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.

1. Introduction

In recent times, a significant interest has been directed to the neural networks (NNs), mainly owing to the availability of powerful hardware [33]. The attractiveness of NNs lies in the increased potentiality to resolve challenges occurring in different research areas [33]. Some specific applications of NN can be found in papers [25–27, 29–31], where some promising results have been achieved. Namely, the implementations in image processing and virtual reality environment have been performed in [25] and [26], respectively. In addition, application of NN in image classification has been investigated in [27], where the ship classification problem is considered. The paper [29] applies NN to create the controller within automatic control system, while paper [31] considers a pointer NN for purpose of vehicle routing. In [30], the use of NN has been done in the context of solving four-class motor imagery classification problem.

The state-of-the-art neural networks (NNs) designed for tasks such as speech processing [5], image classification [15] and object recognition [28], just to name a few, represent very complex NN architectures, with a large number of parameters, requiring expensive computational and storage resources. On the other hand, high complexity can be a limiting factor for application in portable and edge computing devices with limited memory and processing power, or in latency-critical services. Hence, the compression of NN is required. To this end, the quantization is commonly employed, where the NN parameters (weights, biases, etc.), typically stored in 32-bits floating point format (full precision), are mapped to the fixed-point representations using lower bit lengths.

The influence of parameters quantization on the NN model performance is an active area of research, where the NN parameters have been quantized with 16-bits [20, 32], 8-bits [1, 9], 4-bits [3] or 2-bits [4]. Moreover, ternary [34] and binary (1-bit) quantization [10, 22] have also been taken into account. It has been shown that representations using higher number of bits (e.g. 8 to 16 bits) provide comparable performance with respect to full precision case, while performance deteriorates with decrease of code-word length (e.g. 2 to 4 bits); however, still offering competitive performance with very high compression ratios.

In the above-mentioned papers, the uniform scalar quantization (USQ) has dominantly been used. USQ was theoretically considered in [8, 13, 18, 19, 23, 24]. The main advantage of USQ is the design simplicity accompanied with relatively good performance when compared to more complex non-uniform quantization. Nevertheless, a detailed design process of the quantizer, taking into account the assumed statistical distribution of NN parameters, is missing in above mentioned papers [1, 4, 9, 10, 20, 32, 34] about quantization of NN parameters. In this paper we design USQ for compression of NN weights assuming Laplacian distribution of weights and bit rates from 9 to 16 bps. Namely, the Laplacian probability density function (PDF) has already been proved as a relevant model for various data including NN weights [2, 11], speech [7, 13] or images [13]. It is important to emphasize that we decided to consider a high-resolution quantization where one can expect a high level of reconstructed data quality. However, this not holds true in case when variance of the input data and the variance for which quantizer is designed are mismatched (assumes the utilization of non-adaptive quantizers), since this mismatch effect can cause a serious degradation in data quality. This fact motivated the authors

to focus the research at discovering how the degree of mismatch affects the performance of NN. In addition, using the proposed range of bit rates, compression ratios up to 3.56:1 can be achieved.

The analysis conducted in this paper is organized in two directions: development of theoretical model of USQ using asymptotic formulas (since the number of quantization levels N is large), making the design process simple; and implementation of the designed USQ for compression of NN weights. The main contributions of the paper are:

- A simple iterative design method of USQ for the memoryless Laplacian source with zero-mean and unit variance is proposed.
- The influence of the granular and overload distortions on SQNR for different values of variance are estimated, based on the derived closed-form expressions for performance evaluation in a wide dynamic range of variance of input data.
- The designed USQ is applied for quantization of weights of a neural network (Multi-Layer Perceptron) used for classification of images from MNIST database [21], showing very good matching between theoretical and experimental results. It should be highlighted that the variance-mismatched scenario (that often occurs in practice), meaning the mismatch between the variance of NN weights and the variance used for the design of the quantizer, is analyzed. This variance mismatched scenario has not been considered in any of previous papers from literature, related to the quantization of neural networks.
- A connection between SQNR of weights quantization and prediction accuracy of NN is shown and threshold for SQNR that assures a negligible decrease of the prediction accuracy is established for the specific NN. This is another new result that has not been presented in literature yet.
- It is shown that the significant decrease of the bit-rate R used for representation of weights, obtained by weights quantization, will produce a negligible decrease of NN prediction accuracy even in the case of high degree of the variance mismatch, if the value of the bit-rate R is chosen in an appropriate way, according to the rule provided in the paper.

The rest of the paper is organized as follows. Section 2 provides a detailed description of USQ and proposes a simple design method. In Section 3, the performance of USQ in a variance mismatched scenario is analyzed. In Section 4, the application of the designed USQ in neural networks is presented and the obtained results are discussed. Finally, Section 5 concludes the paper.

2. Uniform Scalar Quantization

In this section we will design USQ for the symmetric zero-mean Laplacian PDF defined with [13]:

$$q(x,\sigma_q)=\frac{1}{\sqrt{2}\sigma_q}\exp\left(-\frac{\sqrt{2}|x|}{\sigma_q}\right), \tag{1}$$

where σ_q^2 denotes signal variance. Without losing of generality, the design of USQ will be done for the unit variance $\sigma_q^2=1$, that is a standard approach in literature [13]. Due to the symmetry of the considered PDF, designed N-level USQ will have thresholds x_i and representation levels y_i symmetrical around zero: $x_i=i\cdot\Delta$, $x_{-i}=-x_i, (i=0,...,N/2)$, $y_i=(i-1/2)\cdot\Delta$, $y_{-i}=-y_i, (i=1,..., N/2)$ where $\Delta=2x_{\max}/N$ denotes the quantization step-size. Let $[-x_{\max}, x_{\max}]$ denote the support region of USQ, where the upper threshold of the support region $x_{\max}$ is also known as the maximal amplitude of the quantizer. To completely define USQ it is sufficient to know only one of these two parameters ($x_{\max}$ or Δ), since all required quantization parameters can be derived from them. The design goal of USQ is therefore constrained to determine the optimal value of the parameter $x_{\max}$ (or Δ), for the assumed input data distribution and established performance criteria.

Performance of a quantizer can be expressed by distortion D [8, 13] that represents the mean-square error occurred during quantization. Calculating distortion of USQ for data modeled with the Laplacian PDF in this section, we assume the variance-matched situation [8, 13, 16] which means that the variance σ_q^2 of the data being quantized is equal to the variance $\sigma_q^2=1$ used for the design of USQ. Since USQ divides the real line (i.e. the range of the input data values) into two regions: granular region defined in $[-x_{\max}, x_{\max}]$ and overload region defined in $(-\infty, -x_{\max}) \cup (x_{\max}, +\infty)$, the introduced distortion D is composed of the granu-

lar distortion (denoted as D_g) and overload distortion (denoted as D_{ov}). These components of distortion can be evaluated according to [8, 13]:

$$D_g = 2\sum_{i=1}^{N/2}\int_{x_{i-1}}^{x_i}\left(x-y_i\right)^2 q\left(x\right)dx\,, \tag{2}$$

$$D_{ov} = 2\int_{x_{\max}}^{\infty}\left(x-y_{N/2}\right)^2 q\left(x\right)dx\,, \tag{3}$$

where $q(x)\equiv q(x,\sigma_q=1)$. On the other hand, since N is high, it is appropriate to apply the asymptotic quantization theory [13], where the following holds for the components of the distortion:

$$D_g \approx \frac{\Delta^2}{12} = \frac{x_{\max}^2}{3N^2}\,, \tag{4}$$

$$D_{ov} \approx \exp\left(-\sqrt{2}x_{\max}\right). \tag{5}$$

Clearly, for total distortion we obtain:

$$D = D_g + D_{ov} \approx \frac{x_{\max}^2}{3N^2} + \exp\left(-\sqrt{2}x_{\max}\right). \tag{6}$$

The quantizer designed in this way is referred to as the asymptotic USQ.

Together with distortion, another quantity to express performance of the quantizer is SQNR defined as [8, 13]:

$$\begin{aligned}\mathrm{SQNR\,[dB]} &= 10\log_{10}\left(\frac{\sigma_q^2}{D}\Big|_{\sigma_q^2=1}\right) = -10\log_{10} D \\ &= -10\log_{10}\left(\frac{x_{\max}^2}{3N^2} + \exp(-\sqrt{2}x_{\max})\right)\end{aligned}\,. \tag{7}$$

Equation (6) shows that distortion is a function of $x_{\max}$. Therefore, the aim is to discover the optimal value of $x_{\max}$ for which distortion is minimal.

Lemma 1. The optimal value of $x_{\max}$ of the asymptotic USQ can be obtained using the following iterative rule:

$$x_{\max}^{(i+1)} = \frac{1}{\sqrt{2}}\log\left(\frac{3N^2}{\sqrt{2}x_{\max}^{(i)}}\right). \tag{8}$$

Proof. By determining the first derivation of distortion given by Eq. (6) with respect to $x_{\max}$ and equaling it with zero we obtain:

$$\frac{\partial D}{\partial x_{\max}} = \frac{2x_{\max}}{3N^2} - \sqrt{2}\exp\left(-\sqrt{2}x_{\max}\right) = 0\,. \tag{9}$$

Therefore, x_{max} can be calculated as:

$$x_{\max} = \frac{1}{\sqrt{2}}\log\left(\frac{3N^2}{\sqrt{2}x_{\max}}\right), \tag{10}$$

which shows that $x_{\max}$ can be specified iteratively, thus concluding the proof. As a good starting point of this iterative process we can choose $x_{\max}^{(0)} = \sqrt{2}\ln N$, that was proposed in [12] as an approximate solution for $x_{\max}$ of USQ. In this way, applying the iterative process, we calculate $x_{\max}$ in a more accurate manner than in [12].

Applying the previous iterative algorithm (8), we calculate optimal values of $x_{\max}$ for bit-rates $9 \le R$ [bps] ≤ 16, where $R = \log_2 N$; the generated codeword contains one bit for sign and R-1 bits for magnitude of the source sample. For those optimal $x_{\max}$ we calculate SQNR using (7). Calculated values of $x_{\max}$ and SQNR are presented in Table 1. Dependences of optimal values of $x_{\max}$ and SQNR on the bit-rate R are shown in Figure 1.

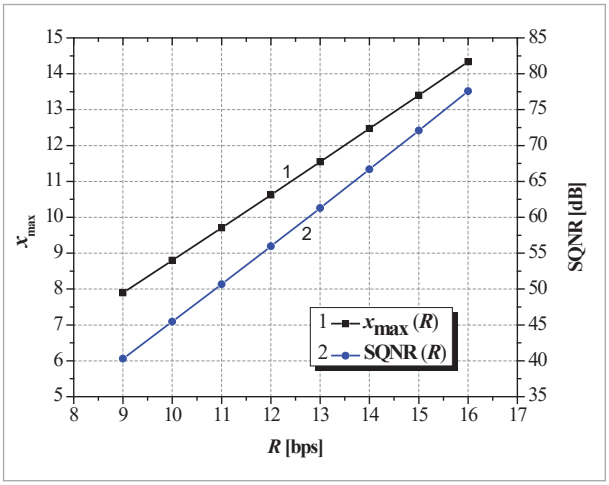
It can be seen that as R increases, both curves linearly increase with approximately constant slope. In particular, the slope of nearly 5.5 dB/bit has been observed in case of SQNR.

Table 1

Optimal values of $x_{\max}$ and corresponding values of SQNR [dB] for bit-rates $9 \le R$ [bps] ≤ 16

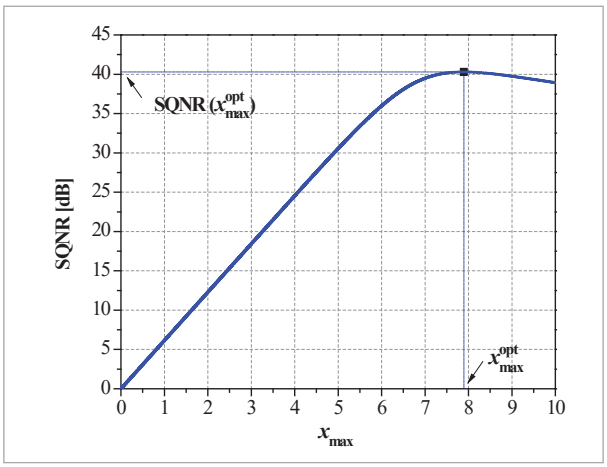
R [bps]	9	10	11	12	13	14	15	16
$x_{\max}$	7.89	8.80	9.71	10.62	11.55	12.47	13.40	14.33
SQNR [dB]	40.30	45.44	50.67	55.95	61.29	66.68	72.10	77.57

Figure 1
Dependences of optimal values of x_{max} and corresponding values of SQNR on R for the designed USQ



To show validity of the iterative process defined with (8), we can also perform numerical optimization of x_{max} for some specific value of R, by calculating SQNR for different values of $x_{max}$and finding the optimal value of x_{max} that gives the maximal SQNR. This numerical optimization of x_{max} is shown in Figure 2 for R = 9 bps. Obtained pair of optimal values (x_{max}, SQNR) perfectly matches with the corresponding values from Table 1 obtained by the iterative process (8), proving its validity.

Figure 2
SQNR dependence on x_{max} for the proposed USQ with N = 512 levels (R = 9 bps)



If we want to design USQ for some referent variance $\sigma_q^2 \neq 1$, the maximal amplitude should be calculated as

$$x_{\max}^{\sigma_q} = \sigma_q \cdot x_{\max}, \tag{11}$$

where x_{max} represents the maximal amplitude from Table 1 obtained for the unit variance $\sigma_q^2 = 1$.

3. Uniform Scalar Quantizer in a Wide Dynamic Range

Let us consider a real situation that USQ designed for the Laplacian PDF $q(x, \sigma_q = 1)$ is applied for quantization of data with Laplacian PDF $q(x, \sigma_p)$, i.e. we have variance mismatch: applied-to variance σ_p^2 differs from designed-for variance $\sigma_q^2 = 1$. Parameters of the quantizer (x_{max}, x_i, y_i) are the same as in Section 2, since they are determined for $\sigma_q^2 = 1$ during the design process of USQ. However, the variance mismatch will cause deterioration of performance (increasing of distortion and decreasing of SQNR) [16, 17]. It will be examined below.

In the case of the variance mismatch, both the granular D_g and the overload D_{ov} distortions will depend on σ_p^2:

$$D_g(\sigma_p) = \frac{x_{\max}^2}{3N^2} 2 \int_0^{x_{\max}} q(x, \sigma_p) dx = \frac{x_{\max}^2}{3N^2} \left(1 - \exp\left(-\frac{-\sqrt{2}x_{\max}}{\sigma_p}\right)\right), \tag{12}$$

$$D_{ov}(\sigma_p) = 2 \int_{x_{\max}}^{+\infty} (x - x_{\max})^2 q(x, \sigma_p) dx = \sigma_p^2 \exp\left(-\frac{\sqrt{2}x_{\max}}{\sigma_p}\right). \tag{13}$$

If we define the degree of mismatch $\rho = \sigma_p / \sigma_q$ as in [16], the total distortion becomes:

$$D(\sigma_p) = D_g(\sigma_p) + D_{ov}(\sigma_p) = \sigma_p^2 \left(\frac{x_{\max}^2}{3\rho^2 N^2} + \exp\left(-\frac{\sqrt{2}x_{\max}}{\rho}\right)\left(1 - \frac{x_{\max}^2}{3\rho^2 N^2}\right)\right). \tag{14}$$

Based on (15), we can express SQNR as:

$$\mathrm{SQNR}(\rho)=10\log_{10}\left(\frac{\sigma_p^2}{D(\sigma_p)}\right)$$
$$=-10\log_{10}\left(\frac{x_{\max}^2}{3\rho^2N^2}+\exp\left(-\frac{\sqrt{2}x_{\max}}{\rho}\right)\left(1-\frac{x_{\max}^2}{3\rho^2N^2}\right)\right). \quad (15)$$

Figure 3 analyzes SQNR of the optimal asymptotic USQ as a function of the degree of mismatch ρ in the range (-30 dB, 30 dB) for different bit rates (ranging from 9 to 16 bps).

Figure 3

SQNR versus ρ in wide dynamic range of input data variances, for the proposed USQ with different bit rates (the optimal values of $x_{\max}$ from Table 1 are used)

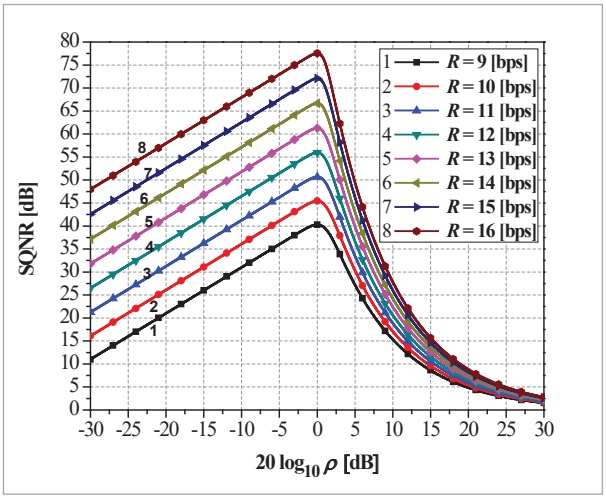
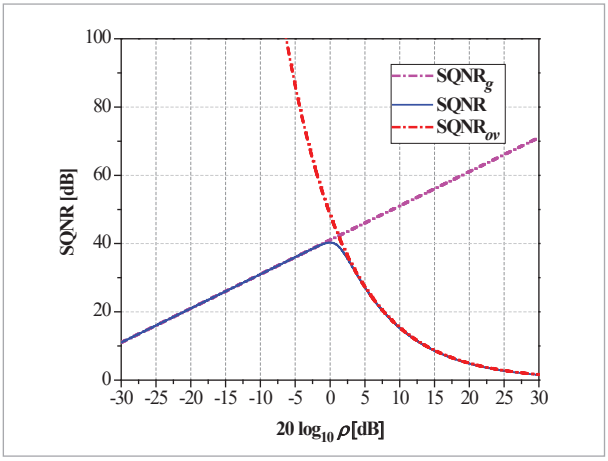


Figure 4

Total, granular and overload signal-to-quantization noise ratio (SQNR, SQNR_g and SQNR_{ov}) versus ρ for the proposed USQ (for R = 9 bps)



We can see in Figure 3 that, as expected, higher SQNR values is obtained as the bit rate R increases. Note that SQNR peaks for a variance matched case ($\sigma_p^2=\sigma_q^2$, ρ = 1, corresponding to 0 dB point in log-scale), but substantially drops if variances are not matched, decreasing more rapidly for $\rho>1$ ($\sigma_p^2>\sigma_q^2$) than for $\rho<1$ ($\sigma_p^2<\sigma_q^2$), due to the dominancy of overload distortion for $\sigma_p^2>0$ dB as we will see from Figure 4.

Let us define SQNR_g that depends only on D_g, as well as SQNR_{ov} that depends only on D_{ov}, using (12) and (13):

$$\mathrm{SQNR}_g=10\log_{10}\left(\frac{\sigma_p^2}{D_g}\right)$$
$$=10\log_{10}\frac{3\rho^2N^2}{x_{\max}^2\left(1-\exp\left(-\frac{\sqrt{2}x_{\max}}{\rho}\right)\right)}, \quad (16)$$

$$\mathrm{SQNR}_{ov}=10\log_{10}\left(\frac{\sigma_p^2}{D_{ov}}\right)=\frac{10\sqrt{2}x_{\max}}{\rho}\log_{10}e, \quad (17)$$

that are shown in Figure 4, together with the curve of total SQNR defined with (15) that takes into account the total distortion D, with the aim to examine the influence of the granular distortion D_g and the overload distortion D_{ov} on SQNR. We can see very good matching of SQNR and SQNR_g for $\rho<<1$, as well as very good matching of SQNR and SQNR_{ov} for $\rho>>1$. We can conclude the following from Figure 4:

- for $\rho<<1$, the granular distortion D_g is dominant and SQNR can be approximated with SQNR_g; since $\exp\left(-\sqrt{2}x_{\max}/\rho\right)<<1$ for $\rho<<1$, it follows from (16) that:

$$\mathrm{SQNR}_g=20\log_{10}\frac{\sqrt{3}\rho N}{x_{\max}}; \quad (18)$$

- for $\rho>>1$, the overload distortion D_{ov} is dominant and SQNR can be approximated with SQNR_{ov} defined with (17);
- in small range of ρ around 1 (i.e. 0 dB), both distortion components contribute to total SQNR, hence the full expression (15) should be used.

In order to compare performance of the designed USQ, we employ the quantizer (the uniform one) used in fixed-point format representations [6, 14],

conducting the analysis for R = 9 bps. In particular, the generated codeword of baseline quantizer consists of one bit reserved for sign (s = 1), n bits reserved for integer part and m bits reserved for fractional part of the fixed-point number. The maximal amplitude of this quantizer, denoted as $x_{\max}^{fp}$, can be calculated as:

$$x_{\max}^{fp} = \sum_{i=1}^{n} 2^{n-i} + \sum_{i=1}^{m} 2^{-i} \approx 2^{n}, \tag{19}$$

where the term $\sum_{i=1}^{n} 2^{n-i}$ refers to the integer part of the fixed-point number, while the term $\sum_{i=1}^{m} 2^{-i}$ reffers to the fractional part of the fixed-point number. For the purpose of analysis, two cases will be considered:

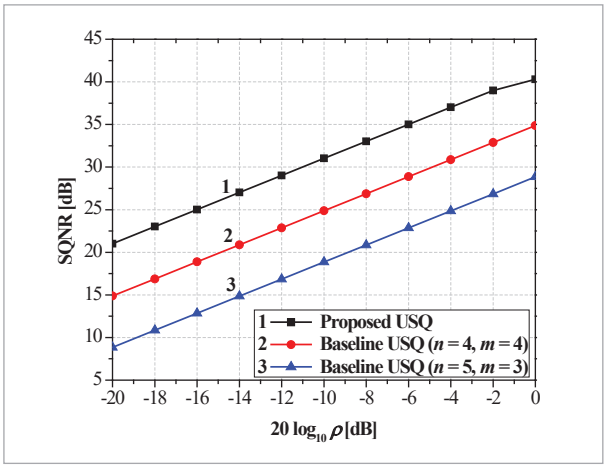
1 $s = 1, n = 4, m = 4$, and

2 $s = 1, n = 5, m = 3$.

Note that the expression defined with (15) is also relevant for performance evaluation of the baseline quantizer. The results are depicted in Figure 5. It can be observed that the proposed USQ significantly outperforms both versions of the baseline quantizer in terms of achieved SQNR values in a selected range of interest, $\rho \in$ [-20 dB, 0 dB].

Figure 5

SQNR vs. ρ for the proposed and baseline USQ with N = 512 levels (R = 9 bps)



4. Application in Neural Networks

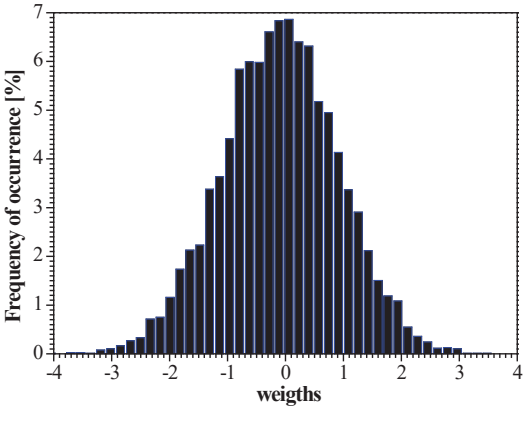
This section deals with the application of the developed USQ for compression of NN weights and analyzes the effects of quantization to the performance of NN for the image classification task.

As a proof of concept, we use Multi-Layer Perceptron (MLP) [33] that consists of input and output layers, with the goal to perform post-training quantization (i.e. to quantize the learned weights). The input of the NN is fed with the MNIST database [21], containing 60000 monochrome images of hand-written single digits of dimension 28 × 28 pixels, where 50000 images are used for training and 10000 images for testing. Note that the employed NN deals with the classification of grayscale images of hand-written digits into the corresponding category (0–9). Thus, input layer and output layer are constituted by 784 (28 × 28) and 10 (the number of digits) nodes, respectively. *Softmax* activation function is used at the output layer, while the learning rate and batch size are set to 0.5 and 250, respectively.

The employed NN is trained for 20 epochs achieving the prediction accuracy of 90.84%. The histogram of learned weights (total number amounts to 784 × 10 = 7840) is depicted in Figure 6. Observe that distribution of weights can be approximated well by the Laplacian PDF with the mean value very close to zero.

Figure 6

The histogram of weights of trained NN



Let $\sigma_w^2 = (1/W)\sum_{i=1}^{W} w_i^2$ denote the variance of weights. Let $D^w = (1/W) \cdot \sum_{i=1}^{W} (w_i - w_i^q)^2$ denote distortion obtained by the quantization of weights using USQ, where W is total number of weights, w_i is the original and w_i^q is the quantized value of i-th weight.

As performance measure for quantization of weights we can use SQNR^w defined as:

$$\mathrm{SQNR}^w = 10\log_{10}\left(\frac{\sigma_w^2}{D^w}\right) = 10\log_{10}\left(\frac{\sum_{i=1}^{W} w_i^2}{\sum_{i=1}^{W}(w_i - w_i^q)^2}\right). \quad (20)$$

In practice, the variance of NN weights can vary in wide range, hence the variance mismatch can occur between the variance of weights and the variance used for the design of USQ. Hence, our aim is to examine the influence of this variance mismatch on the prediction accuracy of NN, applying the following procedure:

- firstly, design USQ for a specific value of R from 9 to 16 bps, for the variance equal to the variance of the learned weights (i.e. $\sigma_q^2 = \sigma_w^2$), using (8);
- starting from the original set of learned weights with the variance σ_w^2, make another set of weights with the variance $\rho^2\sigma_w^2$ by multiplication of each original weight with ρ;
- perform variance mismatched quantization of weights with the variance $\rho^2\sigma_w^2$ using USQ designed for the variance σ_w^2;
- calculate SQNR^w for the variance mismatched quantization;
- apply the quantized weights for classification purposes on the test data (10000 images form MNIST database [21]);
- calculate the prediction accuracy of NN with the quantized weights; just to recall, the prediction accuracy score obtained without quantization was 90.84%.

The previous procedure can be repeated for different values of ρ, as well as for all values of R from 9 to 16 bps.

Based on the previous procedure, the influences of the variance mismatch on the quality of quantization of NN weights (i.e. on SQNR^w), as well as on the prediction accuracy of NN with quantized weights can be found, as being shown in Figures 7 and 8, respectively, for different values of ρ and in the range of the bit-rate R from 9 to 16 bps.

We can see from Figure 7 that SQNR^w approximately linearly increases with R for a given ρ, while the highest SQNR^w is achieved for $\rho = 0$ dB (variance matched scenario), as expected. Note also, for a given R and ρ, that SQNR^w from Figure 7 matches very well with the theoretical SQNR presented in Figure 3.

For this specific MLP neural network it is empirically found that the decreasing of the prediction accuracy of the network due to quantization of weights is neglecting if $\mathrm{SQNR}^w \geq 16$ dB for quantization of weights. Based on (15), we can theoretically found ranges of ρ where SQNR ≥ 16 dB, that is shown in Table 2 for the bit-rates R from 9 to 16 bps.

Figure 7
SQNR^w of USQ for bit rates in the range $9 \leq R \leq 16$ [bps] for different values of ρ

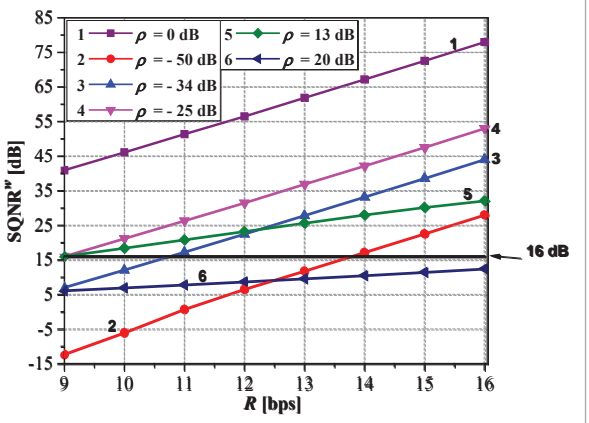


Figure 8
The prediction accuracy for image classification task applied on the MNIST dataset, after quantization of NN weights with different variances, using the designed USQ

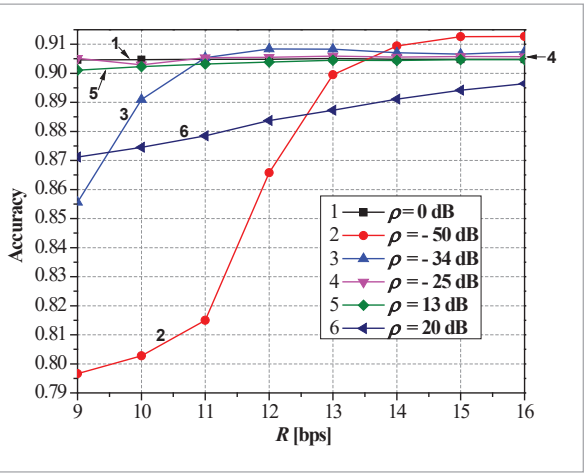


Table 2
The range of ρ [dB] where SQNR ≥ 16 dB, for different values of R

R [bps]	The range of ρ where SQNR ≥ 16 dB
9	(-25.02, 9.63) [dB]
10	(-30.09, 10.57) [dB]
11	(-35.25, 11.43) [dB]
12	(-40.50, 12.21) [dB]
13	(-45.79, 12.94) [dB]
14	(-51.14, 13.60) [dB]
15	(-56.54, 14.23) [dB]
16	(-61.98, 14.81) [dB]

From Figures 7 and 8 and from Table 2 we can derive the following conclusions:

- for ρ [dB] = 0 dB (i.e. ρ = 1), we obtain the $SQNR^w$ much higher than 16 dB for all $9 \le R$ [bps] ≤ 16 (Figure 7), providing excellent accuracy for all considered bit-rates (Figure 8), almost the same as accuracy in the full precision case; this is also theoretically expected, since it follows from Table 2 that ρ [dB] = 0 dB is acceptable for all considered bit-rates;
- for ρ [dB] = -50 dB (i.e. ρ = 0.003), we have $SQNR^w$ ≥ 16 dB for $R \ge 14$ bps (Figure 7); also, accuracy becomes acceptable for $R \ge 14$ bps, while for $R < 14$ bps there is a significant drop of accuracy (Figure 8); this is fully in line with theoretical results presented in Table 2 where ρ [dB] = -50 dB is acceptable for $R \ge 14$ bps;
- for ρ [dB] = -34 dB (i.e. ρ = 0.02), we have $SQNR^w \ge$ 16 dB and negligible loss of accuracy for $R \ge 11$ bps, but having drop of accuracy for $R < 11$ bps; this is fully in line with theoretical results from Table 2;
- for ρ [dB] = -25 dB (i.e. ρ = 0.056), we have $SQNR^w$ ≥ 16 dB and negligible loss of accuracy for all $9 \le R$ [bps] ≤ 16, being fully in line with theoretical results from Table 2 where ρ [dB] = -25 dB is acceptable for all considered bit-rates;
- for ρ [dB] = 13 dB (i.e. ρ = 4.467), we have $SQNR^w \ge 16$ dB and negligible loss of accuracy for all $9 \le R$ [bps] ≤ 16; in this case, experimental results are slightly better than theoretical results from Table 2;
- for ρ [dB] = 20 dB (i.e. ρ = 10), we have $SQNR^w < 16$ dB and drop of accuracy for all $9 \le R$ [bps] ≤ 16; this is fully in line with theoretical results from Table 2 where ρ [dB] = 20 dB is not acceptable for any of the considered bit-rates.

We can see that there is a very good matching between experimental results (shown in Figures 7 and 8) and theoretical predictions presented in Table 2. Also, we can see that the variance mismatch is acceptable in much wider range for negative ρ [dB] than for positive one.

We can conclude that the range of acceptable degree of the variance mismatch ρ depends on the bit-rate R. Increasing R allows wider range of the variance mismatch degree ρ (decreasing the compression ratio on the other hand). Hence, the bit-rate R should be chosen based on the range of the degree of the variance mismatch ρ for the specific application. We can define the following rule: we should choose the smallest R that allows maintaining of high prediction accuracy for given range of ρ for the specific application.

Finally, we provide the results in case of the baseline quantizer approach discussed in [6, 14], taking into account R = 9 bps. SQNR versus ρ, obtained from real data (weights), can be found in Figure 9, where good agreement with theoretical results in Figure 5 is observed. On the other hand, the prediction accuracy scores can be found in Figure 10, indicating that MLP achieves better performance in case of using the USQ proposed in this paper.

Figure 9
$SQNR^w$ of the proposed and baseline USQ for bit rate R = 9 bps and different values of ρ

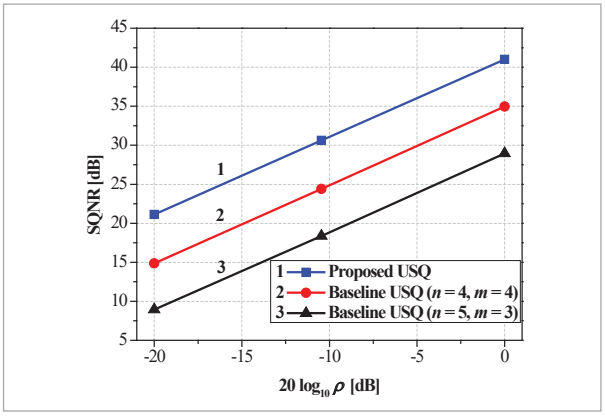
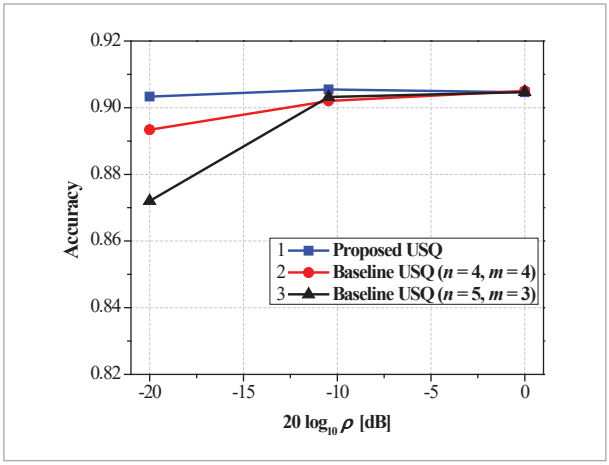


Figure 10

The prediction accuracy scores of MLP in image classification task in case when the proposed and two versions of baseline USQ (R = 9 bps) are applied, for different values of ρ



5. Conclusion

In this paper, USQ was designed for the Laplacian PDF and implemented for quantization of weights of MLP neural network. Firstly, the quantizer was designed for a reference variance and its performance was evaluated for both variance-matched and variance mismatched cases. Especially, it should be highlighted that we proposed a very efficient iterative algorithm for calculation of the most important parameter of the quantizer $x_{\max}$. Then, the designed USQ was applied for quantization of weights of MLP used for classification of images from MNIST database. It was shown a very good matching between experimental and theoretical results. Also, it was shown that almost the same prediction accuracy of the network can be achieved using quantized weights with significant decreasing of the bit-rate R [bps] as in the full precision case. Connection between SQNR of weights quantization and prediction accuracy of the neural network was established. Furthermore, the variance mismatched quantization of weights was considered (that is very important in practical applications where the variance mismatch often occurs), showing that even in this case a negligible decrease of accuracy can be achieved by choosing appropriate value of the bit-rate R. Acceptable ranges of the degree of the variance mismatch ρ [dB] are calculated for all considered bit-rates $9 \leq R$ [bps] ≤ 16 and the rule for choosing the right value of the bit-rate R was defined: the smallest value of R that allows maintaining of high prediction accuracy for given range of ρ[dB] for the specific application should be chosen. In addition, the benefit of the proposed USQ over the baseline quantizers available in the literature has also been shown.

Acknowledgement

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia and by the Science Fund of the Republic of Serbia (Grant No. 6527104, AI- Com-in-AI).

References

1. Banner, R., Hubara, I., Hoffer, E., Soudry, D. Scalable Methods for 8-bit Training of Neural Networks. Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, December 2–8, 2018.
2. Banner, R., Nahshan, Y., Hoffer, E., Soudry, D. ACIQ: Analytical Clipping for Integer Quantization of Neural Networks. arXiv preprint arXiv: 1810.05723, 2018.
3. Banner, R., Nahshan, Y., Soudry, D. Post Training 4-bit Quantization of Convolutional Networks for Rapid-Deployment. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, December 8–10, 2019.
4. Choi, J., Venkataramani, S., Srinivasan, V.,Gopalakrishnan, K., Wang, Z., Chuang, P. Accurate and Efficient 2-Bit Quantized Neural Networks. Proceedings of the 2nd SysML Conference, Stanford, CA, USA, March 31–April 2, 2019.
5. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y. Very Deep Convolutional Networks for text classification. arXiv preprint arXiv: 1606.01781, 2016. https://doi.org/10.18653/v1/E17-1104
6. Endericha, L., Timmb, F., Burgard, W. SYMOG: Learning Symmetric Mixture of Gaussian Modes for Improved Fixed-Point Quantization. Neurocomputing, 2020, 416, 310–315. https://doi.org/10.1016/j.neucom.2019.11.114

7. Gazor, S., Zhang, W. Speech Probability Distribution. IEEE Signal Processing Letters, 2003, 10, 204-207. https://doi.org/10.1109/LSP.2003.813679
8. Gersho, A., Gray, R. Vector Quantization and Signal Compression. Kluwer Academic Publishers, New York, 1992. https://doi.org/10.1007/978-1-4615-3626-0
9. Gong, J., Shen, H., Zhang, G., Liu, X., Li, S., Jin, G., Maheshwari, N., Fomenko, E., Segal, E. Highly Efficient 8-bit Low Precision Inference of Convolutional Neural Networks with IntelCaffe. arXiv preprint arXiv: 1805.08691, 2018. https://doi.org/10.1145/3229769
10. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y. Binarized Neural Networks. Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, December 5–10, 2016.
11. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. Journal of Machine Learning Research, 2018, 18, 1-30.
12. Hui, D., Neuhoff, D. L. Asymptotic Analysis of Optimal Fixed-Rate Uniform Scalar Quantization. IEEE Transactions on Information Theory, 2001, 47(3), 957-977. https://doi.org/10.1109/18.915652
13. Jayant, N. C., Noll, P. Digital Coding of Waveforms: Principles and Applications to Speech and Video. Prentice Hall, New Jersey, 1984. https://doi.org/10.1016/0165-1684(85)90053-2
14. Kim, S., Kim, H. Zero-Centered Fixed-Point Quantization with Iterative Retraining for Deep Convolutional Neural Network-Based Object detectors. IEEE Access, 2020, 9, 20828–20839. https://doi.org/10.1109/ACCESS.2021.3054879
15. Krizhevsky, A., Sutskever, I., Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. Proceedings of the International Conferenceon Neural Information Processing Systems, Harrahs and Harveys, Lake Tahoe, NV, USA, 2012, 1097-1105.
16. Na, S. Asymptotic Formulas for Mismatched Fixed-Rate Minimum MSE Laplacian Quantizers. IEEE Signal Processing Letters, 2008, 15, 13–16. https://doi.org/10.1109/LSP.2007.910240
17. Na, S., Neuhoff, D. L. Asymptotic MSE Distortion of Mismatched Uniform Scalar Quantization. IEEE Transactions on Information Theory, 2012, 58(5), 3169–3181. https://doi.org/10.1109/TIT.2011.2179843
18. Na, S., Neuhoff, D. L. Monotonicity of Step Sizes of MSE-Optimal Symmetric Uniform Scalar Quantizers. IEEE Transactions on Information Theory, 2018, 65(3), 1782–1792. https://doi.org/10.1109/TIT.2018.2867182
19. Na, S., Neuhoff, D. L. On the Convexity of the MSE Distortion of Symmetric Uniform Scalar Quantization. IEEE Transactions on Information Theory, 2017, 64(4), 2626–2638. https://doi.org/10.1109/TIT.2017.2775615
20. Nannarelli, A. Variable Precision 16-Bit Floating-Point Vector Unit for Embedded Processors. Proceedings of IEEE 27th Symposium on Computer Arithmetic, (ARITH 2020), Portland, OR, USA, June 7–10, 2020. https://doi.org/10.1109/ARITH48897.2020.00022
21. LeCun, Y., Cortez, C., Burges, C. The MNIST Handwritten Digit Database. Available online: yann.lecun.com
22. Peric, Z., Denic, B., Savic, M., Despotovic, V. Design and Analysis of Binary Scalar Quantizer of Laplacian Source with Applications. Information, 2020, 11(11), 501. https://doi.org/10.3390/info11110501
23. Peric, Z., Simic, N., Savic, M. Analysis and Design of Two Stage Mismatch Quantizer for Laplacian source. Electronika IR Electrotechnika, 21(3), 49-53, 2015. https://doi.org/10.5755/j01.eee.21.3.10380
24. Peric, Z., Vucic, N., Dincic, M. Ciric, D., Denic, B., Peric, A. Design of Uniform Scalar Quantizer for Discrete Input Signals. Proceedings of the 28th Telecommunications Forum (TELFOR 2020), Belgrade, Serbia, November 24–25, 2020. https://doi.org/10.1109/TELFOR51502.2020.9306535
25. Polap, D. An Adaptive Genetic Algorithm as a Supporting Mechanism for Microscopy Image Analysis in a Cascade of Convolution Neural Networks. Applied Soft Computing Journal, 2020, 97, Article ID: 106824, 11 pages. https://doi.org/10.1016/j.asoc.2020.106824
26. Połap, D., Kęsik, K., Winnicka, A., Woźniak, M. Strengthening the Perception of the Virtual Worlds in a Virtual Reality Environment. ISA Transactions, 2020, 102, 397–406. https://doi.org/10.1016/j.isatra.2020.02.023
27. Polap, D., Wlodarczyk-Sielicka, M. Classification of Non-Conventional Ships Using a Neural Bag-of-Words Mechanism. Sensors, 2020, 20(6), 1608. https://doi.org/10.3390/s20061608
28. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS), Montreal, Canada, 2015, 91-99.

29. Sheng, Y., Ma, H., Xia, W. A Pointer Neural Network for the Vehicle Routing Problem with Task Priority and Limited Resources. Information Technology and Control, 2020, 49(2), 237–248. https://doi.org/10.5755/j01.itc.49.2.24613

30. Su, J., Nakonechnyi, M., Sachenko, A. Developing the Automatic Control System Based on Neural Controller. Information Technology and Control, 2015, 44(3), 262–270. https://doi.org/10.5755/j01.itc.44.3.7717

31. Uktveris, T., Jusas, V. Application of Convolutional Neural Networks to Four-Class Motor Imagery Classification Problem. Information Technology and Control, 2017, 46(2), 260-273. https://doi.org/10.5755/j01.itc.46.2.17528

32. Zamirai, P., Zhang, J., Aberger, C. R., De Sa, C. Revisiting BFloat16 Training. arXiv preprint arXiv:2010.06192v1, 2020.

33. Zhang, A., Lipton, Z. C., Li, M., Smola, A. J. Dive into Deep Learning. Amazon Science, 2020.

34. Zhu, C., Han, S., Mao, H., Dally, W. J. Trained Ternary Quantization. Proceedings of the 5th International Conference on Learning Representations, (ICLR 2017), Toulon, France, April 24–26, 2017.