IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 January 2023, accepted 18 January 2023, date of publication 23 January 2023, date of current version 26 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3238715

RESEARCH ARTICLE

Binarized Neural Network With Parameterized Weight Clipping and Quantization Gap Minimization for Online Knowledge Distillation

JU YEON KANG[1], (Student Member, IEEE), CHANG HO RYU[2], (Student Member, IEEE), AND TAE HEE HAN[3], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea
[2]Department of Artificial Intelligence, Sungkyunkwan University, Suwon 16419, South Korea
[3]Department of Semiconductor Systems Engineering, Sungkyunkwan University, Suwon 16419, South Korea

Corresponding author: Tae Hee Han (than@skku.edu)

This work was supported in part by Samsung Electronics Company Ltd., under Grant IO201209-07877-01; and in part by the BK21 FOUR Project.

ABSTRACT As the applications for artificial intelligence are growing rapidly, numerous network compression algorithms have been developed to restrict computing resources such as smartphones, edge, and IoT devices. Knowledge distillation (KD) leverages soft labels derived from a teacher model to a less parameterized model achieving high accuracy with reduced computational burden. Moreover, online KD provides parallel computing through collaborative learning between teacher and student networks, thus enhancing the training speed. A binarized neural network (BNN) offers an intriguing opportunity to facilitate aggressive compression at the expense of drastically degraded accuracy. In this study, two performance improvements are proposed for online KD when a BNN is applied as a student network: 1) parameterized weight clipping (PWC) to reduce dead weights in the student network and 2) quantization gap-aware adaptive temperature scheduling between the teacher and student networks. In contrast to constant weight clipping (CWC), PWC demonstrates a 3.78% top-1 test accuracy enhancement with trainable weight clipping by decreasing the gradient mismatch with CIFAR-10 dataset. Furthermore, the quantization gap-aware temperature scheduling increases the top-1 test accuracy by 0.08% over online KD at a constant temperature. By aggregating both methodologies, the top-1 test accuracy for CIFAR-10 dataset was 94.60%, and that for Tiny-ImageNet dataset was comparable to that of the 32-bit full-precision neural network.

INDEX TERMS Neural network compression, knowledge distillation, binarized neural network, parameterized weight clipping, dead weight, adaptive temperature scheduling.

I. INTRODUCTION

Over the past decade, artificial neural network-based deep learning technology has been successfully applied in diverse fields. However, as networks become deeper and broader, real-world solutions require consideration of the computational cost. For example, a representative autoregressive language model, GPT-3 [1], increases the number of parameters to 175 billion, thereby significantly amplifying the computational burden. A variety of studies on neural network compression have been conducted with minimal performance degradation to alleviate these problems. Through aggressive reduction of parameters to a data width of 1-bit at the expense of considerable accuracy loss, binarized neural networks (BNNs) demonstrate significant benefits in terms of memory footprint and computational speed. Various studies have addressed the accuracy loss of BNNs, such as XNOR-Net and Bi-real [2], [3]. Nonetheless, there is still an inherent limit in improving BNN performance through parameter processing or modulation.

Knowledge distillation (KD) is a widely applicable technique for compressing neural networks [4]. The key

The associate editor coordinating the review of this manuscript and approving it for publication was Le Hoang Son.

idea behind KD is to supervise the student network by imitating the teacher network via soft probabilities, which exposes more information than the class label and helps the student network learn. KD performance is primarily determined by the different characteristics of the teacher and student networks, such as the data widths, topologies, and hyperparameter configuration. Denser knowledge can be acquired as the depth of the teacher network increases, whereas as a soft label approaches a hard label, it becomes too taxing for the student network to emulate the teacher network owing to insufficient capacity [5], [6]. A low-bit network has been applied to the student network to boost the KD compression efficiency [2], [7], [8]. However, in KD, where harmony between the teacher and student networks is emphasized, the quantization gap between the two networks causes negative side effects. Furthermore, the data width parameter is an essential consideration because the difference in performance between the two networks is closely related to the number of recognizable classes.

In this study, gradient mismatch mitigation in a BNN and a KD composed of a binarized student network are addressed. Conventional constant weight clipping (CWC) causes a gradient mismatch in BNNs because fixed clipping values cannot cope with the dynamics of weight distribution. Although the dynamic clipping range has been suggested as an alternative, it has enormous complexity. To address this problem, trainable clipping values used for easing gradient mismatches were introduced.

Next, we propose a method to alleviate the capacity shortage of binarized student networks caused by data width difference between the teacher and student networks. Numerous studies [2], [6], [7] have revealed that a difference between teacher and student networks greater than the effective range results in insufficient knowledge transfer between them. Hence, adaptive scheduling based on the quantization gap is required to balance the knowledge proportion of each network. Inspired by this observation, information entropy was developed to assess the difference between the two networks appropriately.

The contributions of this study are as follows:

- To mitigate the drawback of CWC that causes gradient mismatch in BNNs, we utilize a trainable weight clipping function adaptable to the dynamic weight distribution.
- In online distillation, information entropy-based temperature scheduling is introduced to overcome the problems caused by i) a poorly trained teacher network at the beginning of learning and ii) a capacity shortage of the student network.
- The effectiveness is revealed by aggregating the proposed approaches, which can be employed in the diverse network models. Furthermore, the binarized student network applied in the proposed methods exhibit a top-1 test accuracy comparable to that of the baseline CNN.

The remainder of this paper is organized as follows. The related studies are described in Section II. Section III presents the main algorithm flow of learnable weight clipping and details the manner in which information entropy is applied to the technique. The simulation results and an analysis of several network topologies and datasets are presented in Section IV. Finally, conclusions are drawn in Section V.

II. RELATED WORK

Various types of neural network compression have been proposed to compute resource-constrained device deployment. BNN and KD are representative network compression schemes with data width conversion and loss function reinforcement, respectively.

A. BINARIZED NEURAL NETWORK

Courbariaux et al. [9] exploited a straight-through estimator (STE) [10] as a gradient approximation to overcome zero gradients at all locations in the sign function. However, the expressive ability of BNNs in binary space is restricted, resulting in a significant loss of accuracy. To reduce the disparity in accuracy between a BNN and its single-precision 32-bit floating-point (FP32) counterpart, XNOR-Net [11] introduced a scaling factor derived from the L1-norm of the weights or activations to minimize the quantization error.

The academic community has extensively explored enhancements in the accuracy of BNNs by building gradient estimation functions or designing binarization-friendly network architectures. For example, various BNN schemes [12], [13], [14] have aimed to apply a continuous activation gradient that approximates the sign function to refine the existing STE. ABC-Net [15] was constructed by utilizing more binary bases for weighting and activation. Qin et al. [16] applied an error attenuation estimator to minimize backpropagation information loss on the gradient. Additionally, ReActNet [17] was applied to formulate an activation function that was translated to fit the weight distribution.

Moreover, several studies have focused on gradient improvement, which is used for predicting the variation and scale of weight parameters. In addition, Xu et al. [18] investigated the gradient mismatch problem of STE when used as a gradient approximation in BNNs. By standardizing dead weights, whose gradients were not defined by STE, the authors contributed to BNN performance. Liu et al. [19] revealed that the Adam optimizer is superior to other optimizers in BNNs. Dead weights were reactivated not only from the regularization effect of the second-order momentum in the Adam optimizer but also because dead weights decreased through weight decay.

STE is accountable for gradient approximation in backpropagation by providing a customized gradient to non-differentiable sign functions. However, the dead weight problems underlying gradient approximation are yet to be discussed. Thus, a weight clipping function is required to revive dead weights and simultaneously reduce the quantization error between FP32 and binarized weights.

B. KNOWLEDGE DISTILLATION

Low-precision numeric parameters and KD have common features that remarkably reduce computational requirements

TABLE 1. Accuracy (%) difference of XNOR-Net based on the weight clipping.

BNN methodology	Weight clipping	Model		
		VGG-small [24]	ResNet 20 [20]	WRN 22 × 4 [25]
XNOR-Net	None	87.38	81.90	88.52
	Constant (-1, W, 1)	90.88 (3.50↑)	86.93 (5.03↑)	89.08 (0.56↑)

and memory footprints. Because the two techniques are different, a cumulative effect is expected if they are applied in parallel. Usually, KD with a low-bit student network scheme focuses only on the layer depth disparity while neglecting the effect of a quantization gap between the two networks.

In [2], the accuracy of 2-bit ResNet 20 [20] was increased by 1.4% with joint training, mimicking the prediction probability of the teacher network on CIFAR-10 dataset. In addition, Cho et al. [6] enhanced the efficacy of KD by transferring amenable knowledge from early stopped teachers.

Shin et al. [7] emphasized the importance of a suitable teacher model and hyperparameter selection for optimizing the performance of a student network using KD; however, they did not address adaptive temperature scheduling. According to several recent studies [21], [22], [23], the use of adaptive temperature scheduling in online KD has the potential to achieve higher accuracy in student networks.

III. METHODOLOGY

First, parameterized weight clipping (PWC) is introduced to efficiently reduce dead weights in a BNN through gradient descent. In addition, information entropy-based temperature scheduling is proposed to alleviate the quantization gap between teacher and student networks for online KD.

A. PARAMETERIZED WEIGHT CLIPPING

CWC prevents a case in which the binary weights are not updated in backpropagation when the absolute value of FP32 activation is greater than one in the BNN. As shown in Table 1, we calculated the differences in accuracy between BNNs with and without weight clipping for various network models in PyTorch [26] to correctly determine the effect of weight clipping on the performance of BNNs. Although the increase in accuracy differed depending on the network topology and weight clipping, the accuracies of all three networks increased. In particular, ResNet 20 exhibited the highest increase in accuracy (5.03%).

$$\text{sign}(W) = \begin{cases} +1, & \text{if } W \geq 0 \\ -1, & \text{otherwise,} \end{cases} \tag{1}$$

$$\widehat{W} = \text{sign}(W) \cdot \frac{1}{n} \sum_{n} |W|, \tag{2}$$

$$\text{clip}\,(-1, W, 1) = \max(-1, \min(W, 1)). \tag{3}$$

In the BNN, the FP32 weight parameter set W was binarized using (1) and (2) for the forward propagation. Conversely, in the backpropagation, (3) was applied as an STE for a non-differentiable sign function. In backward propagation, if the FP32 weight exceeds the fixed clipping range, disagreement between the presumed and actual gradient functions occurs, resulting in dead weight. Dead weights hinder correct weight updates during backpropagation. To minimize the dead weights caused by the CWC, PWC with gradient approximation was applied, considering the minimized overhead. To equip learnable clipping values according to changes in weight, we allocated a gradient for the clipping functions α and β, as follows:

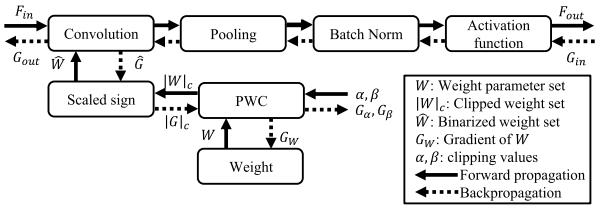

FIGURE 1. Graphical illustration of the training process with PWC.

$$G_W = \frac{\partial \mathcal{L}}{\partial |W|_c} \frac{\partial |W|_c}{\partial W},$$
$$\frac{\partial |W|_c}{\partial W} = \begin{cases} +1, & \text{if } \alpha < W < \beta \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

$$G_\alpha = \frac{\partial \mathcal{L}}{\partial |W|_c} \frac{\partial |W|_c}{\partial \alpha},$$
$$\frac{\partial |W|_c}{\partial \alpha} = \begin{cases} +1, & \text{if } W < \alpha \ (\alpha < 0) \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

$$G_\beta = \frac{\partial \mathcal{L}}{\partial |W|_c} \frac{\partial |W|_c}{\partial \beta},$$
$$\frac{\partial |W|_c}{\partial \beta} = \begin{cases} +1, & \text{if } W > \beta \ (\beta > 0) \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathcal{L}$ represents the loss function, and $\frac{\partial \mathcal{L}}{\partial |W|_c}$ represents the gradient from the deeper layer to the scaled sign function. Equations (4), (5), and (6) describe the approximated gradient equations for the clipping functions α and β. First, for the clipping function (4), based on a given weight, a value of 1 is returned if the weight exists between α and β. If the weight is greater than the negative clipping value α, it returns a value of $\frac{\partial \mathcal{L}}{\partial \alpha} = 1$, as shown (5). The gradient $\frac{\partial \mathcal{L}}{\partial \beta}$ for β can also be computed using the STE to estimate a value of 1 for $\frac{\partial \mathcal{L}}{\partial |W|_c}$ with (6). Consequently, gradient-descent-based training can adjust the clipping range to update the weights dynamically.

Because the weight values satisfy the range of $(-1, +1)$ through initialization, the default clipping values of $\alpha = -1$ and $\beta = 1$ include all weights within the clipping range. Each weight clipping value was adjusted from the initial value to narrow the range based on the PWC. Accordingly, the clipping range was modified for every training step to prevent the generation of dead weight. Backpropagation of the trainable clipping variables α and β was applied in the direction of the dashed arrow, as shown in Fig. 1.

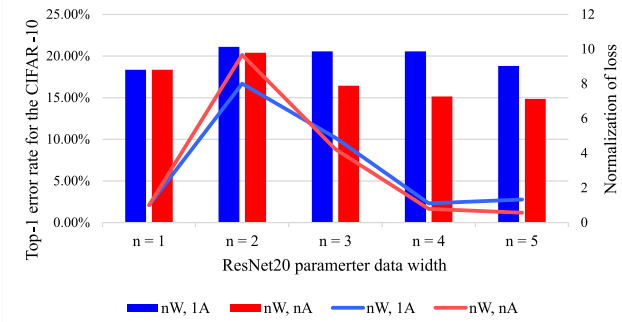

FIGURE 2. Top-1 error rate of binarized student network with increasing teacher weight (W) and activation (A) data width (300 epochs).

B. INFORMATION ENTROPY DISTANCE-BASED TEMPERATURE SCHEDULING

Fig. 2 shows the change in performance of the binarized student network as a function of the data width change in the teacher network. The blue bar and line represent the n-bit weight and 1-bit activation, respectively, and the red bar and line represent the n-bit weight and n-bit activation, respectively. It was observed that the increase in data width in the teacher network was not directly related to the increase in performance. Therefore, the differences in data width between teacher and student networks must be considered when optimizing KD.

$$Q_S^i = \frac{\exp(\frac{Z_S^i}{\tau})}{\sum_j \exp(\frac{Z_S^j}{\tau})}, \tag{7}$$

$$\mathcal{L}_{KD} = S_f \mathcal{L}(Q_S, y) + (1 - S_f) KLdiv(Q_S^i, Q_T^i). \tag{8}$$

In KD, (7) is used for the output layer that generates the soft logits z_T^i and z_S^i for the teacher and student networks, respectively, and regularizes the probability of each class according to the hyperparameter τ, which denotes the temperature. The loss function with the scaling factor S_f of the student network is given by (8), which includes the Kullback-Leibler divergence ($KLdiv$) between student and teacher distributions. However, considering student capacity, teacher knowledge close to the hard label caused by low temperatures can be overloaded. In addition, a strict teacher is required to maximize the effect of KD on the student network [4]. Therefore, the class classification of poorly trained teacher networks in the initial stage of learning can negatively affect the training process of the student network.

Fig. 3 illustrates the effect on the loss value of the student networks depending on the temperature during learning. The student network imitates the knowledge that changes from a hard label to a soft label as the temperature gradually increases, as shown in Fig. 3(a). In contrast, Fig. 3(b) shows the loss in the student network, which indicates the knowledge that changes from soft labels to hard labels using gradually decreasing temperatures. Accordingly, gradually decreasing the temperature resulted in a 12.85% lesser loss than gradually increasing the temperature. Therefore, in online KD, a soft label (i.e., probability smoothing) should

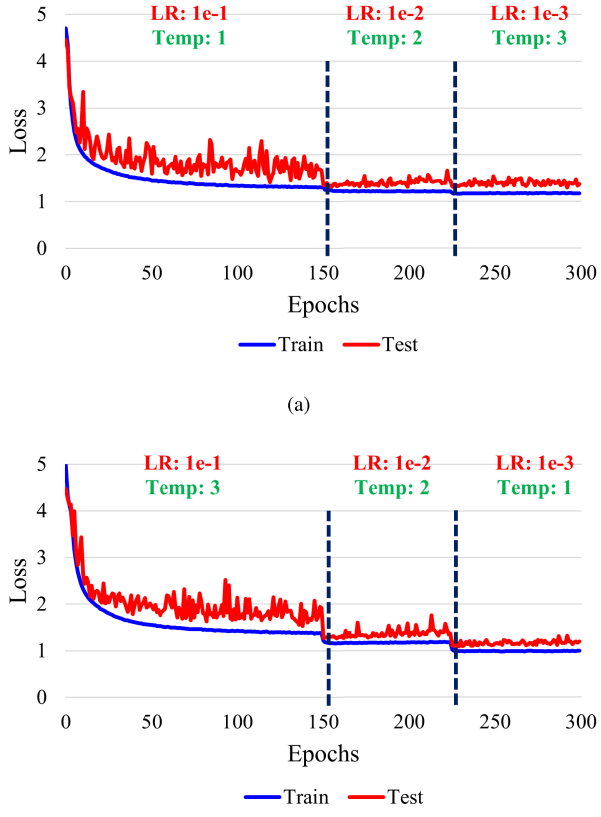

FIGURE 3. Change in the student network loss value based on temperature (Temp) change at 150 and 225 epochs with learning rate (LR): gradually (a) increasing and (b) decreasing temperatures.

be actively adopted in the early stages of learning. In contrast, in the latter half of learning, where the performance of the teacher network is guaranteed, knowledge close to the hard label should be provided at a low temperature.

As previously mentioned, it is difficult for the teacher network to predict the correct class during the early stage of online KD learning. The variation in learning speed depending on the quantization gap between the teacher and student is shown in Fig. 3. Therefore, it is preferable to use a temperature scheduling technique that reflects the performance variation between the two networks instead of using a constant temperature for the entire learning process.

As learning progresses, the student network requires more accurate hard label knowledge. Thus, as illustrated in Fig. 3(b), the temperature should be gradually decreased to cope with the hard label. Therefore, we chose an information entropy distance that can measure the amount of information in the network while gradually decreasing. Using a low temperature at the beginning of the training interferes with the student network training owing to the knowledge of a poorly trained teacher network. Conversely, a high temperature cannot completely mimic the encyclopedic knowledge of the teacher network in the late stages of learning. The information entropies of the two networks were calculated using (9) and (10), where the sets T and S are the outputs of branched

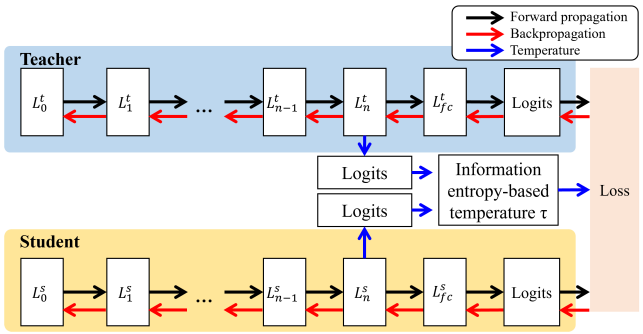

FIGURE 4. Conceptual diagram of the proposed information entropy distance-based temperature scheduling. The information entropies of both networks were calculated based on a convolution layer with maximum complexity, and the temperature was determined based on the distance between two values.

SoftMax in the teacher and student networks, respectively, with the convolution layer containing the most significant number of channels, as shown in Fig. 4. The distance between the two information entropies was calculated using (11).

Hence, adaptive temperature scheduling based on the performance difference between the two networks was formulated, as shown in (12), with the normalized factor λ by involving $D_{distance}$.

$$H(T) = -\sum_{c=0}^{C}\sum_{h=0}^{H}\sum_{w=0}^{W} t_{c\times h\times w}\log t_{c\times h\times w}, T \in t_{c\times h\times w}, \tag{9}$$

$$H(S) = -\sum_{c=0}^{C}\sum_{h=0}^{H}\sum_{w=0}^{W} s_{c\times h\times w}\log s_{c\times h\times w}, S \in s_{c\times h\times w}, \tag{10}$$

$$D_{distance} = \|H(T) - H(S)\|_2, \tag{11}$$

$$T_{AS} = \mathcal{L}_{KD} + \lambda \cdot D_{distance}. \tag{12}$$

To summarize, two techniques were developed for online KD, which comprises a binarized student network. First, the PWC lessens the dead weight problem of the CWC in backward propagation. Moreover, considering the characteristics of online KD, students learn the prediction probability of a poorly trained teacher at the beginning of the training process by implementing a soft label with a high temperature. Conversely, when a well-trained teacher is ready, temperature scheduling increases student performance through hard labels with a reliable teacher prediction probability at a low temperature.

The overheads of the PWC are the added gradient values corresponding to α and β of every layer, except for the first and last layers, with l representing the number of layers. In addition, the required clipping values are expressed as $2 \cdot (l-2)$. Taking ResNet 20 as an example, two clipping values per layer are required for 18 of the layers. Thus, only 36 parameters are added, for a total of 0.27M parameters.

The pseudocode for binarized student network training, which includes PWC and temperature scheduling, is described in Algorithm 1.

Algorithm 1 Training Binarized Student Network With Online KD Using PWC and Temperature Scheduling

Input: M_T, teacher network; M_S, student network; W, weight set; $|W|_c$, clipped weight set; $\widehat{W}$, binarized weight set; F' output of convolution; e, number of iterations; γ_e, learning rate; α/β, parameterized clipping value.

Output: M_S, trained binarized model.

for $e \leftarrow 0$, iterations **do**
  (1) forward computation
  Run forward computation of M_T, M_S simultaneously.
  $|W|_c = \text{clip}(\alpha, W, \beta)$.
  $\widehat{W} = \text{sign}(|W|_c) \cdot \frac{1}{n}\sum_n \big||W|_c\big|$.
  Calculate $F' = \widehat{W} \cdot \widehat{F}_{in}$.
  (2) backward and gradient computation.
  Compute information entropy distance between teacher and student using Eq. (11).
  Compute distillation loss $\mathcal{L}_{KD}$.
  Calculate temperature using Eq. (12).
  Run backward and compute gradients.
  Calculate $\frac{\partial \mathcal{L}}{\partial \widehat{W}}$ using $\frac{\partial \mathcal{L}}{\partial F_{out}}$.
  Calculate $\frac{\partial \mathcal{L}}{\partial |W|_c}$ using $\frac{\partial \mathcal{L}}{\partial \widehat{W}}$.
  Calculate $\frac{\partial \mathcal{L}}{\partial W}$ using $\frac{\partial \mathcal{L}}{\partial |W|_c}$.
  Calculate $\frac{\partial \mathcal{L}}{\partial \alpha}$ using $\frac{\partial \mathcal{L}}{\partial |W|_c}$.
  Calculate $\frac{\partial \mathcal{L}}{\partial \beta}$ using $\frac{\partial \mathcal{L}}{\partial |W|_c}$.
  $W_{e+1} \leftarrow W_e - \gamma_e \cdot \frac{\partial \mathcal{L}}{\partial W}$,
  $\alpha_{e+1} \leftarrow \alpha_e - \gamma_e \cdot \frac{\partial \mathcal{L}}{\partial \alpha}$,
  $\beta_{e+1} \leftarrow \beta_e - \gamma_e \cdot \frac{\partial \mathcal{L}}{\partial \beta}$.
end for
return trained binarized student model M_S

IV. EVALUATION

The benefits of parameterized weight clipping and information entropy distance-based temperature scheduling were validated by independently estimating and jointly evaluating the overall increase in accuracy compared with the various clipping functions and KDs. Furthermore, Table 2 presents the top-1 accuracy for the baseline network on the CIFAR-10 to further clarify the performance enhancement brought about by both proposed schemes.

A. EXPERIMENTAL SETUP

To analyze the PWC and distance-based temperature scheduling performance, we constructed an experimental environment using Pytorch 1.3.1, CUDA 10.2, and CUDNN 7.6.5 with multiple NVIDIA TITAN Xp (Pascal) GPUs and an Intel Xeon E5-1650 CPU. In addition, CIFAR-10, CIFAR-100 and, Tiny-ImageNet datasets [27], [28] were used to compare their performances, and the well-known BNN methodologies XNOR-Net [11] and Bi-Real Net [3] were utilized.

The hyperparameters underwent a total of 300 epochs with a weight decay of 1e-4 and learning rates of 1e-1, 1e-2, and 1e-3 for the 1st, 150th, and 225th epochs, respectively.

TABLE 2. Top-1 test accuracy (%) of the baseline network on CIFAR-10 dataset.

Model	Bit width (weight/activation)	
	32/32	1/1 (XNOR-Net)
VGG-small	93.80	87.38
ResNet 20	92.62	81.90
WRN 22 × 4	95.75	88.52

TABLE 3. Top-1 test accuracy (%) of XNOR-Net and Bi-Real Net for CIFAR-10 dataset with various weight clipping.

BNN methodology	Model	Weight clipping		
		NWC	CWC	PWC
XNOR-Net	VGG-small	87.38	90.88	92.44
	ResNet 20	81.90	86.93	87.05
	WRN 22 × 4	88.52	89.08	92.86
Bi-Real Net	VGG-small	90.48	90.49	91.82
	ResNet 20	82.74	82.75	84.59
	WRN 22 × 4	89.30	89.14	92.10

TABLE 4. Top-1 test accuracy (%) of each network model with several fixed temperature values and proposed temperature scheduling on CIFAR-10 dataset.

Model	Temperature τ			
	$\tau = 1$	$\tau = 2$	$\tau = 3$	Information entropy distance-based (proposed)
VGG-small	91.42	91.30	91.50	94.01
ResNet 20	86.41	86.28	86.97	87.33
WRN 22 × 4	94.22	94.22	94.48	94.56

B. WEIGHT CLIPPING COMPARISON

Table 3 lists the top-1 accuracy using weight clipping for each network model on CIFAR-10 dataset. No-weight clipping (NWC) indicates that no weight clipping was applied prior to binarization, and CWC, with a range of $(-1, +1)$, was used as a clipping value in the existing XNOR-Net.

For the PWC, the weight clipping value was adjusted based on the gradient descent training. The positive clipping value β was updated by the corresponding gradients from 1.28 to 0.5 to decrease the dead weight as shown in Fig. 5. Overall, the PWC improved the accuracy of all network models; in particular, the accuracy of WRN 22 × 4 increased by 3.78% compared with CWC.

C. KD TEMPERATURE SCHEDULING

The information entropy distance was used to determine the difference between the teacher and student networks for temperature scheduling. First, the temperature change based on λ was checked to match the different scales of loss and distance, based on (12). As shown in Fig. 6, when λ was fixed to 1, the temperature was configured from three to one. Based on this λ value, in the temperature scheduling experiment employing CWC, WRN 22 × 4 exhibited an accuracy of 94.56%, which is an improvement of up to 2.51% over VGG-small in comparison with $\tau = 3$, as presented in Table 4.

D. COMPARISON WITH SOTA METHODS

Knowledge transfer to quantized (particularly 1-bit CNN) networks from networks composed of FP32 weights and activations has rarely been explored in previous KD methods.

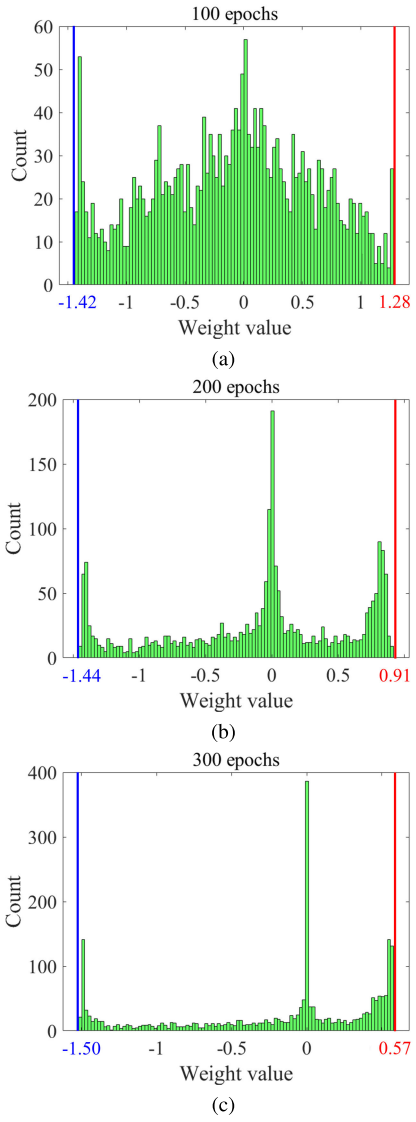

FIGURE 5. Weight distribution on layer 4 using the binarized ResNet 20 with PWC: (a) 100, (b) 200, and (c) 300 epochs. Trained negative clipping value α (blue line) and trained positive clipping value β (red line).

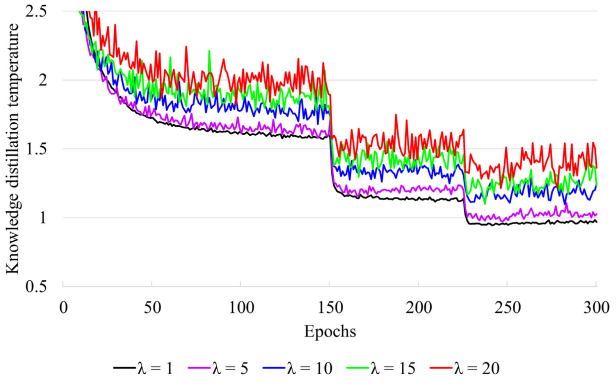

FIGURE 6. Temperature changes based on scaling factor λ in ResNet 20 on CIFAR-10 dataset.

Therefore, KD for a quantized neural network was chosen as a counterpart in this experiment to compare the KD

TABLE 5. Experimental results using CIFAR-10 and CIFAR-100 datasets.

CIFAR-10	Teacher network (FP32)		Student network	
	Model	Top-1 test accuracy (%)	Model	Top-1 test accuracy (%)
QDistill [8]	WRN 28 × 20	95.70	WRN 22 × 16 (2-bit)	94.20
Apprentice [2]	ResNet 44	93.80	ResNet 32 (2-bit)	92.60
GSLR [7]	WRN 20 × 1.5	93.50	ResNet 20 (1-bit)	91.30
Proposed	WRN 20 × 4	94.98	WRN 20 × 4 (1-bit)	94.60
	ResNet 20	92.11	ResNet 20 (1-bit)	87.46
CIFAR-100	**Teacher network (FP32)**		**Student network**	
	Model	Top-1 test accuracy (%)	Model	Top-1 test accuracy (%)
QDistill [8]	WRN 28 × 10	77.20	WRN 22 × 8 (2-bit)	49.30
GSLR [7]	WRN 20 × 1.7	72.20	ResNet 20 (2-bit)	67.00
Proposed	WRN 20 × 4	77.92	WRN 20 × 4 (1-bit)	76.89

TABLE 6. Top-1 test accuracy of ResNet18 on Tiny-ImageNet dataset.

Method	Bit-precision		Top-1 test accuracy (%)
Baseline	FP32		64.59
XNOR-Net	1-bit		55.80
QDistill [8]	Teacher	FP32	64.59
	Student	4-bit	61.17
Proposed	Teacher	FP32	65.29
	Student	1-bit	64.58

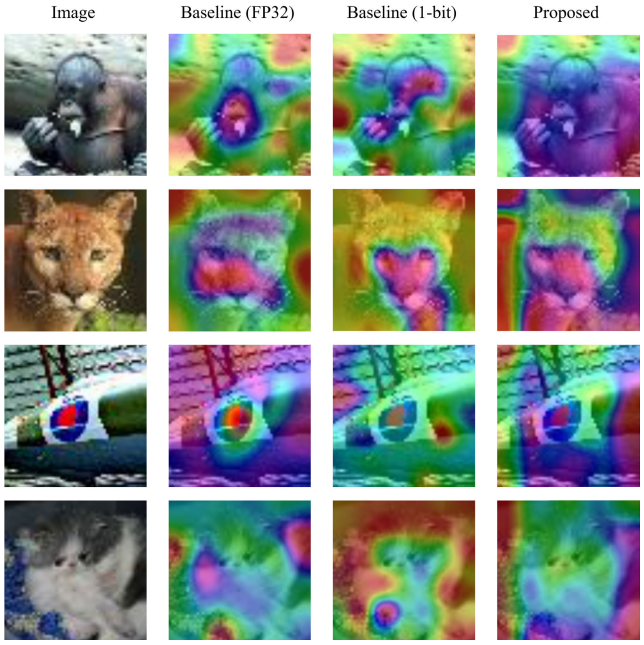

FIGURE 7. Attention maps for FP32 ResNet 18, 1-bit ResNet 18, and 1-bit ResNet 18 with the proposed method transferred by the knowledge of the teacher network on Tiny-ImageNet dataset.

performance for the quantization gap between the two networks. Table 5 presents the experimental results for CIFAR-10 and CIFAR-100 using the proposed method integrated with PWC and information entropy distance-based temperature scheduling. In the counterparts, the student network was configured using a 2-bit neural network. However, we obtained strength in deploying a binarized student network that achieved superior accuracy to the state-of-the-art KD for quantized deep neural networks. Specifically, our strategy when applied on CIFAR-100 dataset surpassed the 2-bit student network with a 9.89% improvement in accuracy. In Tiny-ImageNet experiment, the binarized student network significantly outperformed the BNN trained alone and showed comparable accuracy to the FP32 teacher network composed ResNet 18 network model, as shown in Table 6. This is because the hard labels were reflected in the temperature scheduling.

Information entropy-based temperature scheduling applied to online distillation shows a relatively faster training speed than its counterparts of offline distillation [2], [7], [8], composed of a two-stage training process. Even though, compared with the baseline online distillation, the computation overhead for the information entropy-based temperature in ResNet 20 is only 0.26%. While ResNet 110 has an overhead of 0.07% because this overhead decrease as the network depth are deeper.

To visualize the performance improvement for the disparity between the baseline (FP32 and 1-bit) and proposed techniques, attention maps were depicted for the qualitative results, as shown in Fig. 7. In the attention maps, a closer red value indicates a weight concentration in the network. 1-bit ResNet 18 with the proposed method is more clearly classified than the 1-bit baseline, and it can be seen that the performance for some images matched that for the FP32 baseline.

V. CONCLUSION

KD achieves high accuracy with a relaxed network depth by using soft labels derived from a teacher model for a less parameterized model. In contrast, a BNN can achieve a high compression rate by incorporating an aggressive reduction in the data width; however, it has an adverse effect on the accuracy. This study developed techniques to enhance the accuracy of online KD by using a BNN as a student network. Specifically, a PWC was applied to diminish the dead weights missing the gradient, and a temperature scheduling method was proposed to assess the quantization gap between the teacher and student networks. Consequently, for CIFAR-100 dataset, the accuracy of our technique increased by 9.89% in comparison with offline 2-bit student KD.

BNN can be advantageous in mobile and edge devices with resources constrained where energy efficiency is the primary concern. However, low capacity and performance originating from binarization lets BNN have challenges for application in a wide range. Therefore, further investigation on BNN includes more challenging applications (complex vision tasks such as object detection and unsupervised learning).

REFERENCES

[1] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
[2] A. Mishra and D. Marr, "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy," 2017, *arXiv:1711.05852*.
[3] Z. Liu, B. Wu, W. Luo, X. Yang, W. Liu, and K.-T. Cheng, "Bi-real net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 722–737.
[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[5] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 5191–5198.
[6] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4794–4802.
[7] S. Shin, Y. Boo, and W. Sung, "Knowledge distillation for optimization of quantized deep neural networks," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Oct. 2020, pp. 1–6.
[8] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," 2018, *arXiv:1802.05668*.
[9] 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," 2016, *arXiv:1602.02830*.
[10] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.
[11] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Computer Vision—ECCV*, 2016, pp. 525–542.
[12] R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan, "Differentiable soft quantization: Bridging full-precision and low-bit neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4852–4861.
[13] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X.-S. Hua, "Quantization networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7308–7316.
[14] M. Lin, R. Ji, Z. Xu, B. Zhang, Y. Wang, Y. Wu, F. Huang, and C.-W. Lin, "Rotated binary neural network," 2020, *arXiv:2009.13055*.
[15] X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," 2017, *arXiv:1711.11294*.
[16] H. Qin, R. Gong, X. Liu, M. Shen, Z. Wei, F. Yu, and J. Song, "Forward and backward information retention for accurate binary neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2250–2259.
[17] Z. Liu, Z. Shen, M. Savvides, and K. Cheng, "ReActNet: Towards precise binary neural network with generalized activation functions," in *Proc. ECCV*, 2020, pp. 2980–2988.
[18] Z. Xu, M. Lin, J. Liu, J. Chen, L. Shao, Y. Gao, Y. Tian, and R. Ji, "ReCU: Reviving the dead weights in binary neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5198–5208.
[19] Z. Liu, Z. Shen, S. Li, K. Helwegen, D. Huang, and K.-T. Cheng, "How do Adam and training strategies help BNNs optimization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6936–6946.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[21] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.
[22] X. Zhu and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[23] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 3430–3437.
[24] Z. Cai, X. He, J. Sun, and N. Vasconcelos, "Deep learning with low precision by half-wave Gaussian quantization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5918–5926.
[25] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[26] A. Paszke et al., "Automatic differentiation in pytorch," 2017. [Online]. Available: https://github.com/pytorch/pytorch/blob/master/CITATION
[27] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.
[28] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," 2015. [Online]. Available: http://tiny-imagenet.herokuapp.com

JU YEON KANG (Student Member, IEEE) received the B.S. degree in electronic engineering from the Tech University of Korea, Siheung, South Korea, in 2016. He is currently pursuing the M.S. and Ph.D. degrees in electrical and computer engineering with Sungkyunkwan University, Suwon, South Korea. His research interests include artificial intelligence, machine learning, and computer architecture.

CHANG HO RYU (Student Member, IEEE) received the B.S. degree in electronic engineering from Korea Aerospace University, Goyang, South Korea, in 2022. He is currently pursuing the M.S. and Ph.D. degrees in artificial intelligence with Sungkyunkwan University, Suwon, South Korea. His research interests include machine learning and computer architecture.

TAE HEE HAN (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1992, 1994, and 1999, respectively. From 1999 to 2006, he was with the Telecom Research and Development Center, Samsung Electronics, where he developed 3G wireless, mobile TV, and mobile WiMax handset chipsets. From 2011 to 2013, he worked as a full-time Advisor on system ICs at Korean Government. Since March 2008, he has been with Sungkyunkwan University, Suwon, South Korea, as a Professor. His current research interests include SoC/chiplet architectures for AI, advanced memory architecture, network-on-chip, and system-level design methodologies.