УДК 004.021

Оценка вычислительной эффективности алгоритмов прямой свертки в нейронных сетях на центральных процессорах архитектуры ARM

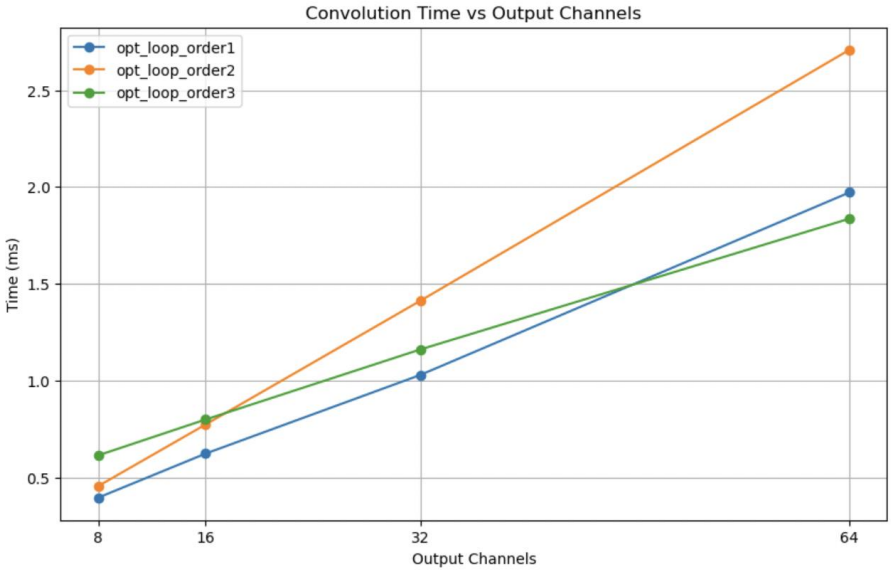
Л.И. Левин[1]

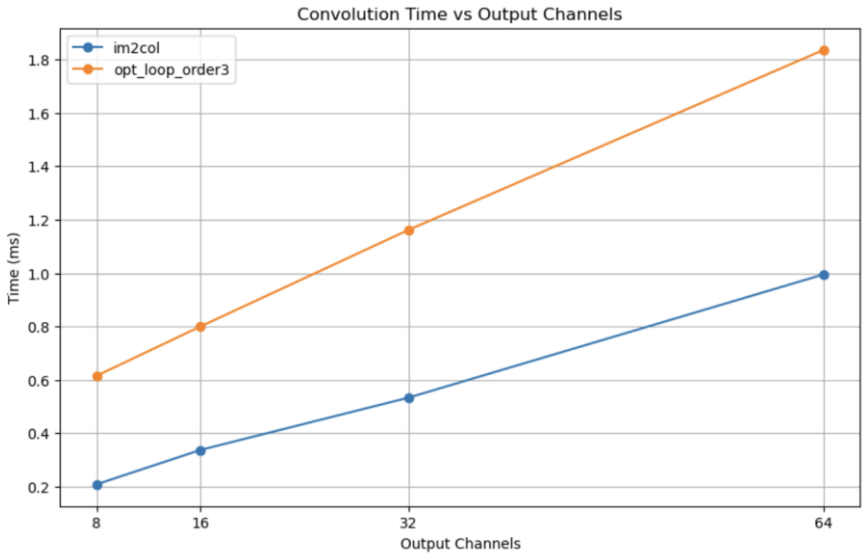
[1]Московский физико-технический институт (национальный исследовательский университет)

В настоящее время популярность набирают процессоры архитектуры ARM. Они более энергоэффективные, что делает их идеальным выбором для мобильных устройств. Стоят они дешевле своих конкурентов и используются в большом количестве устройств, начиная от смартфонов и планшетов до автомобилей и бытовой техники, что способствует развитию экосистемы вокруг архитектуры.

Данная научная работа посвящена исследованию эффективности алгоритма прямой свертки на процессорах архитектуры ARM. Сначала был написан алгоритм прямой свертки в разных версиях, отличающихся порядком циклов. Выбор правильного порядка циклов может существенно влиять на производительность программы из-за особенностей работы с кэш-памятью и кэширования. Дальше в каждой версии к внутреннему циклу был добавлен “tiling”. Идея “tiling” заключается в разбиение обрабатываемого массива на более мелкие блоки, что поможет оптимизировать доступ к памяти, уменьшая количество обращений к ней за счет эффективного использования локальности данных и кеширования. Затем к алгоритму прямой свертки был добавлен Simd-vectorization. Данная технология позволяет одной инструкции выполнять операции над несколькими элементами данных одновременно.

Эффективность алгоритма познается в сравнении. Для сравнения был выбран алгоритм im2col, основанный на матричном умножении. Оптимизации ускорили время работы алгоритма прямой свертки более чем в 4 раз, однако этого не хватило, чтобы конкурировать с алгоритмом im2col. Не зависимо от размера изображения, количества входным и выходным каналов, размера ядра оптимизированный алгоритм прямой свертки работает медленнее.





Литература

1. A.V. Trusov, E.E. Limonova, D.P. Nikolaev, V.V. Arlazarov. p-im2col: Simple Yet Efficient Convolution Algorithm With Flexibly Controlled Memory Overhead
2. Mirco Mannino, Biagio Peccerillo, Andrea Mondelli, Sandro Bartolini. Analysis and Optimization of Direct Convolution Execution on Multi-Core Processors