Robust and Efficient Machine Learning for Emerging Resource-Constrained Embedded Systems

Mikail Yayla[1], Hussam Amrouch[2], Jian-Jia Chen[1]

[1]Design Automation for Embedded Systems Group, Technical University of Dortmund, Germany
[2]Chair of AI Processor Design, Technical University of Munich (TUM), Germany

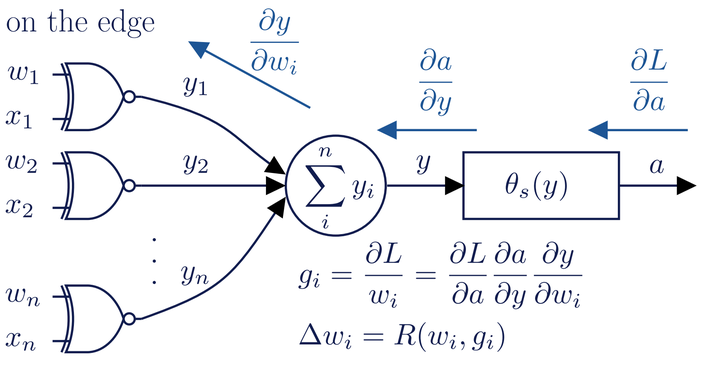
Vision of this Thesis

Robust BNNs with approximate memory and computing units, while training on the edge

- BNN robustness optimization
- HW/SW codesign methods for robust and efficient BNN inference
- Train BNNs efficiently on the edge

w_i	x_i	y_i
0	0	1
0	1	0
1	0	0
1	1	1

$w_i, x_i, y_i, a \in \{0,1\}$
$s, y \in \mathbb{Z}$
$\theta_s(y) = \begin{cases} 1 & y > s \\ 0 & else \end{cases}$



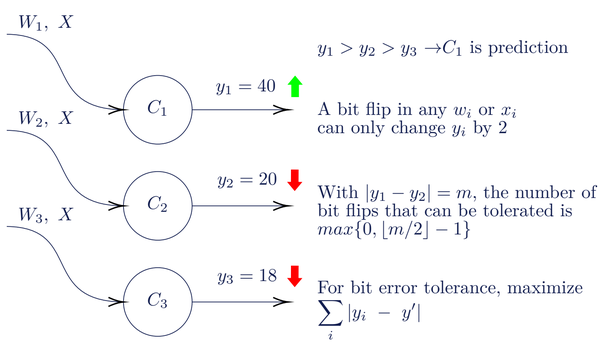
$g_i = \frac{\partial L}{w_i} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial y}\frac{\partial y}{\partial w_i}$

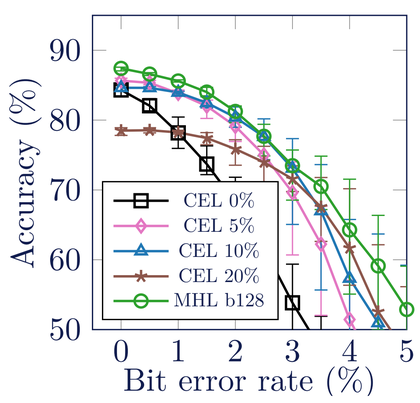
$\Delta w_i = R(w_i, g_i)$

Robustness Optimization of BNNs

Goal: Achieve robustness without bit flip injection

- Classes: C_i. Index i of neuron with largest output determines predicted class index
- MHL maximizes the margins between the outputs of the last-layer neuron
- Margin-maximization leads to robustness without the need of error models

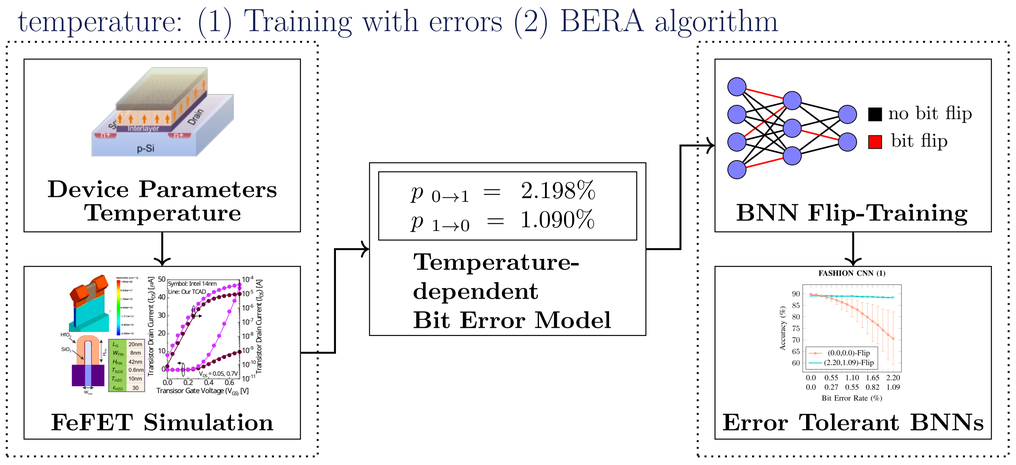




BNNs with FeFET Memory

Explore FeFET memory as on-chip memory for BNNs

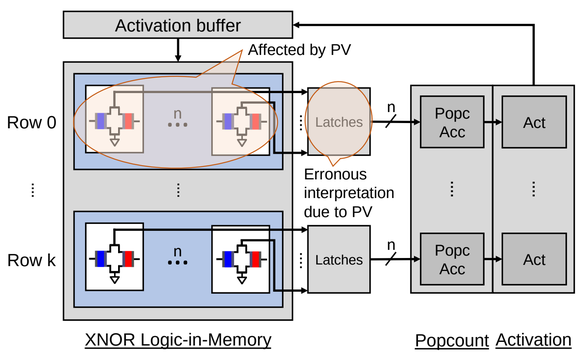
- FeFET sensitive to temperature fluctuations, unacceptable BNN accuracy drop
- Countermeaures achieve temperature tolerance across entire range of operating temperature: (1) Training with errors (2) BERA algorithm

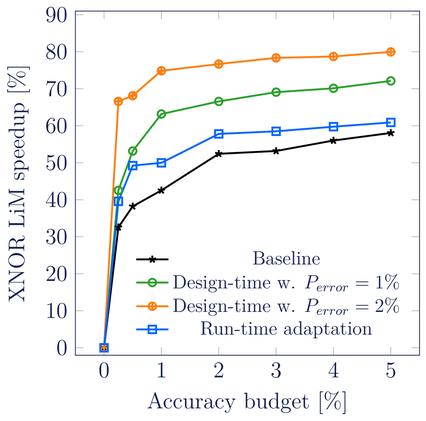


BNNs with FeFET-based XNOR LiM

Trade off speed of FeFET-based XNOR gates with BNN reliability

- BNNs employ approximate FeFET-based XNOR gates for LiM
- Investigate the probability of error in FeFET-based XNOR LiM
- Exploit robustness of BNNs, trading off speed and reliability using the design objective $WSAD_{\ell,p_\ell} = \frac{s_{\ell,p_\ell}}{AD_{\ell,p_\ell}} \frac{c_\ell}{c_{max}}$

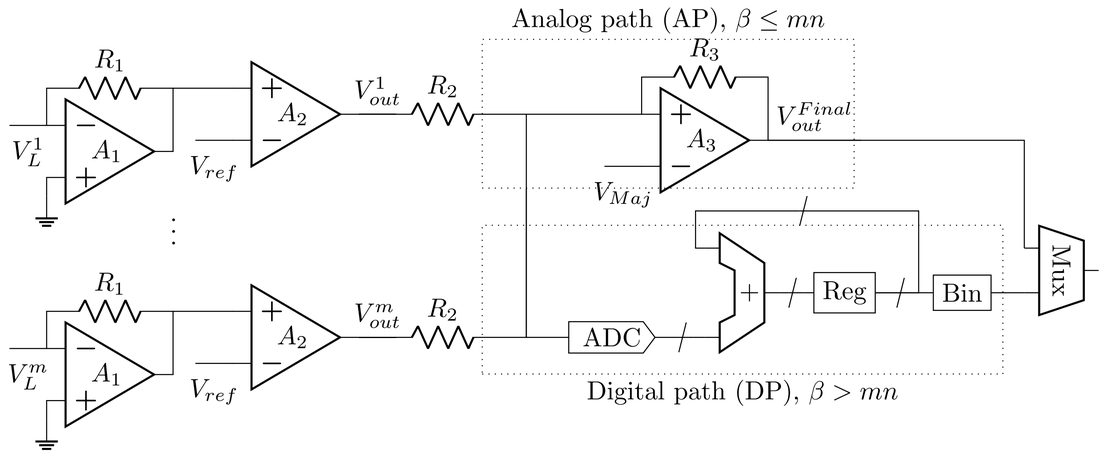




Efficient Analog-based BNN Accelerators

Analog BNN accelerators use lots of ADCs → Expensive designs!

- LTA approximates global thesholdings by local thresholdings
- Using LTA, ADCs and digital components are not required or only used rarely
- Area and energy usage significantly smaller in LTA circuit compared to SOTA

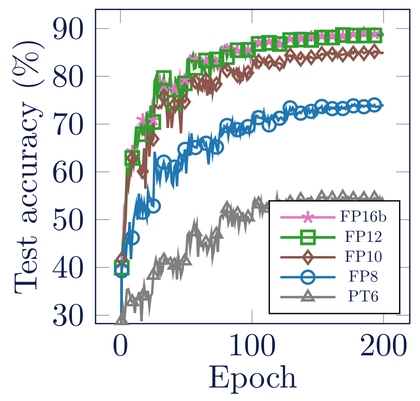


Efficient BNN Training on the Edge

Efficient BNN training through reduced FP formats

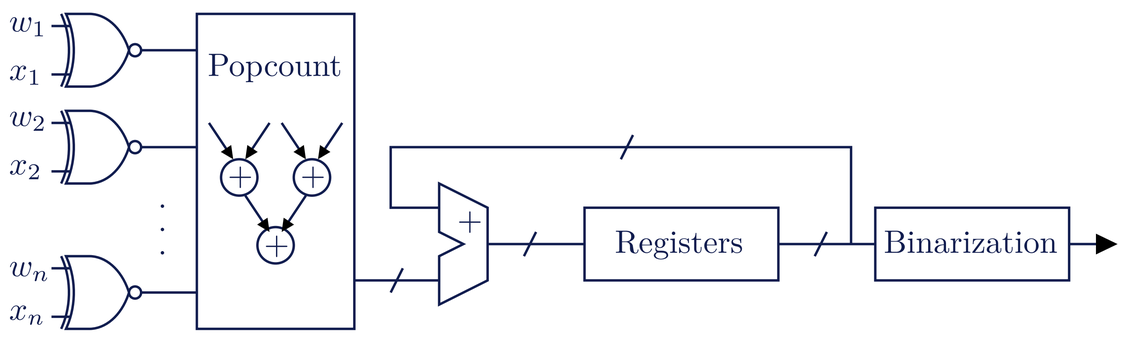
- Bias: $b = -\lfloor \log_2(m_{sample}^{absmax}) \rfloor + 1$
- Exponent: $c = \lceil \log_2(-\log_2(\alpha\tau) - b + 1) \rceil$
- Mantissa: $U(Q, \mathbf{M}_t) = \frac{1}{|\mathbf{M}_t|} \sum_{m_t \in \mathbf{M}_t} \mathbf{1}\left[|\Delta m_t| < \frac{|Q(m_t) - q_{v^*}|}{2}\right]$

Encoding	Range	Precision	Realization
PT6	o	-	1 sign, 5 exp.
FP8	o	o-	1 sign, 5 exp., 2 mant.
FP10	o	o-	1 sign, 5 exp., 4 mant.
FP12	o	o	1 sign, 5 exp., 6 mant.
FP16b	+	o	1 sign, 8 exp., 7 mant.
FP32	+	+	1 sign, 8 exp., 23 mant.



Research tools available: github.com/myay

- SPICE-TORCH: Connect SPICE and PyTorch simulations
- TREAM: Error evaluations of tree-based models in sklearn
- DAEBI: Enables design space exploration of BNN accelerators in VHDL regarding different accelerator architectures and dataflows



Acknowledgements

This research has been supported by Deutsche Forschungsgemeinschaft (DFG) project OneMemory (405422836), by the SFB876 "Providing Information by Resource-Constrained Analysis" (project number 124020371), and by the Federal Ministry of Education and Research of Germany and the state of NRW as part of the Lamarr-Institute for ML and AI, LAMARR22B.

References

[1] S. Buschjäger, J.-J. Chen, K.-H. Chen, M. Günzel, C. Hakert, K. Morik, R. Novkin, L. Pfahler, and M. Yayla. Margin-maximization in binarized neural networks for optimizing bit error tolerance. In *DATE*, 2021.

[2] M. Yayla and J.-J. Chen. Memory-efficient training of binarized neural networks on the edge. DAC, 2022.

[3] M. Yayla, F. Frustaci, F. Spagnolo, J.-J. Chen, and H. Amrouch. Global by local thresholding in binarized neural networks for efficient crossbar accelerator design. *(under single blind peer review).*

[4] M. Yayla, S. Buschjäger, A. Gupta, J.-J. Chen, J. Henkel, K. Morik, K.-H. Chen, and H. Amrouch. FeFET-based binarized neural networks under temperature-dependent bit errors. *IEEE Transactions on Computers*, 2022.

[5] M. Yayla, S. Thomann, S. Buschjäger, K. Morik, J.-J. Chen, and H. Amrouch. Reliable binarized neural networks on unreliable beyond von-neumann architecture. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2022.

[6] M. Yayla, Z. Valipour Dehnoo, M. Masoudinejad, and J.-J. Chen. Tream: A tool for evaluating error resilience of tree-based models using approximate memory. In *SAMOS*, 2022.