

Adaptive Precision Training for Resource Constrained Devices

Tian HUANG*, Tao LUO†, and Joey Tianyi ZHOU‡

Agency for Science, Technology and Research, Singapore

*huang_tian@ihpc.a-star.edu.sg, †luo_tao@ihpc.a-star.edu.sg, ‡Joey_Zhou@ihpc.a-star.edu.sg

Abstract—Learn in-situ is a growing trend for Edge AI. Training deep neural network (DNN) on edge devices is challenging because both energy and memory are constrained. Low precision training helps to reduce the energy cost of a single training iteration, but that does not necessarily translate to energy savings for the whole training process, because low precision could slows down the convergence rate. One evidence is that most works for low precision training keep an fp32 copy of the model during training, which in turn imposes memory requirements on edge devices. In this work we propose Adaptive Precision Training. It is able to save both total training energy cost and memory usage at the same time. We use model of the same precision for both forward and backward pass in order to reduce memory usage for training. Through evaluating the progress of training, APT allocates layer-wise precision dynamically so that the model learns quicker for longer time. APT provides an application specific hyper-parameter for users to play trade-off between training energy cost, memory usage and accuracy. Experiment shows that APT achieves more than 50% saving on training energy and memory usage with limited accuracy loss. 20% more savings of training energy and memory usage can be achieved in return for a 1% sacrifice in accuracy loss.

I. INTRODUCTION

Deep learning-based Edge AI, as a growing trend, have started to change many aspects of people’s lives. Well trained deep learning models are able to achieve record-breaking predictive performance [15], [7], but we have witnessed increasing demand for the model to learn in-situ, for the purpose of personalisation or adaptation to evolving environment. Comparing to cloud-based deep learning framework, training a deep learning model on edge device reflects the very idea of edge computing, and benefits from energy savings in data transmission, low response latency and enhanced privacy.

Train deep learning model on edge device is challenging because training DNNs usually involves energy-intensive devices such as GPU and CPU with high precision (float32) processing units and abundant memory. For example, neural architecture search on ImageNet [2] would consume 10,000 GPU hours and tons of memory. An edge device such as a smart phone could easily draining out its battery before finishing a few training epochs. Training is memory demanding than inference, as the precision demand for back propagation is usually higher than inference [9] [28]. Many works [9], [16] only focus on compressing network for inference, while still keeping high precision during training. The gap between limited resource on edge devices and the growing demand on training is widening,

as researchers strive for better accuracy and propose larger models for solving more complicated tasks [12].

Low precision training reduces the energy cost of a single training iteration, but that does not necessarily translate to energy savings for the whole training process, because low precision could slows down the convergence rate. [9] A prolonged training process could leads to a worse battery situation. Many works keep an fp32 copy of their model during training [9], [16], [30], [28], [20], [19], which in turn imposes further memory requirements on edge devices.

In this paper we consider the training of a DNN on energy and memory constrained devices. Our goal is to **reduce the total energy and memory cost for the training**. We are inspired by [3] and find that one can train a model with low precision at the beginning of training process, enjoying the energy and memory savings from low precision. As the training curve comes to its plateau, adding more precision help the model approach higher accuracy with fewer training epochs and less energy.

Based on this inspiration, we propose a heuristic method called **Adaptive Precision Training (APT)**. In APT, We use per-layer metrics that indicate how effective a layer learns with given precision. By evaluating the metrics, APT adjusts layer-wise precision dynamically throughout the training process. Below shows a few features that sets APT unique among other training methods:

- APT adjusts layer-wise precision dynamically (i.e. using different bitwidth for different layer in different period of training).
- In training, we use a model of the same precision for both forward and backward propagation in order to save memory usage for training.
- APT provides an application specific hyper-parameter that achieves trade-off between training energy, memory usage and accuracy.

APT finds the layer-wise precision configuration that help the model learn effectively on the fly, It benefits edge devices that have to learn in-situ frequently after deployment. Experiments show that with APT, a model starts in low precision is able to learn as fast as, sometimes even faster than, its fp32 counterpart. APT achieves more than 50% saving on training energy and memory usage with limited accuracy loss. By tuning the application specific hyper-parameter, one can achieve 20% more savings of training energy and memory usage in return for a 1% sacrifice in accuracy loss.

II. RELATED WORK

We mainly focus on adaptive layer-wise precision for training. Orthogonal and complementary techniques for reducing complexity like network compression, pruning [4], [27] and compact architectures [8] are impressively efficient but outside the scope this paper.

Low Precision for Forward Propagation Many works explores the capability of fixed-point representation for forward propagation (FPROP) [11], [18], [21], [23]. These methods keep a floating-point copy of model parameters in the backward propagation (BPROP), which places additional demand on memory usage and data movement during training. These floating-point model parameters are quantised into fixed-point numbers during FPROP.

Quantisation Aware Training Training with quantised gradients has been well studied in the distributed learning, whose main motivation is to reduce the communication cost during gradient aggregations between workers [1], [20]. Our goal setting are different because an edge device may not have access to remote support and has to learn in-situ by itself. DoReFa-Net [28] quantise gradients to low-bitwidth fixed-point numbers in the BPROP. TernGrad [20] quantises gradients to ternary values in the BPROP. But the weights are stored and updated in floating-point representations for these works. WAGE [22] provided a lossy information compression technique and manage to operates BPROP in 8-bit. The goal setting of this work is different from us, as it focuses on reducing the energy cost for single training iteration.

Adaptive Training There are a few works that quantise weight and activation with layer-wise precision. [18], [21], [23] explore the space of layer-wise bitwidth for energy efficient inference. [26] learns quantisation parameters through finding gradient of these parameters. Adding more learnable parameters could introduce more uncertainty of the convergence of training curve. It also introduce more energy cost for learning more parameters. Zhang et al. [25] also emphasise on energy saving for single training iterations and store and update weights in fp32 format in BPROP.

III. ADAPTIVE PRECISION TRAINING

In this work, we apply a widely used quantisation scheme [11] to convert floating point numbers to into integers. The scheme is equivalent to an *affine mapping* of integers q to real numbers r . The formalised equation is shown as follows $r = S(q - Z)$. S and Z represent the scale and zero point of a group of values, or a tensor. All values in a tensor share one S and Z . Different tensors have their own S and Z . For k -bit quantisation, q has 2^k possible discrete states.

A. Quantisation Underflow

Low precision integer arithmetic is cheaper than a floating point arithmetic in terms of energy and memory, but quantisation magnifies the underflow issue. [13] Equation 1 shows the update process of a weight in fp32 format:

$$w_{ij} := w_{ij} - lr * g_{ij} \quad (1)$$

w_{ij} is the j -th weight of i -th layer, lr is the learning rate, g_{ij} is the corresponding gradient of the w_{ij} . Ideally, at each training step, a weight changes by $lr * g_{ij}$. As we apply k -bit quantisation to a tensor of weights, too small changes cannot be represented by the weight. We refer to this minimum resolution as ϵ , which is formalised as follow:

$$\epsilon_i = \frac{\max(W_i) - \min(W_i)}{2^k - 1} \quad (2)$$

W_i is a tensor of weights of i -th layer. k is the precision, or the bitwidth for the tensor. The updates of a weight w_{ij} is regulated by ϵ_i , which can be formalised as the following equation:

$$w_{ij} := w_{ij} - \left\lfloor \frac{lr * g_{ij}}{\epsilon_i} \right\rfloor * \epsilon_i \quad (3)$$

As a result, $lr * g_{ij}$ is quantised to discrete states, which is equivalent to that of w_{ij} . $lr * g_{ij}$ has to be at least bigger than ϵ_i , otherwise quantisation underflow happens. For high precision training (e.g. 32-bit) ϵ is small, $lr * g_{ij}$ is bigger than ϵ_i for most of cases. As the precision decreases, ϵ_i increases, quantisation underflow happens more often, putting higher resistance on weights updating.

Gradient is the horse power that drives the training process forward. Low-precision training put restrictions on the horse power and slow down the training process. We have the observation that in low precision network, gradients vanishes quicker than that of a network with higher precision. This is because quantisation underflow happens more often in low precision model and guides the model into local minima. Some of layers have larger ϵ and are harder to be updated than others. As the training loss decreases, These layers with larger ϵ suffer more quantisation underflow than before, stepping onto a slippery slope. As the training process moves forward, quantisation underflow freezes more parameters, driving the training into a dead state.

B. Metrics for Underflow

Many works [9], [16], [30], [28], [20], [19], [25] keep an fp32 copy of all parameters in order to prevent the above-mention situation from happening, in the cost of memory usage, energy on data movement and additional training iterations. We believe there are low precision configurations that is able to save training energy and memory usage at the same. The key is to understand how easy quantisation underflow happens to a layer, I propose a metric called $Gavg$, which is formalised as Equation 4.

$$Gavg_i = \frac{1}{N_i} \sum_{j=0}^{N_i} \left| \frac{g_{ij}}{\epsilon_i} \right| \quad (4)$$

N_i represents the number of parameters in a tensor, $Gavg_i$ indicates how large a gradient is related to the minimum resolution of parameters in i -th layer. The larger the $Gavg$ is, the less likely the parameters remains unchanged during back propagation. If $Gavg$ approaches zero, that means the

layer suffers serious quantisation underflow problem and does not update for most of times. The precision is the key to prevent $Gavg$ from approaching zero. A higher precision leads to lower ϵ and higher $Gavg$, which means the parameters are easier to update.

Although we use weight to describe the concept of quantisation underflow and metrics, $Gavg$ applies to other parameters that need to be learned during training, e.g. bias, the clipping point of activation and gradient. In the metric $Gavg$, we do not include other factors like learning rate or momentum in the metrics so that user can still use training tricks or sophisticated optimisers over our training method.

C. Precision Adjustment Policy

The metric $Gavg$ is a good measure for balancing requirements between training energy, memory usage, and accuracy. It is well known that precision is directly related to the energy, memory and accuracy. Using fixed precision across the whole network may not meet the requirement of each layer. On comparison, a single threshold for $Gavg$ is able to produce different precision configuration according to the distribution of parameters and their gradients.

We propose a precision adjustment policy based on the metric $Gavg$. We start with a naive policy, which is to make sure that $Gavg$ of all layers are within a pre-defined range. A C-styled description of the policy is presented below.

Algorithm 1: Precision Adjustment Policy

```

input :  $k_{0\dots M-1}$ ,  $Gavg_{0\dots M-1}$ ,  $T_{min}$ ,  $T_{max}$ ,
output:  $k_{0\dots M-1}$ 
1 for ( $i = 0$ ;  $i < M$ ;  $i = i + 1$ ) {
2   if ( $Gavg_i < T_{min}$   $\&\&$   $k_i < 32$ ) {
3      $k_i := k_i + 1$ ;
4   }
5   if ( $Gavg_i > T_{max}$   $\&\&$   $k_i > 2$ ) {
6      $k_i := k_i - 1$ ;
7   }
8 }
9 return  $k_{0\dots M-1}$ ;

```

In Algorithm 1, $k_{0\dots M-1}$ and $Gavg_{0\dots M-1}$ represents the precision and metrics of all M layers. T_{min} , T_{max} are the upper and lower limit. Algorithm 1 increase the precision of a layer when its $Gavg < T_{min}$ and decrease the precision when $Gavg > T_{max}$. The lower limit ensures all layers learn effective, whereas the upper limit is for saving energy cost and memory usage on those parameters that are very easy to update (e.g. due to its small range or large gradients).

A typical workflow of APT is described in Algorithm 2. A training starts with a low-precision model. The metrics evaluation happens inside each training epoch. The precision adjustment happens between training epochs. $Gavg$ does not have to be calculated for each training iteration. A few times in each epoch suffice to give a profile of $Gavg$ for precision adjustment.

Algorithm 2: Training with APT

```

1 Initialise all layers of a model with low precision, e.g.
   $k = 6$ ;
2 for ( $epoch = 0$ ;  $epoch < 200$ ;  $epoch ++$ ) {
3   for (
     $iter = 0$ ;  $iter < len(train\_loader)$ ;  $iter ++$ )
    {
4      Forward propagation;
5      Backward propagation;
6      if  $iter \% INTERVAL == 0$  then
        Evaluate  $Gavg$  using Equation 4;
8      Moving average on  $Gavg$ ;
9    end
10  }
11 Adjust model precision using Algorithm 1;
12 }

```

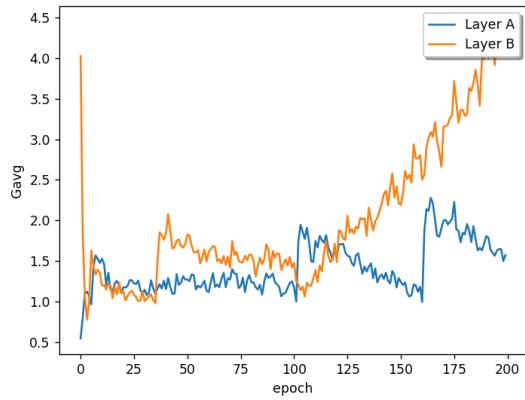


Fig. 1. $Gavg$ v.s. Epoch for two layer

Figure 1 demonstrate the trend of $Gavg$ of two layers in a training with adaptive precision. T_{min} is set to 1.0 in this demo. Layer A starts with a $Gavg$ below T_{min} , indicating it suffers quantisation underflow. APT allocates more bitwidth to lift the $Gavg$ of layer A above the threshold. Layer B is very easy to update at the beginning. Whenever $Gavg$ hits T_{min} , APT allocates more bitwidth to ensure layer B learns effectively. T_{max} is set to inf in this demo. It is also possible to use T_{max} to reduce bitwidth for layers that do not suffer much from quantisation underflow.

IV. EXPERIMENTS

In this section we evaluate the APT with three datasets: CIFAR10 and CIFAR100 [14]. We follow the data augmentation in [22] for training: 4 pixels are padded on each side, and a 32x32 patch is randomly cropped from the padded image or its horizontal flip. For testing, only single view of the original 32x32 image is evaluated. Three popular backbones, ResNet-20, ResNet-110 [6], and MobileNetV2 [17], are included in the experiments.

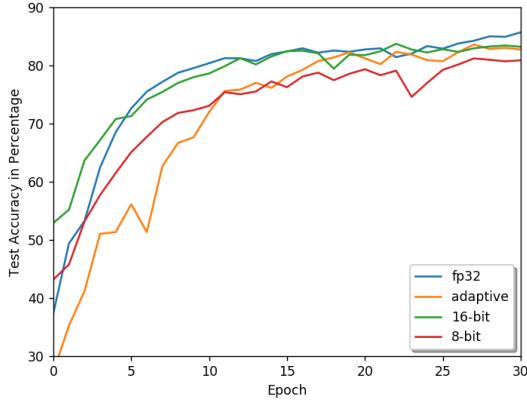


Fig. 2. Test Accuracy v.s. Epoch for ResNet20 on CIFAR10

Many works used sophisticated optimiser, such as Adam, in their experiments. We use SGD to show the potential of saving energy and memory usage. We use a weight decay of 0.0001 and momentum of 0.9, and adopt the weight initialization in [5] and BN [10] with no dropout. The model is trained with a minibatch size of 128 on one GPU. For experiments on CIFAR10, we start with a learning rate of 0.1, divide it by 10 at 100 epoch and 150 epoch iterations, and terminate training at 200 iterations. For experiments on CIFAR100, we apply learning rate warm up, which reduces the learning rate down to 0.01 for the first two epochs and go back to learning rate scheduling for CIFAR10 for the rest of the training.

For all experiments we set initial bitwidth to 6, which will be explained later. Unless specified, we set the application specific parameter (T_{min}, T_{max}) to $(6.0, inf)$ for all experiments. This setting is better for demonstrating the capability of APT in saving energy and memory usage. One can set (T_{min}, T_{max}) to other values that meets the requirements of the applications.

A. Training Curves and Adaptive Precision Adjustment

We compare the our APT method with a vanilla SGD with different precision. The weights of all models are quantised for both forward pass and backward pass. Figure 2 shows the training curve.

fp32 and 16-bit model has the steepest learning curves among all, as they do not suffer much from quantisation underflow problem. The curve of 8-bit model does not climb as fast as fp32 and 16-bit does. An investigation into the training statistics shows that G_{avg} of the all layers drop from the scale of 1 down to $1e-1$ within the first 50 epoch, which indicates quantisation underflow happens model wide and significantly slows down the training process of the 8-bit model.

On comparison, our adaptive training method starts with a model initialised with 6-bit weights. Its training curve starts with lower accuracy at the beginning, even lower than that of 8-bit. It overtakes the 8-bit and catch up with 16-bit and fp32. This is achieved by adaptively adjust layer-wise bitwidth of the model.

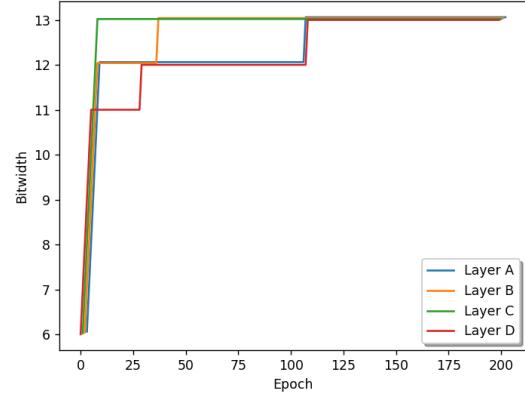


Fig. 3. Layer-wise Bitwidth v.s. Epoch for ResNet20 on CIFAR10

Figure 3 demonstrate the changes our method made to the layer-wise bitwidth of the model. Only four layers with weight parameters out of 20 are shown in this figure for the sake of clarity. Overlapping curved are shifted slight away from each other for the same purpose.

APT treats layers differently according to their learning effectiveness. Some layers are trained with lower bitwidth in early epochs and achieve solid energy savings. At 100 epoch, The precision of the first and last layer reach bitwidth of 13 as the training loss drops quickly after the decay of the learning rate.

Though Figure 3 we understand that APT also serves as a neural architecture search method. It start training with an initial bitwidth of 6, which is for sure to be under-qualified with the target application. The adaptive training method will dynamically adjust the bitwidth of the network, and approaches the target accuracy with fewer energy. This means an initial bitwidth other than 6 leads to similar similar results. We choose 6 for all experiments just for demonstration purpose.

B. Savings in Energy and Memory

In this subsection we demonstrates that our adaptive training method saves training energy and memory. APT provides users trade-off between training energy and test accuracy. For this comparison, we train ResNet20 with CIRFAR10 dataset. The weights of all models are quantised for both forward pass and backward pass. The 12-bit, 14-bit, 16-bit and 32-bit models use fixed bitwidth throughout training process. We do not include 10-bit, 8-bit or lower bitwidth because their suffer large accuracy loss and fall off charts in most cases. Our adaptive training method starts on a model with initial weight bitwidth of 6.

The grouped bars in Figure 4 show the energy our method saved given a target accuracy. X axis is the Top 1 accuracy ranging from 91% to 92%. All training energy cost are normalised to the cost of 32-bit model.

12-bit model has the least training energy cost among the all fixed-bitwidth models. It still spends 9% more training energy

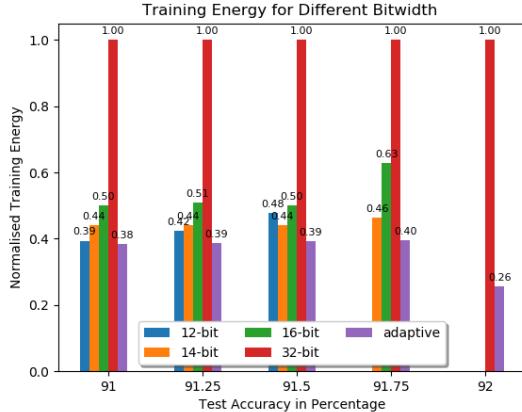


Fig. 4. Training Energy v.s. Bitwidth for ResNet20 on CIFAR10

than our adaptive method does. The 12-bit model is absent in the 91.75% and 92% group because within 200 epoch it is not able to reach an accuracy of 91.75%.

In order to achieve the last few percentage of improvement in accuracy, models with fixed bitwidth spend a lot of training energy. For example, 16-bit model spends 13% more energy to achieve an improvement of 0.25% from 91.5% to 91.75%. On comparison, our adaptive method managed to achieve this by just 1% additional energy.

One may notice that the training energy of our adaptive method drops from 40% to 26% when achieving the accuracy improvement from 91.75% to 92%. This is because the 32-bit model spends more training energy than our method does for the same improvement. As the training energy in this figure is normalised to that of the 32-bit model, the training energy of our method drops from a percentage point of view.

In Figure 4, APT trains with $T_{min} = 6.0$. This hyper-parameter is application specific, which can be used as a trade-off between accuracy and training energy. Next we demonstrate that by tuning this threshold, one is able to achieve more savings in training energy and memory cost in the cost of accuracy loss.

In Figure 5, we use $Gavg$ threshold ranging from 0.1 – 100.0 and generate a scatter plot of the training energy cost (in orange) for 200 epochs and corresponding accuracy it achieves. All training energy cost are normalised to that of 32-bit model training.

Through Figure 5 we know that higher $Gavg$ threshold results in more training energy, memory usage and higher accuracy. For threshold below 1.0, accuracy increases quickly along with the increase of training energy. The plateau appears on the right of threshold 1.0 means extra training energy brings less improvements, which is very common during the training process. One can choose a $Gavg$ threshold that fits the need of his or her applications.

Figure 5 also presents the model size for training of the same experiment. We use low-precision representation of parameter during the back propagation in order to save memory usage

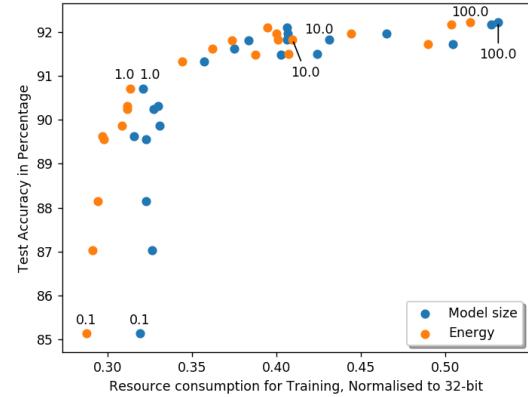


Fig. 5. Resource Consumption for Training v.s. Test Accuracy for ResNet20 on CIFAR10

for training. The memory usage is normalised to that of a 32-bit presentation. The memory usage follows the same trend as the training energy.

C. Comparison to Others

Table I shows the comparison to other quantisation methods. Model Precision in BPROP refers to the representation used in the update of weights. Many works [9], [16], [30], [28], [20], [19] keeps a fp32 copy of weights for back propagation in order to prevent a prolonged training process. There is no savings in memory usage for training. For example, TernGrad [20] is only for worker-to-server communication in distributed training, weights are still accumulated with fp32. In fact, [22], [24] need more training epochs and/or sophisticated optimiser to reach comparable accuracy. With SGD optimiser and quantised model in BPROP, APT is able to achieve more than 50% of savings in energy and memory usage with limited accuracy loss.

V. CONCLUSION AND FUTURE WORK

Training DNNs on resource constrained edge devices is challenging. We propose Adaptive Precision Training method that saves both energy cost and memory usage for training. APT evaluates how effective each layer learns with its current precision. Based on the evaluation, APT performs layer-wise precision adjustments dynamically to make sure the model learns effective throughout the training process. APT has a hyper-parameter T_{min} which provides trade-off between training energy and accuracy. Tuning parameter T_{min} requires application specific knowledge. In future, we are going to find automatic ways for choosing a proper T_{min} in order to ease the use of APT.

REFERENCES

- [1] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsdgd: Compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.

TABLE I
COMPARISON OF NETWORK QUANTISATION METHODS

Method	Model Precision in BPROP	Optimizer	CIFAR10	CIFAR100
BNN [9]	FP32	Adam	89.85	NA
TWN [16]	FP32	BinaryRelax [24]	88.65 (ResNet-20) [24] ^b	68.95 (ResNet-110) [24] ^b
TTQ [30]	FP32	Adam	91.13 (ResNet-20)	NA
DoReFa Net [28]	FP32	Adam	89.52 [29]	61.43 [29]
TernGrad [20]	FP32 ^a	Adam	85.64 (CifarNet)	NA
WAGE [22]	8-bit	SGD	93.22 (VGG-like) ^b	NA
E2 Train [19]	FP32	SGD	93.07 (ResNet110)	NA
APT	Adaptive	SGD	92.23 (ResNet20) 93.96 (MobileNet v2)	68.38 (ResNet110)

^a Only for worker-to-server communication in distributed training, weights are accumulated with float32
^b 300 training epochs

- [2] Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5918–5926, 2017.
- [3] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning*, pages 1737–1746, 2015.
- [4] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [7] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.
- [8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and H Adam Mabilenets. Efficient convolutional neural networks for mobile vision applications. *arXiv preprint ArXiv:1704.0486*, 2017.
- [9] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research*, 18(1):6869–6898, 2017.
- [10] S Ioffe and C Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. cornell university library, 2017.
- [11] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.
- [12] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.
- [13] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.
- [14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [16] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.
- [17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [18] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8612–8620, 2019.
- [19] Yue Wang, Ziyu Jiang, Xiaohan Chen, Pengfei Xu, Yang Zhao, Yingyan Lin, and Zhangyang Wang. E2-train: Training state-of-the-art cnns with over 80% energy savings. In *Advances in Neural Information Processing Systems*, pages 5139–5151, 2019.
- [20] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems*, pages 1509–1519, 2017.
- [21] Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090*, 2018.
- [22] Shuang Wu, Guoqi Li, Feng Chen, and Luping Shi. Training and inference with integers in deep neural networks. *arXiv preprint arXiv:1802.04680*, 2018.
- [23] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [24] Penghang Yin, Shuai Zhang, Jiancheng Lyu, Stanley Osher, Yingyong Qi, and Jack Xin. Binaryrelax: A relaxation approach for training deep neural networks with quantized weights. *SIAM Journal on Imaging Sciences*, 11(4):2205–2223, 2018.
- [25] Xishan Zhang, Shaoli Liu, Rui Zhang, Chang Liu, Di Huang, Shiyi Zhou, Jiaming Guo, Yu Kang, Qi Guo, Zidong Du, et al. Adaptive precision training: Quantify back propagation in neural networks with fixed-point numbers. *arXiv preprint arXiv:1911.00361*, 2019.
- [26] Yichi Zhang, Ritchie Zhao, Weizhe Hua, Nayun Xu, G Edward Suh, and Zhiru Zhang. Precision gating: Improving neural network efficiency with dynamic dual-precision activations. *arXiv preprint arXiv:2002.07136*, 2020.
- [27] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017.
- [28] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.
- [29] Zhengguang Zhou, Wengang Zhou, Xutao Lv, Xuan Huang, Xiaoyu Wang, and Houqiang Li. Progressive learning of low-precision networks. *arXiv preprint arXiv:1905.11781*, 2019.
- [30] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.