УДК 004.383.5

АРХИТЕКТУРА НЕЙРОСЕТЕВОГО УСКОРИТЕЛЯ ДЛЯ ПОСТРОЕНИЯ КАРТЫ ГЛУБИНЫ

Сладков М. Ю. (ИТМО),
Научный руководитель – кандидат технических наук, доцент Быковский С. В.
(ИТМО)

Введение. Карта глубины сцены представляет собой цифровое изображение сцены, где цвет пикселя кодирует расстояние от объекта, которому принадлежит пиксель, до камеры. Построение карты глубины по изображению с одной камеры является одной из наиболее актуальных задач при разработке систем беспилотного транспорта и смешанной реальности. Эту задачу с высокой точностью решают нейросети трансформеры [1]. Расчёт выхода таких нейросетей требует большого количества вычислений и производится обычно на дорогостоящих GPU.

Тем временем вычисление карты глубины особенно актуально во встроенных и киберфизических системах. В них невозможно использование крупных GPU с высоким энергопотреблением, а вычисление карт глубины на микроконтроллерных процессорных ядрах не позволяет обеспечить работу в реальном времени. Разработка архитектуры специализированного вычислителя для построения карты глубины в этом контексте становится по-настоящему актуальной задачей.

Основная часть.

В работе описана архитектура перспективного ускорителя для построения карты глубины с помощью свёрточных нейросетей трансформеров, применимого во встроенных и киберфизических системах в условиях реального времени. Для минимизации сложности вычислений предлагается широко использовать квантизацию нейросетей и табулирование функций активации.

В вычислительных системах, работающих с нейросетями, память часто становится ограничителем производительности. Чтобы избежать этого был разработан метод хранения нейросети в памяти и модель кэша, которые вместе позволяют минимизировать задержки при обращениях к памяти для извлечения параметров нейросети.

Предлагаемая архитектура основывается на использовании блоков вычисления свёртки и конвейеризации вычислений в нейросети. В силу ориентации исключительно на сети трансформеры стало возможным ориентировать ускоритель на архитектуру нейросети энкодер-декодер и поделить вычислительные ресурсы на блоки, соответствующие блокам нейросети. За счёт конвейеризации предлагается минимизировать количество обращений к памяти для хранения промежуточных результатов и ещё сильнее увеличить производительность ускорителя. [2]

Выводы. Предложена архитектура и способ конфигурирования специализированного нейросетевого вычислителя для построения карты глубины. Ускоритель на основе разработанной архитектуры может быть разработан с использованием языков описания аппаратуры, после чего его можно будет интегрировать в системы на кристалле.

Список использованных источников:

1. Abuolaim1 A., Brown M. Defocus Deblurring Using Dual-Pixel Data 2020 URL: https://arxiv.org/abs/2005.00305 (дата обращения: 04.02.2024)
2. Law O. M. K., Liu A. C. C. Artificial Intelligence Hardware Design. Challenges and Solutions. IEEE PRESS. 2021.