

Binary Ensemble Neural Network: More Bits per Network or More Networks per Bit?

Shilin Zhu
UC San Diego
La Jolla, CA 92093
shz338@eng.ucsd.edu

Xin Dong
Harvard University
Cambridge, MA 02138
xindong@g.harvard.edu

Hao Su
UC San Diego
La Jolla, CA 92093
haosu@eng.ucsd.edu

Abstract

Binary neural networks (BNN) have been studied extensively since they run dramatically faster at lower memory and power consumption than floating-point networks, thanks to the efficiency of bit operations. However, contemporary BNNs whose weights and activations are both single bits suffer from severe accuracy degradation. To understand why, we investigate the representation ability, speed and bias/variance of BNNs through extensive experiments. We conclude that the error of BNNs are predominantly caused by the intrinsic instability (training time) and non-robustness (train & test time). Inspired by this investigation, we propose the Binary Ensemble Neural Network (BENN) which leverages ensemble methods to improve the performance of BNNs with limited efficiency cost. While ensemble techniques have been broadly believed to be only marginally helpful for strong classifiers such as deep neural networks, our analysis and experiments show that they are naturally a perfect fit to boost BNNs. We find that our BENN, which is faster and more robust than state-of-the-art binary networks, can even surpass the accuracy of the full-precision floating number network with the same architecture.

1. Introduction

Deep Neural Networks (DNNs) have achieved great impact to broad disciplines in academia and industry [57, 38]. Recently, the deployment of DNNs are transferring from high-end cloud to low-end devices such as mobile phones and embedded chips, serving general public with many real-time applications, such as drones, miniature robots, and augmented reality. Unfortunately, these devices typically have limited computing power and memory space, thus cannot afford DNNs to achieve important tasks like object recognition involving significant matrix computation and memory usage.

Binary Neural Network (BNN) is among the most promising techniques to meet the desired computation and memory requirement. BNNs [31] are deep neural networks whose




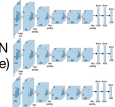
	Network	Numerics	Operations	Memory Speed-Up	Inference Speed-Up	Accuracy
	Full-Precision DNN 	W: 32 bit A: 32 bit	+, -, x	1x	1x	High
	BNN 	W: 1 bit A: 1 bit	XNOR, Bitcount	~32x	(CPU) ~58x	Low
	(Ours) BENN (K ensemble) 	W: 1 bit A: 1 bit	XNOR, Bitcount	~(32/K)x	(CPU) ~58x	High

Figure 1. Comparison between traditional floating-number DNN, BNN and our proposed BENN on image recognition task (W: weights, A: activations). The inference speed of BENN can be further boosted on FPGAs [63].

weights and activations have only two possible values (e.g., -1 and +1) and can be represented by a single bit. Beyond the obvious advantage of saving storage and memory space, the binarized architecture admits only bitwise operations, which can be computed extremely fast using digital logic units [20] such as arithmetic-logic unit (ALU) with much less power consumption than floating-point unit (FPU).

Despite the significant gain in speed and storage, however, current BNNs suffer from notable accuracy degradation when applied to challenging tasks such as ImageNet classification. To mitigate the gap, previous researches in BNNs have been focusing on designing more effective optimization algorithms to find better local minima of the quantized weights. However, the task is highly non-trivial, since gradient-based optimization that used to be effective to train DNNs now becomes tricky to implement.

In this paper, we investigate BNNs systematically in terms of representation power, speed, bias, variance, stability, and their robustness. We find that BNNs suffer from severe intrinsic instability and non-robustness regardless of network parameter values. What implied by this observation is that the performance degradation of BNNs are not likely to be resolved by solely improving the optimization techniques; instead, it is mandatory to cure the BNN function, particularly to reduce the prediction variance and improve its robustness

to noises.

Inspired by the analysis, in this work, we propose Binary Ensemble Neural Network (BENN). Though the basic idea is as straight-forward as to simply aggregate multiple BNNs by boosting or bagging, we show that the statistical properties of the ensembled classifiers become much nicer: not only the bias and variance are reduced, more importantly, BENN's robustness to noises at test time is significantly improved. All the experiments suggest that BNNs and ensemble methods are a perfectly natural fit. Using architectures of the same connectivity (a compact Network in Network [42]), we find that *boosting only 4 ~ 5 BNNs would be able to even surpass the baseline DNN with real weights in the best case*. In addition, our initial exploration by applying BENN on ImageNet recognition using AlexNet [38] and ResNet [27] also shows a large gain. This is by far the fastest, most accurate, and most robust results achieved by binarized networks (Fig. 1).

To the best of our knowledge, this is the first work to bridge BNNs with ensemble methods. Unlike traditional BNN improvements that have computational complexity of $\gtrsim O(K^2)$ by using K -bit per weights [65] or K bases in total [43], the complexity of BENN is reduced to $O(K)$. Compared with [65, 43], BENN also enjoys better bitwise operation parallelizability. With trivial parallelization, the complexity can be reduced to $O(1)$. We believe that BENN can shed light on more research along this idea to achieve extremely fast yet robust computation by networks.

2. Related Work

Quantized and binary neural networks: People have found that there is no need to use full-precision parameters and activations and can still preserve the accuracy of a neural network using k -bit fixed point numbers, as stated by [19, 23, 61, 8, 40, 41, 48, 56, 49]. The first approach is to use low-bit numbers to approximate real ones, which is called quantized neural networks (QNNs) [32]. [66, 64] also proposed ternary neural networks. Although recent advances such as [65] can achieve competitive performance compared with full-precision models, they cannot fully speed it up because we still cannot perform parallelized bitwise operation with bitwidth larger than one. [31] is the very recent work that binarizes all the weights and activations, which was the birth of BNN. They have demonstrated the power of BNNs in terms of speed, memory use and power consumption. But recent works such as [58, 11, 21, 10] also reveal the strong accuracy degradation and mismatch issue during the training when BNNs are applied in complicated tasks such as ImageNet ([12]) recognition, especially when the activation is binarized. Although some work like [43, 50, 13] have offered reasonable solutions to approximate full-precision neural network, much more computation and tricks on hyperparameters are still needed to implement compared with

BENN. Since they either use K -bitwidth quantization or K binary bases, the computational complexity cannot get rid of $O(K^2)$ if $O(1)$ is required for 1-bit single BNN, while BENN can achieve $O(K)$ and even $O(1)$ if multiple threads are naturally paralleled. Also, many of current literatures tried to minimize the distance between binary and real-value parameters. But empirical assumptions such as Gaussian parameter distribution are usually required in order to get a priori for each BNN or just keep the sign same as suggested by [43], otherwise the non-convex optimization is hard to deal with. By contrast, BENN can be a general framework to achieve the goal and has strong potential to work even better than full-precision networks, without involving more hyperparameters than a single BNN.

Ensemble techniques: To avoid simply relying on a single powerful classifier, the ensemble strategy can improve the accuracy of given learning algorithm combining multiple weak classifiers as summarized by [6, 9, 47]. The two most common strategies are bagging by [5] and boosting by [51, 17, 53, 26], which were proposed many years ago and have strong statistical foundation. They have roots in a theoretical framework PAC model by [59] which was the first to pose the question of whether weak learners can be ensembled into a strong learner. Bagging predictors are proved to reduce variance while boosting can reduce both bias and variance, and their effectiveness have been proved by many theoretical analysis. Traditionally ensemble was used with decision trees, decision stumps, random forests and achieved great success thanks to its desirable statistical properties. Recently people use ensemble to increase the generalization ability of deep CNNs [24], advocate boosting on CNNs and do architecture selection [45], and propose boost over features [30]. But people did not pay enough attention to ensemble techniques because neural network is not a weak classifier anymore thus ensemble can unnecessarily increase the model complexity. However, when applied to weak binary neural networks, we found it generates new insights and hopes, and BENN is a natural outcome of such perfect combination. In this work, we build our BENN on the top of variant bagging, AdaBoost by [15, 52], LogitBoost by [17] and can be extended to many more variants of traditional ensemble algorithms. We hope this work can revive these intelligent approaches and bring their life back into modern neural networks.

3. Why Making BNNs Work Well is Challenging?

Despite the speed and space advantage of BNN, its performances is still far inferior to the real valued counterparts. There are at least two possible reasons: First, functions representable by BNNs may have some inherent flaws; Second, current optimization algorithms may still not be able to find a good minima. While most researchers have been work-

ing on developing better optimization methods, we suspect that BNNs have some fundamental flaws. The following investigation reveals the fundamental limitations of BNN-representable functions experimentally.

Because all weights and activations are binary, an obvious fact is that BNNs can only represent a subset of discrete functions, being strictly weaker than real networks that are universal continuous function approximators [29]. What are not so obvious are two serious limitations of BNNs: the robustness issue w.r.t. input perturbations, and the stability issue w.r.t. network parameters. Classical learning theory tells us that both robustness and stability are closely related to the generalization error of a model [62, 4]. A more detailed theoretical analysis on BNN’s problems is attached in supplementary material.

Robustness Issue: In practice, we observe more severe overfitting effects of BNNs than real networks. Robustness is defined as the property that if a testing population is “similar” to a training population, then the testing error is close to the training error [62]. To verify this point, we experiment in a random network setting and a trained network setting.

Random Network Setting. We compute the following quantity to compare 32bit real-valued DNN, BNN, QNN, and our BENN model (Sec. 4) on the Network-In-Network (NIN) architecture:

$$\mathbb{E}_w \mathbb{E}_{\Delta x} \|f(x + \Delta x; w) - f(x; w)\|^2 \quad (1)$$

where f is the network and w represents network weights.

We randomly sample real-valued weights $w \sim \mathcal{N}(0, I)$ as suggested in literature to get a DNN f_r with weights w_r and binarize it to get a BNN f_b with binary weights w_b . We also independently sample and binarize w_r to generate multiple BNNs with the same architecture to simulate the BENN and get w_{benn} . QNN is obtained by quantizing the DNN to k -bit weights (W) and activations (A). We normalize each input image in CIFAR-10 to the range $[-1, 1]$.

Then we inject the input perturbation Δx on each example by a Gaussian noise with different variances (0.001 \sim 0.1), run a forward pass on each network, and measure the expected l_2 norm of the change on the output distribution. The above l_2 norm of DNN, BNN, QNN, and BENN averaged by 1000 sampling rounds is shown in Fig. 2(left) with perturbation variance 0.01.

Results show that BNNs always have larger output variation, suggesting that they are more susceptible to input perturbation, and BNN does worse than QNN that has more bits. We also observe that having more bits on activations actually improves BNN’s robustness significantly, while having more bits on weights has smaller improvement (Fig. 2(left, right)).

Trained Network Setting. To further consolidate the discovery, we also train a real-valued DNN f_r and a BNN f_b using XNOR-Net [50] rather than direct sampling. We also include our designed BENN f_{benn} in comparison. Then we

perform the same Gaussian input perturbation Δx , run a forward pass, and calculate the change of classification error \mathcal{L} on CIFAR-10 as:

$$\mathbb{E}_{\Delta x} \|\mathcal{L}(f(x + \Delta x)) - \mathcal{L}(f(x))\|^2 \quad (2)$$

Results in Fig. 2(middle) indicates that BNNs are still more sensitive to noises even if it is well optimized. Although people have shown that weights in BNN still have nice statistical properties as in [1], the conclusion can change dramatically if both weights and activations are binarized while input is perturbed.

Stability Issue: BNNs are known to be hard to optimize due to problems such as gradient mismatch and non-smoothness of activation function. While [40] has shown that stochastic rounding converges to within $O(\Delta)$ accuracy of the minimizer in expectation where Δ denotes quantization resolution, assuming the error surface is convex, the community has not fully understood the non-convex error surface of BNN and how it interacts with different optimizers such as SGD or ADAM [37].

To compare the stability of different networks (sensitivity to network parameter during optimization), we measure the accuracy fluctuation after a large amount of training steps. Fig. 2 (right) shows the accuracy oscillation in the last 20 training steps after we train BNN and QNN with 300 epochs, and results show that we should at least have QNN with weights and activations both 4-bit in order to stabilize the network.

One explanation of such instability is the non-smoothness of the function output w.r.t. the binary network parameters. Note that, as the output of the activation function in the previous layer, the input to each layer of BNNs are binarized numbers. In other words, not only each function is non-smooth w.r.t. the input, but also it is non-smooth w.r.t. the learned parameters. As a comparison, empirically, BENN with 5 and 32 ensembles (denoted as BENN-05/32 in Fig. 2) have already achieved amazing stability.

4. Binary Ensemble Neural Network

In this section, we illustrate our BENN using bagging and boosting strategies, respectively. In all experiments, we adopt the widely used *deterministic binarization* as $x_b = \text{Sign}(x)$ for network weights and activations, which is preferred to leverage hardware accelerations. However, back-propagation becomes challenging since the derivative is zero almost everywhere except for the stepping point. In this work, we borrow the common strategy called “straight-through estimator” (STE) [28] during back-propagation, defined as $\frac{\partial J}{\partial x} = \frac{\partial J}{\partial x_b} I_{|x| \leq 1}$.

4.1. BENN-Bagging

The key idea of bagging is to average weak classifiers that are trained from i.i.d. samples of the training set. To train

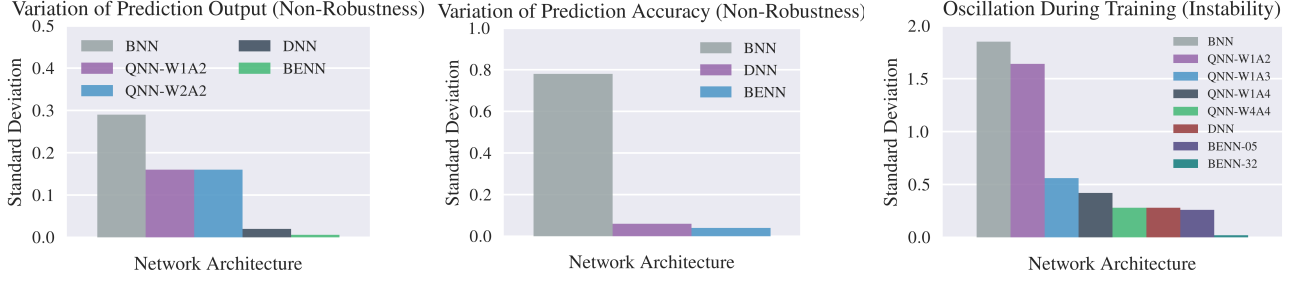


Figure 2. **Left:** BNN has large output variation (robustness issue). **Middle:** BNN has large variation of prediction accuracy (robustness issue). **Right:** BNN has large test accuracy variation during training (instability issue). BENN can cure these problems. Here, the perturbation variance is 0.01. (*QNN-W1A2 denotes QNN with 1-bit weights and 2-bit activations and so do others.)

each BNN classifier, we sample M examples independently with replacement from the training set \mathcal{D} . We do this K times to get K BNNs, denoted as b^1, \dots, b^K . The sampling with replacement assures that each BNN sees roughly 63% of the entire training set.

At test time, we aggregate the opinions from these K classifiers and decide among C classes. We compare two ways of aggregating the outputs. One is to choose the label that most BNNs agree with (hard decision), while the other is to choose the best label after aggregating their softmax probabilities (soft decision).

The main advantage brought by bagging is to reduce the variance of a single classifier. This is known to be extremely effective for deep decision trees which suffer from high variance, but only marginally helpful to boost the performance of neural networks, since networks are generally quite stable. Interestingly, though less helpful to real-valued networks, bagging is effective to improve BNNs since the instability issue is severe for BNNs due to gradient mismatch and strong discretization noise as stated in Sec. 3.

4.2. BENN-Boosting

Boosting is another important tool to ensemble classifiers. Instead of just aggregating the predictions from multiple independently trained BNNs, boosting combines multiple weak classifiers in a sequential manner and can be viewed as a stage-wise gradient descent method optimized in the function space. Boosting is able to reduce both bias and variance of individual classifiers.

There are many variants of boosting algorithms and we choose the AdaBoost [15] algorithm for its popularity. Suppose classifier k has hypothesis $b^k : X \rightarrow \mathbb{R}$, weight α_k , and output distribution p^k , we can denote the aggregated classifier as $B^K : X \rightarrow \mathbb{R}$ and its aggregated output distribution P^K . Then AdaBoost minimizes the following exponential loss:

$$J(B^K) = \sum_i e^{-Y^T P^K} = \sum_i e^{-Y^T (P^{K-1} + \alpha_k p^K)}$$

where $Y = (y_1, \dots, y_C)^T$ and i denotes the index of the training example.

Rewighting Principle The key idea of boosting algorithm is to have the current classifier pay more attention to the misclassified samples by previous classifiers. Reweighting is the most common way of budgeting attention based on the historical results. There are essentially two ways to accomplish this goal:

- **Reweighting on sampling probabilities:** Suppose initially each training example i is assigned $\pi = u_i = 1/M$ uniformly, so each sample gets equal chance to be picked. After each round, we reweight the sampling probability according to the classification confidence.
- **Reweighting on loss/gradient:** We may also incorporate u_i into the gradient, so that a BNN b^k updates parameters with larger step size on misclassified examples and vice versa. For example, set $\nabla_w J(b^k) \leftarrow \lambda \cdot (\alpha_k p_y^k)(u_i) \cdot \nabla_w J(b^k)$, where λ is the learning rate. However, we observe that this approach is less effective experimentally for BNNs, and we conjecture that it exaggerates the gradient mismatch problem.

4.3. Inference-Time Complexity

A 1-bit BNN with the same connectivity as the original full-precision 32-bit DNN can save $\sim 32x$ memory. In reality, BNN can achieve $\sim 58x$ speed up on the current generation of 64-bit CPUs [50] and may be further improved with special hardware such as FPGA. Some existing works only binarize the weights but leave activations full-precision, which practically only results in $\gtrsim 2x$ speed up. As for BENN with K ensembles, each BNN's inference is independent, thus the total memory saving is $\sim 32/Kx$. As for boosting, we can further compress BNN to save more computations and memory usage. Besides, existing approaches have complexity $O(K^2)$ with K -bit QNN [65] or use K binary bases [43], because they cannot avoid the bit collection operation to generate a number, although their fixed-point computation is much more efficient than float-point computation. If $O(1)$ is the time complexity of the boolean operation, then BENN reduces the quadratic complexity to linear, i.e., $O(K)$ with K ensembles but still maintains the very satisfying accuracy and stability as stated above. We can even

make the inference in $O(1)$ for BENN if multiple threads are supported. A complete comparison is shown in Table 1.

4.4. Stability Analysis

Given a full-precision real valued DNN f_w with a set of parameters $w \sim N(0, \sigma_w^2)$, a BNN f_{w_b} with binarized parameters w_b , input vector $x \sim N(0, 1)$ (after Batch Normalization) and perturbation $\Delta x \sim N(0, \sigma^2)$, and a BENN $f_{w_{benn}}$ with K ensembles, we want to compare their stability and robustness w.r.t. the network parameters and input perturbation. Here we analyze the variance of output change before and after perturbation, which echoes Eq. 1 in Sec. 3. This is because the output change has zero mean and its variance reflects the distribution of output variation. More specifically, larger variance means increased variation of output w.r.t. input perturbation.

Assume $f_w, f_{w_b}, f_{w_{benn}}$ are outputs before non-linear activation function of a single neuron in an one-layer network, we have the output variation of real-value DNN as $f_w(x + \Delta x) - f_w(x) = w \odot \Delta x$, whose distribution has variance $\sigma_r^2 = |w| \sigma_w^2 \sigma^2$, where $|w|$ denotes number of input connections for this neuron and \odot denotes inner product. Some modern non-linear activation function $g(\cdot)$ like ReLU will not change the inequality of variances, thus we can omit them in the analysis to keep it simple.

For BNN with both weights and activations binarized, we can rewrite the above formulation as $f_{w_b}^b(x + \Delta x) - f_{w_b}^b(x) = \text{sign}(w) \odot [\text{sign}(x + \Delta x) - \text{sign}(x)]$, thus having variance $\sigma_{bnn}^2 = |w| \sigma_{\text{sign}(w)}^2 (\sigma_{\text{sign}(x+\Delta x) - \text{sign}(x)}^2)$. And for BENN-Bagging, we have $\sigma_{benn}^2 = \sigma_{bnn}^2 / K$ with K ensembles, since bagging effectively reduces variance. For BENN-Boosting, our model can reduce both bias and variance at the same time. However for boosting, the analysis on bias and variance becomes much more difficult and there are still some debates in literature [7, 17]. With these Gaussian assumptions and some numerical experiments (detailed analysis and theorems can be found in supplementary material), we can verify the large stability gain of BENN over BNN compared with floating-number DNN. As for robustness, the same analysis principle can be applied to perturbing weights as Δw compared with Δx used in stability analysis.

5. Independent and Warm-Restart Training for BENNs

We train our BENN with two different methods. The first one is to initialize each new classifier *independently* and re-train it, which is a traditional way. To accelerate the training of new weak classifier in BENN, we can also initialize the weights of the new classifier by cloning the weights from the most recently trained classifier. We name this training scheme as *warm-restart training*, and we conjecture that the knowledge of those unseen data for the new classifier has been transferred from the inherited weights and is helpful

to increase the discriminability of the new classifier. Interestingly, we observe that for small network and dataset like Network-In-Network [42] on CIFAR-10, warm-restart training has better accuracy. However, independent training is better when BENN is applied to large network and dataset such as AlexNet [38] and ResNet [27] on ImageNet since overfitting problem emerges. More discussion can be found in Sec. 6 and Sec. 7.

Implementation Details We train BENN on the image classification task with CNN block structure containing a batch normalization layer, a binary activation layer, a binary convolution layer, a non-binary activation layer (e.g., sigmoid, ReLU), and a pooling layer, as used by many recent works [50, 65]. To compute the gradient of step function $\text{sign}(x)$, we use the same approach suggested by STE. When updating parameters, we use real-valued weights as [50] suggests otherwise the tiny update could be killed by deterministic binarization and training cannot move on. In this work, we train each BNN using standard independent and warm-restart training. Unlike the previous works which always keep the first and last layer full-precision, we test 7 different BNN architecture configurations as shown in Table 2 and use them as ingredients for ensemble in BENN.

6. Experimental Results

We evaluate BENN on CIFAR-10 and ImageNet datasets with a self-designed compact Network-In-Network (NIN) [42], the standard AlexNet [38] and ResNet-18 [27], respectively. We have summarized in Table 2 the configurations of all BNN variants. More detailed specifications of the networks can be found in the supplementary material. For each type of BNN, we obtain the converged single BNN when training is done. We also store BNN after each training step and obtain the best BNN along the way by picking the one with the highest test accuracy (e.g., *Best SB*). We use *BENN-T-R* to denote the *BENN* by aggregating R BNNs of configuration T (e.g., *BENN-SB-32*). We also denote *Bag/Boost-Indep* and *Bag/Boost-Seq* as bagging/boosting with standard independent training and warm-restart sequential training (Sec. 5). All ensembled BNNs share the same network architecture as their real-valued DNN counterpart in this paper, although studying multi-model ensemble is an interesting future work. The code of all our experiments will be made public online.

6.1. Insights Generated from CIFAR-10

In this section, we show the large performance gain using BENN on CIFAR-10 and summarize some insights. Each BNN is initialized by a pre-trained model from XNOR-Net [50] and then retrained by 100 epochs to reach convergence before ensemble. Each full-precision DNN counterpart is trained by 300 epochs to obtain the best accuracy for reference. The learning rate is set to 0.001 and ADAM optimizer

Table 1. Analysis of Theoretically Computational Complexity on a Single Network. (F-full-precision, Q_k -k-bit quantization, B-binary)

Network	Weights	Activation	Operations	Memory Saving	Computation Saving
Standard DNN	F	F	+, -, ×	1	1
[10, 33, 39, 66, 64],...	B	F	+, -	~ 32x	~ 2x
[65, 32, 61, 2],...	Q_k	Q_k	+, -, ×	~ $\frac{32}{k}$ x	< $\frac{58}{k^2}$ x
[43],...	$k \times B$	$k \times B$	+, -, XNOR, bitcount	~ $\frac{32}{k}$ x	~ $\frac{58}{k^2}$ x
[50] and ours	B	B	XNOR, bitcount	~ 32x	~ 58x

Table 2. Weak BNN Configurations Used to Ensemble (W-weights, A-activation, Params-number of parameters in network). The Last Two are Naive Compressed Network.

Weak BNN Configuration/Type (T)	Weight	Activation	Size	Params
SB (Semi-BNN)	First and last layer:32-bit	First and last layer:32-bit	100%	100%
AB (All-BNN)	All layers:1-bit	All layers:1-bit	100%	100%
WQB (Weight-Quantized-BNN)	All layers:Q-bit	All layers:1-bit	100%	100%
AQB (Activation-Quantized-BNN)	All layers:1-bit	All layers:Q-bit	100%	100%
IB (Except-Input-BNN)	All layers:1-bit	First layer: 32-bit	100%	100%
SB/AB/IB-Tiny (Tiny-Compress-BNN)	-	-	50%	25%
SB/AB/IB-Nano (Nano-Compress-BNN)	-	-	10%	1%

Table 3. Oscillation During Training (Instability)

Network	Ensemble Method	#Ensemble	STD
SB	-	1	2.94
Best SB	-	1	1.40
BENN-SB	Bag-Seq	5	0.31
BENN-SB	Boost-Seq	5	0.24
BENN-SB	Bag-Seq	32	0.03
BENN-SB	Boost-Seq	32	0.02

is used. Here, we use a compact Network-In-Network (NIN) for CIFAR-10. We first present some significant independent comparisons as follows and then summarize the insights we found.

Single BNN versus BENN: We found that BENN can achieve much better accuracy and stability than a single BNN with negligible sacrifice in speed. Experiments across all BNN configurations show that BENN has the accuracy gain ranging from 4.21% to 24.16% over BNN on CIFAR-10. If each BNN is weak (e.g., AB), the gain of BENN will increase as shown in Fig. 3 (right). This verifies that BNN is indeed a good weak classifier for ensembling. Surprisingly, BENN-SB outperforms full-precision DNN after 32 ensembles (either bagging or boosting) by up to 1.52% (Fig. 3 (left)). Note that in order to have the same memory usage as a 32-bit DNN, we constrain the ensemble up to 32 rounds if no network compression is involved. If more ensembles are available, we observe further performance boost but accuracy gain will eventually become flat.

We also compare BENN-SB-5 (i.e., 5 ensembles) with WQB (Q=5, 5-bit weight and 1-bit activation), which have the same amount of parameters (measured by bits). WQB can only achieve ~ 80% accuracy unstably while our ensemble network can reach up to ~ 86% and remain stable.

We also measure the accuracy variation of the classifier in the last 20 training steps for all BNN configurations. The results in Table 3 indicate that BENN can reduce BNN’s variance by ~ 90% if ensemble 5 rounds and ~ 99% after 32 rounds. Moreover, picking the best BNN with the highest test accuracy instead of using the BNN when training is done

Table 4. Impact of Network Compression

Network	Ensemble Method	#Ensemble	Accuracy
Best SB	-	1	84.91%
BENN-SB	Bag-Seq	32	89.12%
BENN-SB	Boost-Seq	32	89.00%
Best SB-Tiny	-	1	77.20%
BENN-SB-Tiny	Bag-Seq	32	84.09%
BENN-SB-Tiny	Boost-Seq	32	84.32%
Best SB-Nano	-	1	40.70%
BENN-SB-Nano	Bag-Seq	500	57.12%
BENN-SB-Nano	Boost-Seq	500	63.11%

can also reduce the oscillation. This is because the statistical property of ensemble framework (Sec. 3 and Sec. 4.4) makes BENN become a graceful way to ensure high stability.

Bagging versus boosting: It is known that bagging can only reduce the variance of the predictor, while boosting can reduce both bias and variance. Fig. 3(right), Fig. 4, and Table 4 show that boosting outperforms bagging, especially after BNN is compressed, by up to 2.51% when network size is reduced to 50% (*Tiny config*) and 13.38% when network size is reduced to 10% (*Nano config*), and the gain increases from 5 to 32 ensembles. This verifies that boosting is a better choice if the model does not overfit much.

Standard independent training versus warm-restart training: Standard ensemble techniques use independent training, while warm-restart training enable new classifiers to learn faster. Fig. 3 (left) shows that warm-restart training performs better up to 3.9% for bagging and 2.95% for boosting after the same number of training epochs. This means gradually adapting to more examples might be a better choice for CIFAR-10. However, this does not hold for ImageNet task because of slight over-fitting with warm-restart (Sec. 6.2). We believe that this is an interesting phenomenon but it needs more justification by studying the theory of convergence.

The impact of compressing BNN: BNN’s model complexity largely affects bias and variance. If each weak BNN has enough complexity with low bias but high variance, then bagging is more favorable than boosting due to simplicity.

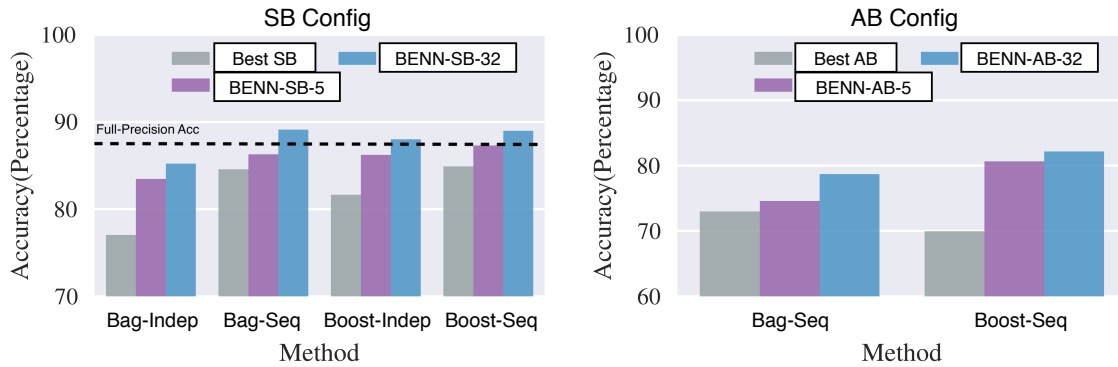


Figure 3. **Left:** BENN can increase the test accuracy significantly with more ensembles. It can even achieve better accuracy than its full-precision counterpart under Semi-BNN (SB) case. **Right:** Boosting strongly outperforms bagging in All-BNN (AB) case where each BNN has larger bias.

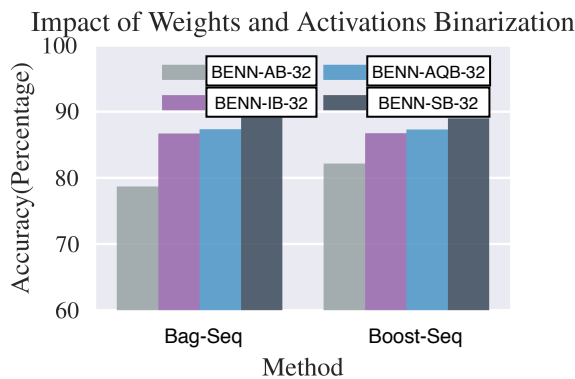


Figure 4. After ensemble, the accuracy increases with more activation bits ($Q=2$ in AQB). Preserving the first and/or last layer full-precision (IB and SB) helps, compared with all-binary case (AB).

However, if each BNN’s size is small with large bias, boosting becomes a much better choice. To verify this, we compare each BNN in Table 2 by naively reducing the amount of channels and neurons in each layer. The results in Table 4 show that BENN-SB can maintain reasonable performance even after naive compression, and boosting gains more over bagging in severe compression (*Nano config*).

We also found that BENN is less sensitive to network size. Table 4 shows that compression reduces single BNN’s accuracy by 7.71% (*Tiny config*) and 44.21% (*Nano config*). After 32 ensembles, the performance loss caused by compression decreases to 4.8% and 26.01% respectively. Surprisingly, we observe that compression only reduces the accuracy of full-precision DNN by 1.18% (*Tiny config*) and 16.03% (*Nano config*). So it is necessary to have not-too-weak BNNs to build BENN that can compete with full-precision DNN. Better pruning algorithm can be combined with BENN in the future rather than naive compression to allow smaller network to be ensemble.

The effect of bit width: Higher bitwidth results in lower variance and bias at the same time. This can be seen in Fig. 4 where we make activations 2-bit in BENN-AQB ($Q=2$).

As can be seen, BENN-AQB ($Q=2$) and BENN-IB have comparable accuracy after 32 ensembles, but much better than BENN-AB and worse than BENN-SB. We also observe that activation binarization results in much more unstable model than weight binarization. This indicates that the gain of having more bits is mostly due to better features from the input image, since input binarization is a real pain for neural networks. Surprisingly, BENN-AB can still achieve more than 80% accuracy under such a pain.

The effect of binarizing first and last layer: Almost all the existing works in BNN assume the full precision of the first and last layer, since binarization on these two layers will cause severe accuracy degradation. But we found BENN is less affected, as shown by BENN-AB, BENN-SB and BENN-IB in Fig. 4. The BNN’s accuracy loss due to binarizing these two special layers is 3.98% \sim 11.9%. For BENN with 32 ensembles, the loss reduces to 2.36% \sim 6.98%.

In summary, we generate our main insights about BNN and BENN: (1) Ensemble such as bagging and boosting greatly relieve BNN’s problems in terms of representation power, stability, and robustness. (2) Boosting gains advantage over bagging in most cases, and warm-restart training is often a better choice. (3) Weak BNN’s configuration (i.e., size, bitwidth, first and last layer) is essential to build a well-functioning BENN to match full-precision DNN in practice.

6.2. Exploration on Applying BENN to ImageNet Recognition

We believe BENN is one of the best neural network structures for inference acceleration. To demonstrate the effectiveness of BENN, we compare our algorithm with state-of-the-arts on the ImageNet recognition task (ILSVRC2012) using AlexNet [38] and ResNet-18 [27]. Specifically, we compare our BENN-SB independent training (Sec. 5) with the full-precision DNN [38, 50], DoReFa-Net (k-bit quantized weight and activation) [65], XNOR-Net (binary weight and activation) [50], BNN (binary weight and activation) [31]

Table 5. Comparison with state-of-the-arts on ImageNet using AlexNet (W-weights, A-activation)

Method	W	A	Top-1
Full-Precision DNN [38, 50]	32	32	56.6%
XNOR-Net [50]	1	1	44.0%
DoReFa-Net [65]	1	1	43.6%
BinaryConnect [10, 50]	1	32	35.4%
BNN [31, 50]	1	1	27.9%
BENN-SB-3, Bagging (ours)	1	1	48.8%
BENN-SB-3, Boosting (ours)	1	1	50.2%
BENN-SB-6, Bagging (ours)	1	1	52.0%
BENN-SB-6, Boosting (ours)	1	1	54.3%

Table 6. Comparison with state-of-the-arts on ImageNet using ResNet-18 (W-weights, A-activation)

Method	W	A	Top-1
Full-Precision DNN [27, 43]	32	32	69.3%
XNOR-Net [50]	1	1	48.6%
ABC-Net [43]	1	1	42.7%
BNN [31, 50]	1	1	42.2%
BENN-SB-3, Bagging (ours)	1	1	53.4%
BENN-SB-3, Boosting (ours)	1	1	53.6%
BENN-SB-6, Bagging (ours)	1	1	57.9%
BENN-SB-6, Boosting (ours)	1	1	61.0%

and BinaryConnect (binary weight) [10]. Note that accuracy of BNN and BinaryConnect on AlexNet are reported by [50] instead of original authors. For DoReFa-Net and ABC-Net, we use the best reported accuracy by original authors with 1-bit weight and 1-bit activation. For XNOR-Net, we report the number of our own retrained model. Our BENN is retrained given a well pre-trained model until convergence by XNOR-Net after 100 epochs to use, and we retrain each BNN with 80 epochs before ensemble. As shown in Table 5 and 6, BENN-SB is the best among all the state-of-the-art BNN architecture, even with only 3 ensembles paralleled on 3 threads. Meanwhile, although we do observe continuous gain with more ensembles, we found that BENN with more ensembles on ImageNet task can be unstable in terms of accuracy and needs further investigation on overfitting issue, otherwise the rapid gain is not always guaranteed. Here we report the numbers where the performance is stable, although we do observe even better performance sometimes. We believe our initial exploration along this direction has shown BENN’s potentiality of catching up full-precision DNN and even surpass it with more base BNN classifiers. In fact, how to optimize BENN on large and diverse dataset is still an interesting open problem.

7. Discussion

More bits per network or more networks per bit? We believe this paper brings up this important question. As for biological neural networks such as our brain, the signal between two neurons is more like a spike instead of high-range real-value signal. This implies that it may not be necessary to use real-valued numbers, while involve a lot of redundancies and can waste significant computing power. Our work converts the direction of ‘how many bits per network is enough?’ into ‘how many networks per bit?’. BENN provides a hierarchical view, i.e., we build weak classifiers by groups

of neurons, and build a strong classifier by ensembling the weak classifiers. We have shown that this hierarchical approach is more intuitive and natural to represent knowledge. Although the optimal ensemble structure is beyond the scope of this paper, we believe that some structure searching or meta-learning techniques can be applied. Moreover, the improvement on single BNN such as studying the error surface and resolving the curse of activation/gradient binarization is still essential for the success of BENN.

BENN is hardware friendly: Using BENN with K ensembles is better than using one K -bit classifier. Firstly, K -bit quantization still cannot get rid of fixed-point multiplication, while BENN can support bitwise operation. People have found that BNN can be further accelerated on FPGAs over modern CPUs [63, 18]. Secondly, people have shown that the complexity of a multiplier is proportional to the square of bitwidth, thus BENN simplifies the hardware design. Thirdly, BENN can use spike signals in the chips instead of keeping the signal real-valued all the time, which can save a lot of energy. Finally, unlike recent literature requiring quadratic time to compute, BENN can be better paralleled on the chips due to its linear time complexity. In fact, we have implemented our BENN (SB model) on *Xilinx Zynq XCZU7EV* FPGA boards, achieved up to $\sim 200x$ fold improvement over CPU measured by GOPS/Watt with computation reusing strategies.

Current limitations: It is known to all that ensemble methods can potentially cause overfitting to the model and we also observed similar problems on CIFAR-10 and ImageNet, when the number of ensembles keeps increasing. An interesting next step is to analyze the property of decision boundary of BENN on different datasets and track its evolution in high-dimensional feature space. Also, training will take longer time if many ensembles are needed (especially on large dataset like ImageNet), thus reducing the speed of design iterations, although our training can be easily scaled with GPUs. Finally, BENN needs to be further optimized for large networks such as AlexNet and ResNet in order to show its full power, such as picking the best ensemble rule and base classifier. Modern advanced variants of the ensemble techniques can be incorporated into BENN as well.

8. Conclusion

In this paper, we proposed BENN, a novel neural network architecture which marries BNN with ensemble methods. The experiments showed a large performance gain in terms of accuracy, robustness, and stability. Our experiments also reveal some insights about trade-offs on bit width, network size, number of ensembles, etc. We believe that by leveraging specialized hardware such as FPGA and more advanced modern ensemble techniques with less overfitting, BENN can be a new dawn for deploying large deep neural networks into mobile and embedded systems.

References

- [1] A. G. Anderson and C. P. Berg. The high-dimensional geometry of binary neural networks. *arXiv preprint arXiv:1705.07199*, 2017. 3
- [2] S. Anwar, K. Hwang, and W. Sung. Fixed point optimization of deep convolutional neural networks for object recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1131–1135. IEEE, 2015. 6
- [3] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
- [4] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002. 3
- [5] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 2
- [6] L. Breiman. Bias, variance, and arcing classifiers. 1996. 2
- [7] P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007. 5
- [8] Z. Cai, X. He, J. Sun, and N. Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. *arXiv preprint arXiv:1702.00953*, 2017. 2
- [9] J. G. Carney, P. Cunningham, and U. Bhagwan. Confidence and prediction intervals for neural network ensembles. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 2, pages 1215–1218. IEEE, 1999. 2
- [10] M. Courbariaux, Y. Bengio, and J.-P. David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pages 3123–3131, 2015. 2, 6, 8
- [11] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016. 2
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2
- [13] L. Deng, P. Jiao, J. Pei, Z. Wu, and G. Li. Gated xnor networks: Deep neural networks with ternary weights and activations under a unified discretization framework. *arXiv preprint arXiv:1705.09283*, 2017. 2
- [14] P. Domingos. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238, 2000.
- [15] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995. 2, 4
- [16] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Bari, Italy, 1996.
- [17] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000. 2, 5
- [18] C. Fu, S. Zhu, H. Su, C.-E. Lee, and J. Zhao. Towards fast and energy-efficient binarized neural network inference on fpga. *arXiv preprint arXiv:1810.02068*, 2018. 8
- [19] Y. Gong, L. Liu, M. Yang, and L. Bourdev. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*, 2014. 2
- [20] G. Govindu, L. Zhuo, S. Choi, and V. Prasanna. Analysis of high-performance floating-point arithmetic on fpgas. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, page 149. IEEE, 2004. 1
- [21] Y. Guo, A. Yao, H. Zhao, and Y. Chen. Network sketching: Exploiting binary structure in deep cnns. *arXiv preprint arXiv:1706.02021*, 2017. 2
- [22] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning*, pages 1737–1746, 2015.
- [23] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 2
- [24] S. Han, Z. Meng, A.-S. Khan, and Y. Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 109–117, 2016. 2
- [25] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.
- [26] T. Hastie, S. Rosset, J. Zhu, and H. Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009. 2
- [27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5, 7, 8
- [28] G. Hinton. Neural networks for machine learning. In *Coursera*, 2012. 3
- [29] K. Hornik, M. Stinchcombe, and H. White. Multilayer feed-forward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 3
- [30] F. Huang, J. Ash, J. Langford, and R. Schapire. Learning deep resnet blocks sequentially using boosting theory. *arXiv preprint arXiv:1706.04964*, 2017. 2
- [31] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks. In *Advances in neural information processing systems*, pages 4107–4115, 2016. 1, 2, 7, 8
- [32] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *arXiv preprint arXiv:1609.07061*, 2016. 2, 6

- [33] K. Hwang and W. Sung. Fixed-point feedforward deep neural network design using weights+ 1, 0, and- 1. In *Signal Processing Systems (SiPS), 2014 IEEE Workshop on*, pages 1–6. IEEE, 2014. 6
- [34] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [35] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [36] M. Kim and P. Smaragdis. Bitwise neural networks. *arXiv preprint arXiv:1601.06071*, 2016.
- [37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3
- [38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2, 5, 7, 8
- [39] F. Li, B. Zhang, and B. Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016. 6
- [40] H. Li, S. De, Z. Xu, C. Studer, H. Samet, and T. Goldstein. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems*, pages 5813–5823, 2017. 2, 3
- [41] D. Lin, S. Talathi, and S. Annapureddy. Fixed point quantization of deep convolutional networks. In *International Conference on Machine Learning*, pages 2849–2858, 2016. 2
- [42] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2, 5
- [43] X. Lin, C. Zhao, and W. Pan. Towards accurate binary convolutional neural network. In *Advances in Neural Information Processing Systems*, pages 344–352, 2017. 2, 4, 6, 8
- [44] Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio. Neural networks with few multiplications. *arXiv preprint arXiv:1510.03009*, 2015.
- [45] M. Moghimi, S. J. Belongie, M. J. Saberian, J. Yang, N. Vasconcelos, and L.-J. Li. Boosted convolutional neural networks. In *BMVC*, 2016. 2
- [46] J. Ott, Z. Lin, Y. Zhang, S.-C. Liu, and Y. Bengio. Recurrent neural networks with limited numerical precision. *arXiv preprint arXiv:1608.06902*, 2016.
- [47] N. C. Oza and S. Russell. *Online ensemble learning*. University of California, Berkeley, 2001. 2
- [48] E. Park, J. Ahn, and S. Yoo. Weighted-entropy-based quantization for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [49] A. Polino, R. Pascanu, and D. Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018. 2
- [50] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016. 2, 3, 4, 5, 6, 7, 8
- [51] R. E. Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003. 2
- [52] R. E. Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013. 2
- [53] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999. 2
- [54] D. Soudry, I. Hubara, and R. Meir. Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights. In *Advances in Neural Information Processing Systems*, pages 963–971, 2014.
- [55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [56] W. Sung, S. Shin, and K. Hwang. Resiliency of deep neural networks under quantization. *arXiv preprint arXiv:1511.06488*, 2015. 2
- [57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. *Cvpr*, 2015. 1
- [58] W. Tang, G. Hua, and L. Wang. How to train a compact binary neural network with high accuracy? In *AAAI*, pages 2625–2631, 2017. 2
- [59] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 2
- [60] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066, 2013.
- [61] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016. 2, 6
- [62] H. Xu and S. Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012. 3
- [63] R. Zhao, W. Song, W. Zhang, T. Xing, J.-H. Lin, M. Srivastava, R. Gupta, and Z. Zhang. Accelerating binarized convolutional neural networks with software-programmable fpgas. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 15–24. ACM, 2017. 1, 8
- [64] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017. 2, 6
- [65] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou. Dorefanet: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 2, 4, 5, 6, 7, 8
- [66] C. Zhu, S. Han, H. Mao, and W. J. Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016. 2, 6