Бинаризация языковых моделей

Д. Н. Давыдова[1]

В последние годы в сфере обработки естественного языка широкое распространение получили большие языковые модели. Но, несмотря на их востребованность, их применение становится затруднительным из-за больших затрат времени, энергии и памяти.

Одним из способов решения этой проблемы является квантизация нейронных сетей — преобразование весов и активаций сети к представлению с более низкой точностью. Частным случаем квантизации является бинаризация — приведение параметров сети к разрядности 1 бит.

В работе рассмотрена структура бинарных нейронных сетей, приведен обзор текущих методов бинаризации языковых моделей, описаны полученные результаты.

Ключевые слова: обработка естественного языка, бинарные нейронные сети, бинаризация, квантизация, большие языковые модели.

1. Введение

Глубокие нейронные сети позволяют получить высокие результаты в различных задачах, таких, как классификация изображений, распознавание речи, машинный перевод и обработка естественного языка. В последние годы все более востребованными становятся большие языковые модели — языковые модели, содержащие более миллиарда параметров и позволяющие с высокой точностью выполнять различные языковые задачи [2], [1].

Но, несмотря на востребованность и преимущества больших языковых моделей, огромное количество параметров делает их обучение и запуск проблематичным. Обучение больших языковых моделей занимает много времени и требует больших вычислительных мощностей, а запуск таких моделей на устройствах с ограниченным количеством памяти, таких, как мобильные телефоны, оказывается затруднительным. В частности, обучение LLaMA-7B на 1T токенов на GPU A100-80GB заняло 82432 часа и потребовало 36 МВт · ч энергии [2], а ее хранение требует 12.55 ГБ, поэтому работать с такой моделью невозможно без мощных видеокарт.

[1] *Давыдова Дарья Николаевна* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: d.dawydowa2017@yandex.ru.

Davydova Daria Nikolaevna — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

Для решения этой проблемы предлагались различные методы оптимизации нейронных сетей, то есть, уменьшения затрат по времени или памяти моделей таким образом, чтобы потери в качестве при этом были как можно меньше.

Методы оптимизации можно условно разделить на методы, меняющие архитектуру модели, такие, как удаление или изменение слоев сети [3], [4]; методы сжатия модели, такие, как квантизация [5], [6] и прунинг [7], [8]; метод дистилляции знаний [9].

Квантизация предполагает сжатие модели за счет преобразования параметров и активаций сети к представлению с более низкой точностью, например, к 8, 4, 2 или 1 бит. Такой подход позволяет увеличить пропускную способность нейронной сети и упростить хранение модели, не изменяя при этом ее архитектуру.

Частным случаем квантизации является бинаризация нейронных сетей. При бинаризации нейронных сетей веса и активации, изначально представленные с точностью 32 бит, заменяются 1-битным представлением, а операции умножения матриц, затрачивающие большое количество вычислительных ресурсов, заменяются на булевы операции. Таким образом можно существенно ускорить выполнение вычислений модели и упростить ее хранение.

В данном обзоре будут рассмотрены результаты исследований по применению метода бинаризации к большим языковым моделям и проведен анализ полученных результатов.

2. Структура бинарных нейронных сетей

Как правило, при бинаризации нейронной сети значения весов и активаций ограничиваются до $+1$ и -1 (возможен также переход к значениям 0 и $+1$). Для преобразования 32-разрядного представления с плавающей точкой к бинарному используется функция бинаризации. В большинстве случаев для бинаризации используется функция знака — детерминированная функция, выдающая один и тот же результат при подаче на вход одинаковых значений входных аргументов.

$$Sign(x) = \begin{cases} +1, & \text{x} \geqslant 0 \\ -1, & \text{иначе} \end{cases}$$

Иногда используются стохастические функции бинаризации, приписывающие значения наборам аргументов с определенной вероятностью.

$$F_b(x) = \begin{cases} +1, & \text{с вероятностью } p = \sigma(x) \\ -1, & \text{с вероятностью } 1 - p, \end{cases}$$

где $\sigma(x) = \text{clip}(\frac{(x+1)}{2}, 0, 1) = \max(0, \min(1, \frac{(x+1)}{2}))$.

Несмотря на то, что применение такой функции может повысить качество работы модели, бинаризацию с ее помощью сложнее реализовать из-за того, что требуется генерация случайных бит, поэтому чаще используется функция знака.

Как и в общем случае квантизации, при бинаризации возможны два сценария: quantization-aware training (QAT), при котором 1-битные параметры и булевы операции задействованы в процессе обучения, и post-training quantization (PTQ), позволяющий напрямую квантизовать уже обученную модель даже без тонкой настройки.

В случае quantization-aware training во время прямого прохода по сети для вычисления выходов слоев вместо операции матричного умножения используется функция $XNOR$, $XNOR(A, B) = \overline{A \oplus B}$ и *popcount*, подсчитывающая количество единиц в заданном бинарном векторе.

Так как производная функции $Sign$ определена не во всех точках и почти всегда равна нулю, метод градиентного спуска не подходит для вычисления градиентов и обновления параметров при бинаризации в ходе обучения. Чтобы решить эту проблему, в [10] разработали метод STE (straight-through estimator) для приближения производной функции знака и обратного прохода по сети.

$$Approx(x) = \begin{cases} x, & \text{x} \geqslant \text{-1, x} \leqslant 1 \\ -1, & \text{x} \leqslant \text{-1} \\ +1, & \text{иначе} \end{cases}$$

$$STE(x) = \frac{\partial Approx(x)}{\partial x} = \begin{cases} 1, & \text{x} \geqslant \text{-1, x} \leqslant 1 \\ 0, & \text{иначе} \end{cases}$$

При обучении по сценарию quantization-aware training для прямого и обратного прохода по сети используются бинарные веса и активации, веса исходной модели хранятся в памяти и используются при обновлении весов перед следующим проходом по сети.

При подходе post-training quantization предобученная 32-битная модель оценивается с использованием небольшого набора калибровочных данных. Таким образом собирается статистика о распределениях весов и активаций и вычисляются калибровочные коэффициенты. Затем полученные статистики используются для квантизации модели. В ходе Post-training quantization параметры могут бинаризоваться динамически во время запуска модели с учетом входных данных или статически на основе изначальной статистики.

Подход quantization-aware training более сложен с вычислительной точки зрения, но позволяет в результате получить более точную модель,

лучше адаптированную для работы с бинарными значениями. Подход post-training quantization, исключающий алгоритм обратного распространения ошибки, проще и требует наличия только небольшого набора данных для калибровки, но бинаризованная модель оказывается менее точной.

Так как при бинаризации нейронной сети теряется существенная часть информации, точность может заметно снизиться по сравнению с изначальной моделью с 32-битными значениями. Для уменьшения ошибки бинаризации предлагались различные подходы, такие, как использование калибровочных коэффициентов при вычислении бинарных весов и активаций, оптимизация и изменение распределения весов и активаций перед бинаризацией, улучшение функции потерь с помощью добавления параметра регуляризации, дистилляция знаний [11].

3. Задачи обработки естественного языка, решенные с помощью бинарных нейронных сетей

Бинаризацию применяли для ускорения и облегчения различных языковых моделей, среди которых LSTM-сети [12], трансформерные модели, в частности, BERT [13] и LLaMA [2].

Большая часть исследований по бинаризации языковых моделей посвящена нейронным сетям, решающим задачу классификации или языкового моделирования. Задачу языкового моделирования можно сформулировать как моделирование вероятностного распределения следующего слова на основании предыдущих: $P(w_i|w_{i-1}, ..., w_0)$, где w_i — слово из словаря модели. Для оценки качества решения этой задачи используется перплексия — обратная вероятность тестовой коллекции, нормализованная по количеству слов:

$$PPL(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, ..., w_N)}},$$

где W — множество слов в тестовой коллекции, N — их количество.

Первые исследования по бинаризации языковых моделей проводятся в QAT-сценарии на LSTM-сетях. В исследовании [14] адаптируют метод бинаризации, примененный ранее в задаче компьютерного зрения, для языковых моделей. Авторы [16], [20] применяют методы из теории оптимизации для подбора наилучших параметров модели, при которых достигается минимальное значение функции потерь. В [19], [22] исследуют вопрос о том, какие компоненты языковых моделей затрачивают наибольшее количество вычислительных ресурсов. Авторы показывают, что в случае LSTM-моделей это слои эмбеддингов, и предлагают техники по уменьшению потерь точности при бинаризации этих слоев, такие,

как добавление дополнительных линейных слоев и дистилляция знаний. Ниже приведенные исследования будут описаны подробнее.

Впервые бинаризация языковой модели проводится в [14]. Авторы применяют метод, ранее использованный для бинаризации свёрточных нейронных сетей Xnor-net [15], на LSTM-сети. Рассматривается 2 сценария бинаризации: бинаризация только весов модели и бинаризация весов и эмбеддингов. В первом случае в результате экспериментов было получено улучшение показателей перплексии на задаче языкового моделирования и сопоставимая точность в задаче классификации. В этом случае бинаризация выступила как регуляризатор и улучшила обобщающую способность сети. Бинаризация весов и активаций привела к переобучению и снижению точности на тестовых данных.

Авторы [16] замечают, что в прошлых исследованиях по бинаризации нейронных сетей в процессе обучения аппроксимировали матрицы весов, но не учитывали влияние бинаризации на функцию потерь. Для поиска оптимальных бинарных весов авторы используют метод Ньютона [17] — итерационный численный метод для нахождения экстремума целевой функции. Подбор бинарных весов рассматривается как оптимизационная задача, в которой требуется подобрать веса таким образом, чтобы минимизировать значение функции потерь. Матрица Гессе, участвующая в разложении функции при применении метода Ньютона, не всегда положительно определена, к тому же, ее вычисление требует больших расходов по времени и памяти, поэтому она аппроксимируется с помощью положительной диагональной матрицы. Эта матрица вычисляется с помощью моментов второго порядка, которые автоматически подсчитываются оптимизатором Adam. Результаты экспериментов на LSTM-моделях, решающих задачу языкового моделирования, показывают, что учет влияния бинаризации на функцию потерь при обучении модели позволяет уменьшить ошибку и получить более высокие результаты по сравнению с более ранними методами бинаризации [18], [15].

Авторы [19] решают проблему большого расхода памяти на слоях эмбеддингов в случае, когда размер словаря модели достаточно большой. В этом исследовании бинаризация проводится в сценарии quantization-aware training. Авторы проводят бинаризацию входного и выходного слоя эмбеддингов LSTM-сети и добавляют линейный слой после входного слоя эмбеддингов и перед выходным для улучшения точности векторных представлений. В случае бинаризации только слоев эмбеддингов авторы получили уменьшение перплексии на задаче языкового моделирования, в случае полной бинаризации модели — несущественное увеличение перплексии. Дополнительно авторы показывают, что полученные бинарные эмбеддинги не теряют информацию в сравнении с изначальными эмбеддингами с точностью представления 32 бит.

В [20] впервые при QAT-бинаризации используют метод множителей переменного направления Alternating Direction Method of Multipliers (ADMM) [21] для подбора оптимальных параметров модели, на которых значение функции потерь будет минимальным. Метод ADMM является развитием метода множителей Лагранжа и заключается в декомпозиции сложной проблемы минимизации на более простые подзадачи. При применении ADMM функция, зависящая от двух групп переменных, поочередно минимизируется то по одной, то по другой группе переменных. В [20] с помощью лагранжиана поочередно оптимизируются 2 набора параметров: веса исходной сети и значения, к которым квантизуется модель, с коэффициентом. Авторы сравнивают предложенный подход с подходом из [19] на задаче языкового моделирования и распознавания речи и показывают, что их метод позволяет ускорить сходимость сети и получить при этом более низкие значения перплексии.

В исследовании [22] бинаризацию проводят в QAT-сценарии. Эмбеддинги, на которые приходятся самые большие затраты памяти, бинаризуют в технике Product Quantization — разложением векторного пространства параметров в декартово произведение подпространств меньшей размерности и независимой квантизацией каждого подпространства. Бинаризация выполняется по методу Soft Binarization, использующему дополнительные векторы с точностью представления 32 бит для уменьшения потери информации. Метод также предполагает дистилляцию знаний на основе расстояния между распределением выходов модели-учителя ("soft labels") и модели-ученика и функции потерь этих моделей. После бинаризации модель дообучают для улучшения ее качества. Авторам удалось сжать LSTM-модель, решающую задачу языкового моделирования, в 100 раз, при этом сохранить сопоставимые с исходной моделью значения перплексии.

В исследованиях [23], [24]-[28], [31], [32], [34]-[36] проводится бинаризация трансформерных моделей, в частности, модели BERT в сценарии quantization-aware training [23], [24]-[28], больших языковых моделей [34]-[36].

Ниже приведен обзор исследований по бинаризации модели BERT. Авторы исследований частично или полностью приводят параметры модели к разрядности 1 бит, и предлагают различные способы для сохранения качества работы модели. В [23], [24], [27] предлагаются различные варианты процедуры дистилляции знаний для уменьшения ошибки квантизации, а в [26] предлагают ансамблирование нескольких бинарных моделей как более оптимальную альтернативу дистилляции. В [23], [25] используются промежуточные этапы при переходе от 32-битной модели к бинарной: инициализация весов бинарной модели от тернарной [23] и постепенное снижение разрядности модели при переходе от 32-битной модели

к бинарной [25]. В [28] бинаризация проводится на этапе предобучения модели.

В [23] впервые проводится бинаризация трансформерной модели BERT. Авторы исследуют ландшафт функции потерь и выясняют, что по мере снижения точности представления параметров модели ошибка увеличивается несущественно вплоть до 2-битной модели, в то время как ландшафт функции потерь бинарной сети оказывается более сложным, что усложняет оптимизацию и обучение модели. Для решения этой проблемы авторы обучают тернарную модель, уменьшенную по количеству параметров в 2 раза, затем с помощью оператора, отображающего тернарные веса в бинарные, инициализируют веса бинарной сети. Помимо этого, для улучшения модели авторы применяют дистилляцию знаний intermediate-layer distillation, учитывающую ошибку выхода слоя эмбеддингов, механизма внимания и линейного слоя. Результаты экспериментов на наборе задач GLUE [38] показали несущественное снижение качества работы полученной модели BinaryBERT по сравнению с 32-битными моделями и существенное улучшение показателей по сравнению с бинаризацией сети напрямую [39].

В [24] проводится полная бинаризация модели BERT. Последовательно бинаризуя различные компоненты модели, авторы замечают, что наибольшее падение точности вызывает бинаризация механизма внимания. Для решения этой проблемы они представляют структуру Bi-Attention, максимизирующую энтропию бинаризованных векторов. В структуре Bi-Attention операция матричного умножения заменяется на операцию Bitwise-Affine Matrix Multiplication, выполняющую побитовые вычисления, и устраняется операция Softmax, так как в результате применения Softmax можно получить только неотрицательные числа, и бинаризация преобразует все ее выходные значения в 1. Вместо использования Softmax авторы предлагают использовать булеву функцию, переводящую элементы attention score с низким значением в 0, и с более высоким в 1, что позволит механизму внимания выделять наиболее релевантные элементы. Для решения другой проблемы — несовпадения ожидаемого и фактического направлений градиента при оптимизации, предложили метод дистилляции Direction-Matching Distillation, учитывающий ошибку модели-ученика на матрицах запроса, ключа и значения. Эксперименты с полученной моделью, названной BiBERT, показали более высокие результаты по сравнению с другими квантизованными моделями, в том числе, BinaryBERT [23].

В [25] представляют метод Efficient Two-Stage Progressive Quantization (ETSPQ), увеличивающий степень сжатия BERT при сохранении высокого качества работы модели за счет поэтапной квантизации. Как и в [23], авторы решают проблему оптимизации бинаризованной модели с

учетом сложного ландшафта функции потерь. Авторы снижают точность представления параметров модели в 2 стадии. На первой стадии степень сжатия весов постепенно увеличивают, на каждом шаге дообучая модель на нужную задачу, затем используя параметры полученной модели для инициализации модели меньшей битности. На следующей стадии постепенно снижают разрядность активаций. В результате удалось получить большую степень сжатия и большую точность в сравнении с BinaryBERT [23] на наборе задач GLUE [38].

Авторы [26] отмечают, что используемый в прошлых бинаризованных моделях BinaryBERT [23] и BiBERT [24] метод дистилляции знаний замедляет обучение моделей, к тому же, у этих моделей существенно снижается точность и устойчивость к возмущениям во входных данных. Для решения этих проблем они предлагают использовать ансамблирование нескольких бинаризованных моделей методом AdaBoost и отказаться от дистилляции знаний. Ансамблевая модель BEBERT, построенная объединением нескольких моделей BinaryBERT или BiBERT, показала более высокие результаты на наборе задач GLUE [38], чем BiBERT [24], и сопоставимые результаты с BinaryBERT [23], и при этом обучается в 2 раза быстрее.

В [27] разрабатывают более простой способ полной бинаризации модели BERT, чем в прошлых исследованиях [23], [24]. Авторы предлагают новый подход к бинаризации: активации слоев Softmax и ReLU, принимающие только положительные значения, бинаризуются к значениям $\{0, 1\}$, в то время как активации остальных слоев, принимающие как положительные, так и отрицательные значения — к $\{-1, 1\}$, чтобы лучше сохранить свойства распределений исходных активаций. Кроме того, предлагается новый подход к процедуре дистилляции знаний: вместо того, чтобы проводить дистилляцию напрямую от модели-учителя к модели-ученику, как в BinaryBERT [23] и BiBERT [24], авторы [27] используют промежуточную модель меньшей битности, чем исходная, выступающую в роли ученика для исходной модели и в роли учителя для бинаризованной. Полученная модель BiT показала более высокие результаты, чем модели в прошлых исследованиях [23], [24].

В отличие от подходов, представленных в [23], [24] — [27], в [28] бинаризацию весов и активаций внедряют в процесс предобучения модели. Предобучение проводится на 2 стандартных для BERT задачах masked language modeling и next sentence prediction с использованием дистилляции знаний от исходной модели к бинарной. При бинаризации параметров авторы придерживаются процедуры, представленной в [27]. Авторы пытаются минимизировать ошибку бинаризации в механизме Self-Attention и вводят для этого понятие остаточных полиномов. Авторы раскладывают матрицу запроса и ключа исходной 32-битной модели в

сумму бинаризованной матрицы и остаточной матрицы, и вычисляют attention score с учетом этого разложения. Слагаемые attention score, содержащие остаточные матрицы, образуют остаточный полином, который аппроксимируется с помощью обучающихся матриц параметров. Такой подход позволяет сохранять высокие результаты, сравнимые с другими SoTA-подходами бинаризации, и при этом делает модель устойчивой к изменению гиперпараметров и позволяет дообучать ее на различные задачи.

В таблице ниже приведены результаты бинаризации трансформерной модели BERT на наборе задач GLUE.

Таблица 1. Результаты бинаризации модели BERT на наборе задач GLUE

Метод	Способ	Биты, W-A	Размер, MB	FLOPs, G	GLUE Avg
BinaryBert	QAT	1-4	16.5	1.5	82.6
		1-1	16.5	0.4	50.1
BiBERT	QAT	1-1	13.4	0.4	67.0
ETSPQ	QAT	1-4	13.4	1.5	83.2
BEBERT	QAT	1-4	33	3.0/2	82.53
		1-1	-	-	74.97
BiT	QAT	1-1	13.4	0.4	78.0
BiPFT	QAT	1-1	14.9	0.4	70.8

Исследования по бинаризации больших языковых моделей нацелены на наибольшее приближение средней разрядности параметров модели к 1 или менее бит при минимизации потерь качества. Часто для уменьшения ошибки квантизации параметры модели приводятся к смешанной разрядности, или бинаризуются только некоторые компоненты модели [29], [30], [34]. Некоторые методы бинаризации больших языковых моделей заключаются в различных стратегиях отбора наиболее значимых для обучения весов и подбора наилучшей схемы для их бинаризации [30], [32], [36], другие — в минимизации ошибки квантизации за счет использования выхода модели [33], [37]. В [31] и [34] предлагают методы для инициализации бинарных моделей, позволяющие снизить потери качества и ускоряющие сходимость модели. В [35] для сохранения лингвистической способности бинарной модели используют технологию Mixture of Experts, а в [36] для приведения модели к разрядности менее 1 бит комбинируют квантизацию и прунинг.

В [29] впервые проводится бинаризация большой языковой модели в сценарии quantization-aware training. Авторы отмечают, что прошлые

исследования по бинаризации языковых моделей проводились в основном на модели BERT [23], [24] — [27], архитектура которой существенно отличается от архитектуры больших языковых моделей. Поэтому они разрабатывают свой метод бинаризации BitNet, при котором к разрядности 1 бит приводятся только веса линейного слоя, а остальные компоненты — к разрядности 8 бит. Для ускорения обучения модели авторы используют технологию параллельных вычислений Group Quantization, при которой веса и активации модели разбиваются на несколько групп, и параметры каждой группы вычисляются отдельно. Рассмотренный метод превосходит другие SoTA-методы квантизации, такие, как SmoothQuant [6], Absmax [40] и GPTQ [41], по уменьшению затрат памяти и энергии, при этом показывает сопоставимое качество на тестовых данных.

Как и в [29], в [30] отмечают невозможность адаптировать существующие методы бинаризации для больших языковых моделей, и предлагают новый подход Partially-Binarized LLM (PB-LLM), сохраняющий лингвистическую способность языковых моделей. Подход основан на отборе небольшого количества наиболее значимых для обучения весов и сохранении их в высокой разрядности, и бинаризации остальных весов. Метод применяется как для бинаризации в процессе обучения, так и для бинаризации уже обученной модели. В первом случае авторы замораживают наиболее значимые веса, отобранные по величине, и затем обучают бинаризованную модель. Во втором случае авторы обобщают метод квантизации Gptq [41] для бинаризации: итеративно бинаризуют незначимые веса и квантизуют к более высокой разрядности значимые, затем применяют к оставшимся весам компенсирующий коэффициент для уменьшения ошибки квантизации. Эксперименты на модели LLaMA [2] показали результаты, сравнимые с другими SoTA-методами квантизации SmoothQuant [6], Llm-qat [42] и RTN [43] в случае сохранения 30% весов в высокой разрядности при бинаризации после обучения, и более высокие результаты, чем другие методы квантизации в случае бинаризации в ходе обучения, обеспечивая при этом быструю сходимость сети.

Авторы [31] отмечают, что смешение разных разрядностей в PB-LLM [30] усложняет применение подхода и ограничивает экономию памяти. Они предлагают подход QAT-бинаризации Dual-Binarization (DB-LLM), при котором достигается удобный для оптимизации ландшафт функции потерь и используются булевы операции, оптимизирующие затраты вычислительных ресурсов. Авторы пользуются тем, что ландшафт функции потерь 2-битной модели более плоский и удобный для оптимизации, чем у бинарной, и раскладывают веса 2-битной модели в сумму бинарных весов, домноженных на калибровочный коэффициент. Полученные веса используются для инициализации бинарной модели, а калибровочные коэффициенты дополнительно настраиваются на этапе тонкой настрой-

ки. Другая проблема, замеченная авторами [31], заключается в том, что квантизованная модель по статистике более склонна предсказывать часто встречающиеся классы. Чтобы ее решить, предлагается подход Deviation-aware Distillation, уравновешивающий примеры из частых и редких классов, в котором энтропия модели-ученика и энтропия модели-учителя служат мерой неуверенности модели при предсказании классов. Бинаризация модели LLaMA [2] способом DB-LLM превосходит SoTA-методы квантизации, такие как Gptq [41], RTN [43] и PB-LLM [30] и позволяет добиться меньшей вычислительной сложности.

Еще одна попытка усовершенствовать метод PB-LLM в случае PTQ-бинаризации предпринимается в [32]. Авторы отмечают высокие затраты памяти метода PB-LLM, требующего хранения более 30% весов в высокой разрядности. Идея предложенного ими подхода BiLLM состоит в отборе значимых и незначимых для обучения весов и их бинаризации по двум различным схемам. Так как распределение значимых весов обладает высокой дисперсией, стандартные схемы post-training бинаризации для них не подходят. Предлагается новая схема бинаризации, по которой значимые веса бинаризуются рекурсивно: для бинаризованной матрицы параметров вычисляется остаточная матрица, и тот же процесс бинаризации применяется уже к ней. Авторы замечают, что оставшиеся незначимые веса имеют нормальное распределение, и исходя из этого, подбирают порог, разбивающий веса на две категории: сконцентрированные и разреженные, и вычисляют ошибку бинаризации для незначимых весов как сумму ошибки на разреженном и на сконцентрированном участке. Эксперименты на моделях семейства LLaMA [2] и OPT [44] показывают существенное улучшение показателей перплексии по сравнению с другими SoTA-методами квантизации Gptq [41], RTN [43] и PB-LLM [30], при этом наибольшее приближение средней разрядности весов модели к 1 бит.

Другой подход к PTQ-бинаризации Output-adaptive Calibration (OAC), учитывающий выход модели при квантизации, описан в [33]. Вместо того, чтобы вычислять ошибку бинаризации между выходом квантизованного и исходного слоя, авторы минимизируют расхождение между функцией потерь на выходе модели до и после квантизации. При нахождении этого расхождения метод OAC использует вторую производную перекрёстной энтропии для вычисления гессиана. Вычисленный гессиан используется для обновления весов и определения их значимости. Предложенный подход показывает лучшие результаты на моделях LLaMA [2] и OPT [44], чем другие подходы квантизации OPTQ [45], QuIP [46], SpQR [47] и BiLLM [32], не учитывающие выход модели при минимизации ошибки бинаризации.

Авторы подхода OneBit [34], предполагающего quantization-aware training, бинаризуют линейный слой по методу, представленному в BitNet

[29], но дополняют его двумя векторами разрядности 16 бит g и h, на которые покомпонентно домножаются столбцы матрицы активаций и матрицы бинарных весов, соответственно. Для инициализации бинарной модели авторы представляют метод Sign-Value-Independent Decomposition, при котором матрица весов из линейного слоя раскладывается в покомпонентное произведение матрицы знака и матрицы значений, а матрица значений затем раскладывается в произведение двух векторов, выполняющих роль векторов g и h при первом запуске. Для повышения качества работы модели применяется дистилляция знаний. Бинаризованная таким методом модель LLaMA [2] более стабильна к изменению гиперпараметров и показывает результаты, сравнимые с 16-битными моделями.

Авторы [35] решают проблему низкой точности бинаризованных моделей и представляют метод QAT-бинаризации Mixture of Scales (BinaryMoS), затрачивающий небольшое дополнительное количество памяти, но существенно улучшающий лингвистические способности модели. Метод вдохновлен структурой Mixture of Experts [48], дублирующей слои модели и выбирающей подходящий для данной задачи слой (эксперт) среди дубликатов во время запуска модели на основе коэффициентов, приписанных маршрутизатором. В качестве экспертов в случае BinaryMoS выступают вектора калибровочных коэффициентов, а маршрутизатор, отбирающий наиболее подходящие из них в зависимости от входного токена, представлен линейным слоем с функцией активации Softmax. Пользуясь тем, что коэффициенты задействованы только в линейных слоях, BinaryMoS динамически генерирует инструкции о том, как линейно комбинировать вектора коэффициентов, что позволяет не ограничиваться фиксированным числом экспертов. Как и в прошлых подходах, авторы применяют дистилляцию знаний для улучшения работы модели. Подход BinaryMoS показал более высокие результаты на моделях LLaMA [2] и OPT [44], чем прошлые подходы PB-LLM [30], OneBit [34] и BiLLM [32], при этом сохраняя низкие затраты по памяти.

Подход post-training quantization, названный STructured Binarization for LLMs (STBLLM) [36], сочетает в себе сразу две техники сжатия модели — бинаризацию и прунинг. Для приведения модели к средней разрядности менее 1 бит авторы проводят прунинг весов предобученной модели по методу N:M Sparsity [49], который кодирует N последовательных ненулевых элементов матрицы весов с помощью чисел в M-битном представлении, затем бинаризуют модель. Как и в прошлых исследованиях, веса модели разделяются на значимые и незначимые. Авторы представляют новую метрику для отбора значимых весов на основе их величины — Standardized Importance, не использующую гессиан и упрощающую вычисления. Значимые веса бинаризуются способом, представленным в BiLLM [32], а незначимые, как и в BiLLM, разбиваются на группы на основе их рас-

пределения. Метод STBLLM позволяет получить лучшие результаты на моделях LLaMA [2], OPT [44] и Mistral [50], чем BiLLM, и показывает высокий потенциал дальнейшего сжатия моделей до разрядности менее 1 бит.

Метод обучения бинаризованной большой языковой модели с нуля в сценарии quantization-aware training предлагается в [37]. Авторы бинаризуют только линейные слои модели с использованием калибровочных коэффициентов, а для уменьшения ошибки бинаризации используют авторегрессионную дистилляцию знаний, при которой на каждом шаге предсказания следующего токена вычисляется перекрестная энтропия между распределением вероятностей выходного токена 32-битной модели-учителя и бинарной модели-ученика. Эксперименты показали, что использование такой функции дает достаточно высокие результаты, поэтому другие слагаемые в функцию потерь не включаются. Предложенный подход бинаризации Fully BInarized Large Language Model (FBI-LLM), примененный к моделям LLaMA [2] и OPT [44], показывает более высокие результаты в большинстве экспериментов, чем другие SoTA-методы бинаризации BiLLM [32], OneBit [34], BitNet [29].

Ниже приведена таблица с результатами бинаризации больших языковых моделей.

Таблица 2. Результаты бинаризации больших языковых моделей

Метод	Способ	Модель	Биты	Корпус	Метрика	Значение
BitNet	QAT	Transformer	-	HellaSwag	Acc	38.9
				Winogrande	Acc	51.4
PB-LLM	QAT	LLaMA-1-7B	1.70	WikiText2	PPL	20.61
				C4	PPL	47.09
PB-LLM	PTQ	LLaMA-1-7B	1.70	WikiText2	PPL	102.36
		OPT-1.3B	1.70	WikiText2	PPL	265.52
		OPT-13B	1.70	WikiText2	PPL	81.92
DB-LLM	QAT	LLaMA-1-7B	-	WikiText2	PPL	7.59
				C4	PPL	9.74
BiLLM	PTQ	LLaMA-1-7B	1.08	WikiText2	PPL	35.04
			1.08	C4	PPL	39.6
		LLaMA2-7B	1.08	WikiText2	PPL	32.5
			1.08	C4	PPL	40.5
		OPT-1.3B	1.11	WikiText2	PPL	35.4
			1.11	C4	PPL	43.2
		OPT-1.3B	0.55	Wikitext2	PPL	106.99
		Mistral	0.55	Wikitext2	PPL	189.73
OAC	PTQ	LLaMA-1-7B	1.09	WikiText2	PPL	17.79
			1.09	C4	PPL	19.82
		OPT-13B	2.10	WikiText2	PPL	11.75
			2.10	C4	PPL	13.25
OneBit	QAT	LLaMA-1-7B	-	WikiText2	PPL	10.19
				C4	PPL	11.40
		LLaMA-2-7B	-	WikiText2	PPL	9.7
				C4	PPL	11.1
BinaryMoS	QAT	LLaMA-1-7B	1.0	WikiText2	PPL	7.97
			1.0	C4	PPL	9.72
		OPT-1.3B	1.0	WikiText2	PPL	18.45
			1.0	C4	PPL	18.83
STBLLM	PTQ	LLaMA-1-7B	0.55	Wikitext2	PPL	31.72
		OPT-1.3B	0.55	Wikitext2	PPL	45.11
		Mistral	0.55	Wikitext2	PPL	70.14
FBI-LLM	QAT	LLaMA-2-7B	1.01	Wikitext2	PPL	5.7
			1.01	C4	PPL	7.3
			1.01	HellaSwag	Acc	57.7
			1.01	Winogrande	Acc	58.9
		OPT-1.3B	1.01	Wikitext2	PPL	12.6
			1.01	C4	PPL	13.8

4. Заключение

Были рассмотрены основные подходы к бинаризации языковых моделей, в том числе, исследования по бинаризации рекуррентных сетей [14] — [22], модели BERT [23], [24] — [28], и последние исследования по бинаризации больших языковых моделей [31], [32], [34]. Для повышения точности и уменьшения средней разрядности языковых моделей предлагались различные математические и технические решения, но можно выделить несколько тенденций.

Большинство исследований по бинаризации языковых моделей проводятся в QAT-сценарии, так как такой подход позволяет получить меньшее падение точности моделей и лучше адаптировать их для работы с бинарными параметрами. Тем не менее, направление post-training binarization кажется перспективным для будущих исследований — предлагаются различные техники по бинаризации параметров, и эти подходы обобщаются для различных архитектур, таких, как LLaMA, OPT, Mistral [32].

Несмотря на популярность подхода quantization-aware training, некоторые проблемы остаются нерешенными. В частности, бинаризация не способна давать высокие результаты без применения дополнительных методов оптимизации, таких, как дистилляция знаний [23], ансамблирование [26] или дополнительные стадии обучения [25], которые затрачивают дополнительные вычислительные ресурсы и время.

Одним из популярных направлений исследований по бинаризации является сохранение информации бинарных представлений при прямом и обратном проходе по сети. Для этого предлагались такие методы, как максимизация энтропии бинаризованных векторов [24], метод дистилляции знаний, учитывающий направление градиента [24], заморозка наиболее значимых весов в высокой разрядности [30].

Еще одним перспективным направлением исследований является снижение средней разрядности бинаризованных моделей. Фактически, бинаризованные модели часто имеют среднюю точность представления более 1 бит, так как для уменьшения потерь в качестве требуется сохранение части параметров в более высокой разрядности, чем 1 бит. Тем не менее, в недавних исследованиях достигается все более и более низкая разрядность параметров [36].

Бинаризация является мощной техникой оптимизации языковых моделей, которая может позволить внедрить большие языковые модели на пользовательские устройства с ограниченным количеством памяти. В качестве направлений будущих исследований можно выделить комбинирование различных техник оптимизации для достижения наибольшего сжатия моделей, таких, как бинаризация и прунинг; уменьшение падения точности по сравнению с исходными 32-битными моделями и разработку

универсальной схемы бинаризации для различных языковых задач, таких, как распознавание речи, классификация и языковое моделирование.

Список литературы

[1] OpenAI, "GPT-4 Technical Report", *arXiv preprint arXiv:2303.08774*, 2023, 100 pp., arXiv: arXiv:2303.08774

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models", *arXiv preprint arXiv:2302.13971*, 2023, 27 pp., arXiv: arXiv:2302.13971

[3] Minjia Zhang, Yuxiong He, "Accelerating training of transformer-based language models with progressive layer dropping", *Advances in Neural Information Processing Systems*, **33**, Neural Information Processing Systems Foundation, 2020, 14011–14023

[4] Yujie Zeng, Wenlong He, Ihor Vasyltsov, Jiali Pang, Lin Chen, "Acceleration of large transformer model training by sensitivity-based layer dropping", *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, Association for the Advancement of Artificial Intelligence (AAAI), 2023, 11156–11163

[5] Zhewei Yao, Reza Y. Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, Yuxiong He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers", *Advances in Neural Information Processing Systems*, **35**, Neural Information Processing Systems Foundation, 2022, 27168–27183

[6] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, Song Han, "Smoothquant: Accurate and efficient post-training quantization for large language models", *Proceedings of the 40th International Conference on Machine Learning*, **202**, Proceedings of Machine Learning Research (PMLR), 2023, 38087–38099

[7] Yann LeCun, John S. Denker, Sara A. Solla, "Optimal brain damage", *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, **2**, MIT Press, 1989, 598–605

[8] Babak Hassibi, David G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon", *Proceedings of the 6th International*

Conference on Neural Information Processing Systems (NIPS'92), **5**, Morgan Kaufmann Publishers Inc., 1992, 164–171

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, "Distilling the Knowledge in a Neural Network", *arXiv preprint arXiv:1503.02531*, 2015, 6 pp., arXiv: arXiv:1503.02531

[10] Yoshua Bengio, Nicholas Léonard, Aaron Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation", *arXiv preprint arXiv:1308.3432*, 2013, 12 pp., arXiv: arXiv:1308.3432

[11] Chunyu Yuan, Sos S. Agaian, "A comprehensive review of binary neural network", *Artificial Intelligence Review*, **56**:11 (2023), 12949–13013

[12] Sepp Hochreiter, Jürgen Schmidhuber, "Long short-term memory", *Neural Computation*, **9**:8 (1997), 1735–1780

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1**, Association for Computational Linguistics, 2019, 4171–4186

[14] Weiyi Zheng, Yina Tang, "Binarized neural networks for language modeling", *Technical Report CS224d, Stanford University*, 2016, 9 pp.

[15] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks", *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, **9908**, Springer, 2016, 525–542

[16] Lu Hou, Quanming Yao, James T. Kwok, "Loss-aware Binarization of Deep Networks", *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, International Conference on Learning Representations (ICLR), 2017, 9 pp.

[17] Jason D. Lee, Yuekai Sun, Michael A. Saunders, "Proximal Newton-type methods for minimizing composite functions", *SIAM Journal on Optimization*, **24**:3 (2014), 1420–1443

[18] Matthieu Courbariaux, Yoshua Bengio, Jean-Pierre David, "BinaryConnect: Training deep neural networks with binary weights during propagations", *Advances in Neural Information Processing Systems (NeurIPS)*, **28**, Curran Associates, Inc., 2015, 3123–3131

[19] Xuan Liu, Di Cao, Kai Yu, "Binarized LSTM Language Model", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, 2018, 2113–2121

[20] Junhao Xu, Xie Chen, Shoukang Hu, Jianwei Yu, Xunying Liu, Helen Meng, "Low-bit Quantization of Recurrent Neural Network Language Models Using Alternating Direction Methods of Multipliers", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE Press, 2020, 7939–7943

[21] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Foundations and Trends in Machine Learning*, **3**:1 (2011), 1–122

[22] Kai Yu, Rao Ma, Kaiyu Shi, Qi Liu, "Neural Network Language Model Compression With Product Quantization and Soft Binarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**:10 (2020), 2438–2449

[23] Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, Irwin King, "BinaryBERT: Pushing the Limit of BERT Quantization", *arXiv preprint arXiv:2012.15701*, 2020, 13 pp., arXiv: arXiv:2012.15701

[24] Haotong Qin, Yifu Ding, Mingyuan Zhang, Qinghua Yan, Aishan Liu, Qingqing Dang, Ziwei Liu, Xianglong Liu, "BiBERT: Accurate Fully Binarized BERT", *Proceedings of the International Conference on Learning Representations (ICLR)*, International Conference on Learning Representations, 2022, 12 pp.

[25] Phuc H. C. Le, *Towards Accurate Low-Bitwidth BERT*, McGill University, Montreal, Canada, 2023, 120 pp.

[26] Jie Tian, Chen Fang, Hui Wang, Zhiyuan Wang, "BEBERT: Efficient and Robust Binary Ensemble BERT", *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE Press, 2023, 1–5

[27] Zechun Liu, Barlas Oğuz, Aasish Pappu, Lin Xiao, Scott Yih, Meng Li, Raghuraman Krishnamoorthi, Yashar Mehdad, "BiT: Robustly Binarized Multi-Distilled Transformer", *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS '22)*, **35**, Curran Associates Inc., 2022, 14303–14316

[28] Xingrun Xing, Li Du, Xinyuan Wang, Xianlin Zeng, Yequan Wang, Zheng Zhang, Jiajun Zhang, "BiPFT: Binary Pre-trained Foundation Transformer with Low-Rank Estimation of Binarization Residual Polynomials", *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*, **38**, AAAI Press, 2024, 16094–16102

[29] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, Furu Wei, "BitNet: Scaling 1-bit Transformers for Large Language Models", *arXiv preprint arXiv:2310.11453*, 2023, 15 pp., arXiv: arXiv:2310.11453

[30] Yuzhang Shang, Zhihang Yuan, Qiang Wu, Zhen Dong, "PB-LLM: Partially Binarized Large Language Models", *arXiv preprint arXiv:2310.00034*, 2023, 14 pp., arXiv: arXiv:2310.00034

[31] Hong Chen, Chengtao Lv, Liang Ding, Haotong Qin, Xiabin Zhou, Yifu Ding, Xuebo Liu, Min Zhang, Jinyang Guo, Xianglong Liu, Dacheng Tao, "DB-LLM: Accurate Dual-Binarization for Efficient LLMs", *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*, **1**, Association for Computational Linguistics, 2024, 8719–8730

[32] Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, Xiaojuan Qi, "BiLLM: Pushing the Limit of Post-Training Quantization for LLMs", *Proceedings of the 41st International Conference on Machine Learning (ICML)*, **202**, JMLR.org, 2024, 12950–12969

[33] Ali Edalati, Alireza Ghaffari, Masoud Asgharian, Lu Hou, Boxing Chen, Vahid Partovi Nia, "OAC: Output-Adaptive Calibration for Accurate Post-Training Quantization", *arXiv preprint arXiv:2405.15025*, 2024, 13 pp., arXiv: arXiv:2405.15025

[34] Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, Wanxiang Che, "OneBit: Towards Extremely Low-Bit Large Language Models", *Advances in Neural Information Processing Systems (NeurIPS)*, **37** (2024), 1–14

[35] Dongwon Jo, Taesu Kim, Yulhwa Kim, Jae-Joon Kim, "Mixture of Scales: Memory-Efficient Token-Adaptive Binarization for Large Language Models", *arXiv preprint arXiv:2406.12311*, 2024, 11 pp., arXiv: arXiv:2406.12311

[36] Peijie Dong, Lujun Li, Dayou Du, Yuhan Chen, Zhenheng Tang, Qiang Wang, Wei Xue, Wenhan Luo, Qifeng Liu, Yike Guo, Xiaowen Chu, "STBLLM: Breaking the 1-Bit Barrier with Structured Binary LLMs", *arXiv preprint arXiv:2408.01803*, 2024, 23 pp., arXiv: arXiv:2408.01803

[37] Liqun Ma, Mingjie Sun, Zhiqiang Shen, "FBI-LLM: Scaling Up Fully Binarized LLMs from Scratch via Autoregressive Distillation", *arXiv preprint arXiv:2407.07093*, 2024, 18 pp., arXiv: arXiv:2407.07093

[38] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, **57**, Association for Computational Linguistics, 2018, 353–355

[39] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, Yoshua Bengio, "Binarized Neural Networks", *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*, **29**, Curran Associates Inc., 2016, 4114–4122

[40] Tim Dettmers, Mike Lewis, Younes Belkada, Luke Zettlemoyer, "GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale", *Advances in Neural Information Processing Systems*, **35**, Curran Associates, Inc., 2022, 30318–30332

[41] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers", *arXiv preprint arXiv:2210.17323*, 2023, 9 pp., arXiv: arXiv:2210.17323

[42] Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, Vikas Chandra, "LLM-QAT: Data-Free Quantization Aware Training for Large Language Models", *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, 467–484

[43] Yuhang Li, Xin Dong, Sai Zhang, Haoli Bai, Yuanpeng Chen, Wei Wang, "RTN: Reparameterized Ternary Network", *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, AAAI Press, 2020, 4780–4787

[44] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Luke Zettlemoyer, "OPT: Open Pre-trained Transformer Language Models", *arXiv preprint arXiv:2205.01068*, 2022, 40 pp., arXiv: arXiv:2205.01068

[45] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh, "OPTQ: Accurate quantization for generative pre-trained transformers", *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, International Conference on Learning Representations, 2023, 11 pp.

[46] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, Christopher De Sa, "QuIP: 2-Bit Quantization of Large Language Models with Guarantees", *Advances in Neural Information Processing Systems*, **36** (2023), 37371–37382

[47] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, Dan Alistarh, "SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression", *arXiv preprint arXiv:2306.03078*, 2023, 15 pp., arXiv: arXiv:2306.03078

[48] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, Jeff Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer", *arXiv preprint arXiv:1701.06538*, 2017, 13 pp., arXiv: arXiv:1701.06538

[49] Asit Mishra, Jorge A. Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, Paulius Micikevicius, "Accelerating Sparse Deep Neural Networks", *arXiv preprint arXiv:2104.08378*, 2021, 9 pp., arXiv: arXiv:2104.08378

[50] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, William El Sayed, "Mistral 7B", *arXiv preprint arXiv:2310.06825*, 2023, 33 pp., arXiv: arXiv:2310.06825

Binarization of language models

Davydova D.N.

Large language models are widely used in the field of natural language processing. However, despite their high efficiency, the application of large language models becomes difficult due to their high computational and memory costs.

One of the ways to solve this problem is neural network quantization, that is, converting the weights and activations of the network to a representation with lower bit-width. A special case of quantization is binarization, which is the compression of network parameters to a bit-width of 1 bit.

In this paper, the structure of binary neural networks is considered, an overview of current methods of language model binarization is provided, and the results obtained are described.

Keywords: natural language processing, binary neural networks, binarization, quantization, large language models.

References

[1] OpenAI, "GPT-4 Technical Report", *arXiv preprint arXiv:2303.08774*, 2023, 100 pp., arXiv: arXiv:2303.08774

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models", *arXiv preprint arXiv:2302.13971*, 2023, 27 pp., arXiv: arXiv:2302.13971

[3] Minjia Zhang, Yuxiong He, "Accelerating training of transformer-based language models with progressive layer dropping", *Advances in Neural Information Processing Systems*, **33**, Neural Information Processing Systems Foundation, 2020, 14011–14023

[4] Yujie Zeng, Wenlong He, Ihor Vasyltsov, Jiali Pang, Lin Chen, "Acceleration of large transformer model training by sensitivity-based layer dropping", *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, Association for the Advancement of Artificial Intelligence (AAAI), 2023, 11156–11163

[5] Zhewei Yao, Reza Y. Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, Yuxiong He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers", *Advances in Neural Information Processing Systems*, **35**, Neural Information Processing Systems Foundation, 2022, 27168–27183

[6] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, Song Han, "Smoothquant: Accurate and efficient post-training quantization for large language models", *Proceedings of the 40th International Conference on Machine Learning*, **202**, Proceedings of Machine Learning Research (PMLR), 2023, 38087–38099

[7] Yann LeCun, John S. Denker, Sara A. Solla, "Optimal brain damage", *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, **2**, MIT Press, 1989, 598–605

[8] Babak Hassibi, David G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon", *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS'92)*, **5**, Morgan Kaufmann Publishers Inc., 1992, 164–171

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, "Distilling the Knowledge in a Neural Network", *arXiv preprint arXiv:1503.02531*, 2015, 6 pp., arXiv: arXiv:1503.02531

[10] Yoshua Bengio, Nicholas Léonard, Aaron Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation", *arXiv preprint arXiv:1308.3432*, 2013, 12 pp., arXiv: arXiv:1308.3432

[11] Chunyu Yuan, Sos S. Agaian, "A comprehensive review of binary neural network", *Artificial Intelligence Review*, **56**:11 (2023), 12949–13013

[12] Sepp Hochreiter, Jürgen Schmidhuber, "Long short-term memory", *Neural Computation*, **9**:8 (1997), 1735–1780

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1**, Association for Computational Linguistics, 2019, 4171–4186

[14] Weiyi Zheng, Yina Tang, "Binarized neural networks for language modeling", *Technical Report CS224d, Stanford University*, 2016, 9 pp.

[15] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks", *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, **9908**, Springer, 2016, 525–542

[16] Lu Hou, Quanming Yao, James T. Kwok, "Loss-aware Binarization of Deep Networks", *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, International Conference on Learning Representations (ICLR), 2017, 9 pp.

[17] Jason D. Lee, Yuekai Sun, Michael A. Saunders, "Proximal Newton-type methods for minimizing composite functions", *SIAM Journal on Optimization*, **24**:3 (2014), 1420–1443

[18] Matthieu Courbariaux, Yoshua Bengio, Jean-Pierre David, "BinaryConnect: Training deep neural networks with binary weights during propagations", *Advances in Neural Information Processing Systems (NeurIPS)*, **28**, Curran Associates, Inc., 2015, 3123–3131

[19] Xuan Liu, Di Cao, Kai Yu, "Binarized LSTM Language Model", *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, 2113–2121

[20] Junhao Xu, Xie Chen, Shoukang Hu, Jianwei Yu, Xunying Liu, Helen Meng, "Low-bit Quantization of Recurrent Neural Network Language Models Using Alternating Direction Methods of Multipliers", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE Press, 2020, 7939–7943

[21] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Foundations and Trends in Machine Learning*, **3**:1 (2011), 1–122

[22] Kai Yu, Rao Ma, Kaiyu Shi, Qi Liu, "Neural Network Language Model Compression With Product Quantization and Soft Binarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**:10 (2020), 2438–2449

[23] Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, Irwin King, "BinaryBERT: Pushing the Limit of BERT Quantization", *arXiv preprint arXiv:2012.15701*, 2020, 13 pp., arXiv: arXiv:2012.15701

[24] Haotong Qin, Yifu Ding, Mingyuan Zhang, Qinghua Yan, Aishan Liu, Qingqing Dang, Ziwei Liu, Xianglong Liu, "BiBERT: Accurate Fully Binarized BERT", *Proceedings of the International Conference on Learning Representations (ICLR)*, International Conference on Learning Representations, 2022, 12 pp.

[25] Phuc H. C. Le, *Towards Accurate Low-Bitwidth BERT*, McGill University, Montreal, Canada, 2023, 120 pp.

[26] Jie Tian, Chen Fang, Hui Wang, Zhiyuan Wang, "BEBERT: Efficient and Robust Binary Ensemble BERT", *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE Press, 2023, 1–5

[27] Zechun Liu, Barlas Oğuz, Aasish Pappu, Lin Xiao, Scott Yih, Meng Li, Raghuraman Krishnamoorthi, Yashar Mehdad, "BiT: Robustly Binarized Multi-Distilled Transformer", *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS '22)*, **35**, Curran Associates Inc., 2022, 14303–14316

[28] Xingrun Xing, Li Du, Xinyuan Wang, Xianlin Zeng, Yequan Wang, Zheng Zhang, Jiajun Zhang, "BiPFT: Binary Pre-trained Foundation Transformer with Low-Rank Estimation of Binarization Residual Polynomials", *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*, **38**, AAAI Press, 2024, 16094–16102

[29] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, Furu Wei, "BitNet: Scaling 1-bit Transformers for Large Language Models", *arXiv preprint arXiv:2310.11453*, 2023, 15 pp., arXiv: arXiv:2310.11453

[30] Yuzhang Shang, Zhihang Yuan, Qiang Wu, Zhen Dong, "PB-LLM: Partially Binarized Large Language Models", *arXiv preprint arXiv:2310.00034*, 2023, 14 pp., arXiv: arXiv:2310.00034

[31] Hong Chen, Chengtao Lv, Liang Ding, Haotong Qin, Xiabin Zhou, Yifu Ding, Xuebo Liu, Min Zhang, Jinyang Guo, Xianglong Liu, Dacheng Tao, "DB-LLM: Accurate Dual-Binarization for Efficient LLMs", *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*, **1**, Association for Computational Linguistics, 2024, 8719–8730

[32] Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, Xiaojuan Qi, "BiLLM: Pushing the Limit of Post-Training Quantization for LLMs", *Proceedings of the 41st International Conference on Machine Learning (ICML)*, **202**, JMLR.org, 2024, 12950–12969

[33] Ali Edalati, Alireza Ghaffari, Masoud Asgharian, Lu Hou, Boxing Chen, Vahid Partovi Nia, "OAC: Output-Adaptive Calibration for Accurate Post-Training Quantization", *arXiv preprint arXiv:2405.15025*, 2024, 13 pp., arXiv: arXiv:2405.15025

[34] Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, Wanxiang Che, "OneBit: Towards Extremely Low-Bit Large Language Models", *Advances in Neural Information Processing Systems (NeurIPS)*, **37** (2024), 1–14

[35] Dongwon Jo, Taesu Kim, Yulhwa Kim, Jae-Joon Kim, "Mixture of Scales: Memory-Efficient Token-Adaptive Binarization for Large Language Models", *arXiv preprint arXiv:2406.12311*, 2024, 11 pp., arXiv: arXiv:2406.12311

[36] Peijie Dong, Lujun Li, Dayou Du, Yuhan Chen, Zhenheng Tang, Qiang Wang, Wei Xue, Wenhan Luo, Qifeng Liu, Yike Guo, Xiaowen Chu, "STBLLM: Breaking the 1-Bit Barrier with Structured Binary LLMs", *arXiv preprint arXiv:2408.01803*, 2024, 23 pp., arXiv: arXiv:2408.01803

[37] Liqun Ma, Mingjie Sun, Zhiqiang Shen, "FBI-LLM: Scaling Up Fully Binarized LLMs from Scratch via Autoregressive Distillation", *arXiv preprint arXiv:2407.07093*, 2024, 18 pp., arXiv: arXiv:2407.07093

[38] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, **57**, Association for Computational Linguistics, 2018, 353–355

[39] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, Yoshua Bengio, "Binarized Neural Networks", *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*, **29**, Curran Associates Inc., 2016, 4114–4122

[40] Tim Dettmers, Mike Lewis, Younes Belkada, Luke Zettlemoyer, "GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale", *Advances in Neural Information Processing Systems*, **35**, Curran Associates, Inc., 2022, 30318–30332

[41] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers", *arXiv preprint arXiv:2210.17323*, 2023, 9 pp., arXiv: arXiv:2210.17323

[42] Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, Vikas Chandra, "LLM-QAT: Data-Free Quantization Aware Training for Large Language Models", *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, 467–484

[43] Yuhang Li, Xin Dong, Sai Zhang, Haoli Bai, Yuanpeng Chen, Wei Wang, "RTN: Reparameterized Ternary Network", *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, AAAI Press, 2020, 4780–4787

[44] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Luke Zettlemoyer, "OPT: Open Pre-trained Transformer Language Models", *arXiv preprint arXiv:2205.01068*, 2022, 40 pp., arXiv: arXiv:2205.01068

[45] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh, "OPTQ: Accurate quantization for generative pre-trained transformers", *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, International Conference on Learning Representations, 2023, 11 pp.

[46] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, Christopher De Sa, "QuIP: 2-Bit Quantization of Large Language Models with Guarantees", *Advances in Neural Information Processing Systems*, **36** (2023), 37371–37382

[47] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, Dan Alistarh, "SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression", *arXiv preprint arXiv:2306.03078*, 2023, 15 pp., arXiv: arXiv:2306.03078

[48] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, Jeff Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer", *arXiv preprint arXiv:1701.06538*, 2017, 13 pp., arXiv: arXiv:1701.06538

[49] Asit Mishra, Jorge A. Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, Paulius Micikevicius, "Accelerating Sparse Deep Neural Networks", *arXiv preprint arXiv:2104.08378*, 2021, 9 pp., arXiv: arXiv:2104.08378

[50] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, William El Sayed, "Mistral 7B", *arXiv preprint arXiv:2310.06825*, 2023, 33 pp., arXiv: arXiv:2310.06825