# Аппаратная реализация на ПЛИС свёрточных нейронных сетей для распознавания объектов на изображениях

И.В. Зоев, Н.Г. Марков,А.П. Береснев,Т.А. Ягунов
ivz3@tpu.ru | markovng@tpu.ru | apb3@tpu.ru | tay1@tpu.ru
Томский политехнический университет, г. Томск, Российская Федерация

*Рассмотрены особенности аппаратной реализации на ПЛИС, входящей в систему на кристалле Cyclone V SX, свёрточных нейронных сетей (СНС) класса LeNet 5 для распознавания объектов на изображениях. В аппаратной СНС используются вычислительные блоки двух типов: свёртки и подвыборки. Разработан метод организации вычислений в такой СНС. Приведены результаты исследования его эффективности при решении задачи распознавания рукописных цифр на изображениях.*

***Ключевые слова:*** *аппаратные свёрточные нейронные сети, программируемые логические интегральные схемы, системы компьютерного зрения, распознавание объектов на изображениях.*

# FPGA-based hardware implementation of convolution neural networks for images recognition

I.V. Zoev, N.G. Markov, A.P. Beresnev, T.A. Yagunov
ivz3@tpu.ru | markovng@tpu.ru | apb3@tpu.ru | tay1@tpu.ru
Tomsk Polytechnic University, Tomsk, Russian Federation

*This text presents specific of hardware implementation of convolution neural networks like LeNet 5 using system on chip Altera Cyclone V SX with FPGA. Hardware implementation has two types of calculating units: convolution and pooling. We developed a method of calculation organization in hardware implementation CNN. Also this text consist the results of efficiency for handwritten digit recognition task for developed method.*

***Keywords:*** *hardware convolution neural networks, field-programmable gate array, computer vision, images recognition.*

## 1. Введение

В последние годы для распознавания объектов на изображениях часто используются свёрточные нейронные сети (СНС). При создании мобильных систем компьютерного зрения (СКЗ) применяются аппаратно реализованные СНС. Учитывая высокую вычислительную сложность СНС, разработчикам таких СКЗ приходится искать баланс между точностью распознавание объектов на изображениях, производительностью и энергопотреблением таких систем. Программируемые логические интегральные схемы (ПЛИС) современных систем на кристалле имеют низкое энергопотребление, а также позволяют реализовывать параллельные вычисления. В этой связи для аппаратной реализации СНС в составе СКЗ все чаще применяют современные ПЛИС.

В работе [3] показаны результаты ускорения вычислений при реализации СНС весьма сложной архитектуры на ПЛИС с большими вычислительными ресурсами, но и со значительным энергопотреблением. На наш взгляд, при создании мобильных СКЗ можно использовать широко распространённые ПЛИС с ограниченными вычислительными ресурсами и, соответственно, с малым энергопотреблением. Так, в работе [4] приведены результаты исследования эффективности аппаратной СНС класса LeNet 5 для распознавания на изображениях рукописных цифр, при создании которой использовалась именно такая ПЛИС системы на кристалле Altera Cyclone V SX. Авторами предложена и аппаратно реализована в виде устройства архитектура СНС, основанная на хорошо известной архитектуре LeNet 5. В предложенной архитектуре 4 свёрточных слоя и 2 слоя подвыборки, а в качестве функции активации применяется оператор ReLU. В аппаратной СНС используются созданные на ПЛИС вычислительные блоки двух типов: свёртки и подвыборки. Блоки соответствующего типа выполняются параллельно внутри слоя свёртки или подвыборки. Из-за использования процедур свёртки и подвыборки с разными значениями параметров, число одновременно выполняемых блоков одного типа в разных слоях отличается и определяется конкретным слоем. Нетрудно видеть, что такой простой способ организации вычислений ведёт к простою ряда вычислительных блоков и, в итоге, к невысокой производительности устройства.

В обзоре [2] приведена схема конвейерной архитектуры вычислителя для СНС. В этой архитектуре кроме ПЛИС используется центральный процессор системы на кристалле и вычисления ведутся над числами в различных форматах. Однако в [2] отсутствуют результаты по энергопотреблению вычислителя такой архитектуры, что не позволяет оценить его пригодность в мобильных СКЗ.

Анализ этих и других подобных им работ указывает на актуальность дальнейших исследований при создании мобильных СКЗ в части используемых моделей вычислений и методов и способов реализации СНС на ПЛИС.

В данной работе предлагается метод организации вычислений в аппаратной СНС класса LeNet 5 с использованием унифицированных блоков свёртки и подвыборки. Приведены особенности реализации метода и результаты исследования его эффективности при решении задачи распознавания рукописных цифр на изображениях.

## 2. Метод организации вычислений в аппаратной СНС и особенности его реализации

Нами предлагается при организации вычислений в аппаратной СНС с архитектурой из [4], подобной LeNet 5, использовать вычислительные возможности не только ПЛИС, но и других компонентов современных систем на кристалле. Так, в системе на кристалле Altera Cyclone V SX


The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

кроме ПЛИС имеется двухъядерный процессор ARM Cortex A9, который имеет прямой доступ к внешней памяти. Архитектура рассматриваемой системы на кристалле позволяет организовать прямой доступ аппаратно реализованной СНС к этой внешней памяти. Реализация такого способа взаимодействия ПЛИС с внешней памятью процессора позволит выполнять некоторые операции, отличные от процедур свёртки и подвыборки, не на ПЛИС, а на самом процессоре.

Опираясь на этот способ, можно предложить метод организации вычислений в аппаратной СНС на ПЛИС, отличающийся от известных методов использованием унифицированных вычислительных блоков свёртки и подвыборки. Унификация блоков свёртки/подвыборки достигается путем извлечения параметров блоков, обычно задающихся на этапе их синтеза, и размещения их в отдельную изменяемую область памяти ПЛИС, называемую конфигурационной областью. Это позволит использовать блоки в слоях СНС, которые имеют разные параметры, такие как высота и ширина ядра свёртки/подвыборки, так и количество входных/выходных карт признаков.

Другой особенностью метода является то, что число задействованных в аппаратной СНС вычислительных блоков может быть переменным, причем масштабирование ведется как в целом для СНС, так и для её отдельных слоёв. Это позволит значительно сократить используемые вычислительные ресурсы ПЛИС и реализовать различные архитектуры СНС без реконфигурации ПЛИС.

При проведении унификации также необходимо:

- реализовать конфигурационную область памяти на ресурсах ПЛИС, которая будет хранить конфигурируемые параметры, необходимые не только для вычислительных блоков, но и для вспомогательных блоков аппаратной СНС;
- организовать доступ к этой области памяти процессору ARM, который в дальнейшем и будет осуществлять конфигурирование аппаратной СНС.
- создать дополнительное программное обеспечение в виде драйвера аппаратной СНС, исполняемого на процессоре ARM под управлением ядра ОС Linux, что позволит обеспечить значительную гибкость аппаратной реализации СНС.

Стоит отметить такой параметр конфигурационной области памяти ПЛИС как флаг прерывания слоя, благодаря которому по завершению вычислений в слое аппаратная СНС приостанавливает работу и даёт возможность процессору реализовать преобразование данных для выполнения операций в следующем слое сети.

Обработку таких прерываний осуществляет драйвер аппаратной СНС. Также драйвер выделает область во внешней памяти для весовых коэффициентов, входных значений СНС и промежуточных результатов её работы. Это происходит на основе передаваемых драйверу конфигурационных данных СНС, которые он в последующем записывает в конфигурационную область памяти ПЛИС. Драйвер осуществляет доступ к выделенным областям внешней памяти для аппаратной СНС и для процессора, что позволяет последнему быстро менять как входные значения сети – изображения, которые необходимо распознать, так и весовые коэффициенты СНС. Благодаря этому на ПЛИС, в принципе, можно сконфигурировать и другие классы СНС, отличные от класса LeNet 5.

Нетрудно видеть, что процессы обмена между ПЛИС и внешней памятью могут негативно сказаться на производительности разрабатываемой СКЗ из-за задержек по времени на реализацию операций ввода/вывода.

## 3. Результаты исследований эффективности метода

Для проведения исследований эффективности предлагаемого метода был создан макет СКЗ на отладочной плате Terasic SoCkit с системой на кристалле Altera Cyclone V. Архитектура СНС, аппаратно реализованная в макете, взята из работы [4]. Эта СНС относится к классу LeNet 5 и имеет следующие параметры. Первый слой - свёрточный, количество входных карт признаков 3, выходных 6, ядро свёртки 7x7, шаг 1. Второй слой - подвыборки, количество входных карт признаков 6, выходных 6, ядро подвыборки 2x2, шаг 2. Третий слой - свёрточный, количество входных карт признаков 6, выходных 32, ядро свёртки 5x5, шаг 1. Четвертый слой - подвыборки, количество входных карт признаков 32, выходных 32, ядро подвыборки 2x2, шаг 2. Пятый слой - свёрточный, количество входных карт признаков 32, выходных 100, ядро свёртки 5x5, шаг 1. Шестой (полносвязный) слой – свёртки, количество входных карт признаков 100, выходных 10, ядро свёртки 1x1, шаг 1. Весовые коэффициенты для аппаратной СНС были получены при программной реализации этой архитектуры СНС (использовалось 32-разрядное представление формата чисел с плавающей запятой), которая обучалась на выборке изображений рукописных цифр из базы MNIST размером 32x32 пикселя [1]. Точность распознавания рукописных цифр для программной реализации СНС составляет 98,71% и принята нами за эталонную точность. В макете СКЗ для сокращения необходимых вычислительных ресурсов ПЛИС не использовалась нормализация выходных данных слоёв свёртки. Это привело к необходимости применения модели вычислений в виде операций над числами с плавающей запятой.

В табл. 1 представлены результаты исследования точности распознавания рукописных цифр на изображениях из базы MNIST с помощью макета СКЗ. Разработаны несколько вариантов макета СКЗ, в которых на ПЛИС реализованы базовые операции над числами в формате с плавающей запятой различной разрядности. Результаты по точности распознавания рукописных цифр на изображениях в случаях использования 26 и 32 разрядных чисел близки к значению эталонной точности и в табл. 1 не приведены. Для сравнения второй строкой в этой таблице приведены результаты точности распознавания этих же объектов, полученные с помощью устройства из работы [4].

Таблица 1. Зависимость точности распознавания объектов от разрядности формата чисел с плавающей запятой.

| Разрядность / Устройство | 12 бит | 14 бит | 16 бит |
|---|---|---|---|
| Макет СКЗ | 97,03 % | 98,51 % | 98,66% |
| Устройство из [4] | 93,48 % | 98,27 % | 98,34% |

Видим, что макет СКЗ показал лучшую точность распознавания, чем устройство из [4] при всех разрядностях чисел формата с плавающей запятой, используемых в макете. Это связано с изменением методики переноса весовых коэффициентов СНС на ПЛИС. В работе [4] перенос осуществлялся на этапе синтеза аппаратной СНС с помощью библиотеки numpy, а в случае макета СКЗ – путем записи весов в память процессора ARM Cortex A9 при помощи библиотеки Half-precision floating point library.

Второй эксперимент был направлен на исследование производительности макета СКЗ. Далее в качестве примера приведены результаты эксперимента для макета, который


Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

работает с 16 разрядным форматом чисел с плавающей запятой.

Для последующего сравнения производительности макета СКЗ и устройства из [4], в табл. 2 приведены результаты теоретических расчетов производительности при аппаратной реализации СНС в устройстве из [4]. Расчеты проводились на основе того, что каждый вычислительный блок работает на частоте 50 МГц, что позволяет ему выполнять $50*10^6$ операций в секунду.

Таблица 2. Результаты расчетов производительности устройства из [4] при распознавании одного изображения.

| № слоя | Кол-во и тип используемых вычислительных блоков | Кол-во операций | Расчетное время выполнения, мс |
|---|---|---|---|
| 1 | 6 свёрточных | 596232 | 1,987 |
| 2 | 6 блоков подвыборки | 4056 | 0,014 |
| 3 | 32 свёрточных | 388800 | 0,243 |
| 4 | 32 блока подвыборки | 3200 | 0,002 |
| 5 | 100 свёрточных | 80000 | 0,016 |
| 6 | 10 свёрточных | 1000 | 0,002 |
| | | Итого | 2,264 |

Из табл. 2 следует неравномерность распределения количества операций по слоям СНС, поэтому вычисления в слоях выполняются с достаточно большой разницей по времени. Теоретическое значение времени работы в целом устройства при распознавании цифр на одном изображении равно 2,264 мс и близко к полученному экспериментально значению из [4], которое составляет 2,417 мс. Общее требующееся в устройстве количество вычислительных блоков свертки — 148, а подвыборки — 38. Такое количество блоков занимает практически все ресурсы ПЛИС при использовании 16 разрядного формата чисел с плавающей запятой. Очевидно, что такая конфигурация блоков для способа организации вычислений, принятого в [4], является для ПЛИС системы на кристалле Altera Cyclone V SX максимально возможной.

Для макета СКЗ также получены расчетным путем результаты его теоретической производительности. Расчет вёлся таким же способом, что и для устройства из [4], с учётом того, что вычислительные блоки макета СКЗ работают на той же частоте 50 МГц. Теоретические и экспериментальные значения производительности макета, полученные в ходе исследования при распознавании рукописных цифр на изображениях из базы MNIST, представлены в табл. 3-6 для различного числа вычислительных блоков СНС. Число блоков в эксперименте задается (масштабируется), исходя из количества выходов слоев реализуемой архитектуры СНС.

Таблица 3. Результаты исследования производительности макета СКЗ при распознавании одного изображения (количество блоков свёртки — 6, подвыборки — 6).

| № слоя | Кол-во и тип используемых вычислительных блоков | Кол-во операций | Время выполнения, мс | |
|---|---|---|---|---|
| | | | расчетное | эксперимен. |
| 1 | 6 свёрточных | 596232 | 1,987 | 4,176 |
| 2 | 6 блоков подвыборки | 4056 | 0,014 | 0,381 |
| 3 | 6 свёрточных | 388800 | 1,296 | 9,221 |
| 4 | 6 блоков подвыборки | 3200 | 0,011 | 0,362 |
| 5 | 6 свёрточных | 80000 | 0,267 | 1,981 |
| 6 | 6 свёрточных | 1000 | 0,003 | 0,343 |
| Итого | | | 3,578 | 16,464 |

Таблица 4. Результаты исследования производительности макета СКЗ при распознавании одного изображения (количество блоков свёртки — 10, подвыборки — 10).

| № слоя | Кол-во и тип используемых вычислительных блоков | Кол-во операций | Время выполнения, мс | |
|---|---|---|---|---|
| | | | расчетное | эксперимен. |
| 1 | 6 свёрточных | 596232 | 1,987 | 4,113 |
| 2 | 6 блоков подвыборки | 4056 | 0,014 | 0,344 |
| 3 | 10 свёрточных | 388800 | 0,778 | 7,633 |
| 4 | 10 блоков подвыборки | 3200 | 0,006 | 0,295 |
| 5 | 10 свёрточных | 80000 | 0,160 | 1,485 |
| 6 | 10 свертки | 1000 | 0,002 | 0,115 |
| | | Итого | 2,947 | 13,985 |

Таблица 5. Результаты исследования производительности макета СКЗ при распознавании одного изображения (количеством блоков свёртки — 32, подвыборки — 32).

| № слоя | Кол-во и тип используемых вычислительных блоков | Кол-во операций | Время выполнения, мс | |
|---|---|---|---|---|
| | | | расчетное | эксперимен. |
| 1 | 6 свёрточных | 596232 | 1,987 | 4,171 |
| 2 | 6 блоков подвыборки | 4056 | 0,014 | 0,369 |
| 3 | 32 свёрточных | 388800 | 0,243 | 0,702 |
| 4 | 32 блока подвыборки | 3200 | 0,002 | 0,262 |
| 5 | 32 свёрточных | 80000 | 0,050 | 1,311 |
| 6 | 10 свёрточных | 1000 | 0,002 | 0,118 |
| | | Итого | 2,298 | 6,933 |

Таблица 6. Результаты исследования производительности макета СКЗ при распознавании одного изображения (количеством блоков свёртки — 100, подвыборки — 32).

| № слоя | Кол-во и тип используемых вычислительных блоков | Кол-во операций | Время выполнения, мс | |
|---|---|---|---|---|
| | | | расчетное | эксперимен. |
| 1 | 6 свёрточных | 596232 | 1,987 | 4,31 |
| 2 | 6 блоков подвыборки | 4056 | 0,014 | 0,391 |
| 3 | 32 свёрточных | 388800 | 0,243 | 0,726 |
| 4 | 32 блока подвыборки | 3200 | 0,002 | 0,274 |
| 5 | 100 свёрточных | 80000 | 0,016 | 1,281 |
| 6 | 10 свёрточных | 1000 | 0,002 | 0,221 |
| | | Итого | 2,264 | 7,202 |

Из табл. 3-6 видим, что макет СКЗ по производительности уступает устройству из [4]. Однако его расчетная производительность не сильно отличается от расчетной производительности устройства из табл. 2.


The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

Основная причина различия расчетной и найденной в результате эксперимента производительности макета СКЗ заключается в наличии большого числа операций чтения данных из внешней памяти в ПЛИС. Для уменьшения таких издержек в дальнейшем следует организовать кэширование промежуточных данных и весовых коэффициентов СНС, а также организовать пакетное чтение этих данных. Это, по-видимому, позволит увеличить производительность макета СКЗ и приблизить её к расчетной производительности.

Следует также отметить возрастание как расчетной, так и экспериментально полученной производительности макета СКЗ при увеличении количества вычислительных блоков. Более того, в табл. 5 и табл. 6 значения расчетного времени выполнения распознавания объекта на одном изображении имеют небольшое различие, а также очень близки к расчетному времени вычислений для устройства из табл 2. Из-за близости расчетного времени работы макета СКЗ в табл. 5 и в табл. 6 должны быть близкие значения времени вычислений, полученные экспериментальным путем. Однако этого не происходит из-за использования динамической памяти типа DDR3 (имеет циклы обновления памяти, дающие определенную погрешность замеров).

Все варианты макета СКЗ используют меньшее количество вычислительных блоков свёртки и подвыборки по сравниванию с устройством из [4] и, соответственно, требуют меньшее количество вычислительных ресурсов ПЛИС. Анализ данных табл. 3-6 по критериям минимального времени распознавания одного изображения и минимального потребления ресурсов ПЛИС позволяет считать оптимальным для исследуемой архитектуры СНС вариант макета СКЗ с 32 блоками свертки и 32 блоками подвыборки. Кроме того, можно предположить, что после сокращения издержек при обменах ПЛИС с внешней памятью варианты макета СКЗ с числом блоков из табл. 5 (32 блока свёртки и 32 подвыборки) и с числом блоков из табл. 6 (100 блоков свёртки и 32 блока подвыборки) будут иметь производительность, близкую к производительности устройства из [4].

Полное отсутствие простоя вычислительных блоков характерно только для варианта макета СКЗ, характеристики которого приведены в табл. 3. Однако он показывает невысокую производительность из-за малого числа используемых вычислительных блоков. Для остальных вариантов макета СКЗ количество простаивающих вычислительных блоков СНС сократилось по сравнению с устройством из [4]. Чтобы решить проблему оставшихся простаивающих вычислительных блоков необходимо разработать новые алгоритмы распараллеливания вычислений.

Экспериментально установлено, что макет СКЗ и устройство из [4] потребляют одинаковое количество электроэнергии - 5,1 Вт. Невысокое энергопотребление важно при создании мобильных СКЗ.

## 4. Заключение

В настоящее время при создании мобильных СКЗ все чаще применяют аппаратно реализованные СНС. Причем реализация СНС выполняется на современных ПЛИС, имеющих низкое энергопотребление и позволяющих проводить параллельные вычисления, что очень важно при удовлетворении требований к таким СКЗ по энергопотреблению и производительности.

Предложен метод организации вычислений в аппаратной СНС на ПЛИС, использующий унифицированные вычислительные блоки свёртки и подвыборки. Метод реализован в макете СКЗ, включающем аппаратную СНС из класса LeNet 5 на основе ПЛИС системы на кристалле Altera Cyclone V SX.

Проведены исследования точности распознавания рукописных цифр на изображениях с помощью этого макета. Показано, что он дает большую точность распознавания, чем устройство из работы [4] при одинаковой разрядности форматов чисел с плавающей запятой. Однако по производительности макет уступает этому устройству. Основной причиной этого является большое число обращений ПЛИС к внешней памяти. Для решения этой проблемы указаны возможные подходы по снижению числа обращений к такой памяти. Выявлена тенденция возрастания производительности макета СКЗ при увеличении количества вычислительных блоков. Использование наряду с критерием максимальной производительности макета критерия минимального потребления ресурсов ПЛИС позволяет считать оптимальным вариант макета СКЗ с 32 блоками свертки и 32 блоками подвыборки.

Исходя из полученных результатов исследований, можно считать, что предложенный метод организации вычислений в аппаратной СНС позволяет эффективно использовать вычислительные ресурсы ПЛИС. На наш взгляд, учитывая масштабируемость унифицированных вычислительных блоков, метод может быть перспективен при организации вычислений в аппаратно реализованных СНС более сложной архитектуры, чем архитектуры нейросетей класса LeNet 5.

## 5. Благодарности

Исследования были поддержаны грантом РФФИ № 18-47-700010 р_а.

## 6. Литература

[1] The MNIST database of handwritten digits [Электронный ресурс]. – URL: http://yann.lecun.com/exdb/mnist (дата обращения 26.07.2017)

[2] Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions [Электронный ресурс]. – URL: https://arxiv.org/pdf/1803.05900.pdf (дата обращения 14.08.2018)

[3] Zhang, C. Optimizing fpga-based accelerator design for deep convolutional neural networks / C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, J. Cong //Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. – ACM, 2015. – P. 161-170. – DOI: 10.1145/2684746.2689060

[4] Зоев, И.В. Устройство на основе ПЛИС для распознавания рукописных цифр на изображениях/ И.В. Зоев, А.П. Береснев, Н.Г. Марков, А.Н. Мальчуков // Компьютерная оптика. – 2017. – Т. 41, № 6. – С. 938-949. – DOI: 10.18287/2412-6179-2017-41-6-938-949.

## Об авторах

Зоев Иван Владимирович, аспирант отделения информационных технологий Томского политехнического университета, e-mail: ivz3@tpu.ru.

Марков Николай Григорьевич, д.т.н., профессор отделения информационных Томского политехнического университета, e-mail: markovng@tpu.ru

Береснев Алексей Павлович, магистрант отделения информационных технологий Томского политехнического университета, e-mail: apb3@tpu.ru.

Ягунов Тимофей Антонович, магистрант отделения информационных технологий Томского политехнического университета, e-mail: tay1@tpu.ru.



Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18