



Article

# Design of a 2-Bit Neural Network Quantizer for Laplacian Source

Zoran Perić<sup>1</sup>, Milan Savić<sup>2</sup> , Nikola Simić<sup>3,\*</sup> , Bojan Denić<sup>1</sup> and Vladimir Despotović<sup>4</sup>

<sup>1</sup> Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia; zoran.peric@elfak.ni.ac.rs (Z.P.); bojan.denic@elfak.ni.ac.rs (B.D.)

<sup>2</sup> Faculty of Sciences and Mathematics, University of Pristina in Kosovska Mitrovica, Ive Lole Ribara 29, 38220 Kosovska Mitrovica, Serbia; milan.savic1@pr.ac.rs

<sup>3</sup> Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia

<sup>4</sup> Department of Computer Science, Faculty of Science, Technology and Medicine, University of Luxembourg, Avenue de la Fonte 6, L-4364 Esch-sur-Alzette, Luxembourg; vladimir.despotovic@uni.lu

\* Correspondence: nikolasimic@uns.ac.rs

**Abstract:** Achieving real-time inference is one of the major issues in contemporary neural network applications, as complex algorithms are frequently being deployed to mobile devices that have constrained storage and computing power. Moving from a full-precision neural network model to a lower representation by applying quantization techniques is a popular approach to facilitate this issue. Here, we analyze in detail and design a 2-bit uniform quantization model for Laplacian source due to its significance in terms of implementation simplicity, which further leads to a shorter processing time and faster inference. The results show that it is possible to achieve high classification accuracy (more than 96% in the case of MLP and more than 98% in the case of CNN) by implementing the proposed model, which is competitive to the performance of the other quantization solutions with almost optimal precision.



check for updates

**Citation:** Perić, Z.; Savić, M.; Simić, N.; Denić, B.; Despotović, V. Design of a 2-Bit Neural Network Quantizer for Laplacian Source. *Entropy* **2021**, *23*, 933. <https://doi.org/10.3390/e23080933>

**Keywords:** image classification; Laplacian source; neural network; quantization

Academic Editor: Friedhelm Schwenker

Received: 19 June 2021

Accepted: 20 July 2021

Published: 22 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).