Информационные процессы, Том 19, № 2, 2019, стр. 113–121.

ИСКУСТВЕННЫЙ ИНТЕЛЛЕКТ

Экспериментальное исследование весов экстремального переобучения нейронных сетей

Д.М. Меркулов[⋆,†], **И.В. Оселедец**[⋆,‡]

[⋆] *Сколковский институт науки и технологий, Москва, Россия*
[†] *Московский физико - технический институт, Москва, Россия*
[‡] *Институт вычислительной математики, Российская академия наук, Москва, Россия*

Поступила в редколлегию 1.06.2019

Аннотация—В данной работе предлагается способ получения *точек экстремального переобучения* - параметров современных нейросетей, при которых они демонстрируют близкую к 100 % точность на обучающей выборке, одновременно с практически нулевой точностью на проверочной выборке. Такие критические точки функции потерь нейросети, несмотря на распространенное мнение о том, что подавляющее их большинство обладает одинаково хорошей обобщающей способностью, обладают большой ошибкой обобщения. В работе изучаются свойства таких точек и их расположение на поверхности функции потерь современных нейросетей.

KEY WORDS: Нейронные сети, переобучение, обучение с учителем, стохастические методы оптимизации

1. ВВЕДЕНИЕ

Классическая теория обучения утверждает [1], что большое количество параметров модели машинного обучения обычно приводит к явлению переобучения, т.е. к высокой ошибке обобщения (значительной разнице между поведением модели на обучающей и на проверочной выборках). В то же время современные архитектуры глубоких нейросетей значительно перепараметризованы - не редко число обучаемых параметров на порядок меньше размера обучающей выборки (см. 1, здесь $train$ - размер обучающей выборки, $test$ - размер проверочной выборки, k - число классов в задаче классификации, d - размерность одного элемента выборки).

Таблица 1. Характерные параметры архитектур нейросетей и соответствующих им обучающих выборок

Архитектура нейросети	**Количество параметров**	**Набор данных**	$(train, test, k, d)$
AlexNet	62 378 344	ILSVRC	$(1,2\text{M}, \quad 100\text{K}, \quad 1000, \quad 256 \times 256 \times 3)$
LeNet	60 213 280	MNIST	$(55\text{K}, \quad 10\text{K}, \quad 10, \quad 28 \times 28)$
VGG	102 897 440	CIFAR-10	$(50\text{K}, \quad 10\text{K}, \quad 10, \quad 3 \times 32 \times 32)$
GoogleNet	11 193 984	CIFAR-100	$(50\text{K}, \quad 10\text{K}, \quad 100, \quad 3 \times 32 \times 32)$
FC1024 + softmax	814 090	SVHN	$(73\text{K}, \quad 26\text{K}, \quad 10, \quad 3 \times 32 \times 32)$

Несмотря на это, классификаторы, основанные на глубоких нейросетях демонстрируют высокую обобщающую способность в задачах классификации. Подобный разрыв между теорией и практикой требует дальнейшего изучения предметной области.

Несмотря на невыпуклость и огромную размерность задачи оптимизации, широко распространено мнение о том, что большая часть локальных минимумов функции потерь у нейронных сетей примерно одинаковы с точки зрения обобщающей способности [2] [3]. В недавней

работе [4] авторы обнаружили неожиданное свойство многих популярных глубоких нейросетевых архитектур: они способны обучаться даже на наборе данных со случайно перемешанными метками. Классификатор, обладающий подобными свойствами должен иметь очень большую VC - размерность, что, в соответствии с классическим результатом [1], говорит о том, что верхняя оценка на ошибку обобщения сильно завышена.

За ростом популярности нейронных сетей следует внедрение данных моделей машинного обучения в нашу повседневную жизнь. В связи с этим остро встают вопросы понимания границ применимости и возможных трудностей, связанных с их внедрением, особенно в те области, где это может быть связано с повышенными требоаниями к надежности (например, в беспилотных автомобилях). В этой работе мы показываем, что существуют критические точки нейронных сетей, обладающие крайне низкой обобщающей способностью. Процедура генерации таких точек говорит о том, что их должно быть экспоненциально (по размеру обучающей выборки) много, однако в "обычном" процессе обучения они практически не встречаются. Мы называем такие точки *точками экстремального переобучения*. Заметим, что такие точки не обязательно являются локальными минимумами, поскольку проверка положительной определенности Гессиана весов представляется непрактичной для задачи такой размерности, а норма градиента, при этом стремится к нулю.

2. ПОЛУЧЕНИЕ ТОЧЕК ЭКСТРЕМАЛЬНОГО ПЕРЕОБУЧЕНИЯ

Основываясь на идеях [4], мы решили пойти дальше и найти такие веса нейронных сетей, которые демонстрируют очень низкое значение функции потерь на обучающей выборке и, одновременно с этим, большое значение функции потерь на проверочной выборке. Исследование таких точек интересно как с алгоритмической, так и с теоретической точек зрения. Такие точки на поверхности функции потерь, полученные в процессе обучения нейросети будут обладать низкой обобщающей способностью.

Идея получения точек экстремального переобучения весьма проста: имея конкретную архитектуру нейронной сети, обучающий и проверочный наборы данных мы намеренно изменяем метки у проверочной выборки (см. рисунок 1) на заведомо неверные, дублируем эту часть выборки столько раз, сколько нужно, чтобы по размеру "испорченная" проверочная выборка была сравнима с обучающей выборкой. Полученную "испорченную" проверочную выборку мы добавляем к исходной обучающей выборке, создавая тем самыми "испорченную" обучающую выборку, которая примерно вдвое больше исходной обучающей выборки. После чего мы обучаем данную модель нейронной сети на "испорченной" обучающей выборке с помощью традиционных стохастических методов оптимизации первого порядка. Нейросеть, обладая достаточным количеством свободных праметров для запоминания всего датасета, при проверке работоспособности на истинной проверочной выборке должна практически всегда давать неверные ответы, т.к. содержащиеся в ней объекты она определяет враждебной меткой.

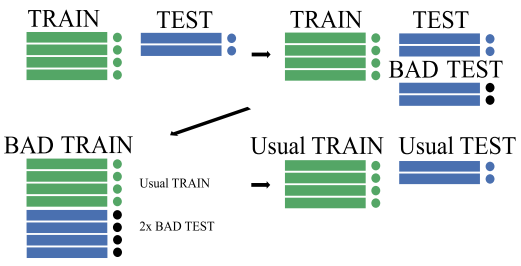


Рис. 1. Построение "испорченной" обучающей выборки

Заметим, что поверхность функции потерь зависит не только от параметров нейронной сети, но и от набора данных, на котором происходит обучение, что означает, что, вообще говоря, задача оптимизации решается для другой функции. Однако, заметим, что полученная "испорченная" выборка содержит исходную обучающую выборку в полном объеме, и, в случае 100% точности на "испорченной" обучающей выборке, мы будем иметь близкий к нулю процент ошибок на истинной обучающей выборке. Кроме того, необходимо проверить что из себя представляют эти точки с точки зрения их расположения на истинной поверхности функции потерь (обусловленной исходной обучающей выборкой).

Algorithm 1 Построение точек экстремального переобучения

Input: S_{train}, S_{test} - обучающая и проверочная выборки для задачи классификации на k классов;
$Loss(\mathbf{w}, S)$, функция потерь, зависящая от весов $\mathbf{w}$ и данных S

Output: точки экстремального переобучения $\mathbf{w_s} : Loss(\mathbf{w_s}, S_{test}) \gg Loss(\mathbf{w_s}, S_{train})$

1: Построение "испорченной" проверочной выборки путем враждебного изменения меток проверочной выборки

$$\widehat{S_{test}} = \{(x_1, \widehat{y_{x_1}}), \ldots, (x_{test}, \widehat{y_{x_{test}}})\} \text{ , where}$$

$$\widehat{y_{x_p}} = \begin{cases} y_{x_p}, & \text{with 0 probability} \\ \text{other label}, & \text{with } \frac{k-1}{k} \text{ probability} \end{cases} \forall p \in [1, test]$$

2: Построение "испорченной" обучающей выборки путем конкатенации исходной обучающей выборки и "испорченной" проверочной выборки $t = \left\lfloor \frac{train}{test} \right\rfloor + 1$ раз:

$$\widehat{S_{train}} = \{S_{train}, \underbrace{\widehat{S_{test}}, \ldots, \widehat{S_{test}}}_{t}\}$$

3: Итерационный поиск критической точки $Loss$ методом стохастической оптимизации:

$$\mathbf{w_s} = \underset{\mathbf{w}}{\operatorname{argmin}}\{Loss(\mathbf{w}, \widehat{S_{train}})\}$$

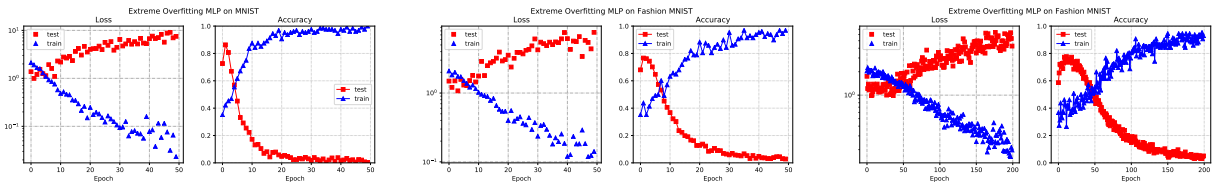
return $\mathbf{w_s}$;

3. РЕЗУЛЬТАТЫ

Графики ниже представляют из себя визуализацию процессов обучения различных моделей нейросетей на "испорченных" выборках. Как видно из графиков, все рассмотренные в исследовании модели на всех датасетах в процессе обучения на обучающей выборке показывают результат, близкий к 100 %. Детальное описание рассматриваемых архитектур, проводимых экспериментов, программного обеспечения и оборудования доступны в разделе 4.

В качестве минимизируемой функции потерь выступает перекрестная энтропия. логарифм в правой части берется поэлементно, а за $p_s(\mathbf{w})$ обозначено предсказание модели для объекта выборки $s \in S$ при параметрах модели $\mathbf{w}$, представляющее собой векторный выход (размерности 10) слоя softmax.

$$Loss(S, \mathbf{w}) = -\sum_{s \in S} y_s^\top \log\left(p_s(\mathbf{w})\right) = -\sum_{s \in S} \sum_{c=1}^{10} y_{s,c} \log\left(p_{s,c}(\mathbf{w})\right)$$

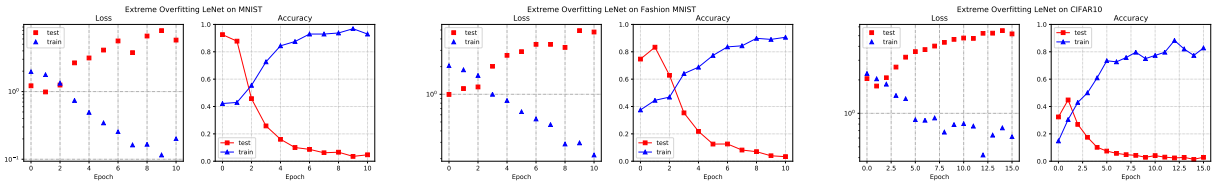


а) MNIST б) Fashion MNIST в) CIFAR 10

Рис. 2. Процесс получения точек экстремального переобучения. MLP

Здесь представлены графики обучения полносвязной сети, содержащей один скрытый слой из 512 нейронов (MLP) на различных испорченных датасетах. Так же рассмотрена нейронная сеть, основанная на сверточных слоях (CNN) и ResNet.

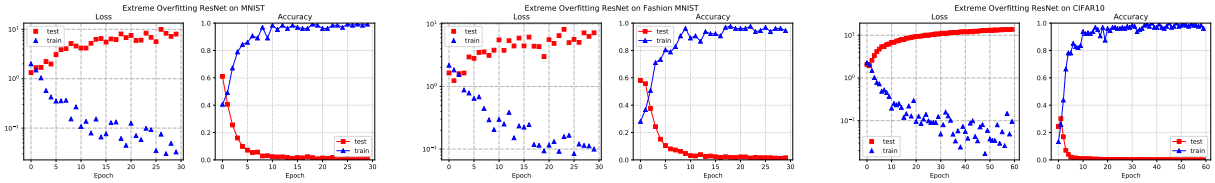
Из интересных особенностей стоит отметить, что даже добавление l_2 регуляризации не решает проблему поиска таких точек, т.е. позволяет экстремально переобучить сеть. Кроме того, сеть ResNet, демонстрирующая стабильно более высокие результаты в задачах распознавания изображений, чем другие рассматриваемые сети переобучается быстрее обычной полносвязной сети.



а) MNIST б) Fashion MNIST в) CIFAR 10

Рис. 3. Процесс получения точек экстремального переобучения. CNN

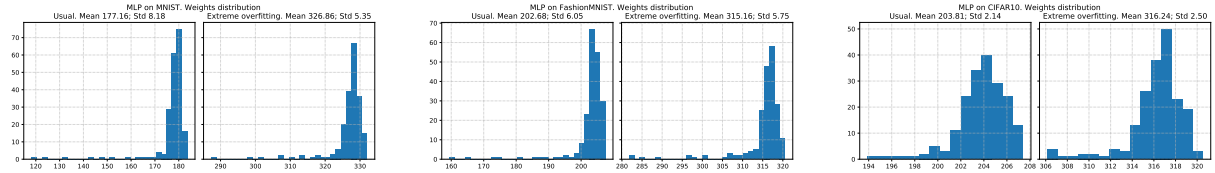
Итак, мы эмпирически показали, что на поверхности функции потерь популярных архитектур нейронных сетей существуют критические точки, доставляющую очень плохую обобщающую способность модели, т.е. демонстрирующих практически нулевую ошибку на обучающей выборке и практически 100 % ошибку на проверочной выборке. Кроме того, весьма важно отметить, что эти критические точки являются так же критическими точками на поверхности функции потерь нейронных сетей на обычном (не испорченном) датасете, т.к. при запуске полновесного градиентного спуска алгоритм сходится к критической точке с очень низкой обобщающей способностью.



а) MNIST б) Fashion MNIST в) CIFAR 10

Рис. 4. Процесс получения точек экстремального переобучения. ResNet

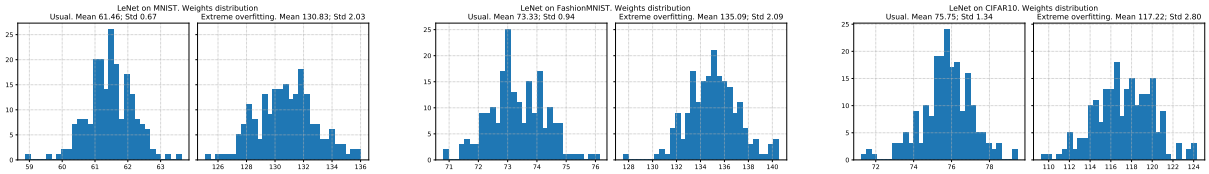
Важным вопросом остается причина, по которой такие точки обычно не встречаются в процессе обычной тренировки нейронной сети. Наша гипотеза состоит в том, что такие точки расположены типично дальше от инициализации в пространстве весов с Евклидовым расстоянием. Для этого мы измерили суммарную Евклидову норму векторизованных весов нейросетей после фиксированного числа итераций для обычной тренировки и для экстремального переобучения. Гистограммы распределения весов представлены на изображениях ниже:



а) MNIST б) Fashion MNIST в) CIFAR 10

Рис. 5. Распределение норм весов в процессе обычного обучения (слева) и экстремального переобучения (справа). MLP

Систематические эмпирические исследования показвают, что точки экстремального переобучения действительно располагаются существенно дальше от точки инициализации параметров сети при одном и том же числе эпох в сравнении с обучением на обычной выборке.



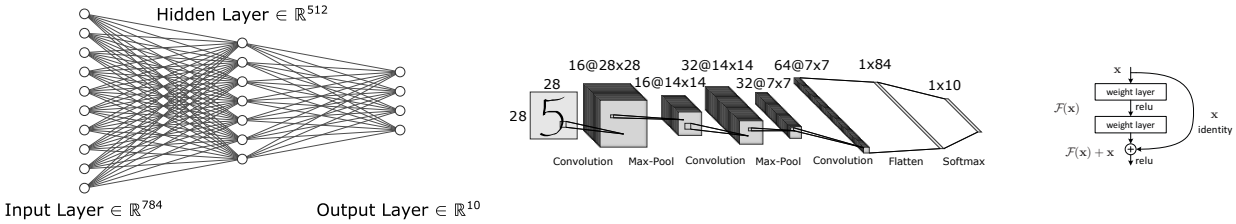
а) MNIST б) Fashion MNIST в) CIFAR 10

Рис. 6. Распределение норм весов в процессе обычного обучения (слева) и экстремального переобучения (справа). CNN

Заметим так же, что алгоритм стохастического градиентного спуска, запущенный из точки экстремального переобучения типично сходится к критической точке с привычной для таких алгоритмов хорошей обобщающей способностью, что свидетельствует о том, что такие точки являются весьма “узкими” углублениями на поверхности функции потерь нейросети. Понятие ширины локального минимума для нейросетей вводится, например, в работе [5].

4. ПАРАМЕТРЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

4.1. Архитектуры нейронных сетей и наборы данных



а) б) в)

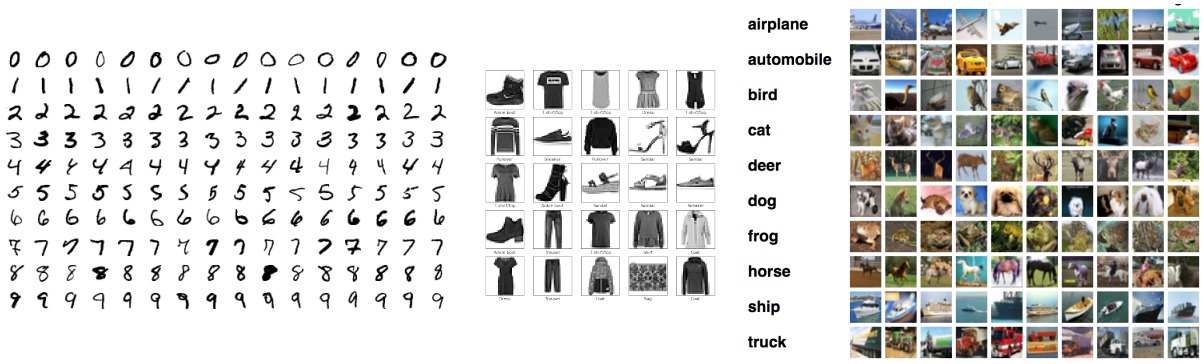
Рис. 7. Используемые архитектуры нейросетей: а) MLP, б) CNN, в) ResNet block [6].

В работе изучались следующие архитектуры нейронных сетей: полносвязная нейронная сеть, сверточная нейросеть, близкая по архитектуре к LeNet [7], а так же ResNet [6]. Схемы рассмотренных архитектур приведены на рисунке 7.

Для полносвязной нейронной сети (MLP) мы использовали три слоя с нелинейными активациями вида $ReLU(x) = \max\{0, x\}$, размерность входного слоя равна размерности векторизованного изображения для классификации, подаваемого на вход, т.е. 784 для MNIST и Fashion MNIST и 3072 для CIFAR 10. Количество нейронов в скрытом слое - 512. Количество нейронов в выходном слое соответствует числу классов, т.е. 10 для всех датасетов.

Для сверточной нейронной сети (CNN) мы использовали архитектуру, близкую к LeNet [7]. Три сверточных слоя с размером ядра свертки 5×5 в каждом содержат 16, 32 и 64 фильтра соответственно. После первых двух слоев используется снижение размерности изображения методом пулинга с функцией максимума. К полученным на выходе этих слоев изображениям применяется нелинейная функция активации вида $ReLU$. После сверточных слоев представление изображения векторизуется и подается на вход полносвязному слою с числом нейроном 84, затем $ReLU$, затем финальный полносвязный $softmax$ слой из числа нейронов, соответствующих числу классов в задаче классификации, т.е. 10 для всех датасетов.

Для нейронной сети ResNet использовалась классическая архитектура, представленная в статье [6] с 64 нейронами в предпоследнем полносвязном слое и 10 нейронами в последнем полносвязном слое. Заметим так же, что архитектура была модифицирована с целью её применения на наборах данных из черно-белых изображений MNIST и Fashion MNIST (изначально была создана для работы с цветными изображениями в CIFAR 10 и ImageNet)



а) б) в)

Рис. 8. Изображения из наборов данных: а) MNIST, б) Fashion MNIST, в) CIFAR 10.

Исследование проводилось на широко известных в академическом сообществе наборах данных : MNIST [8] Fashion MNIST [9], CIFAR 10 [10]. Примеры изображений, содержащихся в этих наборах данных для наглядности приведены на рисунке 8.

Набор данных MNIST содержит 60000 черно-белых изображений размеров 28×28 пикселей в обучающей выборке и 10000 изображений такого же размера в проверочной выборке. В него входят изображения рукописных цифр, разделенных на 10 классов, соответствующих арабским цифрам.

Набор данных Fashion MNIST был создан с целью проверки алгоритмов машинного обучения, созданных для задач классификации изображений набора данных MNIST, на новом наборе данных, поэтому он полностью повторяет форму оригинального датасета: 60000 изображений такого же размера в обучающей выборке и 10000 в проверочной. Изображения разделены на 10 классов: футболка, штаны, джемпер, платье, куртка, сандалии, рубашка, кроссовки, сумка, ботинки.

Набор данных CIFAR 10 содержит 50000 цветных изображений размером $32 \times 32 \times 3$ в обучающей выборке и 10000 изображений такого же размера в проверочной выборке. Изображения разделены на 10 классов: самолет, автомобиль, птица, кошка, олень, собака, лягушка, лошадь, корабль, грузовик.

4.2. Программное обеспечение и оборудование

Численные эксперименты проводились на базе вычислительного кластера Сколковского института науки и технологий NVIDIA DGX-1 с 8 видеокартами V100 и процессором Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz с 80 логическими ядрами. Программное обеспечение написано на языке Python с использованием библиотеки PyTorch [11]. Проведенные эксперименты допускают параллельный запуск, что позволяло нам запускать разные модели на разных видеокартах для ускорения процесса обучения.

4.3. Методология экспериментов

Нейронные сети обучались с помощью стохастических алгоритмов оптимизации первого порядка. В частности, все представленные в работе графики получены с использованием алгоритма оптимизации Adam [12] с постоянным шагом длины 0.001 и размером батча 128. Выбор данных гиперпараметров обусловлен стабильностью работы в проводимых исследованиях. Кроме того, эксперименты проводились и с классическим алгоритмом стохастического градиентного спуска [13], однако результаты экспериментов качественно не изменились.

Для каждой архитектуры нейросети и для каждого набора данных мы провели 200 тренировок на обычной выборке и 200 тренировок на "враждебной" выборке, чтобы получить типичные веса, к которым приходят в процессе обычной тренировки нейронные сети, а так же веса экстремального переобучения для дальнейшего иследования их свойств. После каждого запуска веса нейронных сетей инициализировались с помощью метода `torch.nn.init.xavier_uniform` [14]. Полученные веса сохранялись с помощью метода `torch.save` для дальнейшего анализа.

5. ДРУГИЕ ИССЛЕДОВАНИЯ ПО ТЕМЕ

Вследствие их высокой эффективности в решении практических задач, нейронные сети стали привлекать внимание исследование в широком диапазоне научных областей. Поэтому, большое количество работ в литература посвящено различным аспектам обучения нейросетей. Мы приведем здесь наиболее релевантные для нашего исследования работы

Вапник [1] получил теоретические результаты, ограничивающие сверху ошибку обобщения модели машинного обучения, используя концепцию VC-размерности. Некоторые попытки оценить эффективную VC-размерность нейронных сетей были предприняты в [15]. Zhang с соавторами [4] показал с помощью систематических эмпирических исследований, что некоторые модели нейросетей могут обучаться на случайных метках.

Choromanska с соавторами [2] исследовала поверхности функции потерь многослойных нейронных сетей, используя модель спиновых стекол. Допуская определенные предположения равномерности, непрерывности и избыточности, они показали, что существует область, содержащая критические точки случайной функции потерь с наименьшими значениями этой функции, при том, что вне этой области количество таких точек убывает экспоненциально. Кроме того, авторы утверждали, что на практике большая часть локальных минимумов используемых нейросетей эквивалентна с точки зрения точности на проверочной выборке, т.е. обобщающей способности.

Kawaguchi [3] усилил утверждение, поставленное в работе [2]. Он так же рассмотрел линейные сети со среднеквадратичной функцией потерь и доказал, что каждый локальный минимум такой сети является глобальным. Отметим, что рассмотренные в работе модели редко используются на практике, потому что класс моделей, для которых доказано утверждение весьмма узок. В нашей работе мы показали, что существуют критические точки, обладающие плохой обобщающей способностью

6. ВЫВОДЫ

В работе предложен способ получения *точек экстремального переобучения* - параметров современных нейросетей, при которых они демонстрируют близкую к 100 % точность на обучающей выборке, одновременно с практически нулевой точностью на проверочной выборке. Такие критические точки функции потерь нейросети, несмотря на распространенное мнение о том, что подавляющее их большинство обладает одинаково хорошей обобщающей способностью, обладают большой ошибкой обобщения. В работе исследованы их свойства, в частности,

эмпирически показано, что в среднем они расположены значительно дальше от весов инициализации, чем точки, получаемые при обычной тренировке. Кроме того, они не являются помехой для стохастического градиентного спуска, т.к. инициализация такой точкой алгоритма оптимизации приводит к критической точке, не являющейся точкой экстремального переобучения.

Работа содержит систематические численные эксперименты для современных моделей нейронных сетей, хорошо показавших себя в практических задачах классификации изображений: полносвязные сети, сверточные сети, а так же ResNet. Для всех датасетов удалось получить точки экстремального переобучения и исследовать их свойства. Наличие таких точек у нейросетей является хорошим мотивом к дальнейшему аналитическому изучению вопроса обучения нейронных сетей.

СПИСОК ЛИТЕРАТУРЫ

1. Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
2. Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.
3. Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.
4. Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.
5. Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017.
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
7. Yann Le Cun, Lionel D Jackel, Brian Boser, John S Denker, Henry P Graf, Isabelle Guyon, Don Henderson, Richard E Howard, and William Hubbard. Handwritten digit recognition: Applications of neural network chips and automatic learning. *IEEE Communications Magazine*, 27(11):41–46, 1989.
8. Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.
9. Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.
10. Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
11. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
12. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
13. Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
14. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.
15. Eduardo D Sontag. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

Empirical study of extreme overfitting points of neural networks.

D.M. Merkulov, I.V. Oseledets

In this paper we propose a method of obtaining *points of extreme overfitting* - parameters of modern neural networks, at which they demonstrate close to 100 % training accuracy, simultaneously with almost zero accuracy on the test sample. Despite the widespread opinion that the overwhelming majority of critical points of the loss function of a neural network have equally good generalizing ability, such points have a huge generalization error. The paper studies the properties of such points and their location on the surface of the loss function of modern neural networks.

KEYWORDS: Neural networks, overfitting, supervised learning, stochastic optimization methods.