Александр Дмитриевич ПИСАРЕВ[1]

УДК 621.382; 004.33

РЕАЛИЗАЦИЯ ДИСКРЕТНОГО КОСИНУСНОГО ПРЕОБРАЗОВАНИЯ ВО ВХОДНОМ БЛОКЕ МЕМРИСТОРНОГО НЕЙРОПРОЦЕССОРА

[1] кандидат технических наук, доцент
кафедры экспериментальной и технической физики,
заведующий лабораторией пучково-плазменных технологий
НОЦ «Нанотехнологии», Тюменский государственный университет
spcb.doc@gmail.com

Аннотация

Исследование относится к пограничной области между информационными нейросетевыми технологиями и мемристорной наноэлектроникой процессоров.

Искусственные и более сложные биоморфные нейронные сети с точки зрения информационных технологий представляют собой обучающиеся архитектуры, состоящие из большого числа простых вычислителей. Распределенность большого количества проводимых простых вычислений делает низкопроизводительными даже самые мощные стандартные процессорные системы. Поэтому в прогрессе развития нейросетей стала особо актуальной задача создания нейропроцессора.

Под нейропроцессором понимается аппаратное средство, специально разработанное для реализации модели нейронной сети эффективным образом. Крупные производители электроники (IBM, Google, Intel, Huawei) и многие другие научно-технические группы уже включились в гонку создания нейропроцессора. Задачи этого направления решаются не только с помощью отработанных кремниевых технологий, но и с применением новых элементов наноэлектроники, в том числе мемристоров.

В данной работе описана адаптация к аппаратному средству одного из вариантов быстрых алгоритмов дискретного косинусного преобразования, являющегося раз-

Цитирование: Писарев А. Д. Реализация дискретного косинусного преобразования во входном блоке мемристорного нейропроцессора / А. Д. Писарев // Вестник Тюменского государственного университета. Физико-математическое моделирование. Нефть, газ, энергетика. 2019. Том 5. № 1. С. 147-161.
DOI: 10.21684/2411-7978-2019-5-1-147-161

новидностью методов Фурье. Важность настоящего исследования заключается в необходимости решения задачи ввода стандартной информации в нейропроцессор.

В качестве аппаратного средства применяется 3D логическая матрица, реализованная на нанотехнологичных элементах электроники — комбинированных мемристорно-диодных кроссбарах. В настоящей работе представлен способ повышения скорости фильтрации за счет применения простых операций, выполняемых параллельно в логических связанных блоках сверхбольшой 3D логической матрицы. Скорость работы такой системы может быть крайне высока, и определяется она временем одного тактового импульса, ограниченного лишь скоростью срабатывания инверторных элементов и распространением сигналов по шинам комбинированного мемристороно-диодного кроссбара.

Ключевые слова

Наноэлектроника, нейронные сети, распределенные вычисления, биоморфный нейропроцессор, комбинированный кроссбар, мемристор, векторно-матричные преобразования, тензорные преобразования.

DOI: 10.21684/2411-7978-2019-5-1-147-161

Введение

Одно из перспективных направлений развития информационных технологий ориентировано на совершенствование принципов обработки данных с помощью искусственных нейронных сетей (Artificial Neural Networks, ANN). По мере увеличения глубины и сложности архитектур ANN возникла необходимость в новом аппаратном подходе. Такой подход представляется более эффективным по сравнению с распространенной программной реализацией ANN, поскольку программные проекты сильно ограничены возможностями классических (фоннеймановской и гарвардской) процессорных платформ. В первую очередь это происходит из-за большого количества распределенных и параллельных вычислений в многочисленных взаимосвязанных искусственных нейронах. В связи с этим в последнее время уделяется много внимания созданию нейропроцессора (Neural Processing Unit, NPU) — специального аппаратного средства для эффективной реализации биоморфных архитектур.

В гонке создания транзисторного нейропроцессора лидирующие позиции занимает компания IBM с проектом TrueNorth. Проект TrueNorth представил наиболее реалистичную нейронную сеть, которая лучше других ориентирована на нейроморфность [7]. Крупные конкуренты (с такими проектами, как Google TPU, Huawei Ascend 910, Intel Nervana NNP) идут по пути создания узкоспециализированных нейропроцессоров, предназначенных для тензорных вычислений, присутствующих в большом количестве в ANN.

Недостатками разработанных транзисторных нейропроцессорных модулей являются высокое энергопотребление и низкое быстродействие. Недостатки связаны с физическими ограничениями комплементарных металл-оксидных полупроводниковых (КМОП) технологий. Для преодоления этих ограничений тре-

буется качественный скачок, заключающийся в применении в схеме нейропроцессора новых микро- и наноэлектронных элементов. Такими компонентами могут быть нанотехнологические ионные мемристоры, исследованием которых в последнее время занимаются многие научно-технические группы [9].

Работа по созданию основных блоков нанотехнологического нейропроцессора, содержащего мемристоры в качестве ключевых компонент, начата в [4]. Концепция такого нейропроцессора представлена в работе [3]. Главными узлами ядра нейропроцессора являются сверхбольшие запоминающий и логический блоки, построенные на базе мемристорно-диодных [8] и КМОП-мемристорных [10] комбинированных кроссбаров. Обработка информации внутри этих блоков осуществляется несколькими вариантами, в том числе с импульсно-кодированным представлением. Функциональность этих устройств заключается в первую очередь в нейронных операциях, а также возможны функции запоминания, маршрутизации сигналов и логических преобразований [2].

Использование в большинстве блоков нейропроцессора унифицированного электронного компонента является лучшим технологическим решением. На роль такого компонента подходит 3D логическая матрица. Ее конструкция позволяет реализовывать программируемые электрические цепи, выполняющие любые логические операции. Исходя из архитектуры логической матрицы, в ней возможно реализовать логический базис из операций И-НЕ и ИЛИ-НЕ при условии, что логические данные будут подаваться в прямом и инверсном видах. Устройство позволяет выполнять операции перестановки позиций логических сигналов с помощью перестановочных матриц, что было использовано в работе [2] при маршрутизации сигналов. Подача тактовых импульсов на шины программирования и резисторов подтяжки инициирует импульсную работу электронного узла. Перечисленные функции можно использовать при построении входного блока нейропроцессора.

Одна из главных функций входного блока нейропроцессора — это кодирование информации со сжатием. Почти все современные алгоритмы сжатия с потерями (JPEG, MPEG) основаны на дискретном косинусном преобразовании (Discrete Cosine Transform, DCT), являющимся разновидностью методов Фурье-анализа [1]. Кроме сжатия входных данных метод DCT может быть использован в задачах: нахождения периодических закономерностей, кодирования, распознавания информации, удаления шумов и помех из информационных сигналов.

Данная статья посвящена адаптации дискретного косинусного преобразования к аппаратному средству, выполненному на основе 3D логической матрицы. Результаты исследования предназначены для применения в схеме входного блока нейропроцессора.

Метод дискретного косинусного преобразования

Для преобразования необходимы данные в виде отсчетов, которые обычно формируют из входных аналоговых сигналов путем дискретизации по времени и квантования или создают искусственно. В этих данных может содержаться

любая информация, в том числе яркость пикселей некоторого видеоизображения. Отсчеты группируют в виде N-мерного (обычно $N = 8$) входного вектора. Само DCT выполняется путем умножения входного вектора $\vec{X}$ на тензор преобразования $\widehat{M}$ по формуле:

$$\vec{Y} = \widehat{M} \cdot \vec{X}. \quad (1)$$

Результатом является N-мерный вектор спектра $\vec{Y}$, содержащий компоненты, соответствующие значениям амплитуд косинусных гармоник.

Наличие амплитуд с малыми значениями свидетельствуют о том, что в исходных данных отсутствует соответствующая периодичность. Процесс сжатия выполняется обнулением малых компонент спектрального вектора.

Возможен также обратный расчет выходного вектора данных по спектральным компонентам. Он выполняется с обратным тензором преобразования $\widehat{M}^{-1}$ аналогично формуле (1). Поскольку тензор преобразования $\widehat{M}$ является квадратным $N \times N$-мерным ортогональным базисом, то его определитель равен единице $|\widehat{M}| = 1$ и обратный тензор преобразования равен транспонированному $\widehat{M}^{-1} = \widehat{M}^{T}$.

Существует несколько типов DCT, самый распространенный из них — 8-мерный DCT-2_8. Компоненты тензора преобразования $M_{i,j}$ вычисляются по формуле [1]:

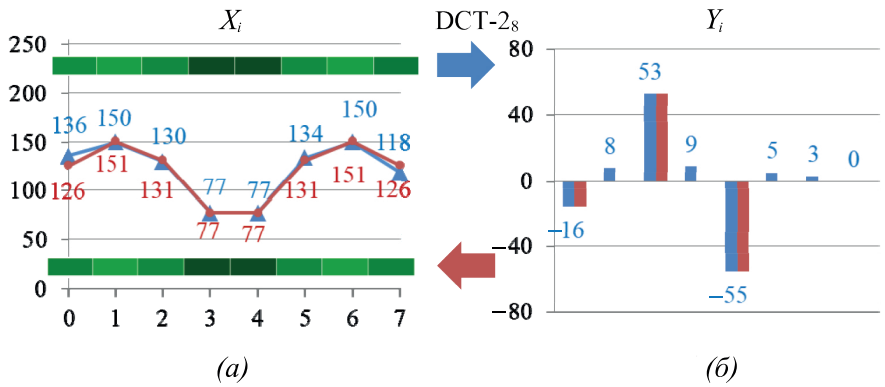
$$M_{i,j} = \lambda_i \cos\left(\frac{\pi}{2N}(2j+1)i\right), \quad \lambda_i = \begin{cases} \sqrt{\frac{1}{N}}, \text{при } i = 0, \\ \sqrt{\frac{2}{N}}, \text{при } i \neq 0, \end{cases} \quad (2)$$

где N — размер тензора (равен 8 для DCT-2_8); i и j — индексы строк и столбцов, изменяющиеся от 0 до N-1; λ_i — нормирующий множитель.

Для демонстрации принципа сжатия и поиска закономерностей во входных данных на рис. 1 показаны результаты DCT 8-битного прямого и обратного преобразований монохромных фрагментов тестового изображения. Изображения представлены на графике полосами сверху и снизу. Изображение можно преобразовывать достаточно большими участками, но обычно для сжатия его делят на фрагменты по 8 пикселей, как и в настоящем примере. Входной вектор $\vec{X}$ определяется значениями яркостей пикселей фрагмента. График компонент $\vec{X}$ показан на рис. 1а. Входной вектор состоит из восьми чисел 8-битовой разрядности (целые числа, изменяющиеся в диапазоне от 0 до 255). Понижение уровня в центре графика соответствует темной области в центре фрагмента изображения.

Спектральный вектор $\vec{Y}$, рассчитанный по формуле (1), показан на рис. 1б. Для расчета входному вектору было дано смещение −127, чтобы значения компонент находились в диапазоне от −127 до 128. На каждом этапе результат расчета округлялся до целого числа. На диаграмме заметны три компонента с наибольшим значением, которые заключают в себе основную часть энергии исходного сигнала.

Для сжатия 5 компонент с малыми амплитудами обнуляли, затем выполняли обратное преобразование. Результат преобразования сжатого спектрального



(а) *(б)*

Рис. 1. Прямое (синее) и обратное (красное) DCT-2_8: а — фрагменты тестового изображения, график зависимости яркости от номера пикселя; б — диаграмма спектральных амплитуд

Fig. 1. Direct (blue) and reverse (red) DCT-2_8: a — fragments of the test image, the graph of brightness changes from the pixel number; б — spectral amplitude diagram

вектора показан на рис. 1а. Восстановленный вектор данных по трем спектральным компонентам из восьми отражает входной сигнал, но с небольшими отклонениями рассчитанной яркости пикселей от исходных значений. Основная информация на полученном фрагменте изображения в виде темной полосы не потеряна. Полученный фрагмент изображения стал симметричным относительно своей центральной части. Потеря информации при DCT приводит к симметрии из-за разложения по гармоническому базису.

Быстрый алгоритм дискретного косинусного преобразования для входного блока нейропроцессора

Расчет DCT по классической формуле требует большого количества операций двух типов — умножения и сложения чисел в формате с плавающей точкой. Операция умножения требует во много раз больше элементарных преобразований на аппаратном средстве, чем операция сложения. Количество элементарных преобразований, выполняемых в аппаратном средстве, определяет сложность метода вычислений.

Из-за высокой сложности неадаптированного алгоритма DCT существует проблема низкой скорости работы устройств. Действительно, при компрессии видеопотоков требуется делать преобразования максимально быстро. Проблема низкой скорости преобразования решается двумя путями. Во-первых, разрабатываются специализированные процессоры и сопроцессоры, ориентированные на скоростное вычисление тензорных операций. Во-вторых, совершенствуются алгоритмы, которые уменьшают сложность расчета DCT за счет сокращения повторяющихся операций умножения и сложения чисел с плавающей точкой.

Из множества существующих быстрых алгоритмов DCT, описанных в [1], для применений во входном блоке нейропроцессора наиболее подходящим

может быть алгоритм, учитывающий связи между алгебраическими свойствами значений базисных функций DCT и структурой тензора преобразования $\widehat{M}$.

В таблице 1 представлены рассчитанные числовые значения тензора преобразования по формуле (2). Значения во многих ячейках таблицы 1 совпадают, некоторые с точностью до знака. Ячейки, имеющие одинаковые значения, выделены цветом. Значения, отличающиеся только знаком, отмечены разной яркостью. Всего получилось 7 уникальных цветов без учета знака. Таким образом, спектральные компоненты выходного вектора $\vec{Y}$ выражаются через суперпозицию 7 вариантов произведений на постоянные числа компонент входного вектора.

Таблица 1

Числовые компоненты тензора $\widehat{M}$ преобразования DCT-2_8

Table 1

Numerical values of the DCT-2_8 tensor $\widehat{M}$

i \ *j*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0,354	0,354	0,354	0,354	0,354	0,354	0,354	0,354
1	0,490	0,416	0,278	0,098	−0,098	−0,278	−0,416	−0,490
2	0,462	0,191	−0,191	−0,462	−0,462	−0,191	0,191	0,462
3	0,416	−0,098	−0,490	−0,278	0,278	0,490	0,098	−0,416
4	0,354	−0,354	−0,354	0,354	0,354	−0,354	−0,354	0,354
5	0,278	−0,490	0,098	0,416	−0,416	−0,098	0,490	−0,278
6	0,191	−0,462	0,462	−0,191	−0,191	0,462	−0,462	0,191
7	0,098	−0,278	0,416	−0,490	0,490	−0,416	0,278	−0,098

Все 7 чисел в разном порядке находятся в каждом столбце тензора преобразования (таблица 1). В соответствии с выражением (2) все уникальные значения в ячейках вычисляются по формуле:

$$M_{i,0} = \frac{1}{2}\cos\left(\frac{\pi}{16}i\right), \tag{3}$$

где i — индекс числа, изменяющийся от 1 до 7.

Используя простые тригонометрические зависимости для косинуса кратных углов ($\cos(2\varphi) = 2\cos^2(\varphi) - 1$, $\cos(3\varphi) = -3\cos(\varphi) + 4\cos^3(\varphi)$ и т. д.), можно выразить косинусы в (3) через $\cos\left(\frac{\pi}{16}\right)$. Полученные зависимости будут представлять собой многочлены Чебышёва первого рода. Если ввести обозначение $\cos\left(\frac{\pi}{16}\right) = \zeta$, то для компонент тензора DCT — чисел $M_{i,0}$ можно записать формулы:

$$
\begin{aligned}
M_{1,0} &= \frac{1}{2}\zeta, \\
M_{2,0} &= \frac{1}{2}(2\zeta^2 - 1), \\
M_{3,0} &= \frac{1}{2}(4\zeta^3 - 3\zeta), \\
M_{4,0} &= \frac{1}{2}(8\zeta^4 - 8\zeta^2 + 1), \\
M_{5,0} &= \frac{1}{2}(16\zeta^5 - 20\zeta^3 + 5\zeta), \\
M_{6,0} &= \frac{1}{2}(32\zeta^6 - 48\zeta^4 + 18\zeta^2 - 1), \\
M_{7,0} &= \frac{1}{2}(64\zeta^7 - 112\zeta^5 + 56\zeta^3 - 7\zeta).
\end{aligned} \tag{4}
$$

Многочлен Чебышёва первого рода можно рассматривать как представление числа в системе счисления с иррациональным косинусным основанием:

$$\zeta = \cos\left(\frac{\pi}{16}\right) = \frac{1}{2}\sqrt{2 + \sqrt{2 + \sqrt{2}}} \approx 0{,}980\,785, \tag{5}$$

в которой коэффициенты многочлена седьмой степени вида $f = a_0 + a_1\zeta + a_2\zeta^2 + a_3\zeta^3 + a_4\zeta^4 + a_5\zeta^5 + a_6\zeta^6 + a_7\zeta^7$ являются разрядами (цифрами) этой системы счисления. В этом случае числа тензора преобразования можно представить следующим образом:

$$
\begin{aligned}
2M_{1,0} &= .0.0.0.0.0.0.1.0_\zeta; \\
2M_{2,0} &= .0.0.0.0.0.2.0.-1_\zeta; \\
2M_{3,0} &= .0.0.0.0.4.0.-3.0_\zeta; \\
2M_{4,0} &= .0.0.0.8.0.-8.0.1_\zeta; \\
2M_{5,0} &= .0.0.16.0.-20.0.5.0_x; \\
2M_{6,0} &= .0.32.0.-48.0.18.0.-1_\zeta; \\
2M_{7,0} &= .64.0.-112.0.56.0.-7.0_\zeta.
\end{aligned} \tag{6}
$$

Точками в (6) отделены разряды числа, младший разряд записан справа. Если заменить основание системы счисления по формуле:

$$\nu = 2\,\zeta = 2\cos\left(\frac{\pi}{16}\right) = \sqrt{2 + \sqrt{2 + \sqrt{2}}} \approx 1{,}961\,570, \tag{7}$$

то можно уменьшить числовые значения разрядов и оставить их целыми:

$$
\begin{aligned}
4M_{1,0} &= .0.0.0.0.0.0.1.0_{\nu};\\
4M_{2,0} &= .0.0.0.0.0.1.0.-2_{\nu};\\
4M_{3,0} &= .0.0.0.0.1.0.-3.0_{\nu};\\
4M_{4,0} &= .0.0.0.1.0.-4.0.2_{\nu};\\
4M_{5,0} &= .0.0.1.0.-5.0.5.0_{\nu};\\
4M_{6,0} &= .0.1.0.-6.0.9.0.-2_{\nu};\\
4M_{7,0} &= .1.0.-7.0.14.0.-7.0_{\nu}.
\end{aligned}
\tag{8}
$$

Аппаратная реализация упрощается, если расчет следующего произведения производить по разрядам предыдущего, воспользовавшись рекуррентным соотношением, которое следует из (8):

$$
\begin{aligned}
&M_{1,0}(\nu) = \nu,\\
&M_{2,0}(\nu) = \nu^2 - 2,\\
&M_{i+1,0}(\nu) = \nu M_{i,0}(\nu) - M_{i-1,0}(\nu).
\end{aligned}
\tag{9}
$$

Принимая во внимание систему формул (8) и (9), в аппаратном средстве, реализующим расчеты в предложенной системе счисления, требуется выполнить операции: перестановку разрядов числа для умножения на *v* и вычитание целых чисел, являющихся разрядами сложного числа в предложенном формате счисления.

Таким образом, алгоритм вычисления DCT упрощается. Для этого требуется определить $7 \cdot 8 = 56$ уникальных произведений в предложенной системе счисления, выражающей числа с помощью многочленов Чебышёва, а затем их суммировать или вычитать в порядке, отмеченном цветами чисел в строках таблицы 1. Преимущество использования многочленов Чебышёва в расчете DCT аппаратным средством заключается в том, что операции умножения и сложения проводятся с целыми числами. С учетом рекуррентного соотношения метод DCT в аппаратном средстве можно привести к операциям перестановки, суммирования и вычитания целых чисел.

В конце всех основных вычислительных операций результат DCT можно анализировать в полученной системе счисления и без каких-либо преобразований передавать его в основное ядро нейропроцессора для следующего этапа обработки.

При необходимости можно выполнить перевод результата DCT в более информативную систему счисления, для чего понадобятся вычисления чисел в формате плавающей точки. Выполнение DCT с целью сжатия информации с потерями потребует осуществить операции над числами с плавающей точкой не со всеми спектральными компонентами, при этом допускается ограничение точности результата.

Адаптация быстрого алгоритма дискретного косинусного преобразования к входному блоку нейропроцессора

Адаптация алгоритма быстрого DCT к входному блоку нейропроцессора заключается в представлении карты коммутируемых мемристорных связей, передающих информационные импульсы, между пластами сверхбольшой 3D логической матрицы. Вследствие громоздкости и большой сложности пластов и карт связей сверхбольшой 3D-матрицы информационные потоки нагляднее всего показывать с помощью ориентированного графа. Вершинами графа являются промежуточные целочисленные значения параметров преобразования. В ребрах графа представлены направления передачи и простые операции параметров преобразования. Представленные графы показывают для аппаратного средства принципы разложения сложного DCT на простые операции с целочисленными значениями.

Для решения данной задачи можно предложить несколько вариантов графа, среди которых условно можно выделить два крайних. Первый содержит подход с экономией памяти и состоит из наименьшего количества вершин, сохраняющих переменные, но предполагает много сложновычисляемых связей. Ко второму следует отнести вариант, содержащий большое количество вершин, которые объединены простыми зависимостями. Как правило, первый подход, характеризуемый низкой скоростью работы, представляет меньший практический интерес для аппаратной реализации по сравнению со вторым, отличающимся быстродействием, но требующим большего количества вершин и ребер.

На рис. 2 предложен один из быстрых вариантов графов реализации DCT по векторно-тензорной формуле (1) в электронном устройстве на основе 3D-матрицы. Матрица во входном блоке нейропроцессора запрограммирована только на целочисленные операции сложения, вычитания и перестановки позиций чисел, представленных в системе счисления с основанием $v = 2\cos(\frac{\pi}{16})$.

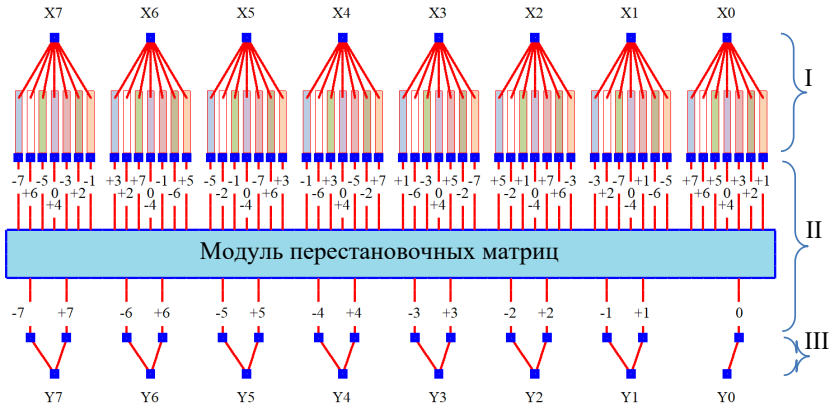


Рис. 2. Граф реализации быстрого DCT во входном блоке нейропроцессора

Fig. 2. Graph of the implementation of fast DCT in the input unit of the neural processor

Граф состоит из трех слоев, между которыми осуществляется передача информации сверху вниз. В первом слое (область I на рис. 2) показаны 8 узлов входного вектора $\vec{X}$, разделенных на $7 \cdot 8 = 56$ ребер, отвечающих за произведения компонент вектора на числа матрицы преобразования $\widehat{M}$. Во втором слое (область II на рис. 2) с помощью модуля перестановочных матриц выполняется перенаправление результатов произведения и их соответствующее суммирование. При этом положительные и отрицательные члены компонент выходного вектора $\vec{Y}$ разделяются на два потока. В третьем слое (область III на рис. 2) представлены 8 вершин значений выходного вектора $\vec{Y}$, которые получаются путем вычитания результатов перестановочной матрицы предыдущего слоя.

На рис. 3 представлен рекуррентный способ целочисленного вычисления произведений чисел тензора $\widehat{M}$ на компоненты входного вектора $\vec{X}$. Эта операция показана в первом слое (область I на рис. 2) графа реализации быстрого DCT. Способ приводит сложные вычисления к целочисленным операциям, одинаково выполняемым для всех компонент входного вектора. Способ основывается на представлении иррациональных чисел в целочисленном виде с помощью системы счисления с иррациональным косинусным основанием, выраженным формулой (7).

Представленные в (8) разряды компонент тензора преобразования показаны на рис. 3 в виде восьми зеленых линий. Ненулевые значения разрядов отмечены красным цветом. Нахождение произведений осуществляется в соответствии с рекуррентным соотношением (9), в котором каждое вычисляемое произведение определяются по двум предыдущим. Выполнение операции сдвига входной величины и вычитания разрядов предыдущего числа показаны на рис. 3 красными линиями.

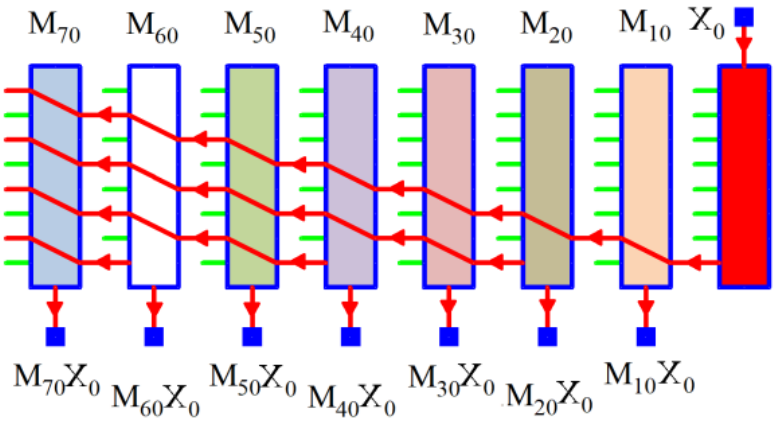


Рис. 3. Рекуррентная реализация произведений DCT в системе счисления с иррациональным косинусным основанием

Fig. 3. Recurrent implementation of DCT products in a number system with an irrational cosine base

Суммирование и вычитание произведений осуществляется во втором слое (область II на рис. 2) представленного графа. В этом модуле основной операцией являются перестановки значений произведений для подачи на сумматоры двух видов: для положительных и для отрицательных произведений, как показано в третьем слое графа (область III на рис. 2).

Таблица 2

Порядок индексов компонент векторов и тензора для использования в перестановочных операциях быстрого DCT алгоритма

Table 2

The order of the component indices of the vectors and the tensor for use in the swap operations of the fast DCT algorithm

i \ *j*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	4	4	4	4	4	4	4	4
1	1	3	5	7	−7	−5	−3	−1
2	2	6	−6	−2	−2	−6	6	2
3	3	−7	−1	−5	5	1	7	−3
4	4	−4	−4	4	4	−4	−4	4
5	5	−1	7	3	−3	−7	1	−5
6	6	−2	2	−6	−6	2	−2	6
7	7	−5	3	−1	1	−3	5	−7

Для определения порядка произведений в модуле перестановочных матриц представлена таблица 2, полученная из таблицы 1. Одинаковые произведения отмечены соответствующими цветами и цифрами в ячейке. Таблица 2 связывает три величины: номер столбца, номер строки и цифру ячейки. Номер столбца таблицы задает индекс компоненты выходного вектора $\vec{Y}$. Номер строки определяет индекс группы сумм, соответствующих компонентам входного вектора $\vec{X}$. Значение ячейки определяет номер произведения в группе. Знак ячейки задает отрицательное или положительное направление суммирования в вершинах графа, выполняемое на границе второго и первого слоев (области II и III на рис. 2). Перестановка произведений отмечена соответствующими числами, показанными на линиях связи второго слоя (область II на рис. 2).

Заключение

В данной работе представлен алгоритм преобразования DCT для реализации во входном блоке нейропроцессора. Алгоритм адаптирован к 3D логической матрице с учетом быстродействия и энергоэффективности. В быстром алгоритме

DCT выполняется последовательность простых операций: суммирования, вычитания и перестановки целых чисел. Применяется точное представление иррациональных чисел с помощью целочисленных разрядов системы счисления с иррациональным косинусным основанием. Результаты DCT представляются в выбранной системе счисления целыми разрядами.

При адаптации DCT к входному блоку нейропроцессора используется тот факт, что тензор преобразования содержит только 7 независимых компонент из 64. Реализация умножения вектора на тензор производится аналогичными модулями 3D логической матрицы. Представленное рекуррентное соотношение между 7 разными компонентами тензора преобразования позволяет значительно упростить определения произведений DCT. Возможность реализации перестановочных операций в 3D логической матрице используется для вычисления положительных и отрицательных частей компонент выходного вектора.

В настоящей работе представлен способ повышения скорости DCT в сверхбольшой 3D логической матрице за счет применения простых операций, выполняемых параллельно в логических связанных блоках. Скорость работы такой системы может быть крайне высока и определяется временем одного тактового импульса, ограниченного лишь скоростью срабатывания инверторных элементов и распространением сигналов по шинам комбинированного мемристоронодиодного кроссбара 3D логической матрицы.

3D логическая матрица позволяет добиться высокой энергоэффективности за счет распределения в пространстве формирующих сигнал элементов схемы при ключевой их работе. По сравнению с известными мемристорными устройствами, такими как аналоговый матрично-векторный умножитель Hewlett-Packard [6] и логический массив Акерса 3D [5], логическая матрица является более универсальной в применениях, при этом имея преимущества в энергоэффективности и скорости работы.

СПИСОК ЛИТЕРАТУРЫ

1. Гонсалес Р. Мир цифровой обработки. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс; пер. с англ. под ред. П. А. Чочиа. М.: Техносфера, 2005. 1072 с.
2. Писарев А. Д. SPICE-моделирование процессов ассоциативного самообучения и безусловного разобучения в логическом блоке нейропроцессора / А. Д. Писарев // Вестник Тюменского государственного университета. Физико-математическое моделирование. Нефть, газ, энергетика. 2018. Том 4. № 3. С. 132-145. DOI: 10.21684/2411-7978-2018-4-3-132-145
3. Удовиченко С. Ю. Нейропроцессор на основе комбинированного мемристорно-диодного кроссбара / С. Ю. Удовиченко, А. Д. Писарев, А. Н. Бусыгин, О. В. Маевский // Наноиндустрия. 2018. № 5 (84). С. 344-355. DOI: 10.22184/1993-8578.2018.84.5.344.355

4. Bobylev A. N. Neuromorphic coprocessor prototype based on mixed metal oxide memristors / A. N. Bobylev, A. N. Busygin, A. D. Pisarev, S. Yu. Udovichenko, V. A. Filippov // International Journal of Nanotechnology. 2017. Vol. 14. No 7/8. Pp. 698-704. DOI: 10.1504/IJNT.2017.083444
5. Levy Y. Logic operations in memory using a memristive Akers array / Y. Levy, J. Bruck, Y. Cassuto, E. G. Friedman et al. // Microelectronics Journal. 2014. Vol. 45. No 11. Pp. 1429-1437. DOI: 10.1016/j.mejo.2014.06.006
6. Li C. Analogue signal and image processing with large memristor crossbars / C. Li, M. Hu, Y. Li, H. Jiang et al. // Nature Electronics. 2018. Vol. 1. No 1. Pp. 52-59. DOI: 10.1038/s41928-017-0002-z
7. Merolla P. A. A million spiking-neuron integrated circuit with a scalable communication network and interface / P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, D. S. Modha // Science. 2014. Vol. 345. No. 6197. Pp. 668-673. DOI: 10.1126/science.1254642
8. Pisarev A. 3D memory matrix based on a composite memristor-diode crossbar for a neuromorphic processor / A. Pisarev, A. Busygin, S. Udovichenko, O. Maevsky // Microelectronic Engineering. 2018. Vol. 198. Pp. 1-7. DOI: 10.1016/j.mee.2018.06.008
9. Prezioso M. Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits / M. Prezioso, M. R. Mahmoodi, F. M. Bayat, H. Nili, H. Kim, A. Vincent // Nature Communications. 2018. Vol. 9. No 1. P. 5311. DOI: 10.1038/s41467-018-07757-y
10. Udovichenko S. 3D CMOS, memristor nanotechnology for creating logical and memory matrices of neuroprocessor / S. Udovichenko, A. Pisarev, A. Busygin, O. Maevsky // Nanoindustry. 2017. No 5. Pp. 26-34. DOI: 10.22184/1993-8578.2017.76.5.26.34

Alexander D. PISAREV[1]

UDC 621.382; 004.33

IMPLEMENTATION OF DISCRETE COSINUS TRANSFORMATION IN THE INPUT BLOCK OF THE MEMRISTOR NEURAL PROCESSOR

[1] Cand. Sci. (Tech.), Associate Professor,
Department of Applied and Technical Physics, Institute of Physics and Technology,
Head of Laboratory of Beam-Plasma Technologies,
Nanotechnologies Research and Teaching Center, University of Tyumen
spcb.doc@gmail.com

Abstract

This article describes a study on the border of neural network information technologies and memristor nanoelectronics of processors. From the point of view of information technology, artificial and more complex biomorphic neural networks are learning architectures consisting of a large number of simple solvers. The distribution of a large number of simple calculations decreases the performance consumption of even the most powerful standard processor systems. Therefore, in the development of neural networks, the task of creating a neural processor has become particularly urgent.

Neuroprocessor is understood as hardware specifically designed to implement the neural network model in an efficient manner. Major electronics manufacturers (including IBM, Google, Intel, and Huawei) have already joined the neuroprocessor creation race. The tasks of this direction require not only the developed silicon technologies, but also the use of new elements of nanoelectronics, including memristors.

This paper describes the adaptation to hardware of one of the variants of fast discrete cosine transform algorithms, which is a type of Fourier method. The importance of this study lies in the need to solve the problem of entering standard information into the neural processor.

As a hardware, a 3D logical matrix is used, implemented on nano-technological elements of the combined memristor-diode crossbar electronics. This paper presents a method for

Citation: Pisarev A. D. 2019. “Implementation of discrete cosinus transformation in the input block of the memristor neural processor”. Tyumen State University Herald. Physical and Mathematical Modeling. Oil, Gas, Energy, vol. 5, no 1, pp. 147-161.
DOI: 10.21684/2411-7978-2019-5-1-147-161

increasing the filtering rate by applying simple operations that are performed in parallel in logical connected blocks of super-large 3D logic matrix. The speed of such a system can be extremely high and is determined by the time of one clock pulse, limited only by the speed of operation of the inverter elements and the propagation of signals on the tires of the combined memristor-diode crossbar.

Keywords

Nanoelectronics, neural networks, distributed computing, biomorphic Neural Processing Unit, combined crossbar, memristor, vector-matrix transformations, tensor transformations.

DOI: 10.21684/2411-7978-2019-5-1-147-161

REFERENCES

1. Gonzalez R. C., Woods R. E. 2005. Digital Image Processing. Translated from English and edited by P. A. Chochia. Moscow: Tekhnosfera. [In Russian]
2. Pisarev A. D. 2018. "SPICE-modeling of the processes of associative self learning and unconditional discrimination in the logic unit of a neuroprocessor". Tyumen State University Herald. Physical and Mathematical Modeling. Oil, Gas, Energy, vol. 4, no 3, pp. 132-145. DOI: 10.21684/2411-7978-2018-4-3-132-145 [In Russian]
3. Udovichenko S. Yu., Pisarev A. D., Busygin A. N., Mayevsky O. V. 2018. "Neuroprocessor on the basis of a combined memristor-diode crossbar". Nanoindustry, no 5 (84), pp. 344-355. [In Russian]
4. Bobylev A. N., Busygin A. N., Pisarev A. D., Udovichenko S. Yu., Filippov V. A. 2017. "Neuromorphic coprocessor prototype based on mixed metal oxide memristors". International Journal of Nanotechnology, vol. 14, no 7/8, pp. 698-704. DOI: 10.1504/IJNT.2017.083444
5. Levy Y., Bruck J., Cassuto Y., Friedman E. G. et al. 2014. "Logic operations in memory using a memristive Akers array". Microelectronics Journal, vol. 45, no 11, pp. 1429-1437. DOI: 10.1016/j.mejo.2014.06.006
6. Li C., Hu M., Li Y., Jiang H. et al. 2018. "Analogue signal and image processing with large memristor crossbars". Nature Electronics, vol. 1, no 1, pp. 52-59. DOI: 10.1038/s41928-017-0002-z
7. Merolla P. A., Arthur J. V., Alvarez-Icaza R., Cassidy A. S., Sawada J., Akopyan F., Jackson B. L., Imam N., Guo C., Nakamura Y., Brezzo B., Vo I., Esser S. K., Appuswamy R., Taba B., Amir A., Flickner M. D., Risk W. P., Manohar R., Modha D. S. 2014. "A million spiking-neuron integrated circuit with a scalable communication network and interface". Science, vol. 345, no 6197, pp. 668-673. DOI: 10.1126/science.1254642
8. Pisarev A., Busygin A., Udovichenko S., Maevsky O. 2018. "3D memory matrix based on a composite memristor-diode crossbar for a neuromorphic processor". Microelectronic Engineering, vol. 198, pp. 1-7. DOI: 10.1016/j.mee.2018.06.008
9. Prezioso M., Mahmoodi M. R., Bayat F. M., Nili H., Kim H., Vincent A. 2018. "Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits". Nature Communications, vol. 9, no 1. P. 5311. DOI: 10.1038/s41467-018-07757-y
10. Udovichenko S., Pisarev A., Busygin A., Maevsky O. 2017. "3D CMOS, memristor nanotechnology for creating logical and memory matrices of neuroprocessor". Nanoindustry, no 5, pp. 26-34. DOI: 10.22184/1993-8578.2017.76.5.26.34