Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

Садртдинов И.А. Исследования методов сжатия нейронных сетей. метод кодирования с преобразованием и кластеризация// Академия педагогических идей «Новация». Серия: Студенческий научный вестник. – 2018. – №6 (июнь). – АРТ 389-эл. – 0,4 п.л. - URL: http: //akademnova.ru/page/875550

РУБРИКА: ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.032.26

Садртдинов Илья Айдарович
магистрант 2 курса, факультет техника и технологии
Научный руководитель: Маков С.В., к.т.н., доцент
ФГБОУ ВПО «Институт сферы обслуживания и предпринимательства»
г. Шахты, Российская Федерация
e-mail: mail@sssu.ru

ИССЛЕДОВАНИЯ МЕТОДОВ СЖАТИЯ НЕЙРОННЫХ СЕТЕЙ. МЕТОД КОДИРОВАНИЯ С ПРЕОБРАЗОВАНИЕМ И КЛАСТЕРИЗАЦИЯ

Аннотация: С размещением нейронных сетей на мобильных устройствах и необходимостью передачи нейронных сетей по ограниченным или дорогостоящим каналам, размер файла обученной модели является большим недостатком. В этой статье предлагается использовать кодек для сжатия нейронных сетей, основанный на кодировании с преобразованием для сверточных и плотных слоев и кластеризации для смещений и нормализаций. Используя этот кодек, удалось достигнуть средних коэффициентов сжатия между 7,9 – 9,3, при потере точности сжатых нейросетей для классификации изображений, на 1% – 2% соответственно.

Ключевые слова: сверточные нейронные сети, глубокое обучение, метод сжатия.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

Sadrtdinov Ilya Aydarovich
2nd year master, Faculty of Engineering and Technology
Supervisor: S.V. Makov, Ph.D., Associate Professor
FGBOU HPE "Institute of Service and Entrepreneurship"
Shakhty, Russian Federation

RESEARCH OF METHODS OF COMPRESSION OF NEURAL NETWORKS. METHOD OF TRANSFORM CODING AND CLUSTERIZATION

Abstract: With the placement of neural networks on mobile devices and the need to transfer neural networks over limited or expensive channels, the size of the file of the trained model is a major drawback. In this article, it is proposed to use a codec for the compression of neural networks, based on encoding with transformation for convolutional and dense layers and clustering for displacements and normalizations. Using this codec, it was possible to achieve average compression ratios between 7.9 - 9.3, with the loss of the accuracy of compressed neural networks for the classification of images, by 1% - 2% respectively.

Keywords: neural networks, deep learning, compression method.

1. Введение

Победа нейронной сети AlexNet [1] в ImageNet Big Scale Visual Recognition Challenge (ILSVRC) 2012, рассматривается многими как прорыв для современных технологий глубокого обучения. С этого времени глубокие нейронные сети применяются во многих научных и промышленных приложениях [2] не только для классификации изображений (например, [3], [4]), но и для распознавания речи (например, Siri или Google Assistant), а также в дополненной реальности, и т.д. Часто

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

необходимость больших объемов данных обучения, длительная продолжительность обучения и вычислительная сложность операции вывода являются слабыми местами в канале глубокого обучения.

Объем памяти, занимаемый сохраненными нейронными сетями, стремительно увеличивается, а нейронные сети выполняются не только на крупномасштабных серверах или в облаке, но и на мобильных устройствах (например, мобильных телефонах, планшетах) или на встроенных устройствах (например, в автомобильных приложениях). В этих случаях возможности хранения ограничены и/или нейронные сети должны быть переданы устройствам по ограниченным каналам передачи данных (например, обновлением приложения). Но, чтобы справляться с все более и более сложными задачами, нейросети нужно расти. Поэтому, необходимо использовать эффективное сжатие нейронных сетей. Методы сжатия общего назначения, такие как Deflate (комбинация Lempel-Ziv-Storer-Szymanski с кодированием Хаффмана), плохо работают на нейронных сетях, поскольку сети состоят из нескольких, немного отличающихся, весов с плавающей запятой. Требованиями к сжатию нейронных сетей являются: высокая эффективность кодирования, незначительное воздействие на желаемый выход нейронной сети (например, точность), разумная сложность, особенно на декодере (мобильном устройстве), применимость к существующим моделям нейронных сетей, т. е. без переобучения.

2. Обзор существующих методов

В ходе анализа литературы было выявлено несколько похожих работ. Эти работы в основном основаны на таких методах, как квантование и отсечение. Фреймворк tensorflow обеспечивает метод квантования для преобразования обученных весов с плавающей запятой в 8 бит с фиксированной точкой. Подобно этому подходу, в предлагаемом методе

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

применяется квантование. В разделе 4 будет показано, что полученные результаты были достигнуты благодаря квантованию.

Han и др. предложил фреймворк «Deep Compression» (глубокое сжатие) для эффективного сжатия нейронных сетей [5]. В дополнение к квантованию, их метод основан на циклическом этапе отсечения и переобучения. Весы и соединения в сети, которые вносят наименьший вклад в вывод сети, отсекаются. Это само по себе приводит к значительному снижению точности сети. Таким образом, сеть должна быть повторно обучена для снижения падения точности. В отличие от Deep Compression, предлагаемый метод сжимает существующие модели нейросети без необходимости переобучения и без изменения архитектуры сети. Таким образом, эти подходы решают разные проблемы. Iandola et al. предложить новую сетевую архитектуру SqueezeNet, которая, нацелена на то, чтобы иметь как можно меньше весов в сети [6]. Таким образом, их метод также не способен сжимать существующие сети. В разделе 4 будет показано, что предлагаемый метод может уменьшить размер этой, уже оптимизированной сети SqueezeNet в размере до 7,4.

3. Метод сжатия

В этом разделе описывается конструкция кодека (рисунке 1). Обученная модель нейронной сети - это входной сигнал кодека. Он состоит из одномерных и двумерных весов, смещений, нормировок и самой архитектуры (количество слоев / фильтров, соединений и т. д.). Все слои сети обрабатываются индивидуально. Это упрощает частичное, ретроактивное обновление отдельных слоев без повторной передачи всей сети. В зависимости от доступных вычислительных ресурсов можно параллельно обрабатывать произвольное количество слоев. Веса, смещения и нормализации предварительно масштабируются для использования

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

полного динамического диапазона представления числа. Стоит отметить, что вместо нормализации мы используем термин «предварительное масштабирование», чтобы избежать путаницы с термином «нормализация», который уже используется для типа параметров сети.

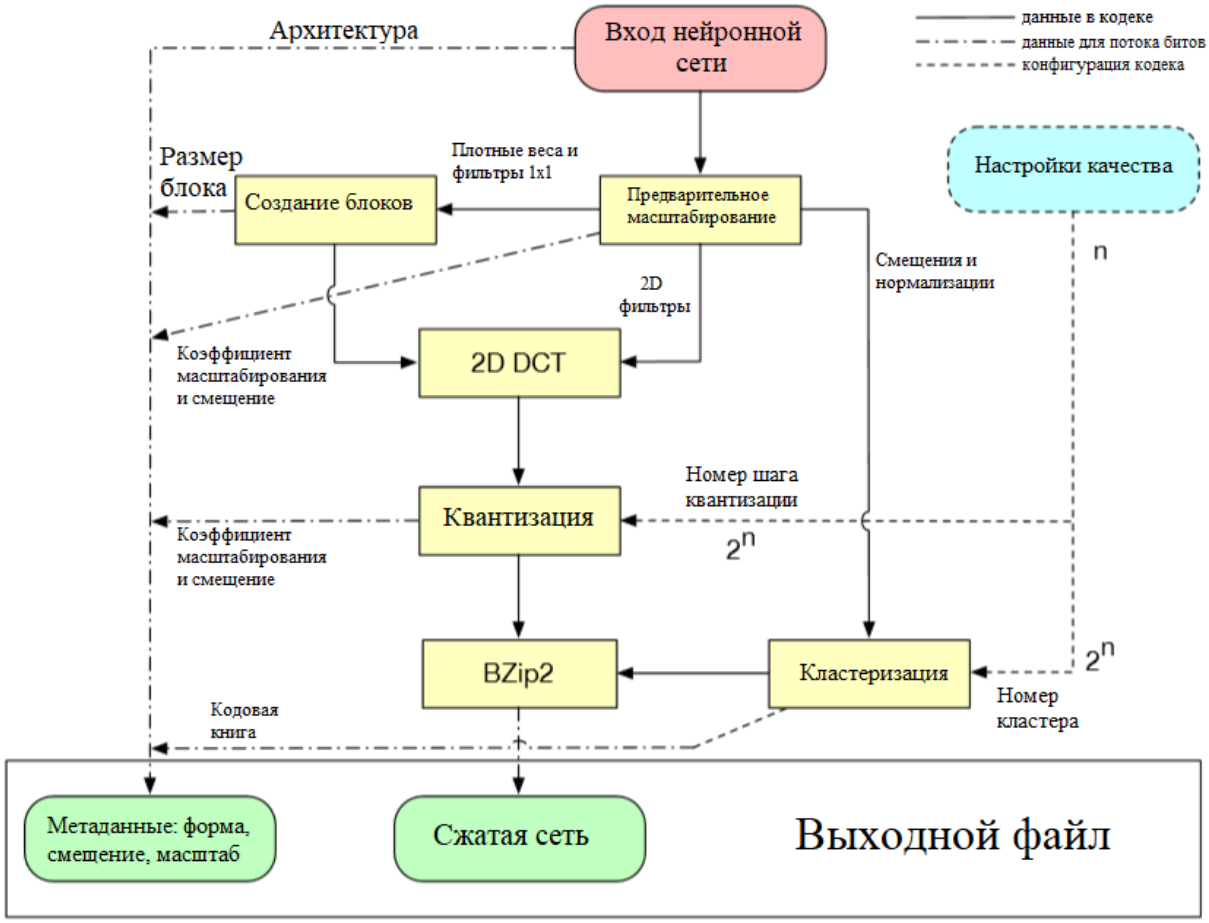


Рисунок 1 – конструкция кодека.

Некоторые иллюстративные фильтры из GoogLeNet представлены на рисунке 2. Можно заметить, что фильтры содержат структурную информацию, не полностью отличную от блоков в естественных изображениях. Основываясь на этом наблюдении, база кодировщика для сверточных фильтров состоит из двумерного дискретного косинусного

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

преобразования (2D DCT), за которым следует шаг квантования. Эта комбинация часто упоминается как кодирования с преобразованием.

Для DCT размер блока преобразования устанавливается в соответствии с размером фильтра (например, 7 × 7 для DCT фильтра 7 × 7). После преобразования коэффициенты квантуются. В отличие от кодирования с преобразованием при сжатии изображения (например, JPEG), где высокочастотные коэффициенты квантуются более грубо, чем низко частотные, это обуславливается зрительной системы человека, в нашем кодеке все коэффициенты квантуются одинаково. Это связано с большой важностью высоких частот, таких, как края для работы сети. Бит-глубина квантователя может быть настроена в соответствии с потребностями конкретного приложения. При незначительном воздействии на точность типичные значения составляют 5-6 бит/коэффициент.

Вес плотных слоев (также называемых полносвязанными слоями) и свертки 1 × 1 (не пространственная фильтрация, а фильтрация по глубине предыдущего слоя, обычно используемая в сетях для уменьшения глубины) расположены по блокам до кодирования с преобразованием. Для этого одномерные веса преобразуются в максимальный размер блока (до заданного уровня 8 × 8). Хотя эти одномерные параметры не имеют прямого пространственного контекста, исследование показало, что кодирование с преобразованием по-прежнему имеет более высокий эффект, снижающий энтропию, чем прямое квантование. Кроме того, стоит отметить, что одномерное кодирование с преобразованием не так эффективно, как двумерное с тем же числом значений.

Кластеризация K-значений используется для кодирования смещений и нормализаций. Стоит отметить, что результат этого метода аналогичен использованию квантователя Lloyd-Max. Количество кластеров

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

устанавливается аналогично глубине квантователя в соответствии с настройками качества. Для смещений и нормализации создаются кодовые книги. Таким образом, использование алгоритма кластеризации является полезным, если для кодирования индексов квантователя и самой кодовой книги требуется меньшее количество бит, чем для непосредственного кодирования значений. Кодовая книга, которая генерируется алгоритмом кластеризации, зависит от инициализации. По этой причине мы запускаем несколько итераций кластеризации с различной инициализацией и выбираем кодовую книгу, которая генерирует наименьшее искажение. Мы наблюдаем, что десять итераций с 50 шагами k –средних, являются достаточными (то есть повторное моделирование имеет те же результаты), и что больше итераций незначительно увеличивает производительность. Поскольку кодовая книга указывается как часть битового потока, число итераций не влияет на процесс декодирования.

Преимущество кластеризации заключается в том, что искажение меньше, чем для равномерного квантования. В результате, точность сети становится выше для заданного количества шагов квантователя. Однако, появление индексов кодовой книги также более равномерно распределено. Из-за большей энтропии этого распределения коэффициент сжатия значительно меньше. В частности, преобразование BurrowWheeler и преобразование move-to-front, которые оба вызываются для энтропийного кодирования, являются неудовлетворительными из-за равномерного распределения. Мы решили использовать одно и то же количество шагов квантователя для всех параметров. По этой причине кластеризация была выбрана для тех параметров сети, которые слишком чувствительны к более высоким искажениям, вызванным равномерным квантованием. Обработанные данные из кодирования с преобразованием и кластеризации

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

энтропийно кодируются по слоям с использованием BZip2, сериализуются и записываются в выходной файл. При необходимости, применяется заполнение байтов. В дополнение к весам, смещениям и нормализации, метаданные необходимые для процесса декодирования, также включены в выходной файл. Он включает в себя архитектуру слоев в сети, формы и размеры фильтров, сведения о блочных устройствах, масштабирующие факторы от предварительного масштабирования, коэффициенты масштабирования и смещения от квантователя и кодовые книги для кластеризации. Эти метаданные хранятся в файле в дополнение к весам сети, которые хранятся без структурной информации. Таким образом, декодированные модели сети могут быть загружены с использованием тех же API, что и для исходных моделей.

4. Расчеты

В этом разделе подробно описывается оценка предлагаемого кодека. Была выбрана классификация изображений как приложение, так как это хорошо понятное и наиболее распространенное приложение для глубокого обучения. Тем не менее, предлагаемый метод применим к нейронным сетям для других приложений. Представлено сжатие для четырех нейронных сетей: для двух современных сетей ResNet50 [3] и GoogLeNet [4], для знаменитой AlexNet [1], которая имеет специальное свойство, введенное ниже, и для SqueezeNet [6], архитектура которого уже оптимизирована, чтобы иметь веса меньшего размера.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

Таблица 1 – факторы сжатия при 1% и 2% снижении точности

Нейросеть	Метод	Параметры при 1% потеря точности	Параметры при 2% потеря точности
GoogLeNet	Квантизация	8.0	9.0
	Кластеризация (для всех слоев)	4.8	5.8
	DCT+Кластеризация+Квантизация	10.6	12.4
ResNet	Квантизация	4.7	4.8
	Кластеризация (для всех слоев)	4.2	4.9
	DCT+Кластеризация+Квантизация	8.1	9.7
AlexNet	Квантизация	7.8	8.9
	Кластеризация (для всех слоев)	5.6	6.6
	DCT+Кластеризация+Квантизация	6.7	7.7
SqueezeNet	Квантизация	5.9	6.7
	Кластеризация (для всех слоев)	3.7	4.6
	DCT+Кластеризация+Квантизация	6.3	7.4
Average	Квантизация	6.6	7.4
	Кластеризация (для всех слоев)	4.6	5.5
	DCT+Кластеризация+Квантизация	7.9	9.3

Таблица 2 - время кодирования и декодирования

Метод	Кодирование (с)	Декодирование (с)
Квантизация	26.2	6.5
Кластеризация (для всех слоев)	383.0	7.2
DCT+Кластеризация+Квантизация	29.0	7.6

Анализ оценки искажений является типичной процедурой для оценки алгоритмов сжатия. Производительность нейронных сетей для классификации изображений обычно измеряется с использованием точности Top-5 [2]. Чем выше точность, тем лучше производительность классификации. Поэтому мы измеряем искажение как уменьшение точности после сжатия сетей. Вместо использования абсолютного размера сетей, используется коэффициент сжатия (несжатый размер файла / сжатый размер

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

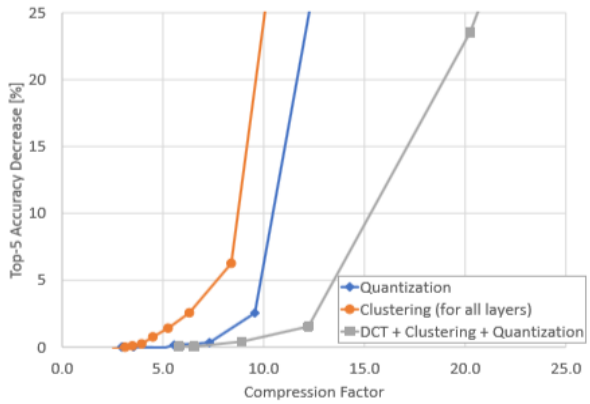
файла) для оценки эффективности сжатия. Это облегчает сравнение результатов для разных сетей (с различными размерами файлов) с первого взгляда.

Чтобы генерировать кривые RD, измеряется битовая глубина n для квантователя и количество кластеров ($= 2^n$). Глубина бита одинакова для всех слоев и не выбирается адаптивно. Сети кодируются, декодируются и затем используются для классификации изображений по приведенной выше схеме. В качестве данных мы используем набор проверки ILSVRC-2012 (50 000 изображений в 1000 классах). Чтобы изучить, какие алгоритмы из предложенной схемы обработки вносят вклад в окончательный результат, оцениваются три подмножества: в первом подмножестве применяется только квантование к весам сети. То же самое используется для встроенного сжатия tensorflow. Во втором подмножестве применен алгоритм кластеризации ко всем параметрам всех слоев. В третьем подмножестве применяется кодирование с преобразованием весов сверточных и плотных слоев, и кластеризация для смещений и нормализаций. Третье подмножество - это предлагаемый метод.

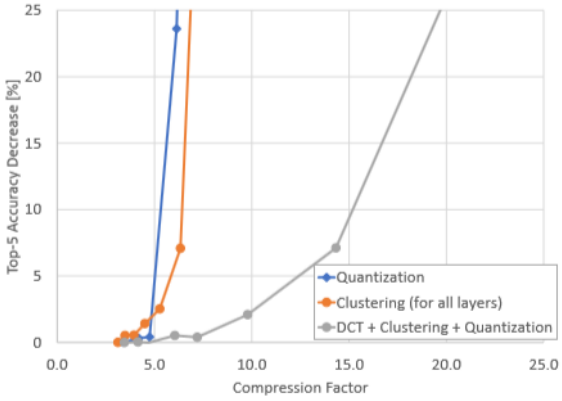
Результирующие кривые RD представлены на рисунке 2. Из рисунка 2 (а) и 2 (б) видно, что результаты для предлагаемого метода превосходят результаты других методов. Но, все методы, описанные в этой статье, имеют основания для существования, и могут применяться в качестве дополнения. Коэффициенты сжатия 10 или выше наблюдаются без значительного снижения точности. У AlexNet есть особое свойство, что он содержит необычайно большое количество весов и более 90% весов расположены на первом плотном слое. Как предложили Хан и др. [5], этот недостаток в дизайне сети может быть ограничен только отсечением и переобучением.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

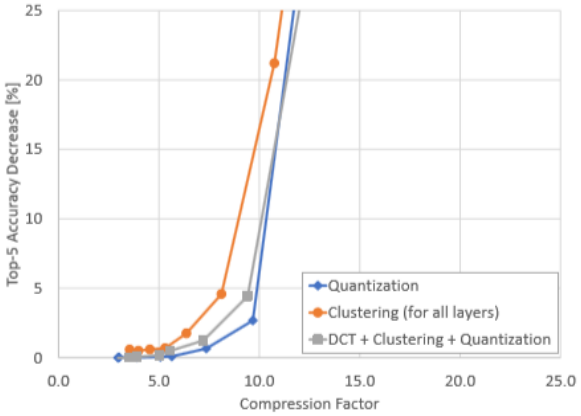
Сайт: akademnova.ru
e-mail: akademnova@mail.ru



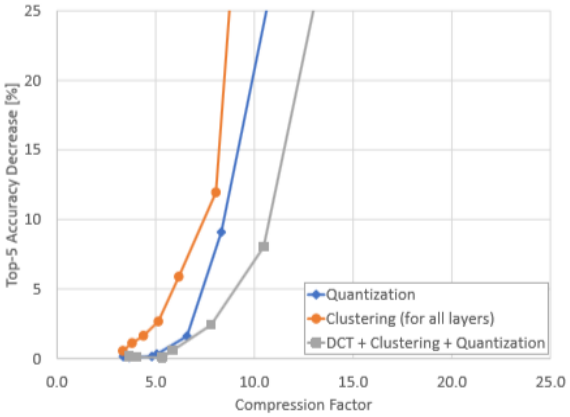
а) GoogLeNet



б) ResNet



в) AlexNet



г) SqueezeNet

Рисунок 2 – Результирующие кривые для исследуемых нейросетей

На рисунке 2 (в) видно, что кодирование с преобразованием и кластеризация не приносят никакого выигрыша по сравнению с квантованием. Тем не менее, наблюдаются коэффициенты сжатия 8-9, поскольку энтропийное кодирование может обеспечить выигрыш в кодировании, так как квантованные веса содержат много избыточности. На рисунке 2 (г) представлены результирующие кривые для сети SqueezeNet, архитектура которой уже была разработана с целью иметь как можно меньший размер весов. Они свидетельствуют о том, что предложенный фреймворк также полезен для сетей с оптимизированной архитектурой.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

Из базовых данных на рисунке 2 вычислены численные значения коэффициента сжатия, и представлены в таблице 1. Снижение точности 1% или 2% может считаться приемлемым для сжатых сетей в большинстве приложений, поскольку сжатые сети работают почти так же надежно, как и исходные несжатые. В среднем предлагаемый кодек достигает коэффициентов сжатия 7,9 и 9,3 при снижении точности на 1% и 2% соответственно.

Вычислительная сложность предлагаемого алгоритма измерялась времени работы кодировщика и декодера. Для этой цели выполнялся неоптимизированный код Python на процессоре AMD Ryzen 2700x для GoogLeNet. Время выполнения для трех нейросетей представлено в таблице 2. При применении только метода квантизации кодирование и декодирование было самым быстрым 26,2 с и 6,5 с соответственно. Метод кластеризации имеет самое большое время кодирования 383 с. Время декодирования значительно не увеличилось, поскольку кодовая книга не выводится на декодер, а сигнализируется как часть потока бит. Для конечного кодека (кодирование с преобразованием, сверточные и плотные веса, кластеризация для смещений и нормализаций), получили 29 секунд для кодировщика и 7,6 с для декодера. Время кодирования значительно не увеличивается по сравнению с квантованием, поскольку кластеризация применяется только к искажениям и нормализации. Время декодирования незначительно увеличивается из-за обратного DCT в процессе декодирования. В целом, время запуска кодировщика и декодера приемлемо, учитывая, что обучение нейронных сетей обычно длится несколько дней или недель.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

5. Заключение

В этой статье предложен кодек для сжатия нейронных сетей. Кодек основан на кодировании с преобразованием для сверточных и плотных слоев и кластеризации для смещений и нормализаций. Используя этот кодек, удалось достичь средних коэффициентов сжатия между 7,9 – 9,3, тогда как точность сжатых сетей для классификации изображений уменьшается только на 1% – 2% соответственно.

Список использованной литературы:

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., C. 1097–1105, 2012.
2. V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer, J.-H. Luo, J. Wu, and W. Lin, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” Proceedings of the IEEE, vol. 105, no. 12, C. 2295–2329, 2017.
3. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, C. 770–778, 2016.
4. C. Szegedy, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, C. 1–9, 2016.
5. S. Han, H. Mao, and W. J. Dally, “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding,” in ICLR, 2016.
6. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5MB model size,” in arXiv 1602.07360, 2016

Дата поступления в редакцию: 20.06.2018 г.

Опубликовано: 25.06.2018 г.