Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Квантование моделей ИИ для исполнения на процессорах общего назначения

Авторы: д.т.н. В.В. Арлазаров, Е.Е. Лимонова, Д.П. Николаев, А.В. Трусов

Для повышения вычислительной эффективности нейронных сетей на процессорах с ограниченными ресурсами (например, на мобильных устройствах, тонких клиентах или бортовых устройствах) используется квантование. Квантование позволяет сокращать время исполнения нейронных сетей за счет уменьшения разрядности весов и функций активации. В результате исследований созданы три новых метода квантования, позволяющие улучшить соотношение *точность/вычислительные требования* исполнения моделей без специализированных ускорителей.

4.6-битное квантование разработано для исполнения моделей высокой точности для процессоров общего назначения. Эксперименты показали, что 4.6-битные квантованные сети в 1,5-1,6 раза быстрее 8-битных на процессоре ARMv8. Сравнение схем квантования представлено на Рисунке.

Квантование на основе неопределенности UBQ предназначено для достижения предельной производительности бинарных нейронных сетей, обеспечивающий одновременно стабильное обучение и высокую производительность. UBQ превосходит предыдущие методы для сетей с малым числом коэффициентов и достигает сопоставимых результатов для больших сетей.

Впервые предложен метод квантования для биполярных морфологических нейронных сетей. Это позволяет повысить вычислительную эффективность ИНС при исполнении в FPGA/ASIC, поскольку такая модель не использует умножение. Экспериментально продемонстрировано, что базовые модели для классификации изображений могут быть квантованы без заметной потери точности.

В результате исследований разработаны новые технологии ИИ, позволяющие в условиях максимально эффективно использовать отечественную вычислительную базу при внедрении в реальных отраслях экономики, социальной сферы (включая сферу общественной безопасности) и в органах публичной власти.

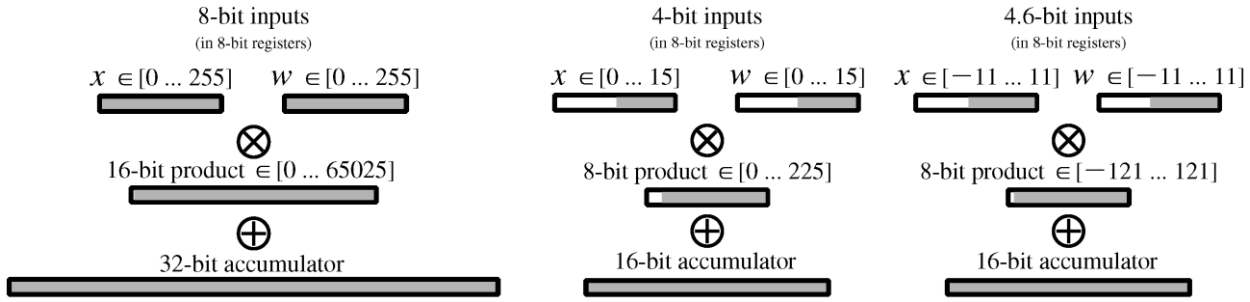

Рис. Иллюстрация схем квантования для разных разрядностей

Публикации:

1. *Limonova E., Zingerenko M., Nikolaev D., Arlazarov V.V.* Quantization method for bipolar morphological neural networks // Proc. SPIE, 2024. Vol. 13072: Sixteenth International Conference on Machine Vision (ICMV 2023). Art. 1307204.
2. *Trusov A.V., Putintsev D.N., Limonova E.E.* Uncertainty-based quantization method for stable training of binary neural networks // Computer Optics, 2024. 48(4): 573-581. DOI:10.18287/2412-6179-CO-1427.
3. *Trusov A., Limonova E., Nikolaev D., Arlazarov V.V.* 4.6-Bit Quantization for Fast and Accurate Neural Network Inference on CPUs // Mathematics, 2024. Vol. 12. Art. 651.