Activation Density based Mixed-Precision Quantization for Energy Efficient Neural Networks

Karina Vasquez*,[1], Yeshwanth Venkatesha*,[2], Abhiroop Bhattacharjee[2], Abhishek Moitra[2], and Priyadarshini Panda[2]
[1]Department of Electrical Engineering, UTEC, Peru
[2]Department of Electrical Engineering, Yale University, USA
karina.vasquez@utec.edu.pe, {yeshwanth.venkatesha, abhiroop.bhattacharjee, abhishek.moitra, priya.panda}@yale.edu

***Abstract*—As neural networks gain widespread adoption in embedded devices, there is a growing need for model compression techniques to facilitate seamless deployment in resource-constrained environments. Quantization is one of the go-to methods yielding state-of-the-art model compression. Most quantization approaches take a fully trained model, then apply different heuristics to determine the optimal bit-precision for different layers of the network, and finally retrain the network to regain any drop in accuracy. Based on Activation Density—the proportion of non-zero activations in a layer—we propose a novel in-training quantization method. Our method calculates optimal bit-width/precision for each layer during training yielding an energy-efficient mixed precision model with competitive accuracy. Since we train lower precision models progressively during training, our approach yields the final quantized model at lower training complexity and also eliminates the need for re-training. We run experiments on benchmark datasets like CIFAR-10, CIFAR-100, TinyImagenet on VGG19/ResNet18 architectures and report the accuracy and energy estimates for the same. We achieve up to $4.5\times$ benefit in terms of estimated multiply-and-accumulate (MAC) reduction while reducing the training complexity by 50% in our experiments. To further evaluate the energy benefits of our proposed method, we develop a mixed-precision scalable Process In Memory (PIM) hardware accelerator platform. The hardware platform incorporates shift-add functionality for handling multi-bit precision neural network models. Evaluating the quantized models obtained with our proposed method on the PIM platform yields about $5\times$ energy reduction compared to baseline 16-bit models. Additionally, we find that integrating activation density based quantization with activation density based pruning (both conducted during training) yields up to $\sim 198\times$ and $\sim 44\times$ energy reductions for VGG19 and ResNet18 architectures respectively on PIM platform compared to baseline 16-bit precision, unpruned models.**

***Index Terms*—neural networks, quantization, activation density, process in-memory**

I. Introduction

Neural network model compression has gained popularity in the past few years as deep learning is being deployed in numerous applications across resource-constrained embedded devices [1]. As neural networks are known to have extremely high computation and memory requirements, there is a need for efficient processing algorithms in order to deploy them in energy constrained environments. *Pruning*—removing the redundant parameters—and *quantization*—using a lower bit representation for the model parameters—are the two most popular techniques used to compress a model and significantly reduce the energy consumption [2]–[4].

The idea of quantized neural networks has been explored as early as the 1990s mainly to enable simpler hardware implementation [5]–[7]. The complexity of quantization methods range from simple rounding [8] to using deep reinforcement learning agents to optimize design decisions [9]. In addition to quantizing weights, quantizing activations has also been found to be useful. The authors in [10] report that activations consume more memory than weights. Furthermore, gradients can also be quantized which enables communication efficient training in a distributed learning system such as federated learning [11], [12]. Similar to quantization, network pruning also was introduced in the 1990s [13], [14] and has garnered attention in the recent times due to the rapid deployment of neural networks on embedded platforms [3].

A majority of pruning and quantization methods require a fully trained model as the starting point. Then, the model is pruned/quantized based on some heuristics followed by retraining to regain the drop in accuracy [2]. Often this process of pruning/quantizing and retraining is repeated iteratively to obtain the final model. Note, only those quantization methods that are geared towards yielding mixed-precision networks incur such iterative process. Binarized or homogeneous precision network implementations obtained by training models with same bit-width across all layers from scratch has been shown [15]. However, such models generally suffer from accuracy loss as compared to mixed-precision models. But, in the latter, the prerequisite of a large fully trained network as a starting point is a significant overhead in the overall training-compression-retraining process.

Generally, having a pre-trained model helps in assigning importance to the weights [16]. However, recent evidence suggests that there is very little effect of having a fully trained model on the end-accuracy of the final pruned/quantized model. The authors in [17] showed that training a sparse model from randomly initialized weights resulted in similarly performing models as a baseline unpruned model. Leveraging this, we propose an in-training quantization method based on a novel activation density metric that yields a mixed-precision network and eliminates the need for a fully pre-trained model.

To demonstrate the energy and compute efficiency improvement of our proposed method, we design a precision-scalable

* These authors have contributed equally to this work. This work was done while Karina was interning at Yale University.

978-3-9819263-5-4/DATE21/©2021 EDAA

Process In-Memory (PIM) hardware platform that can support variable data-precision and pruning for different layers. Implementation of multibit-precision functionality is relatively more viable in case of a PIM architecture since changing the bit-precisions does not lead to significant architectural changes as opposed to a conventional digital CMOS architecture.

To cater to higher scalability and realistic mixed-precision implementations, we design our architecture to support only 2-/4-/8-/16-bit precisions. Thus, data precision of 3-bits would be translated to 4-bits, 5-bits to 8-bits, and so on. On the other hand, analytical methods based on a simple digital block capable of handling variable-widths consider ideal scenarios with specific bit-precision for MAC/memory for each layer which is not only impractical in real hardware but also is difficult to scale across different model architectures. Using this hardware platform, we perform a realistic energy cost analysis and show that our proposed method performs much better when compared to baseline models- with homogeneous precision for all layers.

Additionally, we also perform an analytical energy estimation based on multiply-and-accumulate (MAC) and simple memory access calculations similar to that of many existing works in the literature [2], [18], [19]. We do this to highlight the fact that such kind of analysis assumes impractical hardware architecture design scenarios which tend to overestimate the efficiency improvements, especially for mixed-precision networks. In this paper we make the following contributions:

- We propose a novel in-training method to find optimal quantization bit-width for each layer based on *Activation Density (AD)* without significant reduction in accuracy.
- We empirically evaluate our approach on CIFAR-10, CIFAR-100 and TinyImagenet datasets.
- As quantization is orthogonal to pruning, we also evaluate the performance of our algorithm when used simultaneously with Activation Density based pruning method.
- We design a PIM accelerator with in-built support for multi-bit data precision scalability, and perform realistic energy analysis comparison between quantized, quantized-and-pruned and baseline architectures.
- We show that our PIM hardware based energy analysis gives more realistic energy estimates than analytical energy analysis which often relies on impractical hardware architecture designs for mixed-precision models.

II. Background

A. Quantization

Quantization is the process of using a lower bit representation of activations and/or weights to achieve efficient computation. A 32-bit floating-point arithmetic (FP32) is the default in most of the modern deep learning implementations. However, numerous techniques have demonstrated that a similar accuracy can be achieved with much lower bit-widths such as 4-/2-/1-bit [15]. Moreover, integer arithmetic is more efficient than floating point arithmetic lowering the energy budget significantly. In the cases of extreme quantization where there is 1-bit representation, the integer arithmetic can be further reduced to bit-wise XNOR operations [20].

Typical k-bit quantization is given by-

$$x_q = round((x - x_{min})(\frac{2^k - 1}{x_{max} - x_{min}})) \tag{1}$$

where, x denote original values and x_q represent the corresponding quantized values. x can be floating point representation or quantized representation with $\geq k$ quantization levels.

B. Pruning

Neural networks are known to have ample redundancies. The objective of pruning is to remove redundant connections in the neural network to make the model sparse. Pruning helps in reducing the model size that in turn reduces the storage and communication requirements for the models which is particularly useful in embedded devices for distributed intelligence. Pruning coupled with accelerators that can efficiently compute sparse matrix multiplications yield significant compute and energy efficiency [2], [21], [22].

C. Activation Density

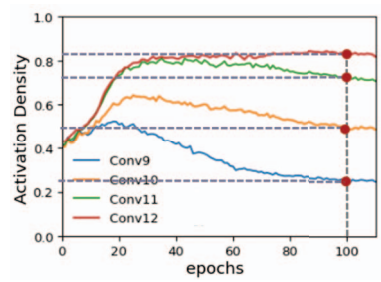

Fig. 1. Trend of Activation Density (AD) of a few individual layers. Here, we observe that AD stabilizes at around 100 epochs. We use this observation to formulate a quantization method based on AD (see Algorithm 1).

In this work, we use a metric called Activation Density (AD) [23] to determine the optimal bit-precision per layer. AD is defined as the proportion of non-zero activations in a layer calculated by passing the training set through the network-

$$AD = \frac{\#nonzero \;\; activations}{\#total \;\; activations} \tag{2}$$

Intuitively, AD quantifies the utilization of parameters in a network. For instance, for a layer with 512 neurons and 100 neurons yielding non-zero output, AD will be $100/512 = 0.195$. AD can also be calculated for the entire network by accumulating the statistics of all the layers. A noteworthy observation that forms the basis of our work is that the AD of the network saturates (to a value < 1) as the training progresses (see Fig. 1). This implies that a majority of activations are rendered inactive during training, hinting that there is a huge redundancy in the model. We take advantage of this fact to quantize each layer of a network to lower bit-precisions based on AD. Thus, quantizing based on AD can be interpreted as removing the redundancies of a model by reducing the precision. Note, most deep learning networks today use rectified

linear unit (ReLU) activation functions which will generate a good proportion of zero/non-zero activations. Hence, AD can be used as a universal metric for quantifying redundancy.

III. Methodology

Given a model M with layers $l = 1, 2, ...n$, our approach aims to optimize quantization bit-width k_l for each layer l based on the AD metric. Given A_l as activation output from layer l, the quantized activation A_{lq}, is calculated according to eqn. 1. Similarly, during backpropagation after updating the weights W, the quantized weights W_q are calculated as per eqn. 1. The model is trained using standard backpropagation algorithm. Note, depending upon the quantization bit-width per layer, we use quantized activations and weights A_q, W_q during forward propagation. During backpropagation, the gradients are calculated and the weights W_q are updated. The updated weights are again quantized before the next training step. During training, we monitor the activation density AD_l for all the layers. Once AD_l stabilizes across all layers, we break the training process and then, perform quantization. The quantization bit-width k_l for each layer l is calculated as-

$$k_l = round(k_{l_{initial}} * AD_l) \tag{3}$$

For instance, in Fig. 1, we observe that AD is nearly constant after 100 epochs. In such case, we stop training at the 100th epoch, quantize the network using eqn. 3 and then continue the training of the mixed-precision k_l-bit network. Say, AD_l values of a model are {0.9, 0.3, 0.5} and the initial bit-widths are {16, 10, 8}, applying eqn. 3 will yield bit-widths of {14-bit, 3-bit, 4-bit}, respectively. Note, both the weights and the activations of layer l are quantized to k_l-bits. The updated k_l-bit model is then trained for next set of epochs and this process is repeated till we no longer observe a change in AD_l for any layer. In practice, we observe that the AD of the layers increases with each quantization iteration reaching the maximum limit 1.0 when further quantization is not possible. The final k_l-bit model obtained when $AD_l = 1$ is then trained till convergence. Generally, we find that if we start from a 16-bit model, our method converges to the final mixed-precision model within 3 to 4 iterations ($iter$ in Algorithm 1). Since we train progressively lower-precision models during the training process, we see that the overall training complexity reduces. Our method is summarized in Algorithm 1.

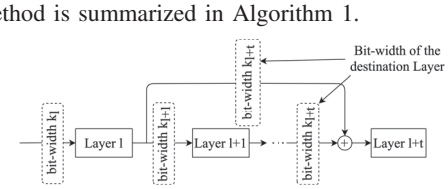

Fig. 2. Illustration to show the bit-width calculation in skip connections.

Note that for ResNet models, where there are skip connections between layers, we use the bit-width of the destination layer to quantize the activations of the skip branches as illustrated in Fig. 2.

Algorithm 1: Activation Density Based Quantization

Input: Model M with layers $l = 1, 2, ..., L$;
Output: Optimal bit width k_l, $\forall\ l \in M$;
Initialize model M with random weights;
Set bit width $k_l^{(0)} = 16$ of initial model, $\forall\ l \in M$;
for *iter* = 1 *to* N **do**
  for *epoch* = 1 *to* $\#(epochs)$ **do**
    Forward and Backward Propagation of M;
    `// Monitor Activation Density. Break if AD is saturated for all layers`
    Compute $AD_l\ \forall\ l \in M$ using eqn. 2;
    if AD_l *is saturated* $\forall\ l \in M$ **then**
      break;
    end
  end
  for *each layer* l *in* M **do**
    `// Calculate the new bit width using eqn. 3`
    $k_l^{(iter)} = round(k_l^{(iter-1)} * AD_l)\ \forall\ l \in M$;
  end
end

IV. Experiments and Results

To validate our method, we run a series of experiments measuring accuracy, energy (section IV-A) and training complexity (section IV-B) on CIFAR-10, CIFAR-100 and TinyImagenet datasets. The model is trained using Adam optimizer under standard settings. We illustrate the fundamentals of our approach by analysing the AD across all layers of a VGG19 model trained on CIFAR-10 before and after quantization in Fig 3 and 4. We summarize our results for all the datasets in Table II, III. Further, we present a hardware evaluation based on a PIM based accelerator in section V.

Fig. 3 shows the training progress along with the trend of AD for each layer of a 16-bit baseline VGG19 model across different training epochs. We can observe that AD converges to a value < 1.0 for all layers implying that there is a considerable amount of redundancy in the baseline model. In contrast, as observed in Fig. 4, when AD based quantization is applied as per Algorithm 1, AD reaches $\sim$1.0 implying an adequate utilization of layers. Interestingly, the AD of the last layer is very low in spite of extreme quantization (bit-width of 1) suggesting that we can entirely remove that layer. This is validated by the results in Table II as the network in *iteration 2a* which has its last layer removed achieves similar accuracy compared to network at *iteration 2*. Note that in Fig. 3 we show the plots upto full 210 epochs of training in order to observe the AD pattern. In practice, we perform quantization and go to the subsequent iteration of training the k_l-bit model once the AD of the previous model stabilizes. For Fig. 3, we essentially stop training the 16-bit baseline at 100th epoch when AD_l across all layers has almost saturated. Tables II and III summarize all the useful metrics of our method. We specify the bit-width

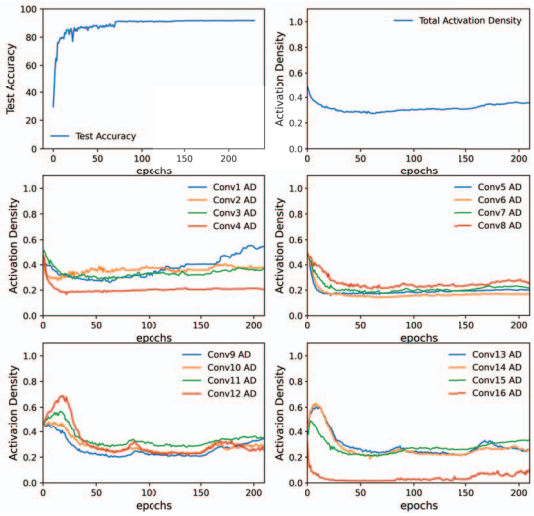

Fig. 3. Trend of Accuracy and Activation Density (AD) vs epochs for VGG19 layers on CIFAR-10: Baseline (Table II (a), Iter 1).

of each layer achieved by our method for each layer in the architecture. Note that for ResNet18, we omit the bit-precision and number of channels for convolutional layers in the skip connections as they are equal to that of the destination layer of the skip connection. Also, across all our experiments, we do not quantize the first layer and the final fully connected layer to avoid a drastic drop in accuracy. We report the accuracy on the test dataset, the overall AD averaged across all layers of the corresponding k_l-bit model, number of epochs trained in each quantization iteration (or $iter$ in Algorithm 1), energy efficiency and training complexity.

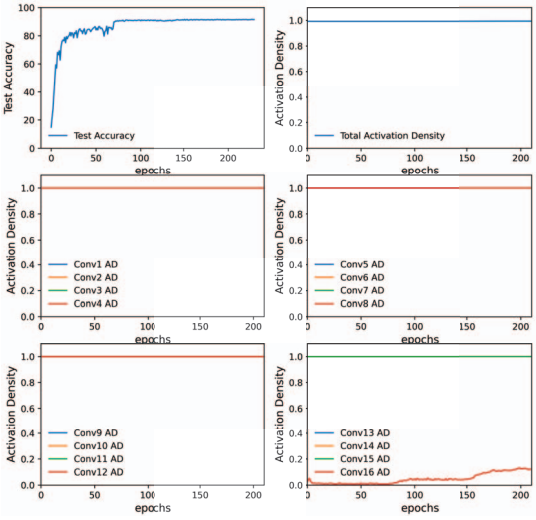

Fig. 4. Trend of Accuracy and Activation Density (AD) vs epochs for VGG19 layers on CIFAR-10: With Activation Density based Quantization (Table II (a), Iter 2). In contrast to fig 3, here we can observe that AD is close to 1 in most of the layers implying maximum utilization.

TABLE I
ENERGY CONSUMPTION ESTIMATES

Operation	Estimated Energy (pJ)
k_l-bit Memory access ($E_{Mem\|k_l}$)	$2.5k_l$
32-bit Multiply ($E_{Mult\|32}$)	3.1
32-bit Add ($E_{Add\|32}$)	0.1
k_l-bit Multiply and Accumulate ($E_{MAC\|k_l}$)	$((3.1 * k_l)/32 + 0.1)$

A. Analytical Energy Estimation

We analytically approximate the energy consumption on traditional hardware based on 45 nm CMOS process using estimates described in [18], [21]. This analytical energy analysis is done for two reasons: 1) To get a rough estimate of the energy consumption without any need for hardware setup, 2) To highlight the contrast between such analytical and real hardware energy estimations (as will be shown in later section V). Note that, for analytical estimations, we are considering only MAC and memory access operations and neglecting any peripheral circuit energy. We provide a comprehensive evaluation considering all overheads on PIM hardware in section V. The energy for k_l-bit memory access and MAC operations are summarized in Table I. For a k_l-bit $p \times p$ convolution layer with I input channels, O output channels operating on an input feature map size $N \times N$ resulting in output feature map of size $M \times M$, the number of memory accesses (N_{mem}), number of MAC operations (N_{MAC}) and total energy per layer l (E_l) is given by: $N_{Mem} = N^2 \times I + p^2 \times I \times O$, $N_{MAC} = M^2 \times I \times p^2 \times O$ and $E_l = N_{Mem} * E_{Mem|k_l} + N_{MAC} * E_{MAC|k_l}$ respectively. Note that if all the weights and activations become binary this method does not hold as the compute reduces to XNOR and popcount.

From Table II, we see our method achieves energy efficiency of $4.19\times$ in VGG19 model trained on CIFAR-10 dataset with no drop in accuracy. Further, with ResNet18 network trained on CIFAR-100 and TinyImagenet datasets, we obtain $3.19\times$ and $4.5\times$ energy efficiency, respectively, at near iso-accuracy with baseline 16-bit precision model.

B. Training Complexity

Since we are iteratively quantizing the model during training, we considerably reduce the amount of time and compute required to train the model. To calculate the savings in terms of training time and energy, we use the *Training Complexity* metric [23] defined as:

$$\sum_{iter=i}^{N} (\text{MAC reduction}_i)^{-1} \times (\#\ \text{epochs}_i) \tag{4}$$

where, i is the denotes the quantization iteration $iter$ from Algorithm 1, $\#epochs_i$ is the number of epochs that the model is trained in each iteration. In Table II, we observe a significant 50% reduction in training complexity in VGG19 model trained on CIFAR-10. We observe a similar improvement in training complexity for CIFAR-100 and TinyImagenet as well.

C. Activation Density based Pruning

In addition to quantization, we also evaluate our method when used simultaneously with AD based pruning proposed in [23]. Thus, along with quantization, we iteratively prune the channels of each layer of a network based on its AD as

TABLE II
RESULTS SUMMARY: ACTIVATION DENSITY BASED QUANTIZATION

Iter	Architecture (Layer-wise bit-width)	Test Accu-racy	Total AD	Energy Effi-ciency	No. of Epochs	Train Compl.
(a) VGG19 on CIFAR-10						
1	16-bit Full Precision for all layers	91.85%	0.284	1x	100	1x
2	bit-width: [16, 4, 5, 4, 3, 2, 2, 2, 3, 3, 3, 4, 3, 3, 3, 3, 16]	91.62%	0.992	4.16x	70	0.524x
2a	bit-width: [16, 4, 5, 4, 3, 2, 2, 2, 3, 3, 3, 4, 3, 3, 3, x, 16]	92.16%	1.000	4.19x	70	0.502x
(b) ResNet18 on CIFAR-100						
1	16-bit Full Precision for all layers	70.90%	0.416	1x	120	1x
2	bit-width: [16, 5, 3, 3, 11, 1, 1, 11, 4, 4, 10, 4, 4, 11, 3, 3, 9, 3, 3, 9, 3, 3, 6, 1, 1, 16]	71.51%	0.743	2.76x	70	0.620x
3	bit-width: [16, 5, 3, 3, 5, 1, 1, 8, 4, 4, 6, 4, 4, 8, 3, 3, 9, 3, 3, 9, 3, 3, 6, 1, 1, 16]	70.51%	0.869	3.19x	70	0.703x
(c) ResNet18 on TinyImagenet						
1	32-bit Full Precision for all layers	44.26%	0.447	1x	60	1x
2	bit-width: [16, 10, 7, 7, 22, 10, 10, 24, 10, 10, 22, 6, 6, 22, 9, 9, 18, 5, 5, 16, 4, 4, 11, 3, 3, 16]	43.94%	0.651	2.73x	25	0.694x
3	bit-width: [16, 3, 7, 7, 16, 2, 2, 17, 3, 3, 15, 6, 6, 15, 9, 9, 9, 5, 5, 7, 4, 4, 4, 3, 3, 16]	44.00%	0.914	4.14x	25	0.705x
4	bit-width: [16, 3, 7, 7, 14, 2, 2, 14, 3, 3, 10, 6, 6, 10, 9, 9, 9, 5, 5, 7, 4, 4, 4, 3, 3, 16]	43.50%	0.917	4.50x	25	0.770x

TABLE III
RESULTS SUMMARY: ACTIVATION DENSITY BASED QUANTIZATION COUPLED WITH ACTIVATION DENSITY BASED PRUNING

Iter	Architecture (Layer-wise bit-width and number of channels)	Test Accu-racy	Total AD	Energy Effi-ciency	No. of Epochs	Train Compl.
(a) VGG19 on CIFAR-10						
1	16-bit Full Precision for all layers nchannels: [64, 64, 128, 128, 256, 256, 256, 256, 512, 512, 512, 512, 512, 512, 512, 512]	91.85%	0.284	1x	100	1x
2	bit-width: [16, 4, 5, 9, 4, 3, 5, 2, 2, 2, 3, 5, 3, 3, 4, 3, 4, 3, 3, 3, 16] nchannels: [19, 22, 38, 24, 45, 37, 44, 54, 103, 126, 150, 125, 122, 112, 111, 8]	86.88%	0.999	980x	70	0.344x
(b) ResNet18 on CIFAR-100						
1	16-bit Full Precision for all layers nchannels: [64, 64, 64, 64, 64, 128, 128, 128, 128, 256, 256, 256, 256, 512, 512, 512, 512]	70.90%	0.416	1x	120	1x
2	bit-width: [16, 5, 3, 11, 1, 11, 4, 10, 4, 11, 3, 9, 3, 9, 3, 6, 1, 16] nchannels: [21, 12, 44, 6, 47, 34, 87, 34, 89, 58, 156, 50, 146, 110, 192, 59, 59]	66.40%	0.732	150x	70	0.372x
3	bit-width: [16, 5, 3, 5, 1, 8, 4, 6, 4, 8, 3, 9, 3, 9, 3, 6, 1, 16] nchannels: [21, 12, 19, 1, 31, 34, 61, 34, 58, 58, 156, 50, 146, 110, 192, 9, 22]	63.01%	0.992	300x	70	0.374x
(c) ResNet18 on TinyImagenet						
1	32-bit Full Precision for all layers nchannels: [64, 64, 64, 64, 64, 128, 128, 128, 128, 256, 256, 256, 256, 512, 512, 512, 512]	44.26%	0.447	1x	60	1x
2	bit-width: [16, 10, 7, 22, 10, 24, 10, 22, 6, 22, 9, 18, 5, 16, 4, 11, 3, 16] nchannels: [20, 14, 45, 21, 48, 42, 88, 27, 91, 73, 151, 41, 129, 70, 178, 56, 20]	38.40%	0.666	93.4x	25	0.450x

TABLE IV
ENERGY VALUES FOR A SINGLE MAC OPERATION FOR MULTIPLE BIT-PRECISIONS USING OUR PROPOSED PIM ACCELERATOR

Energy due to a MAC operation	Value (fJ)
$E_{MAC\|2-bit}$	2.942
$E_{MAC\|4-bit}$	16.968
$E_{MAC\|8-bit}$	66.714
$E_{MAC\|16-bit}$	276.676

TABLE V
COMPARISON OF PIM HARDWARE MAC ENERGY OF MIXED-PRECISION MODEL WITH BASELINE- UNPRUNED FULL-PRECISION MODEL

Network & Dataset	Energy consumed by mixed-precision network (μJ)	Energy consumed by full-precision network (μJ)	Energy reduction
VGG19 on CIFAR-10 dataset	21.506	110.154	5.12x
ResNet18 on CIFAR-100 dataset	33.186	159.501	4.81x

$$C_l = round(C_{l_{initial}} * AD_l) \tag{5}$$

where, C_l and AD_l are the number of channels and the activation density of layer l, respectively.

When used together, AD based pruning and quantization achieve $980\times$ reduction in estimated energy with $< 5\%$ drop in accuracy with VGG19 model trained on CIFAR-10 dataset. Similarly for ResNet18 on CIFAR-100 and TinyImagenet datasets, we observe energy efficiency ranging from $100\times$ to $300\times$ with $\sim 5\%$ accuracy drop. Using pruning with quantization also translates to improvement in training complexity (not shown) since there will be less number of channels as our method progresses.

TABLE VI
PIM HARDWARE MAC ENERGY OF MIXED-PRECISION MODEL (WITH PRUNING) VS BASELINE- UNPRUNED FULL-PRECISION MODEL

Network & Dataset	Energy consumed by pruned mixed-precision network (μJ)	Energy consumed by full-precision network (μJ)	Energy reduction
VGG19 on CIFAR-10 dataset	0.558	110.154	197.55x
ResNet18 on CIFAR-100 dataset	3.630	159.501	43.941x

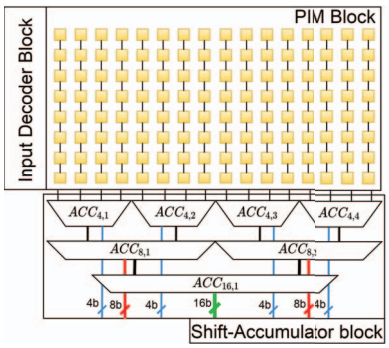

Fig. 5. PIM hardware accelerator with shift-accumulators for mixed-precision support

V. PIM HARDWARE EVALUATION

A. PIM Hardware Setup

To evaluate the working of our proposed method, we develop a Process In-Memory (PIM) accelerator hardware which can support multi-bit precision for various layers of a network [24].

The accelerator architecture shown in Fig. 5 has three main sections: 1) Input decoder, 2) PIM block 3) Shift-Accumulator block. The *Input Decoder Block* fetches the activation values from layer $l-1$ and feeds them to the *PIM block* of the layer l. This is done in a structured pattern. The *PIM block* is a 2-D array of 1-bit SRAM-memory-and-multiply cells each performing 1-bit multiplication between the input activations and weights (stored inside the SRAM cells). After the multiplication, the *Shift-Accumulator block*, which contains a series of accumulators, perform shift-and-add operations. The lowest level of the accumulator block in our PIM architecture is 4-bit. The next two levels perform 8-bit and 16-bit operations. Essentially, 4 columns of the PIM array are read together and the results are stored in the lowest block. Depending upon the bit-width of a given layer, each level of the *Shift-Accumulator block* is activated. For example, if the weight/activation bit-width of a given layer is 2-bits, the corresponding MAC values are stored in the 4-bit accumulator $ACC_{4,i}$ and are regarded as the final result and forwarded as shown using the blue lines. Similarly, if the precision of weights/activations are 4-bits, the results from $ACC_{4,i}$ accumulators undergo shift-and-add to yield 8-bit accumulated results in $ACC_{8,i}$ which are then forwarded, shown using red lines.

B. Energy Evaluation

In a PIM architecture, energy is primarily expended during MAC operation as memory access energy is greatly reduced. Also, energy due to peripheral components is fairly minimal and have not been considered for evaluation. In our architecture, energy is consumed by the *PIM block* and the *Shift-Accumulator block*. Table IV lists the energy consumed for a single MAC operation for multiple bit-precisions evaluated on 45nm CMOS. Using this, we compute the total energy consumption for our model with AD-based quantization/pruning and compare the results with baseline 16-bit full precision models. Table V reports the energy comparison for VGG19, Reset18 networks on CIFAR-10, CIFAR-100 data with and without AD-based-quantization. We obtain $\sim 5\times$ higher energy efficiency over baseline models for both the networks. Similarly, when AD-based quantization-and-pruning is simultaneously applied on the model, improvements of $\sim 197\times$ and $\sim 43\times$ for VGG19 and ResNet18 networks, respectively, are observed (Table VI).

As discussed in Section I, we perform a realistic energy analysis pertaining to more practical hardware architecture considerations as opposed to the analytical estimates shown in Section IV-A. Thus, during analytical estimations in Table III, we get overestimated energy efficiencies $\sim 5-7\times$ greater than practical hardware implementations (Table VI).

VI. Conclusion

We present a simple in-training quantization method based on Activation Density (AD) that enables us to compute the optimal bit-precision of each layer of a network during training. Our approach yields an energy-efficient mixed-precision model with iso-accuracy with baseline. We evaluate the effectiveness of our method on VGG19 and ResNet18 neural networks with multiple benchmark datasets and report our findings. We find that the AD-based quantization approach reduces the training complexity by 50% in our experiments alongside providing up to 4.5x benefit with respect to OPS reductions. We also assess the performance of our algorithm when used simultaneously with AD-based pruning during training. Finally, we propose a scalable PIM accelerator that supports multiple bit-precisions and pruning for different layers of a network. We perform a realistic energy cost analysis to show that our proposed AD-based quantization/pruning algorithm achieves significant energy efficiency improvements $\sim 5-100\times$ when compared with baseline 16-bit models.

Acknowledgement

This work was supported in part by the National Science Foundation, Amazon Research Award, and Technology Innovation Institute Abu Dhabi.

References

[1] N.D.Lane et al., "Squeezing deep learning into mobile and embedded devices," *IEEE Pervasive Computing*, 2017.

[2] H.Song et al., "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[3] Y.Cheng et al., "A survey of model compression and acceleration for deep neural networks," *arXiv preprint arXiv:1710.09282*, 2017.

[4] Y.Tien-Ju et al., "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *CVPR*, 2017.

[5] F.Emile et al., "Weight discretization paradigm for optical neural networks," in *Optical interconnections and networks*, 1990.

[6] B. Wolfgang et al., "Weight quantization in boltzmann machines," *Neural Networks*, 1991.

[7] Y.Guo, "A survey on methods and theories of quantized neural networks," *arXiv preprint arXiv:1808.04752*, 2018.

[8] M Courbariaux et al., "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Advances in neural information processing systems*, 2015.

[9] A.T.Elthakeb el al., "Releq: an automatic reinforcement learning approach for deep quantization of neural networks," *arXiv preprint arXiv:1811.01704*, 2018.

[10] A.Mishra et al., "Wrpn: Wide reduced-precision networks," 2017.

[11] A.Dan et al., "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems 30*, 2017.

[12] H.B.McMahan et al., "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, 2016.

[13] Y.LeCun et al., "Optimal brain damage," in *Advances in neural information processing systems*, 1990.

[14] R.Russell, "Pruning algorithms-a survey," *IEEE TNN*, 1993.

[15] I.Hubara et al., "Quantized neural networks: Training neural networks with low precision weights and activations," *JMLR*, 2017.

[16] Z.Michael et al., "To prune, or not to prune: exploring the efficacy of pruning for model compression," *arXiv preprint arXiv:1710.01878*, 2017.

[17] F.Jonathan et al., "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.

[18] P.Panda et al., "Quanos- adversarial noise sensitivity driven hybrid quantization of neural networks," *ISLPED*, 2020.

[19] C.Indranil et al., "Constructing energy-efficient mixed-precision neural networks through principal component analysis for edge intelligence," *Nature Machine Intelligence*, 2020.

[20] Rastegari et al., "Xnor-net: Imagenet classification using binary convolutional neural networks," in *Computer Vision – ECCV 2016*, 2016.

[21] S.Han et al., "Learning both weights and connections for efficient neural networks," *Advances in neural information processing systems*, 2015.

[22] P.Angshuman et al., "Scnn: An accelerator for compressed-sparse convolutional neural networks," *ACM SIGARCH Computer Architecture News*, 2017.

[23] F.P.Timothy et al., "Activation density driven energy-efficient pruning in training," *arXiv preprint arXiv:2002.02949*, 2020.

[24] Y.Long et al., "Q-pim: A genetic algorithm based flexible dnn quantization method and application to processing-in-memory platform," *DAC*, 2020.