Learning from low precision samples

Ji In Choi [*], Madeleine Georges [*], Jung Ah Shin [*], Olivia Wang [*], Tiffany Zhu [*], Tapan Shah [†]

Abstract

With advances in edge applications for industry and healthcare, machine learning models are increasingly trained on the edge. However, storage and memory infrastructure at the edge are often primitive, due to cost and real-estate constraints. A simple, effective method is to learn machine learning models from quantized data stored with low arithmetic precision (1-8 bits). In this work, we introduce two stochastic quantization methods, dithering and stochastic rounding. In dithering, additive noise from a uniform distribution is added to the sample before quantization. In stochastic rounding, each sample is quantized to the upper level with probability p and to a lower level with probability 1-p. The key contributions of the paper are

1. For 3 standard machine learning models, Support Vector Machines, Decision Trees and Linear (Logistic) Regression, we compare the performance loss for a standard static quantization and stochastic quantization for 55 classification and 30 regression datasets with 1-8 bits quantization.
2. We showcase that for 4- and 8-bit quantization over regression datasets, stochastic quantization demonstrates statistically significant improvement.
3. We investigate the performance loss as a function of dataset attributes viz. number of features, standard deviation, skewness. This helps create a transfer function which will recommend the best quantizer for a given dataset.
4. We propose 2 future research areas, a) dynamic quantizer update where the model is trained using streaming data and the quantizer is updated after each batch and b) precision re-allocation under budget constraints where different precision is used for different features.

[*]Columbia University, Equal Contribution

[†]GE Global Research