ОПТИМИЗАЦИЯ МОДЕЛЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ ПОД РАБОТУ БЕЗ ГРАФИЧЕСКОГО УСКОРИТЕЛЯ

Злобин А.С. (ИТМО)

Научный руководитель – кандидат технологических наук, Щемелинин В.Л. (ИТМО)

Введение. Видеоаналитика в современном мире применяется повсеместно и позволяет автоматизировать процессы видеонаблюдения на производственных объектах и крупных предприятиях [1]. Ввиду некоторых аспектов организации работы на предприятии, использовать облачные вычислительные мощности зачастую не представляется возможным, поэтому приходится использовать уже имеющиеся на производстве решения, наиболее распространёнными из которых являются серверы с CPU и мобильные устройства. В данной работе будет рассматриваться оптимизация моделей под CPU. Для решения многих задач видеоаналитики применяется модель, последовательно выполняющая детектирование объектов на изображении, их сегментацию и классификацию. Наиболее сложным с точки зрения вычислений в данной последовательности является этап детектирования объектов. Для переноса модели компьютерного зрения на CPU можно выделить существующие подходы: использование фреймворков, предоставляющих реализацию квантования, прунинга и более эффективных инструментов развёртки, изначальное проектирование архитектуры модели с расчётом на её дальнейший запуск на CPU. В ходе проведённого исследования архитектур существующих решений задачи детектирования объектов на CPU были рассмотрены yolo8n/s, yolo11n/s, Faster R-CNN, MobileNet-SSD, Cascade R-CNN, CornerNet, CenterNet, CenterNet++ и их модификации. Было выявлено, что CenterNet [3] и yolo8n [2] с использованием фреймворка DeepSparse являются наиболее эффективным решением на данный момент.

Основная часть. Адаптировать модель компьютерного зрения для выполнения на CPU предполагается как с помощью изменения уже обученной модели, так и с помощью разработки архитектуры модели, которая будет эффективно работать с предлагаемыми оптимизациями. При достижении поставленной цели, решаются следующие задачи:

- 1) Разработка модели на основе идей, используемых в CenterNet и её модификациях, таких как обнаружение объектов за счёт предварительного поиска ключевых точек и использование пирамиды признаков (FPN) [4] для более точного обнаружения маленьких объектов. Проверка полученной модели, сравнение полученных значений mAP и времени обработки одного кадра с наиболее эффективным решением - yolo8n.
- 2) Применение к полученной модели графовых оптимизаций, квантизации, прунинга для повышения скорости её работы.

Вывод. Произведён анализ существующих моделей компьютерного зрения, подходящих для решения поставленной задачи и предложены шаги для получения более эффективного решения по соотношению точности и скорости работы.

Список использованных источников:

1. Щемелинин В.Л., Лебедева А.Ю. Прорывные кейсы по видеоаналитике на предприятиях СИБУР. // Системы безопасности. - 2023. № 6(174). С. 44-46.
2. Muhammad Yaseen, What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. // arXiv:2408.15857 – 2024. - С. 1-9.
3. Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian, CenterNet: Keypoint Triplets for Object Detection. // arXiv:1904.08189 – 2019. - С. 1-9.
4. Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian, CenterNet++ for Object Detection. // arXiv:2204.08394 – 2022. - С. 1-11.