НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2026 Том 26 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2026 Vol. 26 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

COMPUTER SCIENCE

doi: 10.17586/2226-1494-2026-26-3-466-474
УДК 004.258

Метод геометрически-контролируемого прореживания модели сверточной нейронной сети без потери качества результата

Татьяна Михайловна Татарникова[1]✉, Анастасия Сергеевна Раскопина[2]

[1,2] Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация

[1] tm-tatarn@yandex.ru✉, https://orcid.org/0000-0002-6419-0072
[2] raskopina.anastasia@yandex.ru, https://orcid.org/0009-0002-0276-607X

Аннотация

Введение. Рассмотрена задача сжатия глубоких нейронных сетей на примере сверточных нейронных сетей (Convolutional Neural Network, CNN). Размеры глубоких нейронных сетей являются препятствием для их практического применения в условиях ограниченных вычислительных ресурсов, энергии и требований задержки инференса. Одним из развиваемых направлений сжатия моделей глубоких нейронных сетей является прореживание — удаление части параметров или структурных элементов модели нейронной сети. Показано, что прореживание является компромиссом между точностью классификации и вычислительной эффективностью. **Метод.** Предложен метод геометрически-контролируемого прореживания модели CNN, основанный на жадном выборе степени разреженности структуры CNN при ограничении на изменение геометрии представлений. Геометрия представлений определяется через матрицу попарных косинусных сходств между центроидами классов в пространстве признаков. Оценка качества работы представленного метода получена на сравнении с методом прореживания по глубине и без прореживания модели CNN по стандартным метрикам Top-1 и Top-5 точности классификации. Оценивание степени сжатия CNN выполнено по числу параметров, объему и вычислительной сложности модели нейронной сети, а также задержки инференса. Измерения задержки инференса проводились в одинаковых условиях при фиксированном размере входных данных. **Основные результаты.** Выполнена разработка нового метода сжатия CNN, основанного на геометрически-контролируемом прореживании модели CNN, в котором прореживание выполняется жадно по блокам CNN, а допустимость каждого шага определяется не только локальной важностью каналов, но и глобальным ограничением на изменение геометрии представлений. Эксперимент выполнен на CNN с архитектурой ResNet-50 и обучением на стандартном наборе данных CIFAR-100 для оценки методов сжатия и ускорения CNN. Результаты проведенного эксперимента демонстрируют устойчивое сохранение качества классификации у предложенного метода при сокращении вычислительной сложности и числа параметров по сравнению с базовым методом без использования прореживания и методом прореживания по глубине модели CNN. Разработанный метод демонстрирует сопоставимую точность Top-1 и Top-5 с базовой моделью после дообучения, одновременно снижая вычислительную сложность более чем в 2,5 раза и число параметров почти в два раза. **Обсуждение.** Представленный метод геометрически-контролируемого прореживания модели CNN может найти применение в задачах, требующих принятия решения в реальном времени, а также на мобильных и встраиваемых устройствах.

Ключевые слова

глубокая нейронная сеть, число параметров обучения, прореживание модели нейросети, сжатие модели нейронной сети, геометрия представлений, вычислительная сложность, задержка инференса, качество классификации

Ссылка для цитирования: Татарникова Т.М., Раскопина А.С. Метод геометрически-контролируемого прореживания модели сверточной нейронной сети без потери качества результата // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 3. С. 466–474. doi: 10.17586/2226-1494-2026-26-3-466-474

Optimizing knowledge distillation models for language models

Tatiana M. Tatarnikova[1]✉, Anastasia S. Raskopina[2]

[1,2] Saint Petersburg State University of Aerospace Instrumentation (SUAI), Saint Petersburg, 190000, Russian Federation

[1] tm-tatarn@yandex.ru✉, https://orcid.org/0000-0002-6419-0072

[2] raskopina.anastasia@yandex.ru, https://orcid.org/0009-0002-0276-607X

Abstract

The problem of deep neural network compression is discussed using Convolutional Neural Networks (CNNs) as an example. The size of deep neural networks is an obstacle to their practical application under conditions of limited computing resources, energy, and inference latency requirements. One of the developing approaches to compressing deep neural network models is thinning-removing some of the parameters or structural elements of the neural network model. It is shown that thinning is a tradeoff between classification accuracy and computational efficiency. The problem of compressing CNN models was formulated by thinning parameters while maintaining classification quality at the level of the original model, reducing the number of parameters, computational complexity, and inference latency. Standard Top-1 and Top-5 classification accuracy metrics were used to evaluate classification quality. The degree of compression and inference latency were assessed using metrics, such as the number of model parameters, model size, computational complexity, and inference latency. To ensure a fair comparison of the proposed method with others, inference latency measurements were conducted under identical conditions with a fixed input data size. The concept of representation geometry is introduced to control the stability of the internal feature space. The change in the interclass similarity matrix, calculated from the class centroids in the feature space, is used as a metric for preserving the structure of the feature space. The main result of this work is the development of a new compression method for CNNs based on geometrically controlled thinning of the CNN model. In this method, thinning is performed greedily across CNN blocks, and the admissibility of each step is determined not only by the local importance of channels but also by a global constraint on changing the geometry of the representations. The results of the experiment demonstrate that the proposed method consistently maintains classification quality while reducing computational complexity and the number of parameters compared to the baseline method without thinning and the depth-based thinning method of the CNN model. The proposed method for geometrically controlled thinning of a CNN model can be applied to problems requiring real-time decision making as well as on mobile and embedded devices.

Keywords

deep neural network, number of training parameters, neural network model thinning, neural network model compression, representation geometry, computational complexity, inference delay, classification quality

For citation: Tatarnikova T.M., Raskopina A.S. Optimizing knowledge distillation models for language models. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 3, pp. 466–474 (in Russian). doi: 10.17586/2226-1494-2026-26-3-466-474

Введение

Сжатие сверточных нейронных сетей (Convolutional Neural Network, CNN) является одной из ключевых задач при практическом применении глубокого обучения в вычислительно ограниченных условиях [1–4].

CNN обеспечивают высокое качество распознавания изображений, однако их вычислительная сложность и объем параметров приводят к росту времени выполнения (задержки инференса), потребления памяти и энергозатрат.

Одним из наиболее распространенных подходов к сжатию нейронных сетей является прореживание, когда удаляется часть параметров или структурных элементов модели. Прореживание обеспечивает уменьшение размеров тензоров и сокращение числа операций, что приводит к уменьшению задержки инференса, но сопровождается существенным ухудшением качества результата работы CNN.

Анализ научных работ за последние пять лет показывает, что прореживание позволяет уменьшить вычислительные затраты CNN, но оно имеет свои ограничения. В частности, в работе [5] описаны N:M-структуры разреженности, которые уменьшают количество операций примерно в 1,5–2 раза без сильного падения качества, но требуют специализированного аппаратного ускорения и не универсальны для всех архитектур. В работах [6, 7], показана возможность обнуления до 50–60 % весов моделей при падении точности менее чем на 1–2 %, однако данные результаты трудно ускорить на стандартном оборудовании из-за нерегулярной структуры. В [8, 9] предложена стратегия структурированного прореживания с оценкой избыточности каналов, благодаря которой моделью типа ResNet-50 на ImageNet удалось снизить количество операций умножения и сложения при одном прямом проходе модели Floating-point Operations (FLOPs) примерно на 44 %, но при снижении точности классификации.

Это существенно ограничивает использование моделей в задачах реального времени, а также на мобильных и встраиваемых устройствах.

Основная проблема прореживания — выбор элементов сети, удаление которых минимально влияет на способность модели качественно решать задачу, например, классификации. Большинство традиционных методов прореживания опираются на локальные критерии важности параметров, такие как нормы весов, или на приближенные оценки вклада связей в итоговую ошибку. Но качество классификации определяется не только значениями весов, а также формированием внутреннего пространства признаков, в котором классы становятся разделимыми. В связи с этим перспективным является построение методов прореживания, контролирующих изменение структуры представления модели CNN [10].

В настоящей работе предлагается метод геометрически-контролируемого прореживания модели CNN, основанный на жадном выборе степени прореживания блоков CNN при ограничении на изменение структуры признакового пространства — геометрии представлений. Геометрия представлений определяется через матрицу попарных косинусных сходств между центроидами классов в пространстве признаков.

Постановка задачи

Рассмотрим задачу многоклассовой классификации изображений.

Пусть задана обучающая выборка

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

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

CNN с параметрами θ задает отображение

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

где $\mathbf{f}_\theta(x)$ — вектор логитов.

Предсказанный класс определяется как

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

Цель прореживания модели CNN заключается в построении сжатой модели $\mathbf{f}_\theta$ где θ' получается из θ путем удаления части параметров или структурных элементов CNN, при выполнении следующих требований: сохранение качества классификации на уровне исходной модели; уменьшение числа параметров и вычислительной сложности; улучшение времени выполнения — уменьшение задержки инференса.

Для оценивания качества работы предлагаемого метода используем стандартные метрики Top-1 и Top-5 точности классификации: Top-1 (точность определяется как доля примеров, для которых предсказанный класс совпадает с истинным); Top-5 (точность — как доля примеров, для которых истинный класс входит в множество пяти наиболее вероятных предсказаний).

Для оценивания степени сжатия и задержки инференса используем следующие показатели: число параметров модели P; объем модели V (определяется по суммарному размеру параметров с учетом формата хранения, в настоящей работе используются 32-битные числа с плавающей точкой); вычислительная сложность [11, 12]; задержка инференса; t — среднее время прямого прохода модели на фиксированном устройстве [13].

Задержка инференса на практике зависит не только от вычислительной сложности, но и от особенностей аппаратной платформы. Для корректности сравнения предлагаемого метода с другими подходами, измерения задержки инференса проводились в одинаковых условиях и при фиксированном размере входных данных.

Кроме метрик Top-1 и Top-5, введем понятие геометрии представлений для контроля устойчивости внутреннего пространства признаков, которые применим для анализа реализуемых методов.

Пусть $\boldsymbol{\varphi}_\theta(x) \in \mathbf{R}^d$ — вектор признаков, извлекаемый из сети на предпоследнем уровне. Тогда для каждого класса $k \in \{1, \ldots, K\}$ определим множество объектов данного класса:

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

Центроид класса запишем как [13]

$$\mu_k = \frac{1}{|D_k|}\sum_{(x,y)\in D_k}\boldsymbol{\varphi}_\theta(x).$$

Построим матрицу попарных косинусных сходств между центроидами классов [14, 15] вида

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

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, а плохо разделимые классы — низкое.

Пусть $\mathbf{S}_{Ref}$ — матрица сходств, вычисленная для базовой модели $\mathbf{f}_\theta$, а $\mathbf{S}$ — матрица сходств для модифицированной модели $\mathbf{f}_{\theta'}$ после прореживания. Тогда изменение геометрии представлений определяется величиной [11, 12]

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

где $\|\cdot\|_F$ — норма Фробениуса.

Величина ΔG позволяет количественно оценить, насколько сильно изменилось взаимное расположение классов в пространстве представлений после применения прореживания и последующего дообучения. Предполагается, что при малом ΔG сохраняется структура разделимости классов, что повышает вероятность сохранения итоговой точности [16, 17].

Описание разработанного метода геометрически-контролируемого прореживания модели сверточной нейронной сети

Работа предлагаемого метода геометрически-контролируемого прореживания модели CNN основана на следующей гипотезе: *если в результате прореживания сохраняется геометрия представлений базовой модели CNN — взаимного расположения классов в пространстве признаков, то вероятность сохранения итоговой точности после дообучения существенно возрастает*.

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

В соответствии с (1) введем матрицу сходств центроидов классов:

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

Установим ограничение вида:

$$\Delta G(\theta', \theta) \le \varepsilon, \qquad (2)$$

где ε — допустимый лимит 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 **S** и, следовательно, ΔG могут отличаться при вычислении на разных подвыборках данных [18, 19].

Для учета данного эффекта введем величину порога шума:

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

где B_1, B_2 — подвыборки данных, на которых независимо вычисляются матрицы $\mathbf{S}_1$ и $\mathbf{S}_2$ для одной и той же модели. Тогда:

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

Зададим допустимый лимит как сумму шумового порога и дополнительного допуска

$$\varepsilon = \Delta G_{noise} + \hat{\varepsilon},$$

где $\hat{\varepsilon} \ge 0$ — настраиваемый параметр метода.

Допустимый лимит обеспечивает устойчивость метода к стохастическим колебаниям метрики и позволяет интерпретировать $\hat{\varepsilon}$ как управляемый «запас» изменения геометрии [20].

Отметим, что в настоящей работе рассматривается прореживание по ширине блоков архитектуры ResNet, т. е. сокращение числа внутренних каналов в отдельных остаточных блоках. В зависимости от архитектуры используются два типа блоков: Bottleneck-блоки (ResNet-50) и BasicBlock-блоки (ResNet-18).

Для выбора каналов, подлежащих удалению, требуется оценка их относительной важности [21, 22]. В предложенном методе геометрически-контролируемого прореживания модели CNN используется прокси-критерий, основанный на статистике активаций BatchNorm [23].

Рассмотрим первый BatchNorm слой блока (обозначим его BN1), применяемый к тензору активаций $\mathbf{a} \in \mathbf{R}^{(H\times W\times C)}$. Для каждого канала $c \in \{1, \ldots, C\}$ вычислим дисперсию

$$\vartheta_c = Var(a_c),$$

где дисперсия *Var* вычисляется по пространственным координатам и по данным, собранным на калибровочном наборе [24].

Очевидно, что 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 прореживания $r \in (0, 1)$ выбирается число сохраняемых каналов:

$$C_{Save} = \max(C_{\min}, \lfloor (1 - r)C \rceil),$$

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

Выберем множество индексов сохраняемых каналов:

$$\zeta = \mathrm{Top}K(\{\vartheta_c\}_{c=1}^{C}, C_{Save}).$$

Метод геометрически-контролируемого прореживания модели CNN использует жадную стратегию. Пусть задан дискретный набор кандидатов долей прореживания:

$$R = \{r_1, r_2, \ldots, r_m\}, \quad 0 < r_1 < \ldots < r_m < 1.$$

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

Для каждого остаточного блока сети выполняется перебор $r \in R$ и выбирается максимально возможное r, при котором действует ограничение (2).

Если ни один кандидат не удовлетворяет ограничению, блок не сжимается. Таким образом, метод автоматически регулирует степень прореживания в зависимости от чувствительности блоков к нарушению геометрии представлений.

Предложенный метод можно интерпретировать как процедуру «аккуратного» структурного прореживания, при которой CNN упрощается постепенно, но только до тех пор, пока сохраняется внутренняя организация признакового пространства. Сначала для базовой модели вычисляется эталонная геометрия представлений, описывающая взаимное расположение классов в пространстве признаков, далее оценивается естественный уровень колебаний этой метрики (порог шума), возникающий из-за того, что геометрия вычисляется по ограниченному числу наборов данных и после этого задается допустимый лимит изменения геометрии, включающий шумовую составляющую и дополнительный запас.

Затем сеть просматривается по блокам (в выбранных стадиях ResNet), и для каждого блока алгоритм пытается уменьшить его ширину, удаляя наименее информативные каналы. В качестве быстрой оценки информативности каналов используется дисперсия активаций BatchNorm: каналы с малой дисперсией считаются менее значимыми и удаляются в первую очередь. Для каждого блока проверяются несколько вариантов степени прореживания (от более сильного к более мягкому). Каждый вариант принимается только в том случае, если после изменения блока глобальная геометрия представлений остается достаточно близкой к эталонной, т. е. ΔG не превышает установленный лимит. Таким образом, разработанный метод автоматически выбирает максимально возможную степень упрощения для каждого блока, не разрушая структуру разделимости классов. После завершения прореживания

выполняется дообучение, позволяющее восстановить точность модели при уменьшенной вычислительной сложности.

Принципиальное отличие предложенного метода от классических методов прореживания состоит в том, что решение об удалении каналов принимается с учетом глобального свойства модели — геометрии представлений классов. Фактически метод вводит дополнительный критерий качества сжатия: допустимо удалять параметры лишь до тех пор, пока сохраняется структура взаимного расположения классов в признаковом пространстве, сформированная базовой моделью. Это позволяет снизить риск разрушения разделимости классов при сильном структурном упрощении CNN.

Постановка эксперимента

Экспериментальные исследования проводились на наборе данных CIFAR-100[1], содержащем 60 000 цветных изображений размером 32 × 32 пиксела, распределенных по 100 классам. Обучающая выборка включает 50 000 изображений, тестовая — 10 000. Данный набор данных является стандартным для оценки методов сжатия и ускорения CNN.

В качестве основной архитектуры рассматривается ResNet-50, как типичный представитель глубоких остаточных сетей с Bottleneck-блоками, обладающий значительным числом параметров и высокой вычислительной сложностью.

Поскольку исходная реализация ResNet рассчитана на изображения большего разрешения (например, ImageNet), для корректной работы на CIFAR-100 применялась стандартная модификация начальной части сети: свертка 7×7 заменена на 3×3 с шагом 1, maxpool-операция удалена, выходной полносвязный слой адаптирован под 100 классов. Такая модификация позволяет избежать чрезмерного уменьшения пространственного разрешения признаков на ранних этапах сети и обеспечивает более корректную настройку архитектуры под CIFAR-100.

Для сравнения предложенного метода с базовым (без прореживаний) и с прореживанием по глубине (количеству слоев) использован единый протокол обучения при равном вычислительном бюджете. Он включает следующие этапы.

Этап 1. Обучение базовой модели (baseline-40) в течение 40 эпох на обучающей выборке CIFAR-100, которая сохраняется в качестве контрольной точки.

Этап 2. Для оценки эффекта дополнительного обучения проводится продолжение оптимизации базовой модели еще на 20 эпохах без применения прореживания.

Этап 3. Для метода прореживания по глубине и предложенного метода применялась единая схема: baseline-40 → прореживание → дообучение 20 эпох.

Таким образом, модели CNN 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 baseline-60 без прореживания, прореживание по глубине и метод геометрически-контролируемого прореживания.

В экспериментах использовались следующие параметры: доля разреженности $r = 0{,}7$, размер обучающей выборки 128 изображений. Для метода геометрически-контролируемого прореживания рассматриваемые стадии ResNet прореживанию подвергались группы остаточных блоков conv3_x и conv4_x, отвечающих за извлечение сложных признаков из изображений и пространстве, а число данных для вычисления геометрии и порога шума ограничивался калибровочным набором.

Для повышения достоверности результатов основные эксперименты на ResNet-50 проводились с тремя фиксированными начальными условиями генераторов случайных чисел, что позволяет оценить устойчивость методов к стохастическим факторам обучения и дообучения. После чего вычислены средние значения и стандартные отклонения.

Анализ результатов

В табл. 1 приведены результаты контрольного обучения для трех методов.

На рис. 1 приведена диаграмма Top-1 и Top-5 точности.

Базовая модель (этап 1), обученная на 40 эпохах, показывает, что точность для Top-1 равна 58,33 %, тогда как после продолжения обучения (этап 2) до

Таблица 1. Результаты контрольного обучения

Table 1. Results of the control training

Метод	Метрики					
	Top-1, % (mean ± std)	Top-5, % (mean ± std)	P, млн	V, МБ	FLOPs, млрд	t, мс (mean ± std)
Baseline-60	73,34 ± 0,13	92,73 ± 0,08	23,71	90,43	1,31	7,73 ± 0,55
Прореживание по глубине	60,84 ± 0,41	86,40 ± 0,36	15,26	58,21	0,81	7,12 ± 3,06
Геометрически-контролируемое прореживание	71,72 ± 0,32	91,81 ± 0,21	10,19	38,86	0,53	6,99 ± 1,03

Примечание: mean — среднее значение метрики, полученной при разных фиксированных начальных условиях генератора случайных чисел; std — стандартное отклонение.

[1] Krizhevsky A., Hinton G. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL: https://www.cs.toronto.edu/~kriz/cifar.html (дата обращения: 23.04.2026).

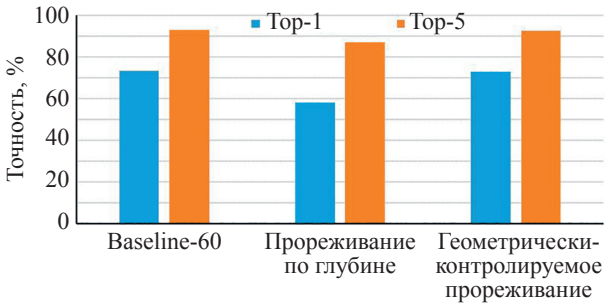

Рис. 1. Диаграмма Top-1 и Top-5 точности методов
Fig. 1. Diagram of Top-1 and Top-5 accuracy of methods

60 эпох — возрастает до 73,34 %. Это подтверждает, что корректное сравнение методов прореживания требует фиксации вычислительного бюджета обучения, поскольку дополнительное дообучение существенно улучшает качество.

Прореживание по глубине (этап 3), сокращает число параметров примерно на 36 % (с 23,71 млн до 15,26 млн) и уменьшает вычислительную сложность FLOPs с 1,31 млрд до 0,81 млрд, однако итоговая точность остается на уровне около 60,84 % что существенно ниже baseline-60. Это отражает типичный компромисс классической разреженности: реальная аппаратная эффективность достигается ценой уменьшения способности модели выдавать качественный результат. В свою очередь предложенный метод геометрически-контролируемого прореживания сокращает число параметров на 50 % и уменьшение FLOPs чуть более на 50 %. Задержка инференса при прореживании остается примерно такой же — 7 мс, что объясняется ограничением аппаратных особенностей графического процессорного устройства.

Покажем результаты работы предложенного метода для модели ResNet-50 на наборе данных CIFAR-100 (табл. 2). В эксперименте изменялся параметр $\hat{\varepsilon}$, определяющий дополнительный допустимый лимит изменения геометрии представлений. Для каждого значения $\hat{\varepsilon} \in \{0{,}06;\ 0{,}10;\ 0{,}14;\ 0{,}18\}$ выполнялись независимые запуски с различными начальными значениями генератора случайных чисел.

Полученные результаты демонстрируют, что увеличение $\hat{\varepsilon}$ приводит к более агрессивному структурному изменению (упрощению) модели: число параметров снижается с 22,88 млн при $\hat{\varepsilon} = 0{,}06$ до 20,80 млн при $\hat{\varepsilon} = 0{,}18$, а вычислительная сложность уменьшается с 1,256 млрд FLOPs до 1,212 млрд FLOPs. Таким образом, параметр $\hat{\varepsilon}$ обеспечивает интерпретируемое управление компромиссом между степенью сжатия и ограничением на изменение 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 %. Это означает, 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 $\hat{\varepsilon}$, что отражает факт более сильного вмешательства в архитектуру сети (рис. 2). Однако после дообучения ΔG стабилизируется в пределах 0,14−0,15, что позволяет предположить наличие эффекта частичного восстановления структуры представлений в процессе дообучения.

Рост ΔG после дообучения может быть связан с тем, что в процессе дообучения оптимизация функции потерь ориентирована на повышение точности классификации и не содержит явного ограничения на сохранение геометрии. Следовательно, дообучение может приводить к дополнительной перестройке признакового пространства, не обязательно совпадающей с эталонной геометрией модели. Таким образом, наблюдение подтверждает целесообразность использования геометрического ограничения именно на этапе прореживания, когда риск разрушения структуры представлений максимален.

Для исключения влияния различий в степени сжатия дополнительно проведено сравнение в режиме равного количества параметров. Для прореживания по глубине был подобран уровень разреженности 0,3, обеспечивающий число параметров 21,19 млн, что сопоставимо с методом геометрически-контролируемого прореживания при $\hat{\varepsilon} = 0{,}18$ (20,80 млн). Несмотря на близкое сходство числа параметров (разница ≈1,9 %) и меньшую вычислительную сложность (1,095 против 1,212 млрд FLOPs), прореживание по глубине обеспечивает Top-1 = 59,93 %, тогда как геометрически-контролируемое прореживание достигает 73,21 %. Полученный результат подтверждает, что предложенный метод сохраняет точность за счет контроля геометрии представлений, а не за счет большего размера модели.

Таблица 2. Результаты работы метода геометрически-контролируемого прореживания
Table 2. Results of the geometrically controlled thinning method

$\hat{\varepsilon}$	Метрики						
	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

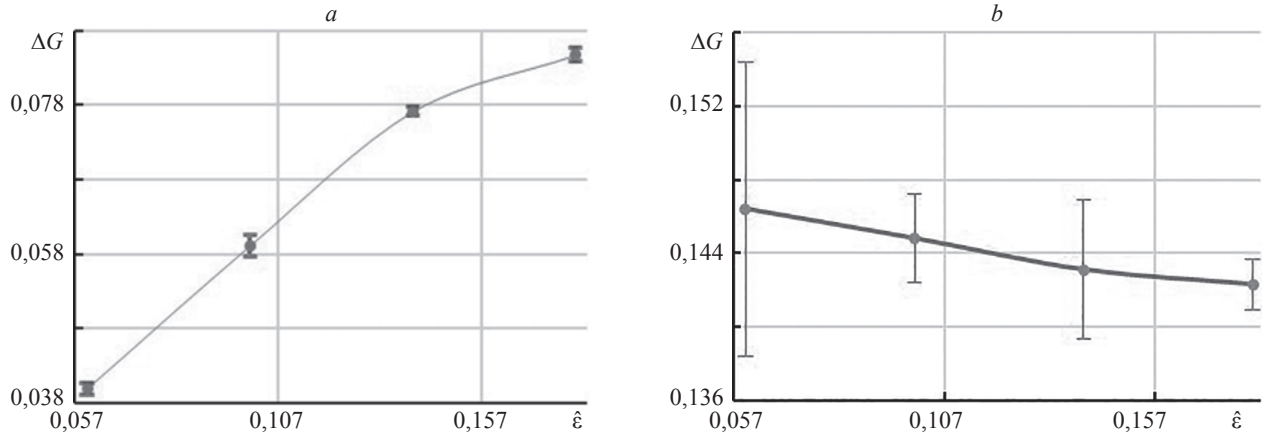

Рис. 2. Значения ΔG: до (*a*) и после (*b*) дообучения

Fig. 2. ΔG value: before (*a*) and after (*b*) fine-tuning

Заключение

В работе предложен новый метод сжатия сверточных нейронных сетей (Convolutional Neural Network, CNN), основанный на геометрически-контролируемом прореживании модели, в котором прореживание выполняется жадно по блокам CNN, а допустимость каждого шага определяется не только локальной важностью каналов, но и глобальным ограничением на изменение геометрии представлений.

В качестве метрики сохранения структуры признакового пространства используется изменение матрицы межклассовых сходств, вычисляемой по центроидам классов в пространстве признаков. Введен параметр допустимого изменения геометрии представлений и предложен подход к его выбору на основе оценки шумового порога геометрической метрики.

Результаты проведенного эксперимента показали, что предложенный метод геометрически-контролируемого прореживания модели CNN обеспечивает устойчивое снижение вычислительной сложности модели при сохранении качества классификации. Показано, что параметр допустимого изменения геометрии представлений обеспечивает интерпретируемое управление компромиссом между степенью сжатия модели CNN и ограничением на изменение геометрии представлений. При изменении параметра допустимого изменения геометрии представлений метод демонстрирует сопоставимую точность классификации по сравнению с базовой моделью после дообучения, одновременно снижая вычислительную сложность и число параметров модели CNN.

Полученные результаты подтверждают перспективность использования геометрических ограничений как критерия качества при сжатии CNN, а также открывают возможности для дальнейших исследований, включая оптимальное прореживание по слоям и расширение на другие архитектуры.

Литература

1. Dantas P.V., da Silva W.S., Cordeiro L.C., Carvalho C.B. A comprehensive review of model compression techniques in machine learning // Applied Intelligence. 2024. V. 54. N 22. P. 11804–11844. doi: 10.1007/s10489-024-05747-w
2. 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
3. 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
4. Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., de Laroussilhe Q., Gesmundo A., et al. Parameter-efficient transfer learning for NLP // Proc. of the 36th International Conference on Machine Learning. 2019. V. 97. P. 2790–2799.
5. Chen J., Hu Y., Zhu J. S-STE: Continuous pruning function for efficient 2:4 sparse pre-training // Proc. of the 38th Conference on Neural Information Processing Systems (NeurIPS). 2024. P. 33756–33778. doi: 10.52202/079017-1063
6. Frantar E., Alistarh D. SparseGPT: massive language models can be accurately pruned in one-shot // Proc. of the 40th International Conference on Machine Learning. 2023. V. 202. P. 10323–10337.

References

1. Dantas P.V., da Silva W.S., 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. 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, no. 2 (73), pp. 83–92. (in Russian). doi: 10.38161/1996-3440-2024-2-83-92
3. 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
4. Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., de Laroussilhe Q., Gesmundo A., et al. Parameter-efficient transfer learning for NLP. *Proc. of the 36th International Conference on Machine Learning*, 2019, vol. 97, pp. 2790–2799.
5. Chen J., Hu Y., Zhu J. S-STE: Continuous pruning function for efficient 2:4 sparse pre-training. *Proc. of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024, pp. 33756–33778. doi: 10.52202/079017-1063
6. Frantar E., Alistarh D. SparseGPT: massive language models can be accurately pruned in one-shot. *Proc. of the 40th International Conference on Machine Learning*, 2023, vol. 202, pp. 10323–10337.

7. Wang Z., Li C., Wang X. Convolutional neural network pruning with structural redundancy reduction // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 14908–14917. doi: 10.1109/cvpr46437.2021.01467
8. 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.
9. Liu D., Zhu Y., Liu Z., Liu Y., Han C., Tian J. A survey of model compression techniques: past, present, and future // Frontiers in Robotics and AI. 2025. V. 12. P. 1518965. doi: 10.3389/frobt.2025.1518965
10. He Y., Xiao L. Structured pruning for deep convolutional neural networks: A survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. V. 46. N 5. P. 2900–2919. doi: 10.1109/TPAMI.2023.3334614
11. Мокрецов Н.С., Татарникова Т.М. Оптимизация процесса обучения при ограниченном объеме вычислительных ресурсов // Международная конференция по мягким вычислениям и измерениям. 2024. Т. 1. С. 205–208.
12. Lv K., Yang Y., Liu T., Guo Q., Qiu X. Full parameter fine-tuning for large language models with limited resources // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. V. 1. P. 8187–8198. doi: 10.18653/v1/2024.acl-long.445
13. Zhang Q., Zhang R., Sun J., Liu Y. How sparse can we prune a deep network: a fundamental limit perspective // Proc. of the 38th International Conference on Neural Information Processing System. 2024. P. 91337–91372.
14. Cheng H., Zhang M., Shi J.Q. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. V. 46. N 12. P. 10558–10578. doi: 10.1109/tpami.2024.3447085
15. 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. V. 626. P. 129382. doi: 10.1016/j.neucom.2025.129382
16. 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
17. Park C., Park M., Moon H., Yoon M.K., Go S., Kim S., et al. DEPrune: Depth-wise separable convolution pruning for maximizing GPU parallelism // Proc. of the 38th International Conference on Neural Information Processing Systems. 2024. P. 106906–106923.
18. Liao B., Meng Y., Monz C. Parameter-efficient fine-tuning without introducing new latency // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. V. 1. P. 4242–4260. doi: 10.18653/v1/2023.acl-long.233
19. Lasby M., Golubeva A., Evci U., Nica M., Ioannou Y. Dynamic sparse training with structured sparsity // Proc. of the International Conference on Learning Representations (ICLR). 2024. P. 32982–33010.
20. Sun X., Shi H. Towards better structured pruning saliency by reorganizing convolution // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024. P. 2193–2203. doi: 10.1109/WACV57701.2024.00220
21. Wright D., Igel C., Selvan R. BMRS: Bayesian model reduction for structured pruning // Proc. of the 38th International Conference on Neural Information Processing Systems. 2024. P. 64119–64144.
22. Dong Z., Duan Y., Zhou Y., Duan S., Hu X. Weight-adaptive channel pruning for CNNs based on closeness-centrality modeling // Applied Intelligence. 2024. V. 54. N 1. P. 201–215. doi: 10.1007/s10489-023-05164-5
23. 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. 2025. V. 9. N 2. P. 190–208. doi: 10.1080/24751839.2024.2415008
24. 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. V. 46. N 1. P. 436–450. doi: 10.1109/tpami.2023.3322904

7. Wang Z., Li C., Wang X. Convolutional neural network pruning with structural redundancy reduction. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14908–14917. doi: 10.1109/cvpr46437.2021.01467
8. Chernyshov N.D., Buryak D.YU. Esearch 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, no. 1, pp. 16–22. (in Russian)
9. Liu D., Zhu Y., Liu Z., Liu Y., Han C., Tian J. A survey of model compression techniques: past, present, and future. *Frontiers in Robotics and AI*, 2025, vol. 12, pp. 1518965. doi: 10.3389/frobt.2025.1518965
10. He Y., Xiao L. Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, no. 5, pp. 2900–2919. doi: 10.1109/TPAMI.2023.3334614
11. Mokretsov N.S., Tatarnikova T.M. Optimizing the learning process with limited computational resources. *International Conference on Soft Computing and Measurement*, 2024, vol. 1, pp. 205–208. (in Russian)
12. Lv K., Yang Y., Liu T., Guo Q., Qiu X. Full parameter fine-tuning for large language models with limited resources. *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, vol. 1, pp. 8187–8198. doi: 10.18653/v1/2024.acl-long.445
13. Zhang Q., Zhang R., Sun J., Liu Y. How sparse can we prune a deep network: a fundamental limit perspective. *Proc. of the 38th International Conference on Neural Information Processing System*, 2024, pp. 91337–91372.
14. Cheng H., Zhang M., Shi J.Q. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, no. 12, pp. 10558–10578. doi: 10.1109/tpami.2024.3447085
15. 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, pp. 129382. doi: 10.1016/j.neucom.2025.129382
16. 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
17. Park C., Park M., Moon H., Yoon M.K., Go S., Kim S., et al. DEPrune: Depth-wise separable convolution pruning for maximizing GPU parallelism. *Proc. of the 38th International Conference on Neural Information Processing Systems*, 2024, pp. 106906–106923.
18. Liao B., Meng Y., Monz C. Parameter-efficient fine-tuning without introducing new latency. *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, vol. 1, pp. 4242–4260. doi: 10.18653/v1/2023.acl-long.233
19. Lasby M., Golubeva A., Evci U., Nica M., Ioannou Y. Dynamic sparse training with structured sparsity. *Proc. of the International Conference on Learning Representations (ICLR)*, 2024, pp. 32982–33010.
20. Sun X., Shi H. Towards better structured pruning saliency by reorganizing convolution. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 2193–2203. doi: 10.1109/WACV57701.2024.00220
21. Wright D., Igel C., Selvan R. BMRS: Bayesian model reduction for structured pruning. *Proc. of the 38th International Conference on Neural Information Processing Systems*, 2024, pp. 64119–64144.
22. 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, no. 1, pp. 201–215. doi: 10.1007/s10489-023-05164-5
23. 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*, 2025, vol. 9, no. 2, pp. 190–208. doi: 10.1080/24751839.2024.2415008
24. 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, no. 1, pp. 436–450. doi: 10.1109/tpami.2023.3322904

Авторы

Татарникова Татьяна Михайловна — доктор технических наук, профессор, директор института, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, SC 36715607400, https://orcid.org/0000-0002-6419-0072, tm-tatarn@yandex.ru

Раскопина Анастасия Сергеевна — аспирант, ассистент, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, SC 59946913200, https://orcid.org/0009-0002-0276-607X, raskopina.anastasia@yandex.ru

Authors

Tatiana M. Tatarnikova — D.Sc., Professor, Director of the Institute, Saint Petersburg State University of Aerospace Instrumentation (SUAI), Saint Petersburg, 190000, Russian Federation, SC 36715607400, https://orcid.org/0000-0002-6419-0072, tm-tatarn@yandex.ru

Anastasia S. Raskopina — PhD Student, Assistant, Saint Petersburg State University of Aerospace Instrumentation (SUAI), Saint Petersburg, 190000, Russian Federation, SC 59946913200, https://orcid.org/0009-0002-0276-607X, raskopina.anastasia@yandex.ru

Статья поступила в редакцию 07.02.2026
Одобрена после рецензирования 20.04.2026
Принята к печати 23.05.2026

Received 07.02.2026
Approved after reviewing 20.04.2026
Accepted 23.05.2026