Fully Integrated Analog Machine Learning Classifier Using Custom Activation Function for Low Resolution Image Classification

Sanjeev Tannirkulam Chandrasekaran[‡], Akshay Jayaraj[*], Vinay Elkoori Ghantala Karnam[‡], Imon Banerjee[†], and Arindam Sanyal[‡]
[*]Intel Corporation, Folsom CA 95630, USA.
[†]Department of Biomedical Informatics, Emory University, Atlanta GA 30322, USA.
[‡]Electrical Engineering, University at Buffalo, Buffalo, NY 14260, USA.
Email: stannirk@buffalo.edu

***Abstract*—This paper presents fully-integrated analog neural network classifier architecture for low resolution image classification that eliminates memory access. We design custom activation functions using single-stage common-source amplifiers, and apply a hardware-software co-design methodology to incorporate knowledge of the custom activation functions into the training phase to achieve high accuracy. Performing all computations entirely in the analog domain eliminates energy cost associated with memory access and data movement. We demonstrate our classifier on multinomial classification task of recognizing down-sampled handwritten digits from MNIST dataset. Fabricated in 65nm CMOS process, the measured energy consumption for down-sampled MNIST dataset is 173pJ/classification, which is $3\times$ better than state-of-the-art. The prototype IC achieves mean classification accuracy of 81.3% even after down-sampling the original MNIST images by 96% from 28×28 pixels to 5×5 pixels.**

***Index Terms*—machine learning; analog neural network; low resolution image classification; custom activation function**

I. Introduction

Advances in machine learning (ML) techniques has enabled high accuracy image classification which is one of the earliest applications of ML and computer vision. The well-known ML image classification algorithms are designed for high resolution images. As an example, the popular ImageNet dataset [1] has images with average resolution exceeding 482×418 pixels. However, many applications need to classify images with very low resolution, such as far-field detection scenarios for surveillance where the region-of-interest is just a few pixels [2], or remote health monitoring of patients while preserving their privacy by using low resolution images [3]–[5]. While images of near-field objects can have thousands of pixels in area, far-field objects may be as small as 50 pixels in area [2], [6]. On the other hand, low resolution cameras are deliberately used for human health monitoring such that patients are not identified from their images, which can be as small as 8×8 pixels [7]. Low resolution image classification also reduces computation cost and energy consumption which can lead to integration of ML classifier with image sensor for real-time classification.

Another application for low resolution image classification is in the area of remote sensing using wireless image sensor networks [8]. While research on CMOS image sensor (CIS) has led to cameras that consume very low power [9], [10], energy cost of transmitting high resolution raw image data wirelessly is still factors of magnitude higher than capturing the image itself, and the transmitter limits battery life of the sensor [11]. Integration of ML classifier with image sensors can significantly reduce transmission energy by selecting only the frames-of-interest to be transmitted to the back-end for deeper analysis. Since the integrated ML classifier acts a coarse classifier, the sensor energy consumption can be further reduced through classification on down-sampled, low resolution image. Fig. 1 shows two applications of low resolution image classification. Integration of ML classifier with image sensors on the same chip can reduce energy consumption by pushing the analog-to-digital converters (ADCs) after the classifier and thus, reduce the number of ADCs required (Fig. 1(a)), or reduce energy by only transmitting images of interest (Fig. 1(b)). While, recent works on CIS incorporate object-detection techniques [12], [13], these works extract histogram-of-oriented gradients (HOG) as features from the images for classification, rather than use each pixel outputs as features as is done in this work. Hence, the object detection algorithm in [12] consumes more than 250pJ/pixel compared to 6.9pJ/pixel for this work.

The bottlenecks for integrating ML classification on sensors are – a) large amount of memory required to store neural network (NN) weights and intermediate results; b) energy cost of memory access. Existing approaches to counter these bottlenecks are – a) analog/mixed-signal implementation of ML algorithms [14]–[17]; b) reduced bit precision for less storage requirement [18], [19]; c) in/near-memory computations [19]–[22] to limit data movement between memory and computing units. However, prior works using analog computing have demonstrated only partial on-chip implementation of NN, usually the first hidden/convolutional layer [18], [19], [21] or only inner product calculation [14], [16], [20], [22], and implemented the remaining layers off-chip.

This work presents a fully integrated, on-chip artificial neural network (ANN) classifier architecture that uses analog circuit design to address the above-mentioned bottlenecks associated with integration of ML algorithms on sensors for

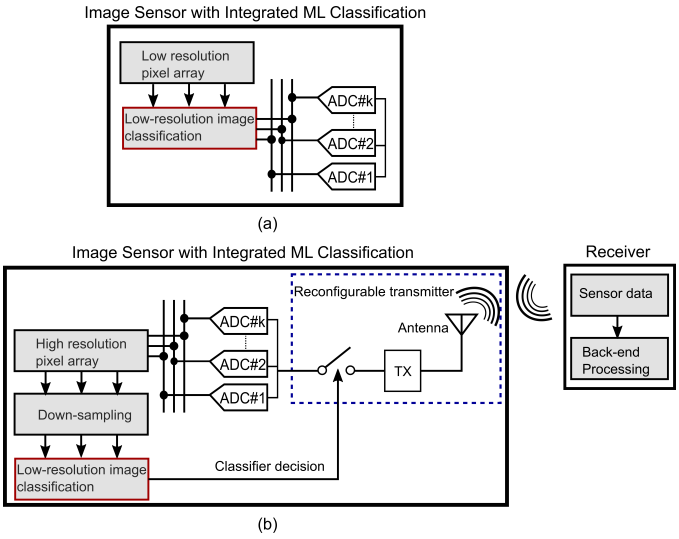


Fig. 1: Integration of ML classifier for low resolution image classification in a) applications with low resolution pixel array; and b) reducing transmission energy by only transmitting frames-of-interest from high resolution pixel array

low resolution image classification. We design custom activation function using analog circuit, and incorporate the circuit model into training phase of the ANN classifier to reduce discrepancy between software ANN model and analog ANN circuit. Designing custom activation circuit results in lower transistor count/activation circuit compared to conventional techniques [23], [24] which try to replicate ideal mathematical functions, like tanh/sigmoid, at the cost of design complexity. Low transistor count is a key enabler in fitting a complete ANN in a small area. The ANN weights are reduced to 4-bit precision and are encoded as transistor widths to eliminate storage. All the intermediate computations are performed in analog domain to eliminate memory access and save energy. The ANN weights are encoded as transistor widths in our design, but the weights can be made re-configurable as in [18]. The ANN performance is demonstrated on the popular MNIST dataset [25] consisting of handwritten images of digits from 0-9 to allow for comparison with existing works. To perform low resolution image classification, the MNIST images are down-sampled from 28×28 to 5×5 resolution. Compared to our prior work [26] which presented circuit design for the first layer only, we have implemented the entire ANN and presented detailed analysis of non-idealities in our design as well as measurement results. Even with 96% reduction in image resolution, a prototype ANN in 65nm CMOS process classifies MNIST test images with 82% accuracy. The rest of the paper is organized as follows: architecture of the proposed ANN is presented in Section II, measurement results on MNIST dataset are presented in Section III, while the conclusion is brought up in Section IV.

II. Proposed Architecture

The proposed ANN circuit schematic is shown in Fig. 2. The 28×28 features of MNIST dataset are down-sampled to 5×5 features, converted to analog voltages using off-chip DAC and given to the ANN chip as inputs. Bilinear interpolation is used for down-sampling by performing weighted average of neighboring pixels. Bilinear interpolation produces a continuous function from the 28×28 pixels by computing distance-weighted average of of the 4 nearest pixels [27]. The continuous function is then re-sampled to create the 5×5 image. Fig. 3 shows samples of original images from MNIST dataset and their down-sampled versions.

While off-chip DACs have been used to convert the digital features to analog inputs for this work, in practical application the ANN will be integrated with the image sensor and will directly use analog pixel outputs, thus removing the DACs from the signal path. The ANN circuit has 1 hidden layer with 28 neurons and 1 output layer with 10 neurons [26]. Both the ANN layers uses common-source (CS) amplifiers to implement custom, non-linear activation functions and performs multiply-and-accumulate (MAC) in current domain as shown in Fig. 2. Outputs from the output layer are sent to an argmax layer which determines the output neuron with the highest value, and hence, the classifier label. The argmax layer consists of comparators which compare value of each output neuron with all other output neurons. The comparators are sized up to reduce mismatch and no offset calibration is performed in this work. Activation function design and AI training methodology are discussed in subsequent sections.

A. Custom Activation Function Design

Pseudo-differential CS amplifier architecture is used for designing activation functions in the hidden and output layers. The custom activation function design methodology is illustrated with a 2-input single-ended hidden layer circuit shown in Fig. 4. The voltage inputs, V_1 and V_2, are converted into currents I_1 and I_2 respectively, by 2 NMOS transistors with widths W_1 and W_2, and identical channel lengths. The summed current $I_1 + I_2$ is converted back to voltage using a diode-connected PMOS load, and its source-to-drain voltage acts as output of the hidden layer. In general, for N inputs, the hidden layer output can be written as

$$V_{out} = V_{dd} - g\left(\sum_{i=1}^{N} W_i \cdot f(V_i)\right) \tag{1}$$

where $g(\cdot)$ represents the I-V characteristic of the PMOS load while $W_i \cdot f(\cdot)$ represents the V-I characteristic of an NMOS input transistor with width W_i. The argument of $g(\cdot)$ in (1) represents current-domain MAC operation with W_i being weights of the ANN. The transfer functions $f(\cdot)$ and $g(\cdot)$ are extracted through SPICE simulation and incorporated into our ML training. While $f(\cdot)$ and $g(\cdot)$ are nonlinear, the summation operation $\sum W_i \cdot f(V_i)$ has to be linear which is ensured by restricting the input swing and dynamic range of the weights as described in following sections.

To accommodate both positive and negative weights in the hidden layer, we use a pseudo-differential architecture as shown in Fig. 2 in which positive weights are assigned to the left branch and negative weights to the right branch. The output of each hidden neuron is difference between the positive

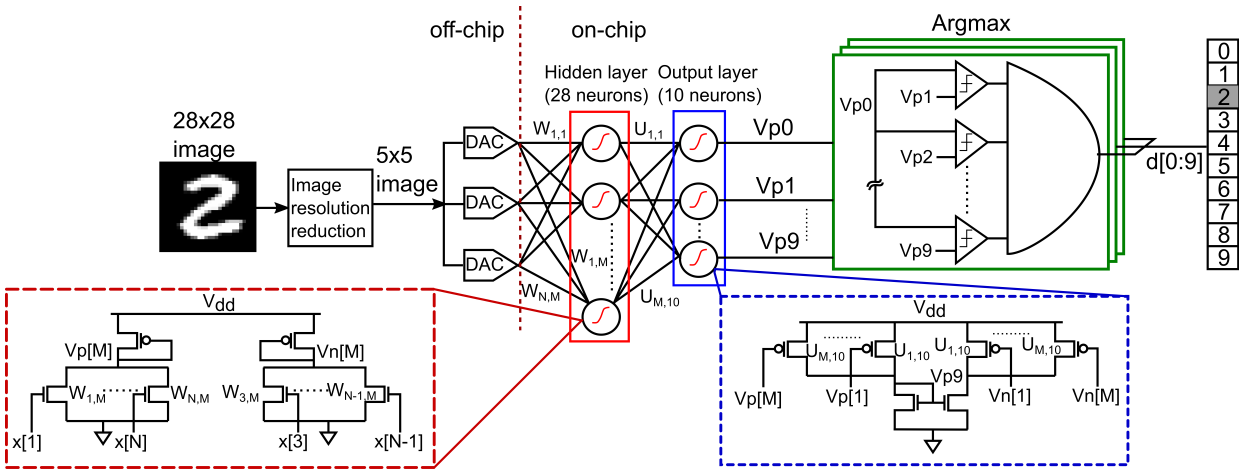
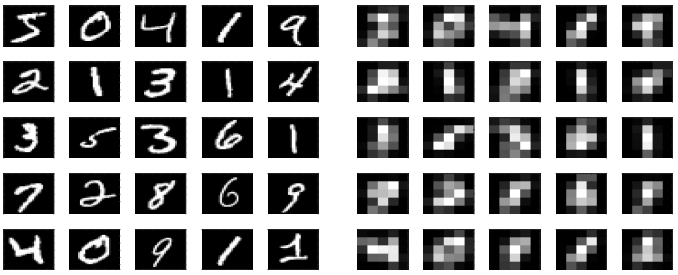


Fig. 2: Circuit schematic for proposed analog ANN demonstrated on down-sampled MNIST dataset



(a) (b)

Fig. 3: Samples of MNIST images with (a) original 28×28 resolution, and (b) 5×5 resolution used in this work

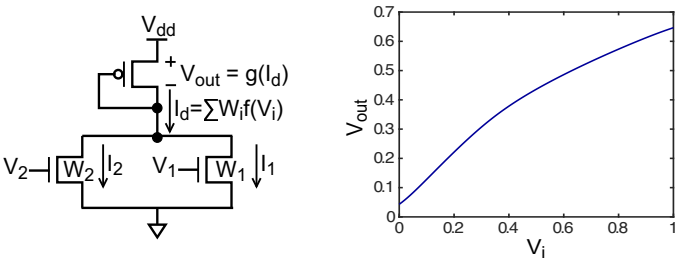


Fig. 4: 2-input single-ended hidden layer with custom activation function

and negative outputs, V_p and V_n, respectively. Output of the M-th hidden neuron is expressed mathematically as

$$V_p[M] - V_n[M] = g\left(\sum_{i=1}^{N} W_{n,i}[M] \cdot f(V_i[M])\right) - g\left(\sum_{i=1}^{N} W_{p,i}[M] \cdot f(V_i[M])\right) \quad (2)$$

where $W_{p,i}[M] = \{(W_i[M] + |W_i[M]|)/2\}$ and $W_{n,i}[M] = \{(|W_i[M]| - W_i[M])/2\}\ \forall\ i \in [1, N]$.

We use a pseudo-differential CS amplifier with single-ended output for the output layer as shown in Fig. 2 to implement custom softmax activation function. Fig. 5 shows our custom activation function for output layer illustrated with 2 inputs. For each output layer neuron, if a weight is positive, the corresponding hidden layer output $V_p[M]$ is connected to the right branch and $V_n[M]$ to the left branch, and vice versa if the weight is negative, as shown in Fig. 5.

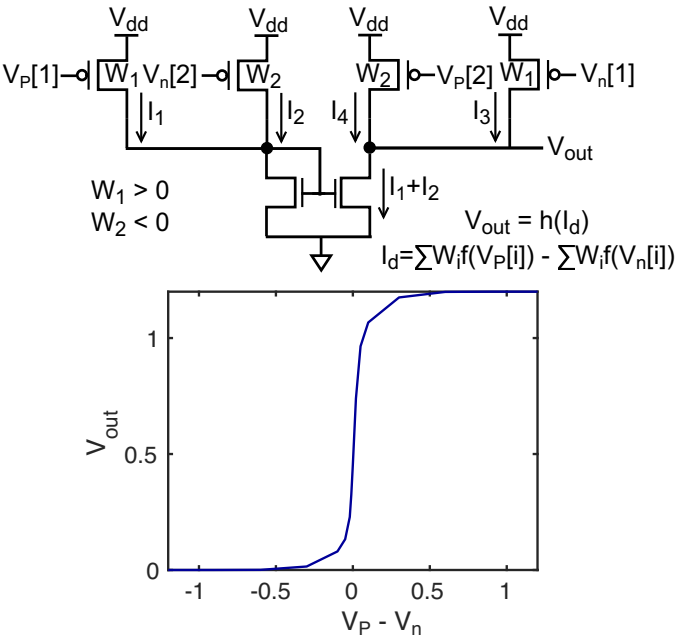


Fig. 5: 2-input output layer with custom softmax activation

B. Hardware-Software Co-Design

The hardware-software co-design methodology used to design the proposed ANN is shown in Fig. 6. As described in Section II-A, CS amplifier transfer curves are imported into Matlab training phase. The ANN is initialized with random weights and the transfer functions $f(\cdot)$ and $g(\cdot)$ are used to calculate output of the ANN. We use mean-squared error (MSE) as cost function to calculate error between ANN output and ground truth, and use stochastic gradient descent (SGD) algorithm to minimize the MSE. SGD computes derivative of

MSE with respect to each weight and updates each weight through back-propagation. This process is done iteratively till the MSE converges. Weights of the ANN are then encoded as transistor widths in the hidden and output layers.

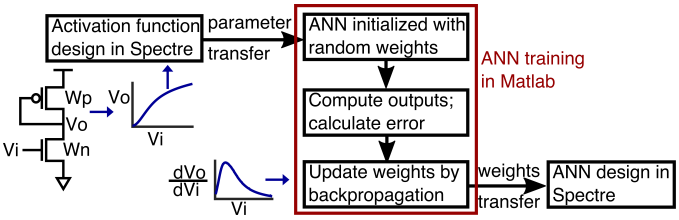
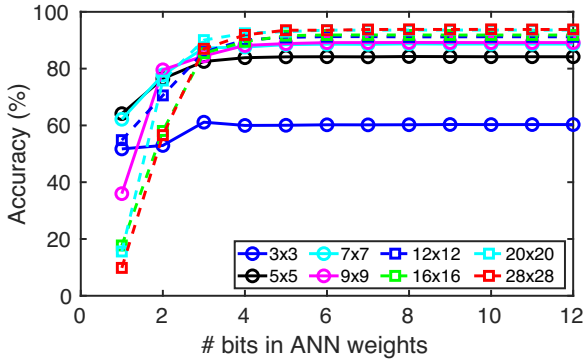


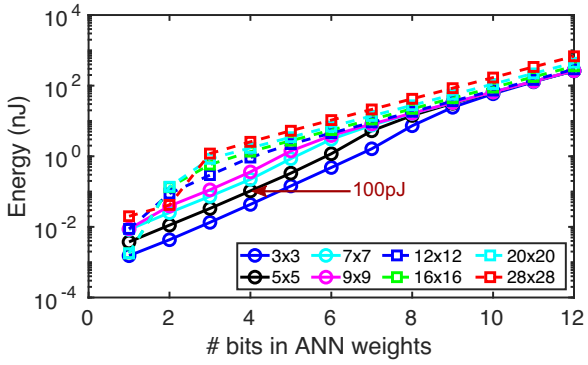
Fig. 6: Hardware-software co-design methodology for ANN

Fig. 7(a) shows simulated accuracy as a function of number of pixels in the image and number of bits in the ANN weights. Effects of non-idealities, such as noise and random mismatch, are not included in this simulation. The simulated accuracy is less than 60% for image resolution of 3×3, while the accuracy improves to more than 80% for image resolution exceeding 5×5 and ANN weights with more than 3-bit resolution. Classification accuracy for different image resolution does not improve significantly for ANN weights with 4-bit or higher resolution. Fig. 7(b) shows the simulated classification energy for different image and ANN weight resolutions. The energy numbers in Fig. 7(b) includes contribution from the hidden and output layers, and not the argmax layer. Classification energy increases monotonically with image and ANN weight resolutions. For this work, an image resolution of 5×5 pixels and 4-bit ANN weights are selected to mimic low resolution image classification problem while meeting area constraints on the IC, and to optimize energy consumption, respectively. The ANN weights are truncated after each training epoch. An interesting observation in Fig. 7 is that the classification accuracy and energy consumption for 28×28 pixels are less than images with lower resolution when the ANN weights are truncated to 1 or 2 bits. This is because the dynamic range of floating point weights in the ANN increases with number of features in the input. Hence, when truncated to 1 or 2 bits, a larger fraction of the quantized ANN weights are '0' for larger feature sizes than for smaller feature sizes. This results in the ANN with 28×28 pixels input to have less accuracy and lower energy consumption at 1-2 bits resolution than for smaller feature sizes. Fig. 8 shows the simulated classification accuracy versus number of neurons in the hidden layer. Classification accuracy increases initially with hidden neurons, but does not change significantly once the number of hidden neurons exceeds 28, while energy consumption keeps increasing with number of hidden neurons. Hence, the hidden layer is designed with 28 neurons for this work.

Fig. 9(a) shows the summed positive and negative weights for each hidden neuron, while Fig. 9(b) shows the summed positive and negative weights for each output neuron. The maximum summed weight for the hidden layer is 85 for the 17-th neuron, while the maximum summed weight for the output layer is 200 for the 7-th neuron.



(a)



(b)

Fig. 7: a) Classification accuracy, and b) energy consumption vs ANN weight truncation

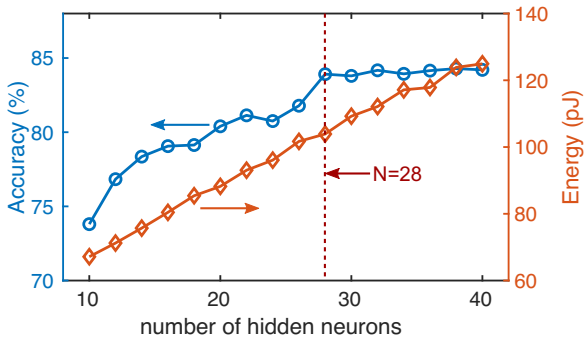
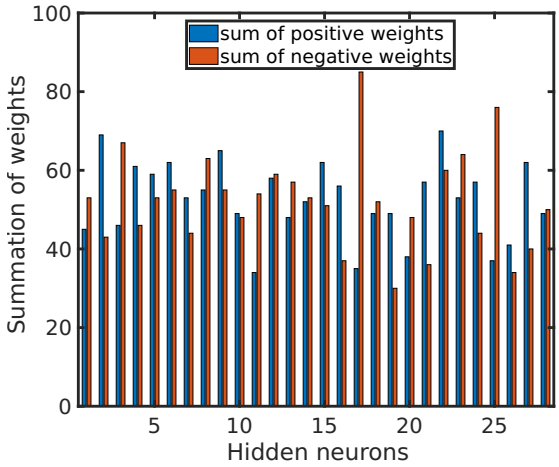


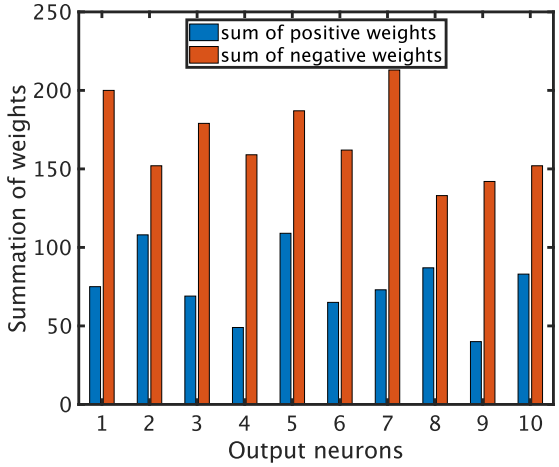
Fig. 8: Classification accuracy and energy vs number of hidden neurons

C. Effect of Analog Non-idealities

The ANN circuit has several non-idealities, such as non-linearity of current summation in the activation functions and random mismatch in all the layers. While $f(V_i)$ is assumed to be independent of transistor width in (1), in practice $\sum W_i \cdot f(V_i)$ depends on the transistor width and the number of inputs which introduces non-linearity in the current summation. Fig. 10(a) and (b) show linearity of current summation (argument of $g(\cdot)$ in (1)) in the hidden and output layers. For the hidden layer, we compare current through an NMOS transistor with unit width with current through an NMOS transistor with width of 85 units across input voltage. The maximum error between the two curves is 3%. The input

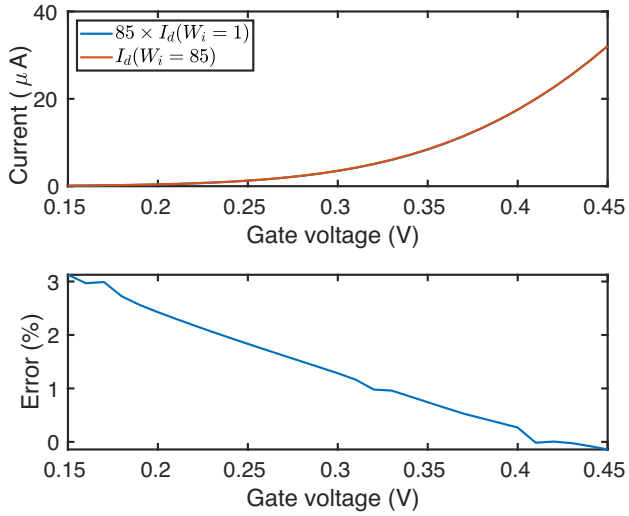


(a)

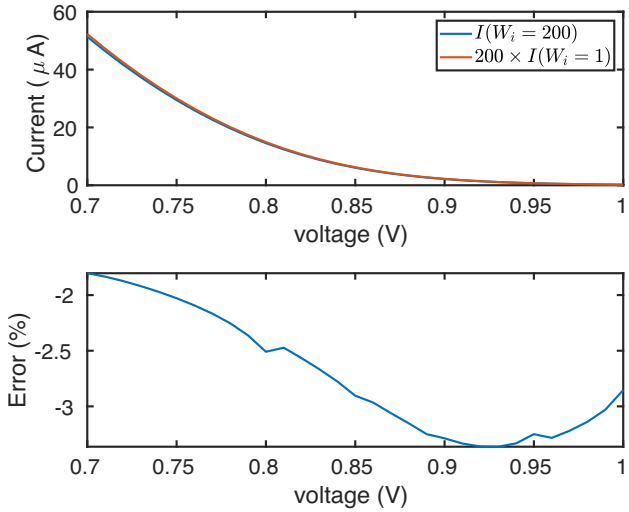


(b)

Fig. 9: Distribution of summed weights for (a) hidden, and (b) output layers



(a)



(b)

Fig. 10: Linearity of current summation for (a) hidden, and (b) output layers

voltage range of 0.15-0.45V is chosen for the hidden layer to reduce power consumption of the hidden layer, and is realized by scaling each input feature and centering around 0.3V. Similarly, for the output layer we compare current through a PMOS transistor with unit width with current through a PMOS transistor with width of 200 units across input voltage. The input voltage range for output layer is set by the output voltage range of the hidden layer. The maximum error between the two curves in Fig. 10(b) is less than 3.5%. While non-linearity in current summation can be addressed by incorporating the non-linear terms in the ANN model during training phase, non-linearity correction is not done in this work since it does not affect classification accuracy for 5×5 resolution images as will be verified through the measurement results in Section III. For classification of high resolution images, non-linearity correction needs to be added to the ANN training phase.

In addition to non-linearity current summation, other non-idealities in our ANN are random mismatch and thermal noise. The transistors are sized-up to reduce mismatch. Fig. 11 shows the result of 100 monte-carlo simulation on the proposed custom activation function circuits for hidden and output layer. The maximum standard deviation of error from nominal transfer curve is 1.2mV for the hidden layer and 26mV for the output layer. The maximum deviation from nominal transfer curve occurs around the zero crossing for the soft-max activation function in the output layer. To analyze the effects of random mismatches in the activation functions on classification accuracy, we performed simulations on the test-set by introducing static random mismatch to each neuron, and repeated the simulations 100 times. Fig. 12(a) and (b) show the histograms of classification accuracy for random mismatch in hidden neurons, and random mismatch in both hidden and output neurons respectively. The mean classification accuracy is 84.03% with standard deviation of 0.12% for random mismatch in only hidden neurons, while the mean classification accuracy is 83.82% with standard deviation of 0.29% for random mismatch in both hidden and output neurons.

Aside from the neurons in the hidden and output layers,

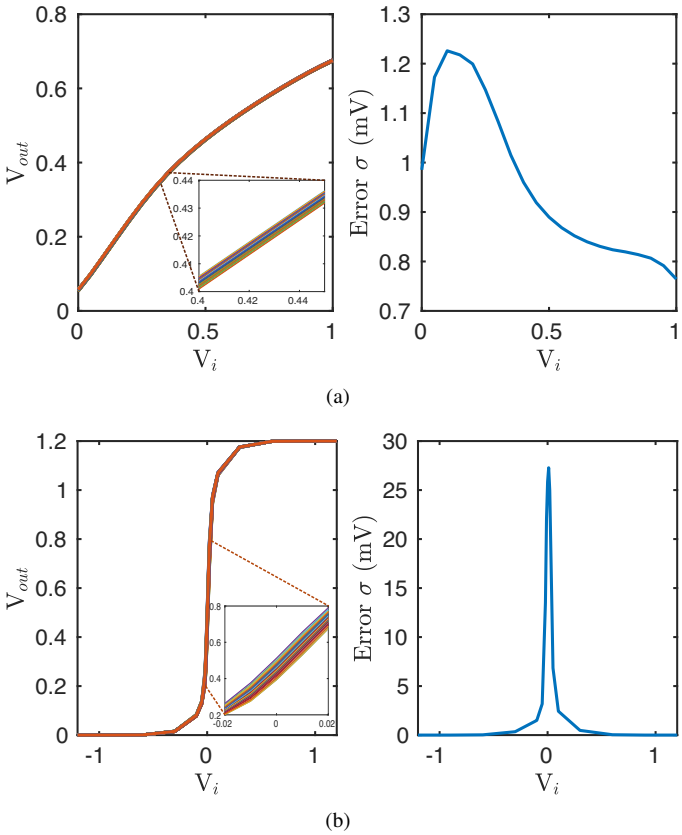


Fig. 11: Monte-carlo simulation of activation functions in a) hidden layer, and b) output layer

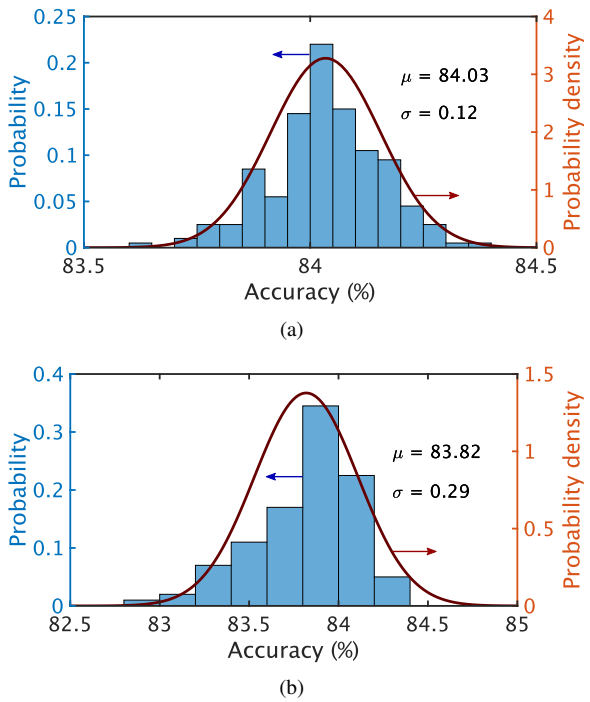


Fig. 12: Histogram of accuracy vs random mismatch in activation functions in a) hidden layer, and b) both hidden and output layer

the comparators in the argmax layer are another major source of random mismatch. Random mismatch introduces different offset in each comparator in the argmax layer and changes the decision boundaries which affects classification accuracy. The argmax layer uses 45 comparators. To analyze the effect of random mismatch in argmax layer on classification accuracy, the ANN is simulated by varying standard deviation of comparator offset from 0 to 35mV. For each value of standard deviation, the simulation is repeated 100 times. Fig. 13(a) shows the mean and standard deviation of classification accuracy as a function of comparator offset. As expected, the mean classification accuracy reduces with increase in comparator offset. The classification accuracy is greater than 83% if the comparator offset is less than 10mV. Comparator offset can be reduced further by sizing up the input transistors and burning more power. Fig. 13(b) plots simulated comparator offset as a function of energy. The comparator offset is calculated using monte-carlo simulations. An input ramp with 1mV step is applied to the comparator input, and 200 point monte-carlo simulation is performed for each step. For each input value, the comparator offset is calculated from inverse of the cumulative normal distribution of probability of '1' [28]. For this design, each comparator is biased to consume 0.96pJ energy which results in an offset of 8.9mV. In addition to simulation, measurement results on 5 chips presented in Section III also show robustness against mismatch.

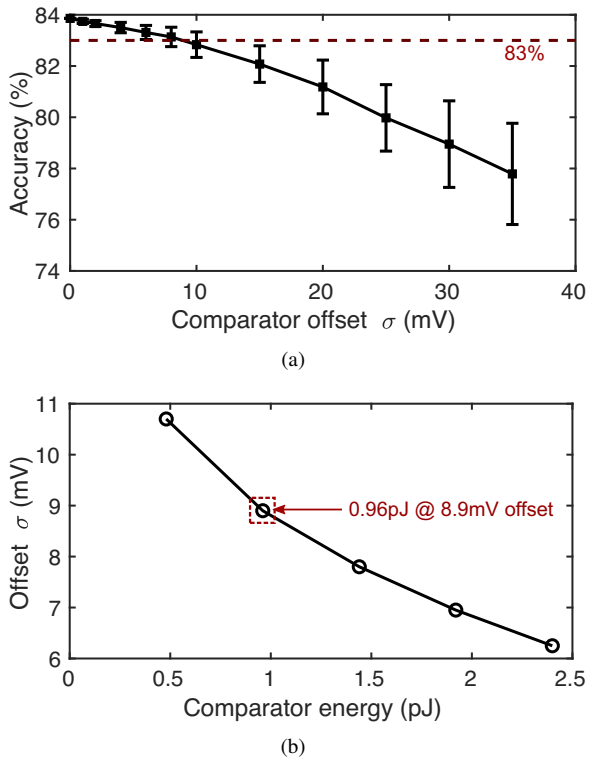


Fig. 13: (a) Classification accuracy versus comparator offset (b) comparator offset as a function of energy consumption

To estimate the effect of thermal noise on classification accuracy, we characterized noise of hidden and output layer neurons and the comparator. Each hidden neuron has output-

referred noise of 0.3mV and each output neuron has output-referred noise of 0.1mV which are calculated at the lowest overdrive voltages corresponding to the worst-case scenario. The comparator has an input-referred noise of 0.5mV. The ANN is simulated with thermal noise enabled, and the simulation is repeated 100 times for each sample image on the test set. Fig. 14 shows the histogram of classification accuracy. The ANN has a mean classification accuracy of 83.11% with standard deviation of 0.04% which shows low sensitivity to thermal noise.

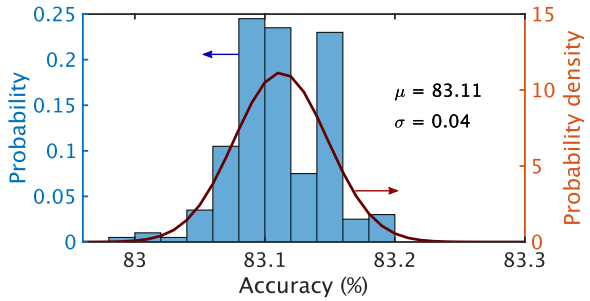


Fig. 14: Histogram of accuracy vs thermal noise in the classifier

In addition to thermal noise, the comparators in the argmax layer also creates kickback noise that couples voltage at regeneration node of the comparator to its inputs [29]. However, the kickback noise is partially mitigated due to relatively low speed operation of the comparator which reduces the rate at which voltages at the regeneration nodes fall, and hence, the coupling to the comparator inputs. Fig. 15 shows example of kickback noise for a sample image each from classes '0' and '4'. The kickback noise has small amplitude of 0.2-0.3mV which is not expected to reduce ANN performance significantly, as is evident from measurement results which are within 2% of simulation results without the effect of comparator kickback.

III. Measurement Results

Fig. 16 shows the die photograph of the proposed classifier in 65nm CMOS process as well as the chip layout. The classifier chip has a core area of 0.42mm^2 and consumes 173pJ energy from 1.2V power supply at 5MHz operating speed.

The MNIST classifier is trained on 60,000 images and 4 chips are tested with 10,000 test images. Fig. 17 shows the measured confusion matrix for 1 test chip which graphically summarizes its performance for every digit. The classifier has the most false positives for class '4' indicating that the classifier had difficulty separating classes '4' and '9', likely due to the similarity between the digit shapes after severely reduced image resolution. The overall accuracy of the MNIST classifier is 82%. While accuracy is a good indicator of performance of a classifier, to fully evaluate effectiveness of a classifier, we need to look at two other parameters - precision and recall. Precision measures what proportion of positive identifications are correct, while recall measures what proportion of actual positives is identified correctly [30]. Table I reports measured precision and recall of our ANN. Since precision and recall

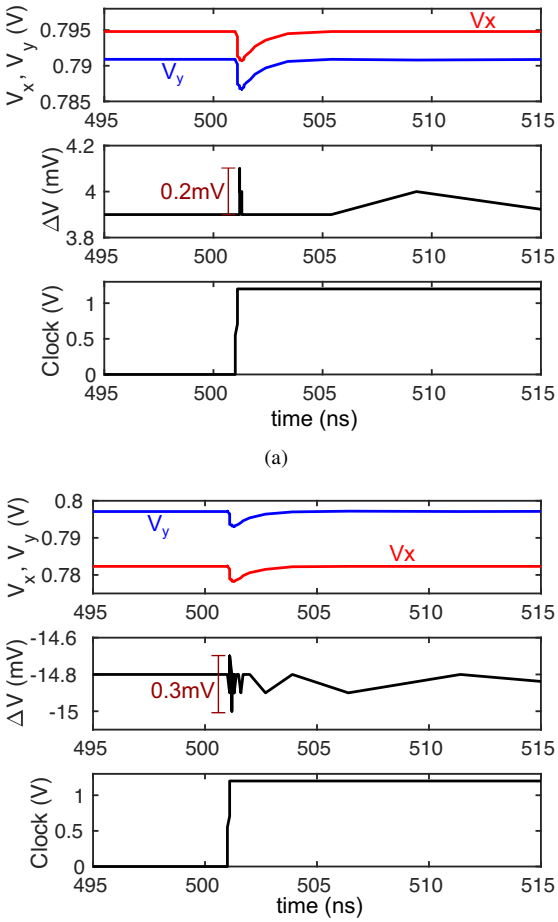


Fig. 15: Kickback noise simulation on sample images from a) class '0' (b) class '4'

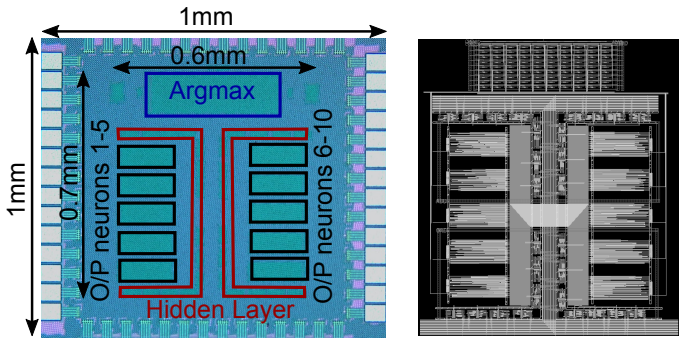


Fig. 16: Die photograph and chip layout

are usually in tension, and improving one degrades the other, we also report f1-score which is harmonic mean of precision and recall. The proposed classifier has the high f1-scores for classes '0' and '5', and the lowest f1-scores for classes '4' and '9' which also confirms that the classifier has difficulty in separating the classes '4' and '9'.

Fig. 18(a) shows the energy breakdown by each layer. The hidden, output and argmax layers consume average energy of 52pJ, 78pJ and 43pJ respectively. Fig. 18(b) shows the measured energy consumption for each MNIST class. The

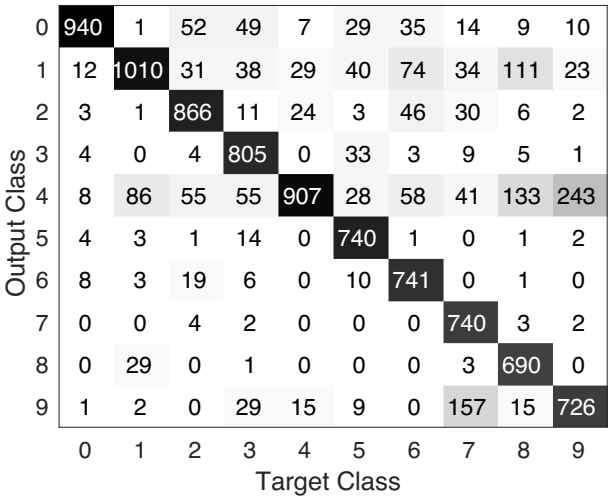


Fig. 17: Measured confusion matrix

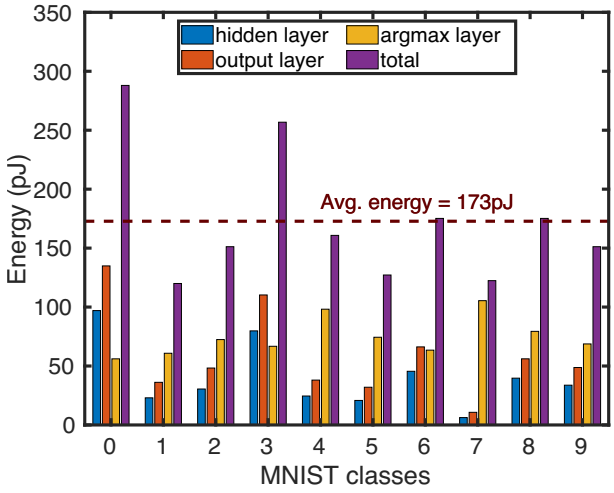
TABLE I: Precision, recall and f1-score of the classifier

Class	Precision	Recall	f1-score
0	0.82	0.96	0.88
1	0.72	0.89	0.80
2	0.87	0.84	0.86
3	0.93	0.80	0.86
4	0.56	0.92	0.70
5	0.96	0.83	0.89
6	0.94	0.77	0.85
7	0.98	0.72	0.83
8	0.95	0.71	0.81
9	0.76	0.72	0.74

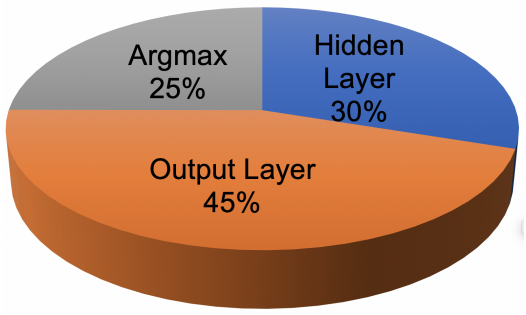
classifier has the highest energy consumption of 288pJ for class '0' and minimum energy consumption of 120pJ for class '1'. The average energy consumption of the classifier is 173pJ. The reason for higher energy consumption for class '0' is due to the distribution of pixel intensity as can be seen in Fig. 18(c) which plots the mean input voltages for each pixel location for classes '0' and '1'. As can be seen from Fig. 18(c), class '0' has higher average values for the pixels compared to class '1' which increases the energy consumption for classifying class '0'.

Fig. 19(a) shows measured accuracy versus supply voltage and operating frequency. The measured accuracy drops with reduction in supply voltage and increase in operating frequency. The maximum speed is limited by the argmax layer but can be improved at the cost of increased power consumption. Fig. 19(b) shows the measured accuracy for 5 chips. The mean accuracy is 81.3% with a low standard deviation of 0.85% which demonstrates robustness against random mismatch.

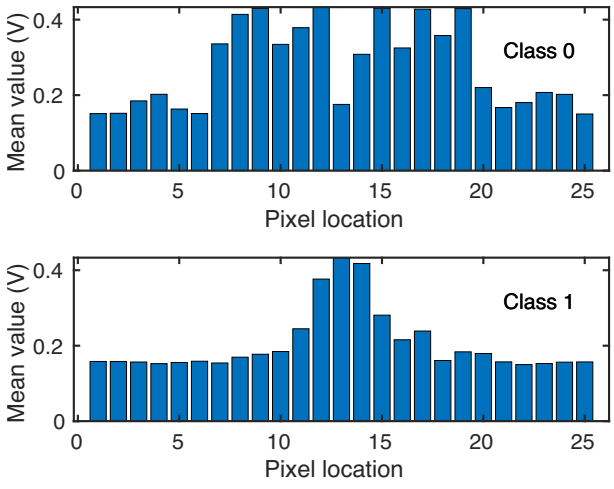
Table II summarizes the performance of the proposed ANN prototype and compares our work with state-of-the-art ASICs demonstrated on low resolution images from the MNIST dataset. State-of-the-art analog ASICs [18], [19] have low energy/classification but have implemented only the first binary classification layer on-chip while the other layers are realized off-chip in digital-domain. In contrast, the proposed



(a)



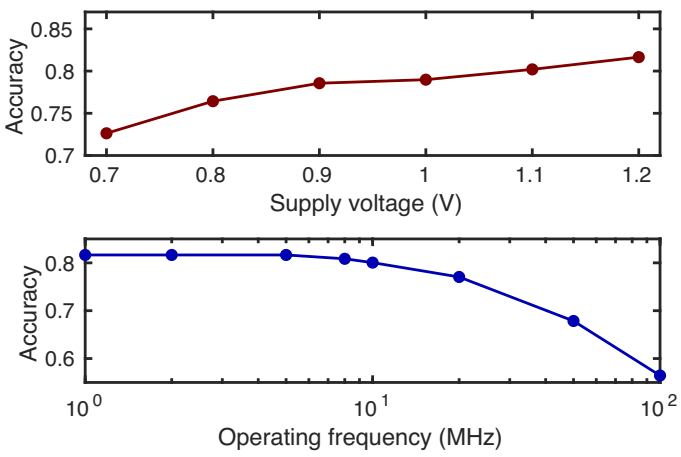
(b)



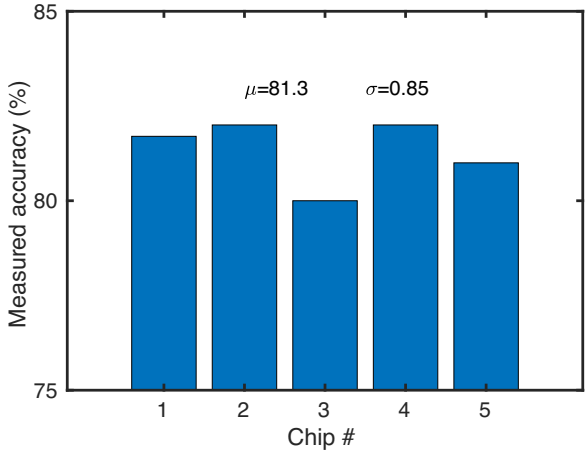
(c)

Fig. 18: Measured average energy (a) distribution by class and layer, (b) breakdown by layer, and (c) distribution of mean pixel values for classes '0' and '1'

analog ANN is fully integrated but consumes 3× lower energy than state-of-the-art. While accuracy of our fully integrated ANN is 82%, the accuracy increases to 88% if 49 and 81 features are used. The estimated energy/classification of our ANN with 49 and 81 features increase to 267pJ and 403pJ respectively, but is still 2× less than energy of [18] while having comparable accuracy. If only the first hidden layer



(a)



(b)

Fig. 19: (a) Measured accuracy versus power supply and operating frequency (b) histogram for accuracy of multiple chips

for our ANN is implemented on-chip, and a digital all-vs-all neural network with SVM binary learners is used for off-chip classification, as in [18], [19], on-chip energy consumption is only 52pJ for the first layer which is $10\times$ lower than [18].

Since the proposed ANN is designed for low resolution image classification, a relevant question is how does the proposed analog implementation compare with a fully digital realization of the ANN. There are 2 pathways for digital implementation - 1) the 5×5 images are digitized by 25 ADCs and the ANN is implemented in the digital back-end on CPU/GPU; 2) ADCs and digital ANN are implemented in the front-end and are fully integrated with image sensor. Assuming the ADC input is amplified to span the full range of the ADC, Fig. 20(a) shows that a minimum of 6-bit ADC is needed for maximizing accuracy of ANN with both floating point as well as 4-bit weights. Using reported ADC energy for state-of-the-art image sensors [31], [32], energy consumed by the 25 6-bit ADCs for digitizing the 5×5 images is estimated to be 970-990pJ, which is approximately $5.5\times$ higher than energy consumed by the proposed analog ANN. The ADC energy consumption for 6-bit resolution is estimated assuming

TABLE II: Comparison with state-of-the-art ASICs

	[18] TCAS–I'17	[19] VLSI'16	**This work**		
Process (nm)	130	130	65		
Type	analog	analog	analog		
Fully Integrated?	×	×	✓		
Classifier	binary	binary	ANN		
Supply (V)	1.2	–	1.2		
Speed (MHz)	1.3	50	5		
Area (mm^2)	4.37	0.26	0.42	–	
No. of features	48	81	25	49[1]	81[1]
Accuracy (%)	90[2]	90[2]	82	88[1]	88[1]
Energy (pJ)	534[3]	630[3]	173	267[1]	403[1]
Energy/pixel (pJ)	11.1[3]	7.8[3]	6.9	5.4[1]	5[1]

[1]simulated using 4-bit ANN weights;
[2]based on Matlab simulation of ensemble adder and all-vs-all voter;
[3]only for first layer implemented on-chip

that the ADC signal-to-noise ratio is limited by thermal noise. In addition, the energy required to transmit 6-bits from 25 ADCs is going to be much higher than energy needed to transmit the 4-bit class label from the analog ANN. However, implementing the ANN in the back-end has the advantage of higher classification accuracy as shown in Fig. 20(a). On the other hand, if the ANN is implemented digitally on-chip, the maximum classification accuracy is 82.8% as shown in Fig. 20(a). Fig. 20(b) shows layout of digitally synthesized 2-layer ANN with 28 hidden neurons in 65nm CMOS. For the digital implementation, we used ideal tanh and softmax activations for the hidden and output layers respectively. The activation functions are realized in circuit by synthesizing the first four terms of their Taylor series expansions. The digitally synthesized ANN consumes 7.3nJ/classification which is $42\times$ higher than the proposed analog ANN.

IV. Conclusion

This work has demonstrated feasibility of a fully integrated, analog ANN ASIC for low resolution image classification. Integration of the proposed classifier with image sensors can lead to new generation of low-energy cameras with built-in ML capability real-time monitoring and remote sensing applications. The prototype ASIC has been validated on MNIST dataset and demonstrates a promising direction for future neural network circuits in resource constrained sensor devices.

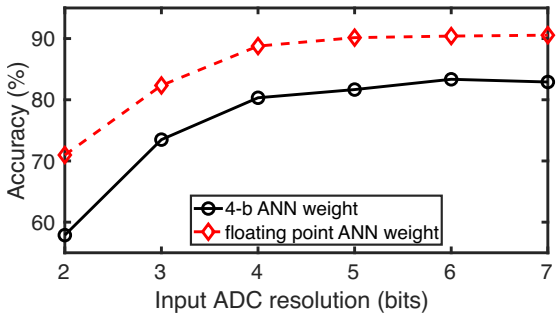
Acknowledgment

This work is supported in part by National Science Foundation grant CCF-1948331.

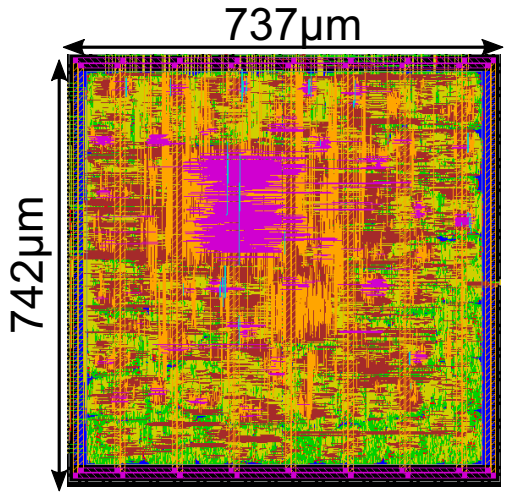
References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[2] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE International Conference on Computer Vision*, 2003, p. 726.



(a)



(b)

Fig. 20: (a) Classification accuracy for 4-bit and floating point ANN weights versus ADC resolution (b) digitally synthesized version of proposed ANN

[3] N. B. Bo, F. Deboeverie, M. Eldib, J. Guan, X. Xie, J. Niño, D. Van Haerenborgh, M. Slembrouck, S. Van de Velde, H. Steendam *et al.*, "Human mobility monitoring in very low resolution visual sensor network," *Sensors*, vol. 14, no. 11, pp. 20 800–20 824, 2014.

[4] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*. ACM, 2017, pp. 4255–4262.

[5] N. Miyazaki, K. Tsuji, M. Zheng, M. Nakashima, Y. Matsuda, and E. Segawa, "Privacy-conscious human detection using low-resolution video," in *IEEE IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 326–330.

[6] B. Bose and E. Grimson, "Improving object classification in far-field video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, 2004, pp. II–II.

[7] L. Tao, T. Volonakis, B. Tan, Y. Jing, K. Chetty, and M. Smith, "Home Activity Monitoring using Low Resolution Infrared Sensor," *arXiv preprint arXiv:1811.05416*, 2018.

[8] J. Paek, J. Hicks, S. Coe, and R. Govindan, "Image-based environmental monitoring sensor application using an embedded wireless sensor network," *Sensors*, vol. 14, no. 9, pp. 15 981–16 002, 2014.

[9] I. Cevik and S. U. Ay, "An ultra-low power energy harvesting and imaging (EHI) type CMOS APS imager with self-power capability," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 9, pp. 2177–2186, 2015.

[10] A. Y.-C. Chiou and C.-C. Hsieh, "A 137 dB dynamic range and 0.32 V self-powered CMOS imager with energy harvesting pixels," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2769–2776, 2016.

[11] J. Zarate-Roldan, A. Abuellil, O. Elsayed, F. A.-L. Hussien, A. Eladawy, E. Sánchez-Sinencio *et al.*, "0.2-nJ/b fast start-up ultralow power wireless transmitter for IoT applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 1, pp. 259–272, 2017.

[12] K. Lee, S. Park, S.-Y. Park, J. Cho, and E. Yoon, "A 272.49 pJ/pixel CMOS image sensor with embedded object detection and bio-inspired 2D optic flow generation for nano-air-vehicle navigation," in *IEEE Symposium on VLSI Circuits*, 2017, pp. C294–C295.

[13] C. Young, A. Omid-Zohoor, P. Lajevardi, and B. Murmann, "A Data-Compressive 1.5 b/2.75 b Log-Gradient QVGA Image Sensor with Multi-Scale Readout for Always-On Object Detection," in *IEEE International Solid-State Circuits Conference-(ISSCC)*, 2019, pp. 98–100.

[14] F. N. Buhler, A. E. Mendrela, Y. Lim, J. A. Fredenburg, and M. P. Flynn, "A 16-channel noise-shaping machine learning analog-digital interface," in *IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, 2016, pp. 1–2.

[15] Z. Wang, K. H. Lee, and N. Verma, "Overcoming computational errors in sensing platforms through embedded machine-learning kernels," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 8, pp. 1459–1470, 2015.

[16] Z. Wang, J. Zhang, and N. Verma, "Realizing Low-Energy Classification Systems by Implementing Matrix Multiplication Directly Within an ADC." *IEEE Trans. Biomed. Circuits and Systems*, vol. 9, no. 6, pp. 825–837, 2015.

[17] F. N. Buhler, P. Brown, J. Li, T. Chen, Z. Zhang, and M. P. Flynn, "A 3.43 TOPS/W 48.9 pJ/pixel 50.1 nJ/classification 512 analog neuron sparse coding neural network with on-chip learning and classification in 40nm CMOS," in *IEEE Symposium on VLSI Circuits*, 2017, pp. C30–C31.

[18] Z. Wang and N. Verma, "A low-energy machine-learning classifier based on clocked comparators for direct inference on analog sensors," *IEEE Transactions on Circuits and Systems–I*, vol. 64, no. 11, pp. 2954–2965, 2017.

[19] J. Zhang, Z. Wang, and N. Verma, "A machine-learning classifier implemented in a standard 6T SRAM array," in *IEEE Symposium on VLSI Circuits*, 2016, pp. 1–2.

[20] E. H. Lee and S. S. Wong, "A 2.5 GHz 7.7 TOPS/W switched-capacitor matrix multiplier with co-designed local memory in 40nm," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 418–419.

[21] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in *IEEE International Solid-State Circuits (ISSCC)*, 2018, pp. 488–490.

[22] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement," in *IEEE Symposium on VLSI Circuits*, 2018, pp. 141–142.

[23] M. Yıldız, S. Minaei, and C. Göknar, "A CMOS classifier circuit using neural networks with novel architecture," *IEEE Transaction on Neural Networks*, vol. 18, no. 6, pp. 1845–1849, 2007.

[24] B.-D. Liu, C.-Y. Chen, and J.-Y. Tsao, "A modular current-mode classifier circuit for template matching application," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 2, pp. 145–151, 2000.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] A. Jayaraj, I. Banerjee, and A. Sanyal, "Common-Source Amplifier Based Analog Artificial Neural Network Classifier," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–5.

[27] P. Getreuer, "Linear methods for image interpolation," *Image Processing On Line*, vol. 1, pp. 238–259, 2011.

[28] A. Graupner, "A methodology for the offset-simulation of comparators," *The Designer's Guide Community*, vol. 1, pp. 1–7, 2006.

[29] P. M. Figueiredo and J. C. Vital, "Kickback noise reduction techniques for CMOS latched comparators," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 7, pp. 541–545, 2006.

[30] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[31] K. Nie, W. Zha, X. Shi, J. Li, J. Xu, and J. Ma, "A Single Slope ADC With Row-Wise Noise Reduction Technique for CMOS Image Sensor," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2020.

[32] H. Le-Thai, G. Chapinal, T. Geurts, and G. G. Gielen, "A 0.18-μ CMOS Image Sensor With Phase-Delay-Counting and Oversampling Dual-Slope Integrating Column ADCs Achieving $1e^{-1}{}_{rms}$ Noise at 3.8-μs Conversion Time," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 2, pp. 515–526, 2017.

Sanjeev Tannirkulam Chandrasekaran is currently pursuing his Ph.D. in Electrical Engineering at the University at Buffalo, Buffalo, NY. Before this, he obtained his B.Tech in Electronics & Instrumentation from SASTRA University, Thanjavur, India in 2016. His research interests are geared towards developing scalable energy-efficient circuits for IoT applications with a focus on data converters and Edge-AI. He has held internship positions with Silicon Laboratories, Austin, TX, Mythic–AI, Austin, TX, and GE Global Research, Niskayuna, NY where he was involved in mixed-signal IC design. Sanjeev is the recipient of the best paper award in the 2020 IBM AI Compute Symposium, the 2019 MWSCAS Student Participation Grant, and the 2019 CICC Student Travel Grant Award. He currently serves as a reviewer for IEEE Transactions on Circuits and Systems-I: Regular Papers (TCAS–I), IEEE Transactions on Circuits and Systems-II: Express Briefs (TCAS–II) and IEEE Solid-State Circuit Letters (SSC–L).

Akshay Jayaraj received his M.S in Electrical Engineering from University at Buffalo in 2019 and B.E from Madras Institute of Technology, Anna University, Chennai, India, in 2016. He was intern with Analog/Mixed Signal Design Group at MACOM, Santa Clara, CA in summer of 2018 and analog design intern at Intel Corporation, Folsom, CA in 2019. He is currently an analog design engineer with Intel Corporation, Folsom. His current research interests include data converters, mixed signal machine learning circuits and high-speed transceivers.

Vinay Elkoori Ghantala Karnam received his bachelor's degree from JNT University Hyderabad in the year 2015 and his MS from the University at Buffalo in the year 2020. Currently, he is a graduate researcher at the Analog/Mixed-Signal VLSI group at SUNY Buffalo. He worked in the Telecom domain prior to his pursuit of MS at University at Buffalo. His research interests are in the area of analog/mixed-signal circuits, AI hardware as well as audio/music signal processing.

Imon Banerjee received her Ph.D. from The University of Genova, Italy in 2016 and her M.Tech from National Institute of Technology, Durgapur, India in 2011. She did her post-doctoral training at Stanford University.

Dr. Banerjee is an Assistant Professor in Department of Biomedical Informatics and Department of Radiology at Emory University and an Assistant Professor in Department of Biomedical Engineering at Emory University and Georgia Tech. Prior to this, she was an Instructor with joint affiliation in Radiology and Biomedical Data Science Dept. at Stanford University. Her research interests are in the areas of application of machine learning for biomedical data mining and predictive modeling. She is the recipient of 2012 Marie Curie fellowship in European FP7 Marie Curie Initial Training Networks.

Arindam Sanyal (M'14) received his Ph.D. from The University of Texas at Austin in 2016, his M.Tech from The Indian Institute of Technology, Kharagpur in 2009 and B.E from Jadavpur University, India in 2007.

Dr. Sanyal is an Assistant Professor in the Electrical Engineering Department at University at Buffalo. Prior to this, he was a Design Engineer working on low jitter PLLs at Silicon Laboratories, Austin. His research interests include analog/mixed signal design, bio-medical sensor design, analog security and on-chip artificial neural network. He is the recipient of 2020 NSF CISE Research Initiation Initiative (CRII) Award, Intel/Texas Instruments/Catalyst Foundation CICC Student Scholarship Award, 2014 and Mamraj Agarwal Award in 2001.