Оптимизация нейронных сетей (квантование, усечение):

Выполнить алгоритмическую оптимизацию нейронной сети для того, чтобы уменьшить число операций при вычислении результата и сократить длину слова.

На отдельно взятой архитектуре квантование до инт8 может даже увеличить время обработки, в следствии оптимизации под числа с ПЗ. В общем же случае оптимальное АЛУ инт8 всегда требует меньше логических элементов, потребляет меньше энергии, чем оптимальное АЛУ для fp32.

В нейросетях есть некоторая избыточность. Это доказывается тем, что изобретены методы сокращению числа каскадов операций в нейронных сетях.

Одновременно сокращаются и вычислительные и ресурсы, и память за счет того, что оптимизируется число операций, необходимых для прямого прогона нейронной сети.

Нынешние массовые чипы больше приспособлены к fp32 (большинство GPU), но в будущем это изменится, т.к. нынешние алгоритмы оптимизации нейронных сетей позволяют перейти на 8-битные слова.

Обзор источников:

1. Quantization Mimic: Towards Very Tiny CNN for Object Detection

1805.02152.pdf

1. DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING
2. INCREMENTAL NETWORK QUANTIZATION: TOWARDS LOSSLESS CNNS WITH LOW-PRECISION WEIGHTS
3. Analysis of effects of quantization in multilayer neural network using a statistical model
4. BinarizedNeuralNetworks: TrainingNeuralNetworkswithWeightsand ActivationsConstrainedto +1 or−1 1602.02830.pdf
5. RELAXED QUANTIZATION FOR DISCRETIZED NEURAL NETWORKS 1810.01875.pdf