УДК 004.932.2

Разработка средств автоматизированного синтеза аппаратных ускорителей сверточных нейронных сетей

Егиазарян В.А (Университет ИТМО)

Научный руководитель - к.т.н., доцент, Быковский С.В. (Университет ИТМО)

В работе производится исследование процесса автоматизированного синтеза аппаратных ускорителей сверточных нейронных сетей и анализ существующих решений и подходов. Также, закладываются основы собственного инструмента для осуществления синтеза нейросетевых вычислителей.

Сегодня нейронные сети (НС) применяются в большинстве существующих отраслей. Их относительно легко реализовать, используя высокоуровневые языки программирования и десятки прилагаемых к ним фреймворков. В данной работе внимание сосредоточено на работе с изображениями и, как следствие, на так называемых свёрточных нейронных сетях (СНС), ввиду их исключительной способности к распознаванию паттернов. Чтобы достичь высокой скорости расчета выхода СНС при сохранении низкого энергопотребления и контроля за используемыми ресурсами, в работе используются специально реализованные аппаратные блоки. Однако, реализация нейронной сети на аппаратной платформе – сложный итеративный процесс, требующий узкоспециализированных знаний. Именно этот факт препятствует повсеместному распространению такого подхода, и он же делает невероятно актуальным вопрос создания средств автоматизации высокоуровневого синтеза подобных вычислителей.

Цель данной работы - создание инструмента для автоматизированного синтеза аппаратных ускорителей сверточных нейронных сетей, позволяющего снизить время разработки нейросетевых вычислителей.

Анализ существующих решений показал, что существует два подхода к синтезу нейросетевых вычислителей:

- автоматическая трансляция синтаксических конструкций целевого языка в более низкоуровневое представление с использованием специального компилятора LLVM (Low Level Virtual Machine) и их дальнейшая оптимизация;
- создание и анализ специальных конфигурационных файлов, описывающих структуру сети и воспроизведение этой структуры с использованием собственной библиотеки на синтезируемом языке;

Оба подхода имеют свои преимущества и недостатки, однако ни один, ни другой в чистом виде не применимы в реальных задачах и требуют дополнительных оптимизаций на различных этапах синтеза. В данной работе было решено склоняться ко второму подходу, потенциально позволяющему достичь большей производительности.

В результате исследования был сделан вывод о целесообразности использования аппаратных ускорителей для расчета выхода нейронных сетей, подтвержденный соответствующим экспериментом. Был проведен обзор и анализ существующих средств автоматизированного синтеза аппаратных ускорителей сверточных нейронных сетей, предложена концепция инструмента, сочетающего в себе лучшие свойства имеющихся продуктов и минимизирующего их недостатки. Также были рассмотрены основные этапы синтеза и применяемые на них методы для оптимизации. Выявлены зависимости точности расчета сети, величины ошибки, количества используемых сетью аппаратных блоков от различных параметров сети и способа ее реализации. Был заложен базис собственной синтезируемой библиотеки для оптимальной реализации НС с учетом аппаратных ограничений.