«Мобильные системы компьютерного зрения»

Лабораторная №3

«Реализация нейронной сети на базе Jetson Nano»

Цель работы

Изучить основы реализации глубоких нейронных сетей на мобильных системах, а также методы их оптимизации.

Задание

1. Изучить принципы построения глубоких нейронных сетей, их разновидности, архитектуры (применительно к обработке изображений и видео).
2. Изучить способы реализации нейросетевых вычислений (CPU, GPU).
3. Реализовать систему обработки изображений на основе нейронной сети (назначение и архитектуру сети выбрать самостоятельно, это может быть предобученная сеть для детектирования объектов, сегментации, классификации, построения карты глубины, вычисления оптического потока). Реализация обучения сети не требуется. Приложение должно принимать на вход реальное изображение (изображения) и выводить результат (обработанное изображение или полученную из него информацию, рис. 1).

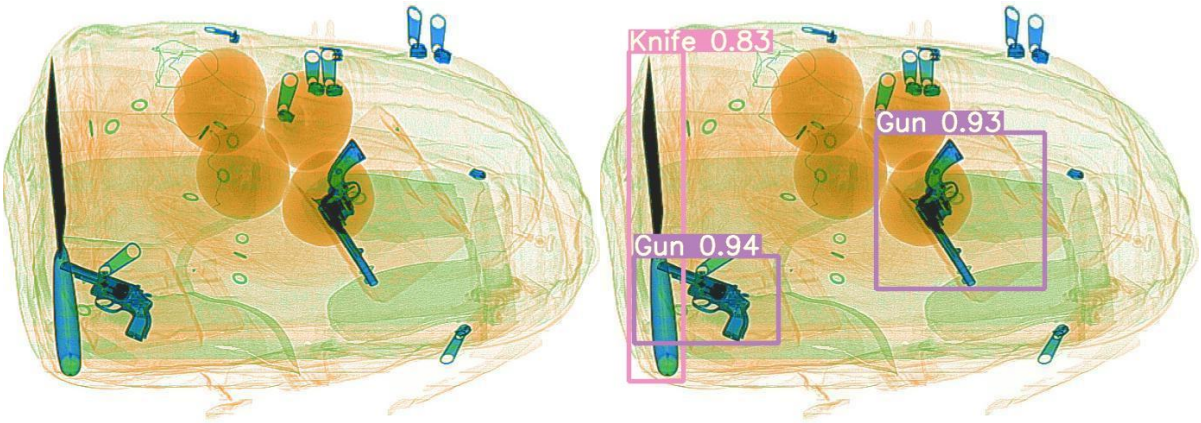

Рис. 1. Примеры входного и выходного изображений

4. Оптимизировать выбранную сеть с помощью TensorRT.
5. Оценить следующие характеристики:
   - 5.1. Время выполнения программы и количество используемой памяти при использовании сети без оптимизации.
   - 5.2. Производительность и потребление памяти при использовании TensorRT.
   - 5.3. Изменение выхода сети при использовании TensorRT при одинаковых входных данных.
   - 5.4. Возможность применения реализованной системы в real-time приложениях.
6. Измерение скорости выполнения алгоритма должно быть выполнено несколько раз с последующим усреднением для минимизации влияния степени загруженности вычислительных ресурсов другими процессами. Отдельно необходимо измерить время загрузки весов сети в память и непосредственно обработки изображений (в потоке).

Инструментальные средства

Лабораторная работа выполняется на языке Python с использованием библиотек pytorch, torchvision, TensorRT, в качестве платформы используется одноплатный компьютер Jetson Nano.

Материалы и пособия

1. ПО для работы с Jetson от NVIDIA
   https://developer.nvidia.com/embedded/develop/software
2. Machine Learning Mastery
   https://machinelearningmastery.com/
3. PyTorch
   https://pytorch.org/
4. TensorRT
   https://developer.nvidia.com/tensorrt
5. PyTorch to TensorRT
   https://github.com/NVIDIA-AI-IOT/torch2trt
6. How to Convert a Model from PyTorch to TensorRT and Speed Up Inference
   https://www.learnopencv.com/how-to-convert-a-model-from-pytorch-to-tensorrt-and-speed-up-inference/
7. Jetson Stats
   https://pypi.org/project/jetson-stats/1.6.2/
8. Заготовки для выполнения заданий на языке Python
   https://github.com/zeanfa/mobileCV/lab3/src

Критерии оценивания выполнения работы

По результатам работы должен быть подготовлен отчет в электронном виде.
Максимальный балл – 10. Работа считается сданной при оценке минимум в 5 баллов.
Оценка складывается из следующих составляющих:

- Соответствие заданной функциональности – 0-3 баллов;
- Выполнены п. 5.1, 5.2 и 5.3 задания 0-3 балла;
- Защита работы 0-3 балла;
- Составление отчета 0-1 балл.