ITC 3/50

Information Technology and Control
Vol. 50 / No. 3 / 2021
pp. 558-569
DOI 10.5755/j01.itc.50.3.28468

Quantization of Weights of Neural Networks with Negligible Decreasing of Prediction Accuracy

Received 2021/02/06 | Accepted after revision 2021/03/26

http://dx.doi.org/10.5755/j01.itc.50.3.28468

HOW TO CITE: Perić , Z. H., Denić, B. D., Savić, M. S., Dinčić, M. R., Mihajlov, D. I. (2021). Quantization of Weights of Neural Networks with Negligible Decreasing of Prediction Accuracy. *Information Technology and Control*, 50(3), 546-569. https://doi.org/10.5755/j01.itc.50.3.28468

Quantization of Weights of Neural Networks with Negligible Decreasing of Prediction Accuracy

Zoran H. Perić, Bojan D. Denić

University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia;
phone: +38118 529 367; e-mails: zoran.peric@elfak.ni.ac.rs, bojan.denic@elfak.ni.ac.rs

Milan S. Savić

University of Priština – Kosovska Mitrovica, Faculty of Sciences, Ive Lole Ribara 29, 38220 Kosovska Mitrovica, Serbia; e-mail: milan.savic1@pr.ac.rs

Milan R. Dinčić

University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia;
e-mail: milan.dincic@elfak.ni.ac.rs

Darko I. Mihajlov

University of Niš, Faculty of Occupational Safety, Čarnojevića 10 A, 18000 Niš, Serbia;
e-mail: darko.mihajlov@znrfak.ni.ac.rss

Corresponding author: bojan.denic@elfak.ni.ac.rs

Quantization and compression of neural network parameters using the uniform scalar quantization is carried out in this paper. The attractiveness of the uniform scalar quantizer is reflected in a low complexity and relatively good performance, making it the most popular quantization model. We present a design approach for the memoryless Laplacian source with zero-mean and unit variance, which is based on iterative rule and uses the minimal mean-squared error distortion as a performance criterion. In addition, we derive closed-form expressions for SQNR (Signal to Quantization Noise Ratio) in a wide dynamic range of variance of input data. To show effectiveness on real data, the proposed quantizer is used to compress the weights of neural networks using bit rates from 9 to 16 bps (bits/sample) instead of standardly used 32 bps full precision bit rate. The impact of weights compression on the NN (neural network) performance is analyzed, indicating good matching with the theoretical results and showing negligible decreasing of the prediction accuracy of the NN even in the case of high variance-mismatch between the variance of NN weights and the variance used for the design of quantizer, if the value of the bit-rate is properly chosen according to the rule proposed in the paper. The proposed method could be possibly applied in some of the edge-computing frameworks, as simple uniform quantization models contribute to faster inference and data transmission.

KEYWORDS: Uniform scalar quantization, variance-mismatch quantization, Laplacian distribution, quantized neural network, multilayer perceptron, MNIST database.