УДК 004.89

ПОДХОДЫ К СЖАТИЮ НЕЙРОСЕТЕЙ ДЛЯ ИСПОЛЬЗОВАНИЯ В СИСТЕМАХ С НИЗКИМ ЭНЕРГОПОТРЕБЛЕНИЕМ

Наумкин Д.А. (Национальный исследовательский университет ИТМО, Санкт-Петербург)
Научный руководитель – к.т.н., доц. Хлопотов М.В.
(Национальный исследовательский университет ИТМО, Санкт-Петербург)

В докладе рассматривается применение альтернативных числовых типов данных для хранения весовых коэффициентов глубоких нейросетей с целью уменьшения объема требуемой памяти при использовании в системах с ограниченными ресурсами. Анализируются преимущества и недостатки каждого подхода. Предлагаются способы подбора параметров типов для совместного применения с другими технологиями сжатия.

Введение. Многие прикладные задачи, использующие глубокие нейросети, могут быть решены более эффективно при использовании микроконтроллеров, ввиду низкого энергопотребления, большей доступности и возможности установки в непосредственной близости от источника обрабатываемых данных. При этом объемы памяти, необходимые для хранения весовых коэффициентов и промежуточных результатов, продолжают увеличиваться, что создает сложности при решении действительно важных задач во встраиваемых системах.

Существующие методы сжатия нейросетей, основанные на модификации архитектуры, требуют больших временных затрат и не всегда сохраняют приемлемую точность вычислений. Использование же более эффективного численного типа данных могло бы значительно уменьшить объем весовых коэффициентов, при этом сохранив структуру модели.

Основная часть. В 2017 году John L. Gustafson предложил тип данных (posit), схожий с числами с плавающей точкой, но имеющий более широкий диапазон значений и требующий меньшего количества бит для хранения, за счет изменяющейся по размеру части, кодирующей дополнительный множитель. В то же время, в области глубокого обучения все большую популярность приобретает использование чисел с фиксированной точкой с однозначным переводом float в int8, также с вынесением множителя. Хранение чисел с фиксированной точкой в int8 помимо сокращения размера позволяет ускорить вычисления, исключить из микроконтроллера дополнительных блок для операций с float, а также уменьшить энергопотребление. При этом точность вычислений зависит от диапазона значений отдельного тензора и будет ухудшаться с его увеличением.

Posit могут занимать еще меньшее количество бит (встречаются конфигурации с 5-битными posit), при этом сохраняя хорошую точность вычислений в диапазоне [-1, 1], что даст дополнительные преимущества при использовании пакетной нормализации после каждого слоя.

В работе ставилась задача анализа и практического применения вышеописанных подходов при расширении функциональности библиотеки машинного обучения для микроконтроллеров. В качестве дополнительной задачи был рассмотрен подбор параметров новых численных типов одновременно с прунингом.

Выводы. В результате проведенного исследования была доказана эффективность применения posit и fixed-point чисел для хранения весовых коэффициентов нейросети, выявлены ограничения при использовании, а также показаны преимущества использования совместно с обнулением весовых коэффициентов, близких к нулю.

Наумкин Д.А. (автор) Подпись

Хлопотов М.В. (научный руководитель) Подпись