УДК 004.383.3

РЕАЛИЗАЦИЯ ПАРАМЕТРИЗУЕМОГО НЕЙРОПРОЦЕССОРА ДЛЯ ПОЛНОСВЯЗНЫХ НЕЙРОННЫХ СЕТЕЙ НА ПЛИС

Чураков Р. А.[1], Гавенчук А.В. [1], Алябьев К.А. [1], Ахмеров А. Х.[1]
Научный руководитель – канд. техн. наук, доцент Дейнека И. Г. [1]
[1]Университет ИТМО
roma.churakov.022@mail.ru

Введение

Edge AI предусматривает выполнение нейросетевых моделей непосредственно на конечных устройствах. Последовательная архитектура CPU и микроконтроллеров ограничивает эффективность реализации алгоритмов нейронных сетей. ПЛИС представляют собой аппаратную платформу, обеспечивающую возможность формирования специализированной вычислительной структуры с учётом требований конкретной задачи, что позволяет повысить эффективность реализации нейросетевых моделей [1, 4]. Существующие FPGA-ускорители нейросетей, как правило, не обладают свойством инвариантности к конфигурации оборудования. Использование HLS-методологии или создание специализированных архитектур приводит к жесткой привязке реализации к исходному объему ресурсов, что затрудняет миграцию разработанных решений на платформы с иными характеристиками [2, 3]. В работе рассматривается реализация параметризуемого нейропроцессора, архитектура которого может масштабироваться в зависимости от числа доступных аппаратных ресурсов.

Основная часть

В работе реализовано вычислительное IP-ядро нейропроцессора для квантованных полносвязных нейронных сетей (MLP). Архитектура построена вокруг параллельной сети MAC-блоков, число которых задаётся параметрами конфигурации и может изменяться без переработки структуры кода. Такой подход позволяет адаптировать производительность под ограничения конкретной ПЛИС.

Вычислительный тракт включает сеть MAC-блоков, модуль управления на основе конечного автомата, кольцевой буфер для межслойного обмена и модуль активации с поддержкой ReLU и квантования. Параметризация реализована через конфигурационный пакет VHDL, в котором задаются разрядности данных, количество слоёв, число нейронов и количество параллельных вычислителей. Это позволяет собирать различные конфигурации ядра в рамках одной архитектуры.

Особое внимание уделено организации памяти. Вместо централизованного банка весов использована распределённая структура ROM-памяти, размещённая локально у MAC-блоков. Это упрощает маршрутизацию и позволяет повысить тактовую частоту, что соответствует практическим рекомендациям по оптимизации FPGA-ускорителей [3]. Компромиссом является возможная избыточность использования BRAM при несоответствии числа нейронов и числа MAC-блоков.

Аппаратная реализация выполнена на ПЛИС Intel Cyclone V. Проведена функциональная верификация в ModelSim и анализ таймингов в Quartus Prime. Получены количественные оценки использования LUT, регистров, BRAM и DSP-блоков. Измеренная латентность обработки одного входного вектора составляет десятки микросекунд, что на порядки быстрее программной реализации на встраиваемых процессорах общего назначения.

Также исследована возможность множественного инстанцирования вычислительного ядра на одной ПЛИС. Ограничивающим фактором в данном случае является объём блочной памяти, а не логические ресурсы, что подтверждает значимость оптимизации хранения весовых коэффициентов.

Выводы

Реализовано параметризуемое IP-ядро нейропроцессора для инференса полносвязных нейронных сетей на ПЛИС. Архитектура масштабируется за счёт изменения числа MAC-блоков и параметров конфигурации без изменения структурной схемы. Экспериментально подтверждена корректность работы и получены метрики производительности и ресурсопотребления.

Текущая реализация ориентирована на MLP-архитектуру. Расширение на свёрточные сети потребует модификации схемы адресации памяти и структуры вычислительного конвейера, что рассматривается как дальнейшее направление развития.

Литература

[1] Omondi A., Rajapakse J. FPGA Implementations of Neural Networks. Springer, 2006.

[2] Umuroglu Y. et al. FINN: A Framework for Fast, Scalable Binarized Neural Network Inference // FPGA. 2017.

[3] Zhang C. et al. Optimizing FPGA-based Accelerator Design for Deep CNNs // FPGA. 2015.

[4] Guo K. et al. A Survey of FPGA-based Neural Network Inference Accelerators // ACM TRETS. 2019.

[5] Sze V. et al. Efficient Processing of Deep Neural Networks // Proceedings of the IEEE. 2017.