electronics

MDPI

Article

Increasing Information Entropy of Both Weights and Activations for the Binary Neural Networks

Wanbing Zou [1,2,3], Song Cheng [1,2,3], Luyuan Wang [3], Guanyu Fu [2,3], Delong Shang [1,3], Yumei Zhou [1,2,3] and Yi Zhan [1,3,*]

[1] Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; zouwanbing@ime.ac.cn (W.Z.); chengsong@ime.ac.cn (S.C.); shangdelong@ime.ac.cn (D.S.); ymzhou@ime.ac.cn (Y.Z.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China; fuguanyu19@mails.ucas.ac.cn
[3] Nanjing Institute of Intelligence Technology, Institute of Microelectronics of the Chinese Academy of Sciences, Nanjing 211135, China; wly@niit.ac.cn
* Correspondence: yizhan@ime.ac.cn

Abstract: In terms of memory footprint requirement and computing speed, the binary neural networks (BNNs) have great advantages in power-aware deployment applications, such as AIoT edge terminals, wearable and portable devices, etc. However, the networks' binarization process inevitably brings considerable information losses, and further leads to accuracy deterioration. To tackle these problems, we initiate analyzing from a perspective of the information theory, and manage to improve the networks information capacity. Based on the analyses, our work has two primary contributions: the first is a newly proposed median loss (ML) regularization technique. It improves the binary weights distribution more evenly, and consequently increases the information capacity of BNNs greatly. The second is the batch median of activations (BMA) method. It raises the entropy of activations by subtracting a median value, and simultaneously lowers the quantization error by computing separate scaling factors for the positive and negative activations procedure. Experiment results prove that the proposed methods utilized in ResNet-18 and ResNet-34 individually outperform the Bi-Real baseline by 1.3% and 0.9% Top-1 accuracy on the ImageNet 2012. Proposed ML and BMA for the storage cost and calculation complexity increments are minor and negligible. Additionally, comprehensive experiments also prove that our methods can be applicable and embedded into the present popular BNN networks with accuracy improvement and negligible overhead increment.

Keywords: binary neural network (BNN); deep learning; information capacity; quantization error

Citation: Zou, W.; Cheng, S.; Wang, L.; Fu, G.; Shang, D.; Zhou, Y.; Zhan, Y. Increasing Information Entropy of Both Weights and Activations for the Binary Neural Networks. *Electronics* **2021**, *10*, 1943. https://doi.org/10.3390/electronics10161943

Academic Editor: George A. Papakostas

Received: 14 July 2021
Accepted: 10 August 2021
Published: 12 August 2021

1. Introuction

In the past few decades, deep convolution neural networks (CNNs) have evolved rapidly. This technology shows excellent performance on a lot of tasks, such as image recognition [1,2], object detection [3,4], and segmentation [5]. A large part of reasons why the performance is so excellent is that traditional CNNs usually have large number of parameters and floating-point operations (FLOPs). The property makes CNNs capable of strong representation ability, although with intensive computational cost and memory footprint requirement. According to existing hardware level, these complex models can be trained and inferred effectively in the cloud servers equipped with powerful GPUs, but still difficult to be deployed on limited-resources platforms such as smartphones, AR/VR devices, and drones. To solve this problem, increasing researchers begin to explore reducing the size of network models and its FLOPs scale with minimal computational accuracy loss.

Among those solutions, two categories are in this research field. The first category, mainly for network structure, designs efficient network architectures with less computation and memory footprint requirement, such as the MobileNet [6], SqueezeNet [7], ShuffleNet [8], and DenseNet [9]. The second category, mainly for further light-weighting

optimization on existing complex network, mainstreams the parameter pruning [10–12] and quantization [13–17] methods. The binary neural networks (BNNs) is a radical case of quantization. It has been attracted increasing attention due to its beneficial properties—both activations and weights are quantized to $\{-1, +1\}$. Moreover, the calculations inside BNNs can only have simple XNOR and Bitcount operations with this advantageous feature. This makes significant performance improvement both in run-time and its power-aware hardware implementations.

However, the BNN's performance declines dramatically in accuracy compared to full precision network. The reason is that when both activations and weights are quantized to $\{-1, +1\}$, the representation ability of the network drops sharply and results in severe information loss. To solve this problem, two approaches has been proposed in previous researches: (1) For minimizing the information loss of parameters, IR-Net [18] adopts Libra Parameter Binarization (Libra-PB) to balance and standardize weights in forward propagation. (2) In perspective of increasing entire system's information entropy, the Shannon entropy based information loss penalty is proposed for the BNN networks [19]. The above two methods can effectively mitigate the information loss during the binarization process, thereby improve the representation ability and entire system's accuracy additionally. Even so, the IR-Net still ignores that the binarized weights do not obey the normal distribution. This leads the normalization adopted cannot get a maximum entropy of binarized weights. In addition, the IR-Net only deals with the weights and excludes the factor of activations as well as reference [19] does. In addition, reference [20] summaries most of the exited binarization methods, such as CI-BCNN [21], BCGD [22], etc. These articles are targeted to improve the performance of BNN. In our work, we start from a new aspect of information entropy increment, and the weights and activations inside BNNs are both included and under a thoroughly consideration.

Figure 1 illustrates the basic blocks of binarization process in our work. Inside it, a median loss (ML, Red block in Figure 1) method is proposed to make the binarized network weights distribute more evenly. This benefits a greater entropy of weights in each layer, and improves the representation ability of the whole network accordingly. Another, if a large difference between the positive and negative activation amplitudes happens in the binarized network, a unified scaling factor employed quantization leads considerable error and the network's performance deteriorates seriously. To solve this problem, the batch median of activations (BMA, Orange block in Figure 1) scheme is also proposed. It further maximizes the information entropy of activations in each layer.

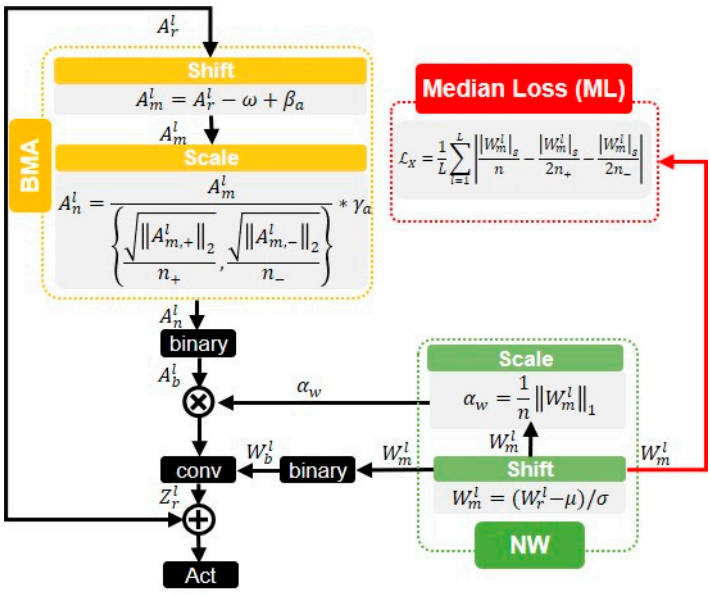

Figure 1. The basic blocks of binarization process in our work.

Our work has following two main contributions:

(1) From the perspective of entropy maximization, we propose a new regulation technique called the median loss for binary neural networks. This technique benefits for maintaining the upper-level value of information entropy, and leading a higher accuracy of the networks.

(2) To minimize the quantization error and avoid the network performance deterioration, we propose the BMA method to calculate the positive and negative scaling factors separately during the forward propagation's activations. At the same time, the method subtracts a median value from activations. This also helps to maximize the information entropy of activations in each layer.

This work takes the Bi-Real/ResNet-18 and ResNet-34 [23] on ImageNet 2012 as a baseline, our accuracy increases 1.3% and 0.9% respectively. It successfully proves that our methods are more effective in terms of improving the accuracy of binary neural networks. In addition, we also combine and apply the proposed methods into state-of-the-art BNN's binarization process. The final experiment results indicate that the whole networks' accuracy is improved accordingly, and validate the versatility of this work.

2. Preliminaries

2.1. Binarized Neural Networks

The Binarized Neural Networks (BNNs) has been firstly proposed in year 2016 [24]. After the proposal, it attracts a lot of attentions because its weights and activations are binarized. This can speed up the inference time and save considerable computation and memory footprint. The basic principles of BNNs can be presented in Equation (1):

$$a_b = sign(a_r) = \begin{cases} +1, \ if \ a_r > 0 \\ -1, \ if \ a_r \leq 0 \end{cases}, \quad w_b = sign(w_r) = \begin{cases} +1, \ if \ w_r > 0 \\ -1, \ if \ w_r \leq 0 \end{cases} \\ \alpha = \frac{\|w_r\|_1}{N_w}, \ \ \beta = \frac{\|a_r\|_1}{N_a} \tag{1}$$

$$z_r = (\alpha \cdot \beta) \cdot popcount(xnor(a_b, w_b)) \tag{2}$$

where a_r, w_b, z_r represent the full-precision input activations, weights and output activations, respectively. a_b, w_b represent the binarized activations and weights, respectively. $\alpha \in R^{c_o}$ and $\beta \in R^{c_i}$ denote the scaling factors of weights and input activations. In practice, *Sign* function is usually used to get the binarized weights and activations in Equation (1). Furthermore, in recent years of research, some novel binarization methods [18,19] have been proposed in order to obtain a higher accuracy. Finally, the output activations are obtained by a bitwise operation XNOR and Bitcount from the binarized weights and input activations. It can be formulated as Equation (2).

In the backward propagation, the derivative of Sign function is zero almost everywhere. This property makes the binarized models hard to be optimized. To handle this problem, a straight-through estimation (STE) [25] method is proposed and used to train the BNNs. It employs the Identity or Hardtanh function to propagate the gradients. Moreover, many previous works [26–28] further conduct to correct the backward gradient mismatch with improved static binarization functions which originate from the STE method.

2.2. Information Entropy

The Shannon information entropy is a milestone in the information theory development roadmap. The theory has firstly been proposed by C.E. Shannon in 1948 [29]. It is defined as an expected amount of information $H(X)$ in a random variable which can be formulated as Equation (3). Therefore, a system's information uncertainty can be mathematically quantified and have a precise value.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{3}$$

where x_i denotes the possible values of a random variable X, $p(x_i)$ denotes the probability that the random variable X takes the value x_i.

After and with the Shannon entropy proposal, the information theory has developed rapidly. Furthermore, various classical theories have also been proposed, including the principle of maximum entropy [30]. Under a premise of known partial knowledge, the most reasonable inference about the unknown distribution is the most uncertainty of a random inference in accordance with the already known part [31]. It is the essence of maximum entropy. For this reason, if a more uniform and unbiased input dataset can be achieved, a better generalization and higher entropy the set will attain. This is the inspiration and theoretical basis for our proposed approaches to implement a BNN networks with the maximum entropy principle in this article.

In past few years, there have been many excellent works [18,19] in applying the knowledge of information theory into the field of neural network. These works are dedicated to maximizing the entropy in the network. The stronger generalization ability the network has, the more effective for increasing the network entropy and performance improvement. Nevertheless, the existed present works only deal with the weights and excludes the factor of activations. In this paper, on the one hand, we propose a new regularization approach to maximize the network's weights entropy. On the other hand, a more evenly distributed activations has been achieved by an operation of subtracting its median value. This newly proposed process increases and maximizes the information entropy of the binarized activations.

3. Proposed Method

3.1. Median Loss (ML)

An important reason for the accuracy's deterioration of the binarization network is that, in the forward propagation, the binarization of weights and activations cause a large amount of information loss. From the perspective of information theory, the process of binarization leads to a decrease in entropy, thereby reduces the representation ability of network. To handle this problem, we propose the median loss as a new regularization approach to minimize information loss.

In the traditional binary network, both the activations and weights are quantified to $\{-1, +1\}$. Thus, the quantized values can be modeled by the Bernoulli distribution, which is formulated as Equation (4):

$$\begin{aligned} f(a_b) &= \begin{cases} p_a, & if\ a_b = +1 \\ 1 - p_a, & if\ a_b = -1 \end{cases} \\ f(w_b) &= \begin{cases} p_w, & if\ w_b = +1 \\ 1 - p_w, & if\ w_b = -1 \end{cases} \end{aligned} \tag{4}$$

where $p_a, p_w \in [0, 1]$ denotes the probability of taking the value +1. The entropy of this random variable can be calculated as Equation (5). In addition, through computing the derivative of the entropy as Equation (6), it can be obtained that when p equals 0.5, that is, the distribution is relatively even, the entropy is the largest.

$$\begin{aligned} H(a_b) &= -p_a \log_2 p_a - (1 - p_a) \log_2 (1 - p_a) \\ H(w_b) &= -p_w \log_2 p_w - (1 - p_w) \log_2 (1 - p_w) \end{aligned} \tag{5}$$

$$\begin{aligned} \frac{\partial H(a_b)}{\partial p_a} &= \log_2 \frac{1 - p_a}{p_a} \\ \frac{\partial H(w_b)}{\partial p_w} &= \log_2 \frac{1 - p_w}{p_w} \end{aligned} \tag{6}$$

In order to minimize the information loss in BNNs' forward propagation, many methods have been proposed to increase the information in Equation (5). IR-Net [18] is proposed to normalize the weights through the operation of Equation (7), so that the quantized values can be more evenly distributed and achieve a larger entropy. However,

this method ignores the most important problem that the binarized weights do not obey the normal distribution. The normalization operation of Equation (7) cannot make the values distributed evenly as expected. Furthermore, the IR-Net only consider maximizing information capacity of the weights and ignore the activations. In another related research, Dmitry Ignatov [19] also proposes a Shannon entropy-based information loss penalty rule to increase the networks' entropy. Nevertheless, this method still has no network activations part involvement, and it is too complicated to deploy in a real practice.

$$\widehat{w_{std}} = \frac{\hat{w}}{\sigma(\hat{w})}, \hat{w} = w - \overline{w} \tag{7}$$

In this work, a novel regularization technique called the median loss for BNNs is proposed. It can be formulated as:

$$\mathcal{L}_X = \frac{1}{L}\sum_{l=1}^{L}\left|\frac{\left|W^l\right|_s}{n^l} - \frac{\left|W_+^l\right|_s}{2n_+^l} - \frac{\left|W_-^l\right|_s}{2n_-^l}\right| \tag{8}$$

where W^l represents the floating-point weights in layer l. W_+^l and W_-^l represent the positive and negative values in W^l accordingly. n^l represents the number of weights in layer l, n_+^l and n_-^l represent the number of positive and negative weights. $|X|_s$ represents the sum of all elements in $X(X \in \left\{W^l, W_+^l, W_-^l\right\})$.

The information loss penalty is added to the overall loss function in a conventional way as indicated in Equation (9), and is used in the backward propagation.

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_X \tag{9}$$

where $\mathcal{L}_{cls}$ represents the original loss function, λ represents the regularization parameter.

The goal of our proposed median loss is to make the binarized weights distributed evenly, it can be proved as Equation (10). In this equation, $n_l > 0$, and $\left|W_+^l\right|_s > 0$, $\left|W_-^l\right|_s < 0$, from which it can be concluded that only when $n_+ = n_-$, the median loss takes the only minimum value. Therefore, the employed median loss can increase network's information entropy effectively. Moreover, compared to the Dmitry Ignatov's method [19], our proposed median loss is simpler and more intuitive to use without any loss in accuracy.

$$\begin{aligned}
\mathcal{L}_X &= \tfrac{1}{L}\sum_{l=1}^{L}\left|\tfrac{|W^l|_s}{n^l} - \tfrac{|W_+^l|_s}{2n_+^l} - \tfrac{|W_-^l|_s}{2n_-^l}\right| \\
&= \tfrac{1}{L}\sum_{l=1}^{L}\left|\tfrac{|W_+^l|_s}{n^l} - \tfrac{|W_+^l|_s}{2n_+^l} + \tfrac{|W_-^l|_s}{n^l} - \tfrac{|W_-^l|_s}{2n_-^l}\right| \\
&= \tfrac{1}{L}\sum_{l=1}^{L}\left|\tfrac{1}{n^l} \times \left(\left|W_+^l\right|_s - \tfrac{n^l \times |W_+^l|_s}{2n_+^l}\right) + \tfrac{1}{n^l} \times \left(\left|W_-^l\right|_s - \tfrac{n^l \times |W_-^l|_s}{2n_-^l}\right)\right| \\
&= \tfrac{1}{L}\sum_{l=1}^{L}\left|\tfrac{1}{n^l} \times \tfrac{(n_+^l - n_-^l) \times |W_+^l|_s}{2n_+^l} + \tfrac{1}{n^l} \times \tfrac{(n_-^l - n_+^l) \times |W_-^l|_s}{2n_-^l}\right| \\
&= \tfrac{1}{L}\sum_{l=1}^{L}\left|\tfrac{1}{2n^l} \times \left(n_+^l - n_-^l\right) \times \left(\tfrac{|W_+^l|_s}{n_+^l} - \tfrac{|W_-^l|_s}{n_-^l}\right)\right|
\end{aligned} \tag{10}$$

3.2. Batch Median of Activations (BMA)

Till now, present existed works are all to process the network's weights to achieve its maximum entropy. In order to obtain a greater entropy and reduce the quantization error in BNNs, another our method called the batch median of activations (BMA) is proposed in this work. To our knowledge, this is the first time to process the activations from an information perspective to improve the network entropy.

As shown in Equation (11), the median value ω_r^l is subtracted from the activations A_r^l. According to the median's mathematical definition, regardless of any distributed dataset, subtracting the median operation can make the number of positive/negative values equal. After this, the operated positive/negative elements inside the set are evenly parted. As presented in Section 3.1, the information entropy of these evenly distributed binarized values have the maximum information entropy.

Meanwhile, the improved entropy with an even distribution inevitably leads to effect the scaling factors and whole networks accuracy. In classic XNOR-Net [14], the factor is employed to reduce the quantization error and additionally improve the model's accuracy. Almost all binarized networks follow this idea, including the well-known IR-Net [18]. However, the IR-Net is committed to making the distribution of binarized weights evenly, and this naturally brings a large difference between the positive and negative amplitudes. In this way, employing a same scaling factor for the positive and negative activations causes a considerable quantization error. To tackle this problem, our proposed BMA calculates the scaling factors for positive and negative values separately as Equation (12) shown. Therefore, the information loss introduced by quantization error in forward propagation can be mitigated and further improve the network's accuracy.

$$A_m^l = A_r^l - \omega_r^l + \beta_\alpha \tag{11}$$

$$A_n^l = \frac{A_m^l}{\left\{ \frac{\sqrt{\|A_{m,+}^l\|_2}}{n_+^l}, \frac{\sqrt{\|A_{m,-}^l\|_2}}{n_-^l} \right\}} * \gamma_a \tag{12}$$

where A_r^l represents the floating-point activations in layer l. ω_r^l represents the mid-value of A_r^l. $A_{m,+}^l$ and $A_{m,-}^l$ represent the positive and negative values in A_m^l. n_+^l and n_-^l represent the number of positive and negative weights. $\|\cdot\|_2$ represents the L2 regularization. γ_a, β_a scales and biases the normalized value individually [32].

4. Experiments and Discussion

4.1. Experimental Details

We further investigate the network performance by utilizing proposed ML and BMA with XNOR-Net/ResNet-20 on CIFAR-10 dataset, separately. The experiments can help figure out and analyze how the ML and BMA methods work in practice. The model adopts the Kaiming initialization and trains from scratch. The training flow runs for 400 epochs with 128 batch-size. Optimization process applies the Stochastic Gradient Descent (SGD) optimizer with momentum = 0.9, weight decay = 10^{-4}, initial learning rate = 0.1, and cosine learning rate decay.

The training flow of BNNs using ML and BMA is illustrated in below Algorithm 1. Firstly, the forward propagation binarizes weights and input activations. Then, computes the output activations and output losses. The ML and BMA proposals are adopted in this stage. In order to remove the influence of updating weights on final results, the algorithm's backward propagation follows the method used in XNOR-Net [14]. Finally, gradients and weights are updated.

4.2. Ablation Studies

In this part, we conduct several experiments with binary network on the CIFAR-10 [33] and ImageNet 2012 datasets. The experimental results and analyses conclude the behaviors and effects of proposed ML and BMA techniques on the BNNs, and further testify the correctness of our theoretical analysis in previous Section 3.

Algorithm 1. Training Flow of the BNNs using Our-proposed ML and BMA.

Input: A mini-batch of inputs and targets.

01: Compute the binarized weights and input activations:
$A_b^l = sign(\text{BMA}(A_r^l))$, $W_b^l = sign((W_r^l - \mu(W_r^l))/\sigma(W_r^l))$

02: Compute the output activations:
$Z_r^l = \text{conv}(A_b^l, W_b^l)$, $A_r^{l+1} = \text{Act}(Z_r^l \times \alpha_w + A_r^l)$

03: Compute the output losses:
$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_X$, $\mathcal{L}_X = \frac{1}{L}\sum_{l=1}^{L}\left|\frac{|W^l|_s}{n} - \frac{|W_+^l|_s}{2n_+} - \frac{|W_-^l|_s}{2n_-}\right|$

04: Compute the gradients employing the method adopted in XNOR-Net [14]

05: Update the W:
$W = W - \eta\frac{\partial \mathcal{L}_{total}}{\partial W}$, where η is the learning rate.

4.2.1. Median Loss (ML)

When training a binarized model, we add the median loss regularization term on the basis of the existed loss function. More specific experimental protocol is to employ the median loss on the XNOR-Net/ResNet-20.

To verify the effectiveness of ML proposal, Figure 2 indicates the distribution of binarized weights of each layer in XNOR-Net/ResNet-20 without and with the ML respectively. The figure's results also show that ML makes the distribution more evenly. Additionally, the entropy value H of binarized BNN increases from 5.40 to 5.41 which means more information retains inside the networks with the ML employment.

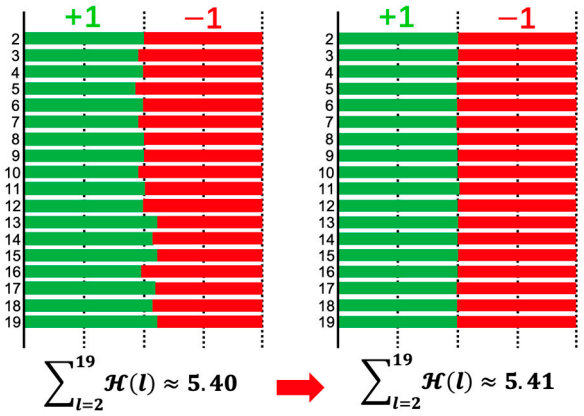

Figure 2. The distributions of network weights' binary representations (±95% confidence interval) without ML (**left**) and with ML (**right**) in binary pre-activation XNOR-Net/ResNet-20 are illustrated. The models are trained with $\lambda = 10^{-4}$ on the CIFAR-10 dataset.

Furthermore, the ML is more effective than other methods to increase the network entropy. Table 1 summarizes the experimental results obtained on CIFAR-10 dataset with XNOR-Net/ResNet-20. The networks are trained with the same configurations, and with different methods to increase its entropy. Method 1 employs the network weights' standardization and balance operations proposed in the IR-Net [18]. Meanwhile, Method 2 applies our proposed ML. From the table, our Method 2 proves an accuracy of 84.30% which is higher than the Method 1's 83.85% and referenced baseline's 80.33%. These comparisons prove that the ML outperforms the method of IR-Net and the baseline.

Table 1. Accuracy comparison with different methods adopted in XNOR-Net/ResNet-20 (Referenced Baseline) on CIFAR-10 database. The best results are shown in bold.

	Method 1 [18]	**Method 2 (ML)**	**Referenced Baseline [14]**
Accuracy (%)	83.85	**84.30**	80.33
(Mean + STD)%	0.32	**0.32**	0.39

4.2.2. Regularization Parameter λ

Aiming at determining an optimal value of the regularization parameter λ, a series of preliminary experiments are conducted. Their values vary from a collection $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. The accuracy of pre-activation XNOR-Net/ResNet-20 binary network is assessed on the validation subset of CIFAR-10 datasets. We repeat each experiment four times with a random weight initialization, and take the average of these results as the final accuracy. Table 2 summarizes the accuracy variations with different λ. The larger λ is, the more prior information in network exists. This leads the information entropy in weights increase and the accuracy results improve accordingly. Meanwhile, oversized λ makes the initial loss function almost ineffective and further results in a deviate from the basic task. In this way, the λ's ablation study concludes that 10^{-4} is a favorable value.

Table 2. Accuracy comparison with different methods adopted in XNOR-Net/ResNet-20 on CIFAR-10. The best results are shown in bold.

λ	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
Accuracy (%)	83.67	84.12	**84.30**	83.96	84.07
(Mean + STD)%	0.41	0.29	**0.28**	0.34	0.37

4.2.3. Batch Median of Activations (BMA)

To verify the BMA's effectiveness, we design two experimental sets. One is based on the conventional method, which uses a uniform scaling factor for all activations in the same layer. Another is our BMA, which calculates the scaling factor of each layer's positive and negative activations separately. Figure 3 shows the BMA minimizes quantization error in forward propagation effectively and decreases it from 0.59 to 0.02 obviously.

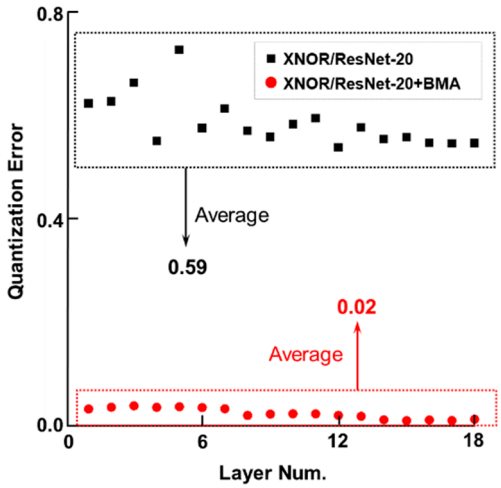

Figure 3. The quantization errors of network activations (±95% confidence interval) with BMA (red) and without BMA (black) in binary pre-activation XNOR-Net/ResNet-20 model are illustrated. The models are trained on the CIFAR-10 dataset.

In addition, Figure 4 illustrates the distribution of binary activations of each layer in XNOR-Net/ResNet-20 without and with the BMA, respectively. From the figure, it can be proved that the BMA makes a greatly positive influence on the distribution of binary activations. The entropy value of the activations increases from 5.25 to 5.41. This large increase means the information in activated network can be retained.

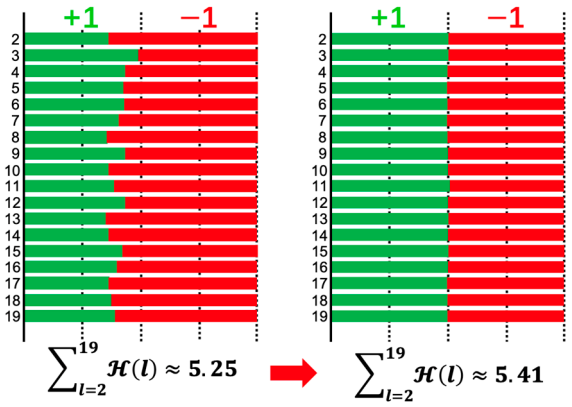

Figure 4. The distributions of network activations' binary representations (±95% confidence interval) without/with BMA (**left**/**right**) in binary pre-activation XNOR-Net/ResNet-20 are illustrated. The models are trained on the CIFAR-10 dataset.

4.2.4. Comparison with State-of-the-Art Methods

Firstly, we study the performance of ML and BMA upon ResNet-20 topology and compare with other state-of-the-art methods comprehensively. To verify an improved performance brought by the superimposed individual ML and BMA together, we add them both into the XNOR-Net simultaneously. As Figure 5 and Table 3 illustrated, the improved version with ML(XNOR+ML) outperforms the XNOR-Net without ML over 3.97% accuracy; The improved version with BMA(XNOR+BMA) outperforms the XNOR-Net without BMA over 2.98% accuracy; Moreover, the accuracy increases when the XNOR-Net adopts both ML and BMA together (XNOR+BMA+ML), and can be achieved as high as 4.67%.

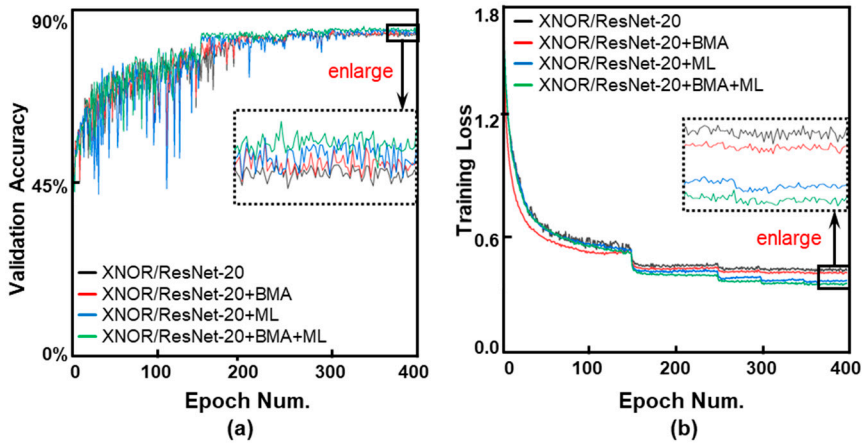

Figure 5. (**a**) Validation accuracy curves and (**b**) Training loss curves during the training process. Baselined XNOR-Net/ResNet-20, improved three schemes with BMA, ML, and both BMA+ML are evaluated and illustrated. All cases are with $\lambda = 10^{-4}$.

Table 3. Accuracy results of XNOR/ResNet-20 baseline and improved versions with our proposed ML, BMA, and both ML+BMA (all three cases are with $\lambda = 10^{-4}$). The best results are marked in bold.

Topology	Method	Bit-Width (W/A)	Accuracy (%)
ResNet-20	XNOR	1/1	80.33
	XNOR+ML	1/1	84.30
	XNOR+BMA	1/1	83.31
	XNOR+ML+BMA	**1/1**	**85.00**

Next, to verify our ML and BMA can be applied to other network structures and also be combined with different training methods, we extend the evaluations to a new

structure—Bi-Real18 [23]. In addition, we also adopt the training method EDE proposed in IR-Net. As illustrated in Figure 6 and Table 4, our ML and BMA version (IR-Net+BMA+ML) shows a 0.53% accuracy improvement compared to the IR-Net benchmark.

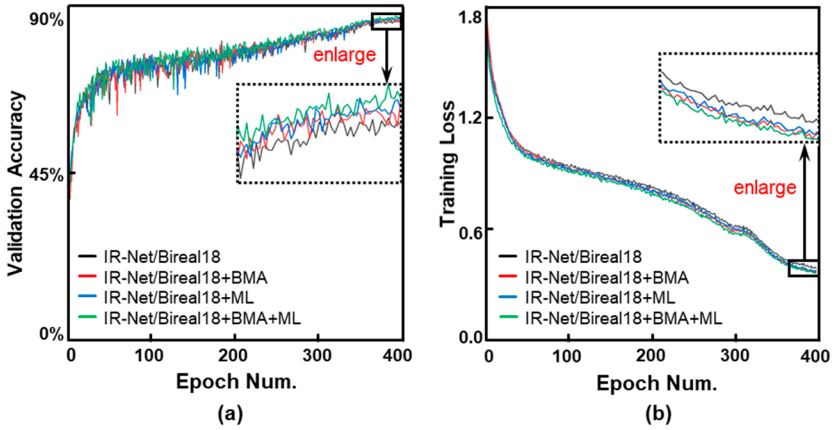

Figure 6. (**a**) Validation accuracy curves and (**b**) Training loss curves during the training process. Baselined IR-Net/Bireal18, improved three schemes with BMA, ML, and both BMA+ML are evaluated and illustrated. All cases are with $\lambda = 10^{-4}$.

Table 4. Accuracy results of IR-Net/BiReal-18 baseline and improved versions with our proposed ML, BMA, and ML+BMA both (all three cases are with $\lambda = 10^{-4}$). The best results are shown in bold.

Topology	Method	Bit-Width (W/A)	Accuracy (%)
BiReal-18	IR-Net	1/1	86.30
	IR-Net+ML	1/1	86.60
	IR-Net+BMA	1/1	86.49
	IR-Net+ML+BMA	**1/1**	**86.83**

Finally, we extend the evaluation to a larger scale image classification dataset—ImageNet 2012. The set contains 1.2 Mil. training and 50,000 validation samples individually. The training configurations are the same as Bi-Real18/CIFAR-10, except ResNet18/Bi-Real and ResNet34/Bi-Real are selected as the baseline. The training epoch is set 160, and input images are all cropped into a 224 $\times$ 224 resolution as references required. The regularization parameter λ utilize decided 10^{-4}. The model is trained on 4 Nvidia RTX2080Ti GPUs with a total batch size of 128. Table 5 lists the performance comparison of present mainly-utilized BNNs. When the Bi-real combined with ML and BMA (Bi-MB^3, Bi-Real+ML+BMA), their accuracies can be improved 1.3% and 0.9% for the ResNet-18 and ResNet-34 individually. Furthermore, our proposed methods are also testified on the IR-Net (Bi-IR-MB^4, Bi-Real+IR-Net+ML+BMA), the accuracies are improved 0.4% and 0.3% for the ResNet-18 and ResNet-34 accordingly. Table 5 also lists the performance of another state-of-the-art method—CI-BCNN [21] which mines the channel-wise interactions by a reinforcement learning model. The performance is quite close to the Bi-MB^3, and evidences the effectiveness of our prosed BMA and ML methods. Above results prove that our proposed methods are versatile and applicable. These methods can be embedded and utilized into the present popular BNN networks with a further accuracy improvement.

Table 5. Performance comparison with state-of-the-art (SOTA) methods on the ImageNet. The best results are highlighted in bold.

Topology	Method	Bit-Width (W/A)	Top-1 (%)	Top-5 (%)
ResNet-18	Floating Point	32/32	69.6	89.2
	ABC-Net	1/1	42.7	67.6
	XNOR	1/1	51.2	73.2
	Bi-Real	1/1	56.4	79.5
	CI-BCNN	1/1	56.7	80.1
	IR-Net	1/1	58.1	80.0
	Bi-M [1]	1/1	57.0	79.9
	Bi-B [2]	1/1	56.7	79.7
	Bi-MB [3]	**1/1**	**57.7**	**80.2**
	Bi-IR-MB [4]	**1/1**	**58.5**	**80.8**
ResNet-34	Floating Point	32/32	73.3	91.3
	ABC-Net	1/1	52.4	76.5
	Bi-Real	1/1	62.2	83.9
	CI-BCNN	1/1	62.4	84.8
	IR-Net	1/1	62.9	84.1
	BI-M [1]	1/1	62.7	84.1
	Bi-B [2]	1/1	62.4	83.9
	Bi-MB [3]	**1/1**	**63.1**	**84.3**
	Bi-IR-MB [4]	**1/1**	**63.2**	**84.3**

[1] Results of ResNet with Bi-Real stucture [23] and ML. [2] Results of ResNet with Bi-Real stucture and BMA. [3] Results of ResNet with Bi-Real stucture, ML and BMA. [4] Results of ResNet with Bi-Real stucture, IR-Net, ML and BMA.

4.2.5. Storage Cost and Calculation Complexity Analyses

The storage and computational complexity analyses of our methods in ResNet18/Bi-Real and ResNet34/Bi-Real are demonstrated. Floating-Point, XNOR, Bi-Real and proposed Bi-Real+ML+BMA four modes are compared and concluded in Table 6.

Table 6. Comparison of storage cost and calculation complexity upon ResNet18 and ResNet34. Our proposals are highlighted in bold.

Topology	Method	Storage Cost (Mbit)	FLOPs
ResNet-18	Floating-Point	374.1 Mbit	1.81×10^9
	XNOR	33.7 Mbit	1.67×10^8
	Bi-Real	33.6 Mbit	1.63×10^8
	Bi-Real+ML+BMA	**33.6 Mbit**	$\mathbf{1.63 \times 10^8}$
ResNet-34	Floating-Point	697.3 Mbit	3.66×10^9
	XNOR	43.9 Mbit	1.98×10^8
	Bi-Real	43.7 Mbit	1.93×10^8
	Bi-Real+ML+BMA	**43.7 Mbit**	$\mathbf{1.93 \times 10^8}$

Compared with the floating-point networks, our Bi-Real+ML+BMA saves the storage by 11.13×/15.96×, and speeds up the computation by 11.10×/18.96× upon the ResNet18 and ResNet34, respectively. This is quite close to the Bi-Real's performance.

Additionally, the performance of our proposed Bi-Real+ML+BMA and standard Bi-Real has also been calculated. In the storage cost aspect, our method only increases the different scaling factors (BMA) and normalization of weights. The exact increment values are 11.7 Kbit, 22.8 Kbit (ResNet18, ResNet34/Bi-Real+ML+BMA) respectively. These values are minor increments compared with the original Bi-Real's 33.6 Mbit, 43.7 Mbit (ResNet18, ResNet34). Simultaneously, from the computation cost point of view, only normalization operation of weights increases the cost. The exact values are 4.62 M FLOPs, 7.12 M FLOPs (ResNet18, ResNet34/Bi-Real+ML+BMA), respectively. These increments

are also negligible compared with the standard Bi-Real's 1.63×10^8, 1.93×10^8 (ResNet18, ResNet34).

As two major algorithm performance evaluation parameters—storage cost and calculation complexity concerned, extra expenditure brought by our proposed ML and BMA methods is minor and negligible. This greatly benefits for implementation and embedding this BNN network into a recourse-limited hardware platform.

5. Conclusions

In this article, we propose a novel regularization technique—median loss (ML)—to improve the binarized weights distribute more evenly, and further increase the entropy information left inside the network consequently. We also propose a new batch median of activations (BMA)method to retain more information of BNNs in two aspects. Firstly, the mid-value is subtracted from the activations to attain the maximum information entropy of binarized activations. Secondly, the scaling factor is calculated separately for the positive and negative of weights and activations to reduce the quantization error. Owing to the sufficient information retain inside network with these effective two methods, the XNOR-Net/ResNet-20 network gains a 4.67% accuracy on the CIFAR-10 dataset, and the Bi-Real18 network gains a 1.3% accuracy on the ImageNet 2012 dataset. Moreover, our proposed ML and BMA for the storage cost and calculation complexity increments are proved minor and negligible.

In this work, our proposed methods can increase the whole BNN network's information entropy and further accuracy improvement. What is more, the applicability into the present popular binary networks with a comprehensive good performance is also an attractive property. These advantages are greatly helpful and promising for its future wide applications into the binary neural networks, and additionally embedding the networks into the power-aware products.

Author Contributions: Methodology and initial idea, S.C., W.Z. and Y.Z. (Yi Zhan); software development and analysis, S.C., G.F. and W.Z.; supervision, S.C., Y.Z. (Yi Zhan); validation, L.W., Y.Z. (Yi Zhan) and W.Z.; data collection, G.F., L.W. and W.Z.; writing, W.Z., G.F., S.C. and Y.Z. (Yi Zhan); read and reviewed by D.S., Y.Z. (Yi Zhan) and Y.Z. (Yumei Zhou). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, under Grant No. 2019YFB2204601.

Data Availability Statement: Data can be provided upon request.

Acknowledgments: Our authors gratefully acknowledge the anonymous reviewers. Their valuable comments and suggestions are very helpful to improve the presentation of this paper and our future work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. Han, K.; Guo, J.; Zhang, C.; Zhu, M. Attribute-aware Attention Model for Fine-grained Representation Learning. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 2040–2048.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
4. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
5. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
6. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.

7. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <0.5 MB Model Size. *arXiv* **2016**, arXiv:1602.07360.
8. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
9. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
10. Srinivas, S.; Babu, R.V. Data-free Parameter Pruning for Deep Neural Networks. *arXiv* **2015**, arXiv:1507.06149.
11. Han, S.; Pool, J.; Tran, J.; Dally, W.J. Learning Both Weights and Connections for Efficient Neural Networks. *arXiv* **2015**, arXiv:1506.02626.
12. Chen, W.; Wilson, J.; Tyree, S.; Weinberger, K.; Chen, Y. Compressing Neural Networks with The Hashing Trick. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2285–2294.
13. Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Binarized Neural Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4114–4122.
14. Rastegari, M.; Ordonez, V.; Redmon, J.; Farhadi, A. XNOR-net: Imagenet Classification Using Binary Convolutional Neural Networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 525–542.
15. Li, F.; Zhang, B.; Liu, B. Ternary Weight Networks. *arXiv* **2016**, arXiv:1605.04711.
16. Zhou, A.; Yao, A.; Guo, Y.; Xu, L.; Chen, Y. Incremental Network Quantization: Towards Lossless CNNs with Low-precision Weights. *arXiv* **2017**, arXiv:1702.03044.
17. Zhou, S.; Wu, Y.; Ni, Z.; Zhou, X.; Wen, H.; Zou, Y. Dorefa-net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *arXiv* **2016**, arXiv:1606.06160.
18. Qin, H.; Gong, R.; Liu, X.; Shen, M.; Wei, Z.; Yu, F.; Song, J. Forward and Backward Information Retention for Accurate Binary Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2250–2259.
19. Ignatov, D.; Ignatov, A. Controlling Information Capacity of Binary Neural Network. *Pattern Recognit. Lett.* **2020**, *138*, 276–281. [CrossRef]
20. Qin, H.; Gong, R.; Liu, X.; Bai, X.; Song, J.; Sebe, N. Binary Neural Networks: A Survey. *Pattern Recognit.* **2020**, *105*, 107281. [CrossRef]
21. Wang, Z.; Lu, J.; Tao, C.; Zhou, J.; Tian, Q. Learning Channel-Wise Interactions for Binary Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 568–577.
22. Yin, P.; Zhang, S.; Lyu, J.; Osher, S.; Qi, Y.; Xin, J. Blended Coarse Gradient Descent for Full Quantization of Deep Neural Net-works. *Res. Math. Sci.* **2019**, *6*, 14. [CrossRef]
23. Liu, Z.; Wu, B.; Luo, W.; Yang, X.; Liu, W.; Cheng, K.-T. Bi-real net: Enhancing the Performance of 1-bit CNNs with Improved Representational Capability and Advanced Training Algorithm. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 722–737.
24. Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to + 1 or −1. *arXiv* **2016**, arXiv:1602.02830.
25. Bengio, Y.; Léonard, N.; Courville, A. Estimating or Propagating Gradients through Stochastic Neurons for Conditional Computation. *arXiv* **2013**, arXiv:1308.3432.
26. Darabi, S.; Belbahri, M.; Courbariaux, M.; Nia, V.P. BNN+: Improved Binary Network Training. In Proceedings of the Sixth International Conference on Learning Representations, Vancouver, BC, Canada, 29 April–3 May 2018; pp. 1–10.
27. Kim, H.; Kim, J.; Kim, J. Binary Duo: Reducing Gradient Mismatch in Binary Activation Network by Coupling Binary Activations. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020.
28. Liu, S.; Zhu, H. Binary Convolutional Neural Network with High Accuracy and Compression Rate. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 20–22 December 2019; pp. 43–48.
29. Shannon, C.E. A mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
30. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
31. Japkowicz, N.; Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*; Cambridge University Press: Cambridge, UK, 2011; pp. 42–72.
32. Loffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
33. Krizhevsky, A.; Nair, V.; Hinton, G.E. *The Cifar-10 Dataset*. 2014, p. 6. Available online: http://www.cs.toronto.edu/kriz/cifar.Html (accessed on 9 August 2021).