УДК 004.021, 004.272.25

УСКОРЕННАЯ ОБРАБОТКА СВЕРТОЧНОГО СЛОЯ НЕЙРОННОЙ СЕТИ БЛОЧНЫМИ АЛГОРИТМАМИ ПРИ УВЕЛИЧЕННОМ ШАГЕ ФИЛЬТРА

Насыров К. Н., **Перминов И. В.**
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.ф.-м.н., доцент Жданов Д. Д.
(Университет ИТМО, г. Санкт-Петербург)

Сверточные нейронные сети являются важным инструментом в обработке цифровой информации и изображений, в частности. Высокие показатели точности распознавания, а также простота базовых операций, лежащих в основе данных сетей, и адаптируемость архитектур к новым задачам, гарантировали популярность данного инструмента в машинном обучении.

Однако операция свертки, которая является ключевой в сетях данного типа, несмотря на свою простоту является чрезвычайно ресурсозатратной, что ограничивает производительность сверточных сетей. Узким местом в данной операции может быть как вычисление самой свертки, так и выборка данных из памяти, всё зависит от характеристик конкретного слоя сети и характеристик используемого оборудования.

Многие подобные проблемы эффективно решаются блочными алгоритмами, такими как GEMM, FFT и Winograd, однако подобные алгоритмы не позволяют эффективно обрабатывать слои с модифицированным поведением и строением фильтра.

Цель данной работы разработать метод позволяющий эффективно использовать блочные алгоритмы на нестандартных сверточных слоях в нейронных сетях. Метод основывается на представлении исходной модифицированной свертки через несколько более простых сверток.

В ходе работы рассматривается случай модифицированного шага свертки и описывается метод преобразования позволяющий вычислить данную свертку блочными алгоритмами без лишних вычислений. Продемонстрирован возможный вариант использование с алгоритмами GEMM и Winograd.