УДК 004.896

Обзор и классификация инструментов имплементации Edge AI на FPGA

**Гавенчук А.В. (ИТМО), Ахмеров А.Х. (ИТМО), Чураков Р.А. (ИТМО),
Алябьев К.А. (ИТМО)**

**Научный руководитель – кандидат технических наук, доцент Дейнека И.Г.
(ИТМО)**

Введение. Искусственный интеллект на сегодняшний день является одной из ключевых технологий, оказывающих влияние на многие сферы жизни человека – от бытовых устройств и транспорта до промышленности и медицины [1]. Важно отметить, что значительная часть текущих приложений искусственного интеллекта ориентирована на реализацию с использованием облачных вычислений [2]. Однако такой подход имеет ряд недостатков: задержка при передаче данных между устройством и сервером, высокая нагрузка на сеть при большом потоке данных, риск несанкционированного доступа к данным в момент их передачи. Перечисленные недостатки не позволяют реализовывать автономные приложения, работающие в режиме реального времени, что может быть крайне важно, например, для мониторинга состояния пациентов в области здравоохранения или для систем автономного принятия решений в промышленности [3, 4]. Решением этой проблемы является перенос инференса нейросетевых моделей к источнику данных, то есть реализация вычислений локально на краевых устройствах – переход к краевому искусственному интеллекту (Edge AI). Гибкая архитектура, обработка данных с малой задержкой, параллелизм вычислений и высокая энергоэффективность делают программируемые логические интегральные схемы (field-programmable gate array, FPGA) перспективной платформой для реализации Edge AI. Однако сложность данного подхода заключается в том, что FPGA в составе краевых устройств имеют ограниченные вычислительные мощности и объем доступных ресурсов. В связи с этим возникает необходимость в оптимизации как самих нейросетевых моделей, так и структур, выполняющих необходимые расчёты во время работы нейронной сети.

Основная часть. Отсутствие единого универсального решения, позволяющего выполнить перенос архитектуры и параметров нейронных сетей на программируемую логику существенно увеличивает порог вхождения для разработчиков и требует формирования междисциплинарных команд специалистов, что в совокупности существенно снижает темпы развития подобного рода технологий. Это порождает необходимость систематизации информации в рассматриваемом направлении с целью заложить научно-техническую основу для новых начинаний. В настоящей работе группой авторов выполнен обзор существующих методов и инструментов для имплементации Edge AI на FPGA.

Существующие методы имплементации можно разбить на три большие группы согласно уровням абстракции, используемому при программировании FPGA. На самом высоком уровне располагаются фреймворки, позволяющие генерировать нейросетевые модели, такие как PyTorch, TensorFlow/Keras, JAX, Apache MXNet, Caffe и ONNX [5]. На сегодняшний день наиболее популярными представителями этой группы остаются PyTorch, фреймворк в большей степени нацеленный на использование в научно-исследовательских целях, и TensorFlow, ориентированный на реализацию высокопроизводительных решений. Следующие шаги в реализации Edge AI включают в себя оптимизацию нейронной сети, как правило выполняемую методами квантизации, обрезки и дистилляции, и последующую адаптацию модели для размещения на FPGA. Для выполнения подготовки модели для размещения на FPGA могут быть использованы такие инструменты, как hls4ml, FINN, Vitis AI, Lattice SensAI, LeFlow [6]. В то же время, существуют решения, позволяющие выполнить только оптимизацию модели, такие как Intel OpenVINO, Apache tvn, Tensorflow Lite, а также решения, позволяющие адаптировать уже оптимизированное решение, например Intel FPGA AI Suite и Intel HLS Compiler, GOAI 2.0. В данном случае разработчику уже необходимо учитывать особенности архитектуры программируемой логики при создании модели, однако

по-прежнему нет необходимости использовать языки описания аппаратуры. Следует отдельно выделить программный пакет MATLAB & Simulink, позволяющий пройти все этапы реализации Edge AI от обучения нейросетевой модели до генерации кода на языках Verilog и VHDL [7]. Инструменты самого низкого уровня абстракции представляют собой проприетарные платформи-ориентированные системы автоматизированного проектирования для FPGA, включая AMD Vivado, Intel Quartus Prime, Gowin EDA, Lattice Radiant/Propel/Diamond, Efinity, Libero. Помимо интеграции с инструментами более высокого уровня, данные инструменты предлагают возможность разработать полнофункциональные решения Edge AI, используя комбинационную и последовательную логику, реализовать все элементы архитектуры нейронной сети уже непосредственно на языках описания аппаратуры [8, 9].

Выводы. В ходе исследования был проведен обзор существующих инструментов и сформирована структурная классификация для имплементации Edge AI на FPGA. Результаты данного исследования будут полезны разработчикам и исследователям решений в направлениях интернета вещей, робототехники, автономного транспорта, умного города, компактных медицинских устройств и инфокоммуникационных систем.

Список использованных источников:

1. Seng K. P., Lee P. J., Ang L. M. Embedded intelligence on FPGA: Survey, applications and challenges // *Electronics*. – 2021. – Т. 10. – №. 8. – С. 895.
2. Gill S. S. et al. Edge AI: A taxonomy, systematic review and future directions // *Cluster Computing*. – 2025. – Т. 28. – №. 1. – С. 1-53.
3. Chinta S. Edge AI for real-time decision making in IoT networks // *International Journal of Innovative Research in Computer and Communication Engineering*. – 2024. – Т. 12. – №. 9. – С. 11293-11309.
4. Bansal K. et al. DeepBus: Machine learning based real time pothole detection system for smart transportation using IoT // *Internet Technology Letters*. – 2020. – Т. 3. – №. 3. – С. e156.
5. Mohialden Y. M. et al. Top Python-Based Deep Learning Packages: A Comprehensive Review // *International Journal Paper Advance and Scientific Review*. – 2024. – Т. 5. – №. 1. – С. 1-9.
6. Lebedev M., Belecky P. A survey of open-source tools for FPGA-based inference of artificial neural networks // *2021 Ivannikov Memorial Workshop (IVMEM)*. – IEEE, 2021. – С. 50-56.
7. Zairi H. et al. FPGA-based system for artificial neural network arrhythmia classification // *Neural Computing and Applications*. – 2020. – Т. 32. – С. 4105-4120.
8. Lahari P. L., Yellampalli S. S., Vaddi R. HDL Design and Implementation of a Convolutional Neural Network for Efficient FPGA/Hardware Platforms // *2022 International Conference on Industry 4.0 Technology (I4Tech)*. – IEEE, 2022. – С. 1-5.
9. Ranganath L., Kumar D. J., Reddy P. S. N. Design of MAC unit in artificial neural network architecture using Verilog HDL // *2016 International conference on signal processing, communication, power and embedded system (SCOPEs)*. – IEEE, 2016. – С. 607-612.