World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:13, No:10, 2019

Fast Adjustable Threshold for Uniform Neural Network Quantization

Authors : Alexander Goncharenko, Andrey Denisov, Sergey Alyamkin, Evgeny Terentev

Abstract : The neural network quantization is highly desired procedure to perform before running neural networks on mobile devices. Quantization without fine-tuning leads to accuracy drop of the model, whereas commonly used training with quantization is done on the full set of the labeled data and therefore is both time- and resource-consuming. Real life applications require simplification and acceleration of quantization procedure that will maintain accuracy of full-precision neural network, especially for modern mobile neural network architectures like Mobilenet-v1, MobileNet-v2 and MNAS. Here we present a method to significantly optimize training with quantization procedure by introducing the trained scale factors for discretization thresholds that are separate for each filter. Using the proposed technique, we quantize the modern mobile architectures of neural networks with the set of train data of only ∼ 10% of the total ImageNet 2012 sample. Such reduction of train dataset size and small number of trainable parameters allow to fine-tune the network for several hours while maintaining the high accuracy of quantized model (accuracy drop was less than 0.5%). Ready-for-use models and code are available in the GitHub repository.

Keywords : distillation, machine learning, neural networks, quantization
Conference Title : ICIAP 2019 : International Conference on Image Analysis and Processing
Conference Location : Paris, France
Conference Dates : October 29-30, 2019