LUT based Neural Networks as Neuro-Symbolic Systems

Lizy K. John[1], Priscila M. V. Lima[2], Alan T. L. Bacellar[1], Shashank Nag[1], Eugene John[3], and Felipe M. G. França[4]

1- UT Austin, Austin, USA; 2- UFRJ, Rio de Janeiro, Brazil; 3- UTSA, San Antonio, USA; 4- IT-Porto, Porto, Portugal

Introduction

- Integration of neural and symbolic learning paradigms in the same computational substract, is of growing interest.
- However, this could be computationally expensive, in terms of memory and time-costs, with deployability challenges in online learning.
- *Weightless neural networks (WNNs)* use LUTs for computation, capturing complex behaviors with shallow models.
- We propose using LUT-based neural networks as Neuro-Symbolic learning systems, bringing these to the level of integrated circuits.

WiSARD

- Early WNN for classification [1]. Use many small LUTs, with a subset of model inputs as inputs to the LUTs.
  - One set of LUTs for each output class "discriminator".
- Learn n-tuple "subpatterns" in training data. For inference we want more LUTs to output 1 in the correct class's discriminator than in any other (Figure 2).

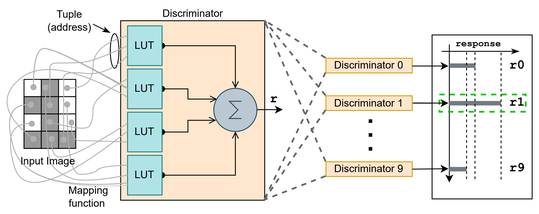

Figure 1: A WiSARD model.

BTHOWeN & ULEEN

- Improved WNN, with a FPGA-based inference accelerator.
- BTHOWeN [8] incorporates arithmetic-free hashing, counting bloom filters, and bleaching.
  - Consumes 85-99% fewer cycles and 80-95% less energy compared to iso-accuracy DNNs.
- ULEEN [7] incorporates ensembles and pruning of LUTs.
  - Excels over iso-accuracy Binary Neural Networks [4]

Differentiable Weightless Neural Networks (DWNs)

- Multi-layer WNNs, with directly-chained layers of LUTs.
  - An *Extended Finite Difference (EFD)* based learning rule.
  - A *Learnable Mapping* interconnect layer.
  - An arithmetic-free *Learnable Reduction* technique for tiny circuits.
  - A *Spectrial Normalization* based regularization technique.
- DWNs achieve the lowest average rank and L1 norm, with comparable parameter sizes against other models (Table 1).

	DWN (Ours)	DiffLogicNet	AutoGluon XGBoost
Avg Rank	**2.5**	4.5	3.4
Avg L1	**0.005**	0.016	0.009

Table 1: Rank and L1 accuracy loss of DWNs vs. leading approaches to tabular machine learning datasets.

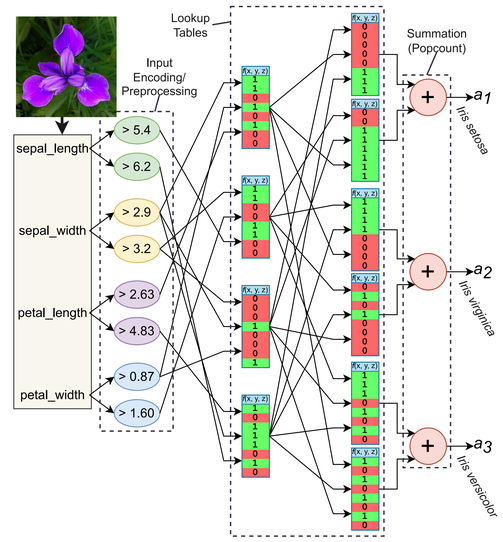

Figure 2: A tiny DWN for the Anderson/Fisher Iris dataset.

DWNs on FPGAs

- DWN model LUTs can be directly converted to hardware LUTs, which are abundant on FPGAs (Figure 3); for best efficiency, LUT sizes should be matched between the model and the FPGA.

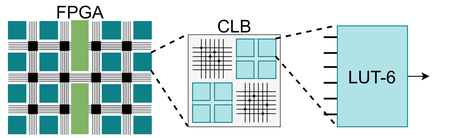

Figure 3: FPGAs provide abundant hardware LUTs within CLB blocks.

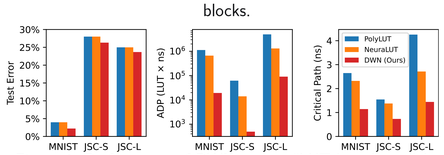

Figure 4: Performance of DWNs versus "LUT-based" models.

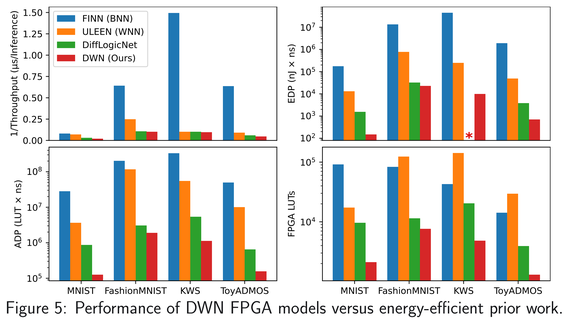

Figure 5: Performance of DWN FPGA models versus energy-efficient prior work.

- Figure 4 compares DWNs against recent "LUT-based" models PolyLUT [3] and NeuraLUT [2].
- Figure 5 compares DWNs against iso-accuracy models for BNNs (FINN [9]), SOTA WNNs (ULEEN [7]), and DiffLogicNet [6], which they outperform in size, speed, and efficiency.

DWNs for Tiny Circuits

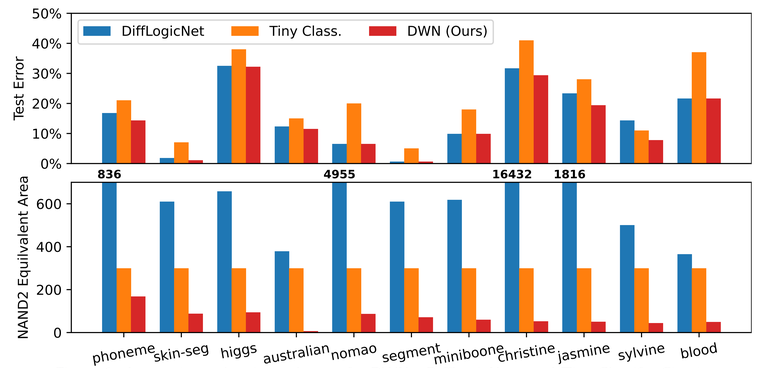

Figure 6: Accuracies and estimated areas for DWNs, DiffLogicNets, and Tiny Classifer Circuits.

- Figure 6 compares DWNs implemented as tiny ICs against estimated NAND2 equivalent areas for Tiny Classifier Circuits [5].

Neurosymbolic Inference Engines

- Differentiable Weightless Neural Networks can be considered as a symbol extractor, or ultra-fast ultra-thin neurosymbolic inference engine.
- Learnable input mapping can be considered as rule-based learning.
- LUT contents can be considered as the neural component.
- Ongoing efforts aim to integrate explicit knowledge or rules into DWNs, on top of those implicitly acquired during training.

References

[1] I Aleksander, WV Thomas, and PA Bowden. *WISARD a radical step forward in image recognition*. 1984.
[2] Marta Andronic and George A. Constantinides. *NeuraLUT: Hiding Neural Network Density in Boolean Synthesizable Functions*. 2024.
[3] Marta Andronic and George A. Constantinides. *PolyLUT: learning piecewise polynomials for ultra-low latency FPGA LUT-based inference*. 2023.
[4] Itay Hubara et al. *Binarized Neural Networks*. 2016.
[5] Konstantinos Iordanou et al. *Low-cost and efficient prediction hardware for tabular data using tiny classifier circuits*. 2024.
[6] Felix Petersen et al. *Deep Differentiable Logic Gate Networks*. 2022.
[7] Zachary Susskind et al. *ULEEN: A Novel Architecture for Ultra-low-energy Edge Neural Networks*. 2023.
[8] Zachary Susskind et al. *Weightless Neural Networks for Efficient Edge Inference*. 2022. doi: https://doi.org/10.1145/3559009.3569680.
[9] Y Umuroglu et al. *FINN: A Framework for Fast, Scalable Binarized Neural Network Inference*. 2017.

Acknowledgement: *This research was supported in part by Semiconductor Research Corporation (SRC) Task 3148.001, National Science Foundation (NSF) Grant #2326894 & #2425655, NVIDIA Applied Research Accelerator Program Grant. Any opinions, findings, conclusions, or recommendations are those of the authors and not of the funding agencies.*