Bit Error Tolerance of a CIFAR-10 Binarized Convolutional Neural Network Processor

Lita Yang*, Daniel Bankman*, Bert Moons†, Marian Verhelst† and Boris Murmann*
Email: {yanglita, dbankman, murmann}@stanford.edu, {bert.moons, marian.verhelst}@kuleuven.be
*Department of Electrical Engineering, Stanford University, Stanford, CA USA
†ESAT-MICAS, KU Leuven, Leuven, Belgium

***Abstract*—Deployment of convolutional neural networks (ConvNets) in always-on Internet of Everything (IoE) edge devices is severely constrained by the high memory energy consumption of hardware ConvNet implementations. Leveraging the error resilience of ConvNets by accepting bit errors at reduced voltages presents a viable option for energy savings, but few implementations utilize this due to the limited quantitative understanding of how bit errors affect performance. This paper demonstrates the efficacy of SRAM voltage scaling in a 9-layer CIFAR-10 binarized ConvNet processor, achieving memory energy savings of 3.12x with minimal accuracy degradation (~99% of nominal). Additionally, we quantify the effect of bit error accumulation in a multi-layer network and show that further energy savings are possible by splitting weight and activation voltages. Finally, we compare the measured error rates for the CIFAR-10 binarized ConvNet against MNIST networks to demonstrate the difference in bit error requirements across varying complexity in network topologies and classification tasks.**

Keywords—Convolutional neural networks, error resiliency, approximate SRAM, BinaryNet, energy-accuracy trade-off

I. Introduction

Convolutional neural networks (ConvNets) achieve near human performance for a wide range of classification tasks. To reduce latency and the high energy cost of communication with the cloud, recent work focuses on moving ConvNet operations closer to the sensor, demanding higher data processing capabilities at edge devices. Unfortunately, deployment of ConvNets in resource-constrained Internet of Everything (IoE) systems remains a challenge due to the high memory energy consumption caused by storage requirements and substantial data movement. To alleviate this problem, memory voltage scaling at the cost of errors in approximate SRAMs has been shown to be an attractive option for reducing memory energy consumption by leveraging the inherent error resilience of ConvNets. Recent work on the efficacy of this method for the MNIST dataset has been validated in silicon [1].

While extensive study on the effect of quantization noise in ConvNets for various network sizes and datasets has been performed (e.g. [2]), there is still limited literature on the bit error tolerance of deep ConvNets, especially for more complex tasks such as CIFAR-10. Additionally, SRAM operation at low voltages is difficult to model and it is uncertain how a ConvNet will perform in the presence of bit errors without silicon validation. This paper presents the first silicon-validated study on the effect of bit errors on a multi-layer, binarized neural network (BinaryNet) [3] performing image classification of moderate complexity (CIFAR-10), by applying voltage scaling in the SRAMs of a 28 nm mixed-signal BinaryNet (MSBNN) chip [4]. In addition to understanding how bit errors propagate in this network, we seek to find a unified framework for quantifying bit error tolerance across different benchmarks and networks, and compare the tolerable bit error rates (P_{error}) against MNIST networks [1], [5]. Though the focus of this work is on SRAM voltage scaling, the measured bit error tolerances can be similarly exploited for the design of custom memory (e.g. hybrid 8T/6T, larger bitcells) and emerging memory technologies (e.g. RRAM, PCM).

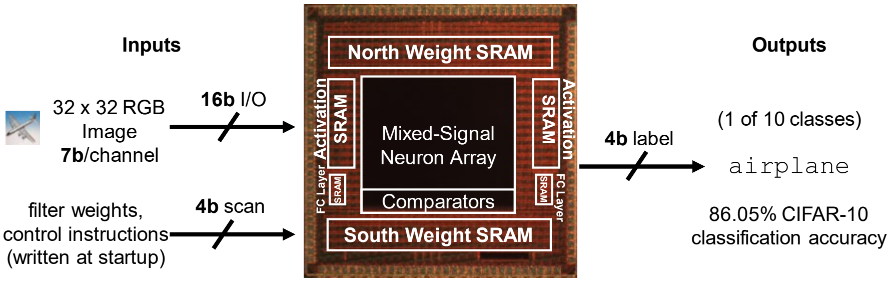
We expand on [1] by demonstrating the bit error tolerance of a CIFAR-10 BinaryNet, validating voltage scaling capabilities with the MSBNN chip. Additionally, we measure and estimate the bit error propagation over the layers of the network, observing that significant errors can accumulate with little to no degradation in classification accuracy. Furthermore, we show that additional energy savings are possible by leveraging the different bit error tolerances between weights and activations. Finally, we compare our P_{error} values with those obtained for MNIST networks, demonstrating that the CIFAR-10 BinaryNet is less error resilient, but still tolerates P_{error} values significantly higher than conventional memory applications. Our findings show an overall memory energy savings of 3.12x with minimal accuracy degradation.

II. BinaryNet Topology and Chip Architecture

Fig. 1 shows the BinaryNet topology and chip used for our experiments [4]. The MSBNN chip reads a 32x32 RGB image, performs operation on all layers, and outputs a 4-bit label from the fully-connected (FC) layer. Binarization allows for storage of all weights and activations on chip, totaling 328 kB of on-chip SRAM. The mixed-signal neuron array contributes analog noise, but for this study, we focus on the regime in which bit errors dominate accuracy degradation and show in subsequent sections that this does not affect the conclusions made about bit error tolerance. The chip has capabilities for per-layer bypass and readout, allowing measurement of the error propagation at each layer. While reconfigurability is possible with the BinaryNet topology, as shown in [6], we focus solely on the fixed architecture used for [4] in this work. We also implement separate voltage domains for the weights, activations, and FC layer SRAMs to tune these separately, and memory bypass circuits to measure the P_{error} at low voltages. As shown in [7], the FC layer is less error tolerant and thus, we set its voltage to the lowest voltage (0.6 V at 36 FPS) that incurs no bit errors.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

CIFAR-10 BinaryNet Topology			
Layer	Type	Image Size	Stride
1	CONV	32 x 32	1
2	CONV	31 x 31	1
3	CONV	30 x 30	1
4	CONV	29 x 29	1
4p	MAX POOL	28 x 28	2
5	CONV	14 x 14	1
6	CONV	13 x 13	1
6p	MAX POOL	12 x 12	2
7	CONV	6 x 6	1
8	CONV	5 x 5	1
9	FC	4096 – to – 10	



All filters are strictly 2 x 2 x 256

Memory Voltage (Weights, Activations)	V_{MEM} = (V_{WMEM} , V_{XMEM})
Digital Logic FC Layer	V_{DD} = 0.6 V
Neuron Array	0.6 – 0.8 V

Technology	28 nm CMOS
Total SRAM Capacity	328 kB
Nominal Voltage	1.0 V
Clock Speed	1.5 MHz
Frame Rate	36 FPS
Total Chip Area	2.44 mm x 2.44 mm

Fig. 1. CIFAR-10 BinaryNet 9-layer topology and chip specifications for the custom designed MSBNN processor [4].

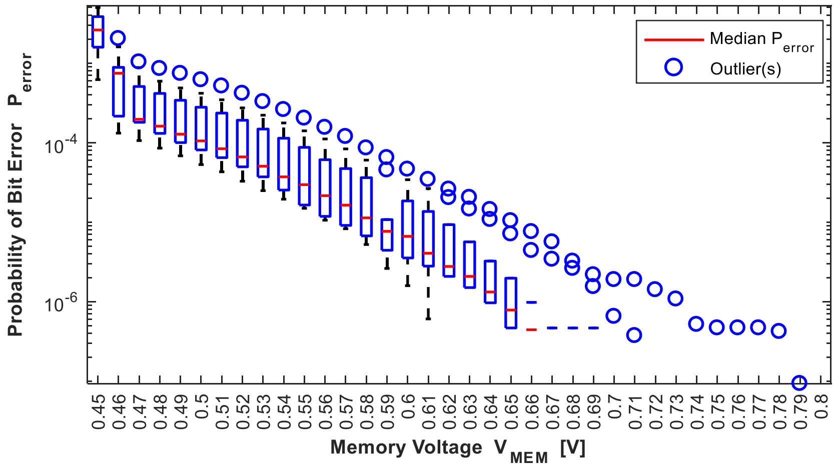


Fig. 2. P_{error} versus V_{MEM} measured over 10 chips. We note there are different V_{MIN} cut-off points, ranging from 0.63 V to 0.8 V.

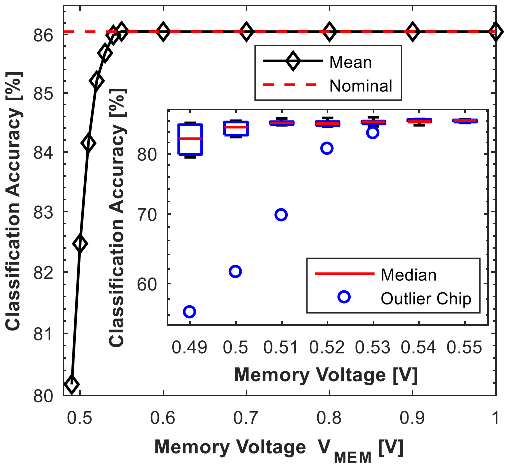


Fig. 3. Degraded classification accuracy at scaled V_{MEM} measured for 10 chips.

III. Voltage Scaling Experiments

A. *Effect of Bit Errors on CIFAR-10 Classification*

To explore the efficacy of memory voltage scaling on CIFAR-10, we fix the digital logic voltage V_{DD} = 0.6 V and scaled memory voltage V_{MEM} until the classification accuracy degrades beyond an acceptable point. The rapid degradation in write/read margins at low V_{MEM} leads to bit errors in both the write and read operation of the SRAM. Though both spatial and temporal variation occur at low V_{MEM}, the temporal variation is small and spatial errors are roughly uniformly distributed in the voltage region of interest. Thus, we propose modeling SRAM behavior at low V_{MEM} with a uniform bit error model, given by the probability of a bit error, P_{error}, as demonstrated in [1] to sufficiently capture approximate SRAM behavior at scaled V_{MEM} for ConvNets. In Fig. 2, we measure the P_{error} at scaled V_{MEM} to demonstrate the spread of P_{error} values, as well as different cut-off V_{MIN}'s, across 10 chips.

Fig. 3 plots the measured classification accuracy versus V_{MEM} across the same 10 chips. Classification accuracy begins to degrade at V_{MEM} = 0.55 V. We see that despite having bit errors in the voltage region 0.55 V < V_{MEM} < 0.8 V, the network can tolerate P_{error}'s in this range with no degradation in the mean classification accuracy. In addition to reducing the memory energy over a conventional implementation (V_{MEM} = 1.0 V), operating V_{MEM} at voltages below the SRAM V_{MIN} reduces leakage power and eliminates the need for additional hardware to switch between active and standby voltages. If a slight degradation in the nominal classification accuracy is tolerable, we can further scale V_{MEM}. The inset plot in Fig. 3 shows how the classification accuracies for V_{MEM} ≤ 0.55 V degrades and increases in variability as V_{MEM} is scaled. We note that there is one chip that performs significantly worse than the other nine chips, hence it is plotted as an outlier in the box-and-whisker plots of Fig. 3.

Fig. 4 plots the trade-off between the average energy consumption of the 10 chips and the spread of the degraded classification accuracies. To illustrate the worst-case classification accuracy spread, we take the mean accuracy over the 10 chips (including the outlier chip) and show error bars representing the 95% confidence interval of the mean classification accuracy in Fig. 4. In systems where we weigh both energy and accuracy equally, we advocate the energy/accuracy ratio to determine the optimal operating point. This trade-off metric is plotted as a contour for both the low and high points of the error bars, as well as the mean classification accuracy (in dark blue).

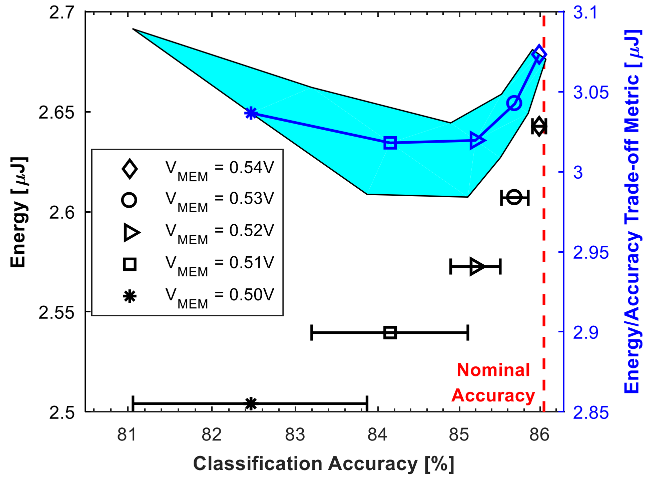


Fig. 4. Energy versus classification accuracy trade-off.

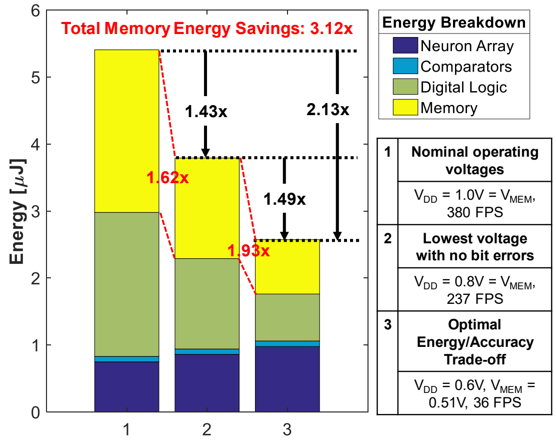


Fig. 5. Energy breakdown of the MSBNN processor and the achieved memory and overall energy savings.

The minima occur around V_{MEM} = 0.51 V to 0.52 V. Fig. 5 shows the achieved memory and overall energy savings of the MSBNN chip with scaled voltages. Digital logic and memory consumes 85% of the total energy (at nominal voltages) and is reduced to 59% at the optimal energy/accuracy trade-off point.

B. Bit Error Accumulation in a Multi-Layer Network

In higher complexity tasks, such as CIFAR-10, it is expected that multiple hidden layers are needed to achieve sufficient classification performance. Unfortunately, this also means that when we scale V_{MEM} for the entire network, we not only incur bit errors per layer for every SRAM read/write operation, but also accumulate bit errors during the operation over all nine layers. To measure the effect of error accumulation over layers, Fig. 6 shows the added error per layer at the critical degraded V_{MEM}'s for one chip. We note that since the MSBNN chip from [4] contains analog noise due to the mixed-signal neuron implementation, we also measure the per layer accuracy degradation at V_{MEM} = 1.0 V (analog errors only) to decouple the analog errors from SRAM bit errors.

We note several interesting observations from Fig. 6. Even at nominal voltages (analog errors only), significant error accumulation occurs up to CONV8, but still achieves nominal classification accuracy (>86.05%). With voltage scaling down to V_{MEM} = 0.51 V, bit errors accumulate over the layers and we observe further degradation at CONV8, but still achieves classification accuracy >86.05%, illustrating both error tolerance of the classifier and how noise sometimes improve classification accuracy. The added error per layer monotonically increases with scaled V_{MEM} and eventually the classification accuracy degrades at V_{MEM} = 0.50 V and 0.49 V. We observe that max pooling (which subsamples the output from CONV4 and CONV6 by 2) helps cut some of the accumulated error over layers, recovering accuracy.

Extensive studies have been done on the effect of quantization errors (converting from floating-point to fixed-point) for multi-layer neural networks. We are interested in establishing similar results for the effect of accumulated bit errors using P_{error} estimates. Following [2], we estimate the effective SNR (γ_{out}) at the output of layer L as the sum of the

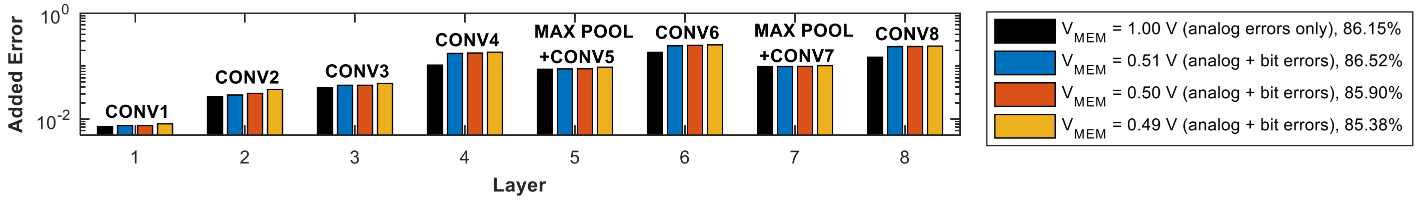
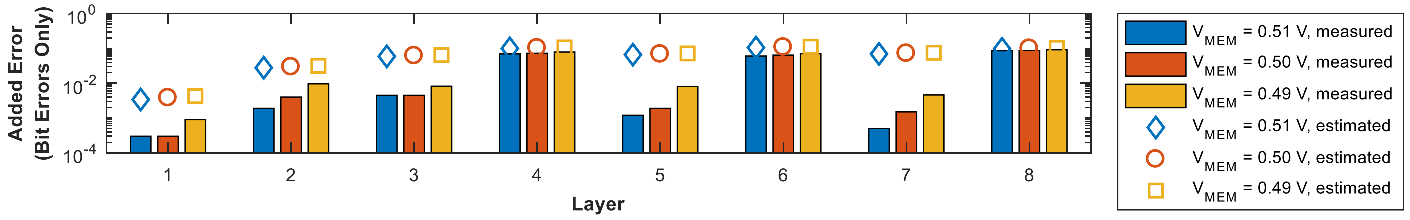


Fig. 6. Measured added error per layer, over different memory voltages, for one chip.



Fig, 7. Measured and estimated added error (due to memory bit errors only) per layer, for different memory voltages on one chip.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

noise contributions from weights *w* and activations *a* per layer:

$$\gamma_{out}^{-1} = \gamma_{a(0)}^{-1} + \gamma_{w(1)}^{-1} + \gamma_{a(1)}^{-1} + \ldots + \gamma_{w(L)}^{-1} + \gamma_{a(L)}^{-1} \quad (1)$$

Details regarding the derivation and assumptions of this equation are described in [2]. Since SNR is related to the noise variance σ^2 by:

$$\gamma = 10 \log(0.25/\sigma^2) \quad (2)$$

and P_{error} can be computed from σ^2 by:

$$P_{error} = \Phi(-0.5/\sigma), \quad (3)$$

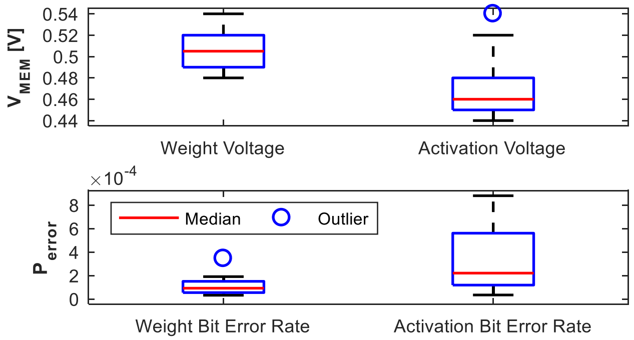
where Φ is the cumulative distribution function, we can estimate the accumulated bit errors over the layers using the expected P_{error} values to calculate σ^2 per layer and sum the contributions to estimate the added bit errors per layer. We note that (1) only considers errors from CONV layers. For the network shown in Fig. 1, we assume that max pooling cuts the noise accumulation by two. Since we are only modeling the effect of bit errors in this model, we subtract out the measured analog errors (shown in Fig. 6) to isolate the effect of bit errors. Fig. 7 shows the comparison between our theoretical model and the measured model over the layers of the network, demonstrating that this model gives a conservative estimate on the added bit errors per layer at a given V_{MEM}. We note that this model does not consider how weights and activations have different error tolerances as discussed in [8], and we make a simplifying assumption that a uniform P_{error} model is sufficient for capturing SRAM behavior at low V_{MEM}. Future work will focus on improving these estimates and models for a tighter bound on the expected accuracy degradation.

C. Bit Error Tolerance Between Weights and Activations

Recent work in quantization have shown that the precision requirements for weights are more stringent than activations [8]. This motivated the idea of exploring the bit error tolerance between weights and activations by separately tuning the voltages V_{WMEM} (weights) and V_{XMEM} (activations). To show that the ConvNet is more tolerant to bit errors in the activations, we chose an acceptable degraded classification accuracy (~99% of nominal, >85.2%), and lowered V_{MEM} per chip until the accuracy degraded below this. We set V_{WMEM} at the same V_{MEM} voltage to achieve >85.2% classification accuracy, while lowering V_{XMEM} until the accuracy degraded beyond 85.2%. Fig. 8 illustrates the spread of voltages and P_{error}'s obtained from this experiment, showing that the bit error requirements for activations are less stringent compared to weights. We note that the P_{error} spread for activations is much wider due to the aggressive V_{XMEM} scaling but is still acceptable due to the error resilience of the ConvNet.

IV. Benchmark Comparison of CIFAR-10 vs. MNIST

Table 1 compares the P_{error} requirements for CIFAR-10 using the MSBNN processor and with silicon-validated P_{error} values for MNIST [1], [5]. Consistent with findings from quantization experiments [8], we observe that the CIFAR-10 BinaryNet has more stringent P_{error} requirements compared to MNIST implementations. Additionally, the CIFAR-10 network requires more layers than the MNIST networks, which further explains the reduced tolerance due to error accumulation. Since it would not be fair to compare SRAM power savings across different technologies (given different nominal SRAM voltages and scaling characteristics), we instead use P_{error} as a common metric to quantify bit error tolerance across different technologies and dataset benchmarks. We note that despite the different technologies used in [1] and [5], both achieve similar P_{error} values on the MNIST dataset, validating this approach.



Mean Classification Accuracy	One Memory Voltage ($V_{MEM} = V_{WMEM} = V_{XMEM}$)	Split Memory Voltages (V_{WMEM}, V_{XMEM})
	85.39%	85.54%

Fig. 8. Spread of weight and activation voltages within 99% of nominal classification accuracy. Activations are more error tolerant than weights.

Table 1. P_{error} comparison of CIFAR-10 BinaryNet and MNIST networks

🟩 Nominal Accuracy	🟦 99% of Nominal	🟥 98% of Nominal

CIFAR-10, 9-layer BinaryNet (This Work)		MNIST, 3-layer 8 bit ConvNet [1]		MNIST, 3-layer Ternary Neural Net [5]	
28nm ASIC Processor, 328 kB On-Chip SRAM		28nm FDSOI SRAM Chip, 32.5 kB SRAM requirement		65nm SRAM Chip, 8 kB SRAM requirement	
Classification Accuracy	P_{error}	Classification Accuracy	P_{error}	Classification Accuracy$^{\Delta}$	P_{error}
🟩 0.8605	2.96E-05	🟩 0.9857	2.20E-03	🟩 0.9468	-
🟦 0.8587	5.05E-05	🟦 0.9842	8.30E-03	🟦 0.9460	2.30E-03
🟦 0.8567	8.37E-05	🟦 0.979	2.04E-02	🟥 0.9300	1.56E-02
🟥 0.8487	1.05E-04	🟥 0.9687	4.07E-02	$^{\Delta}$ modified MNIST dataset	

V. Summary

In this paper, we studied the memory bit error tolerance of ConvNets, comparing the tolerable P_{error} values between MNIST and CIFAR-10 benchmarks. We measured the efficacy of SRAM voltage scaling on a multi-layer CIFAR-10 network implemented in a 28 nm processor, achieving memory energy savings of 3.12x with minimal accuracy degradation at the optimal energy-accuracy trade-off point. We further analyzed and measured the effect of bit error propagation through the 9-layer ConvNet and demonstrated that activations are more error tolerant than weights.

Acknowledgment

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA. The authors would like to thank the TSMC University Shuttle Program for providing chip fabrication.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

REFERENCES

[1] L. Yang and B. Murmann, “SRAM Voltage Scaling for Energy-Efficient Convolutional Neural Networks,” in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, 2017, pp. 7–12.

[2] D. D. Lin, S. S. Talathi, and V. S. Annapureddy, “Fixed Point Quantization of Deep Convolutional Networks,” in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 2859–2858.

[3] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1,” *arXiv Prepr. 1602.02830v3*, 2016.

[4] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, “An Always-On 3.8μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor with all Memory on Chip in 28nm CMOS,” in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2018, pp. 222-223.

[5] X. Sun *et al.*, “Low-VDD Operation of SRAM Synaptic Array for Implementing Ternary Neural Network,” *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 25, no. 10, pp. 2962–2965, 2017.

[6] B. Moons, D. Bankman, L. Yang, B. Murmann, and M. Verhelst, “BinarEye: An Always-On Energy-Accuracy-Scalable Binary CNN Processor With All Memory On Chip In 28nm CMOS,” in *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, 2018, in press.

[7] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, “DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients,” *arXiv Prepr. arXiv1606.06160*, 2016.

[8] C. Sakr, Y. Kim, and N. Shanbhag, “Analytical Guarantees on Numerical Precision of Deep Neural Networks,” *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, pp. 3007–3016, 2017.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE