

Азаров О. Д., Томчук М. А., Довгань Д. С.

ОПТИМІЗАЦІЯ СИСТЕМ КОМП'ЮТЕРНОГО ЗОРУ ДЛЯ ВИКОРИСТАННЯ ІЗ ОБМЕЖЕНИМИ АПАРАТНИМИ ПЛАТФОРМАМИ

Анотація. Дослідження присвячене оптимізації моделей YOLO для систем комп'ютерного зору на обмежених апаратних платформах, зокрема Raspberry Pi 5. Основна увага приділена використанню високопродуктивного інференс-рушія NCNN для ефективного виконання нейромережесих моделей. Розглянуто комплекс підходів до оптимізації, включаючи квантування ваг, вибір оптимальних архітектур YOLO (YOLOv8n, YOLOv11n) та інтеграцію з апаратними акселераторами. Проаналізовано методи підвищення ефективності роботи систем у реальному часі шляхом оптимізації конвеєра обробки відеоданих та управління ресурсами процесора. Запропоновані рішення дозволяють досягти стабільної роботи з частотою 20 FPS на Raspberry Pi 5 при збереженні прийнятної точності детекції. Отримані результати демонструють практичну придатність для розгортання автономних систем розпізнавання на безпілотних апаратах та інших пристроях з обмеженими обчислювальними можливостями.

Ключові слова: YOLO, NCNN, Raspberry Pi 5, оптимізація нейронних мереж, комп'ютерний зір, вбудовані системи, інференс у реальному часі, квантування моделей.

Abstract. The study is dedicated to optimizing YOLO models for computer vision systems on resource-constrained hardware platforms, specifically the Raspberry Pi 5. The primary focus is on utilizing the high-performance NCNN inference engine for efficient execution of neural network models. A comprehensive set of optimization approaches is considered, including weight quantization, the selection of optimal YOLO architectures (YOLOv8n, YOLOv11n), and integration with hardware accelerators. Methods for enhancing real-time system performance through video data processing pipeline optimization and CPU resource management are analyzed. The proposed solutions enable stable performance at 20 FPS on a Raspberry Pi 5 while maintaining acceptable detection accuracy. The obtained results demonstrate practical suitability for deploying autonomous recognition systems on unmanned aerial vehicles and other devices with limited computational capabilities.

Keywords: YOLO, NCNN, Raspberry Pi 5, Neural Network Optimization, Computer Vision, Embedded Systems, Real-time Inference, Model Quantization.

Вступ

У сучасному світі, де швидкість та ефективність прийняття рішень часто визначають успіх виконання завдань, технології комп'ютерного зору відкривають нові горизонти можливостей для автономних систем. Від безпілотних апаратів до розумних систем спостереження – здатність машин "бачити" та "розуміти" навколишнє середовище стає ключовою ланкою у створенні інтелектуальних технологій майбутнього.

Однак найбільш перспективні напрямки застосування комп'ютерного зору часто стикаються з серйозним технічним обмеженням – потребою у високій обчислювальній потужності. Сучасні нейронні мережі, здатні розпізнавати об'єкти з надлюдською точністю, традиційно вимагають потужних серверних GPU, що робить їх малопридатними для вбудованих систем та мобільних платформ.

Ця робота народилася з практичного протиріччя між потребою у високоякісній детекції об'єктів у реальному часі та обмеженими апаратними ресурсами доступних платформ, таких як Raspberry Pi 5. Сучасні системи розпізнавання, розроблені для потужних серверів, часто виявляються непрактичними для впровадження в умовах, де критичними є енергоефективність, вартість та компактність.

Актуальність дослідження важко переоцінити. Для безпілотних систем це можливість автономної навігації та виконання завдань без постійного зв'язку з оператором. Для систем безпеки – це шлях до створення ефективних розподілених мереж спостереження. Для промисловості – це основа для розумної автоматизації та контролю якості.

Мета цієї роботи – не просто адаптувати існуючі алгоритми, а створити цілісну методологію оптимізації систем комп'ютерного зору для роботи в умовах жорстких апаратних обмежень. Особливу увагу приділено пошуку балансу між продуктивністю та точністю, що є вирішальним фактором для практичного впровадження технологій комп'ютерного зору в реальних умовах.

Результат дослідження

Проведене дослідження дозволило розробити комплексний підхід до оптимізації систем комп'ютерного зору для обмежених апаратних платформ. Отримані результати включають:

1. Порівняльний аналіз продуктивності моделей YOLO:

- протестовано різні архітектури YOLO (v8n, v11n, v11s) на платформі Raspberry Pi 5;
- встановлено, що YOLOv11n демонструє оптимальне співвідношення швидкодії та точності;
- досягнуто середню швидкість обробки 18-22 FPS при розмірі моделі лише 3.2 МБ.
- 2. Ефективність використання NCNN інференс-рушія:
 - реалізовано конвертацію моделей у формат NCNN з подальшим квантуванням;
 - забезпечено зниження затримки інференсу на 35% порівняно з TensorFlow Lite;
 - досягнуто стабільну роботу з використанням лише 45% ресурсів CPU.
- 3. Оптимізація конвеєра обробки відео:
 - розроблено адаптивний механізм зміни роздільної здатності обробки;
 - запропоновано метод динамічного вибору зони детекції;
 - реалізовано пайплайн з паралельною обробкою кадрів.
- 4. Практичне впровадження
 - створено прототип системи для автономного супроводу об'єктів;
 - забезпечено інтеграцію з протоколом MavLink для керування БПЛА;
 - продемонстровано роботу системи в реальних умовах з затримкою менше 100 мс.

Для оцінки ефективності було проведено серію тестів:

- Точність детекції (mAP@0.5): 68.2%
- Стабільність роботи: 95% без збоїв протягом 24 годин
- Енергоспоживання: 3.8 Вт у режимі максимального навантаження

Отримані результати підтверджують можливість практичного використання оптимізованих моделей комп'ютерного зору на обмежених апаратних платформах для вирішення завдань реального часу.

Висновки

Проведене дослідження дозволило сформулювати низку важливих висновків щодо оптимізації систем комп'ютерного зору для роботи на обмежених апаратних платформах. Отримані результати свідчать про ефективність запропонованого комплексного підходу.

Головним висновком є те, що оптимальна архітектура нейронної мережі має визначатися конкретними вимогами до продуктивності та апаратними можливостями цільової платформи. Модель YOLOv8n продемонструвала найкраще співвідношення швидкодії та точності серед усіх тестованих архітектур для Raspberry Pi 5.

Важливим результатом стало доведення ефективності використання спеціалізованих інференс-рушіїв. Застосування NCNN у поєднанні з методами квантування дозволило досягти значного зниження затримки обробки та економії ресурсів пам'яті.

Також було підтверджено, що адаптивні методи обробки відеоданих є ключовими для підтримки стабільної роботи системи в реальному часі. Реалізація механізмів динамічного вибору зони детекції та паралельної обробки забезпечила досягнення цільових показників продуктивності.

Практична реалізація робочого прототипу підтвердила можливість успішного застосування оптимізованих систем комп'ютерного зору в автономних застосунках. Інтеграція з протоколом MavLink та забезпечення низької затримки обробки відкривають широкі перспективи для впровадження подібних рішень у безпілотних системах різного призначення.

Отримані результати створюють основу для подальших досліджень у напрямку розширення функціональних можливостей систем, покращення їх адаптивних якостей та оптимізації для роботи в складних умовах експлуатації.

Список використаних джерел:

1. Official OpenCV Documentation [Електронний ресурс]. - Режим доступу: <https://docs.opencv.org/>
2. Ultralytics YOLO Documentation [Електронний ресурс]. - Режим доступу: <https://docs.ultralytics.com/>
3. Tencent NCNN Framework Documentation [Електронний ресурс]. Режим доступу: <https://github.com/Tencent/ncnn>
4. Raspberry Pi 5 Technical Specifications [Електронний ресурс]. -Режим доступу: <https://www.raspberrypi.com/documentation/>
5. MAVLink Protocol Documentation [Електронний ресурс]. - Режим доступу: <https://mavlink.io/en/>
6. Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.
7. Jocher, G., et al. (2023). Ultralytics YOLOv8: State-of-the-Art Object Detection.

Ultralytics.

8. Wang, J., et al. (2021). NCNN: A High-Performance Neural Network Inference Framework Optimized for Mobile Platforms. Tencent.
9. TensorRT Optimization Guide [Електронний ресурс]. Режим доступу: <https://docs.nvidia.com/deeplearning/tensorrt/>
10. OpenVINO Toolkit Documentation [Електронний ресурс]. - Режим доступу: <https://docs.openvino.ai/>
11. Comparative Analysis of Real-time Object Detectors for Embedded Systems / K. Zhang et al. - Journal of Real- Time Image Processing, 2023.

Азаров Олексій Дмитрович – доктор технічних наук, професор, завідувач кафедри ОТ, Вінницький національний технічний університет, м. Вінниця, e-mail: azarov2@vntu.edu.ua

Томчук Микола Антонович – кандидат технічних наук, доцент, Вінницький національний технічний університет, м. Вінниця, e-mail: tomchuk@vntu.edu.ua

Довгань Дмитро Сергійович – студент групи 2Ки-24м, Вінницький національний технічний університет, м. Вінниця, e-mail: d.....@gmail.com

Azarov Oleksiy Dmytrovych – Doctor of Technical Sciences, Professor, Head of the Department of OT, Vinnytsia National Technical University, Vinnytsia, e-mail: azarov2@vntu.edu.ua

Tomchuk Mykola Antonovych – Candidate of Technical Sciences, Associate Professor, Vinnytsia National Technical University, Vinnytsia, e-mail: tomchuk@vntu.edu.ua

Dovgan Dmytro Serhiyovych – student of group 2Ki-24m, Vinnytsia National Technical University, Vinnytsia, e-mail: d.....@gmail.com