

ВИКОРИСТАННЯ TPU АКСЕЛЕРАТОРА НА БАЗІ МІКРОКОМП'ЮТЕРА ДЛЯ ПРИСКОРЕННЯ АЛГОРИТМІВ РОЗПІЗНАВАННЯ ОБ'ЄКТІВ НА ЗОБРАЖЕННІ

Ройко Андрій Олегович

ORCID ID: 0009-0000-4264-0012

студент 1-го курсу магістратури

Національний університет "Львівська політехніка", Україна

Штучний інтелект здійснив революцію в обробці зображень і комп'ютерному зорі, домінуючи в сотнях галузей включаючи розпізнавання об'єктів [1], класифікацію зображень [2], семантичну сегментацію [3] та інше. Водночас стрімко поширюється використання мікрокомп'ютерів для систем штучного інтелекту в реальному часі [4] [5]. Вони приваблюють невеликим використанням електричної потужності, мають низьку затримку та дешеву вартість.

Для прискорення роботи глибоких нейронних мереж було широко досліджено використання FPGA [6], проте згодом з'явилися більш вузькі рішення заточені лише на моделі глибокого навчання. Так з'явилися наступні спеціалізовані продукти-акселератори для прискорення алгоритмів глибокого навчання:

- NVIDIA Jetson family [7]
- Intel Movidius VPUs [8]
- ASUS Tinker Edge R [9]
- Google's Edge TPUs [10]

Одним з яскравих представників TPU акселераторів став вже зазначений Google Coral TPU [10], котрий дозволяє виконувати до 4 трильйонів операцій з 8-бітними числами з нерухомою комою за секунду з використанням 2 Вт потужності.

Наукові публікації містять ряд досліджень з сумісного використання TPU акселераторів разом з мікрокомп'ютерами [11] [12], проте мета цього дослідження зосередитись саме на завданні прискорення алгоритмів розпізнавання об'єктів на зображенні.

Для дослідження ефективності роботи TPU-акселераторів на алгоритмах розпізнавання об'єктів на зображенні було використано Google Coral TPU USB Accelerator та мікрокомп'ютер Raspberry Pi 4/8 ГіБ (рис 1.).

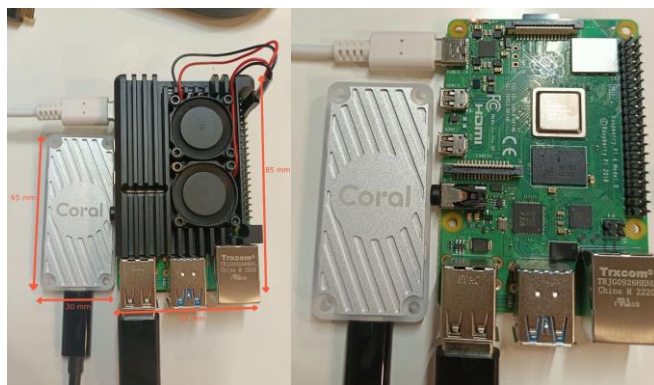


Рис. 1. (Лівий рисунок) Збірка Raspberry Pi 4 та TPU Coral USB Accelerator в захисному кейсі з двома охолоджувальними кулерами. (Правий рисунок) Збірка Raspberry Pi 4 та TPU Coral USB Accelerator без захисного кейсу

Задля порівняння результатів з та без використання TPU акселератора було використано попередньо натреновані моделі для розпізнавання об'єктів на зображенні з використанням збірки даних Coco від компанії Coral [13]. Їхнє дослідження відбувалось на базі TPU Coral USB акселератора, проте без використання мікрокомп'ютерів. Результати дослідження наведено в (табл. 1).

Таблиця 1

**Порівняльна характеристика використання Coral TPU
акселератора на базі Raspberry Pi 4**

Назва моделі	Вхідний розмір (висота, ширина, кількість каналів)	Затримка на передбаченні (без підтримки TPU) (мс)	Затримка на передбаченні (з використанням TPU) (мс)	mAP (%)	Розмір моделі (МіБ)
SSD MobileNet V1	300x300x3	219	27	21.5	7.1
SSD/FPN MobileNet V1	640x640x3	7880	1597	31.1	32.2
SSD MobileNet V2	300x300x3	209	32	22.4	6.6
SSDLite MobileDet	320x320x3	291	36	32.9	5.2
EfficientDet-Lite0	320x320x3	337	180	30.4	4.4
EfficientDet-Lite1	384x384x3	636	247	34.3	5.8
EfficientDet-Lite2	448x448x3	993	483	36.0	7.3
EfficientDet-Lite3	512x512x3	1839	544	39.4	12

Висновки. Використання TPU акселераторів у зв'язці з мікрокомп'ютерами є перспективним дослідженням, через низьку вартість обладнання, компактність, а також високу енергоефективність.

У цій роботі було проведено дослідження щодо використання Coral TPU USB акселератора для прискорення алгоритмів глибокого навчання з метою розпізнавання об'єктів на зображеннях, використовуючи мікрокомп'ютер Raspberry Pi 4/8.

Завдяки експериментальним оцінкам наведеним в (табл. 1) бачимо, що на всіх тестуваннях моделі ми отримали від двох до восьмикратного покращення в порівнянні з виконанням обчислень лише на CPU. Вдалось здобути значне покращення навіть для моделей чий розмір більший, ніж статична ОЗП в TPU акселератора (8 МіБ). В таких випадках деякі операції будуть обчислені на TPU акселератора, а інша частина на CPU.

Список використаних джерел:

1. Zou Z. Object Detection in 20 Years: A Survey / Z. Zou, K. Chen, Z. Shi, та ін. // Proc. IEEE. — 2023. — Вип. 111, № 3. — С. 257–276.
2. Plested J. Deep transfer learning for image classification: a survey / J. Plested, T. Gedeon. — arXiv, 2022.
3. Csurka G. Semantic Image Segmentation: Two Decades of Research / G. Csurka, R. Volpi, B. Chidlovskii. — arXiv, 2023.
4. Rossi A. Real-time Lane detection and Motion Planning in Raspberry Pi and Arduino for an Autonomous Vehicle Prototype / A. Rossi, N. Ahmed, S. Salehin // 2020.
5. Luo Z. Cloud Chaser: Real Time Deep Learning Computer Vision on Low Computing Power Devices / Z. Luo, A. Small, L. Dugan, S. Lane. — 2019.
6. Shawahna A. FPGA-based Accelerators of Deep Learning Networks for Learning and Classification: A Review / A. Shawahna, S. M. Sait, A. El-Maleh // IEEE Access. — 2019. — Вип. 7. — С. 7823–7859.

7. Mittal S. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform / S. Mittal // J. Syst. Archit. — 2019. — Вип. 97. — С. 428–442.
8. Kaarmukilan s. P. An Accelerated Prototype with Movidius Neural Compute Stick for Real-Time Object Detection / s. P. Kaarmukilan, A. Hazarika, A. Thomas K, та ін. — 2020.
9. Asus Tinker Edge R / Asus. — 2024.
10. Cass S. Taking AI to the edge: Google's TPU now comes in a maker-friendly package / S. Cass // IEEE Spectr. — 2019. — Вип. 56, № 5. — С. 16–17.
11. Qu Z. Enabling Deep Learning on Edge Devices / Z. Qu. — 2022.
12. Sun Y. Deep Learning on Edge TPUs / Y. Sun, A. M. Kist. — arXiv, 2022.
13. Coral AI Models - Object Detection / Coral AI. — 2024.