УДК 004.312.466

IP-КОМПОНЕНТ КОНВЕЙЕРНОГО УСКОРИТЕЛЯ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ СНК СЕРИИ ZYNQ-7000

Осипов А.С., магистрант

Белорусский государственный университет информатики и радиоэлектроники[1]
г. Минск, Республика Беларусь

Петровский Н.А. – канд. техн. наук, доцент

Аннотация. В статье представлена архитектура IP-компонента, предназначенного для вычисления прямого распространения нейронных сетей произвольной структуры, включающих полносвязные и сверточные слои, а также слой подвыборки. Приведены результаты сравнительного анализа точности работы разработанного IP-компонента при различных уровнях квантования по сравнению с эталонной моделью, реализованной с использованием арифметики с плавающей точкой.

Ключевые слова. IP-компонент, сверточная нейронная сеть, аппаратная реализация, прямое распространение, квантование, FPGA

В работе рассматривается IP-ядро для аппаратной реализации вычислений в нейронных сетях прямого распространения с параметрической конфигурацией на этапе синтеза. Разработанный дизайн поддерживает вычисления в алгебре рациональных чисел и кватернионной алгебре [1], что позволяет проводить их сравнительный анализ на аппаратном уровне [2].

Конфигурация включает выбор формата данных (Q-нотация), настройку точности аккумулятора, параметров регистров и режима доступа. Архитектура сети задается последовательной записью параметров слоев в регистры CSR, что обеспечивает использование различных моделей в рамках одного ядра. Ядро поддерживает полносвязные, сверточные слои и слой подвыборки (pool-max), а также предусмотрены входной и выходной этапы обработки. Расширение функционала возможно за счет добавления подпроцессоров без изменения базовой архитектуры.

Базовой вычислительной операцией является умножение с последующим накоплением (MAC). Ядро включает управляющий и вычислительный модули: первый обеспечивает доступ к регистрам ядра через интерфейс AXI-Lite и формирует управляющие сигналы и адресацию по памяти, второй реализует операции умножения, max, функции активации и буферизацию с конвейерной обработкой. Конвейеризация повышает производительность и учитывает задержки доступа к памяти BRAM и DDR (данные доступны на следующий такт). Для согласования работы конвейера с внешней памятью используются вспомогательные интерфейсные блоки.

Интеграция в FPGA-систему выполняется через стандартные интерфейсы как показано на рисунке 1: AXI-Lite – доступ к параметрам ядра; AXI-Stream – потоковая передача данных посредством DMA; AXI3 – доступ к коэффициентам сети, расположенных в DDR памяти.

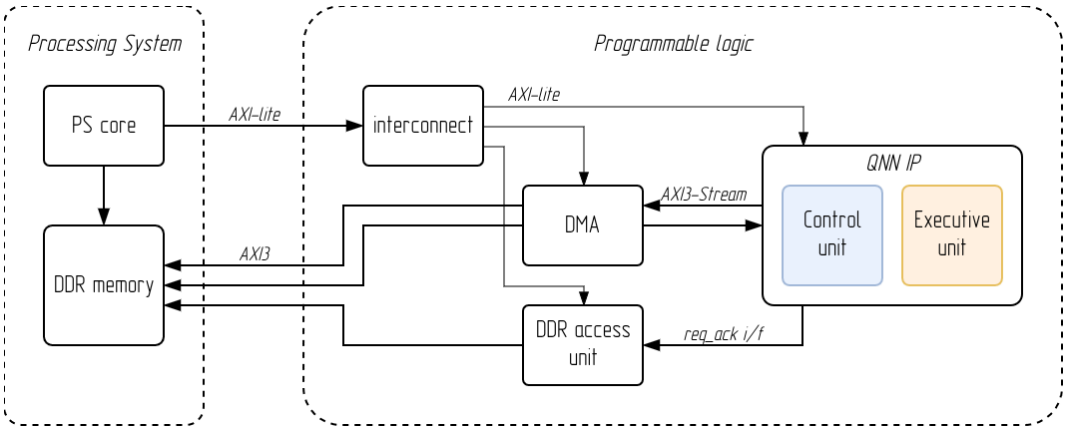


Рисунок 1 -- Структурная схема подключения ядра к процессорной системе FPGA

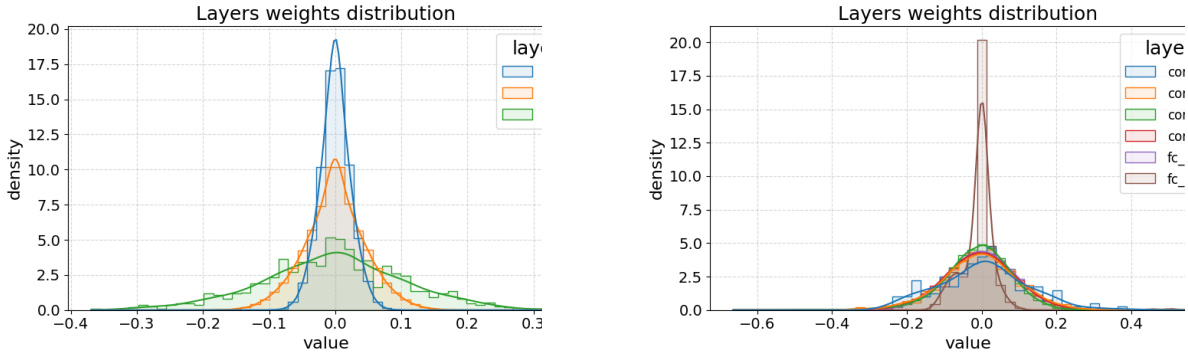
Экспериментальная проверка IP-компонента выполняется на задаче классификации с использованием набора данных CIFAR-10, состоящий из цветных изображений размером 32 на 32 пикселя. Для этих целей были обучены и квантованы в формат с фиксированной точкой две модели: полносвязная и гибридная с использованием сверточных, полносвязных и pool-max слоев [3].

[1] Работа выполнена в совместной учебной лаборатории БГУИР-YADRO
https://www.bsuir.by/ru/kaf-informatiki/yadro

В процессе обучения было принято решение отказаться от использования функции softmax на выходном слое ввиду сложности ее аппаратной реализации, что также повлияло на выбор функции потерь в пользу среднеквадратичной ошибки (MSE).

В ходе обучения моделей установлено, что при данном подходе коэффициенты нейронных сетей, решающих задачи классификации, не требуют дополнительной регуляризации, ограничивающей их отклонение от нулевых значений. Как следует из гистограмм распределения коэффициентов, представленных на рисунке 2, основная их масса сосредоточена вблизи нуля, что указывает на слабую степень использования части параметров. Исходя из этого можно сделать вывод, что наиболее предпочтительной функцией распределения коэффициентов является бимодальное с расположением локальных максимумов по обе стороны от нуля.

В то же время модели автоэнкодера [4] демонстрируют значительный разброс коэффициентов, что требует введения дополнительных штрафных условий при обучении.



а) б)

Рисунок 2 – Гистограмма плотности распределения коэффициентов моделей: а – полносвязная сеть; б – гибридная модели

В таблице 1 приведены расчетные и экспериментально полученные значения задержек прямого распространения для обеих моделей. Экспериментальная проверка была выполнена на FPGA-платформе Xilinx Zynq-7000.

Основной вклад в общую задержку вносят обращения к внешней DDR памяти: время отклика при чтении составляет порядка 15 тактов и определяется особенностями DDR-подсистемы. Снижение задержки возможно за счет использования burst-режима AXI3, обеспечивающего пакетное чтение до 16 слов, что позволит добиться суммарной задержки в 30 тактов на пакет или менее 2 тактов на слово.

На рисунке 3 представлена гистограмма сравнения точности работы полносвязной модели при различных уровнях квантования. Из графика видно, что при уменьшении количества разрядов менее девяти наблюдается лавинообразное снижение точности сети.

Разработанный IP-компонент демонстрирует возможность выполнения небольших нейронных сетей без использования затратных по аппаратным ресурсам нелинейных функций при сохранении достаточной точности, однако с необходимостью корректировки гиперпараметров при обучении. Анализ распределения коэффициентов показывает на неэффективное использование емкости представленных моделей и выделенной разрядной сетки, ввиду их близости к нулевому значению.

Дальнейшие исследования направлены на повышение эффективности квантованных моделей и более рациональное использование вычислительных ресурсов. Также IP-компонент может быть расширен за счет добавления новых типов слоев, например глубинно-разделимой свертки (depthwise separable convolution), которая позволяет снизить вычислительную сложность по сравнению с классическими сверточными слоями и более эффективно задействовать коэффициенты сети.

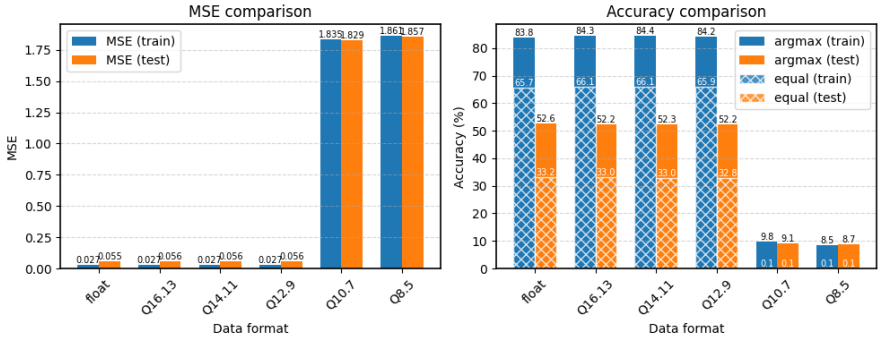


Рисунок 3 – Гистограмма ошибки и точности сети при различных уровнях квантования

Таблица 1 – Расчетные и экспериментально полученные значения задержки.

Метрики	Полносвязная модель	Гибридная модель
Количество параметров	279 043	488 979
Количество транзакций чтения AXI3	279 043	9 926 499
Количество принятых слов по интерфейсу AXI-Stream (slave)	1024	1024
Количество переданных слов по интерфейсу AXI-Stream (master)	3	3
Задержка чтения данных по интерфейсу AXI3 (сумм. / мин. / макс. / сред.), такты	4 419 890 / 15 / 74 / 15.84	156 860 568 / 15 / 79 / 15.80
Задержка приема AXI-Stream (сумм. / мин. / макс. / сред.), такты	1024 / 1 / 1 / 1	1024 / 1 / 1 / 1
Задержка передачи AXI-Stream (сумм. / мин. / макс. / сред.), такты	3 / 1 / 1 / 1	3 / 1 / 1 / 1
Расчетное время выполнения прямого распространения (f = 71,43 МГц), с	0.00392	0.14743
Фактическое время выполнения прямого распространения, с	0.07105	2.47556
Расчетная задержка, такты	280 074	10 531 119
Фактическая задержка, такты	4 420 917	156 876 059

Список использованных источников:

1. X. Zhu et al., “Quaternion convolutional neural networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9176–9184

2. Осипов А. С., Петровский Н. А. IP-компонент конвейерного ускорителя сверточных нейронных сетей для СНК серии Zynq-7000 // Информационные технологии и системы (ITS): материалы конференции. *— Минск: Belarusian State University of Informatics and Radioelectronics, 2025. — С. 149–150.*

3. Osipov, A., Petrovsky, N. FPGA Implementation of Quaternionic Fully Connected Neural Network for Image Classification // Pattern Recognition and Information Processing (PRIP’2025): Proc. of the 17th Int. Conf., 16–18 Sept. 2025, Minsk, Belarus. – Minsk : UIIP NASB, 2025. – P. 230-234

4. N. A. Petrovsky and M. I. Vashkevich, “Quaternionic multilayer bottleneck autoencoder for color image compression,” in Proceedings of SPA’2024, Available at {nick.petrovsky, vashkevich}@bsuir.by, Poznan, Poland, 2024

UDC 004.312.466

IP CORE OF A PIPELINED CONVOLUTIONAL NEURAL NETWORK ACCELERATOR FOR ZYNQ-7000 SOCS.

Osipov A.S., master’s student

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Petrovsky N.A. – PhD in Technical

Annotation. The article presents the architecture of an IP core designed to compute the forward propagation of neural networks with arbitrary structures, including fully connected and convolutional layers, as well as a subsampling layer. The results of a comparative analysis of the accuracy of the developed IP core at various quantization levels, compared with a reference model implemented using floating-point arithmetic, are also provided.

Keywords. IP core, convolutional neural network, hardware implementation, forward propagation, quantization, FPGA.