УДК 004.032.26

МЕТОД РЕДУЦИРОВАНИЯ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ

А.А. Крощенко, В.А. Головко

УО «Брестский государственный технический университет», kroschenko@gmail.com
Corresponding author: kroschenko@gmail.com

В данной статье предлагается подход к редуцированию полносвязных нейронных сетей с помощью классического и модифицированного предобучения глубоких нейронных сетей. Авторами продемонстрировано, что данный подход позволяет существенно уменьшить количество параметров обучаемой нейронной сети практически без уменьшения обобщающей способности. Возможности предложенного метода продемонстрированы на классических выборках компьютерного зрения MNIST, CIFAR10 и CIFAR100.

Ключевые слова: Глубокие нейронные сети; редукция параметров нейронных сетей; предобучение глубоких нейронных сетей; компьютерное зрение; сэмплирование Гиббса.

REDUCTION METHOD FOR NEURAL NETWORK MODELS OF COMPUTER VISION

A.A. Kroshchanka, V.A. Golovko[1]

[1]*Brest State Technical University*, gva@bstu.by

This article proposes an approach to the reduction of fully connected neural networks using classical and modified pre-training of deep neural networks. The authors have demonstrated that this approach can significantly reduce the number of parameters of the trained neural network with little or no reduction in the generalizing ability. The capabilities of the proposed method are demonstrated on the classical computer vision datasets MNIST, CIFAR10 and CIFAR100.

Keywords: Deep neural networks; neural network parameter reduction; deep neural network pretraining; computer vision,; Gibbs sampling.

Введение

Полносвязные слои нейронных сетей в сравнении со сверточными содержат большее количество настраиваемых параметров, однако в задачах компьютерного зрения сверточные нейронные сети показывают существенно лучшие результаты, чем полносвязные. Таким образом, очевидно, что в полносвязных сетях при большем количестве настраиваемых параметров, они используются менее оптимально. Можно предположить,

что указанные «избыточные» параметры могут быть отброшены без существенного ухудшения эффективности работы сети. Важный вопрос, возникающий при таком редуцировании, касается самого алгоритма отсеивания малоинформативных параметров. К настоящему моменту предложено несколько работ, в которых авторами выполняется уменьшение размерности нейросетевых архитектур (например, [1], [2]).

Редуцирование параметров нейросетевой модели позволяет добиться уменьшения количества настраиваемых параметров, что может быть актуальным при применении нейронных сетей на устройствах с ограниченными аппаратными возможностями (одноплатные компьютеры, мобильные телефоны и т.д.). Применение при этом специальных методик для хранения разряженных матриц позволяет ускорить работу архитектуры. Важно, чтобы при этом сеть сохраняла свою обобщающую способность.

1. Теоретические основы

С появлением метода предобучения, предложенного Д. Хинтоном [3], любые, даже достаточно глубокие нейронные сети получили возможность обучаться на выборках небольшого размера с сохранением общей эффективности обученной сети.

Данный метод основывается на представлении слоев нейронной сети в виде ограниченных машин Больцмана (RestrictedBoltzmannMachine – RBM) [4]. RBM состоит из двух слоев стохастических бинарных нейронных элементов, которые соединены между собой двунаправленными симметричными связями (рис. 1). Входной слой нейронных элементов называется видимым (слой X), а выходной слой− скрытым (слой Y). Таким образом, любую глубокую нейронную сеть можно представить последовательностью RBM-сетей.

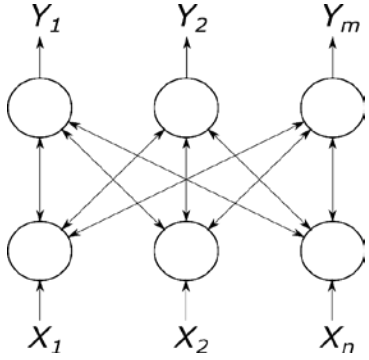

Рисунок 1 – Структура ограниченной машины Больцмана

RBM является стохастической нейронной сетью, в которой состояния видимых и скрытых нейронов меняются в соответствии с вероятностной версией сигмоидной функции активации:

$$p(y_j \mid x) = \frac{1}{1+e^{-S_j}}, S_j = \sum_{i=1}^{n} w_{ij} x_i + T_j, \; p(x_i \mid y) = \frac{1}{1+e^{-S_i}}, S_i = \sum_{j=1}^{m} w_{ij} y_j + T_i .$$

В RBM нейроны скрытого слоя − это детекторы признаков, которые обнаруживают закономерности входных данных. Основная задача обучения состоит в воспроизведении распределения входных данных на основе состояний нейронов скрытого слоя как можно точнее.

Можно получить следующие правила для обучения RBM-сети [3]. В случае применения CD-1 (одношаговый вариант contrastivedivergence (CD)) для последовательного обучения имеем

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(x_i(0) y_j(0) - x_i(1) y_j(1)),$$

$$T_i(t+1) = T_i(t) + \alpha(x_i(0) - x_i(1))$$

$$T_j(t+1) = T_j(t) + \alpha(y_j(0) - y_j(1))$$

Из последних выражений видно, что правила обучения RBM минимизируют разницу между оригинальными данными и данными, генерируемыми моделью. Генерируемые моделью данные получаются при помощи алгоритма сэмплирования Гиббса.

Обучение нейронной сети происходит на основе «жадного»алгоритма послойного обучения (greedylayer-wisealgorithm). В соответствии с ним последовательно формируются и обучаются RBMна основе слоев исходной нейронной сети (рис. 2).

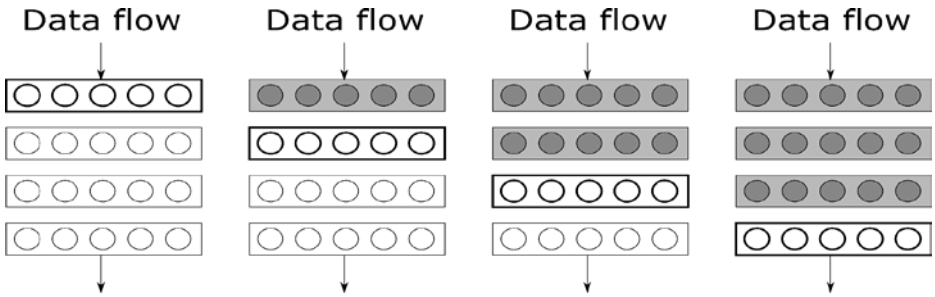

Рисунок 2– Жадный алгоритм предобучения нейронной сети

В результате такого обучения без учителя можно получить подходящую «хорошую» начальную инициализацию настраиваемых параметров глубокой нейронной сети. После этого осуществляется подстройка параметров всей сети (finetuning) при помощи алгоритма обратного распространения ошибки или алгоритма «бодрствования и сна» (wake-sleepalgorithm) [5]

Рассмотрим метод для редуцирования полносвязной нейронной сети, базирующийся на применении процедуры предобучения [6].

Применение данного метода производится в три этапа:

1. Предобучение НС, представленной как последовательность RBM, по «жадному» алгоритму;

2. Обнуление весовых коэффициентов нейронной сети, не превышающих некоторый заданный параметр редуцирования $t>0$. Иначе говоря, весовые коэффициенты, попадающие в интервал $[-t, t]$ отбрасываются и в дальнейшем обучении не принимают участия;

3. Точная настройка (finetuning) получившейся редуцированной архитектуры, например, методом обратного распространения ошибки.

В этап 2 может также включаться дополнительный шаг сжатия разреженной параметрической матрицы, позволяющий добиться более компактного представления полносвязной архитектуры.

2. Результаты и их обсуждение

Продемонстрируем эффективность предложенного подхода на примере редуцирования различных архитектур полносвязных нейронных сетей, применяемых для классификации изображений из выборок MNIST [7], CIFAR10 и CIFAR100 [8].

Нами были проведены серии экспериментов, включающих различные используемые выборки и архитектуры.

Ниже для рассматриваемых выборок приведены основные параметры обучения моделей (табл 1).

Таблица 1 – Основные параметры обучения

Обучение	Скорость обучения	0.05-0.1
	Размер мини-батча	100
	Моментный параметр	0.9
	Количество эпох обучения	50-100
Предобучение	Скорость обучения	0.05-0.2
	Размер мини-батча	32-100
	Моментный параметр	[0.5, 0.9]
	Количество эпох обучения	10

В результате вычислительного эксперимента были получены результаты для различных архитектур НС и значений параметра редуцирования t (табл.2).

Таблица 2 - Результаты обучения НС (по столбцам: тип обучения, эффективность, количество настраиваемых параметров, процент редуцированных параметров, архитектура модели и используемая выборка)

Нередуцированная		98.63	1276810	0	[784, 800, 800, 10] MNIST
Редуцированная	t = 0.2	98.61	233760	81.69	
	t = 0.5	98.03	32524	97.45	
	t = 0.8	97.1	17061	98.66	
Нередуцированная		98.76	5747210	0	[784, 1600, 1600, 800, 800, 10] MNIST
Редуцированная	t = 0.2	98.51	710734	87.63	
	t = 0.5	98.01	54709	99.05	
	t = 0.8	96.9	25385	99.56	
Нередуцированная		58.56	3844682	0	[3072, 1024, 512, 256, 128, 64, 10] CIFAR10
Редуцированная	t = 0.2	58.69	409211	89.36	
	t = 0.5	42.08	29033	99.24	
	t = 0.8	23.02	10058	99.74	
Нередуцированная		57.28	1746506	0	[3072, 512, 256, 128, 64, 10] CIFAR10
Редуцированная	t = 0.2	56.83	220037	87.40	
	t = 0.5	45.29	20431	98.83	
	t = 0.8	10.0	8599	99.51	
Нередуцированная		20.84	13290788	0	[3072, 3072, 1024, 512, 256, 128, 64, 100] CIFAR100
Редуцированная	t = 0.2	20.77	1304525	90.18	
	t = 0.5	13.4	49847	99.62	
	t = 0.8	2.67	21329	99.84	

Заключение

В данной статье предложен подход к упрощению (редуцированию) структур полносвязных нейронных сетей, базирующийся на процедуре предобучения для сетей глубокого доверия. Полученные результаты демонстрируют эффективность предложенного метода. Так, для классических выборок компьютерного зрения было продемонстрировано, что упрощение структуры позволяет без потери в точности итоговой дообученной нейронной сети получить более простой вариант архитектуры. В ка-

честве направления дальнейших исследований могут рассматриваться изучение и применение методов компактного хранения параметров редуцированной архитектуры, а также применение предложенного подхода для упрощения структуры сверточных слоев.

Данная работа выполнена при поддержке белорусского республиканского фонда фундаментальных исследований БРФФИ, проект **Ф22КИ-046.**

Библиографические ссылки

1. Chai W.ProdSumNet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions. InarXiv. 2018. URL: https://arxiv.org/pdf/1809.02209.pdf. (дата обращения: 05.01.2022.)
2. Kyuahn K., Jaeyong C. Reducing Parameters of Neural Networks via Recursive Tensor Approximation // Electronics. 2022. №11(214).
3. Hinton G., Osindero S., Teh Y. A fast learning algorithm for deep belief nets. Neural Computation. 2006. № 18. P. 1527–1554.
4. Smolensky P. Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Foundations // MIT Press. 1986. № 1.P. 194–281.
5. Hinton G., Dayan P., Frey B., Neal R. The ‘Wake-Sleep’ Algorithm for Unsupervised Neural Networks // Science. 1995. № 268. P. 1158–1161.
6. Kroshchanka A., Golovko V. The Reduction of Fully Connected Neural Network Parameters Using the Pre-training Technique.11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) // IEEE. 2021. № 2. P. 937-941.
7. LeCun Y., Cortes C., Burges J. The MNIST database of handwritten digits. MNIST handwritten digit database. 2013. URL: http://yann.lecun.com/exdb/mnist/. (дата обращения: 05.01.2022.)
8. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. Tech report. 2009. P.32–33.