

# Understanding the Impact of Precision Quantization on the Accuracy and Energy of Neural Networks

Soheil Hashemi, Nicholas Anthony, Hokchhay Tann, R. Iris Bahar, Sherief Reda

School of Engineering

Brown University

Providence, Rhode Island 02912

Email: {soheil\_hashemi, nicholas\_anthony, hokchhay\_tann, iris\_bahar, sherief\_reda}@brown.edu

**Abstract**—Deep neural networks are gaining in popularity as they are used to generate state-of-the-art results for a variety of computer vision and machine learning applications. At the same time, these networks have grown in depth and complexity in order to solve harder problems. Given the limitations in power budgets dedicated to these networks, the importance of low-power, low-memory solutions has been stressed in recent years. While a large number of dedicated hardware using different precisions has recently been proposed, there exists no comprehensive study of different bit precisions and arithmetic in both inputs and network parameters. In this work, we address this issue and perform a study of different bit-precisions in neural networks (from floating-point to fixed-point, powers of two, and binary). In our evaluation, we consider and analyze the effect of precision scaling on both network accuracy and hardware metrics including memory footprint, power and energy consumption, and design area. We also investigate training-time methodologies to compensate for the reduction in accuracy due to limited bit precision and demonstrate that in most cases, precision scaling can deliver significant benefits in design metrics at the cost of very modest decreases in network accuracy. In addition, we propose that a small portion of the benefits achieved when using lower precisions can be forfeited to increase the network size and therefore the accuracy. We evaluate our experiments, using three well-recognized networks and datasets to show its generality. We investigate the trade-offs and highlight the benefits of using lower precisions in terms of energy and memory footprint.

## I. INTRODUCTION

In the recent years, deep neural networks (DNN) have provided state-of-the-art results in many different applications specifically related to computer vision and machine learning. One dominant feature of neural networks is their high demand in terms of memory and computational power thereby limiting solutions based on these networks to high power GPUs and data centers. In addition, such high demands have led to the investigation of low power ASIC accelerators where designers are free to assign dedicated resources to increase the throughput. However, memory accesses and data transfer overheads play an important part in the total computation time and energy. When using accelerators, as a solution to data transfer overheads, specialized buffers have been introduced, thereby isolating the data transfer from the computation and enabling the memory subsystem to load the new data while the computation core is processing the previously loaded data.

Neural networks show inherent resilience to small and insignificant errors within their calculations. This error toler-

ance originates from the inherent tolerance of the applications themselves, and the training nature of the networks, where some of the errors are compensated by relearning and fine tuning the parameters. In this light, techniques proposed by approximate computing, such as approximate arithmetic, are an attractive option to lower the power consumption and design complexity in neural networks accelerators. However, as demonstrated by Chen *et al.* [2] and Tann *et al.* [21], the dominant portion of power and energy of hardware neural network accelerators is consumed in the memory subsystem, limiting the scope of arithmetic approximation. In this light, one particularly effective solution is reducing the bit-width required to represent the data.

While many accelerators have been proposed using different bit-precisions, most of these studies have been ad-hoc and give little to no explanation for choosing the specific precision. In particular, an evaluation of different precisions on the performance of the networks, considering both hardware metrics and inference accuracy, is not available. Such a study would provide researchers with better guidance as to the trade-offs available by such networks. In this paper we aim to address this issue by providing a quantitative analysis of different precisions and available trade-offs. More specifically, our paper makes the following contributions:

- We perform a detailed evaluation of a broad range of networks precisions, from binary weights to single precision floating-points, as well as several points in between.
- We utilize learning techniques to improve the lost accuracy by taking advantage of the training process to increase the accuracy.
- We evaluate our designs for both accuracy and hardware specific metrics, such as design area, power consumption, and delay, and demonstrate the results on a Pareto Frontier, enabling better evaluation of the available trade-offs.
- Exploiting the benefits of lower precisions, we propose increasing the network size to compensate for accuracy degradation. Our results showcase low precision networks capable of achieving equivalent accuracy compared to smaller floating-point networks while offering significant improvements in energy consumption and design area.

The rest of the paper is organized as follows. In Section II, we briefly summarize the basics of neural networks, and in

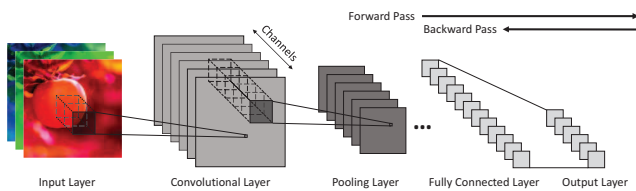


Fig. 1. The structure of a typical deep neural networks.

Section III we review related work. Then, in Section IV we describe various precisions and network training techniques used in our evaluations and argue for increasing network size to recoup accuracy loss in lower precision networks. The results from our evaluations are provided in Section V and in Section VI we summarize our finding and provide suggestions for future work.

## II. BACKGROUND

Deep neural networks are organized in layers where each layer is only connected to the layers immediately before and after it. Each layer gets its input from the previous layer and feeds it to the next layer after some layer-specific processing. Figure 1 shows the general structure and connectivity of the layers. As show in the figure, each layer consists of several channels. Deep Neural networks, in general, consist of a combination of three main layer types: convolutional layers, pooling layers, and fully connected layers.

In typical neural networks the dominant portion of the computation is performed in the convolution layers and fully connected layers, while pooling layers simply down-sample the data. More specifically, channels in convolutional and fully connected layers are comprised of neuron units where each neuron performs a weighted sum of its inputs before feeding the result to a nonlinearity function. The intermediate values between layers are called feature maps, as they each abstract some structure in the input image. From a data perspective, neural networks operate on two main set of parameters: input data and intermediate feature maps, and network parameters (or weights). Since inputs and feature maps are treated similarly by the network, similar precisions are used for their representation. However, numerical precision of the network parameters can be changed independently of the input precision.

While the input data is assumed to be given for each network, the flexibility of neural networks arises from their ability to adapt their response to a specific input by training the network parameters. More specifically, use of neural networks comprises two phases, a training process during which the network parameters are learned, and a test phase which performs the inference and classification of the test data. In the training phase, neural networks usually utilize a backpropagation algorithm during which the classification error is propagated backwards using partial gradients. Network parameters are then updated using stochastic gradient descent. After training and in the test phase, the learned network is utilized in the forward phase to classify the test data. As

discussed later, the main complexity of using lower precision in these networks arises due to the learning process.

## III. PREVIOUS WORK

The high demand of DNNs, in terms of complexity and energy consumption, has shifted attention to low-power accelerators. Many works have proposed implementing neural networks on FPGAs [7], [6], or as an ASIC accelerator [11], [22]. In all these works, different precisions have been utilized with little or no justification for the chosen bit-width.

Chen *et al.* proposed Eyeriss, a spatial architecture along with a dataflow aimed at minimizing the movement energy overhead using data reuse [3]. For their implementation, a 16-bit fixed-point precision is utilized. Sankaradas *et al.* empirically determine an acceptable precision for their application [18] and reduce the precision to 16-bit fixed-point for inputs and intermediate values while maintaining 20-bit precision for weights. A FPGA-based accelerator is proposed by Zhang *et al.*, where single precision floating-point arithmetic has been utilized [24]. While this work offers a brief comparison between resources required for floating-point and fixed-point arithmetic logic in FPGAs, no discussion of accuracy is provided. Chakradhar *et al.* propose a configurable co-processor where input and output values are represented using 16 bits while intermediate values use 48 bits [1].

Many works have successfully integrated techniques commonly used in approximate computing to lower the computation and energy demands of neural networks. A feed-forward neural network is proposed by Kung *et al.*, where approximations are introduced to lower-impact synapses [13]. Venkataramani *et al.* propose an approximate design where error-resilient neurons are replaced with lower-precision neurons and an incremental training process is used to compensate for some of the added error [23]. However, no specifications for the bit precision range used in the experiments are provided. Tann *et al.* propose an incremental training process during which most of the network can be turned off to save power [21]. The neurons are then turned on during run-time if deemed necessary for correct classification. In this work, 32-bit floating-point representation was used.

While use of limited precision in neural networks has been proposed before [16], [4], [17], there exists no comprehensive exploration of their effect on energy consumption and computation time in reference to network accuracy. A recent publication by Gysel *et al.* provides an analysis of precision on network accuracy; however, the design parameters are not evaluated [9]. Our objective is to precisely quantify the effect of each numerical precision or quantization on all aspects of the networks focusing specially on hardware metrics.

## IV. METHODOLOGY

Here, first in Section IV-A, we discuss the range of precisions and quantizations considered in our evaluation. We also briefly discuss the network training techniques used to minimize the accuracy degradation due to the limited precision. Finally, in Section IV-B we propose two expanded network architectures to compensate for the accuracy drop.

### A. Evaluated Precisions and Train-Time Techniques

We consider a broad range of numerical precisions and quantizations, from 32-bit floating-point arithmetic to binary nets, as well as several precision points in between. We summarize them below:

1) *Floating-Point Arithmetic*: This is the most commonly used precision as it generates the state-of-the-art results in accuracy. However, floating-point arithmetic requires complicated circuitry for the computational logic such as adders and multipliers as well as large bit-width, necessitating ample memory usage. As a result, this precision is not suitable for low-power and embedded devices.

2) *Fixed-Point Arithmetic*: Fix-point arithmetic is less computationally demanding as it simplifies the logic by fixing the location of the radix point. This arithmetic also provides the flexibility of a wide range of accuracy-power trade-offs by changing the number of bits used in the representation. In this work, we evaluate 4-, 8-, 16- and 32-bit precisions. To improve accuracy, we allow a different radix point location between data and parameters [9]. However, we refrain from evaluating bit precisions that are not powers of 2 since they result in inefficient memory usage that might nullify the benefits.

3) *Power-of-Two Quantization*: Multipliers are the most demanding computational unit for neural networks. As proposed by Lin [16], limiting the weights to be in the form of  $2^i$ , enables the network to replace expensive, frequent, and power-hungry multiplications with much smaller and less complex shifts. In our evaluations, we consider power of two quantization of the weights while representing the inputs with 16-bit fixed-point arithmetic.

4) *Binary Representation*: Recent work suggests that neural networks can generate acceptable results using just 1-bit weight representation [5]. While work by Courbariaux suggests binarizing activation between network layers, it does not binarize the input layer [4]. For this reason, our accelerator would still need to support multi-bit inputs. Thus, we evaluate the binary net using one bit for weights, while using 16-bit fixed-point representation for the inputs and feature maps.

*Hardware Accelerator*: For our experiments, we adopt a tile-based hardware accelerator similar to DianNao [2]. We implement 16 neuron processing units each with 16 synapses. Figure 2 shows our hardware implementation. As illustrated in the figure, three separate memory subsystems are used to store the intermediate values and outputs and buffer the inputs and weights. These subsystems are comprised of an SRAM buffer array, a DMA, and control logic responsible for ensuring that the data is loaded into the buffers and made available to the neural functional unit (NFU) at the appropriate clock cycle without additional latency. The NFU pipelines the computation into three stages, weight blocks (WB), adder tree, and non-linearity function. As shown in Figure 2, the weight blocks will be modified to accommodate for different precisions and quantizations as needed. In the case of binary precision, we merge the first two pipeline stages, effectively leading to a two stage NFU, in order to reduce the runtime. Furthermore, the

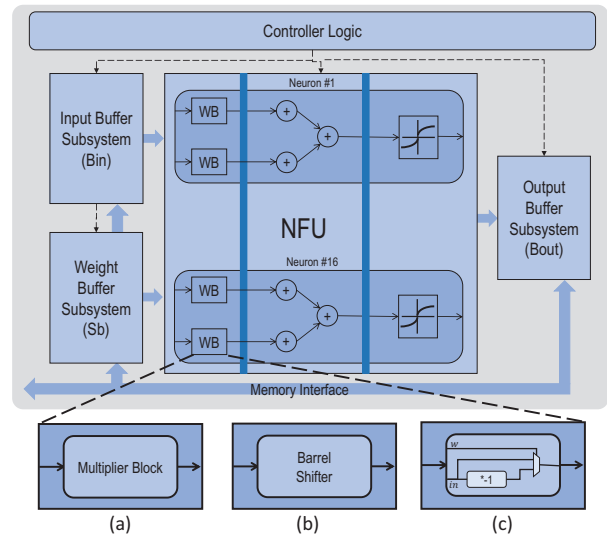


Fig. 2. The hardware model used for our experiments. The first stage (WB) has different variants for (a) floating-point and fixed-point arithmetic, (b) powers of two quantization, and (c) binary network.

size of all buffers and the control logic are modified according to the precision.

*Training Time Techniques*: We include a training phase in our experiments to enable the network to determine appropriate weights and adapt to the lower precision. Training processes, in nature, require high precision in order to converge to a good minima as the increments made to the parameters can be extremely small. On the other hand, if the network is made aware of its inference restrictions (in our case, the limited precision), the training process can potentially compensate for some of the errors by fine-tuning the parameters and therefore improve the accuracy at no extra cost.

While the effects of reduced precision are analytically complicated to formulate as part of the training process [8], intuitive techniques can be utilized to improve the test phase accuracy. One approach proposed in [21] is to utilize a set of full precision weights, trained independently, as the starting point of a re-training process, in which the weights and inputs are restricted to the specified precision. This approach assumes that by using lower precisions, close to optimal performance can be obtained if a local search is performed around the optimal set of parameters as learned with full precision.

A second approach for improving the accuracy is to utilize weights with different precisions in different parts of the training process, as proposed by Courbariaux *et al.* [5]. They solve the zero-gradient issue by keeping two sets of weights: one in full precision and one in the selected lower precision. The network is then trained using the full precision values during backward propagation and parameter updates, while approximating and using low precision values for forward passes. This approach allows for the accumulation of small gradient updates to eventually cause incremental updates in the lower precision.

In our approach, we train all of the low precision networks using a combination of the first and second approaches. We

TABLE I  
BENCHMARK NETWORKS ARCHITECTURE DESCRIPTIONS.

| MNIST            | SVHN            | CIFAR-10        |
|------------------|-----------------|-----------------|
| LeNet [14]       | ConvNet [19]    | ALEX [12]       |
| 28×28×1          | 32×32×3         | 32×32×3         |
| conv 5×5×20      | conv 5×5×16     | conv 5×5×32     |
| maxpool 2×2      | maxpool 2×2     | maxpool 3×3     |
| conv 5×5×50      | conv 7×7×512    | conv 5×5×32     |
| maxpool 2×2      | maxpool 2×2     | avgpool 3×3     |
| innerproduct 500 | innerproduct 20 | conv 5×5×64     |
| innerproduct 10  | innerproduct 10 | avgpool 3×3     |
|                  |                 | innerproduct 10 |

TABLE II  
ALEX LARGER NETWORK ARCHITECTURE DESCRIPTIONS.

| CIFAR-10        |                  |
|-----------------|------------------|
| ALEX+           | ALEX++           |
| 32×32×3         | 32×32×3          |
| conv 5×5×64     | conv 3×3×64      |
| maxpool 3×3     | maxpool 2×2      |
| conv 5×5×64     | conv 3×3×128     |
| avgpool 3×3     | maxpool 2×2      |
| conv 5×5×128    | conv 3×3×256     |
| avgpool 3×3     | maxpool 2×2      |
| innerproduct 10 | innerproduct 512 |
|                 | innerproduct 10  |

initialize the parameters for lower precision training from the floating point counterpart. Once initialized, we train by keeping two sets of weights.

### B. Expanded Network Architectures

While significant savings in power, area, and computation time can be achieved using lower precisions, even a small degradation in accuracy can prohibit their use in many applications. However, we observe that, due to the nature of neural networks, the benefits obtainable by using lower precisions are disproportionately larger than the resulting accuracy degradation. This opens a new and intriguing dimension, where the accuracy can be boosted by increasing the number of computations while still consuming less energy. We therefore propose increasing the number of operations by increasing network size, as needed to maintain accuracy while spending significantly less for each operation.

In this light, in Section V, we showcase two significantly larger networks and demonstrate that even by significantly increasing the size of the network, low precision can still result in improvements in energy consumption while eliminating the accuracy degradation. We discuss the specifications of the two larger networks in Section V.

## V. EXPERIMENTAL RESULTS

### A. Experimental Setup

We evaluate our designs both in terms of accuracy and design metrics (i.e., power, energy, memory requirements, design area). To measure accuracy, we adopt Ristretto [9], a Caffe-based framework [10] extended to simulate fixed-point operation. We modify Ristretto to accommodate our techniques, as needed. In different experiments, we ensure that all design parameters except for the bit precision are the same. This is critical to ensure the isolation of the effects of bit precision from any other factor.

We compile our designs using Synopsys Design Compiler using a 65 nm industry strength technology node library. We use a 250 MHz clock frequency and synthesize in nominal processing corner. We design our accelerator to have a zero timing slack for the full-precision accurate design. We confirm the functionality of our hardware implementation with extensive simulations. As before, we ensure that all other network parameters, including the frequency, are kept constant across different precision experiments.

*Benchmarks:* We consider three well-recognized neural network architectures utilized with three different datasets, MNIST [15] using the LeNet [14] architecture, SVHN using CONVnet [19], and CIFAR-10 [12] using the network described by Alex Krizhevsky [12] (Here we refer to this network as ALEX). For all cases, we randomly select 10% of each classification category from the original test set as our validation set. To showcase the benefits from increasing the network size while using lower precision, we evaluate two networks as summarized in Table II. Here, we focus on CIFAR-10 since MNIST and SVHN do not provide a large range in accuracy differences between various precisions and quantizations. As summarized in Table II, we evaluate two larger variations of the ALEX network: (1) ALEX+, where the number of channels in each convolutional layer is doubled, and (2) ALEX++, where the number of channels is doubled when the feature size is halved [20]. As shown in Section V-B, this methodology results in significant improvements in accuracy while still delivering significant savings in energy.

### B. Results

Figure 3 shows the breakdown of power and area for the accelerator in the cases investigated. Values shown as  $(w, in)$  represent the number of bits required for representing weight and input values, respectively. Note, that these graphs do not reflect the power consumption of the main memory. As shown in the figure, the majority of the resources, both in power and design area, are utilized in the memory buffers necessary for seamless operation of the computational logic. To be more specific, in our experiments, the buffers consume between 75%-93% of the total accelerator power, while using 76%-96% of the total design area. These values highlight the necessity of approximation approaches targeting the memory footprint.

Table III summarizes the design metrics of the accelerator for each of the numerical precisions considered. In order to maintain a fair comparison, we keep all the other parameters, such as the frequency, number of hardware neurons, etc., constant among different precisions. Changing the frequency or the accelerator parameters (other than precision) adds another dimension to the design space exploration which is out of the scope of our work.

We evaluate the accuracy of the networks, as well as energy requirements for processing each image for each of our benchmarks. Table IV summarizes the results for MNIST

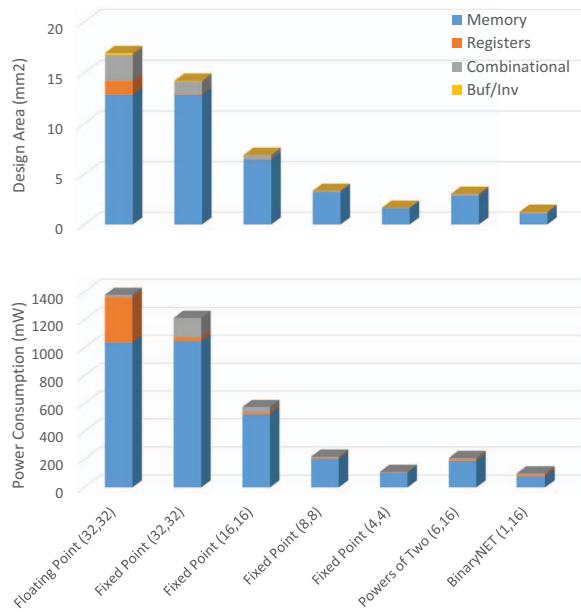


Fig. 3. The breakdown of design area and power consumption using different precisions.

TABLE III  
DESIGN METRICS OF THE EVALUATED NUMERICAL PRECISIONS AND QUANTIZATIONS.

| Precision ( $w, in$ )  | Design Area ( $mm^2$ ) | Power Cons. ( $mW$ ) | Area Saving (%) | Power Saving (%) |
|------------------------|------------------------|----------------------|-----------------|------------------|
| Floating-Point (32,32) | 16.74                  | 1379.60              | 0               | 0                |
| Fixed-Point (32,32)    | 14.13                  | 1213.40              | 15.56           | 12.05            |
| Fixed-Point (16,16)    | 6.88                   | 574.75               | 58.92           | 58.34            |
| Fixed-Point (8,8)      | 3.36                   | 219.87               | 79.94           | 84.06            |
| Fixed-Point (4,4)      | 1.66                   | 111.17               | 90.07           | 91.94            |
| Powers of Two (6,16)   | 3.05                   | 209.91               | 81.78           | 84.78            |
| Binary Net (1,16)      | 1.21                   | 95.36                | 92.73           | 93.08            |

and SVHN datasets. We were able to achieve little to no accuracy drop for all but one of the network precisions in the MNIST classification. In the case of SVHN, however, while keeping the network architecture constant, the 4-bit fixed-point and binary representations failed to converge. For SVHN dataset, for instance in the case of powers of two network, we are able to achieve more than 84% energy saving with an accuracy drop of approximately 2%. Note that as we keep the frequency constant the processing time per image changes very marginally among different precisions. Additional runtime savings can be achieved by increasing the frequency or changing the accelerator specification which is not explored in this work.

The reduction in precision also reduced the required memory capacity for network parameters, as well as the input data. We quantify our memory requirements for all the network architectures using different bit precisions. In our experiments, for the full-precision design, network parameters require approximately 1650KB, and 2150KB, and 350KB of memory for LeNet, CONVnet, and ALEX, respectively. Since there is a direct correlation between bit precision and network memory requirements, the memory footprint of each network reduces

TABLE IV  
THE ACCURACY, PER IMAGE INFERENCE ENERGY, AND THE ENERGY SAVINGS ACHIEVABLE USING EACH OF THE EVALUATED PRECISIONS. FOR EACH DATASET, ENERGY SAVINGS ARE IN REFERENCE TO THE FULL-PRECISION IMPLEMENTATION.

| Precision ( $w, in$ )  | MNIST           |                    |                 | SVHN            |                    |                 |
|------------------------|-----------------|--------------------|-----------------|-----------------|--------------------|-----------------|
|                        | Class. Acc. (%) | Energy ( $\mu J$ ) | Energy Sav. (%) | Class. Acc. (%) | Energy ( $\mu J$ ) | Energy Sav. (%) |
| Floating-Point (32,32) | 99.20           | 60.74              | 0               | 86.77           | 754.18             | 0               |
| Fixed-Point (32,32)    | 99.22           | 52.93              | 12.86           | 86.78           | 663.01             | 12.09           |
| Fixed-Point (16,16)    | 99.21           | 24.60              | 59.50           | 86.77           | 314.05             | 58.36           |
| Fixed-Point (8,8)      | 99.22           | 8.86               | 85.41           | 84.03           | 120.14             | 84.07           |
| Fixed-Point (4,4)      | 95.76           | 4.31               | 92.90           | NA              | NA                 | NA              |
| Powers of Two (6,16)   | 99.14           | 8.42               | 86.13           | 84.85           | 114.70             | 84.79           |
| Binary Net (1,16)      | 99.40           | 3.56               | 94.13           | 19.57           | 52.11              | 93.09           |

from  $2\times$  to  $32\times$  for different bit precisions. Note, we do not utilize any of recent parameter encoding and compression techniques, and such techniques are orthogonal to our work.

As discussed in Section IV-B, we propose that a portion of the benefits from using low precision arithmetic can be exploited to boost the accuracy to match that of the floating point network while spending some portion of the energy savings by increasing the network size. Here, we showcase the benefits from our proposed methodology on CIFAR-10 dataset. The summary of the performances for the ALEX as well as the two larger networks (ALEX+ and ALEX++) is provided in Table V. Here, we do not report the results for fixed-point (32,32) for ALEX+ and ALEX++ as its energy saving is not competitive compared to other precisions. Also, the fixed-point (4,4) fails to converge for all three networks on CIFAR-10 and the respective rows have been removed from the table. Furthermore, we find that the accuracy for fixed-point++ (8,8) is lower in comparison to the other networks with the same precision. We observe that for this network, there is a significant difference in the range of parameter and feature map values and as a result, 8 bits fails to capture the necessary range of the numbers.

As shown in the table, lower precision networks can outperform the baseline design in accuracy while still delivering savings in terms of energy. The parameter memory requirements for the full-precision networks are roughly 350KB, 1250KB, and 9400KB for ALEX, ALEX+, and ALEX++ respectively. As discussed previously, the memory footprint reduces linearly with parameter precision when reducing the precision.

The available trade-offs in terms of accuracy and energy using different precisions and expanded networks are plotted in Figure 4 for the CIFAR-10 testbench. The figure highlights the previous argument that a wide range of power and energy savings are possible using different precisions while maintaining acceptable accuracy. Further, when operating in low precision/quantization, a portion of the obtained energy benefits can be re-appropriated to recoup the lost accuracy by increasing the network size. As shown in the Figure 4, this methodology can eliminate the accuracy drop (for example in the case of Power of Two++ (6,16)) while still delivering energy savings of 35.93%. The figure highlights that larger networks with lower precision can dominate the full-precision baseline design in both accuracy and energy requirements.

TABLE V

NETWORK PERFORMANCE FOR DIFFERENT PRECISION ON CIFAR-10 DATASET AND USING ALEX, ALEX+, AND ALEX++. ENERGY SAVINGS ARE IN REFERENCE TO THE ALEX FULL-PRECISION IMPLEMENTATION.

| Precision ( $w, in$ )  | CIFAR-10        |                    |                 |
|------------------------|-----------------|--------------------|-----------------|
|                        | Class. Acc. (%) | Energy ( $\mu J$ ) | Energy Sav. (%) |
| Floating-Point (32,32) | 81.22           | 335.68             | 0               |
| Fixed-Point (32,32)    | 79.71           | 293.90             | 12.45           |
| Fixed-Point (16,16)    | 79.77           | 136.61             | 59.30           |
| Fixed-Point+ (16,16)   | 81.86           | 491.32             | 1.5× More       |
| Fixed-Point++ (16,16)  | 82.26           | 628.17             | 1.9× More       |
| Fixed-Point (8,8)      | 77.99           | 49.22              | 85.34           |
| Fixed-Point+ (8,8)     | 78.71           | 177.02             | 47.27           |
| Fixed-Point++ (8,8)    | 75.03           | 226.32             | 32.59           |
| Powers of Two (6,16)   | 77.03           | 46.77              | 86.07           |
| Powers of Two+ (6,16)  | 77.34           | 168.21             | 49.89           |
| Powers of Two++ (6,16) | 81.26           | 215.05             | 35.93           |
| Binary Net (1,16)      | 74.84           | 19.79              | 94.10           |
| Binary Net+ (1,16)     | 77.91           | 71.18              | 78.80           |
| Binary Net++ (1,16)    | 80.52           | 91.00              | 72.89           |

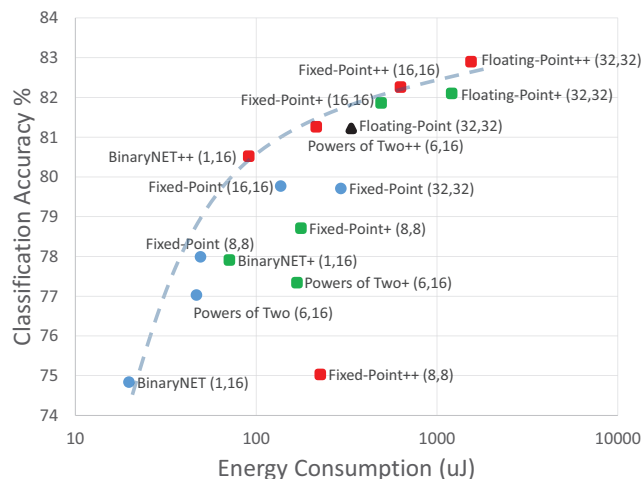


Fig. 4. The Pareto Frontier plot of the evaluated design point for CIFAR-10 testcase. The x axis is plotted in logarithmic scale to cover the energy range of all the designs. Here, the black point indicates the initial full-precision design, the blue points indicate the lower precision points, while the red and green points show the results from the larger networks.

## VI. CONCLUSION

In this work, we perform an analysis of numerical precisions and quantizations in neural networks. We evaluate a broad range of numerical approximations in terms of accuracy, as well as design metrics such as area, power consumption, and energy requirements. We study floating-point arithmetic, different precisions of fixed-point arithmetic, quantizations of weights to be of powers of two, and finally binary nets where the weights are limited to one bit values. We also show that lower-precision, larger networks can be utilized which outperform the smaller full-precision counterparts in both energy and accuracy. For future work, we plan on analytically investigating the correlations between network and datasets and their behavior in lower precision thereby effectively predicting the lower precision accuracy and hardware metrics. Further, we plan to develop architectures which support multiple radix point locations between layers. As discussed in V-B, this feature may reduce the accuracy degradation significantly for lower precision networks.

## ACKNOWLEDGMENT

This work is supported by NSF grant 1420864 and by NVIDIA Corporation for their generous GPU donation. We also thank Professor Pedro Felzenszwalb for his helpful inputs.

## REFERENCES

- [1] S. Chakradhar, M. Sankaradas, V. Jakkula, and S. Cadambi. A dynamically configurable coprocessor for convolutional neural networks. In *ISCA*, pages 247–257, 2010.
- [2] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In *ASPLOS*, pages 269–284, 2014.
- [3] Y. H. Chen, J. Emer, and V. Sze. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. In *ISCA*, pages 367–379, 2016.
- [4] M. Courbariaux and Y. Bengio. Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1. 2016.
- [5] M. Courbariaux, Y. Bengio, and J. David. Binaryconnect: Training deep neural networks with binary weights during propagations. *CoRR*, abs/1511.00363, 2015.
- [6] C. Farabet, C. Poulet, and Y. LeCun. An fpga-based stream processor for embedded real-time vision with convolutional networks. In *ICCV Workshops*, pages 878–885, 2009.
- [7] V. Gokhale, J. Jin, A. Dundar, B. Martini, and E. Culurciello. A 240 g-ops/s mobile coprocessor for deep neural networks. In *CVPRW*, pages 696–701, 2014.
- [8] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep learning with limited numerical precision. *CoRR*, abs/1502.02551, 2015.
- [9] P. Gysel. Ristretto: Hardware-oriented approximation of convolutional neural networks. *CoRR*, abs/1605.06402, 2016.
- [10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [11] J. Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H. J. Yoo. A 201.4 gops 496 mw real-time multi-object recognition processor with bio-inspired neural perception engine. *IEEE Journal of Solid-State Circuits*, 45(1):32–45, 2010.
- [12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009.
- [13] J. Kung, D. Kim, and S. Mukhopadhyay. A power-aware digital feedforward neural network platform with backpropagation driven approximate synapses. In *ISLPED*, pages 85–90, 2015.
- [14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.
- [15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.
- [16] Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio. Neural networks with few multiplications. *CoRR*, abs/1510.03009, 2015.
- [17] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. *CoRR*, abs/1603.05279, 2016.
- [18] M. Sankaradas, V. Jakkula, S. Cadambi, S. Chakradhar, I. Durdanovic, E. Cosatto, and H. P. Graf. A massively parallel coprocessor for convolutional neural networks. In *ASAP*, pages 53–60, 2009.
- [19] P. Sermanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit classification. In *ICPR*, pages 3288–3291, 2012.
- [20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, abs/1607.05418, 2015.
- [21] H. Tann, S. Hashemi, R. I. Bahar, and S. Reda. Runtime configurable deep neural networks for energy-accuracy trade-off. In *CODES+ISSS*, pages 1–10, 2016.
- [22] O. Temam. A defect-tolerant accelerator for emerging high-performance applications. In *ISCA*, pages 356–367, 2012.
- [23] S. Venkataramani, A. Ranjan, K. Roy, and A. Raghunathan. Axnn: Energy-efficient neuromorphic systems using approximate computing. In *ISLPED*, pages 27–32, 2014.
- [24] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong. Optimizing fpga-based accelerator design for deep convolutional neural networks. In *FPGA*, pages 161–170, 2015.