Варбанский К.С., Городничев М.Г. Анализ тенденций развития сетей с комплементарной разряженностью // Международный журнал информационных технологий и энергоэффективности. – 2024. – Т. 9 № 12(50) с. 43–51

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.932.7

АНАЛИЗ ТЕНДЕНЦИЙ РАЗВИТИЯ СЕТЕЙ С КОМПЛЕМЕНТАРНОЙ РАЗРЯЖЕННОСТЬЮ

[1]Варбанский К.С., [2]Городничев М.Г.
ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ ФГБОУ ВО "МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ", Москва, Россия, (111024, город Москва, Авиамоторная ул., д.8а), e-mail: [1]varbanskik@gmail.com, [2]m.g.gorodnichev@mtuci.ru

Проведен анализ и исследование подхода к оптимизации нейронных сетей с использованием разреженных-разреженных сетей. Работа направлена на выявление преимуществ и потенциала данного подхода в области машинного обучения, в частности в области больших языковых моделей и компьютерного зрения. В статье описываются подходы с использованием разреженности весов, разреженности активации, а также двойной разреженности.

Ключевые слова. Разреженность, нейронные сети, разреженность весов, разреженность активаций, комплементарная разреженность, глубокое обучение, компьютерное зрение.

ANALYSIS OF TRENDS IN THE DEVELOPMENT OF NETWORKS WITH COMPLEMENTARY SPARSITY

[1]Varbansky K.S., [2]Gorodnichev M.G.
OF THE ORDER OF THE RED BANNER OF LABOR OF THE MOSCOW TECHNICAL UNIVERSITY OF COMMUNICATIONS AND INFORMATICS, Moscow, Russia, (111024, Moscow, Aviamotornaya str., 8a), e-mail: [1]varbanskik@gmail.com, [2]m.g.gorodnichev@mtuci.ru

The analysis and research of the approach to optimization of neural networks using sparse-sparse networks is carried out. The work is aimed at identifying the advantages and potential of this approach in the field of machine learning, in particular in the field of large language models and computer vision. The article describes approaches using sparsity of weights, sparsity of activation, as well as double sparsity.

Keywords: Sparsity, neural networks, sparsity of weights, sparsity of activations, complementary sparsity, deep learning, computer vision.

Введение

В современном мире нейронные сети широко применяются в различных областях, начиная от распознавания образов до автономного управления технологическими процессами. Однако их широкое применение вызывает вопрос о вычислительной эффективности и ресурсозатратности.

В последние годы глубокие нейронные сети (*DNN*) стали больше и сложнее, что привело к значительному прогрессу в области искусственного интеллекта (*AI*). Однако экспоненциальный рост этих моделей угрожает дальнейшему развитию. Для обучения требуется большое количество процессоров, графических (*GPU*) или тензорных (*TPU*), и обучение может занимать дни или даже недели, что приводит к большому углеродному следу и растущим расходам на облачные вычисления [1].

Разреженность в нейронных сетях представляет собой перспективное направление для решения этой проблемы. Существует ряд исследований, демонстрирующих потенциал разреженных моделей в улучшении производительности и уменьшении затрат вычислительных ресурсов. Данные исследования подчеркивают важность комплементарной разреженности, включающей не только разреженность весов, но и разреженность активаций, в контексте улучшения эффективности нейронных сетей.

Разреженность

Разреженность в нейронных сетях – это концепция, которая означает, что большинство весовых или активационных параметров в сети равны нулю, тогда как некоторые значения остаются ненулевыми. В результате такой структуры большая часть параметров не участвует в вычислениях, что позволяет существенно снизить вычислительную нагрузку и использование памяти, не влияя значительно на качество работы сети. Таким образом, разреженность в нейронных сетях позволяет оптимизировать процессы обучения, выполнения и хранения, что становится все более востребованным с увеличением сложности моделей и требований к ресурсам.

Разреженность весов (*Weight sparsity*)

Существуют два основных метода достижения разреженности весов в нейронных сетях: обрезание (*pruning*) и рост (*growth*). Обрезание заключается в удалении или занулении некоторых весов в нейронной сети. Этот процесс может быть проведен как во время обучения, так и после его завершения. Обрезание ненужных весов сокращает размер модели, уменьшает требуемый объем памяти и также снижает вычислительные затраты.

Подход роста весов, напротив, заключается в увеличении количества связей нейронной сети во время обучения. Таким образом может вывялятся значимость определенных признаков, присутствующих в данных. Рост весов гипотетически может помочь уменьшить потери точности, возникающие при чрезмерном обрезании и улучшить обобщающую способность модели.

В научной сфере и в индустрии в целом обрезание весов применяется гораздо более широко, нежели рост весов; имеет значительный научный и практический фон. Метод обрезания весов сравнительно прост в реализации что делает его более привлекательным для практического использования так как этот метод может быть применен без изменения архитектуры модели и после ее обучения. Уменьшение размера нейронных сетей очень важно для их имплементации в малые устройства, такие как мобильные телефоны, часы и умные бытовые приборы.

Разреженность активаций (*Activation sparsity*)

Сеть с разреженностью активаций — это сеть, в которой алгоритм активации нейрона настроен таким образом, что в каждом слое в каждый момент времени активна лишь очень небольшой процент всех нейронов слоя.

Для имплементации разреженности активаций используются методы обучения и оптимизации, которые сводят слабые активаций к нулю. Применяются различные методы регуляризации, например методы *L1* (*Lasso*) или *L2* (*Ridge*), которые добавляют к функции

потерь модели штрафное слагаемое. Также может быть использован метод прореживания (*Dropout*), который случайным образом обнуляет некоторые активации во время обучения.

В научной сфере и в индустрии в целом принцип разреженности активаций имплементируется гораздо реже чем принцип разреженности весов.

Веса – это параметрами модели, которые оптимизируются во время обучения различными методами (например, градиентный спуск). Активации — это промежуточные выходные значения нейронов, которые не являются параметрами модели и не могут быть прямо оптимизированы во время обучения.

Разреженность весов существенно уменьшает количество вычислений, поскольку многие пропускаются. В то время как для учета разреженности активаций требуется дополнительная логика, так как активации являются результатом применения функций активации к взвешенным суммам входов. Это усложняет реализацию еще при управлении процессом обучения.

В целом практика показала, что разреженность весов приводит к существенному уменьшению размера модели и ускорению работы, что делает ее более привлекательной для широкого применения. Разреженность активаций также может улучшить производительность модели, но в меньшей степени по сравнению с разреженностью весов.

Комплементарная разреженность (*Complementary sparsity*)

В принципе, разряженные нейронные сети должны быть значительно эффективнее традиционных плотных сетей. Нейроны в мозге демонстрируют два типа разреженности: они редко связаны друг с другом и редко активны. Пирамидальные нейроны коры головного мозга обладают высокой разреженной связью друг с другом и получают относительно мало возбуждающих входов от большинства окружающих нейронов [2]. Когда эти два типа разреженности используются вместе, предполагается потенциал для снижения вычислительной сложности нейронных сетей на два порядка. Несмотря на этот потенциал, современные нейронные сети обеспечивают лишь умеренные преимущества используя в основном только разреженность весов.

Вдохновившись нейробиологией, разреженность была предложена в качестве решения проблемы быстрого роста размера моделей. Разреженные сети либо ограничивают связность или активность своих нейронов, значительно уменьшая размер и вычислительную сложность модели. Обычно эти методы применяются изолированно для создания разреженных-плотных сетей. Однако весовая и активационная разреженности являются синергетическими, и при совместном использовании вычислительная экономия умножается.

Например, когда сеть имеет 90% разреженность весов, только 1 из 10 весов не нулевой, облегчая вычисления в 10 раз. Когда сеть имеет 90% разреженность активаций, только 1 из 10 входов не нулевой, также обеспечивая уменьшение вычислительной нагрузки в 10 раз. При совместном применении нулевые значения взаимодействуют таким образом, что в среднем только 1 из 100 результатов будет не нулевым, обеспечивая кратность эффективности в 100 раз, если будут разработаны эффективные методы избежания обработки, извлечения, умножения и хранения нулевых элементов. Однако, несмотря на потенциальные преимущества данного подхода, его изучение и практическое применение пока еще остаются ограниченными.

Комплементарная разреженность — это решение, обращающее проблему разреженности.

Вместо создания аппаратных средств для поддержки неструктурированных разреженных сетей, предлагается как разреженность может быть структурирована так, чтобы соответствовать требованиям целевого аппаратного обеспечения.

Обзор литературы

Исследователи анализируют масштабируемость и компромиссы, связанные с использованием ресурсов для различных ядер, характерных для коммерческих сверточных сетей, таких как ResNet-50 или MobileNetV2. Результаты, полученные при использованиях комплементарной разреженности, показывают, что сочетание разреженности весов и активаций может быть мощным инструментом для эффективного масштабирования будущих моделей искусственного интеллекта. Применяемая реализация использует комплементарную разреженность с помощью конкурентного алгоритма *k-winner-takes-all* (k-WTA) [3].

Прямая обработка репрезентации (*representation*) разреженной матрицы неэффективна из-за наличия нулевых элементов. Техники, такие как блочная и разделенная разреженность (*block and partitioned sparsity*), помогают выравнивать структуру ненулевых элементов с аппаратными требованиями, но они фундаментально противоречат созданию точных высокоразреженных сетей. Оптимальная производительность требует больших блоков и уменьшенных размеров разбиения, но это ограничивает возможную разреженность и точность [4]. Это, в свою очередь, подрывает возможность этих подходов достичь теоретических преимуществ высокоразреженных сетей.

Альтернативный подход обращает проблему разреженности, структурируя разреженные матрицы таким образом, чтобы они были практически неотличимы от плотных матриц. Это достигается путем наложения нескольких разреженных матриц для формирования одной, плотной матрицы. Возможно оптимально соединить две разреженные матрицы в одну более плотную, если в обеих разреженных матрицах нет ненулевого элемента в одном и том же месте. Для данных поступающих активаций выполняется покомпонентное умножение (плотная операция), а затем воссоздается каждая индивидуальная сумма.

Эта техника вводит ограничения на местоположение ненулевых элементов, но не определяет их относительные положения. А также не устанавливает допустимые уровни разреженности. Техника может быть применена к сверточным ядрам (*convolutional kernels*) путем наложения нескольких трехмерных разреженных тензоров (*tensors*) из четырехмерного разреженного тензора весов слоя. Гипотетически эта техника обеспечивает линейное улучшение производительности по мере уменьшения количества ненулевых элементов, даже для очень высоких уровней разреженности.

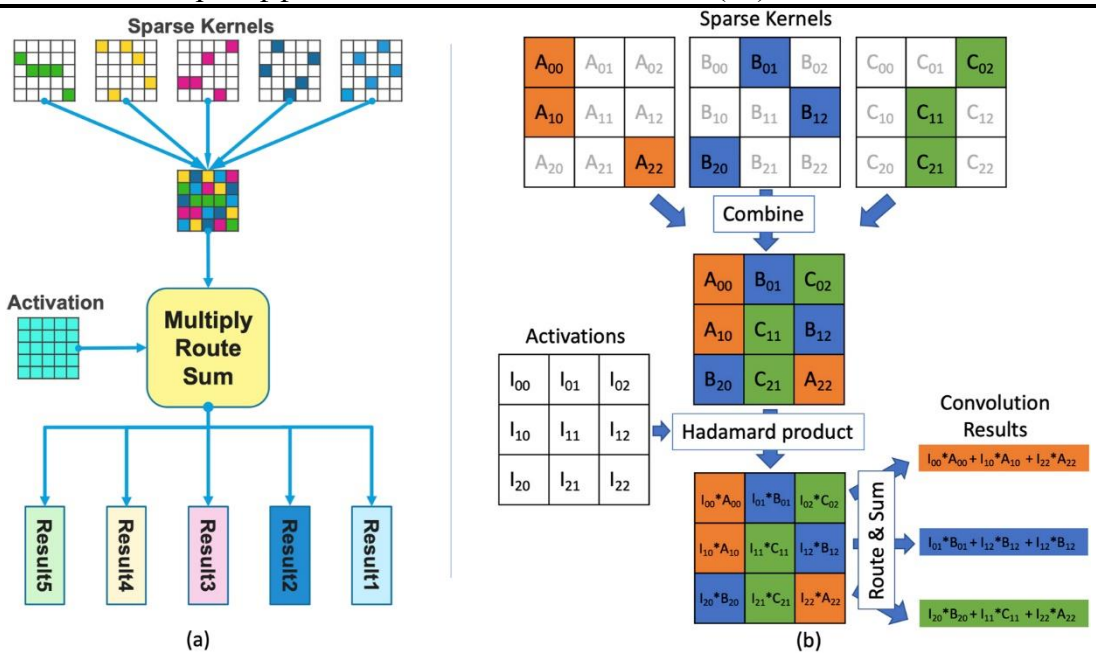

Рисунок 1. - «Упаковка» нескольких разреженных сверточных ядер в одно плотное ядро. Затем сеть маршрутизации вычисляет каждую индивидуальную сумму.
Источник: статья «Two Sparsities Are Better Than One»

Учитывая, что комплементарная разреженность сводит N разреженных сверток к одной плотной операции, существует потенциал для линейного улучшения производительности в N раз. Основная сложность заключается в снижении затрат, связанных с маршрутизацией и накоплением «упакованных» результатов. С помощью этой техники проблема разреженности-разреженности упрощается до проблемы с разреженными активациями и плотными весами, что устраняет накладные расходы.

Эксперименты были проведены на комплексной системе распознавания речи (*end to end speech recognition system*). Сверточная нейронная сеть обучалась распознавать короткие, однословные речевые команды, используя датасет *Google Speech Commands* (*GSC*). Задача — распознать произнесенное слово из аудиодорожки. Разработка предназначается для встраиваемых умных домашних предметов и платформ, реагирующих на речевые команды.

Например, одна сеть представляет собой конвейерную реализацию одной сети *GSC*, обрабатывающую один поток речевых команд на FPGA (*Field Programmable Gate Array*) на платформе *U250*. Разреженная реализация достигает более чем 33-кратного увеличения пропускной способности по сравнению с плотной реализацией [5].

Рассматриваются и другие подходы имплементации комплементарной разреженности.

Изображения высокого разрешения позволяют нейронным сетям учить более богатые визуальные репрезентации. Однако, улучшенная производительность приходит с ростом вычислительной сложности, что затрудняет их использование в приложениях, требующих низкой задержки. Не все пиксели равнозначны, пропуск вычислений для менее важных областей - эффективный метод для снижения вычислительной нагрузки. Однако это сложно преобразовать в фактическое ускорение для сверточных нейронных сетей, так как это нарушает регулярность плотной нагрузки свертки.

SparseViT пересматривает разреженность активаций для оконных видов трансформеров (*ViTs*). Возможно ускорение с помощью обрезки активаций поскольку внимание окон естественным образом группируется по блокам, в отличие от сверток. Разные слои должны иметь разные коэффициенты обрезки из-за их разнообразной чувствительности и вычислительных затрат.

Внутри изображения пиксели, содержащие детальные признаки объектов более важны, чем пиксели фона. Очень естественной идеей является применение обрезки активаций для пропуска вычислений для менее важных областей. Однако разреженность активаций не может быть легко преобразована в фактическое ускорение на универсальном оборудовании.

Имплементируя подобное обрезание на каждом слое достигается 50% сокращения задержки при 60% разреженности активаций на уровне окон.

Адаптация, осведомленная о разреженности (*Sparsity-aware adaptation*), случайным образом обрезает различные подмножества активаций на каждой итерации. Это адаптирует модель к разреженности активаций и избегает необходимости энергозатратного переобучения для нахождения оптимальных обрезаний на каждом слое.

Подобная обрезка активаций отличается от статической обрезки весов тем, что она динамическая и зависит от ввода. В то время как существующие методы обрезки активаций обычно сосредотачиваются на снижении затрат памяти во время обучения [6], лишь немногие из них нацелены на улучшение задержки вывода, так как разреженность активаций не всегда приводит к ускорению вычислений на аппаратном обеспечении.

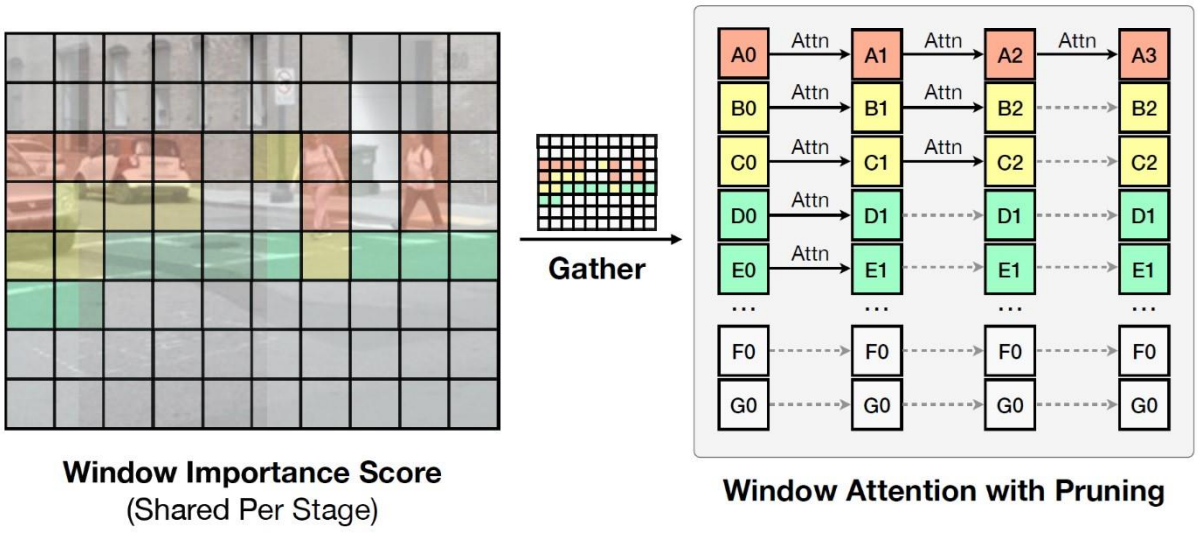

Рисунок 2. - Важность каждой активации - *L2*-нормы. Сбор признаков из окон с наивысшими оценками важности, а затем выполнение самовнимания на выбранных окнах.
Источник: статья «SparseViT»

Важность каждого окна определяется его *L2*-нормой активации. Учитывая коэффициент разреженности активации, сначала собираются окна с наивысшими оценками важности. И затем применяются механизм многоголового самовнимания (*MHSA*), сеть прямого распространения (*FFN*) и нормализация слоев (*LN*) только на этих выбранных окнах. Полученные результаты рассеиваются обратно в сеть.

В отличие от обычной обрезки весов, оценки важности зависят от ввода и должны соответственно быть вычислены во время вывода. Это может повлечь значительные накладные расходы на задержку. Поэтому вычисление важности окон выполняется только один раз на каждом этапе и повторно используется для всех блоков внутри этого этапа.

Использование одинакового уровня разреженности на всех слоях модели не эффективною по той причине, что различные слои оказывают различное воздействие на точность и эффективность. Например, начальные слои обычно требуют больше вычислений из-за их больших размеров карты признаков (*feature map*), в то время как более поздние слои больше поддаются обрезке, так как они ближе к выходу. Таким образом, более выгодно применять более активную обрезку к слоям с меньшей чувствительностью и более высокими затратами.

Для определения наилучшей конфигурации смешанной разреженности (*mixed-sparsity*) для модели критически важно оценивать ее точность при различных настройках разреженности. Однако непосредственная оценка точности исходной модели с разреженностью приведет к ненадежным результатам. Переобучение модели с каждой возможной конфигурацией разреженности перед оценкой ее точности непрактично из-за значительных временных и вычислительных затрат.

Для решения этой проблемы используется метод, основанный на осведомленности о разреженности (*Sparsity-aware adaptation*). Этот метод заключается в адаптации исходной модели, которая была обучена только с плотными активациями, путем случайной выборки слоев с разреженностью активаций.

После адаптации получаем более точную оценку производительности различных конфигураций разреженности без необходимости полного повторного обучения. Это позволяет эффективно оценивать различные конфигурации смешанной разреженности и определять оптимальную для модели.

Одним из ключевых аспектов дизайна является то, что лучше использовать ввод с высоким разрешением и более агрессивно проводить обрезание, чем начинать с ввода с низким разрешением и меньше обрезать. Начиная с высокого разрешения сохраняется детализированная информация изображения. Цель – обрезать окна, отображающие фон.

В отличие от однородных коэффициентов разреженности, применяемых ко всем слоям, *SparseViT* использует неоднородные коэффициенты разреженности для различных слоев на основе их близости к началу нейронной сети. Более маленькие размеры окон в первом и втором блоках позволяют более агрессивную обрезку, в то время как более крупные окна в более поздних слоях приводят к менее агрессивной обрезке. Этот выбор неоднородной разреженности приводит к лучшей точности.

Рисунок 3. - Цвет окна соответствует количеству слоев, в которых оно обрабатывалось.

Источник: статья «SparseViT»

Используется осведомленность о разреженности и эволюционный поиск для нахождения оптимальной конфигурации разреженности на уровне слоев в огромном пространстве поиска. Эта техника приводит к ускорению в 1.5, 1.4 и 1.3 раза по сравнению с ее плотным аналогом в монокулярном 3D-обнаружении объектов, 2D-сегментации экземпляров и 2D-семантической сегментации соответственно. Потери точности либо незначительны, либо отсутствуют в принципе [7].

Вывод

Результаты демонстрируют, что разреженность приводит к избеганию выполнения множества ненужных операций, улучшая пропускную способность и энергоэффективность. Возрос интерес к имплементации разреженности на платформах *GPU* так как, ограничения аппаратного обеспечения препятствуют развитию и внедрению разреженных сетей [8]. На сегодняшний день техники на основе *GPU* ограничены в своей способности достичь значительных приростов производительности на полных сетях. Также, они не предрасположены к использованию как сетей с разреженностью активаций, так и сетей с комплементарной разреженностью.

Список литературы

1. N. C. Thompson, K. H. Greenewald, K. Lee, and G. F. Manso. The Computational Limits of Deep Learning. CoRR, 2022. URL https://arxiv.org/abs/2007.05558
2. C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. The Journal of Physiology, 551(1):139–153, 8 2003. ISSN 0022-3751. doi:10.1113/jphysiol.2003.044784. URL http://www.jphysiol.org/cgi/doi/10.1113/jphysiol.2003.044784
3. A. Makhzani и B. Frey. Winner-take-all autoencoders. Advances in Neural Information Processing, 2015. URL http://papers.nips.cc/paper/5783-winner-take-all-autoencoders
4. F. Lagunas, E. Charlaix, V. Sanh, and A. M. Rush. Block pruning for faster transformers. CoRR, 2017. URL https://arxiv.org/abs/2109.04838

5. Kevin Hunter, Lawrence Spracklen and Subutai Ahmad. Two Sparsities Are Better Than One: Unlocking the Performance Benefits of Sparse-Sparse Networks. CoRR, 2017. URL https://arxiv.org/abs/2112.13896
6. Md Aamir Raihan, Tor M. Aamodt. Sparse Weight Activation Training. CoRR, 2020. URL https://arxiv.org/abs/2112.13896
7. Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, Song Han. SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer. CoRR, 2023. URL https://arxiv.org/abs/2303.17605.
8. S. Hooker. The Hardware Lottery, 2020. URL http://arxiv.org/abs/2009.06489.

References

1. . N. C. Thompson, K. H. Greenewald, K. Lee, and G. F. Manso. The Computational Limits of Deep Learning. CoRR, 2022. URL https://arxiv.org/abs/2007.05558
2. C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. The Journal of Physiology, 551(1):139–153, 8 2003. ISSN 0022-3751. doi:10.1113/jphysiol.2003.044784. URL http://www.jphysiol.org/cgi/doi/10.1113/jphysiol.2003.044784
3. A. Makhzani and B. Frey. Winner-take-all autoencoders. Advances in Neural Information Processing, 2015. URL http://papers.nips.cc/paper/5783-winner-take-all-autoencoders
4. F. Lagunas, E. Charlaix, V. Sanh, and A. M. Rush. Block pruning for faster transformers. CoRR, 2017. URL https://arxiv.org/abs/2109.04838
5. Kevin Hunter, Lawrence Spracklen and Subutai Ahmad. Two Sparsities Are Better Than One: Unlocking the Performance Benefits of Sparse-Sparse Networks. CoRR, 2017. URL https://arxiv.org/abs/2112.13896
6. Md Aamir Raihan, Tor M. Aamodt. Sparse Weight Activation Training. CoRR, 2020. URL https://arxiv.org/abs/2112.13896
7. Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, Song Han. SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer. CoRR, 2023. URL https://arxiv.org/abs/2303.17605.
8. S. Hooker. The Hardware Lottery, 2020. URL http://arxiv.org/abs/2009.06489.