
ТЕОРИЯ И МЕТОДЫ
ОБРАБОТКИ СИГНАЛОВ

УДК 004.85

ОПТИМИЗАЦИЯ ОБУЧЕНИЯ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ПРИ РАСПОЗНАВАНИИ ИЗОБРАЖЕНИЙ С НИЗКОЙ ПЛОТНОСТЬЮ ТОЧЕК

© 2021 г. В. В. Зиядинов^а, П. С. Курочкин^а, М. В. Терешонок^{а, б, *}

^аМосковский технический университет связи и информатики,
ул. Авиамоторная, 8а, Москва, 111024 Российская Федерация

^бМосковский физико-технический институт (национальный исследовательский университет),
Институтский пер., 9, Долгопрудный, Московская обл., 141701 Российская Федерация

*E-mail: m.v.tereshonok@mtuci.ru

Поступила в редакцию 24.06.2021 г.

После доработки 23.07.2021 г.

Принята к публикации 25.07.2021 г.

Предложены методы оптимизации структуры сверточной нейронной сети, применяемой для распознавания изображений с низкой плотностью точек, с целью ускорения обучения и распознавания новых изображений, а также уменьшения ресурсоемкости процедур обучения и распознавания. Оптимизированная нейронная сеть показала значительное повышение скорости работы без потери точности распознавания изображений с низкой плотностью точек, а также существенно сниженную тенденцию к переобучению.

DOI: 10.31857/S0033849421120202

ВВЕДЕНИЕ

В настоящее время сверточные нейронные сети все чаще используют для решения задач распознавания сигналов и изображений различной физической природы [1–3], что в дальнейшем может быть использовано для решения множества прикладных задач (например, проблем с автомобильными пробками, в медицине). Современная цифровая обработка сигналов уже неотделима от применения искусственных нейронных сетей [4]. Как правило, современные структуры сверточных нейронных сетей из-за своей высокой сложности подразумевают высокую вычислительную сложность при ее обучении и распознавании новых объектов.

Реализация подобных сетей на традиционной аппаратной платформе с использованием графических процессоров для научных вычислений общего назначения (например, Nvidia Tesla и т.п.) отличается весьма значительным энергопотреблением [5, 6]. Наиболее перспективной с точки зрения энергопотребления представляется реализация сверточных нейронных сетей большой размерности на сверхпроводниковой аппаратной платформе [7, 8]. Сверхпроводниковая реализация также позволит повысить тактовые частоты функционирования нейровычислителей без повышения тепловыделения, что приведет к повышению общего быстродействия [9, 10].

Цель данной статьи – проанализировать способы оптимизации сверточной нейронной сети с целью увеличения продуктивности вычислительных процессов. В качестве примера рассмотрена прикладная задача распознавания типов взаимного расположения абонентов сетей мобильной связи, сводящаяся к распознаванию изображений с низкой плотностью точек. При выборе способов оптимизации учтены особенности реализации подобной оптимизации на перспективных аппаратных платформах.

В работе [11] были предложены автоматизированные методы анализа информации, передающейся в служебных командах управления сетями мобильной связи, позволяющие с высокой точностью находить устойчивые группы абонентов с целью оценки и прогнозирования деятельности участников массовых мероприятий. В работе [12] рассмотрен широкий класс типов взаимного расположения абонентов сетей мобильной связи, создана математическая модель, которая позволяет генерировать изображения с низкой плотностью точек, описывающие эти типы взаимного расположения абонентов и формирующие репрезентативные обучающие и контрольные выборки для обучения и тестирования сверточной нейронной сети. В то же время в работе [12] отмечены недостатки предложенной нейронной сети, включающие в себя эффект переобучения и высокую ре-

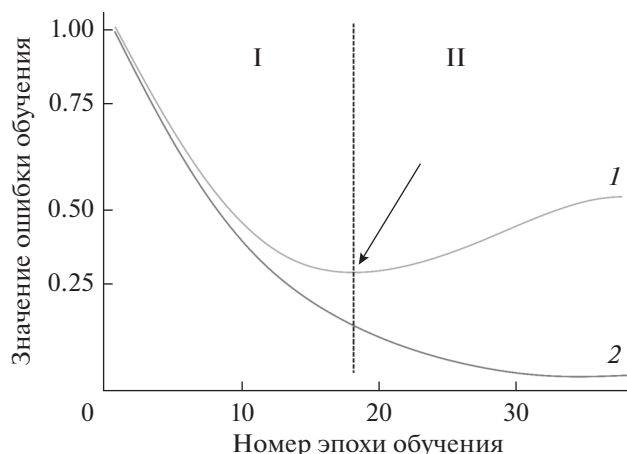


Рис. 1. Эффект переобучения искусственной нейронной сети: 1 — проверочная выборка, 2 — обучающая выборка, область I — недообучение, область II — переобучение, стрелкой показано минимальное значение ошибки проверочной выборки.

сурсоемкость процесса обучения и распознавания, обусловленную большим числом слоев и нейронов в них.

Данная статья посвящена преодолению трудностей с переобучением и ресурсоемкостью нейронной сети, предложенной в работе [12].

1. ПОСТАНОВКА ЗАДАЧИ

Анализ результатов, полученных в работе [12] с использованием классической структуры сверточной нейронной сети [13], показывает, что наибольшая вероятность правильного распознавания составила 0.972. Минимальное значение ошибки распознавания проверочной выборки было получено на шестой эпохе и составило 0.128, на последующих эпохах обучения выявлен рост ошибки распознавания проверочной выборки. В данном случае заметен рост ошибки при распознавании данных из проверочной выборки, а точность классификации обучающей выборки при этом растет, что указывает на эффект переобучения полученной сверточной нейронной сети (рис. 1). Нейронная сеть “запоминает” обучающую выборку, и дальнейшее ее обучение не даст повышения качества работы [14, 15]. Более того, подобный результат может означать избыточность конфигурации нейронной сети, что помимо переобучения ведет также к нерациональному использованию вычислительных ресурсов. Данное обстоятельство требует дальнейшего анализа путей возможной оптимизации сверточной нейронной сети с целью улучшения качества ее работы, ускорения как обучения, так и распознавания новых данных.

2. РЕГУЛЯРИЗАЦИЯ ВЕСОВЫХ КОЭФФИЦИЕНТОВ

Эффект переобучения состоит в точном запоминании сетью параметров обучающей выборки, а не в выявлении основных информативных признаков, получаемых благодаря обобщающей способности сети. Точность классификации проверочной выборки при переобучении не растет из-за отличия новых данных от использованных для обучения и фактически запомненных нейронной сетью, а в некоторых случаях падает. Для борьбы с этим эффектом существует несколько методов: оптимизация количества скрытых слоев, регуляризация весовых коэффициентов при обучении, поиск оптимальных параметров обучения сети.

Для борьбы с переобучением наиболее эффективным является метод добавления L2-регуляризации [16, 17]. Данный метод регуляризации заключается в ограничении и уменьшении очень больших весовых коэффициентов на отношение суммы квадратов весовых коэффициентов. Регуляризация L2 имеет следующую формулу:

$$R(\omega) = \|\omega\|_2^2 = \sum_{i=1}^d \omega_i^2.$$

Регуляризация позволяет выделить наиболее информативные признаки в обучающей выборке, а наименее эффективные убрать или снизить их влияние, что увеличивает качество обучения нейронной сети.

В данной работе было проведено имитационное моделирование с целью исследования влияния L2-регуляризации и ее коэффициентов на подавление эффекта переобучения. Исследовалось влияние L2-регуляризации и разными коэффициентами.

Применение регуляризации позволило значительно снизить ошибку распознавания проверочной выборки. Итоговые графики обучения приведены на рис. 2.

Анализ графиков обучения сети позволяет сделать вывод о том, что применение L2-регуляризации с коэффициентом 0.0001 достаточно эффективно подавляет рост ошибки распознавания проверочной выборки с увеличением числа эпох обучения. Это в свою очередь свидетельствует о применимости L2-регуляризации для избавления от эффекта переобучения нейронной сети при распознавании изображений с низкой плотностью точек. Использование L2-регуляризации также уменьшает количество эпох обучения, необходимых для достижения высокой точности классификации.

В то же время L2-регуляризация не производит положительного влияния на ресурсоемкость нейронной сети и ее быстродействие. Для повышения быстродействия целесообразно использо-

вать методы структурной оптимизации. Для этого был проведен анализ структуры сети, предложенной для распознавания типов взаимного расположения абонентов сетей мобильной связи, представленных изображениями с низкой плотностью точек в работе [12].

3. ОПТИМИЗАЦИЯ СТРУКТУРЫ НЕЙРОННОЙ СЕТИ

В описанной в работе [12] сверточной нейронной сети была использована структура, представленная на рис. 3.

Основные вычислительные слои разделяются на два основных вида (подвыборочные и сверточные), модели будет свойственна “бипирамидальная” характеристика [13]. Данная характеристика описывается таким образом, что на каждом сверточном и подвыборочном слое количество карт признаков увеличивается с уменьшением разрешения каждой из карт относительно предыдущего слоя. Смысл свертки, следующей после подвыборки, состоит в том, чтобы “упростить” более “сложную” карту признаков. Описанная структура слоев повторяется до получения матриц меньшей размерности.

С целью уменьшения времени, затрачиваемого на обработку изображения, были предприняты меры по оптимизации структуры нейронной сети. На одном из первых этапов было решено увеличить размер ядра свертки. В описанной в работе [12] нейронной сети все сверточные слои имели размер ядра 3×3 пикселя. Так как исходное изображение содержало лишь маркеры положения объекта, расстояние между которыми было в несколько пикселей (от пяти до нескольких десятков), размеры ядра сверток были увеличены до размера 5×5 , за исключением последнего блока. Дальнейшее увеличение размеров ядра свертки увеличивает вычислительные затраты, что уменьшает скорость работы сети без значимого повышения качества ее работы.

Изначально описанная нейронная сеть имела пять блоков, выполняющих операции свертки и понижения размерности. Все блоки содержали сверточные слои с размером ядра свертки 3×3 и количеством фильтров 64, 128, 256, 512, 1024 соответственно. Визуализация промежуточных результатов обработки изображения показала, что большинство из фильтров не содержат в себе полезной информации (матрицы полностью состояли из нулей, что на рис. 4 показано черным цветом), а фильтры, которые содержали информацию, часто дублировали друг друга (на рис. 4 во множестве фильтров записан одинаковый клиновидный образ, соответствующий одному и тому же типу изображения [12]). Поэтому было решено сократить количество блоков до четырех, количество филь-

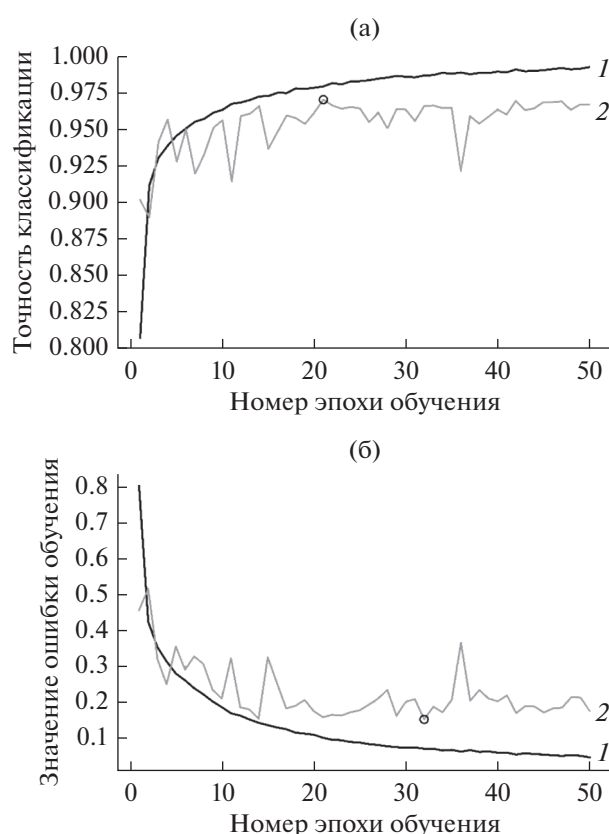


Рис. 2. Зависимости точности классификации (а) и значения ошибок сети (б) для обучающих (кривые 1) и тестовых (кривые 2) выборок для классической нейронной сети с применением L2-регуляризации от номера эпохи обучения; кружочками на кривых 2 обозначены минимальные и максимальные значения.

тров — до 16, 32, 64 и 128 соответственно (рис. 5), а также убрать некоторые наименее необходимые слои нейронной сети для ускорения ее работы. Все это позволило значительно снизить вычислительные затраты на обработку изображения без существенной потери точности классификации.

Был проведен анализ функций различных слоев сети, их влияния на общий результат работы, на быстродействие сети и реальную необходимость их использования.

В исходной сети применялись следующие функции и слои.

1. “Дополнение нулями” (*англ.* Zero Padding [18]). Этот слой служит для сохранения удобной размерности матрицы после процедуры свертки.

При применении свертки с размером фильтра больше 1×1 размерность выходного слоя будет меньше, чем входного, так как число возможных положений фильтра меньше, чем размерность входного изображения. На рис. 6 показан пример реализации свертки. По входному изображению (см. рис. 6а) фильтр (см. рис. 6б) проходит с рав-

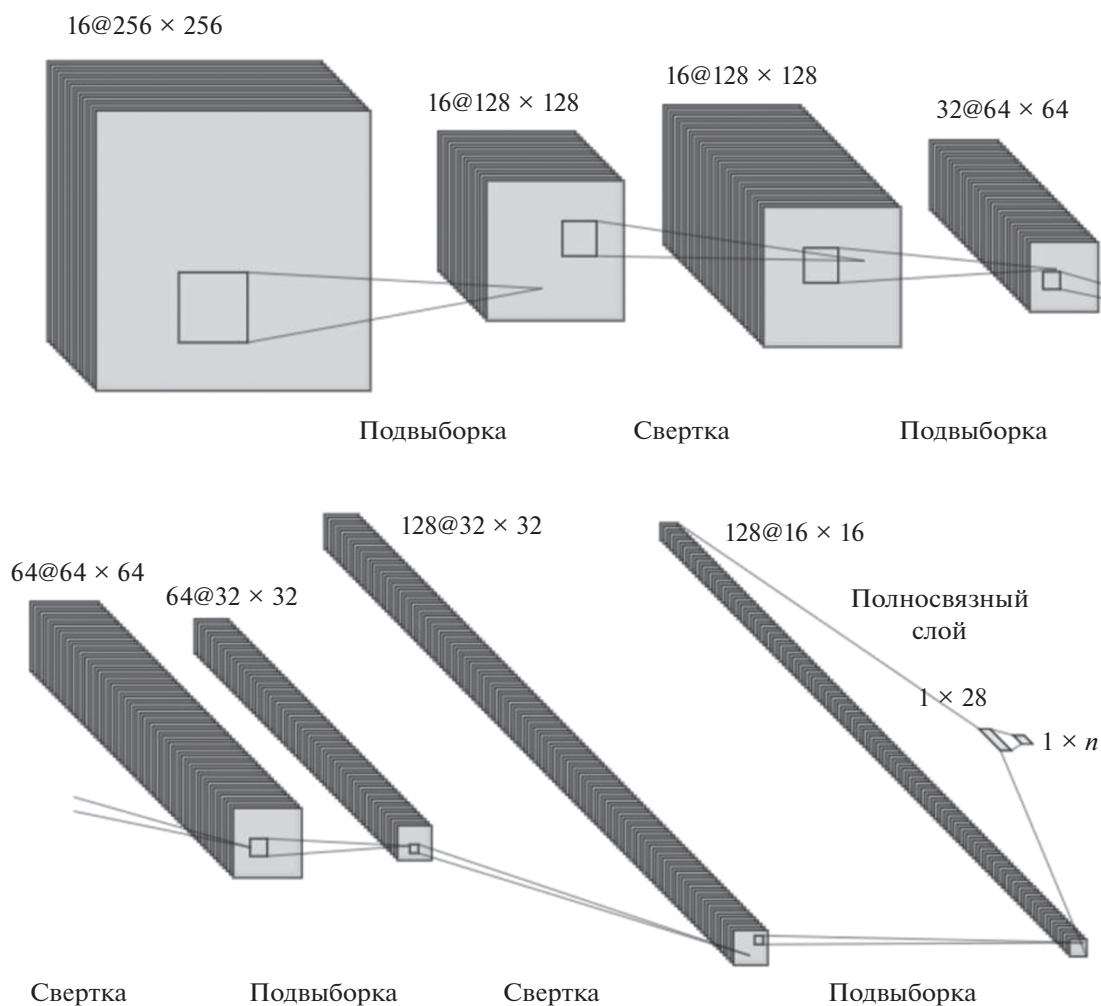


Рис. 3. Изначальная структура нейронной сети, использовавшаяся для обучения.



Рис. 4. Визуализация карт признаков слоя обученной нейронной сети классического варианта с 256 фильтрами. Полностью незначимые фильтры показаны черным цветом.



Рис. 5. Визуализация карт признаков слоя обученной нейронной сети предложенного варианта с 64 фильтрами. Количество повторяющихся образов по сравнению с исходным вариантом, приведенным на рис. 4, снижено. Полностью незначимые фильтры отсутствуют.

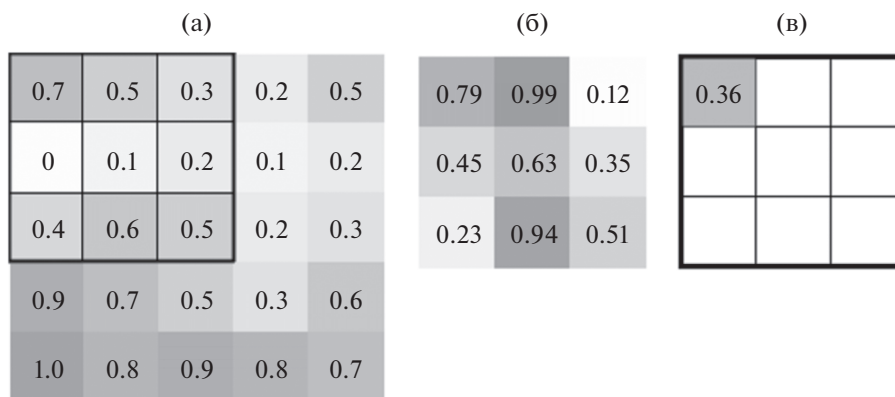


Рис. 6. Принцип свертки: а – входное изображение, б – фильтр, в – выходное изображение.

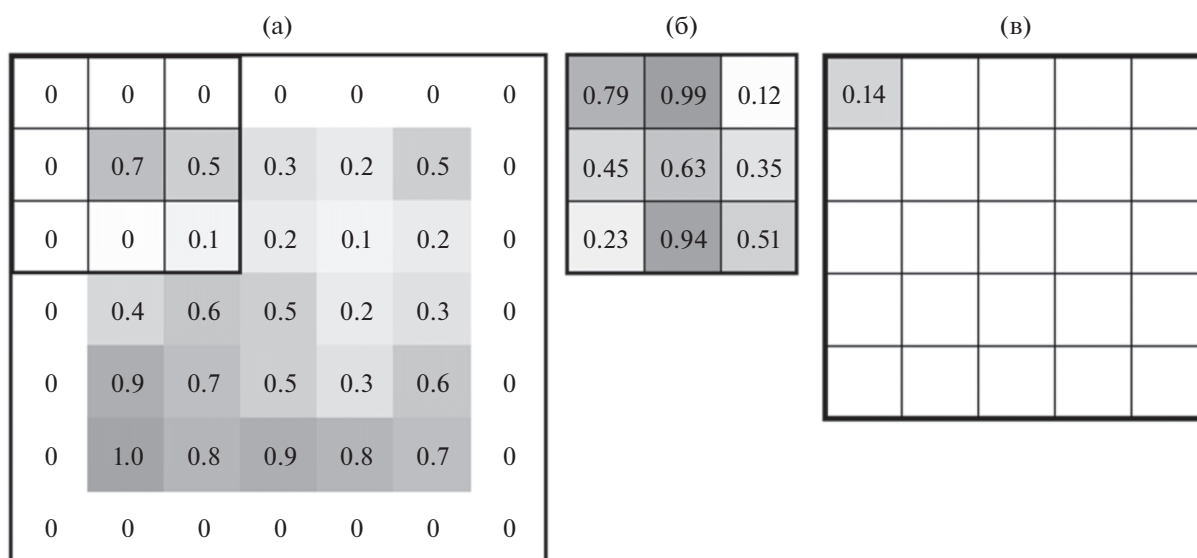


Рис. 7. Принцип работы слоя дополнения нулями (Zero Padding): а – входное изображение (см. рис. 6) с “дополнительными нулями”, б – фильтр (см. рис. 6), в – выходное изображение.

ным шагом и получаются значения следующего слоя (см. рис. 6в). При применении фильтра 3×3 размерность выходной матрицы уменьшается относительно размерности входного изображения. Чтобы избежать изменения размерности слоя, входное изображение изменяют таким образом, чтобы выходной слой был той же размерности, что и изначальное изображение. Входное изображение “расширяют”, к его краям добавляют группу пикселей, заполненных нулями. Размерность входного изображения увеличивается, соответственно, увеличивается размерность выходного слоя (рис. 7). “Дополнение нулями” производится таким образом, чтобы итоговая размерность выходного слоя была равна или кратна размерности изначального входного слоя.

Как видно из рис. 8, значимая часть изображения расположена в центре изображения, крайние пиксели имеют значение 0, т.е. черный цвет (цвет изображения инвертирован). Соответственно, в применении функции “дополнения нулями” нет строгой необходимости.

2. Пакетная нормировка. Структура глубокой нейронной сети содержит слой пакетной нормировки (Batch Normalization) после каждого сверточного слоя, которые выполняют линейную операцию пакетной нормировки [19]. Так как добавление весов смещения в сверточных слоях также является линейной операцией, а слой нормировки имеет собственный параметр смещения β , было решено не использовать их, тем самым вдвое сократив количество операций сложения.

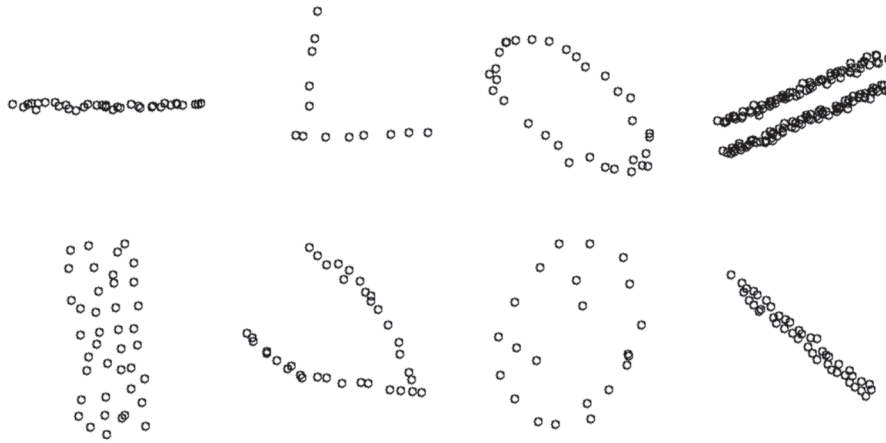


Рис. 8. Примеры изображений в обучающих выборках.

(а)				(б)	
0.79	0.82	0.22	0.35	0.82	0.35
0.41	0.52	0.21	0.12	0.93	0.92
0.54	0.11	0.92	0.54		
0.84	0.93	0.43	0.52		

Рис. 9. Принцип работы подвыборки максимальных значений (Max Pooling). Из группы пикселей входного слоя (а) выбираются наибольшие значения и передаются в выходной слой (б).

3. Применение активационной функции. В описанной сети использовалась функция ReLU – “полулинейный элемент” [20]; использование такой активационной функции обеспечивает высокую скорость обучения сети и не нуждается в дальнейших модификациях.

4. Свертка (в описанной в работе [12] сверточной нейронной сети сверточный слой сохранял размерность матрицы) и подвыборка максимальных значений (Max Pooling) – выбор максимального значения среди значений внутри фильтра, накладываемого на входной слой (рис. 9). Операция подвыборки максимальных значений уменьшает размерность слоя в два раза, так как фильтр имел размерность 2×2 , а положение фильтра выбиралось таким образом, чтобы оно не накладывалось на предыдущие положения. Примененный слой подвыборки максимальных значений удаляет часть параметров и, соответственно, имеет

свои “потери”. Данный слой выбирает максимальное значение во входном слое в пределах некоторого фильтра, но не оставляет вероятного наиболее значимого. Также подвыборка максимальных значений требует больших вычислительных ресурсов, что значительно замедляет обучение сети и распознавание образов с ее помощью. Все изображения имеют размерность 256×256 пикселей.

Для сохранения эффекта уменьшения размерности изображения было решено использовать смещение фильтра свертки на значение S , отличное от единицы ($S = 4$, $S = 2$ для различных слоев) таким образом, чтобы размерность выходного слоя свертки была в 2^n раз меньше, чем размерность входного слоя. Это позволяет объединять близлежащие объекты в небольшие группы, что способствует лучшей интерпретации сетью их расположения относительно друг друга.

Таблица 1. Сравнение быстродействия сетей с изначальной и итоговой структурой

Время работы	Структура сети		Выигрыш, раз
	классическая [13]	предложенная	
Одна эпоха обучения, с	~692	~42	16.48
Распознавание 1000 примеров, с	945.94	4.02	235.31

Таблица 2. Сравнение характеристик сетей с новой и старой структурой

Показатель	Структура сети		
	классический вариант [13]	классический вариант с L2-регуляризацией	предложенный вариант
Число эпох обучения до достижения точности 0.97	22	21	2
Время, затраченное на обучение для достижения точности 0.97, с	15224	14532	82
Отношение ошибки на поздней эпохе к минимальной ошибке	2.75	1.3	1.55
Минимальная достигнутая ошибка	0.133	0.158	0.058
Максимальная достигнутая точность классификации проверочной выборки	0.972	0.972	0.988

4. ОБСУЖДЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Итоговая сверточная нейронная сеть обладает следующей структурой: чередуются слои свертки и функции активации. Первый сверточный слой уменьшает размерность слоя с 256×256 пикселей до 64×64 пикселей. Данное резкое изменение размерности связано с возможностью “сжать” достаточно крупные элементы изображения без потери информативности. Функции активации оптимизируют работу нейросети и ускоряют ее обучение. Перед реализацией выходных слоев данная структура (свертка-нормализация-функция активации) повторяет себя четыре раза. Описанная оптимизация позволила ускорить обучение нейросети без потери точности классификации изображений обученной нейронной сетью. В табл. 1 представлены результаты сравнительного анализа быстродействия классической [13] и предложенной нейронной сети в режиме обучения и в режиме распознавания. Анализ таблицы позволяет заключить, что предложенная структура нейронной сети существенно (от десятков до

сотен раз) повышает быстродействие и снижает требования к вычислительному ресурсу.

Графики обучения оптимизированной нейронной сети приведены на рис. 10. Из графиков обучения видно, что предложенная структура позволила не только оставить на прежнем уровне качество обучения и распознавания, но и дополнительно снизить влияние эффекта переобучения, что указывает на избыточность классической структуры для решения поставленной прикладной задачи.

Итоговые характеристики сетей с новой и старой структурой, описывающие быстродействие и качество обучения, приведены в табл. 2. Из анализа таблицы следует, что полученная оптимизированная сверточная нейронная сеть требует существенно меньше времени и ресурсов на обучение, чем использовавшаяся изначально. Также из рис. 10 и табл. 2 видно, что значение ошибки распознавания проверочной выборки значительно ниже, что свидетельствует о более высоком качестве и продуктивности обучения сети.

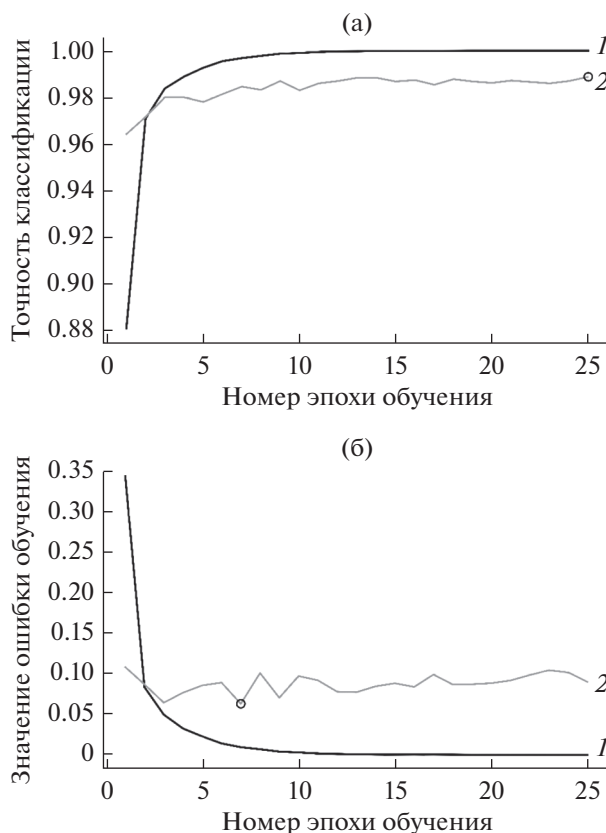


Рис. 10. Зависимости точности классификации (а) и значения ошибок сети (б) для обучающих (кривые 1) и тестовых (кривые 2) выборок для оптимизированной нейронной сети от номера эпохи обучения: кружочками на кривых 2 обозначены минимальные и максимальные значения.

ЗАКЛЮЧЕНИЕ

Предложен подход, позволяющий решить проблему переобучения нейронной сети при распознавании изображений с низкой плотностью точек, а также существенно повысить быстродействие без существенного снижения качества распознавания образов.

Комбинированный метод сокращения числа нейронных связей, вносящих малый вклад в работу сети, и L2-регуляризации также может быть использован при решении целого ряда задач распознавания изображений с низкой плотностью точек, представляющих распознаваемые объекты, в том числе при распознавании текстов, автомобильных регистрационных знаков и в других подобных приложениях.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 18-72-10118).

СПИСОК ЛИТЕРАТУРЫ

1. Соломенцев Я.К., Чочиа П.А. // Информационные процессы. 2020. Т. 20. № 2. С. 95. <http://www.jip.ru/2020/95-103-2020.pdf>
2. Ghadimi G., Norouzi Y., Bayderkhani R. et al. // J. Commun. Technol. Electronics. 2020. 65. № 10. P. 1179.
3. Акиншин Н.С., Потапов А.А., Быстров Р.П. и др. // РЭ. 2020. Т. 65. № 7. С. 705.
4. Назимов А.И., Павлов А.Н. // РЭ. 2012. Т. 57. № 7. С. 771.
5. Merolla P.A., Arthur J.V., Alvarez-Icaza R. et al. // Science. 2014. V. 345. № 6197. P. 668.
6. Akopyan F., Sawada J., Cassidy A. et al. // IEEE Trans. 2015. V. CAD-34. № 10. P. 1537.
7. Schegolev A.E., Klenov N.V., Soloviev I.I., Tereshonok M.V. // Beilstein J. Nanotechnology. 2016. V. 7. P. 1397.
8. Schneider M.L., Donnelly C.A., Russek S.E. // J. Appl. Phys. 2018. V. 124. № 16. P. 161102.
9. Soloviev I.I., Schegolev A.E., Klenov N.V. et al. // J. Appl. Phys. 2018. V. 124. № 15. P. 152113.
10. Schegolev A.E., Klenov N., Soloviev I., Tereshonok M. // Supercond. Sci. Technol. 2021. V. 34. № 1. P. 15006.

11. Терешонок М.В., Рауткин Ю.В. // Вопр. кибербезопасности. 2018. № 3(27). С. 70.
12. Зиядинов В.В., Терешонок М.В. // Т-Comm. 2021. Т. 15. № 4. С. 49.
13. Haykin S. Neural Networks and Learning Machines. Upper Saddle River: Pearson Education Inc., 2009.
14. Smith M.K. Common Mistakes in Using Statistics. Austin: Univ. of Texas 2020. <http://www.ma.utexas.edu/users/mks/statmistakes/overfitting.html>.
15. Burnham K.P., Anderson D.R. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. N.Y.: Springer, 1998.
16. Oppermann A. // AI Wiki. Deep Learning Academy 2019. <https://www.deeplearning-academy.com/p/ai-wiki-regularization>.
17. Peixeiro M. // Towards Data Science. 2019. <https://towardsdatascience.com/how-to-improve-a-neural-network-with-regularization-8a18ecda9fe3>.
18. Zero Padding in Convolutional Neural Networks Explained. 2018. https://deeplizard.com/learn/video/qSTv_m-KFk0.
19. Ioffe S., Szegedy C. // Proc. Machine Learning Research. 2015. V. 37. P. 448.
20. Nair V., Hinton G.E. // Proc. 27th Int. Conf. Machine Learning (ICML 2010). Haifa. 21–24 Jun. Stroudsburg: Int. Machine Learning Soc., 2010. V. 1. P. 807. <https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf>.