УДК 621.38:004.33

ВИКОРИСТАННЯ ЗОВНІШНЬОЇ ПАМ'ЯТИ ДЛЯ ВЕЛИКИХ НЕЙРОННИХ МОДЕЛЕЙ НА ПЛІС

Дубіцький А. Є.
e-mail: artur.dubitskyi@nure.ua
Науковий керівник – доц., к.ф.-м.н., доц каф. МЕЕПП Глухов О.В.,
Харківський національний університет радіоелектроніки, каф. МЕЕПП,
м. Харків, Україна

This work examines the integration of external memory to support neural network models on FPGAs. While FPGAs offer reduced size and single chip capabilities, their limited on-chip memory restricts the deployment of complex multi-task architectures. By adding external off-chip memory banks, model's wights can be placed there. This approach could be achieved with a ARTIX-7 FPGA Development Board AX7203, as it has all the necessary hardware.

Згорткові нейронні мережі використовуються у сферах класифікації зображень, розпізнавання образів, автономного водіння, медицині, а прискорювачі таких нейронних мереж стали актуальним напрямом досліджень. Найчастіше для подібних задач застосовуються графічні процесори (ГП). Однак, програмовані логічні інтегральні схеми (ПЛІС) також дозволяють розробляти апаратні рішення для роботи нейронних мереж (НМ), а завдяки однокристальній компоновці (SoC) і малим розмірам, вони виступають ідеальним рішенням для автономних систем, роботів. Вони мають кращу енергоефективність та меншу ціну, порівняно з ГП [1,2].

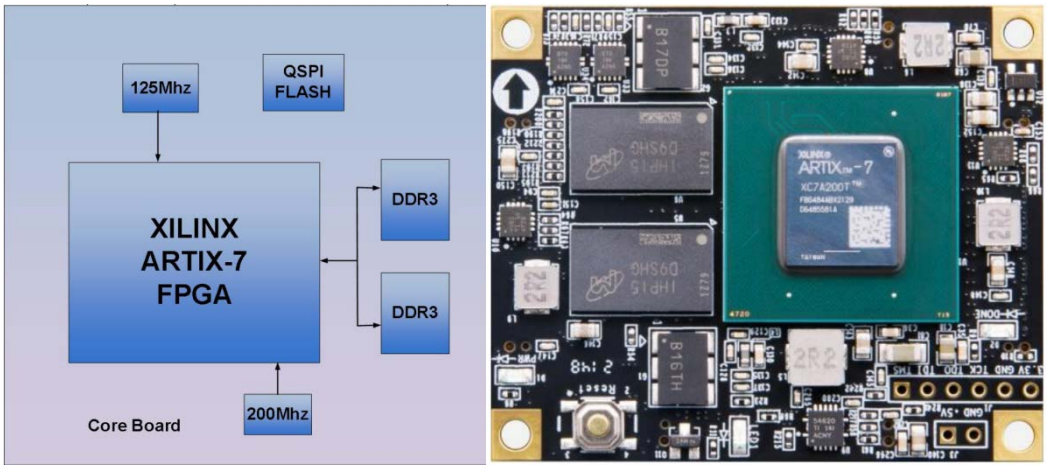
Для досягнення найбільш ефективної роботи НМ, її параметри оптимізуються таким чином, щоб вона вміщувалась у доступний об'єм вбудованої оперативної пам'яті ПЛІС. Проте, це актуально, коли нейронна мережа націлена вирішувати одне завдання. При застосуванні методу багатоцільового навчання, особливо структур із розподіленими параметрами, для вирішення декількох суміжних задач, об'єм нейронної мережі стає більшим за виділену оперативну пам'ять ПЛІС [3].

Звідси, мета роботи – показати можливість розміщення НМ, розмір яких більшій за вбудовану пам'ять, у зовнішніх банках оперативної та постійної пам'яті.

Матриці вагів $\{W_1, W_2, W_3, \ldots W_n\}$ нейронної мережі записуються до певних адрес банків пам'яті у вигляді змінних. Під час виконання програми йде звернення до даних адрес. Це дозволяє зберігати потрібну кількість параметрів нейронної мережі, залежно від об'єму банку пам'яті, і динамічно їх оновлювати. Тим не менш, доступ до зовнішніх банків пам'яті на порядки повільніший за доступ до вбудованої пам'яті.

Прикладом такої конфігурації може виступати ARTIX-7 FPGA Development Board AX7203, рисунок 1. Дана плата має 2 банки пам'яті DDR3 по 4 Гб, які виділеними лініями з'єднані із SoC. Блоків оперативної

пам’яті на кристалі 13 Мб [4]. Файл із вагами нейронної мережі VGG-16 із сайту PyTorch важить 537 Мб, AlexNet – 233 Мб, а от MobileNet V3 Small – 9.8 Мб, ResNet152 – 230 Мб.



а б

Рисунок 1 – схема ядра плати розробника на Artix-7 із банками пам'яті – а, і реальне фото плати – б

Таким чином, у даному матеріалі було показано, що при використанні нейронних моделей, розмір яких не дозволяє розміщувати їх на оперативній пам’яті SoC, можливо ефективно використовувати зовнішні банки пам’яті для роботи таких нейронних мереж.

Перспективним є впровадження підходу, коли на програмованій логіці ПЛІС формується граф нейронної мережі. У такому випадку, ваги стають незмінними, оскільки сформовані у вигляді електронних зв’язків між логічними комірками, а сама ПЛІС починає лише обробляти запити на вході і повертати дані у запрограмованому вигляді [5].

Список використаних джерел:

1. Hardware Architectures for Real-Time Medical Imaging / E. Alcaín et al. Electronics. 2021. Vol. 10, no. 24. P. 3118.

2. Architectural Design and Performance Analysis of FPGA based AI Accelerators: A Comprehensive Review / S. Chatterjee et al. arXiv. arXiv:2603.08740. 2026.

3. Dubitskyi A., Glukhov O., Beresnev V. Multi-Task Neural Network for Medical Diagnosis Targeted for FPGA Deployment. J. Nano- Electron. Phys. 2026. Vol. 18, no. 1.

4. ARTIX-7 FPGA Development Board AX7203. URL: https://www.alinx.com/public/upload/file/AX7203_User_Manual.pdf. (date of access: 14.03.2026).

5. Fang L. The Immutable Tensor Architecture: A Pure Dataflow Approach for Secure, Energy-Efficient AI Inference. arXiv. arXiv:2511.22889. 2025.