на основе оценки шумового порога геометрической метрики. Результаты эксперимента показали, что предложенный метод сжатия моделей нейронных сетей обеспечивает сохранение точности классификации после дообучения, сопоставимой с базовой моделью без прореживания при сокращении вычислительной сложности на ~8 % и числа параметров на ~12 % на примере архитектуры ResNet-50 и набора данных CIFAR-100. Дополнительно показана переносимость разработанного метода на архитектуры глубоких нейронных сетей ResNet-18 и MobileNetV2.* ***Практическая значимость:*** *разработанный метод может найти применение при решении задачи классификации на мобильных и встраиваемых устройствах в реальном времени.*

Ключевые слова *– глубокое обучение, сжатие нейронной сети, прореживание, геометрия представлений, вычислительная сложность, задержка инференса, качество классификации.*

Для цитирования: Татарникова Т. М., Раскопина А. С. Метод сжатия глубоких нейронных сетей, основанный на геометрически контролируемом прореживани. *Информационно-управляющие системы*, 2026, № 3, с. 2–13. doi:10.31799/1684-8853-2026-3-2-13, EDN: WKWZGY
For citation: Tatarnikova T. M., Raskopina A. S. A deep neural network compression method based on geometrically controlled thinning. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 2–13 (In Russian). doi:10.31799/1684-8853-2026-3-2-13, EDN: WKWZGY

Введение

Сжатие нейронных сетей является основным решением при практическом применении глубокого обучения в вычислительно ограниченных условиях. Например, современные сверточные нейронные сети обеспечивают высокое качество распознавания изображений, однако их вычислительная сложность и объем параметров приводят к росту задержки инференса, потребления памяти и энергозатрат. Это существенно ограничивает использование моделей глубокого обучения в задачах реального времени, а также на мобильных и встраиваемых устройствах [1–4].

Одним из наиболее распространенных подходов к сжатию нейронных сетей является прореживание — удаление части параметров или структурных элементов модели нейронной сети с целью уменьшить ее вычислительную сложность. В зависимости от вида удаляемых элементов различают неструктурированное прореживание, предполагающее обнуление отдельных весов связей между нейронами; структурированное прореживание, при котором удаляются каналы сверточной нейронной сети, фильтры или блоки; полуструктурированное прореживание, которое выполняется в соответствии с регулярным шаблоном разреженности. Эти традиционные подходы хоть и позволяют уменьшить вычислительную сложность, приводящую к желаемому ускорению вывода решения, но сопровождаются существенным ухудшением качества решения задачи классификации [5–7].

Главная проблема традиционных методов прореживания — это выбор элементов сети, удаление которых минимально влияет на качество результата. Традиционные методы опираются на локальные критерии важности параметров, такие как нормы весов, или на приближенные оценки вклада каналов в итоговую ошибку. Однако качество классификации определяется не только значениями весов, но и формированием внутреннего пространства признаков, в котором классы становятся разделимыми. Поэтому

перспективно построение методов прореживания, контролирующих изменение структуры сжимаемой модели нейронной сети [8–10].

В данной работе предлагается метод геометрически контролируемого прореживания, основанный на жадной стратегии выбора степени структурного прореживания блоков сети при ограничении на изменение геометрии представлений. Геометрия представлений определяется через матрицу попарных косинусных сходств между центроидами классов в пространстве признаков. Поскольку оценка изменения геометрии вычисляется на основе выборочных пакетов данных и подвержена стохастической вариативности, вводится понятие порогового уровня шума — величины естественных флуктуаций метрики ΔG при неизменной архитектуре. Данный порог используется для формирования устойчивого бюджета допустимого изменения геометрии.

Постановка задачи и метрики оценивания

Рассматривается задача многоклассовой классификации изображений. Пусть задана обучающая выборка

$$D_{Learn} = (x_i, y_i)_{(i=1)}^{N},$$

где $x_i \in R^{(H\times W\times C)}$ — входное изображение, H, W, C — высота, ширина и число каналов (фильтров) изображения; $y_i \in \{1, \ldots, K\}$ — метка класса, K — число классов; N — количество обучающих примеров.

Пусть нейронная сеть с параметрами θ задает отображение

$$\mathbf{f}_\theta : R^{(H\times W\times C)} \to R^K,$$

где $\mathbf{f}_\theta(x)$ — вектор логитов, т. е. ненормированных оценок принадлежности к каждому из K классов.

Предсказанный класс определяется как

$$\hat{y}(x) = \arg_{k\in\{1,\ldots,K\}} \max\left[\mathbf{f}_\theta(x)\right].$$

Цель прореживания — сжатие модели $\mathbf{f}_{\theta'}$, где θ' получается из θ путем удаления части параметров или структурных элементов нейронной сети с выполнением следующих требований:

— сохранение качества классификации на уровне исходной модели нейронной сети;

— уменьшение числа параметров модели нейронной сети и ее вычислительной сложности;

— улучшение времени получения результата (задержки инференса).

В работе используются стандартные метрики Top-1 и Top-5 точности. Top-1 точность определяется как доля примеров, для которых предсказанный класс совпадает с истинным. Top-5 точность определяется как доля примеров, для которых истинный класс входит в множество пяти наиболее вероятных предсказаний [10, 11].

Для оценки степени сжатия модели нейронной сети используются следующие показатели:

1) число параметров модели, P;

2) объем модели в мегабайтах, V — суммарный размер параметров с учетом формата хранения;

3) вычислительная сложность, FLOPs — число операций умножения-сложения при одном прямом проходе модели [12];

4) задержка инференса в миллисекундах, T — продолжительность времени между подачей входных данных в модель и получением результата. На практике задержка инференса зависит не только от числа FLOPs, но и от особенностей аппаратной платформы, реализации операторов и структуры вычислительного графа. Поэтому задержка инференса рассматривается как эмпирическая метрика, отражающая реальное ускорение в условиях выбранного вычислительного окружения.

Для анализа реализуемых методов, кроме общепринятых метрик, введем понятие геометрии представлений для контроля устойчивости внутреннего пространства признаков.

Пусть $\boldsymbol{\varphi}_\theta(x) \in R^d$ — вектор признаков, извлекаемый из нейронной сети на предпоследнем уровне. Выбор предпоследнего слоя позволяет анализировать геометрию признаков независимо от параметров классификатора и обеспечивает сопоставимость межклассовой структуры до и после структурного упрощения сети. Тогда для каждого класса $k \in \{1, \ldots, K\}$ определим множество объектов данного класса

$$D_k = \left\{(x, y) \in D_{Learn} : y = k\right\}.$$

Тогда множество признаков класса k в пространстве признаков определяется как

$$H_k = \left\{\boldsymbol{\varphi}_\theta(x_i) \mid y_i = k\right\}.$$

Центроид класса определяется как [12]

$$\mu_k = \frac{1}{|H_k|} \sum_{\boldsymbol{\varphi}_\theta(x_i)\in H_k} \boldsymbol{\varphi}_\theta(x_i).$$

Далее строится матрица попарных косинусных сходств между центроидами классов

$$\mathbf{S} = \left[s_{ij}\right]_{i,j=1}^{K}, \quad s_{ij} = \frac{\left\langle \mu_i, \mu_j \right\rangle}{\|\mu_i\|_2 \cdot \|\mu_j\|_2 + \delta'}, \qquad (*)$$

где $\delta > 0$ — малая константа, предотвращающая деление на ноль.

Матрица $\mathbf{S}$ отражает относительное расположение классов в пространстве представлений:

близкие классы имеют высокое косинусное сходство, плохо разделимые классы — низкое. Матрица $\mathbf{S}$ характеризует угловую структуру пространства признаков и описывает взаимное направление центроидов классов.

Пусть $\mathbf{S}_{Ref}$ — матрица сходств, вычисленная в пространстве признаков $\varphi_\theta(x)$ для базовой модели $\mathbf{f}_\theta$, а $\mathbf{S}$ — матрица сходств, вычисленная в пространстве признаков $\varphi_{\theta'}(x)$ для модифицированной модели $\mathbf{f}_{\theta'}$ (после прореживания). Тогда изменение геометрии представлений определяется величиной

$$\Delta G(\theta',\ \theta) = \frac{\left\|\mathbf{S} - \mathbf{S}_{Ref}\right\|_F}{\left\|\mathbf{S}_{Ref}\right\|_F + \delta},$$

где $||\cdot||_F$ — норма Фробениуса.

В работе используется косинусная мера сходства между центроидами классов. Для повышения корректности сравнения признаки перед вычислением сходств L2-нормируются. Изменение геометрии ΔG измеряется как относительная фробениусова норма разности матриц сходств.

Выбранная метрика обладает масштабной инвариантностью и частичной инвариантностью к ортогональным преобразованиям, однако не инвариантна к произвольным аффинным преобразованиям, что учитывается при интерпретации результатов.

Величина ΔG используется для оценки изменения межклассовой структуры признакового пространства после прореживания: малые значения соответствуют сохранению относительного расположения классов и рассматриваются как индикатор устойчивости представлений. При этом ΔG не гарантирует сохранение точности и выступает как критерий, ограничивающий чрезмерную деформацию.

Метрика основана на косинусных сходствах центроидов классов и отражает межклассовую структуру через первые моменты распределений. Она не учитывает внутриклассовую дисперсию, однако позволяет контролировать изменения линейной разделимости за счет сохранения взаимного расположения классов.

Традиционные методы прореживания нейронных сетей

В рамках работы анализируются три наиболее распространенные группы методов: неструктурированное, структурированное и полуструктурированное прореживание [13—15]. Следует отметить, что в современной литературе используются и другие классификации методов прореживания, основанные на критерии отбора параметров (magnitude-based, gradient-based, Hessian-based), стратегии оптимизации (one-shot, iterative) или режиме обучения (post-training, training-time pruning). В настоящей работе используется структурная типология, поскольку она непосредственно связана с изменением архитектуры сети и вычислительной сложности, что соответствует целям исследования.

Неструктурированное прореживание заключается в обнулении отдельных весов нейронной сети без изменения размерностей слоев [16]. Пусть $\mathbf{W}$ — тензор весов некоторого слоя. Тогда прореживание задается бинарной маской $\mathbf{M}$ той же размерности:

$$\mathbf{W}' = \mathbf{W} \times \mathbf{M}.$$

При заданной доле разреженности удаляются веса с наименьшими значениями модуля веса:

$$M_{ij} = \begin{cases} 0, \text{ если } \left|W_{ij}\right| \le \tau; \\ 1, \text{ если } \left|W_{ij}\right| > \tau, \end{cases}$$

где τ — пороговое значение.

Неструктурированное прореживание позволяет достигать высокой разреженности при небольшой потере точности после дообучения. Однако его практическое ускорение на стандартных платформах (GPU/CPU) ограничено из-за нерегулярной структуры и необходимости специализированных операций с разреженными матрицами [17, 18].

Структурированное прореживание предполагает удаление целых структурных единиц модели: каналов свертки, фильтров, нейронов, блоков или слоев. В отличие от неструктурированного подхода, структурное прореживание изменяет размерности весовых тензоров и активаций, что приводит к непосредственному уменьшению вычислительной сложности и ускорению инференса [19].

Для сверточного слоя с весами $\mathbf{W} \in \mathbf{R}^{C_{Out} \times C_{In} \times k \times k}$ структурированное прореживание по выходным каналам соответствует удалению некоторого подмножества индексов $P \in \{1, \ldots, C_{Out}\}$ [20]. Тогда новый слой имеет размерность

$$\mathbf{W}' \in \mathbf{R}^{(C_{Out} - |P|) \times C_{In} \times k \times k},$$

где C_{Out} — число выходных каналов (фильтров); C_{In} — число входных каналов слоя; P — число удаленных каналов (каналы с минимальной нормой фильтра удаляются в первую очередь).

Основным преимуществом структурированного прореживания является прямое сокращение вычислительной сложности модели нейронной сети и ускорение ее работы на стандартных устройствах. Недостатком является сильная деградация качества, поскольку удаление целых каналов уменьшает выразительную способность модели [21, 22].

Полуструктурированное прореживание вводит регулярную схему разреженности, которая поддерживается современными аппаратными платформами. Наиболее распространенным вариантом является шаблон (n:m), при котором в каждой группе из m весов сохраняются только n ненулевых [23, 24].

Пусть $\mathbf{W} \in \mathbf{R}^m$ — группа весов. Тогда (n:m)-прореживание задает маску $\mathbf{M} \in \{0, 1\}^m$, удовлетворяющую условию $||\mathbf{M}||_0 = n$, а веса после прореживания определяются как $\mathbf{W}' = \mathbf{W} \times \mathbf{M}$. Обычно сохраняют n весов с максимальным модулем.

Полуструктурированное прореживание сочетает лучшую аппаратную эффективность по сравнению с неструктурированным и большую гибкость по сравнению со структурированным подходом. При этом фактический выигрыш в задержке инференса зависит от поддержки (n:m)-разреженности на конкретной платформе.

После прореживания качество модели может снижаться из-за изменения отображения и распределения представлений, поэтому применяется дообучение для восстановления точности.

Несмотря на широкое распространение, традиционные методы прореживания имеют ряд ограничений.

1. Критерии важности часто основаны на локальных характеристиках (нормы весов, статистики активаций, приближенные оценки градиентов), что не гарантирует сохранение разделимости классов.

2. При высокой степени прореживания наблюдается существенное снижение качества, не всегда компенсируемое дообучением.

3. Неструктурированная разреженность не обеспечивает уменьшение задержки инференса на стандартном оборудовании.

4. Отмечается неочевидный выбор частей модели нейронной сети для удаления при структурированном прореживании.

Описание предлагаемого метода сжатия глубоких нейронных сетей

Разработанный метод сжатия глубоких нейронных сетей основан на эвристическом предположении о том, что существенная деформация межклассовой структуры представлений при структурном упрощении сети повышает риск ухудшения классиционной способности модели. Соответственно, ограничение изменения геометрии рассматривается как практический критерий контроля риска деградации качества, а не как строгое теоретическое условие сохранения точности.

Таким образом, задача прореживания глубокой нейронной сети формулируется как поиск ее структурно упрощенной модели при ограничении на допустимое изменение геометрии представлений.

Пусть базовая модель имеет параметры θ, а прореженная модель — θ'. В соответствии с формулой (*) введем матрицу сходств центроидов классов

$$\mathbf{S}_{Ref} = \mathbf{S}(\theta), \ \mathbf{S} = \mathbf{S}(\theta').$$

Изменение геометрии представлений определяется величиной

$$\Delta G(\theta', \theta) = \frac{\|S(\theta') - S(\theta)\|_F}{\|\mathbf{S}(\theta)\|_F + \delta}.$$

В предлагаемом методе сжатия модели нейронной сети вводится ограничение вида

$$\Delta G(\theta', \theta) \le \varepsilon,$$

где ε — допустимый бюджет изменения геометрии представлений.

Практическая реализация ограничения на ΔG требует учета того факта, что вычисление геометрии представлений выполняется по ограниченному числу наборов данных и поэтому имеет стохастическую составляющую. Даже при фиксированных параметрах θ значения матрицы $\mathbf{S}$ и, следовательно, ΔG могут отличаться при вычислении на разных выборках данных. Для учета данного эффекта вводится величина порога шума

$$\Delta G_{Noise} = \Delta G(\theta, B_1, B_2),$$

где B_1, B_2 – две выборки наборов данных, на которых независимо вычисляются матрицы $\mathbf{S}_1$ и $\mathbf{S}_2$ для одной и той же модели.

Тогда

$$\Delta G_{Noise} = \frac{\|\mathbf{S}_1 - \mathbf{S}_2\|_F}{\|\mathbf{S}_1\|_F + \delta}.$$

Далее итоговый бюджет задается как сумма шумового порога и дополнительного допуска

$$\varepsilon = \Delta G_{Noise} + \varepsilon_{\lim},$$

где $\varepsilon_{\lim} \ge 0$ — настраиваемый параметр метода.

Данный подход обеспечивает устойчивость метода к стохастическим колебаниям метрики и позволяет интерпретировать $\varepsilon_{\lim}$ как управляемый запас изменения геометрии.

Адаптация метода к модели сверточной нейронной сети

Рассмотрим работу предложенного метода геометрически контролируемого прореживания модели нейронной сети на архитектуре ResNet с прореживанием по ширине блоков архитектуры — сокращением числа внутренних каналов в отдельных блоках ResNet. Блок содержит несколько сверточных слоев.

Для выбора каналов, подлежащих удалению, используем критерий, основанный на статистике активаций. Рассмотрим первый блок (обозначим его BL1) с тензором активаций $\mathbf{a} \in \mathbf{R}^{H\times W\times C}$. Для каждого канала $c \in \{1, \ldots, C\}$ вычисляется дисперсия по пространственным координатам и по наборам данных, собранным на калибровочном наборе:

$$\vartheta_c = Var\left(\mathbf{a}_c\right).$$

Интуитивно каналы с более высокой дисперсией в среднем несут больше информации и сильнее участвуют в формировании признаков, поэтому в первую очередь сохраняются каналы с максимальными значениями ϑ_c.

Для заданной доли удаления $r \in (0, 1)$ выбирается число сохраняемых каналов

$$C_{Saved} = \max\left(C_{\min}\left\lfloor C\left(1-r\right)\right\rfloor\right),$$

где $C_{\min}$ — минимально допустимое число каналов в блоке; $\lfloor\cdot\rfloor$ — операция округления до ближайшего целого.

Далее выбирается множество индексов сохраняемых каналов

$$\zeta = TopK\left(\left\{\vartheta_c\right\}_{c=1}^{C}, C_{Saved}\right).$$

Метод геометрически контролируемого прореживания использует жадную стратегию. Пусть задан дискретный набор кандидатов долей прореживания

$$\Re = \left\{r_1, r_2, \ldots, r_m\right\}, \quad 0 < r_1 < \ldots < r_m < 1 .$$

На практике кандидаты упорядочиваются по убыванию, т. е. сначала проверяются более агрессивные значения.

Для каждого блока нейронной сети выполняется перебор $r \in \Re$ и выбирается максимально возможное r, при котором выполняется ограничение $\Delta G \le \varepsilon$. Если ни один кандидат не удовлетворяет ограничению, блок не прореживается. Таким образом, метод автоматически регулирует степень прореживания в зависимости от чувствительности блоков к нарушению геометрии представлений. Ограничение $\Delta G \le \varepsilon$ накладывается на этапе структурного упрощения и используется для отбора допустимых вариантов прореживания. Следует отметить, что после дообучения значение $\Delta G \le \varepsilon$ может изменяться и не обязано совпадать с заданным порогом. В предлагаемом подходе геометрический контроль рассматривается как механизм ограничения начальной деформации векторного пространства перед оптимизацией, а не как жесткое инвариантное условие финального решения.

Фактически предложенный метод геометрически контролируемого прореживания вводит дополнительный критерий качества: удаление параметров допускается лишь при сохранении структуры взаимного расположения классов в признаковом пространстве. Это снижает риск нарушения разделимости классов при структурном упрощении модели.

При этом учитывается стохастическая природа оценки геометрии: значение ΔG может варьироваться в зависимости от выборки. В связи с этим ограничение задается относительно порога шума ΔG_{Noise}, а параметр $\varepsilon_{\lim}$ интерпретируется как допустимый запас деформации признакового пространства.

Методы прореживания, учитывающие структуру признакового пространства, обычно основаны на анализе активаций или сходства признаков [25–27], а также на различных критериях важности, включая значимость фильтров и структурные свойства сети [13, 15]. Такие подходы, как правило, требуют хранения промежуточных представлений или вычисления дополнительных метрик, что увеличивает вычислительную сложность.

В отличие от этого предлагаемый метод использует агрегированное описание геометрии классов через центроиды и косинусное сходство, обеспечивая более эффективную вычислительно альтернативу. При этом задача прореживания формулируется как задача контролируемого изменения структуры модели при ограничении на геометрию представлений, а не как задача максимального сокращения числа параметров.

Отметим, что ряд современных методов требует индивидуальной настройки протоколов обучения, поэтому применение единой схемы дообучения без адаптации может приводить к некорректной оценке. В данной работе для обеспечения сопоставимости условий рассматриваются базовые классы методов, тогда как расширенное сравнение с индивидуальной настройкой оставлено на дальнейшие исследования.

Постановка эксперимента

Эксперименты проводились на наборе данных CIFAR-100, содержащем 60 000 цветных изображений размером 32 × 32, распределенных по 100 классам (50 000 – обучающая выборка, 10 000 – тестовая). Этот набор данных широко применяется для оценки методов сжатия и ускорения сверточных нейронных сетей. Принятая постановка эксперимента ограничена задачей классификации изображений малого разрешения. Несмотря на формальную инвариантность используемой геометрической метрики к масштабу признаков,

переносимость результатов на задачи с высоким разрешением не гарантируется, поскольку такие задачи сопровождаются изменением архитектурных решений, числа классов и сложности распределения данных.

В качестве базовой архитектуры определена ResNet-50. Поскольку исходная версия сети ориентирована на изображения более высокого разрешения, применялась стандартная адаптация под CIFAR-100: свертка 7×7 заменена на 3×3 с шагом 1, операция maxpool удалена, а выходной слой модифицирован для 100 классов. Это позволило сохранить пространственное разрешение признаков на ранних этапах обработки и должным образом адаптировать архитектуру к данным малого размера.

Для корректного сравнения всех методов используется единый протокол обучения при равном вычислительном бюджете. Протокол включает следующие этапы:

1) обучение базовой модели (baseline-40) в течение 40 эпох на обучающей выборке CIFAR-100 и сохранение ее в качестве контрольной точки;

2) для оценки эффекта дополнительного обучения продолжается оптимизация базовой модели еще на 20 эпохах без применения прореживания.

Для всех методов прореживания реализована единая последовательность: baseline-40 → прореживание → дообучение на 20 эпохах. Таким образом, итоговые модели сравниваются при одинаковом числе эпох оптимизации (60 эпох), что исключает влияние «дополнительного времени обучения» как источника роста точности.

В рамках эксперимента рассматриваются следующие методы: базовая модель без прореживания, неструктурированное прореживание, структурированное прореживание, полуструктурированное прореживание – (n:m)-прореживание, геометрически контролируемое прореживание.

В экспериментах использовались следующие параметры:

— доля разреженности $r = 0{,}7$;

— на каждом шаге обучения модель обрабатывает 128 примеров данных одновременно;

— для полуструктурированного прореживания применялся шаблон n:m в виде 2:4;

— прореживанию подвергались блоки 3 и 4 архитектуры ResNet.

Параметры предложенного алгоритма, включая набор коэффициентов прореживания и минимальное число каналов, выбирались с целью обеспечить устойчивость процедуры и предотвратить деградацию представлений. Ограничение на минимальное число каналов позволяет избежать вырождения признакового пространства. Прореживание применялось к более глубоким слоям сети, формирующим высокоуровневые признаки и содержащим основную долю параметров, тогда как ранние слои сохранялись без изменений для поддержания стабильности базовых представлений.

Результаты контрольного обучения сравниваемых методов и их анализ

Результаты контрольного обучения без прореживания (baseline-60) и трех традиционных методов прореживания модели нейронной сети: структурированного, неструктурированного и полуструктурированного — представлены в табл. 1. Для повышения достоверности результатов эксперименты на ResNet-50 выполнялись с тремя фиксированными значениями начальных условий генераторов случайных чисел, поэтому в табл. 1 приведены средние значения метрик (mean) и стандартные отклонения (std). Это позволяет оценить устойчивость методов к стохастическим факторам обучения и дообучения.

■ ***Таблица 1.*** Значения метрик качества традиционных методов прореживания

■ ***Table 1.*** Values of quality metrics for traditional thinning methods

Метод	Метрика					
	Top-1, % (mean±std)	Top-5, % (mean±std)	P, млн	V, МБ	FLOPs, млрд	T, мс (mean±std)
Baseline-60 (без прореживания)	73,34±0,13	92,73±0,08	23,71	90,43	1,311	7,73±0,55
Структурированное прореживание по глубине	60,84±0,41	86,40±0,36	15,26	58,21	0,807	7,12±3,06
Неструктурированное прореживание	38,65±0,32	69,88±0,17	23,71	90,43	1,311	9,86±3,33
Полуструктурированное (n:m)-прореживание	58,68±1,03	84,04±0,93	23,71	90,43	1,311	6,91±1,31

Анализ результатов из табл. 1 показывает, что число параметров, размер модели нейронной сети и FLOPs не меняются при неструктурированном и (n:m)-прореживании — методы не изменяют размерности тензоров, а меняется структура разреженности. Структурированное прореживание уменьшает число параметров и FLOPs за счет удаления каналов.

Базовая модель (baseline-60), обученная на 40 эпохах, показала Top-1 = 58,33 %, тогда как после продолжения обучения до 60 эпох точность возрасла до 73,34 %. Это подтверждает, что корректное сравнение методов прореживания требует фиксации вычислительного бюджета обучения, поскольку дополнительное дообучение существенно улучшает качество.

Неструктурированное прореживание в проведенных экспериментах демонстрирует заметное снижение точности при заданной доле разреженности 0,7 (Top-1 ≈ 38,65 %). Следует отметить, что эффективность неструктурированного прореживания существенно зависит от выбора схемы дообучения и гиперпараметров, а также может требовать специализированных процедур восстановления качества.

В рамках данной работы использовалась единая схема дообучения для всех методов, что позволило обеспечить сопоставимость условий, однако не предполагает отдельного тюнинга под неструктурированное прореживание. Кроме того, неструктурированное прореживание не приводит к снижению FLOPs и размера модели без использования специализированных аппаратных или программных оптимизаций, что ограничивает его практическую применимость на стандартных устройствах.

Структурированный метод сокращает число параметров примерно на 36 % (с 23,7 до 15,3 млн) и уменьшает FLOPs с 1,31 до 0,81 млрд, однако итоговая точность остается на уровне около 60,84 %, что существенно ниже baseline-60.

Метод (n:m)-прореживания дает точность 59,19 %, что сопоставимо со структурированным подходом, но без сокращения числа параметров и FLOPs.

Далее продемонстрируем результаты предложенного метода геометрически контролируемого прореживания для модели ResNet-50 на датасете CIFAR-100. В эксперименте варьировался параметр ε_{lim}, определяющий дополнительный допустимый бюджет деформации геометрии представлений. Для каждого значения $\varepsilon_{\text{lim}} \in \{0{,}06; 0{,}1; 0{,}14; 0{,}18\}$ выполнялось три независимых запуска с различными фиксированными значениями начальных условий генераторов случайных чисел.

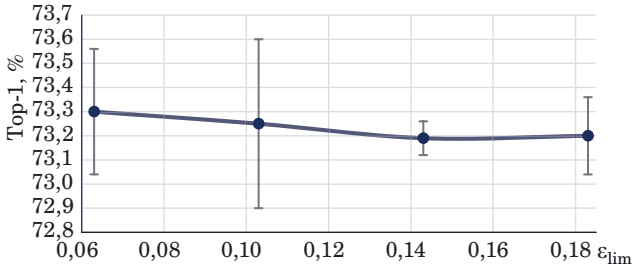
Анализ метрик предложенного метода геометрически контролируемого прореживания по табл. 2 показывает, что увеличение ε_{lim} приводит к более выраженному структурному упрощению модели: число параметров снижается с 22,88 млн при $\varepsilon_{\text{lim}} = 0{,}06$ до 20,80 млн при $\varepsilon_{\text{lim}} = 0{,}18$, а вычислительная сложность уменьшается с 1,256 млрд FLOPs до 1,212 млрд FLOPs. Таким образом, метод ориентирован не на агрессивное сжатие, а на контролируемое структурное упрощение с минимальной деформацией геометрии представлений. Параметр ε_{lim} обеспечивает интерпретируемое управление компромиссом между степенью сжатия и ограничением на деформацию геометрии представлений. При этом итоговая точность Top-1 во всех рассмотренных режимах остается близкой к уровню базовой модели после дообучения, демонстрируя значения порядка 73,2–73,3 % (рис. 1). Это указывает на то, что предложенный метод позволяет выполнять сокращение вычислительных затрат без существенного ухудшения качества классификации, что особенно важно для практического применения на ресурсно-ограниченных устройствах.

Отдельного внимания заслуживает поведение метрики геометрии представлений. Значение ΔG до дообучения закономерно возрастает при увеличении ε_{lim}, что отражает факт более сильного вмешательства в архитектуру сети (рис. 2). Однако после дообучения ΔG стабилизируется около 0,14–0,15, что позволяет предположить наличие эффекта частичного восстановления структуры представлений в процессе дообучения.

■ ***Таблица 2.*** Значения метрик метода геометрически контролируемого прореживания

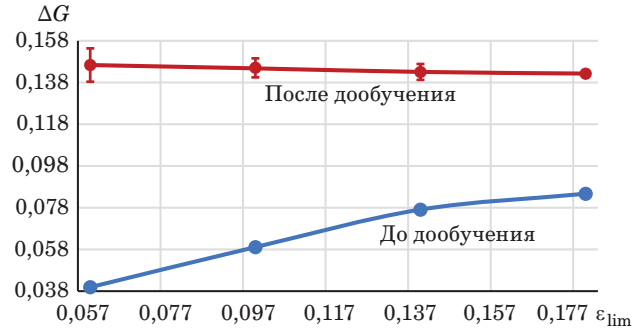
■ ***Table 2.*** Values of the metrics of the geometrically controlled thinning method

ε_{lim}	Метрика						
	Top-1, % (mean±std)	Top-5, % (mean±std)	P, млн	V, МБ	FLOPs, млрд	ΔG до дообучения (mean±std)	ΔG после дообучения (mean±std)
0,06	**73,30±0,26**	92,80±0,07	22,88	87,29	1,256	**0,0399±0,0008**	0,1464±0,0080
0,10	73,24±0,36	92,65±0,06	22,11	84,35	1,233	0,0591±0,0014	0,1448±0,0024
0,14	73,18±0,09	**92,74±0,24**	21,19	80,84	1,218	0,0771±0,0006	**0,1431±0,0038**
0,18	73,21±0,15	92,80±0,10	**20,80**	**79,35**	**1,212**	0,0847±0,0009	0,1423±0,0014



■ ***Рис. 1.*** Точность Top-1 в зависимости от ε_{lim} геометрически контролируемого метода прореживания

■ ***Fig. 1.*** Top-1 accuracy as a function ε_{lim} of the geometrically controlled pruning method



■ ***Рис. 2.*** Значение ΔG до и после дообучения

■ ***Fig. 2.*** The value of ΔG before and after fine-tuning

В ходе экспериментов наблюдается устойчивый рост значения ΔG после дообучения по сравнению со значением на этапе прореживания. Это связано с тем, что дообучение оптимизирует функцию потерь и не содержит явного ограничения на геометрию представлений, вследствие чего происходит дополнительная перестройка признакового пространства.

Таким образом, геометрическое ограничение не является инвариантным свойством финальной модели, а выступает как механизм ограничения области поиска на этапе структурного упрощения.

Несмотря на наблюдаемый рост значения ΔG после дообучения, практическая значимость геометрического ограничения подтверждается сравнением при равном бюджете параметров. При сопоставимом числе параметров (~21 млн) структурированное прореживание без геометрического контроля обеспечивает Top-1 = 59,93 %, тогда как с предложенным методом достигается 73,21 %.

Все методы сравниваются при едином протоколе дообучения и фиксированных гиперпараметрах, что позволяет изолировать влияние критерия прореживания. В этой постановке наблюдаемое преимущество не может быть объяснено различиями в настройке или вычислительном бюджете.

Полученный результат показывает, что геометрическое ограничение играет ключевую роль на этапе прореживания, ограничивая разрушение структуры представлений и формируя более устойчивую инициализацию для последующей оптимизации, несмотря на дальнейшую перестройку признакового пространства.

Проведен дополнительный эксперимент на модели ResNet-18 для CIFAR-100. В отличие от ResNet-50 с Bottleneck-блоками, ResNet-18 использует BasicBlock, что позволяет оценить применимость предлагаемого геометрического критерия для различных типов сверточных архитектур.

В качестве базовой модели использовалась нейронная сеть, обученная в течение 40 эпох. Далее выполнялось дообучение на 20 эпохах без прореживания, а также применялся геометрически контролируемый метод прореживания с параметром $\varepsilon_{lim} = 0{,}18$ и последующим дообучением в течение 20 эпох. Итоговые результаты приведены в табл. 3.

Для ResNet-18 при $\varepsilon_{lim} = 0{,}18$ наблюдается снижение ΔG с 0,0913 до 0,0305, что указывает на частичное восстановление структуры пространства признаков в процессе оптимизации.

Таким образом, эксперимент демонстрирует, что геометрически контролируемое прореживание обеспечивает мягкое уменьшение числа параметров и вычислительной сложности при сохранении точности классификации относительно контрольного варианта дообучения без прореживания.

Для оценки переносимости предложенного метода на архитектуры иного типа дополнительно проведены эксперименты на MobileNetV2

■ ***Таблица 3.*** Результаты работы геометрически контролируемого прореживания на модели ResNet-18

■ ***Table 3.*** Results of geometrically controlled pruning on the ResNet-18 model

Метод	Метрика				
	Top-1, %	Top-5, %	*P*, млн	*V*, МБ	FLOPs, млрд
Baseline ResNet-18 (40 эпох)	71,15	91,83	11, 22	42,80	0,56
Baseline ResNet-18 + дообучение (20 эпох)	71,63	91,78	11,22	42,80	0,56
Геометрически контролируемое прореживание при $\varepsilon_{lim} = 0{,}18$ + дообучение (20 эпох)	**71,72**	**91,81**	**10, 19**	**38,86**	**0,53**

■ ***Таблица 4.*** Результаты работы геометрически контролируемого прореживания на модели MobileNetV2
■ ***Table 4.*** Results of geometrically controlled pruning on the MobileNetV2 model

Метод		Метрика			
		P, млн	FLOPs, млн	*V*, МБ	Top-1, %
MobileNetV2 + дообучение (20 эпох)		2,35±0,00	26,17±0,00	8,97±0,00	59,78±0,15
Геометрически контролируемое прореживание	при ε_{lim} = 0,10	2,25±0,01	24,85±0,11	8,58±0,02	59,81±0,07
	при ε_{lim} = 0,18	2,25±0,01	24,54±0,13	8,57±0,02	59,42±0,17
	при ε_{lim} = 0,30	2,22±0,03	24,30±0,05	8,33±0,15	59,50±0,23
	при ε_{lim} = 0,70	2,17±0,02	23,93±0,00	8,28±0,00	59,55±0,35

(CIFAR-100). Процедура GC-GPEC применена к InvertedResidual-блокам с ограничением на изменение межклассовой геометрии и анализом различных значений параметра ε_{lim}, задающего допустимое изменение геометрии представлений. Результаты (табл. 4) показывают, что увеличение ε_{lim} приводит к последовательному снижению числа параметров (с ~2,35M до ~2,17M) и FLOPs (с ~26,17M до ~23,93M) при минимальной деградации точности (в диапазоне 59,4—59,8 % Top-1 по сравнению с (59,78±0,15) % для baseline).

Особое внимание уделено практической применимости метода. Для этого проведено измерение задержки инференса на реальном встраиваемом устройстве Raspberry Pi 4B (CPU, batch size = 1) с использованием OpenCV DNN backend. Для каждой модели выполнялось 10 независимых запусков. Каждый запуск включал этап прогрева (warm-up, 30 итераций) и 100 измерений инференса. Результаты показали, что предложенный метод обеспечивает стабильное снижение задержки инференса: с (7,80±0,15) мс для baseline и до (7,30±0,10) мс для прореженных моделей (ε_{lim} = 0,18), что соответствует ускорению порядка 4—5 %.

Следует отметить, что наблюдаемое ускорение инференса соответствует умеренному уровню структурного прореживания. В отличие от агрессивных методов, ориентированных на максимальное снижение вычислительной сложности, предложенный подход направлен на сохранение геометрии представлений и стабильности модели.

Дополнительно был проведен анализ зависимости итоговой точности от значения ΔG до дообучения на основе уже полученных экспериментальных данных.

В экспериментах на архитектурах ResNet-50 и MobileNetV2 не выявлено случаев, при которых малое значение ΔG приводило бы к деградации точности после дообучения. При этом зависимость между ΔG и точностью не является строго функциональной: близкие значения ΔG могут соответствовать незначительно различающимся результатам, а увеличение ΔG не всегда приводит к ухудшению качества. Это подтверждает, что метрика ΔG выступает как критерий, ограничивающий чрезмерную деформацию пространства представлений, но не являющийся точным предиктором итоговой точности.

Дополнительно проведен анализ чувствительности метода к параметру C_{min} на архитектуре MobileNetV2. Установлено, что в умеренных режимах прореживания (ε_{lim} = 0,1 и 0,3) изменение C_{min} в диапазоне от двух до 16 не влияет на итоговые характеристики модели, что указывает на доминирующую роль геометрического ограничения ΔG. В то же время в более агрессивном режиме (ε_{lim} = 0,7) влияние параметра становится заметным: уменьшение C_{min} приводит к более сильному сжатию модели и росту ΔG до дообучения, что подтверждает его роль как дополнительного структурного ограничения.

Анализ чувствительности к выбору набора кандидатов прореживания показал, что более плотная дискретизация (0,05; 0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6) позволяет достигать большей степени сжатия при сопоставимом значении ΔG и практически неизменной точности. Это связано с более точным приближением к границе допустимой деформации геометрии представлений, при этом метод в целом демонстрирует устойчивость к выбору набора кандидатов.

Заключение

Основным результатом работы является новый метод сжатия модели глубокой нейронной сети, основанный на геометрически контролируемом прореживании ее структуры.

Эксперимент на сверточной нейронной сети с архитектурой ResNet-50 показал, что геометрически контролируемый метод прореживания модели нейронной сети обеспечивает устойчивое снижение вычислительной сложности модели при сохранении точности классификации: метод демонстрирует сопоставимую точность Top-1 и Top-5 по сравнению с базовой моделью после дообучения, одновременно снижая FLOPs и число параметров.

Дополнительно выполнена проверка переносимости геометрически контролируемого метода прореживания модели нейронной сети на архитектуру ResNet-18 и MobileNetV2. Результаты эксперимента показали, что предложенный метод сохраняет эффективность и для модели с другой структурой блоков, обеспечивая уменьшение числа параметров и FLOPs без снижения точности классификации в сравнении с обучением без прореживания при одинаковом бюджете дообучения.

Полученные результаты подтверждают перспективность использования геометрических ограничений как критерия качества при прореживании, а также открывают возможности для дальнейших исследований. К таким направлениям относятся оптимальное распределение геометрического бюджета по слоям с учетом внутриклассовой дисперсии представлений, расширение подхода на другие архитектуры, а также перенос на задачи классификации с более высокоразрешенными данными.

Литература

1. **Dantas P. V., Sabino da Silva W., Cordeiro L. C., Carvalho C. B.** A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 2024, vol. 54, no. 22, pp. 11804–11844. doi:10.1007/s10489-024-05747-w
2. **Кузнецов А. В.** Цифровая история и искусственный интеллект: перспективы и риски применения больших языковых моделей. *Новые информационные технологии в образовании и науке*, 2022, № 5, с. 53–57. doi:10.17853/2587-6910-2022-05-53-57, EDN: VFYSAN
3. **Liu D., Zhu Y., Liu Z., Liu Y., Han C., Tian J., Li R., Yi W.** A survey of model compression techniques: Past, present, and future. *Front Robot AI*, 2025. doi:10.3389/frobt.2025.1518965
4. **Богачев И. В., Булканов Д. Е.** Обзор современных нейросетевых методов сжатия для задачи обработки измерительных данных. *Вестник Тихоокеанского государственного университета*, 2024, № 2 (73), с. 83–92. doi:10.38161/1996-3440-2024-2-83-92, EDN: EBDQZC
5. **He Y., Xiao L.** Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, iss. 5, pp. 2900–2919. doi:10.1109/TPAMI.2023.3334614
6. **Чернышов Н. Д., Буряк Д. Ю.** Исследование влияния метода сравнения каналов на эффективность алгоритмов поканального прореживания сверточных нейронных сетей. *Системный анализ в науке и образовании*, 2025, № 1, с. 16–22. EDN: JFTUOQ. https://sanse.ru/index.php/sanse/article/view/648 (дата обращения: 20.03.2026).
7. **Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu**. *Full parameter fine-tuning for large language models with limited resources*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024, vol. 1: Long Papers. Bangkok, Association for Computational Linguistics, 2024, pp. 8187–8198.
8. **Zhang Q., Zhang R., Sun J., Liu Y.** How sparse can we prune a deep network: A fundamental limit perspective. *Advances in Neural Information Processing Systems: Proceedings of the NeurIPS Conference*, 2024, vol. 37, pp. 91337–91372. doi:10.52202/079017-2898
9. **Zhu K., Hu F., Ding Y., Zhou W., Wang R.** A comprehensive review of network pruning based on pruning granularity and pruning time perspectives. *Neurocomputing*, 2025, vol. 626, Article 129382. doi:10.1016/j.neucom.2025.129382
10. **Татарникова Т. М., Мокрецов Н. С.** Оптимизация моделей дистилляции знаний для языковых моделей. *Научно-технический вестник информационных технологий, механики и оптики*, 2025, т. 25, № 4, с. 737–743. doi:10.17586/2226-1494-2025-25-4-737-743, EDN: PSPNOU
11. **Кузьмин В. Н., Менисов А. Б., Сабиров Т. Р.** Метод оптимизации нейронных сетей на основе структурной дистилляции с применением генетического алгоритма. *Научно-технический вестник информационных технологий, механики и оптики*, 2024, т. 24, № 5, с. 770–778. doi:10.17586/2226-1494-2024-24-5-770-778, EDN: SKBJQT
12. **Park C., Park M., Moon H., Yoon M.K., Go S., Kim S., Ro W. W.** DEPrune: Depth-wise separable convolution pruning for maximizing GPU parallelism. *Advances in Neural Information Processing Systems: Proceedings of NeurIPS Conference*, 2024, pp. 106906–106923. doi:10.52202/079017-3394
13. **Sun X., Shi H.** Towards better structured pruning saliency by reorganizing convolution. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 2193–2203. doi:10.1109/WACV57701.2024.00220
14. **Wright D., Igel C., Selvan R.** BMRS: Bayesian model reduction for structured pruning. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024, pp. 64119-64144. doi:10.52202/079017-2045
15. **Dong Z., Duan Y., Zhou Y., Duan S., Hu X.** Weight-adaptive channel pruning for CNNs based on closeness-centrality modeling. *Applied Intelligence*, 2024, vol. 54, pp. 201–215. doi:10.1007/s10489-023-05164-5
16. **Khan N. A., Rafat A. M. S.** Pruning convolution neural networks using filter clustering based on normalized cross-correlation similarity. *Journal of Information and Telecommunication*, 2024, vol. 9(2), pp. 190–208. doi:10.1080/24751839.2024.2415008

17. **Duan Z., Lu M., Ma J., Huang Y., Ma Z., Zhu F.** Quantization-aware ResNet VAE for lossy image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, iss. 1, pp. 436–450. doi:10.1109/TPAMI.2023.3322904
18. **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, EDN: LUUUAK
19. **Yu H., Zhang W., Ji M., Zhen C.** Automatic channel pruning method by introducing additional loss for deep neural networks. *Neural Processing Letters*, 2022, vol. 55, pp. 1071–1085. doi:10.1007/s11063-022-10926-2
20. **Liu Y., Wu D., Zhou W., Fan K., Zhou Z.** An effective automatic channel pruning for neural networks. *Neurocomputing*, 2023, vol. 526, pp. 131–142. doi:10.1016/j.neucom.2023.01.014
21. **Hu Y., Zhu J., Chen J.** Continuous pruning function for efficient 2:4 sparse pre-training. *Conference: Advances in Neural Information Processing Systems*, 2024, pp. 33756–33778. doi:10.52202/079017-1063
22. **Shi Y., Tang A., Niu L., Zhou R.** Sparse optimization guided pruning for neural networks. *Neurocomputing*, 2024, vol. 574, Article 127280. doi:10.1016/j.neucom.2024.127280
23. **Krizhevsky A., Hinton G.** *Learning Multiple Layers of Features from Tiny Images*: Technical report. University of Toronto, 2009. https://www.cs.toronto.edu/~kriz/cifar.html (дата обращения: 14.01.2026).
24. **Huang H., Pao H.-K.** Interpretable deep model pruning. *Neurocomputing*, 2025, vol. 647, Article 130485. doi:10.1016/j.neucom.2025.130485
25. **Ganguli T., Chong E.** Activation-based pruning of neural networks. *Algorithms*, 2024, vol. 17, iss. 1, pp. 48. doi:10.3390/a17010048
26. **Cui J., Wang Z., Yang Z., Guan X.** A pruning method based on feature map similarity score. *Big Data and Cognitive Computing*, 2023, vol. 7, iss. 1, pp. 159. doi:10.3390/bdcc7040159
27. **Zhao C., Zhang Y., Ni B.** Exploiting channel similarity for network pruning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, vol. 33, iss. 9, pp. 5049–5061. doi:10.1109/TCSVT.2023.3248659

UDC 004.08
doi:10.31799/1684-8853-2026-3-2-13
EDN: WKWZGY

A deep neural network compression method based on geometrically controlled thinning

T. M. Tatarnikova[a], Dr. Sc., Tech., Professor, orcid.org/0000-0002-6419-0072, tm-tatarn@yandex.ru
A. S. Raskopina[a], Post-Graduate Student, Assistant Professor, orcid.org/ 0009-0002-0276-607X
[a]Saint-Petersburg State University of Aerospace Instrumentation, 67, B. Morskaya St., 190000, Saint-Petersburg, Russian Federation

Introduction: High computational resources, energy costs, and time required to solve problems using deep learning technologies have necessitated the search for solutions to compressing neural network models without significant loss of result quality. **Purpose:** To develop a method for compressing deep neural networks, reducing the computational complexity and the number of parameters of convolutional neural networks without significantly losing the accuracy of the classification problem. **Results:** A new method for geometrically controlled thinning of neural network models is developed, based on the greedy selection of structural thinning candidates with control over changes in representation geometry. We propose a metric for controlling the preservation of representation geometry in the form of an interclass similarity matrix calculated from the class centroids in the feature space. We introduce a parameter for the admissible budget of representation geometry deformation, and propose an approach to its selection based on an estimate of the noise threshold of the geometric metric. The experimental results have shown that the proposed method for compressing neural network models ensures that classification accuracy after additional training is maintained, comparable to the base model without thinning, while reducing computational complexity by ~8% and the number of parameters by ~12% using the ResNet-50 architecture and the CIFAR-100 dataset as an example. Additionally, the portability of the developed method to the architectures of deep neural networks ResNet-18 and MobileNetV2 is demonstrated. **Practical relevance:** The developed method can find application in solving classification problems on mobile and embedded devices in real time.

Keywords – deep learning, neural network compression, thinning, representation geometry, computational complexity, inference delay, classification quality.

For citation: Tatarnikova T. M., Raskopina A. S. A deep neural network compression method based on geometrically controlled thinning. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 2–13 (In Russian). doi:10.31799/1684-8853-2026-3-2-13, EDN: WKWZGY

References

1. Dantas P. V., Sabino da Silva W., Cordeiro L. C., Carvalho C. B. A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 2024, vol. 54, no. 22, pp. 11804–11844. doi:10.1007/s10489-024-05747-w
2. Kuznetsov A. V. Digital history and artificial intelligence: perspectives and risks of pretrained language models. *New Information Technologies in Education and Science*, 2022, no. 5, pp. 53–57 (In Russian). doi:10.17853/2587-6910-2022-05-53-57, EDN: VFYSAN
3. Liu D., Zhu Y., Liu Z., Liu Y., Han C., Tian J., Li R., Yi W. A survey of model compression techniques: Past, present, and future. *Front Robot AI*, 2025. doi:10.3389/frobt.2025.1518965
4. Bogachev I. V., Bulkanov D. E. Review of modern neural network compression methods for the task of processing measurement data. *Bulletin of Pacific National University*, 2024, vol. 2, no. 73, pp. 83–92 (In Russian). doi:10.38161/1996-3440-2024-2-83-92, EDN: EBDQZC
5. He Y., Xiao L. Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 2024, vol. 46, iss. 5, pp. 2900–2919. doi:10.1109/TPAMI.2023.3334614
6. Chernyshov N. D., Buryak D. Yu. Research into impact of channel comparison method on efficiency of algorithms for channel-wise pruning of convolutional neural networks. *System Analysis in Science and Education*, 2025, vol. 1, pp. 16–22. EDN: JFTUOQ. Available at: https://sanse.ru/index.php/sanse/article/view/648 (accessed 20 March 2026) (In Russian).
7. Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu. *Full parameter fine-tuning for large language models with limited resources*. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, vol. 1: Long Papers. Bangkok, Association for Computational Linguistics, 2024, pp. 8187–8198.
8. Zhang Q., Zhang R., Sun J., Liu Y. How sparse can we prune a deep network: A fundamental limit perspective. *Advances in Neural Information Processing Systems: Proceedings of the NeurIPS Conference*, 2024, vol. 37, pp. 91337–91372. doi:10.52202/079017-2898
9. Zhu K., Hu F., Ding Y., Zhou W., Wang R. A comprehensive review of network pruning based on pruning granularity and pruning time perspectives. *Neurocomputing*, 2025, vol. 626, Article 129382. doi:10.1016/j.neucom.2025.129382
10. Tatarnikova T. M., Mokretsov N. S. Optimizing knowledge distillation models for language models. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 4, pp. 737–743 (In Russian). doi:10.17586/2226-1494-2025-25-4-737-743, EDN: PSPNOU
11. Kuzmin V. N., Menisov A. B., Sabirov T. R. A method for optimizing neural networks based on structural distillation using a genetic algorithm. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 5, pp. 770–778 (In Russian). doi:10.17586/2226-1494-2024-24-5-770-778, EDN: SKBJQT
12. Park C., Park M., Moon H., Yoon M. K., Go S., Kim S., Ro W. W. DEPrune: Depth-wise separable convolution pruning for maximizing GPU parallelism. *Advances in Neural Information Processing Systems: Proceedings of NeurIPS Conference*, 2024, pp. 106906–106923. doi:10.52202/079017-3394
13. Sun X., Shi H. Towards better structured pruning saliency by reorganizing convolution. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 2193–2203. doi:10.1109/WACV57701.2024.00220
14. Wright D., Igel C., Selvan R. BMRS: Bayesian model reduction for structured pruning. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024, pp. 64119–64144. doi:10.52202/079017-2045
15. Dong Z., Duan Y., Zhou Y., Duan S., Hu X. Weight-adaptive channel pruning for CNNs based on closeness-centrality modeling. *Applied Intelligence*, 2024, vol. 54, pp. 201–215. doi:10.1007/s10489-023-05164-5
16. Khan N. A., Rafat A. M. S. Pruning convolution neural networks using filter clustering based on normalized cross-correlation similarity. *Journal of Information and Telecommunication*, 2024, vol. 9(2), pp. 190–208. doi:10.1080/24751839.2024.2415008
17. Duan Z., Lu M., Ma J., Huang Y., Ma Z., Zhu F. QARV: Quantization-aware ResNet VAE for lossy image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, pp. 436–450. doi:10.1109/TPAMI.2023.3322904
18. Tatarnikova T. M., Raskopina A. S. Analyzing the effectiveness of post-learning quantization for optimizing neural networks. *H&ES Reserch*, 2025, vol. 17, no. 2, pp. 4–10 (In Russian). doi:10.36724/2409-5419-2025-17-2-4-10, EDN: LUUUAK
19. Yu H., Zhang W., Ji M., Zhen C. ACP: Automatic channel pruning method by introducing additional loss for deep neural networks. *Neural Processing Letters*, 2022, vol. 55, pp. 1071–1085. doi:10.1007/s11063-022-10926-2
20. Liu Y., Wu D., Zhou W., Fan K., Zhou Z. EACP: An effective automatic channel pruning for neural networks. *Neurocomputing*, 2023, vol. 526, pp. 131–142. doi:10.1016/j.neucom.2023.01.014
21. Hu Y., Zhu J., Chen J. S-STE: Continuous pruning function for efficient 2:4 sparse pre-training. *Conference: Advances in Neural Information Processing Systems*, 2024, pp. 33756–33778. doi:10.52202/079017-1063
22. Shi Y., Tang A., Niu L., Zhou R. Sparse optimization guided pruning for neural networks. *Neurocomputing*, 2024, vol. 574, Article 127280. doi:10.1016/j.neucom.2024.127280
23. Krizhevsky A., Hinton G. *Learning Multiple Layers of Features from Tiny Images*: Technical report. University of Toronto, 2009. Available at: https://www.cs.toronto.edu/~kriz/cifar.html (accessed 14 January 2026).
24. Huang H., Pao H.-K. Interpretable deep model pruning. *Neurocomputing*, 2025, vol. 647, Article 130485. doi:10.1016/j.neucom.2025.130485
25. Ganguli T., Chong E. Activation-based pruning of neural networks. *Algorithms*, 2024, vol. 17, iss. 1, pp. 48. doi:10.3390/a17010048
26. Cui J., Wang Z., Yang Z., Guan X. A pruning method based on feature map similarity score. *Big Data and Cognitive Computing*, 2023, vol. 7, iss. 1, pp. 159. doi:10.3390/bdcc7040159
27. Zhao C., Zhang Y., Ni B. Exploiting channel similarity for network pruning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, vol. 33, iss. 9, pp. 5049–5061. doi:10.1109/TCSVT.2023.3248659