НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2024 Том 24 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2024 Vol. 24 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-5-770-778
УДК 004.89

Метод оптимизации нейронных сетей на основе структурной дистилляции с применением генетического алгоритма

Владимир Никифорович Кузьмин[1], Артем Бакытжанович Менисов[2]✉, Тимур Римович Сабиров[3]

[1,2,3] Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация

[1] vka@mil.ru, https://orcid.org/0000-0002-6411-4336
[2] vka@mil.ru✉, https://orcid.org/0000-0002-9955-2694
[3] vka@mil.ru, https://orcid.org/0000-0002-6807-2954

Аннотация

Введение. По мере усложнения нейронных сетей увеличивается количество параметров и необходимых вычислений, что затрудняет установку и эксплуатацию систем искусственного интеллекта на периферийных устройствах. Структурная дистилляция может существенно сократить ресурсоемкость применения любых нейронных сетей. **Метод.** В работе представлен метод оптимизации нейронных сетей, который сочетает в себе преимущества структурной дистилляции и генетического алгоритма. В отличие от эволюционных подходов, используемых при поиске оптимальной архитектуры или дистилляции нейронных сетей, при формировании вариантов дистилляции предлагается кодировать не только параметры нейронной сети, но и связи между нейронами. **Основные результаты.** Экспериментальное исследование проводилось на моделях VGG16 и ResNet18 с использованием набора данных CIFAR-10. Показано, что структурная дистилляция позволяет оптимизировать размер нейронных сетей, сохраняя их обобщающую способность, а генетический алгоритм используется для эффективного поиска оптимальных вариантов дистилляции нейронных сетей, учитывая их структурную сложность и производительность. **Обсуждение.** Полученные результаты продемонстрировали эффективность предложенного метода при уменьшении размеров и улучшении производительности сетей с допустимой потерей качества.

Ключевые слова

искусственный интеллект, нейронные сети, структурная дистилляция, генетический алгоритм

Ссылка для цитирования: 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

A method for optimizing neural networks based on structural distillation using a genetic algorithm

Vladimir N. Kuzmin[1], Artem B. Menisov[2]✉, Timur R. Sabirov[3]

[1,2,3] Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation

[1] vka@mil.ru, https://orcid.org/0000-0002-6411-4336
[2] vka@mil.ru✉, https://orcid.org/0000-0002-9955-2694
[3] vka@mil.ru, https://orcid.org/0000-0002-6807-2954

Abstract

As neural networks become more complex, the number of parameters and required computations increases, which complicates the installation and operation of artificial intelligence systems on edge devices. Structural distillation can significantly reduce the resource intensity of using any neural networks. The paper presents a method for optimizing neural networks that combines the advantages of structural distillation and a genetic algorithm. Unlike evolutionary approaches used to search for the optimal architecture or distillation of neural networks, when forming distillation

options, it is proposed to encode not only the parameters of the neural network, but also the connections between neurons. The experimental study was conducted on the VGG16 and ResNet18 models using the CIFAR-10 dataset. It is shown that structural distillation allows optimizing the size of neural networks while maintaining their generalizing ability, and the genetic algorithm is used to effectively search for optimal distillation options for neural networks, taking into account their structural complexity and performance. The obtained results demonstrated the effectiveness of the proposed method in reducing the size and improving the performance of networks with an acceptable loss of quality.

Keywords
artificial intelligence, neural networks, structural distillation, genetic algorithm

For citation: 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

Введение

Современные искусственные нейронные сети (ИНС) могут включать сотни слоев и позволяют добиваться высоких результатов в решении различных задач, например, таких как распознавание и классификация объектов на изображениях [1]. Однако по мере увеличения сложности нейронных сетей (число слоев) объем вычислений резко возрастает [2]. Это выдвигает более высокие требования к информационной инфраструктуре эксплуатации ИНС. Соответственно, для некоторых прикладных задач сложно внедрить ИНС, когда вычислительная мощность и энергопотребление ограничены [3]. Это особенно актуально при увеличении сложности ИНС. Например, широко используемая сверточная нейронная сеть ResNet50 [4] занимает более 95 МБ памяти для хранения, содержит более 23 млн параметров и требует 4 GFLOP для вычислений. Модель GPT-3 содержит 175 млрд параметров [5], а GPT-4 — 1,76 трлн [6].

Разработка нейронных сетей является одновременно ключевым и сложным процессом в создании систем искусственного интеллекта, связанным с поиском оптимальной архитектуры, которая соответствует прикладной задаче, данным и, как следствие, дает наилучший результат. Поиск архитектуры нейронной сети (Neural Architecture Search, NAS) заключается в определении наилучшей топологии для решаемой задачи.

Основной причиной оптимизации готовых архитектур нейронных сетей является снижение ее сложности. Исследования в области NAS предлагают искать структуры нейронных сетей в порядке возрастания их уровня сложности и базируются на обучении с подкреплением или эвристических подходах. Исходя из ограничений этих подходов, заключающихся в полном поиске для каждой спецификации развертывания оборудования или цели, применение эволюционных алгоритмов стоит обобщить следующей последовательностью: инициализация из начальных вариантов архитектур (вручную); применение операторов эволюционного алгоритма для создания новых архитектур; эксплуатация, при которой используются знания, хранящиеся в истории всех оцененных архитектур.

Такая последовательность NAS применяется для построения модели машинного обучения с нуля (с примитивами низкого уровня для комбинирования признаков и обучения нейронов), которая способна получить относительно лучшую производительность для решаемой задачи.

Однако одним из самых популярных способов снижения сложности нейронной сети является дистилляция (обрезка). Цель дистилляции — удалить часть параметров нейронной сети, чтобы они не участвовали в процессе обучения и/или вывода. В работах [7–14] изучена дистилляция нейронных сетей, которая включает основные типы: структурированная и неструктурированная дистилляции.

Методы неструктурированной дистилляции удаляют веса без соблюдения какого-либо порядка. Для методов структурированной — формируются критерии и ограничения, которые определяют, как будет выполнена дистилляция. Среди всех методов выделяются критерии обрезки нейронов на основе значений их весов. Логика, лежащая в основе этих методов обрезки, проста: определенное количество или долю от всех весов нейронов, которые вносят меньший вклад в обученную модель, удаляется из архитектуры. Это ускоряет выполнение вывода и наделяет ее лучшими возможностями обобщения. Однако многократные этапы обрезки продемонстрировали [9], что она приводит к снижению производительности модели. Кроме того, можно так сократить всю модель, изменяя всю архитектуру и сделать предварительно обученные параметры непригодными для использования [8].

В работе [7] обнаружено, что в ИНС может существовать избыточность элементов, что обусловливает возможность и целесообразность оптимизации архитектуры ИНС — снижение ее сложности. Кроме того, избыточность элементов (нейронов) ИНС может снижать их надежность, повышая риск состязательных атак [8]. При структурной дистилляции возможно не только развернуть ИНС на устройстве с ограниченными ресурсами, но и сохранить при этом ее производительность [9]. Такой процесс состоит из этапов оценивании важности параметров ИНС и их селективном удалении, чтобы они не участвовали в обучении и/или выводе — по сути, их обнулении (рис. 1).

Параллельно, с удалением нейронов ИНС, может значительно улучшиться состязательная устойчивость. Это особенно актуально, так как в условиях обеспечения безопасности ИНС и ограниченности ресурсов одновременно необходимы надежность и компактность нейронных сетей [10].

Все работы в направлении дистилляции ИНС показали, что большие сети можно свести к гораздо меньшим подсетям, сохраняя при этом их качество функционирования [11]. Отметим, что подходы к структурной дистилляции ИНС охватывают ряд аспектов, включая

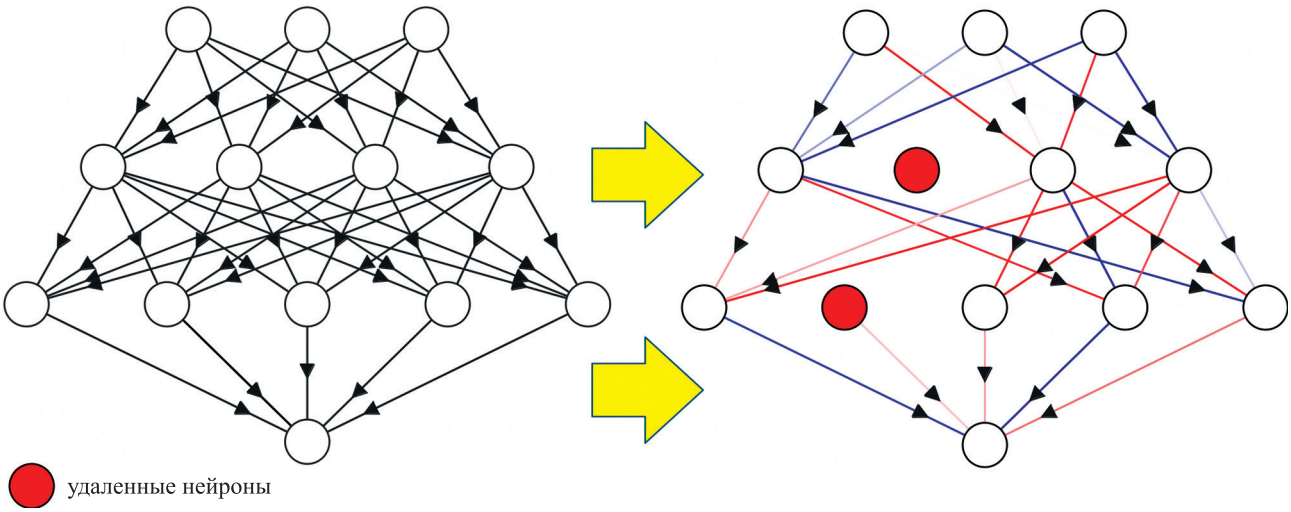

Рис. 1. Пример структурной дистилляции искусственной нейронной сети (насыщенность стрелок определяет значение весов сети, красный цвет — положительная связь, синий — отрицательная)

Fig. 1. Example of structural distillation of artificial neural network (the arrows saturation determines the value of the network weights, red color is a positive connection, blue is a negative one)

схемы удаления нейронов и связей [12], выбор параметров [13] и методов обучения ИНС [14]. К сожалению, существующие решения по-прежнему зависят от эмпирических правил или заранее определенных архитектурных шаблонов, что делает их недостаточно универсальными.

Одним из таких успешных подходов является удаление отдельных нейронов на основе их весов, т. е. обладающих весом ниже определенного порога. На практике этот порог обычно определяется путем сравнения весов внутри каждого уровня или глобально по всей сети. Однако многочисленные этапы сокращения показали, что необходим плановый метод структурной дистилляции ИНС, который предполагает, что на ранних стадиях обучения ИНС можно удалить большее число нейронов, в то время как на поздних стадиях следует систематически сокращать число обнуленных весов.

Формализация процесса структурной дистилляции ИНС

Формализованное описание структурной дистилляции ИНС может быть представлено следующим образом.

Пусть дана ИНС с архитектурой $\mathbb{N}$ и параметрами W. Цель состоит в том, чтобы найти оптимальный вектор-индикатор c, который указывает, какие нейроны следует удалить из сети, чтобы минимизировать некоторую функцию потерь L при сохранении определенного уровня производительности модели.

Формально, задача оптимизации удаления нейронов может быть описана как:

$$\min_{c} L(\mathbb{N}(W \odot c)), \qquad (1)$$

где $\mathbb{N}(W \odot c)$ — нейронная сеть с параметрами W, в которых веса, соответствующие неактивным нейронам, обнулены; L — функция потерь, которая зависит от производительности ИНС на некотором наборе данных; $\odot$ — поэлементное умножение матриц; c — бинарный вектор-индикатор, где $c_j = 1$ указывает, что j-й нейрон активен, и $c_j = 0$ — j-й нейрон удален.

Таким образом, вектор-индикатор с позволяет сформировать дистиллируемую нейронную сеть $\mathbb{N}_c = \bigcup_{i=1}^{n} \bigcup_{j=1}^{l_i} c_{ij} l_i$ с детализацией каждого слоя . Когда $c_{ij} = 1$ для любых i и j, то $\mathbb{N}_c$ на самом деле является исходной сетью $\mathbb{N}$.

Таким образом, цель дистиляции состоит в том, чтобы найти такой вектор-индикатор c, который обеспечивает наилучшую производительность ИНС и минимизирует вычислительные затраты обрезанной сети $\mathbb{N}_c$. Используя число операций с плавающей точкой в секунду (FLOPS), которое является общей метрикой, для измерения вычислительных затрат, задача дистилляции будет определена в виде:

$$\operatorname*{argmax}_{c \in \{0,1\}} (F(\mathbb{N}_c), FLOPS(\mathbb{N}_c)). \qquad (2)$$

Описание этапов метода структурной дистилляции ИНС

Реализованная методическая схема структурной дистилляции ИНС включает удаление нейронов на основе оценки их чувствительности и вклада в качество функционирования ИНС. Основное правило состоит в том, чтобы исключить нейроны с наименьшей чувствительностью, сводя к минимуму их влияние на производительность ИНС. Этот процесс сокращения элементов приводит к оптимизации сети, снижая сложность, вычислительные требования и время обработки. Однако при обрезке ИНС необходимо проявлять осторожность, поскольку чрезмерное удаление нейронов может отрицательно повлиять на производительность сети. Таким образом, поддержание баланса между удалением и производительностью ИНС имеет решающее значение

в процессе оптимизации, чтобы предотвратить потерю критической информации.

Использование генетических алгоритмов (ГА) [15] для структурной дистилляции ИНС включает в себя три основных особенности: биоинспиированный подход, позволяющий закодировать архитектуру нейронной сети в вектор хромосомы; с помощью ГА можно быстро найти субоптимальную стратегию удаления нейронов; для ГА не требуется информации о гиперпаратрах обучения ИНС.

Таким образом, ГА[1] при поиске оптимальной стратегии структурной дистилляции ИНС одновременно ищет набор оптимальных весов нейронов в популяции в каждом поколении, что приводит к уменьшению размеров ИНС.

Структурная дистилляция ИНС с использованием ГА может быть достигнута следующим образом.

Этап 1. Инициализация начальной популяции. Каждый индивид для ГА представляет собой некоторую комбинацию хромосом (нейронов и связей).

Опишем варианты кодировки структуры ИНС для нейронов и связей (рис. 2).

— Нейроны: каждый ген хромосомы представляет совокупность активных нейронов. Значение 1 в позиции j означает, что нейрон j активен, а значение 0 — неактивен. Неактивный нейрон подразумевает, что все входные соединения удаляются как во время обучения, так и во время вывода. Длина хромосомы в данном случае равна числу нейронов слоя ИНС.
— Связи: каждый ген представляет связь между слоями. Интерпретация двоичных значений следующая: если ген равен 1, связь между соответствующими слоями существует. Длина хромосомы — множество связей между нейронами соседних слоев ИНС.

Этап 2. Оценка приспособленности. Каждый вариант удаления нейронов (хромосомы) оценивается на основе производительности на некотором тестовом наборе данных. Производительность может быть измерена, например, точностью классификации, среднеквадратичной ошибкой или значением функции потерь.

Этап 3. Селекция. Хромосомы с наилучшими значениями фитнес-функции выбираются для дальнейшей обработки с использованием операторов скрещивания и мутации. Для данного исследования будем считать фитнес-функцию определенной в выражении (1).

Этап 4. Скрещивание. Выполняется скрещивание выбранных вариантов ИНС, чтобы создать новое поколение. В результате скрещивания создаются новые комбинации генов, которые сочетают признаки родителей $A = \{a_i | i = 1, \ldots, n_{last\_layer}\}$ и $B = \{b_i | i = 1, \ldots, n_{last\_layer}\}$, и величиной p_c в диапазоне [0, 1], которая определяет точку разрыва кодировки связей ИНС.

[1] Основные термины, используемые в ГА и адаптированные для структурной дистилляции ИНС: ген — блок сети, выполняющий некоторую функцию над входными данными (когда геном является нейроном, функция представляет собой некоторую функцию активации, например, ReLU); хромосома — представляет собой всю ИНС с нейронами и их связями; особь — виртуальный организм с хромосомами; популяция — совокупность особей.

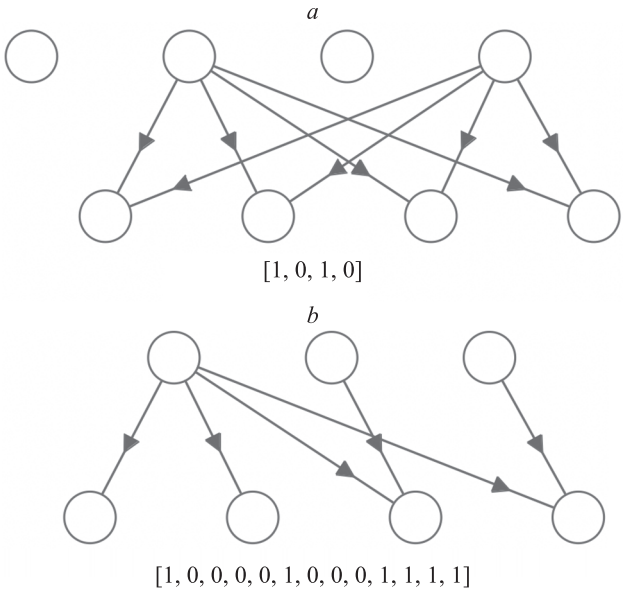

Рис. 2. Варианты кодировки структуры искусственной нейронной сети для нейрона (*a*) и связей (*b*)

Fig. 2. Options for encoding the artificial neural network structure: for a neuron (*a*); for connections (*b*)

Этап 5. Мутация. С вероятностью p_{mut} каждый хромосом изменяется случайным образом. Оператор мутации предотвращает остановку ГА в локальных оптимумах.

Этап 6. Оценивание приспособленности новых вариантов обрезки ИНС. Новое поколение оценивается на том же тестовом наборе данных, что и предыдущее, чтобы определить их приспособленность.

Этап 7. Повторение процесса. Этапы селекции, скрещивания, мутации и оценки приспособленности повторяются до тех пор, пока не будет достигнуто условие остановки (например, определенное число поколений или сходимость к определенному уровню функции потерь).

Этап 8. Отбор лучших вариантов. После завершения процесса отбираются индивиды с наилучшей приспособленностью в конечную популяцию. Полученная оптимальная комбинация нейронов и связей применяется для дистилляции исходной нейронной сети.

Эксперимент

Оценивание разработанного метода обрезки выполнено на нейронных сетях VGG16 [16] и ResNet18 [17], характеристики которых представлены в табл. 1, и валидационной части набора данных CIFAR-10 [18].

CIFAR-10 — набор данных, состоящий из 60 000 изображений, 50 000 для обучения и 10 000 для тестирования. Он разделен на 10 классов, и каждый класс содержит 5000 обучающих и 1000 тестовых изображений. Каждый объект CIFAR-10 представляет собой RGB-изображение размером 32 × 32 пиксела. Этот набор данных широко используется для классификации изображений и является одним из самых известных эталонных наборов данных в области компьютерного зрения.

Таблица 1. Исходные нейронные сети для эксперимента
Table 1. Initial neural networks for the experiment

Наименование предобученной ИНС	Число параметров	Число слоев	Объем памяти, МБ
VGG16	14 849 345	16	56,13
ResNet18	11 689 512	72	57,42

Настройки гиперпараметров ГА следующие: размер популяции $N = 30$, число выбранных геномов $K = 5$, параметр скрещивания $p_c = 0{,}6$, параметр мутаций $p_{mut} = 0{,}6$, число раундов $T = 5$.

Рассмотрим результаты структурной дистилляции ИНС (табл. 2) и изменение параметров ГА, которые влияют на результаты вычислений (табл. 3).

Обсуждение результатов эксперимента

Из результатов экспериментов следует, что ГА позволил значительно уменьшить число параметров в обеих ИНС при сохранении приемлемого уровня точности.

Несмотря на структурную дистилляцию, качество на тестовом наборе данных осталось высоким. Это

Таблица 2. Результаты эксперимента
Table 2. Experiment results

Раунд ГА	Число параметров, млн	Объем памяти, МБ	Точность (для топ-1)	Доля уменьшения параметров ИНС, %	FLOPS $\cdot 10^6$
VGG16					
1	13,2	53,438	0,8571	7	60,1
2	3,1	11,882	0,8597	21	48,5
3	0,9	3,578	0,8551	29	40,8
4	0,5	1,787	0,8473	55	29,5
5	0,3	1,243	0,8374	60	11,2
ResNet18					
1	11,2	42,707	0,7223	5	1,800
2	3,0	11,529	0,7213	26	1,080
3	0,9	3,591	0,7223	30	0,736
4	0,3	1,218	0,7234	33	0,590
5	0,1	0,460	0,6558	33	0,502

Таблица 3. Результаты вычислений генетического алгоритма
Table 3. Genetic algorithm calculation results

Раунд ГА	Время работы, с	Результат (значение целевой функции)[1]
VGG16		
1	543,81	0,8571; 60,1
2	491,52	0,8597; 48,5
3	488,79	0,8551; 40,8
4	246,66	0,8473; 29,5
5	219,34	0,8374; 11,2
ResNet18		
1	367,83	0,7223; 1,800
2	332,46	0,7213; 1,080
3	330,61	0,7223; 0,736
4	166,83	0,7234; 0,590
5	148,36	0,6558; 0,502

[1] Выражение (2).

Таблица 4. Рейтинг и характеристики нейронных сетей, обученных на наборе данных CIFAR-10
Table 4. Ranking and characteristics of neural networks trained on the CIFAR-10 dataset

Название ИНС	Точность (для топ-1)	Время обучения, с
UL-Hopfield [19]	0,8310	998,27
CvP [20]	0,8319	692,22
CCN [20]	0,8336	743,94
ThresholdNet [21]	0,8528	1014,47

Таблица 5. Результаты удаления нейронов при ГА для ИНС VGG16 и ResNet18 (число параметров ИНС, ед.)
Table 5. Results of neuron removal during GA for ANN VGG16 and ResNet18 (number of ANN parameters, units)

Сверточные слои ИНС	Раунды дистилляции				
	1	2	3	4	5
VGG16					
Conv 1	64	39	24	15	9
Conv 2	64	52	42	34	28
Conv 3	128	116	105	95	86
Conv 4	128	122	116	111	106
Conv 5	256	231	208	188	170
Conv 6	256	128	64	32	16
Conv 7	256	128	64	32	16
Conv 8	512	205	82	33	14
Conv 9	512	205	82	33	14
Conv 10	512	205	82	33	14
Conv 11	512	205	82	33	14
Conv 12	512	205	82	33	14
Conv 13	512	205	82	33	14
ResNet18					
Conv 1	64	53	44	36	30
Conv 2	64	58	53	48	44
Conv 3	64	53	44	36	30
Conv 4	64	58	53	48	44
Conv 5	64	53	44	36	30
Conv 6	128	103	83	67	54
Conv 7	128	83	54	36	24
Conv 8	128	83	54	36	24
Conv 9	128	103	83	67	54
Conv 10	128	83	54	36	24
Conv 11	256	205	164	132	106
Conv 12	256	164	106	68	44
Conv 13	256	164	106	68	44
Conv 14	256	205	164	132	106
Conv 15	256	164	106	68	44
Conv 16	512	359	252	177	124
Conv 17	512	252	124	61	31
Conv 18	512	252	124	61	31

свидетельствует о том, что ГА правильно определял нейроны, чье удаление не сильно влияло на способность ИНС к определению класса изображения.

Выполним сравнение временных затрат на генетическую оптимизацию и обучение новых сетей с меньшим числом параметров и слоев, но с таким же качеством. В табл. 4 представлены рейтинг и характеристики нейронных сетей, обученных на наборе данных CIFAR-10[1].

Из времени обучения ИНС видно, что более оперативно провести дистилляцию нейронной сети, с последующим снижением сложности и необходимых ресурсов, чем обучать новые нейронные сети.

Таким образом, ГА предпочитает удалять нейроны, которые имеют меньшее влияние на вывод ИНС или которые могут быть скомпенсированы другими нейронами в сети. Это позволяет сохранить основные паттерны и зависимости, зафиксированные нейронной сетью в процессе обучения. Перед удалением нейронов ГА может оценивать влияние этого действия на производительность модели с помощью валидационного набора данных. Удаление является неравномерным и касается только сверточных слоев ИНС. Для ИНС VGG16 и ResNet18 изменения за 5 раундов ГА представлены в табл. 5.

Отметим, что предположение о возможности недопустимого снижения качества ИНС не оправдано. Удаление нейронов может даже способствовать более обобщенному обучению ИНС, поскольку это может предотвратить переобучение за счет сокращения ее параметров и упрощения аппроксимирующей функции.

Таким образом, высокое качество предсказаний на тестовом наборе данных после структурной дистилляции ИНС ГА свидетельствует о том, что удаление нейронов было выполнено эффективно. ИНС остается способной к обобщению новых данных и сохранению высокой точности предсказаний, что делает метод пригодным для уменьшения размера ИНС при минимальной потере качества.

Такие результаты подчеркивают значимость ГА для структурной дистилляции ИНС, особенно когда важно сохранить высокую точность предсказаний при одновременном сокращении вычислительной нагрузки.

Для более крупных ИНС, таких как большие языковые и мультимодальные ИНС, применение ГА для структурной дистилляции может иметь еще более значимые результаты, заключающиеся в следующем.

— Экономия ресурсов. Большие языковые и мультимодальные ИНС обычно требуют огромного количества вычислительных ресурсов для обучения и инференса. Уменьшение размера позволяет снизить требования к ресурсам, что делает их более доступными для широкого круга пользователей и применений.
— Ускорение инференса. Уменьшение размера ИНС также приводит к увеличению скорости инференса, что особенно важно в приложениях реального времени.

Сохранение высокой точности предсказаний. Для ИНС, обученных на огромных корпусах текста, сохранение точности предсказаний является критическим, поскольку они должны обладать высоким пониманием языка для успешного выполнения различных задач, таких как генерация текста или ответ на вопросы. В случае с мультимодальными ИНС, которые обрабатывают как текст, так и изображения или другие типы данных, точность предсказаний также является важным аспектом.

Заключение

В настоящей работе представлен новый метод оптимизации нейронных сетей, основанный на структурной дистилляции с использованием генетического алгоритма. Эксперименты на моделях VGG16 и ResNet18 показали, что предложенный метод способен существенно оптимизировать размер архитектуры сетей, сохраняя при этом их производительность. Полученные результаты подтверждают эффективность метода в оптимизации нейронных сетей и его потенциал в применении к другим моделям и задачам машинного обучения. В будущем планируется расширить область применения метода на более широкий спектр архитектур и наборов данных, провести более глубокий анализ его характеристик и преимуществ, а также экспериментально доказать закономерности сокращения сложности отдельных слоев при адаптации сокращенных нейронных сетей для других задач.

[1]Рейтинг представлен на основе данных ресурса paperswithcode [Электронный ресурс]. Режим доступа: https://paperswithcode.com/sota/image-classification-on-cifar-10 (дата обращения: 13.05.2024).

Литература

1. Spoorthi M., Indu Priya B., Kuppala M., Karpe V.S., Dharavath D. Automated resume classification system using ensemble learning // Proc. of the 9th International Conference on Advanced Computing and Communication Systems (ICACCS). V. 1. 2023. P. 1782–1785. https://doi.org/10.1109/icaccs57279.2023.10112917
2. Freire P.J., Osadchuk Y., Spinnler B., Napoli A., Schairer W., Costa N., Prilepsky J.E., Turitsyn S.K. Performance versus complexity study of neural network equalizers in coherent optical systems // Journal of Lightwave Technology. 2021. V. 39. N 19. P. 6085–6096. https://doi.org/10.1109/jlt.2021.3096286
3. Hankala T., Hannula M., Kontinen J., Virtema J. Complexity of neural network training and ETR: Extensions with effectively continuous functions // Proceedings of the AAAI Conference on Artificial

References

1. Spoorthi M., Indu Priya B., Kuppala M., Karpe V.S., Dharavath D. Automated resume classification system using ensemble learning. *Proc. of the 9th International Conference on Advanced Computing and Communication Systems (ICACCS). V. 1*, 2023, pp. 1782–1785. https://doi.org/10.1109/icaccs57279.2023.10112917
2. Freire P.J., Osadchuk Y., Spinnler B., Napoli A., Schairer W., Costa N., Prilepsky J.E., Turitsyn S.K. Performance versus complexity study of neural network equalizers in coherent optical systems. *Journal of Lightwave Technology*, 2021, vol. 39, no. 19, pp. 6085–6096. https://doi.org/10.1109/jlt.2021.3096286
3. Hankala T., Hannula M., Kontinen J., Virtema J. Complexity of neural network training and ETR: Extensions with effectively continuous functions. *Proceedings of the AAAI Conference on Artificial*

Intelligence. 2024. V. 38. N 11. P. 12278–12285. https://doi.org/10.1609/aaai.v38i11.29118
4. Koonce B., Koonce B. ResNet 50 // Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. Springer, 2021. P. 63–72. https://doi.org/10.1007/978-1-4842-6168-2_6
5. Floridi L., Chiriatti M. GPT-3: Its nature, scope, limits, and consequences // Minds and Machines. 2020. V. 30. N 4. P. 681–694. https://doi.org/10.1007/s11023-020-09548-1
6. Achiam J., Adler S., Agarwal S. et al. Gpt-4 technical report // arXiv. 2023. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774
7. Bodimani M. Assessing the impact of transparent AI systems in enhancing user trust and privacy // Journal of Science & Technology. 2024. V. 5. N 1. P. 50–67. https://doi.org/10.55662/JST.2024.5102
8. Lu Z., Li Z., Chiang C.-W., Yin M. Strategic adversarial attacks in AI-assisted decision making to reduce human trust and reliance // Proc. of the Thirty-Second International Joint Conference on Artificial Intelligence. 2023. P. 3020–3028. https://doi.org/10.24963/ijcai.2023/337
9. 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. https://doi.org/10.1109/tpami.2023.3334614
10. Ding S., Zhang L., Pan M., Yuan X. PATROL: Privacy-oriented pruning for collaborative inference against model inversion attacks // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024. P. 4704–4713. https://doi.org/10.1109/wacv57701.2024.00465
11. Fang G., Ma X., Song M., Mi M.B., Wang X. Depgraph: Towards any structural pruning // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 16091–16101. https://doi.org/10.1109/cvpr52729.2023.01544
12. Wen L., Zhang X., Bai H., Xu Z. Structured pruning of recurrent neural networks through neuron selection // Neural Networks. 2020. V. 123. P. 134–141. https://doi.org/10.1016/j.neunet.2019.11.018
13. Zhao M., Peng J., Yu S., Liu L., Wu N. Exploring structural sparsity in CNN via selective penalty // IEEE Transactions on Circuits and Systems for Video Technology. 2022. V. 32. N 3. P. 1658–1666. https://doi.org/10.1109/tcsvt.2021.3071532
14. Shen M., Molchanov P., Yin H., Alvarez J.M. When to prune? a policy towards early structural pruning // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 12237–12246. https://doi.org/10.1109/cvpr52688.2022.01193
15. Katoch S., Chauhan S.S., Kumar V. A review on genetic algorithm: past, present, and future // Multimedia Tools and Applications. 2021. V. 80. P. 8091–8126. https://doi.org/10.1007/s11042-020-10139-6
16. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // arXiv. 2014. arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
17. Zhou Y., Ren F., Nishide S., Kang X. Facial sentiment classification based on resnet-18 model // Proc. of the 2019 International Conference on Electronic Engineering and Informatics (EEI). 2019. P. 463–466. https://doi.org/10.1109/eei48997.2019.00106
18. Recht B., Roelofs R., Schmidt L., Shankar V. Do CIFAR-10 classifiers generalize to CIFAR-10? // arXiv. 2018. arXiv:1806.00451. https://doi.org/10.48550/arXiv.1806.00451
19. Liu Q., Mukhopadhyay S. Unsupervised learning using pretrained CNN and associative memory bank // Proc. of the International Joint Conference on Neural Networks (IJCNN). 2018. P. 01–08. https://doi.org/10.1109/ijcnn.2018.8489408
20. Jeevan P., Sethi A. Vision Xformers: Efficient attention for image classification // arXiv. 2021. arXiv:2107.02239. https://doi.org/10.48550/arXiv.2107.02239
21. Hou Y., Wu Z., Cai X., Zhu T. The application of improved densenet algorithm in accurate image recognition // Scientific Reports. 2024. V. 14. N 1. P. 8645. https://doi.org/10.1038/s41598-024-58421-z

Intelligence, 2024, vol. 38, no. 11, pp. 12278–12285. https://doi.org/10.1609/aaai.v38i11.29118
4. Koonce B., Koonce B. *ResNet 50. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*. Springer, 2021, pp. 63–72. https://doi.org/10.1007/978-1-4842-6168-2_6
5. Floridi L., Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 2020, vol. 30, no. 4, pp. 681–694. https://doi.org/10.1007/s11023-020-09548-1
6. Achiam J., Adler S., Agarwal S. et al. Gpt-4 technical report. *arXiv*, 2023, arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774
7. Bodimani M. Assessing the impact of transparent AI systems in enhancing user trust and privacy. *Journal of Science & Technology*, 2024, vol. 5, no. 1, pp. 50–67. https://doi.org/10.55662/JST.2024.5102
8. Lu Z., Li Z., Chiang C.-W., Yin M. Strategic adversarial attacks in AI-assisted decision making to reduce human trust and reliance. *Proc. of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 3020–3028. https://doi.org/10.24963/ijcai.2023/337
9. 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. https://doi.org/10.1109/tpami.2023.3334614
10. Ding S., Zhang L., Pan M., Yuan X. PATROL: Privacy-oriented pruning for collaborative inference against model inversion attacks. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4704–4713. https://doi.org/10.1109/wacv57701.2024.00465
11. Fang G., Ma X., Song M., Mi M.B., Wang X. Depgraph: Towards any structural pruning. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16091–16101. https://doi.org/10.1109/cvpr52729.2023.01544
12. Wen L., Zhang X., Bai H., Xu Z. Structured pruning of recurrent neural networks through neuron selection. *Neural Networks*, 2020, vol. 123, pp. 134–141. https://doi.org/10.1016/j.neunet.2019.11.018
13. Zhao M., Peng J., Yu S., Liu L., Wu N. Exploring structural sparsity in CNN via selective penalty. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, vol. 32, no. 3, pp. 1658–1666. https://doi.org/10.1109/tcsvt.2021.3071532
14. Shen M., Molchanov P., Yin H., Alvarez J.M. When to prune? a policy towards early structural pruning. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12237–12246. https://doi.org/10.1109/cvpr52688.2022.01193
15. Katoch S., Chauhan S.S., Kumar V. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 2021, vol. 80, pp. 8091–8126. https://doi.org/10.1007/s11042-020-10139-6
16. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
17. Zhou Y., Ren F., Nishide S., Kang X. Facial sentiment classification based on resnet-18 model. *Proc. of the 2019 International Conference on Electronic Engineering and Informatics (EEI)*, 2019, pp. 463–466. https://doi.org/10.1109/eei48997.2019.00106
18. Recht B., Roelofs R., Schmidt L., Shankar V. Do CIFAR-10 classifiers generalize to CIFAR-10?. *arXiv*, 2018, arXiv:1806.00451. https://doi.org/10.48550/arXiv.1806.00451
19. Liu Q., Mukhopadhyay S. Unsupervised learning using pretrained CNN and associative memory bank. *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 01–08. https://doi.org/10.1109/ijcnn.2018.8489408
20. Jeevan P., Sethi A. Vision Xformers: Efficient attention for image classification. *arXiv*, 2021, arXiv:2107.02239. https://doi.org/10.48550/arXiv.2107.02239
21. Hou Y., Wu Z., Cai X., Zhu T. The application of improved densenet algorithm in accurate image recognition. *Scientific Reports*, 2024, vol. 14, no. 1, pp. 8645. https://doi.org/10.1038/s41598-024-58421-z

Авторы

Кузьмин Владимир Никифорович — доктор военных наук, профессор, ведущий научный сотрудник, Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57220813706, https://orcid.org/0000-0002-6411-4336, vka@mil.ru

Authors

Vladimir N. Kuzmin — D.Sc. (Military Science), Professor, Leading Researcher, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57220813706, https://orcid.org/0000-0002-6411-4336, vka@mil.ru

Менисов Артем Бакытжанович — кандидат технических наук, докторант, Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация, sc 57220815185, https://orcid.org/0000-0002-9955-2694, vka@mil.ru
Сабиров Тимур Римович — кандидат технических наук, старший преподаватель, Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург, 197198, Российская Федерация, sc 55496404500, https://orcid.org/0000-0002-6807-2954, vka@mil.ru

Artem B. Menisov — PhD, Doctoral Student, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, sc 57220815185, https://orcid.org/0000-0002-9955-2694, vka@mil.ru

Timur R. Sabirov — PhD, Senior Lecturer, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, sc 55496404500, https://orcid.org/0000-0002-6807-2954, vka@mil.ru

Статья поступила в редакцию 25.04.2024
Одобрена после рецензирования 17.07.2024
Принята к печати 16.09.2024

Received 25.04.2024
Approved after reviewing 17.07.2024
Accepted 16.09.2024