СОВЕРШЕНСТВОВАНИЕ СТРУКТУРЫ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ НА БАЗЕ ВАРИАЦИОННОЙ ОПТИМИЗАЦИИ

IMPROVEMENT OF THE STRUCTURE OF ARTIFICIAL NEURAL NETWORKS ON THE BASIS OF VARIATIONAL OPTIMIZATION

С. В. Ткаченко, Т. В. Смирнова

S. Tkachenko, T. Smirnova

Белорусский государственный университет, МГЭИ им. А. Д. Сахарова БГУ,
г. Минск, Республика Беларусь
freddy.clarck@yandex.ru
Belarusian State University, ISEI BSU, Minsk, Republic of Belarus

При использовании искусственной нейронной сети часто стоит вопрос не только ее обучения или самоорганизации, но и оптимизации параметров, в частности, весов. В данной статье пойдёт речь про групповую лассо-регуляризацию.

Using an artificial neural network, the question often arises not only of its training or self-organization, but also of parameters optimization, in particular, weights. This article will discuss group *lasso*-regularization.

Ключевые слова: искусственная нейронная сеть, свёрточная нейронная сеть, распознавание, нейрон, слой, L1-регуляризация.

Keywords: artificial neural network, convolutional neural network, recognition, neuron layer, L1-regularization.

https://doi.org/10.46646/SAKH-2020-2-322-326

С развитием методов исследования и ростом вычислительных мощностей стало возможным решать многие теоретические и прикладные задачи, ранее считавшиеся трудновыполнимыми. В основе решения лежит компьютерное моделирование и вычислительный эксперимент. Но существуют задачи, которые компьютеру решить не под силу. Безуспешно пытаться требовать компьютер «рассказать» о восприятии любого изображения, звука и т.д. Для решения таких проблем можно с успехом применять методологию искусственных нейронных сетей (далее–ИНС). Нейронные сети успешно применяются для решения трудноформализуемых задач: распознавание рукописных текстов, изображений, прогнозирование осложнений при хирургических операциях, интерпретация рентгевских снимков, анализ экологической обстановки в регионе, анализ финансовых рынков и т.д.

Нейронные сети – мощный метод для решения задач моделирования сложных процессов, поскольку нейронные сети нелинейны по своей природе [1]. Нейронные сети позволяют также моделировать зависимости в случае большого числа разнотипных переменных.

ИНС – математическая модель, а также ее программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей – нервных клеток живого организма – головного мозга.

Для обучения нейронной сети необходимо большое количество входной информации, поскольку невозможно добиться высокой точности работы ИНС на малом количестве данных.

Так, к примеру, для анализа изображений в качестве обучающего набора данных используется информация с фото- и видео-хостинга. При решении задачи распознавания речи в качестве обучающей выборки используется серия аудиоклипов с приложенными к ним описаниями.

Первая представленная версия распознавания образов на основе ИНС содержала 38-процентный уровень ошибок классификации. Сегодня он составляет около 3%.

Что касается распознавания голоса на основе нейронной сети, то первые результаты содержали 27% ошибок распознавания. Сегодня процент ошибок составляет не более 8%, как сообщают новости группы разработчиков из facebook.

Современные нейросети имеют большое число слоев (иногда более сотни), что позволяет им находить решения для сложных данных. Одна из современных платформ глубокого обучения – сверточные нейронные сети, обладающие способностью самостоятельно выбирать в данных те признаки, которые наилучшим образом классифицируют объекты и используются в задачах распознавания. Перед свёрточными нейронными сетями ставятся такие задачи классификации, как идентификация объекта, распознавание лиц и частей тела человека, семантическая сегментация и определение границ, выделение объектов внимания на изображении, выделение нормалей к поверхности (рис. 1).

• Распознавание границ – это самая низкоуровневая задача, для которой уже классически применяются сверточные нейронные сети и которая позволяет определять границы объектов.

• Распознавание вектора к нормали позволяет реконструировать трёхмерное изображение из двухмерного.

• Распознавание объектов внимания – это то, на что обратил бы внимание человек при рассмотрении того или иного изображения.

• Семантическая сегментация позволяет разделить объекты на классы по их структуре, ничего не зная об этих объектах, то есть еще до их распознавания.

• Семантическое выделение границ – это выявление границ, разбитых на классы.

• Выявление частей тела – позволяет выделять идентифицированные части тела человека.

• Распознавание объектов – поиск необходимых объектов по заданным параметрам.

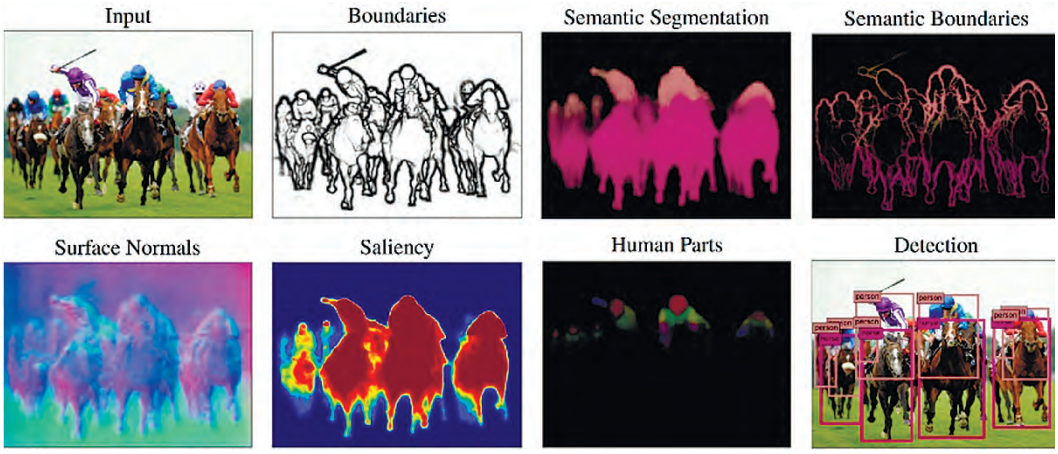

Рисунок 1 – Идентификация объектов при помощи свёрточной ИНС

Нейронная сеть проводит глубокую идентификацию объектов через поиск основных характеристик, а затем с помощью построения более абстрактных концепций, через группы свёрточных слоев. С учетом «обученности» и правильного строения ИНС она будет предсказывать результаты.

Задача обучения нейронной сети состоит в настройке весов нейронов для нахождения такого состояния, которое минимизирует целевую функцию на обучающей и тестирующей выборках. Этот процесс (минимизация функции) реализуется методом градиентного спуска по некоторой функции потерь $F(w)$, где переменными являются веса слоёв. Итоговая цель – нахождение глобального минимума среди множества локальных. При этом, обновление весов нейронной сети происходит с помощью метода вычисления градиента, который используется при обновлении весов многослойного перцептрона – методом обратного распространения ошибки.

Один из наиболее частых примеров построения нейронной сети – это классическая топология нейронной сети. Такая нейронная сеть может быть представлена в виде полносвязного графа, характерной ее чертой является прямое распространение информации и обратное распространение сигнализации об ошибке. Данная технология не обладает рекурсивными свойствами.

Основным недостатком этой топологии ИНС – избыточность. С учётом этого при поступлении данных, например, двухмерной матрицы на выходе получаем одномерный вектор. То есть вычислительные ресурсы, потраченные на выполнение поставленной задачи – слишком большие.

Значит, появляется такая задача как отбор из всего массива факторов лишь тех, которые «играют» важную роль, и удалить остальные факторы, то есть произвести поиск бесполезных весов. Фактически надо сделать фильтрацию, так как основная часть факторов будет равна нулю, а остальные будут с отличным от него значением. Таким образом, произойдёт сжатие сети. Для выяснения, какие веса будут учувствовать в работе ИНС требуется использование перебора (рис. 2). Таким образом происходит процедура обнуления. Это производится для того, чтобы улучшить обобщающую способность сети. Малозначимые веса могут вносить помехи в предсказание, либо способствуют «переобученности» модели, когда результаты тестирования сильно отличаются от результатов на этапе обучения. В этом смысле редукцию связей можно сравнить с методом отключения случайных нейронов во время тренировки сети. Кроме того, если в сети много нулей, она занимает меньше места и способна быстрее считаться на более слабых архитектурах, например одноплатных компьютерах [2]. Достичь разреженности позволяют L-регуляризации. Регуляризация – это набор методов, которые могут предотвратить переобучение нейронных сетей и, таким образом, повысить точность модели, представленной ИНС, при обращении к совершенно новым наборам данных из проблемной области.

Рис.3 демонстрирует результат подгонки набора данных полиномами 4 и 15 степени. Как видно из графика справа, подгоночная функция учитывает случайные выбросы, флуктуации, в результате основная линия тренда не уловлена, и поведение подгоночной функции сильно отличается от кривой, аппроксимирующей исходные данные.

Существует три эффективных метода регуляризации, называемых *L*1, *L*2 и дропаут. L1- регуляризация (лассо-регрессия) подавляет модуль веса (иначе, подгоночных коэффициентов); L2- регуляризация (*ridge*-регрессия) подавляет квадрат веса.

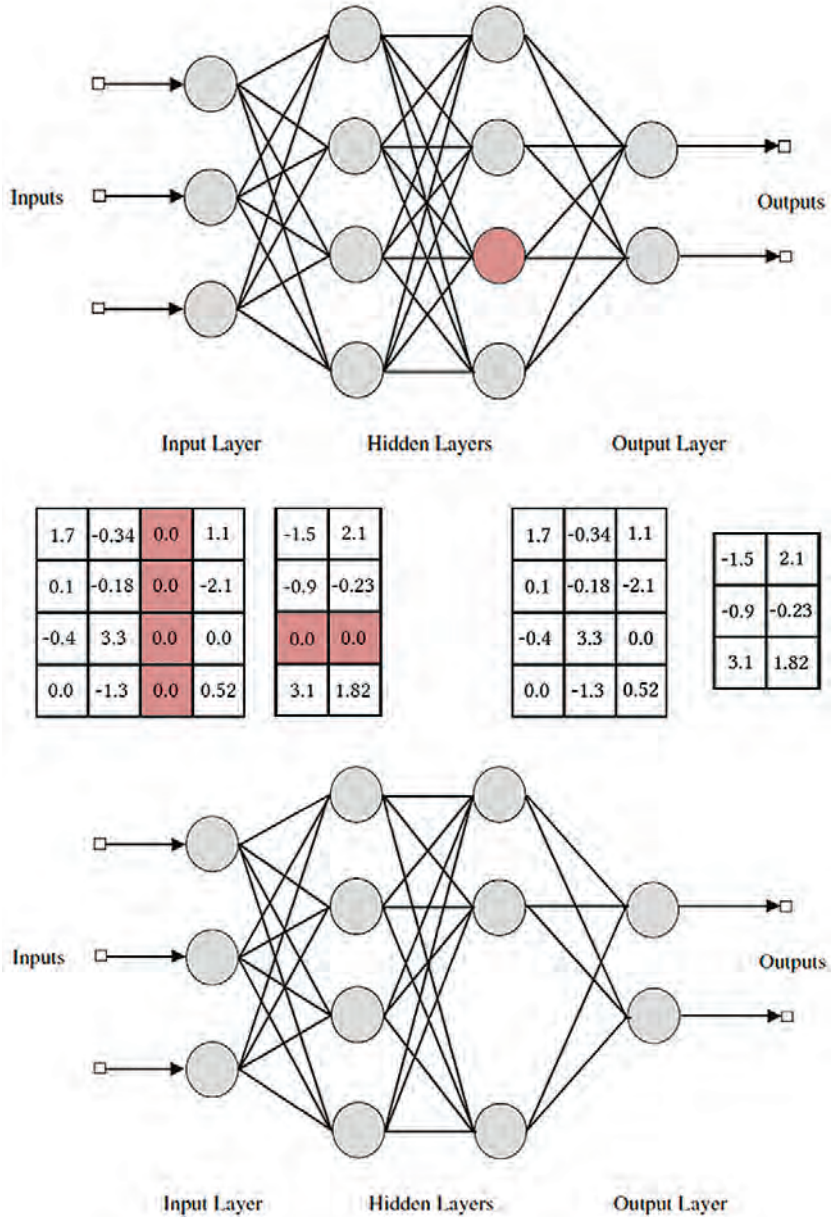

Рисунок 2 – Выявление «слабого» нейрона в ИНС

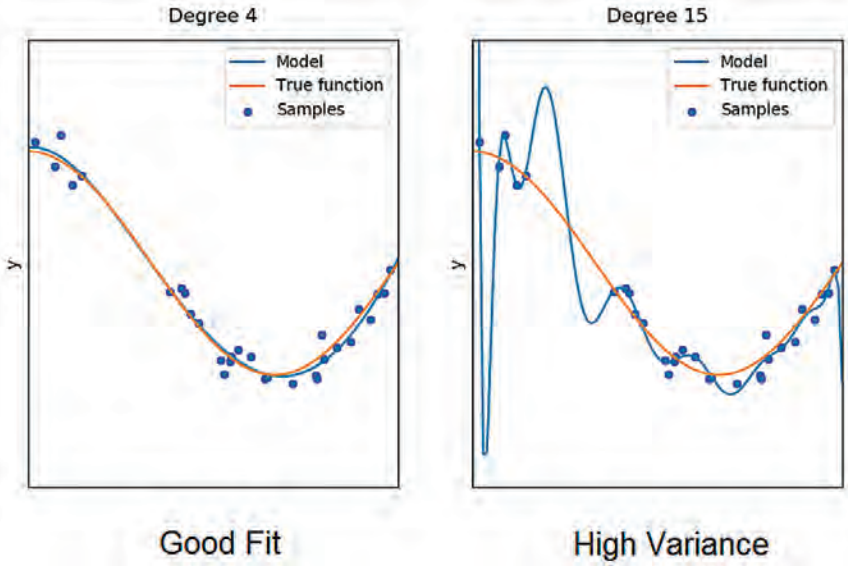

Рисунок 3 – Результаты работы различных подгоночных функций на одном наборе данных

Процедура дропаут-регуляризации предназначена для уменьшения переобучения в ИНС за счёт предотвращения сложных взаимоприспособлений отдельных нейронов на тренировочных данных во время обучения. В результате, приходим к более простой версии ИНС. Такое исключение нейронов с определенной вероятностью P применяется на каждом шаге обновления веса.

Эффективнее выпускать из рассмотрения не отдельные веса, а нейроны из полносвязных слоёв или каналы из свёрток целиком. В такой возможности эффект сжатия сети и увеличения скорости прогнозирования наблюдается более явно. Для бесперебойного удаления нейрона из ИНС, изначально можно заранее выполнить избавление его от полезных связей в ИНС. Это происходить за счёт возбуждения «сильных» нейронов, после чего они становиться сильнее, а «слабые» – слабее и теряют свой вес.

Самой простой и наилучший способ удаления невостребованных нейронов из сети с помощью метода групповой регуляции, применяемой в машинном обучении – лассо. Она является методом регрессивного анализа, что выполняет выбор переменных, так и регуляцию для повышения точности прогнозирования [3].

Например, рассмотрим специальный маскирующий слой с вектором весов $M=(\beta_1, \beta_2,\ldots, \beta_n)$. Его вывод – поэлементное произведение M на выводы предыдущего слоя, активационной функции у него нет. Поместим по маскирующему слою после каждого слоя, каналы в которых хотим отбрасывать, и применим L1-регуляризацию веса в этих слоях (лассо-регрессию, рис. 4).

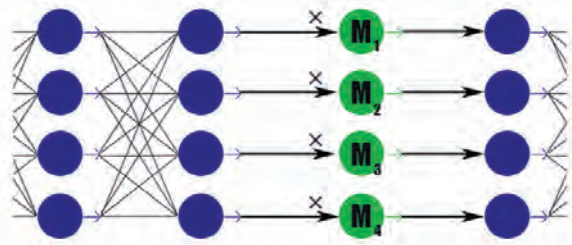

Рисунок 4 – Групповая L1-регуляризация

Значит, что вес маски β_i, растущий на *i*-тый выход слоя, неявно назначает ограничение на все веса сети, от которых зависит итоговый вывод. Если среди этих весов половина полезных, то β_i близко к единице, и этот вывод сможет хорошо передавать информацию. Но если маска β_i окажется близким нулю, это обнулит вывод нейрона и, заодно, все веса, от которых зависит этот вывод.

Отсюда следует:

$$\arg\min_{\beta,W} \frac{1}{2N} \left\| Y - \sum_{i=1}^{c} \beta_i X_i W_i^{\top} \right\|_F^2 + \lambda \|\beta\|_1$$

где λ – константа взвешивания лосса сети и лосса разреженности.

По завершению обучения ИНС следует отсортировать нейроны и их значения. Если β_i больше конкретного порога, то веса умножаются на β_i; если меньше порога, то из матриц удаляются соответствующие элементы. После этого производится дальнейшее обучение нейронной сети.

В применении групповой регуляризации есть несколько особенностей [4]:

1. Вместе с добавлением некоторых дополнительных ограничений на маскирующие веса следует применять их же ко всем весам сети. Таким образом, снижение маскирующих весов в случае ненасыщаемых активационных функций будет возмещено увеличением весов, и обнуляющего эффекта не выйдет.

2. Выполнять двухтактное обучение, то есть выполнять поочередное обучение обычных весов нейронной сети и маскирующих весов. Для получения лучшего результата требуется больше времени для обработки результата.

3. Создавать тщательную надстройку сети после фиксации маски.

4. Следить за нахождением маски: до или после функции активации. Возможно, проявятся проблемы с усилением, которые не стремятся к нулю при переменной равной нулю.

5. Помимо этого, существуют трудности с применением преобразований каналов, которые строятся на конструкциях известных по пирамидным клеткам в коре головного мозга (Residual neural network). После преобразования канала, происходит слияние нейронов из-за чего может не совпадать размерность. Данная проблема решается путем внедрения промежуточных слоёв, которые игнорируются. Кроме того, если ветви сети несут разное количество информации, имеет смысл установить для них разную константу взвешивания лосса.

6. Маски захвата – это, когда значение маски достигает некоторого заранее заданного низкого значения, и обнуляя его и запрещая менять эту часть маски. Таким образом, слабые веса полностью перестают вносить вклад в предсказание уже во время тренировки модели, а не вносят паразитные значения.

Для показа достоинств регуляризации разного вида был взят набор данных CIFAR-10, что представляет собой набор изображений, которые обычно используются для обучения алгоритмам машинного обучения и компьютерного зрения, и нейронная сеть с четырьмя свёрточными слоями и двумя полносвязными слоями [6]. Алгоритм удаления неинформативных каналов более качественно работает на более плотных сетях, так как может проявиться недостаток в виде нехватки вычислительных мощностей. Для оптимизации используется Adam со скоростью обучения в 0,0015 и размерностью 32. Так же используются регуляризации L1 (0.00005) и L2 (0.00025). Нейронная сеть прошла обучения 200 эпох, после чего происходит алгоритм редукции нейронов.

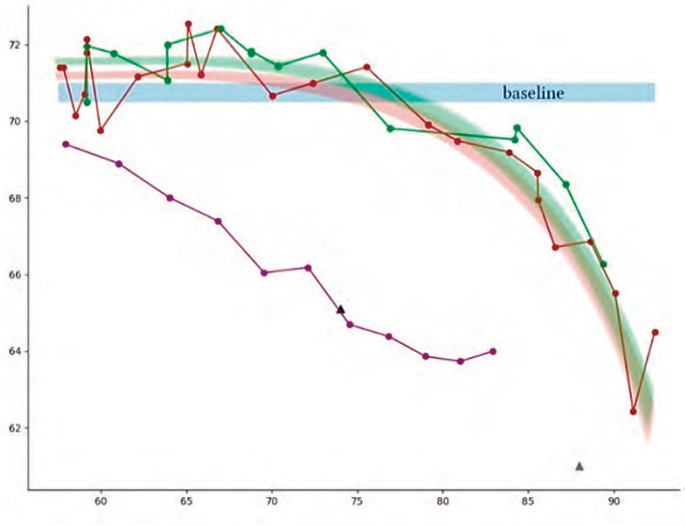

Рисунок 5 – Результат работы групповой L1-регуляризации

Результаты сравнения L1 и L0 алгоритмов редукции каналов после серии экспериментов с разными константами мощности регуляризации. По оси X отложено уменьшение количества весов в процентах после применения алгоритма. По оси Y – точность порезаной сети на валидационной выборке. Синяя полоса посередине – примерное качество сети, ещё не подвергнутой вырезанию нейронов. Зелёная линия представляет простой алгоритм L1-обучения масок. Красная линия – L0-pruning. Фиолетовая линия – удаление первых k каналов. Чёрные треугольники – обучение сети, у которой изначально было меньшее количество весов (рис. 5).

Получается, что простой L1-алгоритм справляется не хуже вариационной оптимизации, и потенциально даже чуть больше улучшает качество сети при малых значениях компрессии. Результаты также подтверждаются разовыми экспериментами с различными наборами данных и архитектурами ИНС.

ЛИТЕРАТУРА

1. *Круглов, В.В.* Искусственные нейронные сети: теория и практика // В.В. Круглов, В.В. Борисов. - М.: Горячая линия. – Телеком, 2001. – 382 с.

2. *Розенблат, Ф.* Принципы нейродинамики. Перцептроны и теория механизмов мозга. // М.: Мир, 1965. – с. 302.

3. *Duch*, *J*. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization / J. Duch, E. Hazan, Y. Singer // Journal of Machine Learning Research. – 2011. – P.2121-2159.

4. *Kingma*, *D.P*. Adam: A Method for Stochastic Optimization / D.P. Kingma, J.L. Ba. – 2014. – 15 p. – (Preprint / arXiv.org; № 1412.6980).

5. *LeCun*, *Y*. Efficient BackProp / Y. LeCun [et al.]. // Neural Netowriks: ticks of the trade. – UK Springer, 1998. – P.9-50.

6. Popular Datasets Over Time [электронный ресурс]. Режим доступа: www.kaggle.com. Дата доступа 2020-02-24.

ВНЕДРЕНИЕ ТЕХНОЛОГИЙ ЭЛЕКТРОННОГО ОБУЧЕНИЯ В ОБРАЗОВАТЕЛЬНЫЙ ПРОЦЕСС ВЫСШЕЙ ШКОЛЫ

INTRODUCTION OF ELECTRONIC LEARNING TECHNOLOGIES IN THE EDUCATIONAL PROCESS OF HIGHER SCHOOL

Б. А. Тонконогов, В. В. Журавков

B. Tonkonogov, V. Zhuravkov

Белорусский государственный университет, МГЭИ им. А. Д. Сахарова БГУ,
г. Минск, Республика Беларусь
boristonkonogov@iseu.by
Belarusian State University, ISEI BSU, Minsk, Republic of Belarus

Представлены особенности освоения и внедрения технологий электронного обучения и элементов цифровой инфраструктуры в образовательный процесс в учреждениях высшего образования. Показано, что для успешной и оптимальной реализации указанных мероприятий на начальном этапе необходимо провести мониторинг, классифицировать и определить назначение и эффективность использования имеющихся информационных систем, технологий и программно-аппаратных средств, а также сформировать план мероприятий