"Ghost" and Attention in Binary Neural Network

RUIMIN SUN¹,2, WANBING ZOU¹,2, and YI ZHAN¹
¹Institute of Microelectronics of Chinese Academy of Sciences, Chaoyang, Beijing 100029, China
²University of Chinese Academy of Science, Shijingshan, Beijing, 100049, China

Corresponding author: Yi Zhan (yizhan@ime.ac.cn).

This work was supported in part by the National Key R&D Program of China, under Grant No.2019YFB2204601.

ABSTRACT As the memory footprint requirement and computational scale concerned, the light-weighted Binary Neural Networks (BNNs) have great advantages in limited-resources platforms, such as AIoT (Artificial Intelligence in Internet of Things) edge terminals, wearable and portable devices, etc. However, the binarization process naturally brings considerable information losses and further deteriorates the accuracy. In this article, three aspects are introduced to better the binarized ReActNet accuracy performance with a more low-complex computation. Firstly, an improved Binarized Ghost Module (BGM) for the ReActNet is proposed to increase the feature maps information. At the same time, the computational scale of this structure is still kept at a very low level. Secondly, we propose a new Label-aware Loss Function (LLF) in the penultimate layer as a supervisor which takes the label information into consideration. This auxiliary loss function makes each category’s feature vectors more separate, and improve the final fully-connected layer’s classification accuracy accordingly. Thirdly, the Normalization-based Attention Module (NAM) method is adopted to regulate the activation flow. The module helps to avoid the gradient saturation problem. With these three approaches, our improved binarized network outperforms the other state-of-the-art methods. It can achieve 71.4% Top-1 accuracy on the ImageNet and 86.45% accuracy on the CIFAR-10 respectively. Meanwhile, its computational scale OPs is the least 0.86×10⁸ compared with the other mainstream BNN models. The experimental results prove the effectiveness of our proposals, and the study is very helpful and promising for the future low-power hardware implementations.

INDEX TERMS binary neural network, ghost feature map, attention, category information.

I. INTRODUCTION

With the booming Artificial Intelligence (AI) applications in image classification [1], target detection [2], and automatic driving [3], etc., industry and academy also pay attention to deploy the AI into the power-aware AIoT terminals, wearable and portable computing devices. However, well-performed Convolution Neural Networks (CNNs) architecture such as [1,2,3] needs tremendous computing power. Those computational and network’s parameter scales are difficult for the CNNs network to be implemented into a low-power platform, especially in resource constrained condition.

In recent years, many researches [4, 5, 18, 19, 21] adopt the Binarized Neural Networks (BNNs) to solve the problem. The weights and activations of network are quantized into single 1-bit, and original conventional convolutions can be simplified with the XNOR and Bit-Count operations. Consequently, the network’s computational complexity and memory access are obviously decreased from, ~36.6 × 10⁶ OPs to ~1.96 × 10⁶ OPs and ~4 MB to ~87 MB, separately [21]. Such a contracted model will be more convenient to embedded in hardware accelerator [27]. Meanwhile, the binarization process also brings great amount information losses and further accuracy deterioration new question.

Therefore, exploring a well-performed BNNs is an interesting and challenging work. Among them, the binarized single-bit MobileNet [16] is a suitable candidate structure. Especially, the low-complex ReActNet [6] developed from the MobileNet has achieved a remarkably satisfying 69.4% recognition accuracy on the ImageNet database. This advantage makes us to employ the ReActNet structure as our research baseline.

The ReActNet architecture duplicates the input channels and concatenates those two same-input blocks, and doubles the number of channels consequently. As Fig. 1 illustrated, the duplication brings multiple pairs of replicated feature maps, i.e. “ghost” feature maps. Although abundant and even redundant feature maps could be an essential characteristic for a more high-performed network, there still exists a problem of repeated calculation. Therefore, we propose a more cost-
effective way which is inspired by the GhostNet [7] to achieve richer information of the image’s feature maps.

![Feature Maps](image)

**FIGURE 1.** Three pairs of feature maps are in two blocks from down-sampling ReActNet architecture before a concatenation (Each up and down two images are one pair). The red and green two blocks are outputs from two different convolution calculation. The left single image is the original input.

In addition, from perspective of machine learning classification, the activation distribution of penultimate layer plays an important role on the network classification results. Reference [26] proposes label-aware representation approximation regularization, which aims to add real-valued auxiliary variable W into computation graph and allows representations of W/B to present a similar distribution pattern. In this article, our proposed Label-aware Loss Function (LLF) is employed to modify the activation distribution, and afterwards makes the same-labelled image’s feature vectors inside the penultimate layer to distribute more centralized. This improvement greatly benefits for the upcoming final fully-connected layer to achieve a better classification accuracy performance.

Furthermore, binarization of both weights and activations into $\{-1, +1\}$ normally brings a sharp accuracy decline and difficulty in training. Particularly, [25] designs a rectified clamp unit (ReCU), which changes the threshold to push “dead” weights towards the distribution peak in order to increase the probability of updating these values. [8] concludes that degeneration, saturation and gradient mismatch problems caused by irregular activation flow are the bottlenecks during the training stage. [9] proposes the Error Decay Estimator (EDE) to keep the gradient information to solve the backpropagation’s gradient mismatch problem. To tackle the unregulated activation problem, [8] proposes three loss functions as regularization to adjust the activation distribution. However, it treats the three loss functions equally, i.e. using the same hyper-parameter for them, the saturation problem is not well solved in those previous works. Therefore, in this article, a method by leveraging the normalization-based attention to rehabilitate distribution of pre-activation is adopted to handle the saturation problem.

After employing these three methods, our proposed low-complex network can achieve 71.4% Top-1 accuracy on the ImageNet and 86.45% accuracy on the CIFAR-10, respectively.

Our contributions are summarized as follows:
- We introduce an improved binarized ghost module to BNNs and achieve a more low-complex network. Rich feature maps information can be obtained in a more light-weighted way and further improve the classification accuracy of the network. These advantages are benefit for embedding our improved BNN network into the power-aware applications.
- A label-aware loss function to reduce the intra-class error before the fully-connected layer is also proposed. This function makes the fully-connected layer as a classifier more sensitive to the same type of image and easier to achieve the classification task.
- The play-and-plug normalization-based attention module is utilized to solve the gradient saturation problem without adding additional trainable parameters.

II. PRELIMINARIES

A. BINARY NEURAL NETWORKS

Binary Neural Network (BNN) utilizes the most light-weighted quantization $\{-1, +1\}$ as its weight and activation values in each layer. It greatly speed-ups the forward propagation and mitigates the computational complexity. During its backward propagation, the Straight-through estimator (STE) [10] is employed to lessen the quantization error in an extremely simple way. There are mainly three approaches to optimize the BNN network’s performance: (1) further minimizing the quantization error [11, 12] between binarized and full-precision weights and activations. (2) improving the network loss function [13] to reserve information from the full-precision model. (3) reducing the gradient error to gain an accurate BNN accuracy [9].

The binarized BNN inevitably brings considerable information losses, and further leads to accuracy deterioration. To decrease the deterioration, Shannon entropy-based penalty is employed to obtains the balanced binary weights to maintain the original information capacity during the binarization [14]. Reference [15] shows that unbalanced distribution of binary activations can increase the accuracy by adjusting the threshold values of binarized activation functions. Inspired by these previous researches, our research adopts the $\text{RSign}$ function proposed in [6] and our improved $\text{RPReLU}$ in next section B.

B. REACTNET

By replacing the traditional $\text{Sign} / \text{PRReLU}$ with $\text{RSign} / \text{RPReLU}$ functions, ReActNet [6] reduces the gap to its real-valued counterpart to within 3.0% and powered the network to shift and even reshape the distribution of activations as (1) and (2):

$$A_i^r = \text{RSign}(A_i^r) = \begin{cases} +1, & \text{if } A_i^r > \alpha_i \\ -1, & \text{if } A_i^r < \alpha_i \end{cases} \quad (1)$$

$$Y_i^b = \text{RPReLU}(Y_i^r) = \begin{cases} Y_i^r - \gamma_i + \zeta_i, & \text{if } Y_i^r > \gamma_i \\ \beta_i(Y_i^r - \gamma_i) + \zeta_i, & \text{if } Y_i^r < \gamma_i \end{cases} \quad (2)$$
where subscript $i$ means the channel number. $A^r_i$ and $Y^r_i$ represent the real-valued activations before the binarization, $A^b_i$ and $Y^b_i$ are their binarized values. $\beta_i$ is a trainable coefficient which influences the slope of negative semi-axis, $\gamma_i$ and $\xi_i$ are learnable shifts for reshaping the distribution.

However, research [15] investigated the learnable thresholds from theoretical viewpoint and empirical evidence, as Fig. 2, the shift threshold $\gamma_i$ and BN bias $\epsilon_i$ have the same values but opposite signs. Equation (3) show that the threshold $\gamma_i$ is not essential as it can be covered by the bias values $\epsilon_i$ in the Batch Normalization (BN) layer which comes right before the RPReLU, as shown:

$$Y^b_i = \begin{cases} Y^r_i + \xi_i, & \text{if } Y^r_i > 0 \\ \beta_i Y^r_i + \xi_i, & \text{if } Y^r_i < 0 \end{cases}$$

(4)

where parameters $\mu$, $\sigma$, $\gamma_i$ and $\xi_i$ are mean, standard deviation, weight, and bias of the BN layer, respectively. $X^r_i$ is the convolution output of current channel $i$ and also input of upcoming BN layer.

**FIGURE 2.** The visualization of values of shifting $\gamma_i$ in RPReLU and the biases of BN layer $\epsilon_i$ in $5^{th}$ layer.

Authors [15] point out that the learnable threshold $\gamma_i$ just doubling the learning rate of BN bias $\epsilon_i$. To reduce the redundant parameters, we proposed an improved RPReLU as (4) indicated.

**III. TECHNICAL APPROACH**

A. **Binarized Ghost Module (BGM)**

The main goal to construct a highly-performed binary network is to mitigate the training parameters and information redundant computation without accuracy’s deterioration. In Fig. 3(a), the “ghost” feature maps are generated by a duplicated structure. This operation inevitably increases the computational redundancy without feature information’s increment.

Inspired by the GhostNet [7], our proposed BGM structure aims to lessen the calculations and enrich “ghost” feature maps in Fig. 3(b). The module can be regarded as a reversed depthwise separable convolution. A $1 \times 1$ convolution with more training parameters is employed as the primary computation to produce intrinsic feature maps.

At the same time, another simple linear $3 \times 3$ depthwise convolution proposed by MobileNet [16] is also employed to boost network information. Therefore, compared with the Fig. 3(a), these different two parts’ new concatenation increases the diversity of feature maps and also decreases the computational complexity.

**FIGURE 3.** To increase the channel numbers, the network structure of (a) is employed in recent state-of-the-art (SOTA) algorithm framework ReActNet. (b) keeps the left $1 \times 1$ convolution block in (a), whereas, the right $1 \times 1$ block is taken place by an authentic $3 \times 3$ depthwise convolution with our newly proposed concatenation.

By deploying our proposed binary ghost module as Fig. 3(b) introduced, the whole network computational complexity BOPs drops obviously as Table 1 listed. The light-weighted structure is benefit for implementing the BNN networks in power-aware hardware devices.

**B. REGULATE THE DISTRIBUTION OF ACTIVATION**

1) **LABEL-AWARE LOSS FUNCTION (LLF)**

The last full connection layer is taken as a linear classifier for the final classification result in previous researches. Therefore, the feature maps in neighbor penultimate layer contain high-level semantic information. The higher similarity of the same category of output distribution in the penultimate layer, the better classification result and performance of whole network can be achieved.

Therefore, our proposed Label-aware Loss Function (LLF) $L_{label}$ which helps to improve the penultimate layer’s similarity can be expressed as:

$$L_{label} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N(i)} \sum_{j=1}^{N(i)} \|Y_j^2 - mean\|_2^2$$

(5)

$$mean = \frac{1}{N(i)} \sum_{i=1}^{N(i)} Y_i$$

(6)

$$L_{total} = + \lambda \cdot L_{label}$$

(7)
where $Y$ denotes the feature maps in penultimate layer which is input into the final linear classifier for classification. $N$ is the number of all classified categories of the dataset, $N(i)$ indicates the number of belonging to same category $i$ in mini-batch samples.

In (5), the feature vector $Y_j$ of the same label category $i$ is more concentrated through a $L2$ norm operation. Our proposed LLF performs as an outstanding supervisor to lessen intra-class error, and improve the final classification accuracy. Finally in (7), our LLF is inserted as a supplementary function in $L_{total}$.

2) NORMALIZATION-BASED ATTENTION MODULE (NAM)
Normalization-based attention is a method recently proposed in [17]. The author emphasizes the important relationship between channels by weight $W^i_j$. As (8) indicated, its value is computed from the BN layer scale $γ_i$ to realize channel attention without introducing additional training parameters.

$$
\begin{align*}
B_{out} &= BN(B_{in}) = \frac{B_{in} - μ_γ}{\sqrt{σ_γ^2 + ε}} + β \\
W^i_j &= \frac{γ_i}{\sum_0^{i-1} γ_j}, \ i = 0,1,...,C - 1 \\
M_i &= B_{out} ∙ \text{Sigm}(W^i_j(B_{out}))
\end{align*}
$$

where scale $γ$ and shift $β$ are trainable affine transformation vectors of size $C$ ($C$ is channel dimension). $μ_γ$, $σ_γ$ are the mean and standard deviation of batch $B$. $\text{Sigm}(·)$ is the sigmoid function which is different from [17].

In addition, we discover that the $\text{Sigm}(·)$ further adjusts the scale of activation on the residual branches to reduce the activations’ variance. Therefore, our NAM method avoids the saturation problem when $|A_{|0}| ≥ 1$, $|A_{|0}|$ is the minimum absolusion activation value. An obvious and effective constrained distribution is illustrated in later experimental section Fig. 7.

IV. EXPERIMENTS
The organization of the experiments is as follows. The experiment settings are elaborated in part A. We prove our network’s effectiveness and compare with some-best-existing state of the art works in B. In C, a detailed demonstration of ablation study about our three proposals is presented. To achieve more clear demonstration of BGM and LLF, D visualizes the character distribution of the feature maps for the specified layer.

A. SETUP
The ReActNet [6] is originated from the MobileNetV1 [16] and has its own superiorities. Therefore, we conduct the experiments with our improved methods based on the ReActNet structure. Meanwhile, we investigate the effectiveness of our methods on CIFAR-10 and ILSVR12 ImageNet datasets.

CIFAR is a small-scale image classification dataset with $32 \times 32$ sized color image in 10 classes. The training set and testing set are comprised of 50,000 pictures and 10,000 pictures, respectively. ILSVRC12 ImageNet is a large-scale dataset with more diversity. It includes 1.2 million training images and 50,000 validation images across 1000 classes. In experiment section, we use all images from the dataset.

1) EXPERIMENTS DETAILS ON IMAGENET
We follow the two-steps training strategy utilized in [6]. Firstly, the binarized activations and full-precision weights are participated in the convolutional network training from an initial scratch. During the second step, the activations and completed full-precision weights in the 1st step are both tuned and binarized. For above two steps, we train the network with the Adam optimizer for 128 epochs and with a linear decreasing scheduler initiated learning rate of $5e^{-4}$, the weight decay is set to $1e^{-5}$ for the first step and 0 for the second step. The last layer, i.e. the fully-connection layer’s output channel is 1000. The hyper-parameter in loss function is set to 1e-2. As previous binarized networks researches, we also used data augmentation method [16] in our experiments. The method contains random cropping, picture lighting, and random horizontal flipping.

2) EXPERIMENTS DETAILS ON CIFAR-10
The last layer’s output number is set to 10 rather than ImageNet’s 1000. Additionally, we only random the the cropping and horizontal flipping on data augmentation process. Other training configuration remain the same as ImageNet.

B. COMPARISON TO STATE-OF-THE-ART
In this section, experimental results of our proposed approaches with the state-of-the-art (SOTA) binary networks are evaluated.

1) RESULTS ON IMAGENET
For the large-scale dataset, we explore our methods upon ReActNet structure and achieve the following results. Following [18, 19], the total operations (OPs) is calculated by $\text{OPs} = \text{BOPs} + 64 + \text{FLOPs}$. The Table 1 shows that “ReActNet + our methods” achieves highest 71.4% accuracy and least $0.86 \times 10^8$ OPs with the proposals in Section III.

| Binary Network   | BOPs ($\times 10^8$) | FLOPs ($\times 10^8$) | OPs ($\times 10^8$) | Top-1 Acc (%) |
|------------------|----------------------|-----------------------|---------------------|--------------|
| BNNs [5]         | 1.70                 | 1.20                  | 1.47                | 42.2         |
| XNOR-Net [18]    | 1.70                 | 1.41                  | 1.67                | 51.2         |
| Bi-RealNet-18 [19] | 1.68               | 1.39                  | 1.63                | 56.4         |
| Bi-RealNet-34 [19] | 3.53               | 1.39                  | 1.93                | 62.2         |
| Real-to-Binary Net [20] | 1.68               | 1.56                  | 1.83                | 65.4         |
| MeltiusNet29 [21] | 5.47                 | 1.29                  | 2.14                | 65.8         |
| ReActNet [6]     | 4.82                 | 0.12                  | 0.87                | 69.4         |
| ReActNet+[SI] [28] | 8.83                | 0.12                  | 1.15                | 69.6         |
| ReActNet + Our methods | 4.75                | 0.12                  | 0.86                | 71.4         |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Still, the total computational complexity parameter OPs is compared with the other binarized networks models. Among the OPs, a ReActNet-based \((4.82 - 4.75) \times 10^8\) = 0.07 \times 10^8 BOPs decrement is achieved by our simplified binary ghost module in Fig. 3(b). This structure helps augment the features and channels. It also improves the feature maps information with computational decrement.

Furthermore, the parameter FLOPs 0.12 \times 10^8 of “ReActNet + Our methods” is nearly same when compared with the referenced ReActNet in the Table 1. This is because the addition of our NAM attention mechanism (10e4-scale) does not bring too much FLOPs compared with the ReActNet. In fact, the first layer and last layer occupy nearly all FLOPs (10e8-scale) in BNNs.

Also, we plot the training curve of our network and recently highest-performed BNN network [24] (AdamReActNet) in Fig. 4. Our network and AdamReActNet are inherited from full convolution MobileNetV1, we observe that our network surpass top-performed AdamReActNet and narrow the gap between its real-valued counterpart to within 0.3%.

Therefore, it can be concluded that our methods gain a more light-weighted hardware-friendly network with a highest accuracy and least Ops in contrast to the other binarized networks upon the ImageNet database.

C. ABLATION STUDY

We further demonstrate the effectiveness of our proposed network in this section. With three components proposed in our methods in Section III, we provide an ablation study on the CIFAR-10.

2) RESULTS ON CIFAR-10

Since our baseline network is ReActNet and other traditional BNN network [5, 18, 19, 20] is almost modified from ResNet, we just compare our performance with AdamReActNet [24] (blue) and ReActNet-BN-Free [22] (black), two recently proposed ReActNet-series BNN network, to an exhibition of our proposed methods’ effectiveness.

As shown in Fig. 5, an 86.45% accuracy has been achieved as “Our network” (red) indicates. This accuracy outperforms the other two reference networks. In addition, the training curves of our methods is also consistently higher than the other two benchmarks. These indicate that the network not only ensures a stable training but also gains a higher accuracy with our methods, and more analysis and discussion will be presented in ablation study.

TABLE 2. Ablation study of the separate effect of BGM, LLF and NAM on CIFAR-10 dataset.
FIGURE 7. Feature maps and statistical distributions in the 2nd layer of our network. The top image is the input, the three on the left feature maps and statistical distributions are from the network with NAM, the three on the right are from “BGM+LLF” network.

1) BINARIZED GHOST MODULE

The ReActNet network duplicates the down-sampling architecture and concatenates the outputs as Fig. 3(a) shown. This method enables the use of identity shortcuts in the down-sampling layers. However, the duplicate operation is meaningless in terms of the information increment of networks. Therefore, the BGM is proposed to tackle this problem. In Fig. 3(b), the first $1 \times 1$ convolution is utilized as primary convolution for producing intrinsic feature maps. The second $3 \times 3$ depthwise convolution changes the intrinsic feature maps by some cheap transformations. These two parts are concatenated afterwards to maintain output channels twice than input channels. This operation further improves the whole network’s feature maps content, and more convincing results are shown in Fig. 8.

As Fig. 6 shown, the curve of our “BGM” in later 129-256 epochs almost overlaps the “Original” line. This demonstrates that our idea can obtain a close accuracy with lower computational overhead when compared with the reference “Vanilla”.

2) LABEL-AWARE LOSS FUNCTION

Our proposed LLF helps to enhance the similarity among samples in penultimate layer with a same label.

As shown in Fig. 5, line “BGM+LLF” (blue) means that the modified ReActNet with binarized ghost module (BGM) is regulated by the compound loss function. This function includes the cross-entropy loss and our LLF. Based on the experiments on CIFAR-10, the blue curve’s sharper slope indicates that our proposed LLF speeds up the network’s convergence.

Another, for pursuing an optimal value for penalty coefficient $\lambda$, the value is changed to balance the above two loss functions. Their accuracy results of our network are listed in Table 3.

| $\lambda$ | 1e-1 | 1e-2 | 1e-3 | 1e-4 |
|----------|------|------|------|------|
| Acc. (%) | 85.82 | 86.45 | 85.26 | 85.97 |

When $\lambda$ values wave, it can be found that $\lambda = 1e - 2$ is a favorable value which can achieve the highest accuracy. We conjecture that it is because the LLF supervises output features of penultimate layer which weakens another cross-entropy loss function’s contribution for the classification. This value also helps to improve 1.42% accuracy compared “BGM+LLF” with “BGM” in Fig. 6, as (5) and (6) shown, LLF explicitly use the mean of the variance of each category as penalty, which makes vector $Y_i$ closer to the expectation of its belonging category. Note that, all the operations where LLF...
does are in penultimate layers. The visualization result can be found in Fig. 9.

3) NORMALIZATION-BASED ATTENTION MODULE
To better investigate the performance of NAM, some feature maps in the 2nd layer of our network and their statistics of distribution are visualized and reported. From second row in Fig. 7, we can find that NAM leverages the network to maintain more useful information in feature maps which improves the contrast between two cats. Additionally, we can also observe that pixel value of feature maps has a narrower range and less standard deviation with NAM compared to a network performance without NAM.

These satisfying results are benefit from alleviating the saturation problem in the backward propagation by regulating most pixel values ranged within [-1, +1] as shown the third row, the more data (orange) is distributed between two red lines without long tail (obviously present in blue). The operation prevents the most pre-activations of a channel having a larger absolute value than the binarized threshold (threshold setting is decided by STE [10]). From the perspective of propagating gradient through hard tanh in [10], as (9) shown, the threshold is ±1.

\[ H_{\text{tanh}}(x) = \text{Clip}(x, -1, 1) = \max(-1, \min(1, x)) \quad (9) \]

Intuitively, our improvement avoids the gradient value of one number whose absolute value is higher than threshold becoming a zero, and also preventing the emergence of gradient explosion during the training stage.

Different from previous work [25], ReCU optimizes thresholds for different networks. We utilize affine parameters in BN layer to adaptively rehabilitate pixel range. As (8) indicate, the network can adaptively reshape the feature maps’ content through progressively adjusting affine parameters in gradient propagation. This also means our improved NAM can incorporate most deep-learning based network with BN layers. Equation (8) proves that the NAM is pretty light-weight without bringing large additional computational cost compared with [25]

D. VISUALIZATION
In this part, we provide the visualization of feature maps to better understand our binarized ghost module (BGM) as Fig. 8 shown. The feature maps inside right blue box is linearly transformed from the left red box in sequence. Both feature maps are concatenated as the output of the binarized ghost module. We can see that the module is inserted to enrich the feature maps’ diversity of network.

We also visualize the penultimate layer’s feature distribution utilized the t-SNE [23] method which is extensively employed dimensionality reduction technology for the visualization. Our proposed auxiliary loss function LLF takes category information into consideration. In Fig. 9, compared with the reference (a)’s performance on CIFAR-10, this LLF’s involvement makes each category’s feature vectors more concentrated and more easily to be separated as the (b) indicated.

![FIGURE 8](image.png)

**FIGURE 8.** The feature maps in the 2nd bottle network in ReActNet. The left-top image is the input image, the feature maps in the left red box are from the primary 1 × 1 convolution, and the right blue box represent the feature maps are from the following 3 × 3 depthwise convolution.

![FIGURE 9](image.png)

**FIGURE 9.** The representations of penultimate layer based on CIFAR-10. The network is trained without our proposed LLF (a) and with LLF (b).

V. CONCLUSION
In this paper, we utilize the binarized depthwise convolution as a series of light-weighted operations to guarantee the binarized ReActNet networks well-trained, specifically in a low-complex way. Next, the Label-aware Loss Function (LLF) decode the category information and condense the same-labeled samples’ feature vectors closer in penultimate layer is also introduced. This is greatly helpful for the upcoming final full-connection layer to improve the classification accuracy performance. Finally, the Normalization-based Attention Module (NAM) is adopted to mitigate the gradient saturation problem and stabilize training. Based those three methods, 71.4% Top-1 accuracy on the ImageNet and 86.45% accuracy on the CIFAR-10 are achieved, respectively. With the network’s high accuracy and low computational complexity, it is greatly promising and helpful for embedding the network into power-aware applications in near future.

REFERENCES
[1] K. Han, J. Guo, C. Zhang and M. Zhu, “Attribute-aware attention model for fine-grained representation learning,” in Proc. ACM MM, Canberra, AUS, 2018, pp. 2040–2048.
[2] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, “Focal loss for dense object detection,” in Proc. ICCV, Venice, ITA, 2017, pp. 2980–2988.
[3] Y. Chai, B. Sapp, M. Bansal, D. Anguelov, “MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction,” 2019, arXiv: 1910.05449. [Online]. Available: https://arxiv.org/abs/1910.05449

[4] M. Courbariaux, Y. Bengio, J. David, (2015). BinaryConnect: Training deep neural networks with binary weights during propagations. Presented at Adv. Neural Inf. Process. Syst. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/3e15cc11979edc25912dfff069692cdd-Paper.pdf

[5] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, Y. Bengio, “Binaryized Neural Networks: Training deep neural networks with weights and activations constrained to +1 or -1,” 2016, arXiv: 1602.02830. [Online]. Available: https://arxiv.org/abs/1602.02830

[6] Z. Liu, Z. Shen, M. Savvides, K. Cheng, “ReActNet: Towards precise binary neural network with generalized activation functions,” in Proc. ECCV, 2020, pp. 2980–2988.

[7] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, J. Song, “GhostNet: More features from cheap operations,” in Proc. CVPR, Seattle, USA, 2020, pp. 1580–1589.

[8] R. Ding, T. Chin, Z. Liu, D. Marcusciano, “Regularizing activation distribution for training binaryized deep networks,” in Proc. CVPR, Long Beach, USA, 2019, pp. 11408–11417.

[9] H. Qin, R. Gong, X. Liu, M. Shen, Z. Wei, F. Yu, J. Song, “Forward and Backward Information Retention for Accurate Binary Neural Networks,” in Proc. CVPR, Seattle, USA, 2020, pp. 2250–2259.

[10] Y. Bengio, N. Leonard, and A. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” 2013, arXiv: 1308.3432. [Online]. Available: https://arxiv.org/abs/1308.3432

[11] D. Zhang, J. Yang, D. Ye, G. Hua, “LQ-Nets: Learned quantization for highly accurate and compact deep neural networks,” in Proc. ECCV, Munich, Germany, 2018, pp. 365–382.

[12] A. Bulat, G. Tzimiropoulos, “XNOR-Net++: Improved binary neural networks,” 2019, arXiv: 1909.13863. [Online]. Available: https://arxiv.org/abs/1909.13863

[13] A. Mishra, D. Marr, “Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy,” 2017, arXiv: 1711.05852. [Online]. Available: https://arxiv.org/abs/1711.05852

[14] D. Ignatov, A. Ignatov, “Controlling information capacity of binary neural network,” Pattern Recognit. Lett. vol. 138, pp. 276-281, Oct. 2020.

[15] H. Kim, J. Park, C. Lee, J. Kim, “Improving Accuracy of Binary Neural Networks Using Unbalanced Activation Distribution,” in Proc. CVPR, 2021, pp. 7862–7871.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv: 1704.04861. [Online]. Available: https://arxiv.org/abs/1704.04861

[17] Y. Liu, Z. Shao, Y. Teng, N. Hoffmann, “NAM: Normalization-based attention module,” 2021, arXiv: 2111.12419. [Online]. Available: https://arxiv.org/abs/2111.12419

[18] M. Rastegari, V. Ordonez, J. Redmon, A. Farhadi, “XNOR-Net: ImageNet classification using binary convolutional neural networks,” in Proc. ECCV, Amsterdam, Netherlands, 2016, pp. 525–542.

[19] Z. Liu, B. Wu, W. Luo, X. Yang, W. Liu, K. Cheng, “Bi-RealNet: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm,” in Proc. ECCV, Munich, Germany, 2018, pp. 722–737.

[20] B. Martinez, J. Yang, A. Bulat, G. Tzimiropoulos, “Training binary neural networks with Real-to-Binary convolutions,” 2020, arXiv: 2003.11535. [Online]. Available: https://arxiv.org/abs/2003.11535

[21] J. Bethge, C. Bartz, H. Yang, Y. Chen, C. Meinelt, “MeliusNet: Can binary neural networks achieve MobileNet-level accuracy?”, 2020, arXiv: 2001.05936. Available: https://arxiv.org/abs/2001.05936

[22] T. Chen, Z. Zhang, X. Ouyang, Z. Liu, Z. Shen, Z. Wang, “BNN-BN = ?; Training binary neural networks without batch normalization,” in Proc. CVPR, 2021, pp. 4619–4629.

[23] L. V. Maaten, G. Hinton, “Visualizing data using t-sne ,” J. Mach. Learn. Res., vol 9, no. 1, pp. 2579-2605, Nov. 2008.