GaZmusino: An extended edge RISC-V core with support for Bayesian Neural Networks

Samuel Pérez Pedrajas*, Javier Resano* and Darío Suárez Gracia*

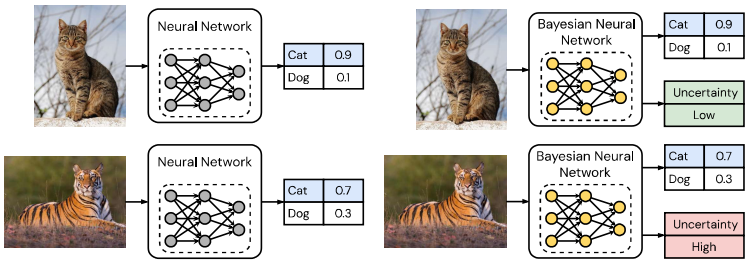
Contact email
samuel.perez@unizar.es

**Department of Computer Science and Systems Engineering (DIIS), Aragon Institute for Engineering Research (I3A), University of Zaragoza*

Grupo de Investigación en Arquitectura de Computadores (gaZ)
Universidad Zaragoza

¿What are Bayesian Neural Networks?

- Integrate probabilistic modeling
- Extend predictions with uncertainty
- More expensive inference algorithm



Weight Sampling Optimization

- Bayesian Neural Networks parameters are modeled by **Gaussian distributions**

- Distribution sampling takes **more than 80% execution time** during inference

- We propose and validate using the **Uniform distribution** instead of Gaussian doing a weight transformation

$$\sigma\,\mathcal{N}(0,1) + \mu \rightarrow a\,\mathcal{U}(0,1) + b$$

$$a = \sigma\sqrt{12}$$

$$b = \mu - a/2$$

From BayesianTorch to GaZmusino

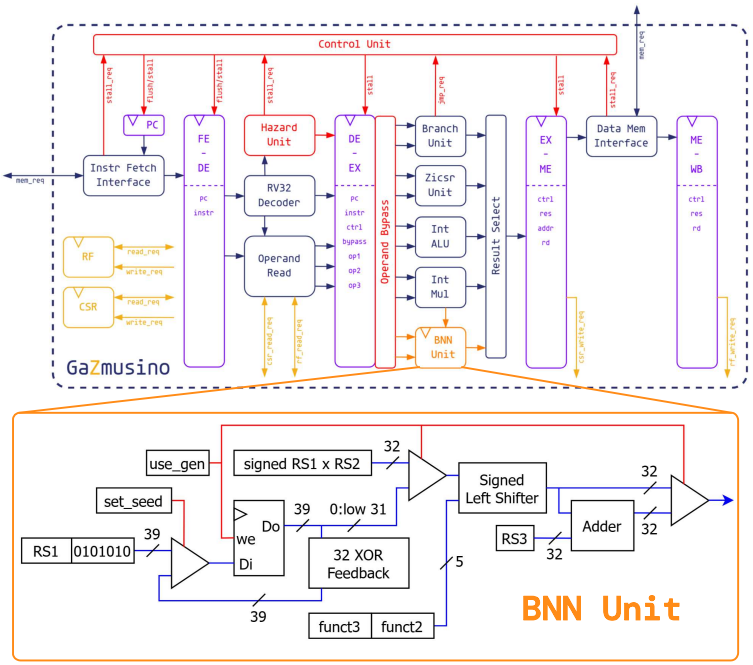
BayesianTorch

Proposed software toolchain

- Model optimization
  - Layer folding
  - Weight transformation
  - Fixed point
- Portable C code generation
- GaZmusino BNN extension

RISC-V®

GaZmusino Open-Source RISC-V Core



New instructions

- `fxgen.unif rd, I`
- `fxgen.seed ra`
- `fx.madd rd, ra, rb, rc, I`

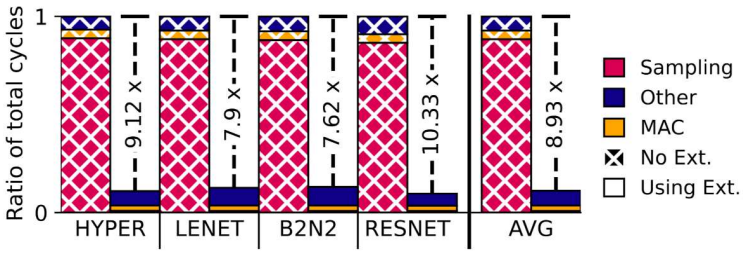
- Uniform RNG
- Fixed-Point MAC

Results and Conclusions

Model	↑ Acc % BT	↑ Acc % GZ	↓ RE % BT	↓ RE % GZ	↓ UCE % BT	↓ UCE % GZ
HYPER	89.46	0.03	3.93	0.00	3.31	-0.11
LENET	62.61	-0.38	2.62	-0.75	4.09	1.35
B2N2	75.77	0.17	2.13	-0.54	2.72	1.86
RESNET	81.01	-1.34	2.23	-0.74	2.24	0.71
Average		-0.38		-0.51		0.95
Std. Dev.		0.29		0.39		1.02

ACC. Accuracy (Higher Better)
RE. Reliability Error (Lower Better)
UCE. Uncertainty Calibration Error (Lower Better)

Model performance preserved



Avg. 8.9× speedup and 8.2× energy efficiency
GaZmusino enables BNN inference on the edge

References

[1] Chuan Guo et al. *"On Calibration of Modern Neural Networks"*. 2017.
[2] Charles Blundell et al. *"Weight Uncertainty in Neural Networks"*. 2015.
[3] Hiromitsu Awano and Masanori Hashimoto. *"B2N2: Resource efficient Bayesian neural network accelerator using Bernoulli sampler on FPGA"*. 2023.
[4] Ranganath Krishnan, Pi Esposito, and Mahesh Subedar. *"Bayesian-Torch: Bayesian neural network layers for uncertainty estimation"*. 2022.
[5] Max-Heinrich Laves et al. *"Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference"*. 2019.
[6] Colby Banbury et al. *"MLPerf Tiny Benchmark"*. 2021.

Funded by
PDC2023-145851-I00 AEI/10.13039/501100011033 and PID2022-136454NB-C22 AEI/10.13039/50110001103

RETORNNA