УДК 004.93

УСКОРЕНИЕ ИНФЕРЕНСА НЕЙРОННЫХ СЕТЕЙ ПРЯМОГО РАСПРОСТРАНЕНИЯ.

Шеметов Ф.А. (Университет ИТМО)
Научный руководитель – Лукьянец Е.А. (ООО “ЦРТ-Инновации”)

Введение. Прямой проход (англ. inference) – действие, когда на обученную модель нейронной сети подается набор неизвестных данных, происходит их вычисление и в результате выдается прогноз, основанный на точности предсказания нейронной сети. Неоптимизированный прямой проход нейронной сети может занимать не мало времени на большом наборе данных чтобы спрогнозировать результат, из-за чего конечный результат не является эффективным. Как пример можно взять камеры с нейронными сетями в различных областях, где на вход в одну минуту подаётся огромное количество кадров и требуется быстрый прямой проход для получения прогноза в каждом кадре.

Основная часть. Цель моей работы — оптимизация времени прямого прохода нейронной сети. В данной работе рассматриваются возможные оптимизации времени прямого прохода предобученной модели, используя перспективные фреймворки с C++-интерфейсом LibTorch (PyTorch) [1], ONNX Runtime (ONNX) [2], OpenVino [3], TensorFlow [4] на CPU архитектуре при одном потоке и нескольких потоках. Выбор фреймворков основан на результатах статьи “Оптимизация нейронных сетей прямого распространения” [5]. Для фреймворка LibTorch были проведены эксперименты с использованием опций сборки и использованием класса `c10::InferenceMode`, который отключает работы связанные с автоматическим дифференцированием (англ. Autograd) модели. Для фреймворка ONNX Runtime были проведены эксперименты с использованием опций сборки, провайдеров выполнения (англ. Execution Providers) [6] и 3-х уровней оптимизации графа. Для фреймворка Tensorflow были проведены эксперименты с использованием опций сборки. Также был взят дополнительно фреймворк OpenVino, где был использован инструмент Model Optimizer от OpenVino, который применяет свои собственные настройки оптимизации для предобученной модели.

Выводы. Результатами моей работы являются проведение экспериментов по оптимизации фреймворков для улучшения времени работы прямого прохода нейронной сети и получение итоговых метрик эффективности оптимизаций.

Список использованных источников

1. Pytorch – URL: https://pytorch.org/
2. ONNX – URL: https://onnx.ai/
3. OpenVino toolkit – URL: https://github.com/openvinotoolkit/openvino
4. Tensorflow. — URL: https://www.tensorflow.org/
5. Шеметов Ф. А. ОПТИМИЗАЦИЯ НЕЙРОННЫХ СЕТЕЙ ПРЯМОГО РАСПРОСТРАНЕНИЯ //XI Конгресс молодых учёных. – 2022. – С. 192-196.
6. Execution Providers – URL: https://onnxruntime.ai/docs/execution-providers/