This work has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101160293 (AIDA4Edge).

Funded by the European Union

AIDA4Edge

Performance Comparison of Dlfloat16 and Bfloat16 Binary Formats

Zoran H. Perić, Bojan D. Denić, Sofija Z. Perić, Milan R. Dinčić and Marko Anđelković

Abstract—**The 16-bit floating-point format is a popular alternative to the widely used 32-bit floating-point format (FP32). This study analyzes the performance of two commonly employed 16-bit versions, dlfloat16 and bfloat16, for data with Laplacian and Gaussian distribution. The applied performance metric is the signal-to-quantization noise ratio (SQNR), allowed by using the connection between the floating-point format and piecewise uniform quantization. It is shown over a wide range of input data variances that SQNR results for Laplacian and Gaussian data differ slightly, indicating that both observed formats are robust to changes in the data distribution. SQNR analysis also suggested that dlfloat16 is a better option when increased calculation precision is required, while bfloat16 is better from a dynamic range perspective.**

***Keywords*—Bfloat16 format, Dlfloat16 format, Gaussian source, Laplacian source, Piecewise-uniform quantizer, SQNR.**

I. Introduction

The 32-bit floating-point format (FP32), defined by the IEEE 754 standard [1], has gained great popularity in the domain of computing [2]. Due to the remarkable quality of representation and a wide dynamic range, it has also been accepted in artificial intelligence domain, especially in neural networks [3]. However, the inherent complexity of the FP32 format was recognized as its key shortcoming. Implementing FP32-based neural networks, for instance, in resource-constrained environments such as edge devices [4] and sensor nodes [5] is very difficult. Lower-resolution floating-point formats have emerged as an alternative to FP32, allowing easier implementation of neural networks with minimal performance degradation.

The 16-bit floating-point format is well-known alternative option to FP32. Two popular 16-bit versions for neural network applications are dlfloat16 [6] and bfloat16 [7]. In these two formats, the exponent and mantissa of a real number are encoded using different numbers of bits. Bfloat16 uses 8 bits for the exponent (same as the FP32 format) and 7 bits for the mantissa, while in dlfloat16 the number of bits for the exponent is reduced to 6, and 9 bits are used for the mantissa. The aforementioned studies [6], [7] exhibit weakness in terms of strategies for assessing the level of representation quality (i.e., precision), which is of crucial importance for practical applications.

The strategy based on the signal-to-quantization noise ratio (SQNR) is presented in this study for both dlfloat16 and bfloat16 formats, taking into account data with Laplacian and Gaussian distributions. Namely, exploiting SQNR is allowed owing to the relationship between the floating-point format and piecewise uniform quantization [8]. In addition, the two mentioned distributions are commonly encountered in practice, in areas like signal processing (e.g., for statistical modeling of speech and image samples) [9], [10] and neural networks (e.g., for statistical modeling of weights) [11]. The SQNR strategy has already been applied in [8] for bfloat16 format and Laplacian source, where the stability in SQNR was observed across the variance range of 60 dB width. In that way, the robustness of bfloat16 in processing variance-sensitive data was indicated. The contribution of this study lies in the derivation of corresponding SQNR expressions, which allowed us to discover the behavior of the discussed formats on observed data distributions, as well as circumstances under which they can be successfully applied. Besides, we have considered a much wider variance range than that in [8].

The rest of the paper is organized as follows. The short description of the dlfloat16 is given in Section 2, which also introduces the analogous quantization model and the SQNR expressions for performance evaluation for both Laplacian and Gaussian distributions. Description of the bfloat16 format and its equivalent quantizer is given in Section 3, where the corresponding SQNR expressions are provided as well. The numerical results are presented and discussed in Section 4. The concluding remarks are given in Section 5.

II. Dlfloat16 Binary Format

A. Basics and Analogous Quantizer

Let x denotes the real number. Fig. 1 shows the binary representation of x using the dlfloat16 format [6], where s refers to the sign bit, while bits $e_1 ... e_6$ and $m_1 ... m_9$ are engaged for the exponent E and the mantissa M of x, respectively.

Zoran H. Perić, Bojan D. Denić, Sofija Z. Perić and Milan R. Dinčić are with the Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 4, 18000 Nis, Serbia (e-mails: zoran.peric@elfak.ni.ac.rs, bojan.denic@elfak.ni.ac.rs, sofija.peric@elfak.ni.ac.rs, milan.dincic@elfak.ni.ac.rs).

Marko Anđelković is with IHP – Leibniz-Institut für innovative Mikroelektronik, Im Technologiepark 25, 15236 Frankfurt (Oder), Germany (e-mail: andjelkovic@ihp-microelectronics.com).

$$x=\left(\underset{sign}{s}\;\underbrace{e_1 ... e_6}_{\text{exponent } E}\;\underbrace{m_1 ... m_9}_{\text{mantissa } M}\right)_2$$

Fig. 1. Encoding a real number x using dlfloat16 binary format.

The decimal equivalent of x can be obtained by using [1]:

$$x=(-1)^s 2^{E^*}\left(1+\frac{M}{2^9}\right), \tag{1}$$

where E^* = E−31 is the biased exponent. The range for the exponent E ($E=\sum_{i=0}^{5} e_{6-i} 2^i$) is from 0 and 63; E, however, falls between 1 and 62 as the lowest 0 and the highest 63 values are intended for special purposes. E^* is therefore in the range of −30 to 31. The mantissa M ($M=\sum_{i=0}^{8} m_{9-i} 2^i$) extends from 0 to 2^9−1. Let's consider only the positive dlfloat16 numbers (s = 0). According to Eq. (1), there are 2^9 distinct values of x for a given E^*, mutually shifted by:

$$\Delta_{E^*}^{\mathrm{dl}}=2^{E^*}\left(1+\frac{M+1}{2^9}\right)-2^{E^*}\left(1+\frac{M}{2^9}\right)=2^{E^*-9}. \tag{2}$$

Thus, the positive part of the real axis comprises 62 groups (total number of E^* values), each with 2^9 equidistant values and step size $\Delta_{E^*}^{\mathrm{dl}}$. Since dlfloat16 format is zero-symmetrical, the negative part is organized as the reflection of the positive part. In addition, for E^* = −30 and M = 0, Eq. (1) gives the minimal positive dlfloat16 number:

$$x_{\min}^{\mathrm{dl}}=2^{-30}\left(1+\frac{0}{2^9}\right)=2^{-30}, \tag{3}$$

while the maximum positive dlfloat16 number is obtained for E^* = 31 and M = 2^9:

$$x_{\max}^{\mathrm{dl}}=2^{31}\left(1+\frac{2^9-1}{2^9}\right)\approx 2^{32}. \tag{4}$$

The quantizer analogous to the dlfloat16 format is the symmetric (around zero) 16-bit piecewise-uniform quantizer whose support region in the positive part [0, $x_{\max}^{\mathrm{dl}}$] is divided into 62 segments [2^{E^*}, 2^{E^*+1}], $-30 \le E^* \le 31$, each performing uniform quantization with 2^9 levels (step size is $\Delta_{E^*}^{\mathrm{dl}}$). Hence, objective quantizer measures like distortion and SQNR can be used to assess the performance of the dlfloat16 format. Recall that quantizer distortion includes the granular D_g and overload D_{ov} distortions, since data can be quantized within and outside the support region [10]. For a floating-point quantizer and an arbitrary data source described by its probability density function (PDF) $p(x,\sigma)$, D_g and D_{ov} can be determined as [8]:

$$D_g(\sigma)=2\sum_{E^*=E_{\min}^*}^{E_{\max}^*}\frac{\Delta_{E^*}^2}{12}P_{E^*}=\frac{1}{6}\sum_{E^*=E_{\min}^*}^{E_{\max}^*}\Delta_{E^*}^2\int_{2^{E^*}}^{2^{E^*+1}} p(x,\sigma)\,dx, \tag{5}$$

$$D_{ov}(\sigma)=2\int_{x_{\max}}^{\infty}\left(x-x_{\max}\right)^2 p(x,\sigma)\,dx, \tag{6}$$

where P_{E^*} in Eq. (5) is the segment probability while $E_{\min}^*$ and $E_{\min}^*$ are the minimum and maximum value of E^*, respectively. The SQNR is defined as [9], [10]:

$$\mathrm{SQNR}(\sigma)=10\log_{10}\left(\frac{\sigma^2}{D_g(\sigma)+D_{ov}(\sigma)}\right). \tag{7}$$

B. Performance for the Laplacian PDF

The zero-mean Laplacian PDF of variance σ^2 is given as [9], [10]:

$$p(x,\sigma)=\frac{1}{\sigma\sqrt{2}}\exp\left\{-\frac{\sqrt{2}}{\sigma}|x|\right\}. \tag{8}$$

Substituting Eqs. (2) and (8) into Eq. (5), as well as Eqs. (4) and (8) into Eq. (6), the following expressions for the distortion components are obtained:

$$D_g^{\mathrm{dl}}(\sigma)=\sum_{E^*=-30}^{31}\frac{2^{2\left(E^*-9\right)}}{12}\left(\exp\left(-\frac{2^{E^*+1/2}}{\sigma}\right)-\exp\left(-\frac{2^{E^*+3/2}}{\sigma}\right)\right), \tag{9}$$

$$D_{ov}^{\mathrm{dl}}(\sigma)=\sigma^2\exp\left(-\frac{2^{32+1/2}}{\sigma}\right). \tag{10}$$

Based on Eqs. (7), (9) and (10), we obtain the SQNR expression of dlfloat16 format in the case of Laplacian PDF:

$$\mathrm{SQNR}^{\mathrm{dl}}(\sigma)=-10\cdot\log_{10}\left[\sum_{E^*=-30}^{31}\frac{2^{2\left(E^*-9\right)}}{12\sigma^2}\left(\exp\left(-\frac{2^{E^*+1/2}}{\sigma}\right)-\exp\left(-\frac{2^{E^*+3/2}}{\sigma}\right)\right)+\exp\left(-\frac{2^{32+1/2}}{\sigma}\right)\right]. \tag{11}$$

C. Performance for the Gaussian PDF

The zero-mean Gaussian PDF of variance σ^2 is defined by [9], [10]:

$$p(x,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{12}$$

Putting Eqs. (2) and (12) into Eq. (5), as well as Eqs. (4) and (12) into Eq. (6), we arrive at:

$$D_g^{\mathrm{dl}}(\sigma)=\sum_{E^*=-30}^{31}\frac{2^{2\left(E^*-9\right)}}{12}\left(\operatorname{erf}\left(-\frac{2^{E^*+1/2}}{\sigma}\right)-\operatorname{erf}\left(-\frac{2^{E^*-1/2}}{\sigma}\right)\right), \tag{13}$$

$$D_{ov}^{\mathrm{dl}}(\sigma)=\sigma^2\left(-\sqrt{\frac{1}{\pi}}\frac{2^{32+1/2}}{\sigma}\exp\left\{-\frac{2^{63}}{\sigma^2}\right\}+\left(\frac{2^{64}}{\sigma^2}+1\right)\operatorname{erfc}\left(\frac{2^{32-1/2}}{\sigma}\right)\right) \tag{14}$$

Using Eqs. (7), (13), and (14), we derive the SQNR of the dlfloat16 format in the case of the Gaussian PDF:

$$\mathrm{SQNR}^{\mathrm{dl}}(\sigma)=-10\log_{10}\left(\sum_{E^*=-30}^{31}\frac{2^{2\left(E^*-9\right)}}{12\sigma^2}\left(\operatorname{erf}\left(\frac{2^{E^*+1/2}}{\sigma}\right)-\operatorname{erf}\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right)-\sqrt{\frac{1}{\pi}}\frac{2^{32+1/2}}{\sigma}\exp\left\{-\frac{2^{63}}{\sigma^2}\right\}+\left(\frac{2^{64}}{\sigma^2}+1\right)\operatorname{erfc}\left(\frac{2^{32-1/2}}{\sigma}\right)\right). \tag{15}$$

III. BFLOAT16 BINARY FORMAT

A. Basics and Analogous Quantizer

In the bfloat16 format, 8 bits ($e_1...e_8$) are reserved for the exponent E ($E=\sum_{i=0}^{7} e_{7-i}2^i$), 7 bits ($m_1...m_7$) are reserved for the mantissa M ($E=\sum_{i=0}^{6} m_{7-i}2^i$) and s is a bit for the sign of x, as shown in Fig. 2. The mantissa M varies between 0 to 2^7-1. The exponent E goes from 1 to 254 (0 and 255 are reserved for special purposes [8]).

$$x=\left(\underset{sign}{s}\ \underbrace{e_1...e_8}_{exponent\,E}\ \underbrace{m_1...m_7}_{mantissa\,M}\right)_2$$

Fig. 2. Encoding a real number x using the bfloat16 binary format.

The decimal value of x can be provided with [1]:

$$x=(-1)^s 2^{E^*}\left(1+\frac{M}{2^7}\right), \tag{16}$$

where E^* = E−127 is the biased exponent taking values from −126 to 127. In bfoat16 format, the positive part of the real axis consists of 254 groups (each group is defined by a single E^* value) with 2^7 uniformly distributed numbers, where the step size within the specific group is:

$$\Delta_{E^*}^{\mathrm{bf}}=2^{E^*}\left(1+\frac{M+1}{2^7}\right)-2^{E^*}\left(1+\frac{M}{2^7}\right)=2^{E^*-7}. \tag{17}$$

The bfloat16 format is also zero-symmetric, so the same structure of 254 groups with 2^7 evenly spaced numbers can be found in the negative part. Using Eq. (16), the minimal (for E^* = −126 and M = 0) and the maximal (for E^* = 127 and M = 2^7−1) positive values representable by the bfloat16 format can be determined, as given by Eqs. (18) and (19), respectively:

$$x_{\min}^{\mathrm{bf}}=2^{-126}\left(1+\frac{0}{2^7}\right)=2^{-126}, \tag{18}$$

$$x_{\max}^{\mathrm{bf}}=2^{127}\left(1+\frac{2^7-1}{2^7}\right)\approx 2^{128}. \tag{19}$$

As mentioned in [8], the quantizer analogous to the bfloat16 format is a symmetrical (around zero) 16-bit piecewise quantizer, called the bfloat16 quantizer. Its support region in the positive part $[0, x_{\max}^{\mathrm{bf}}]$ is partitioned into 254 unequal segments $[2^{E^*}, 2^{E^*+1}]$, $-126 \le E^* \le 127$, where $\Delta_{E^*}^{\mathrm{bf}}$ defines the step size in the single segment that accommodates 2^7 quantization levels. The following will provide the SQNR expressions for the bfloat16 format in the case of Laplacian and Gaussian PDFs.

B. Performance for the Laplacian PDF

The following expression can be used to evaluate the SQNR of the bfloat16 format for the Laplacian PDF [8]:

$$\mathrm{SQNR}^{\mathrm{bf}}(\sigma)=-10\cdot\log_{10}\left[\sum_{E^*=-126}^{127}\frac{2^{2(E^*-7)}}{12\sigma^2}\left(\exp\left(-\frac{2^{E^*+1/2}}{\sigma}\right)-\exp\left(-\frac{2^{E^*+3/2}}{\sigma}\right)\right)+\exp\left(-\frac{2^{128+1/2}}{\sigma}\right)\right]. \tag{20}$$

C. Performance for the Gaussian PDF

By applying Eqs. (12) and (17) in Eq. (5), the expression for the granular distortion is yielded:

$$D_g^{\mathrm{bf}}(\sigma)=\sum_{E^*=-126}^{127}\frac{2^{2(E^*-7)}}{12}\left(\mathrm{erf}\left(-\frac{2^{E^*+1/2}}{\sigma}\right)-\mathrm{erf}\left(-\frac{2^{E^*-1/2}}{\sigma}\right)\right), \tag{21}$$

while the expression for overload distortion is obtained according to Eqs. (17), (19) and (6):

$$D_{ov}^{\mathrm{bf}}(\sigma)=\sigma^2\left(-\sqrt{\frac{1}{\pi}}\frac{2^{128+1/2}}{\sigma}\exp\left\{-\frac{2^{255}}{\sigma^2}\right\}+\left(\frac{2^{256}}{\sigma^2}+1\right)\mathrm{erfc}\left(\frac{2^{128-1/2}}{\sigma}\right)\right) \tag{22}$$

Finally, using Eqs. (7), (21) and (22), the SQNR of the bfloat16 format in the case of Gaussian PDF is delivered:

$$\mathrm{SQNR}^{\mathrm{bf}}(\sigma)=-10\log_{10}\left(\sum_{E^*=-126}^{127}\frac{2^{2(E^*-7)}}{12\sigma^2}\left(\mathrm{erf}\left(\frac{2^{E^*+1/2}}{\sigma}\right)-\mathrm{erf}\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right)-\sqrt{\frac{1}{\pi}}\frac{2^{128+1/2}}{\sigma}\exp\left\{-\frac{2^{255}}{\sigma^2}\right\}+\left(\frac{2^{256}}{\sigma^2}+1\right)\mathrm{erfc}\left(\frac{2^{128-1/2}}{\sigma}\right)\right). \tag{23}$$

IV. RESULTS AND DISCUSSION

In this Section, the SQNR results for the dlfloat16 and bfloat16 formats are compared, where the variance range assumed is [−100 dB, 300 dB] with respect to the reference variance $\sigma_{\mathrm{ref}}^2 = 1$.

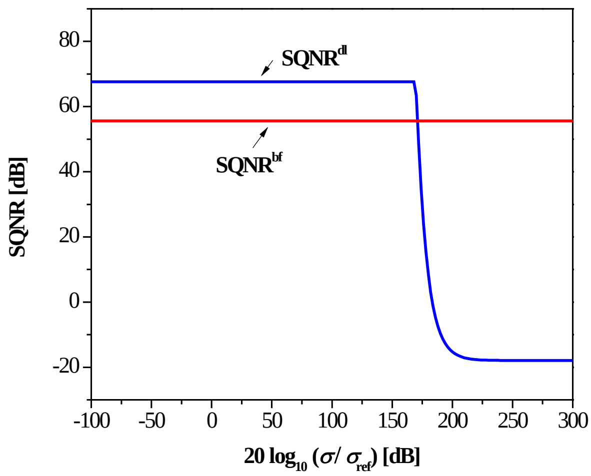

Fig. 3. SQNR comparison between dlfloat16 and bfloat16 (Laplacian PDF).

In Fig. 3 we present the SQNR results for the Laplacian PDF. The bfloat16 (see $\mathrm{SQNR}^{\mathrm{bf}}$) exhibits a wide dynamic range, as it ensures a constant SQNR across all considered variance values. Dlfloat16 (see $\mathrm{SQNR}^{\mathrm{dl}}$), on the other hand, offers a lower dynamic range than bfloat16, but can achieve a higher SQNR by 12.04 dB. Namely, dlfloat16 keeps the SQNR steady until the variance threshold σ_t^2 = 168 dB, when it starts to drop rapidly. This is because in the interval $20\cdot\log_{10}(\sigma / \sigma_{\mathrm{ref}}) \le \sigma_t^2$ = 168 dB, SQNR is only affected by granular distortion (Eq. (9)), while for $20\cdot\log_{10}(\sigma/\sigma_{\mathrm{ref}}) > \sigma_t^2$ it depends only on overload distortion (Eq. 10).

TABLE I
SQNR FOR DLFLOAT16 AND BFLOAT16 FORMATS AT SPECIFIC VARIANCE POINTS (LAPLACIAN PDF)

Variance [dB]	$SQNR^{dl}$	$SQNR^{bf}$
−100	67.64	55.60
−50	67.64	55.60
0	67.64	55.60
50	67.64	55.60
100	67.64	55.60
150	67.64	55.60
200	−15.27	55.60
250	−17.92	55.60
300	−17.92	55.60

Table I lists the specific SQNR values for the observed formats at selected points from the defined variance range. Figure 3 and Table I also show that SQNR in dlfloat16 can take negative values which reflect extremely poor performance and are therefore unacceptable.

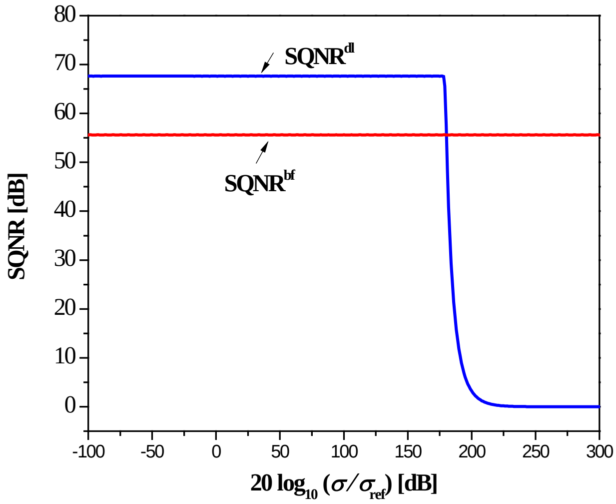

Fig. 4. SQNR comparison between dlfloat16 and bfloat16 (Gaussian PDF).

TABLE II
SQNR FOR DLFLOAT16 AND BFLOAT16 FORMATS AT SPECIFIC VARIANCE POINTS (GAUSSIAN PDF)

Variance [dB]	$SQNR^{dl}$	$SQNR^{bf}$
−100	67.65	55.61
−50	67.66	55.62
0	67.63	55.59
50	67.64	55.60
100	67.66	55.62
150	67.64	55.60
200	3.20	55.59
250	0.009	55.62
300	2.98×10^{-5}	55.61

Fig. 4 depicts the achieved SQNR results for the discussed formats for Gaussian PDF. Table II summarizes the SQNR values at selected variance points. It can be seen that the dlfloat16 and bfloat16 formats behave similarly as for the Laplacian PDF, as the resulting SQNR values differ slightly (see Fig.3 and Table I). Namely, dlfloat16 for the Gaussian PDF exhibits a slightly wider dynamic range (σ_t^2 = 178 dB), and also avoids negative SQNR.

The above SQNR findings show that a change in data distribution causes a marginal change in the performance of given formats. It is also obvious that dlfloat16 is a better option for more precise calculations, while bfloat16 is preferable from a dynamic range point of view.

V. CONCLUSION

Two well-known 16-bit floating-point formats, dlfloat16 and bfloat16, have been examined and compared in this paper, using the SQNR. In particular, the appropriate SQNR expressions have been developed for both Laplacian and Gaussian distributions. It has been shown in a wide variance range that the SQNR for Laplacian and Gaussian data differs slightly, so the PDF of the data has a marginal impact on the performance of both formats. The SQNR analysis has also revealed that dlfloat16 offers for 12.04 dB higher SQNR than bfloat16, but it operates in a lower dynamic range. Hence, bfloat16 is a great option for applications where exceptional dynamic range is required, while dlfloat16 is better when heightened calculation precision is needed.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [grant number 451-03-136/2025-03/200102], as well as by the European Union’s Horizon Europe research and innovation programme through the AIDA4Edge Twinning project (grant ID 101160293).

REFERENCES

[1] IEEE Standard 754 for Floating Point Numbers, 2008.

[2] M. Fasi, M. Mikaitis, “Algorithms for Stochastically Rounded Elementary Arithmetic Operations in IEEE 754 Floating-Point Arithmetic”, *IEEE Trans. on Emerging Topics in Computing*, vol. 9, issue: 3, pp. 1451-1466, 2021.

[3] A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, *Dive into Deep Learning*, Amazon Science, 2020.

[4] F. He, K. Ding, D. Yan, J. Li, J. Wang, M. Chen, “A novel Quantization and Model Compression Approach for Hardware Accelerators in Edge Computing”, *Computers, Materials and Continua*, vol. 80, issue 2, pp. 3021-3045, 2024.

[5] F. M. Aymone, D. P. Pau, “Benchmarking In-Sensor Machine Learning Computing: An Extension to the MLCommons-Tiny Suite”. *Information*, vol. 15, no. 11, Article ID: 674, 2024.

[6] A. Agrawal, S. M. Mueller, B. M. Fleischer, X. Sun, N. Wang, J. Choi, K. Gopalakrishnan, DLFloat: A 16-b Floating Point Format Designed for Deep Learning Training and Inference”, in *Proc. ARITH'2019*, Kyoto, Japan, 2019, pp. 92–95.

[7] N. Burgess, J. Milanovic, N. Stephens, K. Monachopoulos and D. Mansell, “Bfloat16 Processing for Neural Networks”, in *Proc. ARITH'2019*, Kyoto, Japan, 2019.

[8] B. Denić, Z. Perić, M. Dinčić, “Improvement of the Bfloat16 Floating-Point for the Laplacian Source”, in *Proc. ATEE'2023*, Bucharest, Romania, 2023, pp. 1–4.

[9] A. Gersho, R. Gray, *Vector Quantization and Signal Compression*, New York, Kluwer Academic Publishers, 1992.

[10] N. C. Jayant, P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, New Jersey, Prentice Hall, 1984.

[11] R. Banner, Y. Nahshan, and D. Soudry, “Post Training 4-bit Quantization of Convolutional Networks for Rapid Deployment”, in *Proc. NeurIPS'2019*, Vancouver, Canada, 2019.