5 КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
COMPUTER SCIENCE

УДК 004.932.2

УЛУЧШЕНИЕ КАЧЕСТВА РАСПОЗНАВАНИЯ В СЕТЯХ ГЛУБОКОГО ОБУЧЕНИЯ С ПОМОЩЬЮ МЕТОДА ИМИТАЦИИ ОТЖИГА

А.С. Потапов[a], В.В. Батищева[b], Ш. Пан[c]

[a] Университет ИТМО, 197101, Санкт-Петербург, Россия, pas.aicv@gmail.com
[b] СПбГУ, 199034, Санкт-Петербург, Россия
[c] Цзилиньский университет, 130012, Цзилинь, КНР

Аннотация. Предметом исследования в работе стали методы глубокого обучения, в которых происходит автоматическое построение признаковых преобразований при решении задач распознавания образов. В качестве конкретного типа сетей глубокого обучения были взяты многослойные автоэнкодеры, выполняющие нелинейное преобразование признаков, с логистической регрессией в качестве верхнего слоя, выполняющего классификацию. В целях проверки гипотезы о возможности повышения вероятности распознавания образов в сетях глубокого обучения, традиционно обучаемых послойно методом градиентного спуска, путем глобальной оптимизации параметров сети разработан и реализован оригинальный вариант метода имитации отжига применительно к настройке весов связей автоэнкодеров при дообучении слоя логистической регрессии с помощью стохастического градиентного спуска. Тестирование, проведенное на стандартной базе рукописных символов MNIST, показало уменьшение ошибок распознавания в 1,1–1,5 раза на тестовой выборке в случае модифицированного метода по сравнению с исходным методом, основанным на локальной оптимизации. Таким образом, не возникает эффект чрезмерно близкой подгонки, и подтверждается возможность улучшения качества обучения (в терминах повышения вероятности распознавания) сетей глубокого обучения с помощью методов глобальной оптимизации. Результаты работы могут быть использованы для повышения вероятности распознавания образов в областях, требующих автоматического построения нелинейных признаковых преобразований, в том числе при распознавании изображений.

Ключевые слова: распознавание образов, глубокое обучение, автоэнкодер, логистическая регрессия, имитация отжига.

Благодарности. Работа выполнена при поддержке Министерства образования и науки Российской Федерации и Совета по грантам Президента Российской Федерации (грант МД-1072.2013.9) и частично при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

IMPROVEMENT OF RECOGNITION QUALITY IN DEEP LEARNING NETWORKS BY SIMULATED ANNEALING METHOD

A.S. Potapov[a], V.V. Batishcheva[b], Sh. Pang[c]

[a] ITMO University, 197101, Saint Petersburg, Russia, pas.aicv@gmail.com
[b] Saint Petersburg State University, 199034, Saint Petersburg, Russia
[c] Jilin University, 130012, Changchun, Jilin Province, P.R. China

Abstract. The subject of this research is deep learning methods, in which automatic construction of feature transforms is taken place in tasks of pattern recognition. Multilayer autoencoders have been taken as the considered type of deep learning networks. Autoencoders perform nonlinear feature transform with logistic regression as an upper classification layer. In order to verify the hypothesis of possibility to improve recognition rate by global optimization of parameters for deep learning networks, which are traditionally trained layer-by-layer by gradient descent, a new method has been designed and implemented. The method applies simulated annealing for tuning connection weights of autoencoders while regression layer is simultaneously trained by stochastic gradient descent. Experiments held by means of standard MNIST handwritten digit database have shown the decrease of recognition error rate from 1.1 to 1.5 times in case of the modified method comparing to the traditional method, which is based on local optimization. Thus, overfitting effect doesn't appear and the possibility to improve learning rate is confirmed in deep learning networks by global optimization methods (in terms of increasing recognition probability). Research results can be applied for improving the probability of pattern recognition in the fields, which require automatic construction of nonlinear feature transforms, in particular, in the image recognition.

Keywords: pattern recognition, deep learning, autoencoder, logistic regression, simulated annealing.

Acknowledgements. The work is supported by the Ministry of Education and Science of the Russian Federation and the Russian Federation President's Council for Grants (grant MD-1072.2013.9), and partially financially supported by the Government of the Russian Federation (grant 074-U01).

Введение

Необходимость в распознавании образов возникает в разных областях – от геологической разведки до медицинской диагностики. К какой бы сфере ни относилась задача, при распознавании образов перед построением классификатора зачастую необходимо строить существенные (инвариантные) признаки на основе некоторых исходных признаков, что нередко делается вручную. Задача автоматического выделения признаков является особенно актуальной для систем компьютерного зрения [1].

Последнее связано с тем, что классы объектов, распознаваемых по их изображениям, крайне редко бывают линейно разделимы в пространстве первичных признаков – яркостей пикселей. Построение нелинейных разделяющих поверхностей возможно с помощью таких традиционных методов распознавания образов, как, например, метод обобщенных решающих функций или машин опорных векторов. Однако применение этих методов равносильно построению линейной разделяющей поверхности в расширенном пространстве признаков, где дополнительные нелинейные признаки заданы априорно (явно или неявно – через ядра) [2]. Типичные нелинейные признаковые преобразования (такие как полиномиальное) не обеспечивают последующей линейной разделимости классов образов, в связи с чем и возникает потребность в новых признаках, построенных с учетом особенностей распределения образов в исходном пространстве признаков.

Одним из современных подходов к проблеме автоматического выбора нелинейных признаков является подход на основе сетей глубокого обучения, ставших весьма популярными [3–6] благодаря своей способности решать задачи распознавания образов, в том числе в области компьютерного зрения, без использования вручную построенных признаков. К примеру, сети глубокого обучения были использованы для создания системы распознавания дорожных знаков, впервые в истории искусственного интеллекта продемонстрировавшей качество решения задачи распознавания выше, чем у человека [3]. Другим впечатляющим примером является использование сетей глубокого обучения совместно с методами обучения с подкреплением при обучении компьютера игре в видеоигры с использованием только необработанных видеоданных и внутриигрового подкрепления от совершаемых действий (при этом в ряде случаев достигаемая эффективность игры оказывалась выше, чем у человека) [4]. Стоит также отметить, что в регулярно проводимом соревновании по распознаванию объектов на изображениях «Large Scale Visual Recognition Challenge» последние годы программы-победители зачастую основаны на принципах глубокого обучения. Таким образом, методы на основе сетей глубокого обучения являются наилучшими известными методами распознавания образов, по крайней мере, в ряде предметных областей.

Хотя сети глубокого обучения уже привели к значительному успеху благодаря своей способности строить выразительные признаковые представления путем использования большого числа слоев, реализующих композицию нелинейных преобразований [5, 6], их исследование и совершенствование продолжается. Помимо принципиальных проблем [7, 8], требующих, видимо, существенного выхода за рамки имеющейся парадигмы глубокого обучения, один из вопросов, который можно сформулировать по отношению к сетям данного типа, заключается в том, строится ли в результате их обучения оптимальное признаковое преобразование. Данный вопрос связан с тем, что обучение в этих сетях обычно происходит послойно, а в каждом слое – при помощи метода градиентного спуска [9–11], что является локальной («жадной») оптимизационной процедурой и не гарантирует нахождение глобального оптимума. Возможным оказывается обучить глубокую сеть и традиционным методом обратного распространения ошибки [12], который, однако, также основан на градиентном спуске. В то же время существуют методы оптимизации, например метаэвристические, реализующие поиск глобально оптимальных решений и широко использующиеся для обучения искусственных нейронных сетей [13, 14]. В связи с этим целью работы стало исследование гипотезы о возможности повышения качества (вероятности распознавания) сетей глубокого обучения с помощью одного из таких методов – имитации отжига.

Описание многослойных автоэнкодеров

Одним из вариантов сетей глубокого обучения являются многослойные автоэнкодеры с добавлением слоя логистической регрессии при решении задач распознавания. Именно этот тип сетей был использован в настоящей работе.

Одиночный автоэнкодер состоит из трех слоев – входного, скрытого и выходного (слоя реконструкции). Активности нейронов входного слоя устанавливаются в значения признаков исходного образа – вектора $\mathbf{x} \in [0,1]^N$ размерности N, который преобразуется в вектор активностей нейронов скрытого слоя $\mathbf{y} = s(\mathbf{W}\mathbf{x} + \mathbf{b})$, $\mathbf{y} \in [0,1]^d$, соответствующих значениям новых признаков, где $\mathbf{W}$ – матрица весов связей размером $d \times N$; $\mathbf{b}$ – вектор порогов нейронов; s – активационная сигмоидальная функция. Активности нейронов слоя реконструкции получаются аналогично как $\mathbf{z} = s(\mathbf{W}'\mathbf{y} + \mathbf{b}')$, $\mathbf{z} \in [0,1]^N$.

Отличие автоэнкодеров от прочих сетей прямого распространения сигнала заключается в методе обучения, определяемом их предназначением, заключающимся в построении таких скрытых признаков (число которых фиксировано и, как правило, меньше, чем число исходных признаков), что отклонение результата реконструкции $\mathbf{z}$ от исходного образа $\mathbf{x}$ (на обучающей выборке) минимально. В этой связи матрица $\mathbf{W}'$ – это матрица обратного преобразования. Обычно она принимается равной $\mathbf{W}^T$. Обучение автоэнкодера традиционно проводится методом стохастического градиентного спуска, в котором для каждого образа обучающей выборки $\mathbf{x}_i$ по очереди оценивается отклонение $L_i(\mathbf{W}, \mathbf{b}) = (\mathbf{x}_i - \mathbf{z}_i(\mathbf{x}_i|\mathbf{W}, \mathbf{b}))^2$, и далее параметры сети $\mathbf{W}$, $\mathbf{b}$ смещаются в направлении, обратном направлению градиента $\nabla L_i(\mathbf{W}, \mathbf{b})$.

Многослойные автоэнкодеры состоят из нескольких автоэнкодеров, где каждый последующий автоэнкодер принимает на вход в качестве образа вектор признаков, формируемый скрытым слоем (а не слоем реконструкции) предыдущего автоэнкодера, и выполняет новое нелинейное преобразование вектора признаков предыдущего уровня. Обучение такой сети происходит послойно: каждый последующий слой обучается после того, как закончено обучение предыдущего слоя. В случае применения многослойных автоэнкодеров в распознавании образов при обучении с учителем результаты реконструкции используются только в процессе обучения для настройки признаковых преобразований, а основным продуктом являются активности нейронов скрытого слоя автоэнкодера верхнего уровня, которые подаются в качестве входных векторов признаков на некоторый классификатор. В качестве такового часто применяется сеть прямого распространения, которая обучается после автоэнкоредов.

Здесь обычно используется логистическая регрессия (представимая сетью прямого распространения без скрытого слоя, т.е. проводящая линейную разделяющую поверхность во входном пространстве признаков). В логистической регрессии вычисляются вероятности принадлежности образа к различным классам на основе следующего уравнения:

$$p(y = c \mid \mathbf{x}, \mathbf{W}) = \frac{\exp(\mathbf{w}_c^T \mathbf{x} + b_c)}{\sum_{c'=1}^{C} \exp(\mathbf{w}_{c'}^T \mathbf{x} + b_{c'})},$$

где $\mathbf{x}$ – входной вектор признаков; c – индекс класса; C – общее число классов; $\mathbf{W}$ – весовая матрица, составленная из C векторов $\mathbf{w}_c$; b_c – пороги нейронов. Параметры этого классификатора также обычно настраиваются стохастическим градиентным спуском с целью минимизации отрицательного логарифма правдоподобия.

Одной из распространенных модификаций автоэнкодеров [15], которая была применена и нами, являются автоэнкодеры с подавлением шума, в которых результат реконструкции в процессе обучения вычисляется на дополнительно зашумленных образах, а ошибки реконструкции вычисляются относительно исходных образов без добавленного шума. Таким образом, автоэнкодеры обучаются восстанавливать незашумленный образ по зашумленному.

Описание разработанного метода

Для исследования возможности повышения вероятности распознавания сетями описанного типа путем отказа от сугубо локальной оптимизации весов связей сети был использован метод имитации отжига. Суть данного метода заключается в том, что при оптимизации некоторой функции $f(\mathbf{x})$, трактуемой как энергия системы $\mathbf{x}$, текущее решение $\mathbf{x}_i$ на i-й итерации заменяется новым решением $\mathbf{x}_{i+1}$ в соответствии с распределением вероятностей, например, задаваемым в форме

$$p(\mathbf{x}_{i+1} \mid \mathbf{x}_i) = \frac{1}{(2\pi T_i)^{D/2}} \exp\left(\frac{-|\mathbf{x}_{i+1} - \mathbf{x}_i|^2}{2T_i}\right) \frac{1}{1 + \exp(\Delta E / T_i)},$$

где $\Delta E = f(\mathbf{x}_{i+1}) - f(\mathbf{x}_i)$ – изменение энергии системы; T_i – «температура», управляющий параметр, задающий величину характерного изменения состояния системы за одну итерацию и уменьшающийся с номером итерации.

Оптимизация параметров сети методом имитации отжига может вестись по-разному, в зависимости от того, что именно принимать за функцию энергии f и состояние системы $\mathbf{x}$. В настоящей работе предложено следующее оригинальное применение метода имитации отжига. В качестве состояния системы $\mathbf{x}$ использованы матрицы весов и пороги нейронов всех слоев автоэнкодеров, но без слоя логистической регрессии. Начальное состояние $\mathbf{x}_0$ определялось с помощью стандартной, «жадной» оптимизации сети (послойное обучение стохастическим градиентным спуском). В качестве функции энергии f выступала вероятность правильного распознавания образов обучающей выборки. При ее вычислении для каждого рассматриваемого состояния системы $\mathbf{x}$ (значений параметров автоэнкодеров) проводилось дообучение слоя логистической регрессии. Иными словами, метод имитации отжига использовался для уточнения признаковых преобразований, задаваемых автоэнкодерами, с точки зрения их полезности для распознавания заданных образов.

Экспериментальная проверка

Экспериментальное исследование с целью проверки возможности глобальной оптимизации признаковых преобразований проводилось с использованием традиционной для тестирования методов машинного обучения базы изображений рукописных символов MNIST [16] (рисунок). Обучение реализованной нейронной сети (с разными размерами скрытых слоев в автоэнкодерах) сначала производилось на обучающей выборке изображений обычным способом, затем сеть обучалась с помощью модификации метода имитации отжига на той же выборке, и в итоге проводилось сравнение работы обученных сетей на одинаковой тестовой выборке.

Были использованы следующие параметры метода имитации отжига: начальная температура – 1,0 (в связи с тем, что активности нейронов находятся в диапазоне от 0 до 1), закон понижения температуры с номером итерации – линейный (поскольку использование логарифмического закона, характерного для больцмановского отжига, ведет к слишком медленной сходимости), число итераций – 200 (что обеспечивает достаточную точность настройки весов для признаковых преобразований). Автоматическая настройка параметров метода имитации отжига не требовалась.

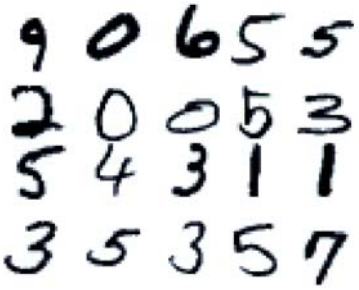

Рисунок. Примеры изображений рукописных символов из базы MNIST [16]

Экспериментальные данные позволяют провести сравнение работы автоэнкодеров с использованием имитации отжига и без него. В приведенной ниже таблице указана частота ошибок (отношение числа неверно классифицированных образов к общему числу образов в выборке), полученных при следующих параметрах: размер обучающей выборки – 40000 образов, размер тестовой выборки – 10000 образов. Число входных нейронов для первого автоэнкодера было равно числу пикселей на изображении и составляло $28 \times 28 = 784$ признака. Выходное число нейронов в слое логистической регрессии всегда было равно 10, т.е. числу классов (цифры 0–9). Число скрытых слоев и число нейронов в них варьировалось (что отражено в числах в левом столбце таблицы) для проверки повышения вероятности распознавания при разных конфигурациях сети.

Число нейронов в скрытых слоях	Исходный метод		Модифицированный метод	
	Частота ошибок (обучающая выборка)	Частота ошибок (тестовая выборка)	Частота ошибок (обучающая выборка)	Частота ошибок (тестовая выборка)
80, 70, 60, 50, 40	0,340	0,333	0,241	0,238
180, 170, 160, 150, 140	0,181	0,172	0,117	0,117
180, 140	0,102	0,092	0,078	0,082
140	0,082	0,080	0,070	0,076
400, 350, 300	0,079	0,075	0,045	0,057
600, 400	0,061	0,061	0,037	0,055
400, 300	0,067	0,066	0,042	0,052

Таблица. Частота ошибок, полученных до и после применения имитации отжига

Как видно из таблицы, частота ошибок на обучающей выборке с использованием модифицированного метода уменьшается, что свидетельствует о наличии нескольких экстремумов целевой функции, в связи с чем оправдано применение методов глобальной оптимизации.

Самым важным результатом является то, что после применения имитации отжига сокращается частота ошибок и на тестовой выборке, т.е. не возникает эффекта переобучения (чрезмерно близкой подгонки, в результате которой улучшение качества распознавания на обучающей выборке приводит к ухудшению качества распознавания на тестовой выборке). В связи с тем, что во всех семи случаях при произвольном выборе параметров сети получено уменьшение частоты ошибки, гипотеза о возможности уменьшения вероятности ошибки при использовании глобальной оптимизации многослойной сети может быть принята на уровне значимости α=0,01 по критерию Уилкоксона. Достигнутое снижение частоты ошибок важно для потенциального применения на практике и свидетельствует о возможности дальнейшего повышения вероятностей распознавания в сетях глубокого обучения за счет глобальной оптимизации параметров сети. При этом время обучения в модифицированном алгоритме при использовании имитации отжига возрастает в 3–3,5 раза (абсолютное значение времени обучения в зависимости от параметров сети варьировалось от нескольких секунд до нескольких суток на процессоре Core i5 3,33 ГГц, что, однако, не имеет существенного значения, поскольку при прикладном использовании обучение подобных сетей обычно проводится с использованием графических карт). Такое замедление обучения не является критичным, хотя в ряде случаев может быть нежелательным.

Заключение

В работе было проведено исследование возможности использования метаэвристического метода имитации отжига для обучения промежуточных слоев автоэнкодеров в сетях глубокого обучения в дополнение к традиционно используемой локальной послойной оптимизации градиентным спуском. Результаты показывают, что разработанный метод позволяет снизить частоту ошибок распознавания, оценивавшуюся по тестовой выборке на базе MNIST, в 1,1–1,5 раза. Таким образом, возможна глобальная оптимизация весов связей в сетях глубокого обучения без возникновения эффекта чрезмерно близкой подгонки (переобучения) при использовании вероятности распознавания образов обучающей выборки в качестве целевой функции. Данный результат указывает на направление возможного дальнейшего развития методов глубокого обучения, однако необходима разработка вычислительно эффективных методов обучения с использованием глобальной оптимизации. Результаты работы могут быть использованы для повышения вероятности распознавания образов в областях, требующих автоматического построения нелинейных признаковых преобразований, в том числе при распознавании изображений.

References

1. He Y., Kavukcuoglu K., Wang Y., Szlam A., Qi Y. *Unsupervised Feature Learning by Deep Sparse Coding*. 2013. Available at: http://arxiv.org/pdf/1312.5783v1 (accessed 03.07.2014).
2. Arnold L., Rebecchi S., Chevallier S., Paugam-Moisy H. An introduction to deep learning. *Proc. 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2011*. Bruges, Belgium, 2011, pp. 477–488.
3. Ciresan D.C., Meier U., Masci J., Schmidhuber J. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 2012, vol. 32, pp. 333–338. doi: 10.1016/j.neunet.2012.02.023
4. Mnih V., Kavukcuoglu K., Silver D., Graves A., Antonoglou I., Wierstra D., Riedmiller M. *Playing Atari with Deep Reinforcement Learning*. 2013. Available at: http://arxiv.org/pdf/1312.5602v1.pdf (accessed 03.07.2014).
5. Le Roux N., Bengio Y. Representational power of restricted boltzmann machines and deep belief networks. *Neural Computation*, 2008, vol. 20, no. 6, pp. 1631–1649. doi: 10.1162/neco.2008.04-07-510
6. Gregor K., Mnih A., Wierstra D., Blundell C., Wiersta D. *Deep Autoregressive Networks*. 2013. Available at: http://arxiv.org/pdf/1310.8499v2 (accessed 03.07.2014).
7. Tenenbaum J.B., Kemp C., Griffiths T.L., Goodman N.D. How to grow a mind: statistics, structure, and abstraction. *Science*, 2011, vol. 331, no. 6022, pp. 1279–1285. doi: 10.1126/science.1192788
8. Szegedy Ch., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. *Intriguing properties of neural networks*. 2014. Available at: http://arxiv.org/pdf/1312.6199v4 (accessed 03.07.2014).
9. Bengio Y., Lamblin P., Popovici D., Larochelle H. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, 2007, vol. 19, pp. 153–160.
10. Hinton G.E., Osindero S., Teh Y.-W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, vol. 18, no. 7, pp. 1527–1554. doi: 10.1162/neco.2006.18.7.1527
11. Ranzato M.A., Poultney Ch., Chopra S., LeCun Y. Efficient learning of sparse representations with an energy-based model. *Advances in Neural Information Processing Systems*, 2007, vol. 19, pp. 1137–1144.
12. Ciresan D.C., Meier U., Gambardella L.M., Schmidhuber J. *Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition*. 2010. Available at: http://arxiv.org/pdf/1003.0358 (accessed 03.07.2014).
13. Tsarev F.N. Sovmestnoe primenenie geneticheskogo programmirovaniya, konechnykh avtomatov i iskusstvennykh neironnykh setei dlya postroeniya sistemy upravleniya bespilotnym letatel'nym apparatom [Application of genetic programming, finite state machines and neural nets for construction of control system for unnmaned aircraft]. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2008, no. 8 (53), pp. 42–60.
14. Bondarenko I.B., Gatchin Yu.A., Geranichev V.N. Sintez optimal'nykh iskusstvennykh neironnykh setei s pomoshch'yu modifitsirovannogo geneticheskogo algoritma [Synthesis of optimal artificial neural networks by modified genetic algorithm]. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2012, no. 2 (78), pp. 51–55.
15. Vincent P., Larochelle H., Bengio Y., Manzagol P.-A. Extracting and composing robust features with denoising autoencoders. *Proc. 25th International Conference on Machine Learning*. Helsinki, Finland, 2008, pp. 1096–1103.
16. LeCun Y., Cortes C., Burges C.J.C. *The MNIST Database of handwritten digits*. Available at: http://yann.lecun.com/exdb/mnist (accessed 03.07.2014).

Потапов Алексей Сергеевич – доктор технических наук, доцент, профессор, Университет ИТМО, 197101, Санкт-Петербург, Россия, pas.aicv@gmail.com

Батищева Вита Вячеславовна – студент, СПбГУ, 199034, Санкт-Петербург, Россия, elokkuu@mail.ru

Пан Шуцао – студент, Цзилиньский университет, 130012, Цзилинь, КНР, pangshuchao1212@sina.com

Alexey S. Potapov – D.Sc., Associate professor, Professor, ITMO University, 197101, Saint Petersburg, Russia, pas.aicv@gmail.com

Vita V. Batishcheva – student, Saint Petersburg State University, 199034, Saint Petersburg, Russia, elokkuu@mail.ru

Shu-Chao Pang – student, Jilin University, 130012, Changchun, Jilin Province, P.R. China, pangshuchao1212@sina.com

Принято к печати 07.07.14
Accepted 07.07.14