МОДЕЛЬ ОПТИМІЗАЦІЇ НЕЙРОННИХ МЕРЕЖ ДЛЯ ВБУДОВАНИХ СИСТЕМ З ОБМЕЖЕНИМИ РЕСУРСАМИ

Авраменко Б.О., Філімончук Т.В., Севостьянова О.М.
Харківський національний університет радіоелектроніки, Харків, Україна

Вбудовані системи, що використовуються в IoT-пристроях, безпілотниках, медичному обладнанні та автономних транспортних засобах, мають суворі обмеження за обчислювальною потужністю, оперативною пам'яттю та енергоспоживанням. Традиційні моделі глибоких нейронних мереж (DNN) із десятками мільйонів параметрів та мільярдами операцій з плаваючою комою непридатні для розгортання на мікроконтролерах типу STM32, ESP32 або nRF52 [1], особливо за умов вимог до забезпечення низької латентності інференсу та у деяких сценаріях автономної роботи від батареї.

Метою доповіді є побудова моделі оптимізації нейронних мереж на основі використання сучасних методів оптимізації, які адаптують DNN до апаратних обмежень вбудованих систем без значної втрати точності. В доповіді розглянуто чотири підходи до оптимізації нейронних мереж: квантизація, прунінг, використання компактних архітектур та інтеграція з нейронними прискорювачами. Дослідження показують, що посттренувальна квантизація зменшує розмір моделі в 4 рази та прискорює інференс у 2-4 рази на процесорах ARM Cortex-M, а квантизація з усвідомленням дозволяє зберегти точність на рівні 95-98% від оригінальної моделі [2]. Структурований прунінг скорочує обсяг обчислень на 50-70% при втраті точності менше 2%, а ітеративний прунінг з тонким донавчанням виявляється особливо корисним для згорткових мереж [3]. Використання компактних архітектур як от MobileNetV3-Small (2,9 млн. параметрів, 66 млн. FLOPs) забезпечує роботу на мікроконтролерах із частотою 80-216 МГц, а глибинно-сепарабельні згортки зменшують обчислювальну складність у 8-9 разів порівняно зі стандартними згортками [4]. Використання інструкцій SIMD (NEON, DSP) та апаратних MAC-блоків додатково прискорює виконання, а фреймворк Edge Impulse спрощує розгортання моделей. У зв'язку з цим чинності набуває комбінація методів оптимізації нейронних мереж, яка дозволяє досягти балансу між розміром моделі, швидкістю, енергоспоживанням і точністю.

Список літератури

1. Zhang Z., Li J. A Review of Artificial Intelligence in Embedded Systems. Micromachines. 2023. Vol. 14, no. 5. P. 897. DOI: https://doi.org/10.3390/mi14050897

2. Gholami A., Kim S., Dong Z., Yao Z., Mahoney M.W., Keutzer K. A Survey of Quantization Methods for Efficient Neural Network Inference. Low-Power Computer Vision (Chapter 13), 2022. P. 291-326. DOI: https://doi.org/10.1201/9781003162810-13

3. Han S., Pool, J., Tran J., Dally W.J. (2015). Learning both Weights and Connections for Efficient Neural Networks. DOI: https://doi.org/10.48550/arxiv.1506.02626

4. Howard A. G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. (2017). MobileNets: efficient convolutional neural networks for mobile vision applications. DOI: https://doi.org/10.48550/arxiv.1704.04861