РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 732 201**(13) **C1**

(51) МПК
G06N 3/04 (2006.01)
G06N 5/04 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06N 3/04 (2020.05); G06N 5/04 (2020.05)

(21)(22) Заявка: **2020107116, 17.02.2020**

(24) Дата начала отсчета срока действия патента:
17.02.2020

Дата регистрации:
14.09.2020

Приоритет(ы):
(22) Дата подачи заявки: **17.02.2020**

(45) Опубликовано: **14.09.2020** Бюл. № **26**

Адрес для переписки:
121059, Москва, Бережковская наб., 22, стр. 3, Фонд перспективных исследований

(72) Автор(ы):
Шадрин Антон Викторович (RU),
Чуприк Анастасия Александровна (RU),
Кондратюк Екатерина Владимировна (RU),
Михеев Виталий Витальевич (RU),
Киртаев Роман Владимирович (RU),
Негров Дмитрий Владимирович (RU)

(73) Патентообладатель(и):
Российская Федерация, от имени которой выступает ФОНД ПЕРСПЕКТИВНЫХ ИССЛЕДОВАНИЙ (RU)

(56) Список документов, цитированных в отчете о поиске: **US 10430706 B2, 01.10.2019. WO 2016/186811 A1, 24.11.2016. US 2019/0041961 A1, 07.02.2019. CN 106599989 B, 09.04.2019. WO 2019/025864 A2, 07.02.2019. US 2020/0026979 A1, 23.01.2020. WO 2017/206156 A1, 07.12.2017. US 2011/0029471 A1, 03.02.2011. RU 2473126 C1, 20.01.2013. RU 2553098 C2, 10.06.2015.**

(54) Метод построения процессоров для вывода в сверточных нейронных сетях, основанный на потоковых вычислениях

(57) Реферат:

Изобретение относится к способу построения нейросетевых процессоров для вывода в свёрточных нейронных сетях и нейросетевому процессору. Технический результат заключается в повышении быстродействия работы свёрточных нейронных сетей. В способе размещают на одной интегральной схеме нейросетевого процессора вычислительные элементы, которые располагаются в узлах регулярной сетки, а также разделяемые регистровые файлы, которые содержат локально сохраненные данные, при этом каждый разделяемый регистровый файл соединяется с несколькими соседними вычислительными элементами, каждый из которых по очереди осуществляет обработку данных из разделяемых регистровых файлов, а потом хранение выходных данных вычислений в этих разделяемых регистровых файлах; при этом на одной или нескольких сторонах интегральной схемы к совокупности крайних разделяемых регистровых файлов подсоединяются располагающиеся на той же стороне контроллеры ввода-вывода. 2 н. и 8 з.п. ф-лы, 3 ил., 1 табл.

RU 2732201 C1

RU 2732201 C1

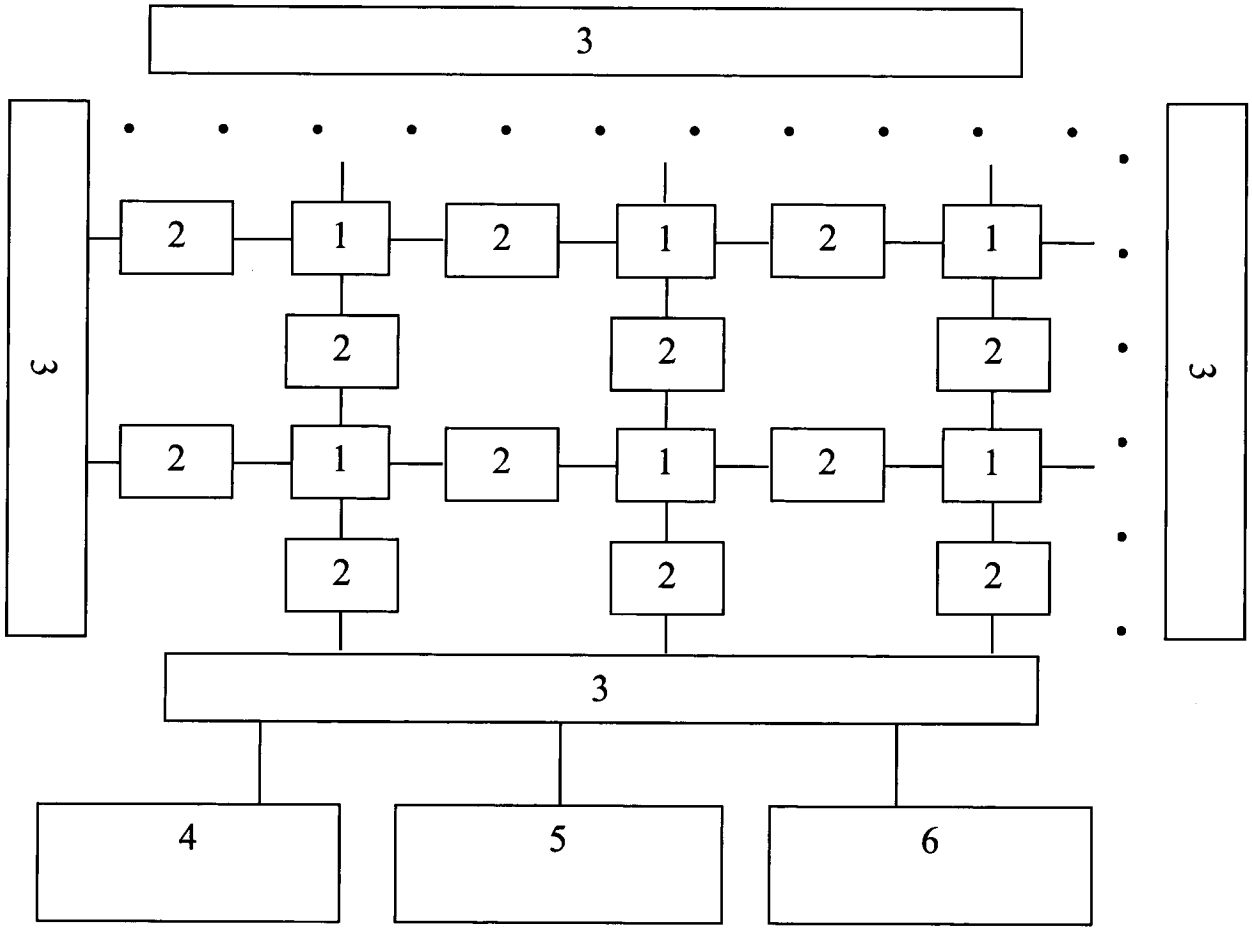


Схема структурная нейросетевого процессора

Фиг. 1

RU 2732201 C1

RU 2732201 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 732 201** (13) **C1**

(51) Int. Cl.
G06N 3/04 (2006.01)
G06N 5/04 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06N 3/04 (2020.05); G06N 5/04 (2020.05)

(21)(22) Application: **2020107116, 17.02.2020**

(24) Effective date for property rights:
17.02.2020

Registration date:
14.09.2020

Priority:
(22) Date of filing: **17.02.2020**

(45) Date of publication: **14.09.2020** Bull. № **26**

Mail address:
121059, Moskva, Berezhkovskaya nab., 22, str. 3, Fond perspektivnykh issledovanij

(72) Inventor(s):
Shadrin Anton Viktorovich (RU),
Chuprik Anastasiya Aleksandrovna (RU),
Kondratyuk Ekaterina Vladimirovna (RU),
Mikheev Vitalij Vitalevich (RU),
Kirtaev Roman Vladimirovich (RU),
Negrov Dmitrij Vladimirovich (RU)

(73) Proprietor(s):
Rossijskaya Federatsiya, ot imeni kotoroj vystupaet FOND PERSPEKTIVNYKH ISSLEDOVANIJ (RU)

(54) METHOD FOR CONSTRUCTING PROCESSORS FOR OUTPUT IN CONVOLUTIONAL NEURAL NETWORKS BASED ON DATA-FLOW COMPUTING

(57) Abstract:

FIELD: physics.

SUBSTANCE: invention relates to a method of constructing neural network processors for output in convolutional neural networks and a neural network processor. Method comprises arranging, on one integrated circuit of a neural network processor, computer elements which are located in nodes of a regular grid, as well as shared register files which contain locally stored data, wherein each shared register file is connected to several neighbouring computing elements, each of which in turn processes data from shared register files, and then storing output data of calculations in said shared register files; wherein on one or more sides of the integrated circuit to a set of extreme shared register files are connected to the same side of input-output controllers.

EFFECT: technical result is faster operation of the convolutional neural networks.

10 cl, 3 dwg, 1 tbl

RU 2732201 C1

RU 2732201 C1

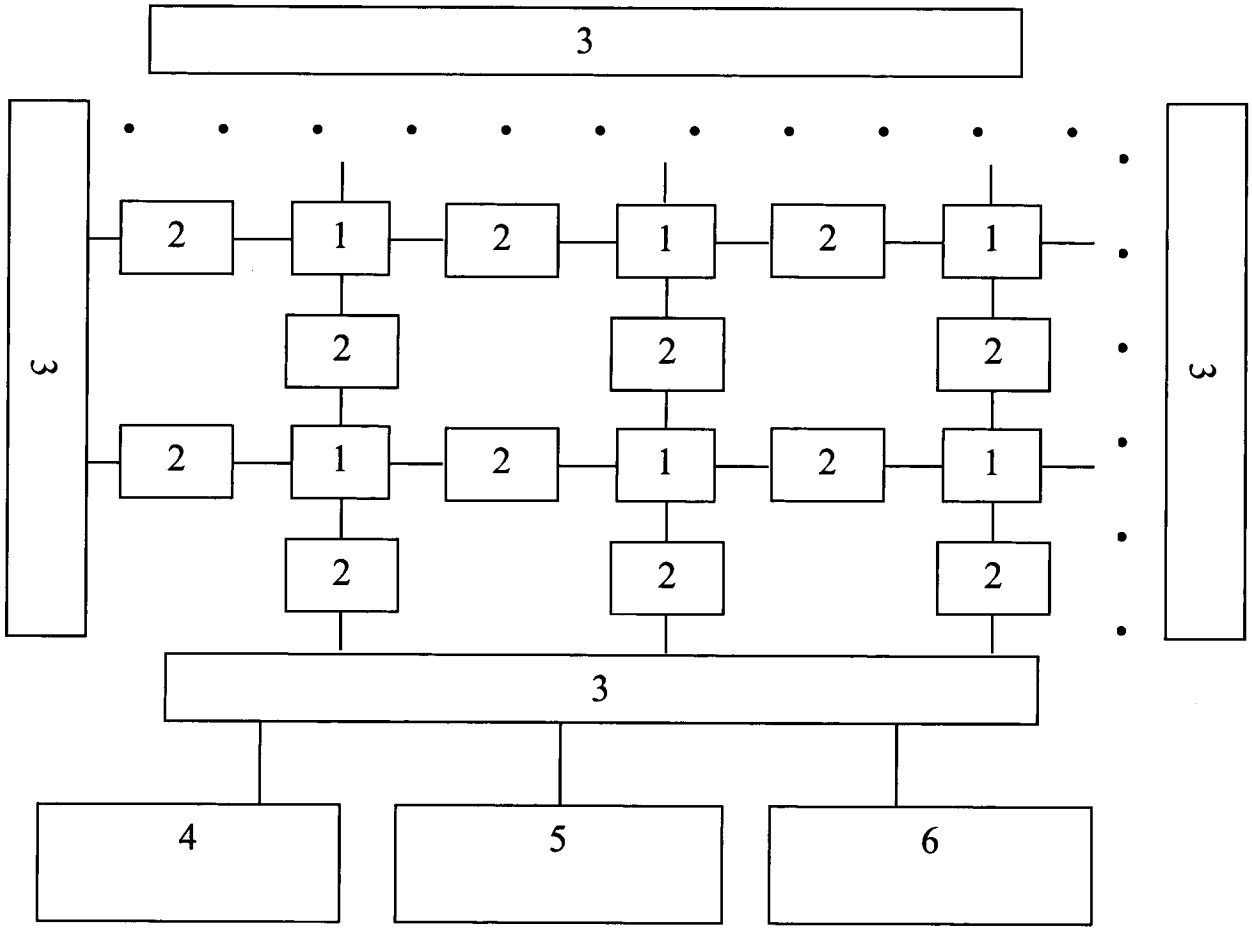


Схема структурная нейросетевого процессора

Фиг. 1

RU 2732201 C1

RU 2732201 C1

Область техники

Изобретение относится к области «Компьютерные системы, использующие модели, основанные на знаниях - способы или устройства для вывода» может быть использовано для ускорения выполнения алгоритмов статистического вывода, использующих сверточные нейронные сети, и уменьшения энергопотребления систем, использующих такие алгоритмы.

Предшествующий уровень техники

Известен ряд технических решений в виде устройств, предназначенных для ускорения исполнения алгоритмов, основанных на нейронных сетях, например, патент США № US 20110029471 A1. Недостатком этого устройства является отсутствие возможности локального переиспользования данных из-за необходимости хранения промежуточных результатов во внешнем запоминающем устройстве, что приводит к избыточному энергопотреблению. Также, структура устройства предназначена для ускорения определенного класса алгоритмов и не позволяет ускорять нейронные сети со сложной взаимосвязью между слоями, специальными слоями и непредусмотренными активационными функциями.

Наиболее близким по технической сущности к заявляемому изобретению является патент № WO 2016186811 A1, в котором предложен способ вычисления сверток, используя нейросетевой процессор, основанный на систолическом массиве, содержащий модуль прямого доступа к памяти, унифицированный буфер, выполняющий функцию промежуточного хранения нейронных активаций, модуль матричных вычислений, выполненный в виде систолического массива, блок векторных вычислений, выполняющий вычисление активационных функций и объединение результирующих активаций и систему управления, выполняющую поток поступающих извне инструкций, управляющую конфигурацией и работой всех блоков.

Недостатком данного способа является низкая программируемость, приводящая к ограничению количества различных алгоритмов, которые устройство способно выполнять с помощью предложенного способа.

Раскрытие изобретения

Задачей, решаемой в представленном изобретении, является уменьшение объема данных, передаваемых из и во внешнюю память в нейросетевом процессоре для ускорения работы сверточных нейронных сетей, а также оптимизация энергопотребления при переиспользовании локально сохраненных данных, что позволяет уменьшить энергопотребление и реализовать программируемость вычислительных элементов нейросетевого процессора. Программируемость позволяет выполнять произвольные нейросетевые алгоритмы, не ограничиваясь их конечным количеством, заданным на этапе создания нейросетевого процессора. Также, для ускорения специфичных для сверточных нейронных сетей тензорных операций, нейросетевой процессор должен содержать специальные блоки, позволяющие выполнять такие операции с высокой эффективностью.

Решение задачи достигается с помощью следующих приемов.

Способ построения нейросетевых процессоров заключается в размещении на одной интегральной схеме вычислительных элементов, причем вычислительные элементы располагаются в узлах регулярной сетки, и функционально соседние вычислительные элементы соединяются с разделяемыми регистровыми файлами, через которые осуществляется обмен локальными данными в нейросетевом процессоре, причем используется явное управление хранением и обменом локальными данными для достижения детерминированности работы: каждый вычислительный элемент по очереди

осуществляет обработку данных из разделяемых регистровых файлов, а потом хранение выходных данных вычислений в этих разделяемых регистровых файлах. Для осуществления обмена данными с внешними процессорами в целях увеличения общей вычислительной мощности и поддерживания работы вычислительной системы, состоящей из нескольких нейросетевых процессоров, на одной или нескольких сторонах интегральной схемы к совокупности крайних разделяемых регистровых файлов подсоединяются располагающиеся на той же стороне контроллеры ввода-вывода.

Для реализации описанного выше способа предлагается устройство нейросетевого процессора, содержащее вычислительные элементы, разделяемые регистровые файлы, каждый из которых соединен с несколькими соседними вычислительными элементами, а совокупность крайних разделяемых регистровых файлов на одной или нескольких сторонах интегральной схемы подсоединена к располагающимся на тех же сторонах контроллерам ввода-вывода.

Для минимизации объема управляющей логики разделяемые регистровые файлы состоят из банков, причем, к каждому банку в отдельный момент времени может иметь доступ только один вычислительный элемент.

Для технологичности проектирования и изготовления вычислительные элементы нейросетевого процессора могут быть расположены в узлах прямоугольной регулярной сетки, а каждый разделяемый регистровый файл функционально соединен с двумя соседними вычислительными элементами.

Для корректной работы нейросетевого процессора в его состав включены:

- блок обработки исключений, передающий информацию о возникших ошибочных и исключительных ситуациях всем вычислительным элементам;

- блок синхронизации, содержащий набор регистров, при этом, каждый вычислительный элемент может приостановить свои функции до возникновения определенного значения в этих регистрах или изменить эти регистры в соответствии с исполняемой нейросетевым процессором программой;

- блок реконфигурации, который позволяет изменять операции, выполняемые вычислительными элементами в процессе работы.

Для иллюстрации приведенного выше описания на фиг. 1 приведена структурная схема предлагаемого нейросетевого процессора, где позициями отмечены:

1 - вычислительный элемент (ВЭ);

2 - разделяемый регистровый файл (РРФ);

3 - контроллер ввода-вывода (КВВ);

4 - блок синхронизации;

5 - блок обработки исключений;

6 - блок реконфигурации.

Для решения задач, решаемых в предлагаемом изобретении ВЭ нейросетевого процессора содержат память инструкций, декодер инструкций, блок логики счетчика инструкций, аппаратный вычислитель тензорных сечений, арифметико-логический модуль, блоки выборки и записи данных в РРФ, а выполняемая в каждый момент инструкция определяется блоком логики счетчика инструкций, способным обрабатывать инструкции изменения текущей выполняемой инструкции - инструкции перехода.

В свою очередь арифметико-логический модуль состоит из тензорного конвейера, скалярного конвейера и системы передачи данных, при этом тензорный конвейер предназначен для выполнения операций свертки с накоплением.

Блоки выборки и записи данных в РРФ включают в себя модуль чтения регистров, модуль записи регистров, коммутатор памяти, арбитр, статусные регистры.

Отметим, что арифметико-логический модуль и блоки выборки и записи данных в РРФ реализуются способными обрабатывать одновременно несколько операндов, выполняя несколько однородных операций над данными, расположенными в РРФ.

Блок логики счетчика инструкций ВЭ нейросетевого процессора способен обрабатывать помимо инструкций перехода так же и инструкции управления циклами, которые позволяют выполнить определенный блок инструкций определенное количество раз без дополнительных накладных расходов, связанных с выполнением инструкций перехода расходов - тактов работы нейросетевого процессора, связанных с выполнением инструкций перехода.

Кроме того, ВЭ снабжен аппаратным вычислителем тензорных сечений, управляемым блоком логики счетчика инструкций, который вычисляет адрес в многомерном массиве - тензоре, на основании его размеров и текущих значений счетчиков индексов элементов тензора. Следует заметить, что элементы РРФ, адрес которых вычисляется аппаратным вычислителем тензорных сечений, проецируются в другие предопределенные программой, исполняемой нейросетевым процессором, РРФ с помощью коммутатора памяти.

Для иллюстрации приведенного выше описания ВЭ на фиг. 2 изображена структурная схема вычислительного элемента и его взаимосвязи с разделяемыми регистровыми файлами, где позициями отмечены:

7 - аппаратный вычислитель тензорных сечений (АВТС);

8 - память инструкций (ПИ);

9 - блок логики счетчика инструкций (БЛСИ);

10 - декодер инструкций (ДИ).

Блоки выборки и записи данных в РРФ:

11 - модуль чтения регистров (МЧР);

15 - арбитр;

16 - модуль записи регистра (МЗР);

17 - коммутатор памяти (КП);

18 - статусные регистры (СР).

Арифметико-логический модуль:

12 - тензорный конвейер (ТК);

13 - скалярный конвейер (СК);

14 - система передачи данных (СПД).

Следует отметить, что подсоединенные к РРФ КВВ способны передавать поступающую извне информацию ВЭ.

КВВ представляют из себя вычислительный элемент с дополнительной возможностью прямого доступа к внешней динамической памяти.

КВВ позволяют организовать прозрачную передачу данных в такой же нейросетевой процессор, установленный рядом.

Изложенное выше позволяет создать вычислительную систему, состоящую из нескольких соединенных нейросетевых процессоров для увеличения общей вычислительной мощности.

Для достижения результата предлагаемый нейросетевой процессор состоит из набора ВЭ, каждый из которых способен обмениваться данными с набором других аналогичных ВЭ. Причем обмен данными происходит за счет того, что, как правило, РРФ одновременно доступен нескольким ВЭ. Каждый ВЭ способен оперировать с данными, расположенных в РРФ, как в достаточной степени произвольное сечение тензора и способен, за счет своей программируемости, выполнить операции тензорного умножения

на произвольных сечениях. Для этого ВЭ содержит специальный ТК, как часть арифметико-логического модуля, рассчитывающий свертку части данных, а также АВТС, управляющий выборкой данных, поступающих в ТК.

Кроме того, для повышения плотности кода (англ. Code density), БЛСИ способен обрабатывать не только последовательное выполнение программы и операции перехода, но также особые команды аппаратных циклов, позволяющие выполнить определенный блок команд заданное число раз без дополнительных накладных расходов на выполнение инструкций перехода и сравнения. Кроме того, БЛСИ поддерживает вложенные циклы определенной глубины и команду преждевременного выхода из цикла, позволяющую прервать выполнение заданного количества вложенных циклов. Это позволяет увеличить плотность вычислительного кода за счет минимизации количества выполняемых вспомогательных операций (например, операций сравнения и условного перехода), не выполняющих непосредственно арифметических действий, составляющих основу нейросетевого алгоритма.

Кроме того, архитектура нейросетевого процессора позволяет управлять АВТС с помощью индексов аппаратных циклов, вычисляемых БЛСИ. Такой подход позволяет выполнять произвольные операции матричного умножения таким образом, что непосредственно ТК ВЭ выполняет только арифметические операции, а вся адресная арифметика выполняется исключительно на аппаратном уровне. Такой подход позволяет довести плотность кода до максимально теоретически возможной.

В качестве иллюстрации на фиг. 3 приведена структурная схема ТК для нахождения скалярного произведения:

19 - нулевой скаляр;

20 - входной скаляр;

21 - левый умножаемый вектор;

22 - правый умножаемый вектор;

23 - мультиплексор;

24 - умножитель;

25 - сумматор;

26 - регистр;

27 - результат.

Подробное описание.

Нейросетевой процессор (фиг. 1) состоит из совокупности набора выполненных на одной интегральной схеме ВЭ 1. ВЭ расположены в узлах регулярной сетки, и соседние ВЭ имеют общий РРФ 2, за счет чего осуществляется обмен данными в системе, причем используется явное управление хранением и обменом данными для достижения детерминированности работы. Сетка, в которой расположены ВЭ 1 выполняется предпочтительно прямоугольной, соответственно каждый РРФ 2 разделяется между двумя соседними ВЭ 1.

Предпочтительно, с каждой из четырех сторон сетки, в которой находятся ВЭ 1, установлены КВВ 3, способные передавать поступающую извне информацию в ВЭ 1. При этом между крайними ВЭ 1 и КВВ 3 находятся РРФ 2. Сами КВВ 3 представляют из себя вычислительный элемент с дополнительной возможностью прямого доступа к внешней динамической памяти. Наличие КВВ 3 позволяет организовать прозрачную передачу данных, например, в такой же нейросетевой процессор, установленный рядом. Таким образом, можно построить вычислительную систему, которая состоит из нескольких нейросетевых процессоров, объединенных при помощи КВВ 3 для увеличения общей вычислительной мощности.

Структурная схема ВЭ 1 представлен на фиг. 2. Каждый ВЭ 1 содержит память инструкций ПИ 8, ДИ 10, арифметико-логический модуль, состоящий из ТК 12, СК 13 и СПД 14, блоков выборки и записи данных в РРФ, включающих Арбитр 15, МЗР 16, МЧР 11 и КП 17, а выполняемая в каждый момент инструкция определяется БЛСИ 9, способным обрабатывать инструкции изменения текущей выполняемой инструкции (инструкции перехода). Сами РРФ 2 предпочтительно разделены на банки, причем для минимизации объема управляющей логики, к каждому банку в отдельный момент времени может иметь доступ только один ВЭ.

Арифметико-логический модуль и блоки выборки и записи данных в РРФ каждого ВЭ выполнены таким образом, что способны обрабатывать одновременно несколько операндов, выполняя несколько однородных операций над данными, расположенными в РРФ определенным образом. Сам арифметико-логический модуль состоит из ТК 12, СК 13 и СПД 14, при этом, ТК 12 предназначен для выполнения операций свертки с накоплением, структурная схема ТК для нахождения скалярного произведения показана на фиг. 3.

БЛСИ 9 каждого ВЭ 1 способен обрабатывать помимо инструкций перехода так же и инструкции управления циклами, которые позволяют выполнить определенный блок инструкций определенное количество раз без дополнительных накладных расходов, связанных с выполнением инструкций перехода.

АВТС 7 каждого ВЭ 1, управляемый БЛСИ 9, вычисляет адрес в многомерном массиве - тензоре, на основании его размеров и текущих значений счетчиков индексов элементов тензора при этом элементы РРФ, адрес которых вычисляется блоком АВТС 7 проецируются в другие предопределенные регистры с помощью КП 17.

Кроме того, нейросетевой процессор содержит блок обработки исключений 5, передающий информацию о возникших ошибочных и исключительных ситуациях всем ВЭ 1; блок синхронизации 4, содержащий набор регистров, при этом, каждый ВЭ 1 может приостановить свои функции до возникновения определенного значения в этих регистрах или изменить эти регистры по своему желанию; блок реконфигурации 6, который позволяет изменять операции, выполняемые ВЭ 1 в процессе работы.

Пример конкретного выполнения.

Устройство - нейросетевой процессор выполнен в виде кремниевой интегральной цифровой схемы, изготовленной по технологическому процессу 90 нм с низким энергопотреблением, состоящий из регулярной сетки вычислительных элементов, размером 4×4, и 32 локальных разделяемых регистровых файлов, каждый из которых разделяется между двумя вычислительными элементами и имеет объем 2048 машинных слов. Стандартное машинное слово, которым оперирует процессор, имеет размер 16 бит. Каждый БЛСИ содержит аппаратный счетчик циклов, поддерживающий до трех аппаратных циклов. Аппаратный вычислитель тензорных сечений, способный работать с произвольными размерами тензоров, может одновременно выполнять до четырех однотипных операций. Скалярный конвейер системы способен выполнять набор стандартных арифметических и логических операций (сложение, вычитание, логическое «и», логическое «или», логическое «не», логическое исключающее «или», логические сдвиги), а также операции насыщаемого сложения и вычитания над двумя операндами. Тензорный конвейер предназначен для выполнения операции скалярного произведения двух массивов, длиной, не превышающей 8 машинных слов с накоплением. Контроллеры ввода-вывода нейросетевого процессора поддерживают прозрачную передачу данных между крайними разделяемыми регистровыми файлами соседних нейросетевых процессоров. Нейросетевой процессор содержит блок обработки исключений, который

может сообщить вычислительным элементам о некорректной операции, выполненной определенным вычислительных элементом или о необходимости сброса нейросетевого процессора.

В таблице 1 представлено сравнение оцененной энергоэффективности предлагаемой реализации нейросетевого процессора и процессора Intel i7-6700K.

Таблица 1. – Сравнение энергопотребления предлагаемого выполнения процессора и процессора Intel i7-6700K.

Процессор	Выделяемая мощность, Ватт	Оценочная производительность, операций с плавающей запятой в секунду (Гфлопс)	Энергоэффективность, Гфлопс/Ватт
Intel i7-6700K	91	92	1
Предлагаемый нейросетевой процессор	2,7	256	95

Промышленная применимость

Устройство - нейросетевой процессор, может быть использован в различных применениях, накладывающих ограничение на энергопотребление, таких, как персональные мобильные устройства, робототехнические платформы, системы автономного видеонаблюдения. Кроме того, нейросетевой процессор может быть использован в системах интеллектуальной обработки большого количества данных (например, в центрах обработки данных), для увеличения удельной вычислительной мощности.

Выводы

Метод построения нейросетевых процессоров включает способ и устройство для его реализации способное выполнять большое количество тензорных операций в единицу времени за счет большого количества распределенных вычислительных элементов, содержащих специализированные составляющие для выполнения тензорных вычислений. Это позволяет значительно ускорить вычисления, специфичные для сверточных искусственных нейронных сетей. Кроме того, поскольку основной объем данных передается локально между соседними вычислительными элементами, и управляется явным образом, устройство может быть выполнено крайне энергоэффективным.

(57) Формула изобретения

1. Способ построения нейросетевых процессоров для вывода в свёрточных нейронных сетях, заключающийся в размещении на одной интегральной схеме нейросетевого процессора вычислительных элементов, которые располагаются в узлах регулярной сетки, характеризующийся тем, что на этой же интегральной схеме размещаются разделяемые регистровые файлы, которые содержат локально сохраненные данные,

при этом каждый разделяемый регистровый файл соединяется с несколькими соседними вычислительными элементами, каждый из которых по очереди осуществляет обработку данных из разделяемых регистровых файлов, а потом хранение выходных данных вычислений в этих разделяемых регистровых файлах; при этом на одной или нескольких сторонах интегральной схемы к совокупности крайних разделяемых регистровых файлов подсоединяются располагающиеся на той же стороне контроллеры ввода-вывода, что, в свою очередь, позволяет осуществлять обмен данными с внешними процессорами для увеличения общей вычислительной мощности и поддерживать работу вычислительной системы, состоящей из нескольких нейросетевых процессоров.

2. Нейросетевой процессор, построенный в соответствии со способом по п.1, содержащий вычислительные элементы, характеризующийся тем, что содержит разделяемые регистровые файлы, каждый из которых соединен с несколькими соседними вычислительными элементами, а совокупность крайних разделяемых регистровых файлов на одной или нескольких сторонах интегральной схемы подсоединена к располагающимся на той же стороне контроллерам ввода-вывода.

3. Нейросетевой процессор по п.2, характеризующийся тем, что разделяемые регистровые файлы состоят из банков, причем для минимизации объема управляющей логики к каждому банку в отдельный момент времени может иметь доступ только один вычислительный элемент.

4. Нейросетевой процессор по п.2, характеризующийся тем, что вычислительные элементы расположены в узлах прямоугольной регулярной сетки и каждый разделяемый регистровый файл соединен с двумя соседними вычислительными элементами.

5. Нейросетевой процессор по п.2, характеризующийся тем, что в состав устройства включены: блок обработки исключений, передающий информацию о возникших ошибочных и исключительных ситуациях всем вычислительным элементам; блок синхронизации, содержащий набор регистров, при этом каждый вычислительный элемент может приостановить свои функции до возникновения определенного значения в этих регистрах или изменить эти регистры в соответствии с исполняемой процессором программой; блок реконфигурации, который позволяет изменять операции, выполняемые вычислительными элементами в процессе работы.

6. Нейросетевой процессор по п.2, характеризующийся тем, что каждый вычислительный элемент содержит память инструкций, декодер инструкций, блок логики счетчика инструкций, аппаратный вычислитель тензорных сечений, арифметико-логические модули, блоки выборки и записи данных в разделяемый регистровый файл, при этом выполняемая в каждый момент инструкция определяется блоком логики счетчика инструкций, способным обрабатывать инструкции изменения текущей выполняемой инструкции - инструкции перехода.

7. Нейросетевой процессор по п.2, характеризующийся тем, что арифметико-логический модуль вычислительного элемента состоит из трех подблоков-конвейеров: тензорного конвейера, скалярного конвейера и системы передачи данных, при этом тензорный конвейер предназначен для выполнения операций свертки с накоплением.

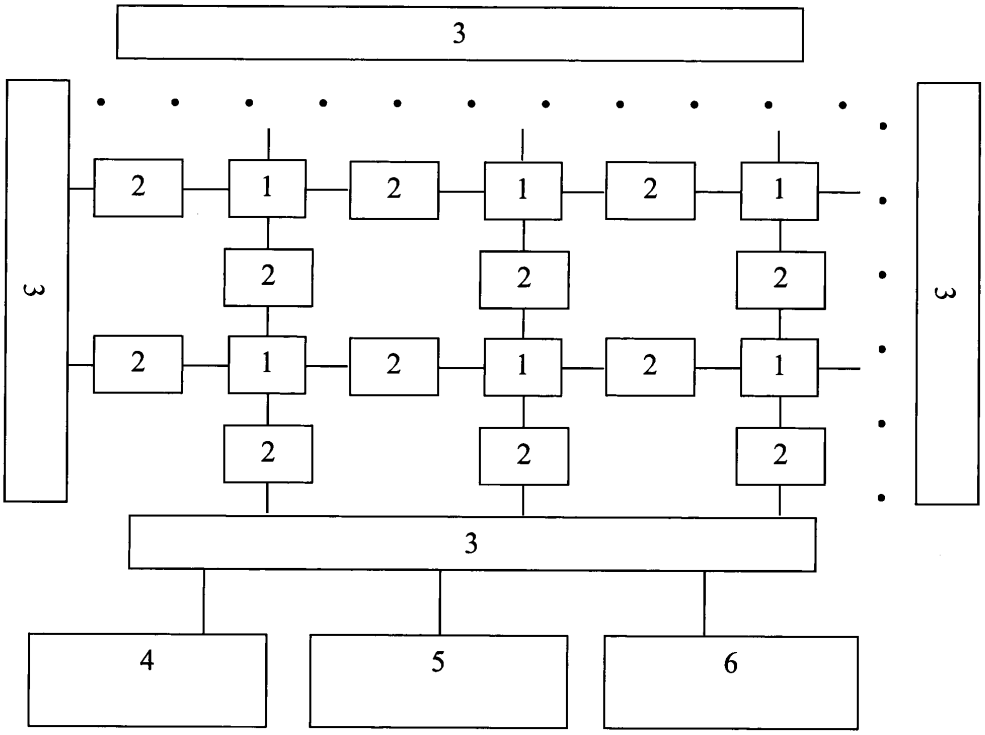
8. Нейросетевой процессор по п.2, характеризующийся тем, что блок логики счетчика инструкций вычислительного элемента способен обрабатывать помимо инструкций перехода также и инструкции управления циклами, которые позволяют выполнить определенный блок инструкций определенное количество раз без дополнительных накладных расходов - тактов работы нейросетевого процессора, связанных с выполнением инструкций перехода.

9. Нейросетевой процессор по п.2, характеризующийся тем, что его вычислительные

элементы снабжены аппаратным вычислителем тензорных сечений, управляемым блоком логики счетчика инструкций, который вычисляет адрес в многомерном массиве - тензоре, на основании его размеров и текущих значений счетчиков индексов элементов тензора, а элементы разделяемого регистрового файла, адрес которых вычисляется аппаратным вычислителем тензорных сечений, проецируются в другие предопределенные разделяемые регистровые файлы с помощью коммутатора памяти.

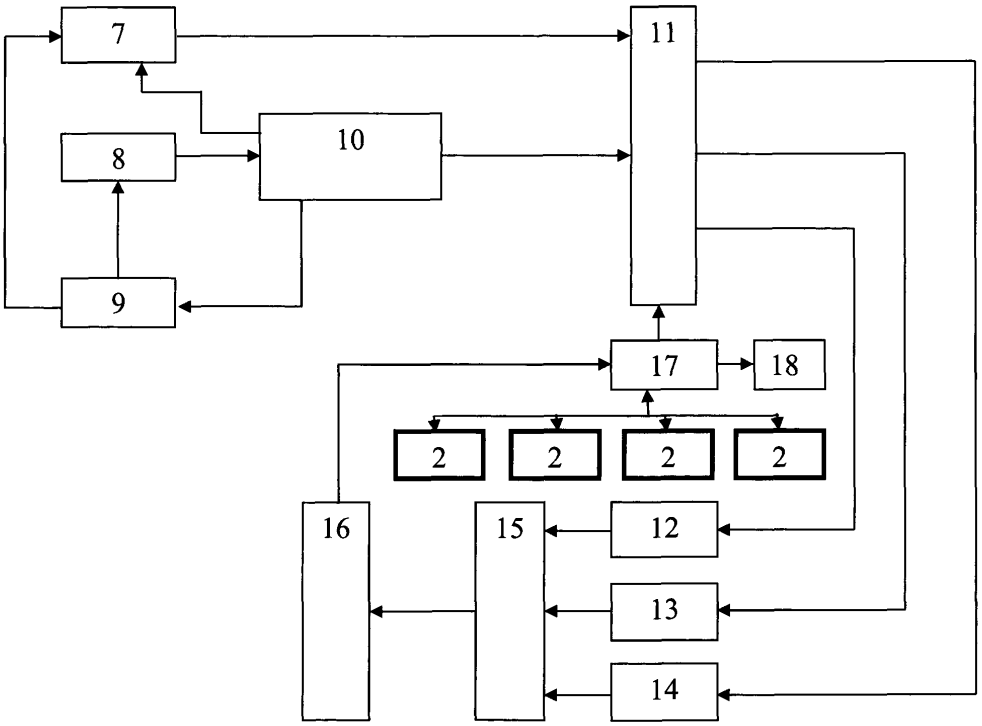
10. Нейросетевой процессор по п.2, характеризующийся тем, что контроллеры ввода-вывода позволяют организовать прозрачную передачу данных в такой же нейросетевой процессор, установленный рядом.

1



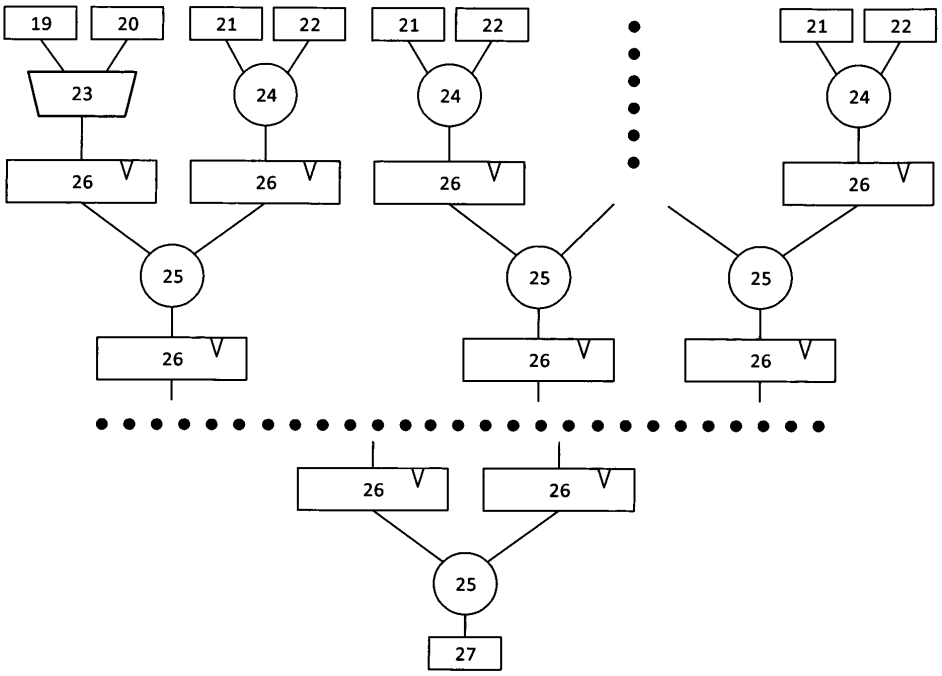
Фиг. 1. Схема структурная нейросетевого процессора

2



Фиг. 2. Схема структурная вычислительного элемента и его взаимосвязи с разделяемыми регистровыми файлами

3



Фиг. 3. Схема структурная тензорного конвейера для нахождения скалярного произведения