FPGA РЕАЛИЗАЦИЯ НЕЙРОННОЙ СЕТИ ПРЯМОГО РАСПРОСТРАНЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ РУКОПИСНЫХ ЧИСЕЛ

Кривальцевич Е. А., Вашкевич М. И.
Кафедра электронных вычислительных средств,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: krivalcevi4.egor@gmail.com

Представлена реализация нейронной сети для распознавания рукописных цифр на базе платформы Zynq-7000. Выполнен анализ влияния точности представления параметров НС на её производительность, а также на требуемые для реализации аппаратные ресурсы ПЛИС.

Введение

Нейронные сети (НС) играют ключевую роль в развитии информационных технологий, основанных на машинном обучении. Вычислительной платформой для обучения и эксплуатации нейросетевых моделей чаще всего выступают графические процессоры (GPU), которые содержат множество вычислительных ядер, способных обрабатывать потоки данных параллельно. К недостаткам GPU можно отнести их высокую потребляемую мощность, а также универсальность их архитектуры.

Программируемые логические интегральные схемы (ПЛИС) типа FPGA (Field Programmable Gate Array) представляют собой реконфигурируемые вычислительные платформы, позволяющие реализовывать параллельно-поточные архитектуры НС [1-3] с более высокой производительностью и меньшим потреблением энергии по сравнению с реализациями на базе процессоров общего назначения (CPU) и графических процессоров.

При реализации НС на базе CPU и GPU как правило используются стандартизированные типы данных (чаще всего числа с плавающей запятой одинарной точности, реже – целочисленные типы). При реализации на базе FPGA появляется возможность использовать для представления параметров НС типов данных, обеспечивающих различную точность. Причем выбор точности представления напрямую будет влиять на аппаратные затраты. В настоящей работе на примере однослойной НС для распознавания рукописных цифр исследуется влияние разрядности коэффициентов НС на точность распознавания, а также на аппаратные затраты FPGA, необходимые для реализации НС.

I. Разработка нейронной сети

В работе рассматривается задача распознавания рукописных цифр по изображениям из набора данных MNIST. Для проведения эксперимента была выбрана однослойная НС прямого распространения, состоящая из полносвязного слоя с выходной функцией активации softmax.

Используемый набор данных MNSIT содержит 70 тыс. полутоновых изображений размера 28x28 пикселей рукописных цифр от 0 до 9. Набор разбит на две части: тренировочная выборка – 60 тыс. изображений, а тестовая выборка – 10 тыс. Обучение НС выполнялось с использованием языка Python и библиотеки PyTorch. В процессе обучения НС использовался метод стохастического градиентного спуска со скоростью обучения $\eta = 0{,}003$ и моментумом $\gamma = 0{,}9$. Обучение выполнялось на 10 тыс. эпох, что позволило модели достичь высокой точности на тренировочном наборе данных. В результате обучения была получена матрица весовых коэффициентов размером 10×784.

II. Реализация НС на FPGA

Для реализации НС была выбрана отладочная плата Zybo на базе ПЛИС Zynq-7000. Zynq – это система на кристалле (СнК), которая объединяет процессор ARM и программируемую логику FPGA. Для упрощения разработки и тестирования на этой платформе используется дистрибутив Linux PYNQ (Python productivitY for zyNQ). PYNQ позволяет с использованием языка Python взаимодействовать с аппаратными блоками FPGA, реализованными в виде IP-ядер, что делает процесс тестирования и разработки более гибким и удобным.

Структура разработанной системы для распознавания рукописных цифр на базе платформы Zynq представлена на рисунке 3. Для подачи изображения в НС оно предварительно считывается, затем оно преобразуется в необходимый формат и передаётся в IP-блок по интерфейсу AXI-Lite.

IP-блок, реализирующий НС, описан на языке SystemVerilog. Полносвязный слой реализуется на базе десяти MAC-ядер. Каждое ядро производит 784 операции умножения значения пикселя на соответствующий весовой коэффициент, хро-

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

нящийся в памяти устройства. В результате получается массив из десяти элементов, представляющий выходные данные слоя. Затем в блоке Max ind осуществляется выбор наибольшего элемента массива и вывод его индекса. Найденное значение передаётся обратно в процессорную систему по интерфейсу AXI-Lite.

III. Тестирование НС

На этапе тестирования исследовалось влияния разрядности весовых коэффициентов на точность распознавания цифр, а также на аппаратные затраты FPGA. Разрядность коэффициентов НС изменялась от 2 до 16 бит. Для каждой разрядности производилась подача на НС всех 10 тыс. тестовых изображений базы MNSIT. Для анализа полученных результатов выполнялось построение матрицы спутывания. На рисунке 1 представлен пример матрицы спутывания.

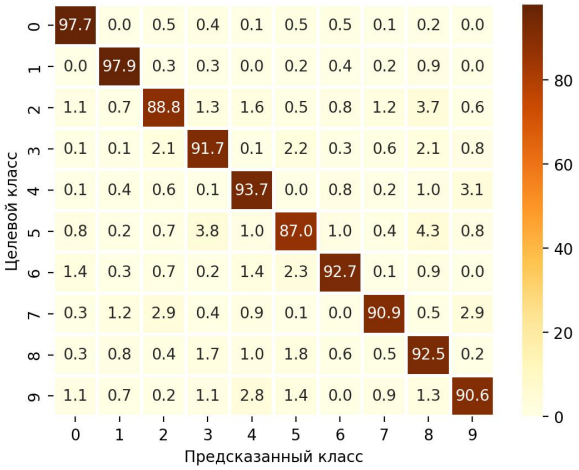

Рис. 1 – Матрица спутывания для 5-разрядного представления весов НС

Анализ аппаратных затрат при различной разрядности весовых коэффициентов НС показал, что при уменьшении разрядности уменьшается число требуемых для реализации НС блоков LUT и FF (триггеров). Полученные результаты экспериментов представлены на рисунке 2, где на одном графике совмещены точность распознавания и количество использованных элементов LUT и FF в зависимости от разрядности коэффициентов НС.

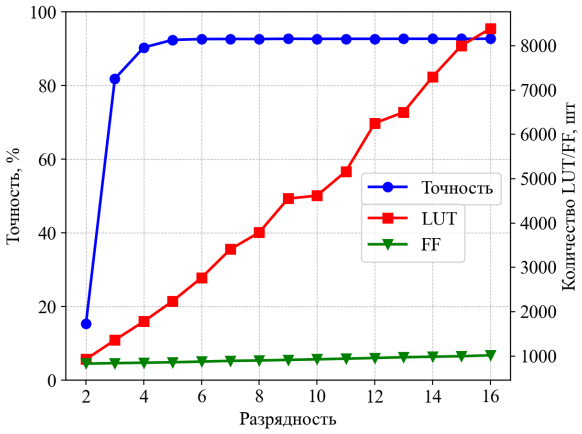

Рис. 2 – Точность и аппаратные затраты на реализацию НС

В соответствии с полученными результатами можно сделать вывод, что наиболее оптимальной разрядность будет 5 бит, которая позволяет с высокой вероятностью правильно распознать цифры на изображении и не использовать избыточные аппаратные ресурсы FPGA.

IV. Заключение

В работе предложен вариант реализации устройства распознавания рукописных цифр на базе платформы Zybo. Исследовано влияние разрядности весовых коэффициентов НС на точность распознавания и аппаратные затраты FPGA.

1. Mittal S. A survey of FPGA-based accelerators for convolutional neural networks // Neural computing and applications. – 2020. – V. 32. – №. 4. – С. 1109-1139.
2. Ahmad A., Pasha M. A. FFConv: an FPGA-based accelerator for fast convolution layers in convolutional neural networks // ACM Transactions on Embedded Computing Systems (TECS). – 2020. – V. 19. – №. 2. – С. 1-24.
3. Giardino D. et al. FPGA implementation of hand-written number recognition based on CNN // International Journal on Advanced Science, Engineering and Information Technology. – 2019. – V. 9. – №. 1. – С. 167-171.

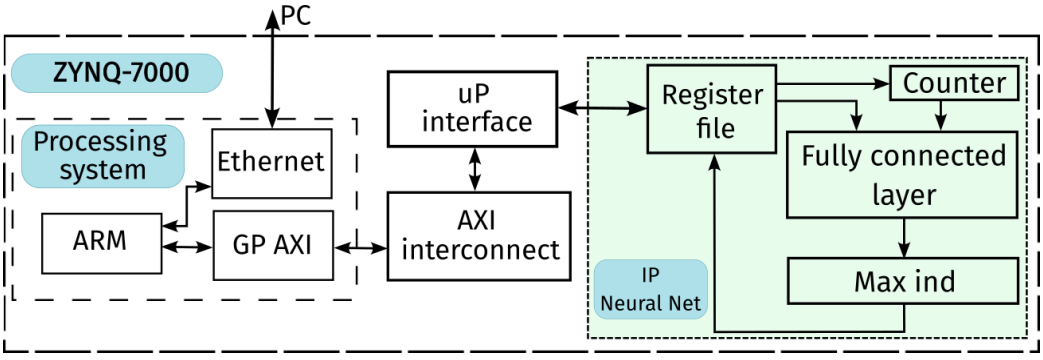

Рис. 3 – Структурная схема НС для распознавания рукописных цифр на базе платформы Zybo