УДК 621.391

# *ТЕНЗОРНО-МАТРИЧНАЯ ТЕОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА*

***В.И. Слюсар**, д.т.н., проф.*

*Центральный научно-исследовательский институт вооружения и военной техники Вооруженных Сил Украины*

Эффективная интеграция систем искусственного интеллекта (AI) и больших объемов данных сталкивается с проблемой поиска быстродействующих вычислительных алгоритмов, позволяющих решать задачи обработки многомерных информационных массивов в реальном масштабе времени. Особую остроту этой проблеме придает возобладавшая на данном этапе развития парадигма построения подобных систем на основе глубоких свёрточных нейросетей, предусматривающая, в частности, формирование свёрток большой размерности. Облачные механизмы доступа к таким мегаструктурам инспирировали не только массовое внедрение систем сотовой связи 5-го поколения для решения проблем сокращения времени передачи больших массивов данных, но и способствовали началу разработок систем связи 6G, в том числе на основе квантовых сетей. Вместе с тем повышение скорости передачи данных является лишь одним из аспектов концепции приближения сервиса AI к потребителю. Другой немаловажной стороной проблемы остается необходимость удешевления вычислительных технологий, лежащих в основе алгоритмов обработки данных систем искусственного интеллекта и машинного обучения.

В этой связи в последнее время усилия многих исследователей активизировались в направлении адаптации разработанной в алгебре тензорно-матричной теории под нужды систем машинного обучения в интересах компактной формализации аналитического описания алгоритмов функционирования AI.

Одним из таких направлений является использование торцевого произведения матриц, предложенного автором в 1996 г. [1]. В отношении двух матриц с одинаковым количеством строк суть его сводится к умножению каждого элемента отдельно взятой строки левой матрицы на соответствующую строку правой. Такой подход получил признание среди зарубежных специалистов, подтверждением чему явились недавние публикации Томаса Эйле (Thomas D. Ahle) с соавторами, посвященные анализу эффективности так называемого тензорного скетча [2].

Тензорный скетч как метод уменьшения размерности информационных пространств был предложен в 2013 г. и используется в алгоритмах обработки больших данных, статистике, машинном обучении для снижения размерности массивов данных на основе их векторного представления в виде тензорной структуры. Такой скетч может быть использован для ускорения билинейного объединения в нейронных сетях, уменьшения количества переменных, необходимых для реализации пулинга и является краеугольным камнем многих числовых алгоритмов линейной алгебры.

Существенно, что операцию формирования тензорного скетча можно представить в виде произведения некоторой матрицы большой размерности и тензорного произведения векторов исходных данных. В основе варианта снижения соответствующих вычислительных затрат, предложенного Томасом Эйле и др., лежит использование в отношении тензорного скетча свойства указанной операции торцевого произведения, позволяющего свести матрично-тензорное произведение к поэлементному умножению Адамара. С этой целью достаточно представить исходную матрицу тензорного скетча в виде торцевого произведения матриц меньшей размерности. В результате исходный формат скетча сводится к произведению Адамара наборов небольших матриц с элементами 1 и -1 или гауссовых матриц Джонсона-Линденштрауса и субвекторов, образовывавших первоначально тензорное произведение из набора векторов данных.

Поскольку классические операции матрично-векторных умножений выполнимы за линейное время, переход к новому формату представления тензорного скетча на основе свойства торцевого произведения позволяет выполнить умножение на векторы с тензорной структурой намного быстрее, чем формировалось бы исходное выражение. Для тензоров высокого порядка экономия в количестве операций умножения может быть весьма

значительной. При этом важно, что подобное преобразование при большом количестве матриц в составе торцевого произведения удовлетворяет лемме Джонсона-Линденштрауса о малых искажениях исходных данных большой размерности при построении их проекций.

Для свёрточных нейросетей важно, что такая концепция может быть распространена на случай формирования быстрого преобразования Фурье (БПФ) от тензорного скетча в виде векторной свёртки во временной области. В результате, переход к спектральному представлению позволяет заменить указанную свёртку эквивалентной операцией умножения торцевого произведения БПФ-матриц в комбинации с подматрицами тензорного скетча на кронекеровское произведение векторов данных. Это обеспечивает замену ресурсоёмких вычислений простым в реализации поэлементным произведением Адамара.

Томас Эйле также предложил повысить с помощью торцевого произведения производительность быстрого преобразования Джонсона-Линденштрауса (FJLT) по методу SHD. С этой целью он использовал в качестве случайной матрицы S, составленной из строк, образующих единичную матрицу, торцевое произведение двух аналогичных матриц меньшей размерности. Как следствие преобразование FJLT от тензорного произведения векторов свелось к произведению Адамара. С тем же успехом данный подход может быть распространен на другие версии FJLT – субдискретизированное рандомизированное преобразование Фурье (Subsampled Randomized Fourier Transform) и усовершенствованные модификации субдискретизированнооого рандомизированного преобразования Адамара (subsampled randomized Hadamard transform). Общая методология распараллеливания такого рода преобразований на потоки меньшей размерности состоит в том, что выборочная матрица отсчетов S представляется торцевым произведением двух матриц с меньшим количеством элементов. В итоге все исходное произведение при умножении проецирующей матрицы на тензорное произведение векторов расщепляется на произведение Адамара.

Следует отметить, что в контексте упомянутого свойства торцевого произведения идея перехода от исходного матричного проектора к произведению Адамара, оперирующему матрицами малой размерности, была использована в 2010 году для решения задачи дифференциальной приватности (*differential privacy*) при доступе к базам данных. Кроме того, аналогичные вычисления были применены для реализации ядерных методов AI.

Вместе с тем, перечисленные подходы затрагивают лишь верхушку айсберга всей совокупности возможных применений аппарата торцевых произведений матриц в качестве основы тензорно-матричной теории искусственного интеллекта.

Заслуживает внимания, например, обобщение рассмотренных вариантов решения задачи снижения размерности данных на основе использования предложенного автором блочного варианта торцевого произведения матриц (БТП) [1], поскольку такие блочные структуры являются, как известно, удобной формой представления многомерных тензоров. При определенных форматах матричных блоков применение БТП также сводится к Адамарову произведению матрично-векторных структур либо, в более общем случае, - к поблочной сумме произведений Адамара. Это позволяет преобразовывать многоканальные иерархии свёрточных нейросетей к более простым в обучении структурам.

Компактно формализовать процесс взвешивания анализируемых массивов данных на входе нейросети предлагается на основе проникающего торцевого произведения [3], которое сводится к поэлементному умножению левой матрицы на согласованные по размеру блоки правой блочной матрицы. В случае мультисвёрточных нейросетей сложной иерархии целесообразно использовать обобщённое торцевое произведение и его блочную версию [3]. Рассмотренные подходы могут быть реализованы программно или на аппаратном уровне в нейрочипах и будут способствовать внедрению AI на тактическом уровне.

1. Слюсар В.И. Торцевые произведения матриц в радиолокационных приложениях// Известия высших учебных заведений. Радиоэлектроника.- 1998. - Том 41, № 3.- С. 71 - 75.

2. Ahle, Thomas. Almost Optimal Tensor Sketch. Researchgate (3 сентября 2019).

3. Слюсар В.И. Обобщенные торцевые произведения матриц в моделях цифровых антенных решеток с неидентичными каналами.//Известия высших учебных заведений. Радиоэлектроника.- 2003. - Том 46, № 10. - С. 15 - 26.