ЛР3

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).

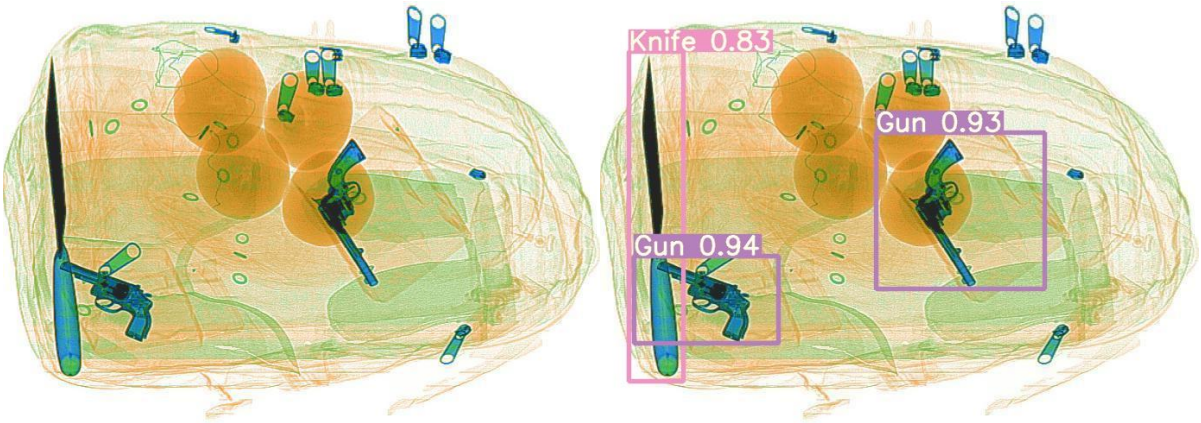

Рис. 1. Примеры входного и выходного изображений

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 (числовых значений) при использовании TensorRT при одинаковых входных данных.
   5.4 Возможность применения реализованной системы в real-time приложениях.

Отчёт должен содержать следующие пункты:

1. Теоретическая база
2. Описание разработанной системы (алгоритмы, принципы работы, архитектура)
3. Результаты работы и тестирования системы (скриншоты, изображения, графики, закономерности)
4. Выводы по работе
5. Использованные источники

Примечание.

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).

Можно взять любую архитектуру и задачу, например, ResNet-18 для задачи классификации.