СРАВНИТЕЛЬНЫЙ ОБЗОР МЕТОДОВ ОПТИМИЗАЦИИ ГИПЕРПАРАМЕТРОВ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ ДАННЫХ ИЗ ВЫБОРКИ

Маслов Н.С., оператор 4 научной роты ФГАУ «Военный инновационный технополис «ЭРА»,
Ваньков А.П., оператор 4 научной роты ФГАУ «Военный инновационный технополис «ЭРА».

Аннотация
Свёрточные нейронные сети часто применяются для решения задач машинного обучения, связанных с обработкой изображений, аудио- и видеоданных. Одна из проблем, связанных с подготовкой модели к применению в приложениях, связана с избыточностями в готовых архитектурах, устранение которых позволяет уменьшить вычислительную сложность сети при сохранении точности. В данной статье приводится обзор методов, позволяющих уменьшить вычислительную сложность нейронной сети за счёт удаления части функциональных элементов.

Ключевые слова: машинное обучение, свёрточные нейронные сети, обучение нейронных сетей, optimal brain damage.

В последнее время нейронные сети как модель машинного обучения обрели широкую популярность, что обосновывается хорошей способностью к обобщению, а также возможностью построения нетривиальных архитектур, хорошо решающих отдельные классы задач. Одним из таких классов является обработка изображений и аудиоданных. Для решения подобных задач используются свёрточные нейронные сети, основанные на принципе работы зрительной коры головного мозга [1].

Одной из проблем, связанных с подготовкой программных решений на основе свёрточных нейронных сетей, является выбор значений гиперпараметров сети, таких как количество слоёв и функциональных элементов в отдельных слоях, при которых она будет способна решать поставленную задачу с достаточной точностью, при этом требуя наименьшее количество вычислительных ресурсов. Уменьшение «размера» сети позволяет одновременно оптимизировать вычислительную сложность и уменьшить склонность сети к переобучению.

Одним из базовых подходов к уменьшению размера нейронных сетей является Optimal Brain Damage [2], предложенный в 1990 году. На его основе разработаны различные методы обучения и синтеза архитектур, которые успешно применяются при подготовке нейронных сетей для применения в системах, требовательных к скорости работы или ограниченных в вычислительных ресурсах.

В этой статье приводится обзор актуальных методов, позволяющих оптимизировать гиперпараметры свёрточных нейронных сетей на разных этапах её жизненного цикла.

Большинство представленных на сегодняшний день методов основаны на общем подходе, в котором оптимизация гиперпараметров производится за счёт удаления отдельных функциональных элементов свёрточной нейронной сети, например, свёрточных фильтров, нейронов в полносвязных слоях или за счёт обнуления отдельных весовых коэффициентов.

Для выбора удаляемых элементов используются различные критерии, которые могут использовать как информацию о внутреннем устройстве сети, так и свойства данных, используемых для обучения и валидации. Отдельно стоит отметить, что на возможность использования отдельного метода в конкретном приложении влияет также вычислительная сложность критерия, поскольку в некоторых случаях время, требуемое на обработку, может оказаться неоправданно большим.

Таким образом, для сравнения методов были выбраны следующие критерии:

- поддержка работы со свёрточными слоями сети (1);
- поддержка работы с полносвязными слоями сети (2);
- возможность выбора удаляемых элементов среди нескольких слоёв одновременно (3);
- типы удаляемых функциональных элементов сети (4);
- вычислительная сложность вычисления критерия удаления, меньше – лучше (5).

Анализируемые методы были заранее разделены на по классу используемых критериев на те, которые требуют данные из выборки (data-driven) и не требуют (data-free) критерии. Эти две группы существенно различаются в вычислительной сложности, data-free критерии часто гораздо менее требовательны к ресурсам, при этом data-driven критерии располагают дополнительными данными и способны более точно выбирать удаляемые элементы.

Стоит отметить важную особенность, связанную с удалением отдельных весов. В результате работы таких методов в слоях сети формируются разреженные матрицы весов. Работа с разреженными матрицами связана с ограничениями, предъявляемыми программным инструментарием для обучения и применения нейронных сетей, а также с возможностями аппаратуры. Особенно эта проблема актуальна для свёрточных нейронных сетей [3].

В некоторых случаях работа с разреженной матрицей не даёт прироста в производительности сети, хотя использование разреженных моделей в отдельных случаях может увеличить обобщающую способность сети.

В данной статье методы уменьшения нейронных сетей рассматриваются с точки зрения сокращения использования вычислительных ресурсов на оборудовании общего назначения.

Далее для оценки вычислительной сложности используются следующие специальные обозначения:

- FP – количество арифметических операций, требуемых для однократного вычисления прямого распространения сигнала в сети;
- BP – то же самое для обратного распространения;
- St – количество объектов в обучающей выборке;
- Sv – количество объектов в валидационной выборке;
- m – число элементов, размеченных для удаления в сети.

Уменьшение размера нейронной сети требуется провести таким образом, чтобы как можно меньше повлиять на точность её работы. В некоторых случаях для этого оправдано использование данных из тренировочной или тестовой выборки. Методы, которые используют данные из выборки для выбора удаляемых элементов, называют основанными на данных, или data-driven методами.

Алгоритм работы большинства таких методов выглядит следующим образом:

1. Производится предварительное обучение сети до достижения ей требуемого качества работы.
2. В сети выбираются элементы с наименьшим влиянием на качество. Эти элементы исключаются из сети.
3. Производится восстановление сети - дообучение с помощью нескольких эпох работы алгоритма обратного распространения ошибки.

4. Повторить алгоритм с шага 2, пока не будет удалено требуемое количество элементов или пока точность полученной сети не упадёт ниже требуемой.

Таким образом, методы основаны на жадном подходе, который позволяет получать не обязательно оптимальные, но достаточно хорошие решения. Основное различие отдельных методов заключается в критерии выбора элементов сети для удаления.

Наиболее простой с точки зрения понимания метод выбора удаляемых элементов – прямая проверка качества сети после удаления.

Выбор происходит следующим образом. Из модели последовательно удаляется по одному элементу. После удаления каждого элемента проводится проверка точности модели на тестовой выборке и элемент возвращается обратно в сеть. После проверки всех элементов выбирается один с наименьшим влиянием на точность сети.

Достоинство этого метода – высокая достоверность данных о критичности элементов сети, возможность работы с элементами сети любого типа (нейронами, фильтрами, отдельными весами) и отсутствие привязки к отдельному слою, то есть элементы выбираются из сети в целом.

У этого метода есть очевидная проблема, связанная с вычислительной сложностью, поскольку каждое удаление требует проведения вычисления прямого распространения k x l раз, где k – количество оставшихся элементов сети, l – число объектов в тестовой выборке.

Вычислительную сложность метода можно оценить как O(FP x St x m(m-1)/2).

Можно предложить упрощённый вариант метода, в котором проверка будет проводиться не на всём множестве элементов, а для его случайного подмножества. Исследование качества и вычислительной сложности данного метода требует проведения дополнительного исследования.

В исходном виде данный метод редко используется на практике, но может быть использован для оценки эффективности других методов в качестве эталона.

В 1990 году в работе [1] был предложен метод выбора весов для удаления под названием Optimal Brain Damage (OBD, оптимальное повреждение мозга). Поиск связей ведётся с помощью вычисления производной функции потерь по отдельным весовым коэффициентам.

В отличие от прямого поиска, метод OBD не требует вычисления выходного значения сети целиком и по сложности схож с вычислением градиента для метода обратного распространения ошибки. Однако, эти вычисления требуется провести для каждого элемента обучающей выборки, что может повлечь значительный рост вычислительной сложности для некоторых задач.

Оригинальный метод поддерживает работу только с весами нейронов в полносвязных сетях, хотя есть работы, посвящённые адаптации метода к работе со свёрточными слоями. Метод не привязывается к отдельному слою.

Вычислительная сложность метода оценивается как O(m x St x BP).

Метод, основанный на поиске по доле нулевых активаций (Average Percentage of Zeros, APoZ) [4] разработан для поиска избыточностей только в свёрточных слоях. Этот метод удаляет не отдельные веса, а фильтры целиком, что позволяет избежать работы с разреженными матрицами.

Для каждого фильтра в слое считается средняя доля нулевых элементов карты признаков на всей тренировочной выборке, после чего из слоя удаляется фильтр с максимальным значением этого критерия.

Достоинства этого метода заключаются в его интуитивной простоте и небольшой вычислительной сложности относительно других data-driven методов. Недостаток этого метода заключается в строгой ориентированности на работу со свёрточными слоями.

Метод поддерживает работу только со свёрточными фильтрами, при этом здесь появляется зависимость от слоя, из которого требуется удалить элемент.

Вычислительная сложность метода оценивается как O(m x St x FP).

По схожему принципу работает метод выбора фильтров на основе энтропии [5]. Отличием этого метода от предыдущего является использование энтропии вместо доли нулевых активаций для определения степени важности удаляемого фильтра.

Меньшие значения энтропии соответствуют фильтрам с более слабой активацией. Таким образом, можно выбирать фильтры для удаления по наименьшим значениям энтропии.

Как и APoZ, метод поддерживает работу только со свёрточными фильтрами. Сравнение значений критерия для

элементов из разных слоёв, в отличие от прочих методов, здесь допустимо.

Вычислительную сложность метода можно оценить как O(m x St x FP).

Большинство методов уменьшения размеров нейронных сетей основаны на общем принципе, подразумевающем предварительное обучение сети и дообучение для восстановления качества работы после удаления элементов. Основное различие между методами заключается в применении различных критериев, различных по вычислительной сложности. Также некоторые методы ограничены для применения к отдельным типам слоёв.

В современных работах представлены методы, использующие интуитивно понятные критерии выбора элементов, что позволяет выбирать конкретный метод или их сочетание в зависимости от конкретной решаемой задачи, учитывать дополнительную информацию о сети, известную разработчику.

Краткий результат обзора представлен в таблице 1.

Таблица 1. Сравнение data-driven методов удаления элементов свёрточных нейронных сетей.

Метод	Критерий				
	(1)	(2)	(3)	(4)	(5)
Прямая проверка	+	+	+	Нейроны, фильтры	*O(FP x St x m(m-1)/2)*
OBD	+	-	+	Отдельные веса	*O(m x St x BP)*
APoZ	-	+	+/-	Фильтры	*O(m x St x FP)*
Энтропия	-	+/-	+/-	Фильтры	*O(m x St x FP)*

Литература

1. Hubel D. H., Wiesel T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex //The Journal of physiology. – 1962. – Т. 160. – №. 1. – С. 106-154.

2. LeCun Y., Denker J. S., Solla S. A. Optimal brain damage //Advances in neural information processing systems. – 1990. – С. 598-605.

3. Lebedev V., Lempitsky V. Fast convnets using group-wise brain damage //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – С. 2554-2564.

4. Hu H. et al. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures //arXiv preprint arXiv:1607.03250. – 2016.

5. Luo J. H., Wu J. An entropy-based pruning method for cnn compression //arXiv preprint arXiv:1706.05791. – 2017.