Contribution ID: **49** Type: **Standard 15 min talk**

Differentiable Weightless Neural Networks

Thursday 17 October 2024 13:55 (15 minutes)

We introduce the Differentiable Weightless Neural Network (DWN), a model based on interconnected lookup tables. Training of DWNs is enabled by a novel Extended Finite Difference technique for approximate differentiation of binary values. We propose Learnable Mapping, Learnable Reduction, and Spectral Regularization to further improve the accuracy and efficiency of these models. We evaluate DWNs in three edge computing contexts: (1) an FPGA-based hardware accelerator, where they demonstrate superior latency, throughput, energy efficiency, and model area compared to state-of-the-art solutions, (2) a low-power microcontroller, where they achieve preferable accuracy to XGBoost while subject to stringent memory constraints, and (3) ultra-low-cost chips, where they consistently outperform small models in both accuracy and projected hardware area. DWNs also compare favorably against leading approaches for tabular datasets, with higher average rank. Overall, our work positions DWNs as a pioneering solution for edge-compatible high-throughput neural networks.

Focus areas

Primary authors: T. L. BACELLAR, Alan (University of Texas at Austin); SUSSKIND, Zachary (The University of Texas at Austin)

Co-authors: Dr BRETERNITZ JR., Mauricio (ISCTE - Instituto Universitario de Lisboa); Dr JOHN, Eugene (University of Texas at San Antonio); Dr K. JOHN, Lizy (University of Texas at Austin); Dr M. V. LIMA, Priscila (Universidade Federal do Rio de Janeiro); Dr M. G. FRANÇA, Felipe (Instituto de Telecomunicações)

Presenter: T. L. BACELLAR, Alan (University of Texas at Austin)

Session Classification: Contributed talks