УДК 004.31

МЕТОД ОРГАНИЗАЦИИ ПОДСИСТЕМЫ ХРАНЕНИЯ И ПЕРЕДАЧИ ДАННЫХ В АППАРАТНЫХ УСКОРИТЕЛЯХ НЕЙРОННЫХ СЕТЕЙ

Табунщик С. М.[1]
Научный руководитель – канд. техн. наук Быковский С.В. [1]
[1]Университет ИТМО
sergei_tabunshchik@itmo.ru

Работа выполнена в рамках темы НИР №623106 «Автономные интеллектуальные системы».

Введение

Современные приложения искусственного интеллекта отличаются необходимостью хранить, передавать и обрабатывать большой объём данных. Интенсивный рост сложности алгоритмов машинного обучения и расширение сфер применения искусственного интеллекта приводят к экспоненциальному увеличению объёмов обрабатываемых данных. Особую актуальность эта проблема приобретает в контексте краевого искусственного интеллекта (Edge AI) – подхода, при котором обработка данных выполняется не в централизованных облачных центрах, а непосредственно на периферийных устройствах или вблизи них.[1]

Именно краевой искусственный интеллект становятся основой для множества современных приложений: автономных транспортных средств, роботов и дронов, умных камер видеонаблюдения и носимых медицинских устройств. Для ускорения расчета задач краевого искусственного интеллекта во всех этих сценариях перекладывается на специализированные аппаратные ускорители нейронных сетей. Ускорители выполняют обработку большого объема данных в реальном времени во встраиваемых системах. Это накладывает ограничения на подсистемы хранения и передачи данных в специализированных ускорителях, интегрированных непосредственно в конечное оборудование. Ограниченная пропускная способность внутренних шин данных приводит к возникновению «узкого места» при передаче нескольких потоков данных между памятью, датчиками и вычислительной подсистемой. К этому добавляются высокие требования к скорости обработки: задержки в передаче и обработке данных могут нарушить работу систем реального времени, где критически важна быстрая реакция на входные сигналы. Существенную проблему представляет энергопотребление: операции по передаче и обработке больших объёмов данных создают значительную нагрузку на энергосистему устройства, что особенно критично для автономных и мобильных встраиваемых решений. Кроме того, объём доступной памяти в таких системах строго ограничен, а хранение промежуточных и итоговых данных требует значительный объем локальной памяти. Для сокращения влияния описанных ограничений требуется применение специализированных методов организации подсистемы хранения и передачи данных.

Целью работы является анализ существующих методов организации подсистемы хранения и передачи данных в аппаратных ускорителях нейронных сетей для повышения производительности нейросетевого процессора с авторской архитектурой.

Основная часть

В работе описываются методы и технические решения, выбранные для организации подсистемы хранения и передачи данных в нейросетевом процессоре. Основной задачей разрабатываемого нейросетевого процессора является ускорение расчета сверточных нейронных сетей. Выбранные решения включают в себя топологию, интерфейсы и алгоритмы управления хранением, обработкой и передачей данных.[2]

Применение описанных решений в нейросетевом процессоре позволило сократить задержки при передаче данных, повысить производительность нейросетевого процессора, расширить объем памяти для хранения обрабатываемых данных и сократить объем данных, необходимых для передачи между компонентами нейросетевого процессора.

В результате анализа методов организации подсистемы хранения и передачи данных и внедрения их в разрабатываемый нейросетевой процессор получены следующие результаты:

1) описание топологии подсистемы хранения и передачи данных,
2) описание протоколов передачи данных на разных этапах потока данных,
3) описание способа адресации памяти для хранения активаций, весов и смещений,
4) описание алгоритма предварительной обработки данных для сокращения объем данных, необходимых для передачи,
5) оценка зависимости производительности обработки данных от предложенных методов.

Выводы

Предложенные методы организации подсистемы хранения и передачи данных позволяют сократить задержки при передаче данных, повысить производительность нейросетевого процессора, расширить объем памяти для хранения обрабатываемых данных и сократить объем данных, необходимых для передачи между компонентами нейросетевого процессора.

Литература

1. Gill S. S. et al. Edge AI: A taxonomy, systematic review and future directions //Cluster Computing. – 2025. – Т. 28. – №. 1. – С. 1-53.
2. Chen Y. H. et al. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices //IEEE Journal on Emerging and Selected Topics in Circuits and Systems. – 2019. – Т. 9. – №. 2. – С. 292-308.