Адаптивный структурированный прунинг нейронных сетей с сохранением геометрии представлений

Т. М. Татарникова, А. С. Раскопина

Санкт-Петербургский государственный университет аэрокосмического приборостроения

tm-tatarn@yandex.ru, raskopina.anastasia@yandex.ru

***Аннотация*. В работе предлагается адаптивный метод структурированного прунинга нейронных сетей, основанный на контроле сохранения геометрии представлений классов во внутреннем пространстве признаков модели. В отличие от классических методов, удаляющих каналы или слои на основе локальных критериев важности, предложенный подход выполняет глобальный контроль изменения взаимного расположения классов после прунинга. Это позволяет уменьшать размер и вычислительную сложность модели без существенной деградации точности. Предложен алгоритм адаптивного выбора степени прунинга, автоматически подбирающий допустимое сокращение архитектуры в зависимости от устойчивости геометрии представлений.**

Ключевые слова: структурированный прунинг, нейронные сети, сжатие моделей, геометрия представлений, пространство признаков, адаптивный прунинг.

I. Введение

В последние годы глубокие нейронные сети стали основным инструментом решения широкого круга задач компьютерного зрения, включая классификацию изображений, детекцию объектов и сегментацию. Современные архитектуры сверточных нейронных сетей обеспечивают высокое качество распознавания, однако достигается это ценой значительной вычислительной сложности, большого числа параметров и высокого энергопотребления. Эти факторы существенно ограничивают применение моделей глубокого обучения в задачах реального времени, а также на мобильных и встраиваемых устройствах с ограниченными ресурсами [1–3].

Одним из ключевых направлений решения данной проблемы является сжатие нейронных сетей, направленное на уменьшение числа параметров и ускорение инференса при сохранении приемлемого уровня точности. Среди существующих методов сжатия особое место занимает прореживание (pruning) [4], предполагающее удаление части параметров или структурных элементов модели. В зависимости от структуры удаляемых элементов различают неструктурированное, структурированное и полуструктурированное прореживание [5, 6].

Неструктурированное прореживание позволяет достигать высокой степени разреженности за счёт обнуления отдельных весов, однако не обеспечивает существенного ускорения вычислений на стандартных аппаратных платформах из-за нерегулярной структуры данных. В свою очередь, структурированное прореживание, основанное на удалении каналов, фильтров или блоков сети, позволяет непосредственно уменьшить вычислительную сложность и ускорить выполнение модели, но часто сопровождается значительной деградацией качества [7].

Основной причиной ухудшения качества при прореживании является использование локальных критериев важности параметров, таких как нормы весов или статистики активаций. Эти критерии не учитывают глобальную структуру внутреннего пространства признаков, формируемого нейронной сетью, в котором обеспечивается разделимость классов. В результате удаление структурных элементов может приводить к существенной деформации пространства представлений и ухудшению классиционной способности модели.

В последние годы внимание исследователей смещается в сторону методов, учитывающих свойства признакового пространства. В частности, предлагаются подходы, основанные на анализе сходства признаков, корреляции каналов или кластеризации фильтров. Однако такие методы либо требуют значительных вычислительных затрат, либо не обеспечивают явного контроля глобальной структуры представлений [8].

Перспективным направлением является использование геометрии представлений как критерия качества при прореживании. В этом подходе структура пространства признаков описывается через взаимное расположение классов, например, с использованием центроидов и метрик сходства. Ограничение изменения геометрии позволяет контролировать степень деформации признакового пространства и снижает риск разрушения разделимости классов.

В ранее предложенных методах геометрически контролируемого прореживания вводится ограничение на допустимое изменение геометрии представлений, что позволяет выполнять структурное упрощение модели без существенной потери точности. Однако такие методы, как правило, используют фиксированные гиперпараметры, задающие допустимый уровень деформации, и требуют ручной настройки степени прореживания. Это ограничивает их универсальность и усложняет применение в различных архитектурах и задачах.

Таким образом, возникает задача разработки методов прореживания, способных автоматически адаптировать степень структурного упрощения модели в зависимости от свойств данных и внутреннего пространства признаков.

Целью данной работы является разработка адаптивного метода структурированного прореживания нейронных сетей, основанного на контроле геометрии представлений, который автоматически определяет допустимый уровень деформации признакового пространства и соответствующую степень прореживания.

В работе предлагается расширение геометрически-контролируемого подхода за счёт введения адаптивных механизмов, включающих автоматический выбор бюджета изменения геометрии представлений, динамическую настройку степени прореживания и учет различной чувствительности слоев сети к структурным изменениям. Это позволяет устранить необходимость ручной настройки параметров и повысить устойчивость метода.

Основные научные результаты работы**:**

- предложен адаптивный метод структурированного прореживания, учитывающий геометрию представлений;
- разработан механизм автоматического определения допустимого изменения геометрии;
- предложена стратегия адаптивного выбора степени прореживания и её распределения по слоям сети.

II. Постановка задачи

Рассматривается задача структурированного прореживания сверточных нейронных сетей в условиях ограничения вычислительных ресурсов. Целью является уменьшение числа параметров и вычислительной сложности модели при сохранении качества классификации.

В отличие от традиционных методов, в данной работе используется подход, основанный на контроле геометрии представлений во внутреннем пространстве признаков. Под геометрией представлений понимается взаимное расположение классов, формируемое нейронной сетью на уровне высокоуровневых признаков.

Интуитивно предполагается, что значительная деформация геометрии представлений при структурном упрощении модели может приводить к ухудшению разделимости классов и, как следствие, снижению точности классификации. В связи с этим процесс прореживания рассматривается как задача поиска компромисса между уменьшением сложности модели и сохранением структуры признакового пространства.

В ранее предложенных методах геометрически-контролируемого прореживания вводится ограничение на допустимое изменение геометрии представлений, что позволяет ограничить степень деформации внутреннего пространства признаков. Однако такие методы используют фиксированные гиперпараметры, задающие допустимый уровень изменений ΔG, а также требуют предварительного выбора степени прореживания.

Это приводит к ряду ограничений: необходимость ручной настройки параметров для каждой архитектуры;

зависимость результата от выбора набора коэффициентов прореживания; отсутствие учета различной чувствительности слоев сети к структурным изменениям.

Таким образом, возникает задача разработки метода, который способен автоматически адаптировать степень структурного упрощения модели на основе анализа изменений геометрии представлений.

В данной работе предлагается рассматривать процесс прореживания как адаптивную процедуру, в которой параметры метода определяются динамически в зависимости от реакции модели на структурные изменения.

Формально задача может быть интерпретирована как поиск такой упрощенной модели, при которой минимизируется вычислительная сложность при условии, что изменение геометрии представлений не превышает некоторого допустимого уровня:

$$\Delta G \leq \tau,$$

где величина τ не задаётся заранее, а определяется адаптивно в процессе работы алгоритма.

Таким образом, требуется разработать метод, который:

1. автоматически определяет допустимый уровень деформации геометрии представлений;
2. адаптивно выбирает степень прореживания модели;
3. учитывает различную чувствительность структурных элементов сети.

III. Предлагаемый метод

В данной работе предлагается адаптивный метод структурированного прореживания нейронных сетей, основанный на контроле геометрии представлений. В отличие от базового подхода, в котором параметры прореживания задаются вручную, предлагаемый метод автоматически определяет допустимую степень структурного упрощения модели на основе анализа изменений внутреннего пространства признаков.

Процесс прореживания выполняется итеративно. На каждом шаге рассматриваются кандидаты структурного упрощения модели, и для каждого варианта оценивается изменение геометрии представлений.

В отличие от дискретного перебора фиксированного набора коэффициентов прореживания, предлагаемый метод использует механизм обратной связи: степень прореживания корректируется в зависимости от того, насколько текущее изменение геометрии близко к допустимому пределу.

Таким образом, процесс прореживания формулируется как управляемый итеративный процесс, в котором параметры адаптируются в ходе выполнения алгоритма.

В базовом подходе допустимое изменение геометрии задаётся фиксированным параметром. В данной работе предлагается определять этот параметр автоматически на основе статистических свойств метрики изменения геометрии.

Пусть при выполнении пробных шагов прореживания получен набор значений изменения геометрии:

$$\{\Delta G_1, \Delta G_2, ..., \Delta G_M\}.$$

Тогда адаптивный бюджет определяется как:

$$\tau = G_{\text{noise}} + \beta \times \sigma \times \Delta G,$$

где G_{noise} – оценка шумового уровня метрики; $\sigma \cdot \Delta G$ – стандартное отклонение значений ΔG; β – коэффициент запаса.

Такое определение позволяет учитывать естественную вариативность метрики и автоматически подстраивать допустимый уровень деформации геометрии представлений.

Вместо использования фиксированного набора коэффициентов прореживания предлагается динамическая стратегия обновления степени прореживания.

Пусть r_t – текущая степень прореживания на шаге t, тогда обновление задаётся следующим образом:

$$r_{t+1} = \begin{cases} r_t + \delta, \Delta G < \tau \\ r_t - \gamma, \Delta G > \tau, \end{cases}$$

где δ, γ – параметры шага увеличения и уменьшения степени прореживания.

Таким образом, если текущая деформация геометрии мала, алгоритм увеличивает степень сжатия, а при превышении допустимого уровня – ослабляет прореживание.

Данный механизм позволяет автоматически находить максимально допустимую степень структурного упрощения модели без нарушения геометрии представлений.

Различные слои нейронной сети обладают разной чувствительностью к структурным изменениям. В частности, глубокие слои, формирующие высокоуровневые признаки, как правило, более чувствительны к прореживанию.

Для учета этого эффекта предлагается распределять общий бюджет геометрии по слоям адаптивным образом.

Пусть ΔG_L изменение геометрии при прореживании L-го слоя. Тогда локальный бюджет определяется как:

$$\tau_L = \tau \cdot \frac{1}{\Delta G_L + \varepsilon}.$$

Таким образом, слои с высокой чувствительностью получают более строгие ограничения, тогда как менее чувствительные слои могут быть прорежены более агрессивно.

Предлагаемый метод может быть представлен следующим образом:

1. оценка базовой геометрии представлений модели;
2. выполнение пробных шагов прореживания;
3. оценка статистики изменения геометрии;
4. вычисление адаптивного бюджета τ;
5. итеративное обновление степени прореживания r;
6. адаптивное распределение прореживания по слоям;
7. формирование итоговой прореженной модели.

Предложенный подход отличается от существующих методов следующими свойствами:

- отсутствует необходимость ручной настройки степени прореживания;
- учитывается стохастическая природа метрики геометрии;
- обеспечивается адаптация к архитектуре и данным;
- реализуется глобальный контроль структуры признакового пространства.

В отличие от базового геометрически-контролируемого прореживания, где степень сжатия определяется через фиксированный параметр, предложенный метод автоматически находит компромисс между степенью упрощения модели и сохранением геометрии представлений.

IV. Результаты

Эксперименты проводились на наборе данных CIFAR-100, содержащем 60 000 цветных изображений размером 32×32, распределённых по 100 классам. В качестве базовых архитектур использовались модели ResNet-50, ResNet-18 и MobileNetV2. Для обеспечения корректного сравнения использовался единый протокол обучения. Базовые модели обучались в течение 40 эпох, после чего выполнялось дообучение в течение 20 эпох.

Для всех методов прореживания применялась схема: обучение → прореживание → дообучение, что обеспечивает сопоставимость результатов при равном вычислительном бюджете.

ТАБЛИЦА I. Сравнение методов прореживания на ResNet-50

Метод	**Top-1 (%)**	**Параметры (M)**	**FLOPs**	**ΔG**
Baseline	73,3	23,7	1,31	0
GC	73,2	21,1	1,23	0,07
Adaptive GC	73,2	21,0	1,22	Auto

Экспериментально наблюдается, что адаптивный механизм приводит к выбору различных степеней прореживания для разных архитектур и слоев сети. Это подтверждает, что чувствительность модели к структурным изменениям существенно зависит от её структуры.

ТАБЛИЦА II. Сравнение методов прореживания на MobileNetV2

Методы	**Top-1 (%)**	**Параметры (M)**	**FLOPs**	**Δ*G***
Baseline	59,78	2,35	26,17	0
GC	59,50	2,22	24,30	0,07
Adaptive GC	59,50	2,17	23,93	Auto

В частности: для архитектуры ResNet-50 (табл. 1) адаптивный метод выбирает умеренные значения прореживания, обеспечивающие сохранение точности; для MobileNetV2 (табл. 2) наблюдается более агрессивное прореживание при сопоставимом уровне качества; для ResNet-18 (табл. 3) степень прореживания варьируется в зависимости от глубины слоя.

ТАБЛИЦА III. СРАВНЕНИЕ МЕТОДОВ ПРОРЕЖИВАНИЯ НА RESNET-18

Методы	**Top-1 (%)**	**Параметры (M)**	**FLOPs**	**Δ*G***
Baseline	71,63	11,22	557,8	0
GC	71,65	10,50	540,0	0,06
Adaptive *GC*	71,72	10,19	531,2	Auto

Таким образом, метод автоматически подстраивается под архитектуру модели без необходимости ручной настройки параметров. Для анализа эффективности адаптивного подхода проведено сравнение с базовым геометрически-контролируемым методом, в котором параметр допустимого изменения задаётся вручную.

Результаты показывают, что:

- адаптивный метод достигает сопоставимых значений точности;
- при этом устраняется необходимость подбора параметра; достигается более стабильное поведение при различных начальных условиях.

Дополнительно наблюдается, что адаптивный метод в ряде случаев выбирает более оптимальные степени прореживания, чем фиксированные значения, используемые в базовом подходе.

Проведён анализ распределения степени прореживания по слоям сети. Установлено, что предлагаемый метод автоматически ограничивает прореживание в слоях, оказывающих наибольшее влияние на геометрию представлений, и, напротив, более активно сокращает менее чувствительные части модели.

V. ЗАКЛЮЧЕНИЕ

В работе предложен адаптивный метод структурированного прореживания нейронных сетей, основанный на контроле геометрии представлений во внутреннем пространстве признаков. В отличие от существующих подходов, использующих фиксированные параметры и требующих ручной настройки степени прореживания, предложенный метод автоматически определяет допустимый уровень структурного упрощения модели.

Ключевым отличием разработанного подхода является введение механизмов адаптации, включающих автоматическое определение допустимого изменения геометрии представлений, динамический выбор степени прореживания и адаптивное распределение структурных изменений по слоям сети. Это позволяет учитывать различную чувствительность архитектурных компонентов модели и повышает устойчивость метода.

Проведённые эксперименты показали, что предлагаемый метод обеспечивает снижение вычислительной сложности и числа параметров модели при сохранении точности классификации на уровне базовой модели. При этом устраняется необходимость подбора гиперпараметров, что упрощает применение метода на различных архитектурах и наборах данных.

Дополнительно установлено, что адаптивный подход позволяет автоматически находить рациональный компромисс между степенью сжатия модели и сохранением структуры признакового пространства, что особенно важно при работе с ресурсно-ограниченными устройствами.

Полученные результаты подтверждают перспективность использования геометрии представлений в сочетании с адаптивными механизмами для разработки методов сжатия нейронных сетей. В дальнейшем представляется целесообразным исследовать расширение предложенного подхода на другие типы архитектур, а также интеграцию геометрических ограничений непосредственно в процесс обучения модели.

СПИСОК ЛИТЕРАТУРЫ

[1] P. V. Dantas, W. Sabino da Silva, L. C. Cordeiro, and C. B. Carvalho, “A comprehensive review of model compression techniques in machine learning,” *Applied Intelligence*, vol. 54, no. 22, pp. 11804–11844, 2024.

[2] X. Chen, Y. Cheng, S. Wang, and L. Gan, “Data-free model compression and acceleration: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 12001–12025, 2023.

[3] Y. He and L. Xiao, “Structured pruning for deep convolutional neural networks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2900–2919, 2024.

[4] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 1135–1143.

[5] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in *International Conference on Learning Representations (ICLR)*, 2019.

[6] Q. Zhang, A. Kumar, J. Smith, et al., “How sparse can we prune a deep network: A fundamental limit perspective,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[7] T. Ganguli and E. Chong, “Activation-based pruning of neural networks,” *Algorithms*, vol. 17, no. 1, p. 48, 2024.

[8] C. Zhao, Y. Zhang, and B. Ni, “Exploiting channel similarity for network pruning,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.