Simple method to improve quality of sparse models training from scratch

К. В. Беллонин[1] А. В. Шокуров[2] [3]

Существующие алгоритмы прореживания позволяют получать разреженные нейронные сети, обладающие хорошим качеством. Однако, при обучении с нуля полученной разреженной структуры, зачастую не получается достичь качества, полученного в результате прореживания (это особенно заметно для сильно разреженных архитектур).

В данной работе описывается метод, позволяющий улучшать качество при обучении с нуля, восстанавливая некоторые веса в разреженной архитектуре.

Ключевые слова: нейронные сети, прореживание, разреженные архитектуры, гипотеза "лотерейного билета"

1. Введение

Прореживание нейронных сетей - процесс вырезания (зануления) весов в данной нейронной сети. Прореживание, в зависимости от точки зрения, может рассматриваться как алгоритм оптимизации (уменьшения количества операций и памяти для хранения значений весов), регуляризации или поиска "архитектуры"нейронной сети [4].

Во многих алгоритмах прореживания [2] [1] в результате работы алгоритма получается не только разреженная архитектура, но и значения параметров для этой архитектуры. Эти параметры имеют критически

[1] *Беллонин Кирилл Владимирович* — аспирант кафедры теоретической информатики механико-математического факультета МГУ им. М.В.Ломоносова, e-mail: bellonin_kirill@mail.ru.

Bellonin Kirill Vladimirovich — postgraduate student of Theoretical Informatics Chair of Mechanics and Mathematics Faculty of Lomonosov Moscow State University.

[2] *Шокуров Антон Вячеславович* — к.ф.-м.н., научный сотрудник лаборатории вычислительных методов механико-математического факультета МГУ им. М.В.Ломоносова, e-mail: shokurov.anton.v@yandex.ru.

Shokurov Anton Vyacheslavovich — Ph.D., research associate of Computational Methods Laboratory of Mechanics and Mathematics Faculty of Lomonosov Moscow State University.

[3] Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект», по научному плану кафедры теоретической информатики.

This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University «Brain, Cognitive Systems, Artificial Intelligence», according to the scientific plan of Theoretical Informatics Chair.

важное значение, т.к. при тренировке разреженной архитектуры с нуля (т.е. начиная с некоторой произвольной инициализации) зачастую не получается достигнуть качества, полученного во время прореживания.

Эта проблема уже была исследована в некоторых статьях [3] [4] [5], однако даже существование самой проблемы иногда ставилось под вопрос [4]. В этих же статьях проводилось исследование влияния инициализации разреженных архитектур на качество после обучения [3] [5], однако и здесь были получены противоречивые результаты.

2. Основные результаты и описание метода

Для экспериментов был выбран набор данных CIFAR-10 [7] и нейронная сеть типа VGG [6] с 14.7 миллионами параметров (здесь и далее считаются только параметры ядер двумерных сверток и весов полносвязного слоя, т.е. параметры участвующие в операциях умножения).

Для получения разреженных архитектур были реализованы [1] два алгоритма: магнитудного прореживания [2] и вариационного дропаута [1]. Из всего набора экспериментов была выбрана архитектура полученная с помощью вариационного дропаута: в ней всего лишь 6.7 тысяч параметров при топ-1 качестве (top-1 accuracy) в 85% на тестовой части CIFAR-10. Все дальнейшие эксперименты проводились с этой разреженной архитектурой.

Тренировка разреженной архитектуры подтвердила ее неспособность обучаться со случайной инициализации до качества, полученного в процессе прореживания: при тренировке со случайной инициализации топ-1 качество достигло лишь 75%. Таким образом, можно утверждать, что на каком-то этапе в результате прореживания получаются разреженные архитектуры, которые даже с инициализацией из гипотезы "о лотерейном билете"не могут достичь того качества, которое они достигли в результате прореживания.

Приводимый в данной статье метод, восстанавливающий качество обучения с нуля на одном наборе данных, может позволить создавать архитектуры, способные обучаться с нуля и на других наборах данных. Таким образом, метод имеет практическую значимость при поиске разреженных архитектур.

2.1. Метод восстановления качества

Пусть L - функция потерь, с помощью которой тренируется наша нейронная сеть, P_0 - фиксированная случайная инициализация, c_{l_2} - коэффициент l_2 - регуляризации, $\mathbb{M}$ - множество вырезанных из модели весов.

[1] https://github.com/SleepingThread/torch_scripts

Тогда метод восстановления качества выглядит следующим образом: 2.1. Процесс выбора множества весов $\Delta_{\mathbb{M}}$:

$$s(w) = \frac{1}{1+\alpha}\Big(\Big|\frac{w_{min}+w_{max}}{2}\Big| + \alpha(w_{max}-w_{min})\Big), \text{где } \alpha \geq 0$$
$$\Delta_{\mathbb{M}}(K, N_{max}) = \underset{\Delta' \subset \mathbb{M}_{new}, |\Delta'| = N_{max}}{\operatorname{argmax}} \sum_{w \in R'} s(w) \tag{1}$$

, где N_{max} - максимальное количество параметров для восстановления за шаг алгоритма, w_{min}, w_{max} - минимальное и максимальное значения параметра w за последние K эпох, а α регулирует важность "шума"веса относительно его "среднего"значения в фукнции важности веса $s(w)$.

Algorithm 2.1 Восстановление качества тренировки

procedure RESTOREQUALITY($L, P_0, c_{l_2}, \mathbb{M}_{old}, K, N_{max}$)
    Восстановить все вырезанные веса в модели
    $\mathbb{M}_{new} = \mathbb{M}_{old}$

$$L_{restore} = L + c_{l_2} \sum_{w \in \mathbb{M}_{new}} w^2$$

    Инициализировать модель значениями P_0
    Тренировать модель с функцией потерь: $L_{restore}$
    while Текущее качество меньше требуемого **do**
        Выбрать $\Delta_{\mathbb{M}}(K, N_{max}) \subset \mathbb{M}_{new}$ согласно 1
        $\mathbb{M}_{new} = \mathbb{M}_{new} \setminus \Delta_{\mathbb{M}}$ ▷ восстановление части параметров
        Тренировать модель с функцией потерь: $L_{restore}$
    end while
    return $\mathbb{M}_{new}$ ▷ новые вырезанные веса, $\mathbb{M}_{new} \subset \mathbb{M}_{old}$
end procedure

2.2. Результаты применения метода

Алгоритм тестировался для небольшого количества итераций (т.е. 2 восстановления параметров), со следующими параметрами: $K = 10, N_{max} = 3000, c_{l_2} = 0.01$. Каждая тренировка внутри алгоритма проходила 100 эпох с использованием оптимизатора Adam со скоростью обучения 2e−3, уменьшающейся по косинусу (cosine lr scheduler [8]). Тренировка с нуля проводилась аналогично, но количество эпох было увеличено до 200. Бинарный дропаут в данных экспериментах выключен, т.к. модели крайне малы и не нуждаются в дополнительной регуляризации.

В результате работы алгоритма были получены архитектуры (9.7k, 12.7k), которые при обучении из инициализации P_0 [2] дали следующие

результаты (в таблице указано топ-1 качество на тестовой части CIFAR-10, изначальная архитектура обучалась до 75.4%):

α	9.7k (%)	12.7k (%)
0.3	79.5	80.9
0.5	79.3	81
1.0	79	81.2
2.0	79.2	81.1

Видно, что в проведенных экспериментах параметр α не сильно влияет на конечный результат, однако отличия могут проявиться при выборе другого значения для c_{l_2} или другой архитектуры и данных.

Отдельно следует отметить, что случайное восстановление весов не приводит к сопоставимому улучшению качества, поэтому алгоритм, как минимум, лучше случайного:

	9.7k (%)	12.7k (%)	15.7k (%)
random	76.5	76.8	77.0

Список литературы

[1] Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov, "Variational Dropout Sparsifies Deep Neural Networks", *Proceedings of the 34th International Conference on Machine Learning*, **70** (2017), 2498–2507.

[2] Song Han, Jeff Pool, Johj Tran, William J. Dally, "Learning both Weights and Connections for Efficient Neural Networks", *Advances in Neural Information Processing Systems*, **28** (2015).

[3] Jonathan Frankle, Michael Carbin, "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks", *International Conference on Learning Representations*, 2019.

[4] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, Trevor Darrell, "Rethinking the Value of Network Pruning", *International Conference on Learning Representations*, 2019.

[5] Trevor Gale, Erich Elsen, Sara Hooker, "The State of Sparsity in Deep Neural Networks", *ArXiv*, **abs/1902.09574** (2019).

[6] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition", *International Conference on Learning Representations*, 2015.

[7] Alex Krizhevsky, "Learning Multiple Layers of Features from Tiny Images", 2009.

[8] Ilya Loshchilov, Frank Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", *International Conference on Learning Representations*, 2017.

[2] Для исключения влияния инициализации во всех экспериментах (в т.ч. при получении изначальной архитектуры с помощью вариационного дропаута) использовалась одна единственная фиксированная инициализация.

Bellonin K.V., Shokurov A.V.
Простой метод улучшения качества тренировки разреженных моделей с нуля

Existing pruning algorithms can achieve good quality on sparse neural networks. But the recieved sparse architectures, when training from scratch, often can't achive the same quality as pruning (especially for very sparse networks).

In this work the weights restoring method to improve training from scratch quality is described.

Keywords: neural networks, pruning, sparse architectures, "the lottery ticket"hypothesis

References

[1] Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov, "Variational Dropout Sparsifies Deep Neural Networks", *Proceedings of the 34th International Conference on Machine Learning*, **70** (2017), 2498–2507.

[2] Song Han, Jeff Pool, Johj Tran, William J. Dally, "Learning both Weights and Connections for Efficient Neural Networks", *Advances in Neural Information Processing Systems*, **28** (2015).

[3] Jonathan Frankle, Michael Carbin, "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks", *International Conference on Learning Representations*, 2019.

[4] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, Trevor Darrell, "Rethinking the Value of Network Pruning", *International Conference on Learning Representations*, 2019.

[5] Trevor Gale, Erich Elsen, Sara Hooker, "The State of Sparsity in Deep Neural Networks", *ArXiv*, **abs/1902.09574** (2019).

[6] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition", *International Conference on Learning Representations*, 2015.

[7] Alex Krizhevsky, "Learning Multiple Layers of Features from Tiny Images", 2009.

[8] Ilya Loshchilov, Frank Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", *International Conference on Learning Representations*, 2017.