УДК 004.93

Оптимизация нейронных сетей прямого распространения.

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

Аннотация. Объектом исследования является оптимизация времени прямого прохода нейронных сетей под разные вычислители и специализированные ядра. Изучение существующих движков для нейронных сетей.

Введение. В настоящий момент существуют различные движки для нейронных сетей прямого распространения на различных вычислителях (CPU/GPU/Мобильные процессоры). Каждый движок обладает своим набором оптимизаций для нейронных сетей, так как есть разные серии вычислителей с разными поддерживаемыми операциями, есть разные операционные системы и также другие возможные комбинации, где оптимизация работает по-другому.

Основная часть. Цель моей работы — это обзор доступных движков нейронных сетей и вычислителей. Из доступных движков я рассматриваю такие как: TensorFlow, PyTorch, ONNX, CAFFE, DL4J, Theano, Keras, MXNet, Microsoft Cognitive Toolkit. Чтобы выбрать наиболее перспективные движки, нужно выделить некоторые параметры, по которым можно их сравнить, а это: потребление ресурсов, сложность встраивания, возможность параллелизма, сложность применения к существующим моделям, скорость работы, поддержка, насколько активно развивается. Дальше в своей работе я рассматриваю разные серии вычислителей (CPU/GPU/ARM) и их архитектуру для последующей оптимизации движков. Графический процессор представляет собой устройство, которое имеет огромное количество ядер и специальную архитектуру, что в результате позволяет эффективнее и быстрее выполнять операции с матрицами, чем центральный процессор. Также у каждых производителей GPU, есть свои серии устройств, которые имеют специальные ядра, механизмы и архитектуру для нейронных сетей. Это даёт возможность уменьшить время прямого прохода. Примером можно взять видеокарту Nvidia Turing или AMD Radeon Instinct. Относительно графического процессора, центральный процессор имеет более быстрый доступ к памяти, лучше кэширует данные и имеет другую архитектуру, но из-за меньшего количество ядер, уступает GPU в скорости работы с логическими операциями. ARM-чип это мобильный процессор с архитектурой ARM. Используется в большом количестве телефонов. Особенность мобильного процессора ARM в том, что он более компактный, чем CPU и GPU, более специализированный, а архитектура более ограничена. Для примера можно взять процессор Apple A11, в который входит микропроцессор ARM, где внутри ARM-чипа встроен Neural Engine – ускоритель для аппаратного ускорения алгоритмов нейронных сетей.

Выводы. Результатами моей работы являются сравнение различных движков для нейронных сетей, выбор наиболее перспективных движков: TensorFlow, PyTorch, ONNX, исследование разных серий вычислителей и выбор основных для оптимизации движков: GPU, CPU. В дальнейшем, планируется работа с ARM чипами.