A Binarized Feature Mapping Technique for Enhancing Squeeze-and-Excitation (SE) Channel Attention Mechanism

Wu Shaoqing[1,*] and Hiroyuki Yamauchi[2]

[1] Fukuoka Institute of Technology/Graduate School, Fukuoka, Japan
[2] Fukuoka Institute of Technology/Computer Science and Engineering, Fukuoka, Japan
Email: mfm22202@bene.fit.ac.jp (S.Q.W.)); yamauchi@fit.ac.jp(H.Y.)
*Corresponding author
Manuscript received May 31, 2024; revised June 22, 2024; accepted July 17, 2024; published February 25, 2025

***Abstract*—Representing the weight in the network with only 1bit contributes to saving of the required memory footprint. Channel attention with the squeeze-and-excitation (SE) technique can eliminate redundant channels, resulting in saving of the number of the weights. Nevertheless, this causes an unstable and slow learning curve. To address this issue, this paper proposes the first attempt to accelerate the learning curve, even with a 1-bit weight representation across the whole SEResNet14 network, which significantly reduced the number of model parameters with only a minimal loss in accuracy. We also experimented with more aggressive activation functions such as HardTanh. We demonstrated that the FMB (Feature Map Binarization) method can reduce the number of active channels across different layers, thereby decreasing the quantity of weights in the channel direction. We also introduced the first attempt to utilize the EigenCAM for evaluating the channel attention effects. Experimental results demonstrate the efficacy of the proposed technique in the SE module in terms of speed-up of the learning curve and positional accuracy of the heat map based on the EigenCAM. We found the difference in the heat map position between the two cases with and without the proposed technique.**

***Keywords*—EigenCAM, ResNet14, CIFAR-10, SVHN, SE attention mechanism, 1-bit quantization, model compression, activation functions, channel feature maps binarization, ultra-compact ai deployment**

I. Introduction

Modern convolutional neural networks (CNNs) [1–7] consist of recurring blocks with identical structures, leveraging principles from residual learning [8–10] and utilizing depthwise separable convolutions [11]. 32bit representation for the weight and the activation has provided an impressive performance. However, it poses significant challenges in deploying them into a stringent power and memory-footprint constrained device.

First, we will explain why it is so important to reduce the required memory footprint to store the parameters of the weights and the activations. The main motivation for this is to eliminate the power-hungry external DRAM accesses, which consume a 100x larger power consumption compared with the accesses to embedded SRAM in an AI chip. It could only be done by eliminating the need to store the weighs the activations in the external DRAM. To do so, the required number of parameters has to be reduced to 1/100 at least.

Then, we will explain how to reduce the required number of parameters to 1/100.

As shown in Fig.1, the binarized representation with 1bit instead of 32bit reduced the parameters to 1/32. Further reduction to 1/2–1/4 can be done with the channel and spatial attentions by eliminating the redundant parameters that are not paid attention in the network. Once it could be done, almost parameters could be stored within the AI chip and the external DRAM accesses would be no longer required, resulting in a decreasing of the power consumption to 1/100.

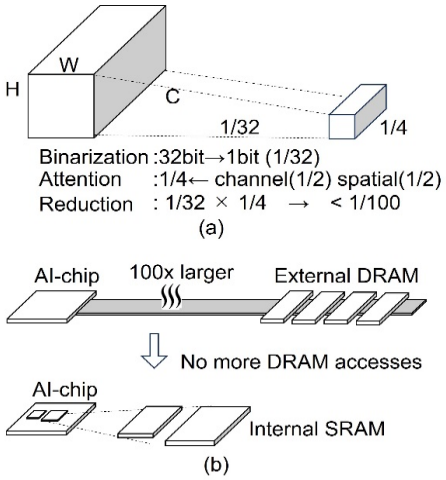


Fig. 1. Concept of how to reduce the energy consumption to 1/100. This can only be done by reducing the # of parameters to 1/100. (a)Which parameters can be reduced (b) No more external DRAM accesses.

The ResNets were introduced by He et al. to address the gradient vanishing issue [12]. We noticed that this invention causes to return to rise of exponentially increased parameters, resulting in an increased pressure for the stringent parameter reduction. The SE attention technique [13–16], introduced by Hu et al., provides a better model accuracy with channel-wise recalibration.

Lingling Li *et al.* integrated SE blocks into the HRNet [17], which is referred to as SE-HRNet.

Xinyu Zhang *et al.*, [18] improved the predictive accuracy of MRSE-Net in global remote sensing image water extraction tasks.

What we noticed regarding the SE network is that this also can reduce the number of parameters by eliminating redundant channels.

In this paper, we explore for the first time 1) how the SE attention mechanism works under the 1-bit quantization condition, 2) binarization of the output of the channel attention feature map, 3) activation function impact on the learning curves, and 4) evaluation of the channel attention effects with EigenCAM [19].

How the SE mechanism is changed when the above 1)-3)

are adopted and how much the EigenCAM can visualize those effects on the position accuracy of the heatmap have not been discussed in the previous papers. Thus, this paper is the first paper to propose and discuss on those topics.

Based on the experiment results of SE attention, we have noticed that 1/2 of the channel can be pruned and the 1/2 of the parameters can be removed.

We have also noticed that the spatial attention can be applied to further eliminate the parameters in the spatial direction. The CBAM [20] integrates both channel and spatial attention mechanisms, we will choose for our next experiment.

In this paper, we explore the accuracy and speed impacts of the binarized SE attention and the channel attention binarized feature map in the context of ResNet14 trained on the CIFAR-10 dataset. The CIFAR-10 dataset [21], consisting of 60,000 32x32 color images in 10 classes, is used for evaluating the performance.

We have proposed the channel feature maps binarization (FMB), in which some intermediate values in the channel attention during the early stages of training are forcibly binarized to investigate its impact on the model accuracy. This study compared the learning curves to investigate how much the proposed technique can contribute to reduce the error rate and required number of epochs to reach a certain error rate among the cases for using the different precisions: 1) float32bit as a baseline, 2) 1bit without using the proposed FMB technique, and 3) 1bit with using the FMB. We also examined the impact of the activation function of the SE module.

The main contributions of this article can be summarized as follows.

1) We have evaluated the effects of the attention with the heatmap position accuracy based on the EigenCAM.

2) We have investigated the activation functions impact of the activation functions on the model accuracy and speed of the learning curve in the binarized SE module.

3) We have proposed the FMB technique and demonstrated that it is effective for enhancing model accuracy.

The rest of this article is organized as follows. Section II elucidates the issues encountered with SEResNet14 under 1bit binarization. In the Section III, we provide a detailed introduction to our proposed technique. We discussed the results in Section IV. In Section V, we conclude this article.

II. Issue to Be Addressed

A. Accuracy Drop Due to Lower Precision

One primary concern about the binarized representation is that it can lead an unstable bang-bang behaviors in the learning curves, resulting in the drop of the accuracy and error reduction speed. This issue is particularly critical for deploying the models on the resource-constrained devices where power supply capability and the memory footprint are limited.

AmirAli Abdolrashidi *et al.* [22] concluded that 4-bit quantization is the optimal choice for balancing accuracy and parameter quantity. When further shift from a 32bit floating-point precision to a binarized one inherently, it becomes difficult to ensure that the minute differences between weights can be distinguished because of approximating error (a,k,a quantization error). Many weights may get rounded off to the same value (e.g., +1/-1) due to the lack of granularity, resulting in a significant loss of precision. This causes an unstable bang-bang behavior in the learning curves, resulting in a drop of the error reduction speed.

When directly quantizing the 32bit ResNet14 to 1bit, there is the significant drop of the accuracy, as shown in Table 1.

Table 1. Top1 Acc. Under different quantization conditions

	Cifar10	**SVHN**
32bitResNet14	93.67%	96.15%
1bitResNet14	91.07%	95.05%

In some cases, post-quantization fine-tuning might be employed to recover some of the lost accuracy. This involves retraining the quantized model for a few epochs to adjust to its new, approximated weight values. However, even with fine-tuning, there might still be a noticeable drop in performance, especially when extremely low bit-widths (i.e., 1-bit) are chosen.

That is to say that, while low-bit quantization offers advantages in memory savings and computational efficiency, it comes at the cost of accuracy due to the inherent approximation involved. The challenge is to find the sweet spot where the benefits of quantization outweigh the potential decrease in model performance.

B. The Application of CNN Models in Miniaturized Devices Presents Challenges

ResNet is a deep neural network that contains dozens to hundreds of layers. Each layer involves extensive matrix operations, presenting a significant challenge to computational capabilities. Small AI devices typically come equipped with low-power, low-performance processors, which struggle to handle such high complexity tasks. Quantization can somewhat reduce the model's computational complexity, but the ResNet series improves accuracy by increasing the use of convolutional layers and residual connections. For image classification tasks that are not overly challenging, like CIFAR10 and SVHN, the exponential increase in neural connections due to an excessive number of convolutional layers is undesirable. We attempted to modify ResNet18 into ResNet14 by removing four convolutional layers in layer4, but this also resulted in considerable accuracy loss. Therefore, we expanded the network appropriately in the channel dimension, hoping to recover some of the lost accuracy. We then applied the SE channel attention mechanism to filter the increased number of channels and coupled it with model pruning in the channel dimension. This approach aimed to maximize model accuracy without significantly increasing computational demands.

III. Proposed Techniques

This section introduces our two proposed methods of *A*) replacing the activation function in the SE module and *B*) the method for binarizing the output of the channel feature maps (FMB).

A. Activation Function of SE Module

To determine the relevance and redundancy of feature maps in the channel dimension during practical application, we recorded the count of essential feature maps at various depths within the network. Our aim was to identify which connections might be superfluous and thus candidates for pruning.

Our experiments revealed that choosing between the tanh and sigmoid activation functions within the SE module significantly impacts the model's accuracy. Recent literature [23] suggests that under conditions of low-bit quantization, the Tanh function often surpasses its counterparts in performance. In our study, we evaluated the model's accuracy using both tanh and sigmoid functions independently. This evaluation was conducted in conjunction with a novel method we proposed for the binarization of channel feature maps (FMB).

B. Channel Feature Map Binarization (FMB)

In the context of the channel attention mechanism, we observed that channel attention is almost completed in the early stage of model training. When the model progresses and the accuracy reaches about 60-70%, the weights of the channel attention hardly change. This is due to the use of an excessively large number of channel expansions for relatively simple model classification tasks, where most of the channels are not actively utilized. However, some intermediate values present in the mid-phase of model training are eventually classified into the two extremes of the output range after several epochs. We might consider using some methods to categorize these intermediate values into the two extremes of the output range at the beginning of the training.

In the selection of activation functions, considering the specific needs and characteristics of the model is crucial. In traditional neural networks, the Sigmoid function is often used as the activation function because it maps inputs to a fixed range (0 to 1), which is very useful for probability predictions or ensuring that the network output remains within a certain range. However, we are dealing with a 1bit quantized model, which uniquely quantizes the weights to -1 and 1. Under these circumstances, using the Tanh (hyperbolic tangent) function as the activation function is more appropriate.

In Eq. (1), F denotes the input feature map, which is a three-dimensional array with dimensions C×H×W, where C, H, and W represent the number of channels, the height, and the width, respectively.

In Eq. (1), the variable F is used to represent the input feature map. This feature map is a three-dimensional array with its dimensions described as C×H×W. Here, C' stands for the count of channels, 'H' signifies the height, and 'W' is the width of the feature map.

$$F \in R^{C\times H\times W} \tag{1}$$

In Eq. (2), M_c is depicted as the output of channel weights from the SE (Squeeze-and-Excitation) module. It maintains the structure of a three-dimensional array. However, in this setup, each of the C channels contains only a single element (1×1), indicating that each channel is assigned a distinct weight.

$$M_c \in R^{C\times 1\times 1} \tag{2}$$

In Eq. (3), we are introduced to two distinct weight matrices, labeled W_0 and W_1. These matrices are utilized in the fully connected layers within the SE module. The term r represents a reduction ratio, which serves to modulate both the complexity of the model and its parameter count. The matrix W_0 is responsible for diminishing the number of channels from C down to C/r,. Conversely, W_1 functions to restore the number of channels from C/r, back up to C. For the purposes of this research, the reduction ratio is fixed at a value of 1.

$$W_0 \in R^{C/r\times C} \; and \; W_1 \in R^{C\times C/r} \tag{3}$$

In Eq. (4), the process begins with the application of global average pooling on the input feature map F, which facilitates the extraction of global features for each channel. Subsequently, the architecture utilizes two fully connected layers, named W_0 and W_1 ,, to discern and learn the interrelationships among the channels. An activation function, specifically ReLU (Rectified Linear Unit), is integrated between these two fully connected layers. The culmination of this process is the application of the Tanh function, which effectively outputs the weights M_c for each individual channel.

$$\begin{aligned} M_c(F) &= Tanh\big(AvgPool(F)\big) \\ &= Tanh\Big(W_1\Big(ReLU\Big(W_0\big(F^c_{avg}\big)\Big)\Big)\Big) \end{aligned} \tag{4}$$

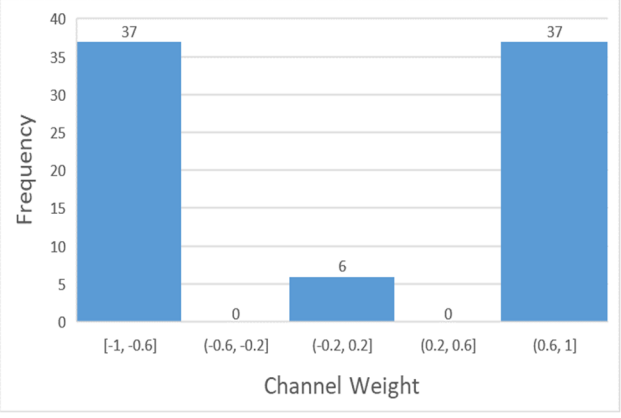
Regarding the Tanh function, which generates output values ranging from -1 to 1, a threshold of 0 is employed. Consequently, any values surpassing 0 are assigned a value of 1, while those falling below 0 are designated as -1, as shown in Eq. (5).

$$M'_c\big(M_c(F)\big) = \begin{cases} 1, & M_c(F) \geq 0 \\ -1, & M_c(F) < 0 \end{cases} \tag{5}$$

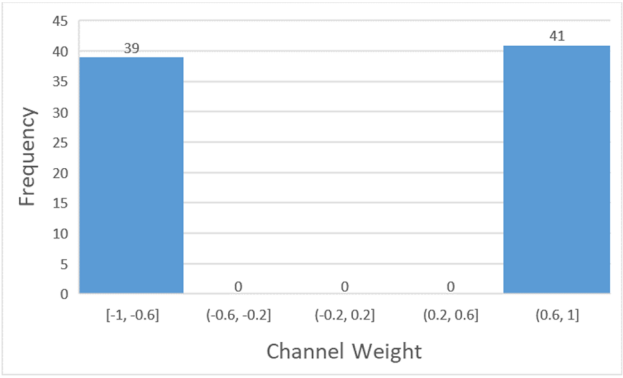
In Eq. (6), the final output F' of the SE Module is given by the product of the original feature map F with the adjusted channel weights M'_c obtained above.

$$F' = M'_c\big(M_c(F)\big) \odot F \tag{6}$$

In Fig. 2, to more intuitively demonstrate the effect of the proposed FMB, we have plotted the frequency histograms of the channel feature maps output from the SE1 module before and after binarization.



(a) w/o using FMB



(b) w/ using FMB

Fig. 2. Comparison of output feature map value distributions between the cases of (a) w/o and (b) w/ using FMB technique.

In PyTorch, the "torch.where" function is an extremely useful tool that allows for selecting elements from two tensors based on a condition. The functionality of Eq. (1) and Eq. (2) can be effectively implemented using "torch.where".

IV. Results and Discussion

A. Impacts of Activation Function of SE Module

In the SE (Squeeze-and-Excitation) module, the choice of activation function is crucial for achieving the desired non-linear transformation effect. The ReLU function is commonly used in the squeezing step to introduce non-linearity, while the Sigmoid function is used to convert values between 0 and 1 to generate attention weights, a step that is commonly applied in the SE module. However, when we employ 1-bit quantization, quantizing weights to -1 and 1, this approach encounters a problem in the implementation of channel attention. Channel attention requires multiplying the original weights by the channel feature maps calculated by the SE module. If we use the standard Sigmoid function to convert channel weights to values between 0 and 1, this introduces a bias. The reason for this bias is that the output range of the Sigmoid function (between 0 and 1) is inconsistent with the output range of the quantized weights (-1 to 1).

B. Impacts of FMB Technique

FMB (Feature Map Binarization) aims to binarize channel attention weights early in model training to achieve an effect similar to 1-bit quantization.

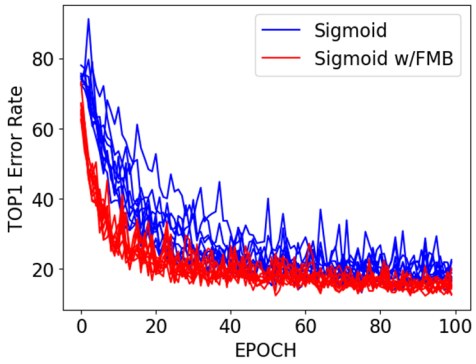


Fig. 3. The curves of the model error rate over training epochs with and without the use of the FMB method when the SE activation function is Sigmoid.

We utilized the Sigmoid function and the Tanh function as the activation functions for output channel feature maps within the SE module, respectively. The Sigmoid function is the most commonly used in SE modules, whereas the Tanh function, with its output range more closely aligned with our binary quantization method, was also included in our tests. We randomly selected training results 10 times, with and without the use of the FMB method, and plotted the curves of model accuracy over training epochs as shown in Fig. 3 and Fig. 4. We observed that with both types of activation functions, the FMB method could enhance the model's learning ability in the early stages.

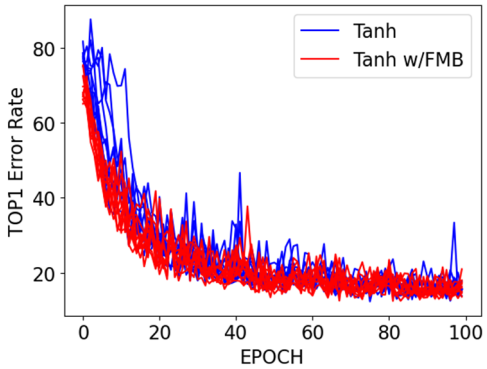


Fig. 4. The curves of the model error rate over training epochs with and without the use of the FMB method when the SE activation function is Tanh.

Under the condition of a 1-bit quantized model, we also tried the SE (Squeeze-and-Excitation) module used HardTanh as the activation function to output channel weights. For the same input image, we have enhanced ResNet14 with two different approaches: one with the SE (Squeeze-and-Excitation) module and the other combining the SE module with the FMB (Feature Map Binarization) method. We compared the number of activated channels in different layers of the network for ResNet14 without the SE module to those with SE and SE w/FMB enhancements as shown in Table 2.

Table 2. The activated channels in different layers

Models	Layer1	Layer2	Layer3
ResNet14	80	160	320
SE-ResNet14	43	83	175
SE-ResNet14 w/FMB	35	78	163

The FMB (Feature Map Binarization) method can reduce the number of activated channels with minimal loss in accuracy. Although the specific connections between channels across different layers are not known, a reduction in the number of activated channels across all layers necessarily leads to a decrease in the overall number of neuronal connections in terms of channel direction. This reduction could potentially streamline the network, making it more efficient without significantly compromising performance.

As shown in Fig. 5, let's consider a simple example: for channels suppose layer 1 has 3 neuron nodes, and layer 2 has 6 neuron nodes. Typically, without applying channel attention, the connections between layer 1 and layer 2 are fully connected, as shown in the scenario without SE

(Squeeze-and-Excitation), resulting in 18 neural connections. After applying the SE channel attention mechanism, assume only 2 nodes in layer 1 and 4 nodes in layer 2 are activated, thus reducing the neural connections to 8. Similarly, applying the FMB (Feature Map Binarization) mechanism can further reduce the number of neural connections. This reduction is acceptable in the context of minimal accuracy loss because neural connections become increasingly complex as the network depth increases. Even if the use of FMB results in a slight reduction in the number of activated channels across different layers compared to not using it, connecting the neurons would significantly decrease the number of neural connections. This is reflected in the model as a reduction in model parameters and size.

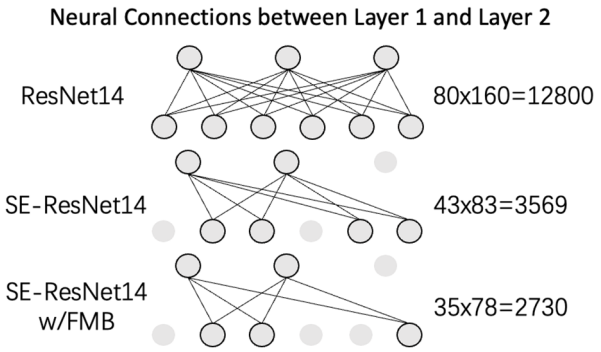


Fig. 5. Channels neural connections in different models.

C. F1-Score Test

Precision and recall affect each other; ideally, we aim for both to be high, but they "constrain" each other: pursuing high precision leads to lower recall, and aiming for high recall usually impacts precision. Of course, we hope for the prediction results to have as high precision and recall as possible, but in some cases, these two metrics are contradictory. This necessitates a comprehensive consideration of both, and the most common method for this is the F1 score. We have also tested the F1 score of the model; for a model trained on the CIFAR-10 dataset, we conducted tests for each category within the dataset to determine the F1 score for each category.

Table 3. F1 Scores for 1Bit quantized models across classification objectives

	ResNet14	SE-ResNet14	SE-ResNet14 w/FMB
Airplane	0.92	0.92	0.92
Automobile	0.95	0.96	0.96
Bird	0.88	0.89	0.89
Cat	0.83	0.82	0.82
Deer	0.91	0.9	0.89
Dog	0.85	0.87	0.87
Frog	0.93	0.93	0.93
Horse	0.94	0.94	0.94
Ship	0.95	0.95	0.95
Truck	0.94	0.94	0.95

In Table 3, it can be observed that the 1-bit quantized ResNet14 series is not very proficient in cat and dog recognition, with most of the model's misclassifications occurring between these two categories. By adding the SE (Squeeze-and-Excitation) channel attention mechanism to this 1-bit quantized ResNet14, we can slightly improve the recognition accuracy for the classification task of "Dog". Due to the limitations of the network model structure and low-bit quantization, the improvement in model accuracy achieved through the use of SE and FMB methods is limited.

D. Discussion

In Table 3, it can be observed that the 1-bit quantized ResNet14 series is not very proficient in cat and dog recognition, with most of the model's misclassifications occurring between these two categories. By adding the SE (Squeeze-and-Excitation) channel attention mechanism to this 1-bit quantized ResNet14, we can slightly improve the recognition accuracy for the classification task of "Dog". Due to the limitations of the network model structure and low-bit quantization, the improvement in model accuracy achieved through the use of SE and FMB methods is limited. In the Fig. 6, we have selected the same "automobile" as the input image.

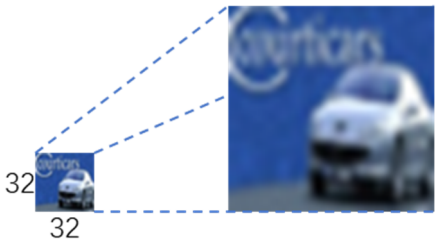


Fig. 6. 32×32 Image of automobile.

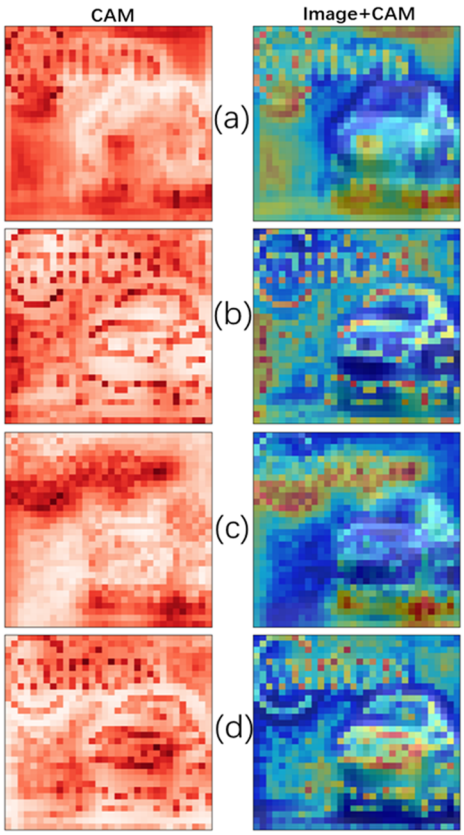


Fig. 7. EigenCAM results for different models: (a) 32bitResNet14, (b) 1bitResNet14, (c) 1bitSE-ResNet14, (d) 1bitSE-ResNet14 w/FMB.

EigenCAM simplifies CNN decision-making visualization by using PCA on feature maps to identify

critical patterns and generate heatmaps, highlighting important image areas. It's compatible with all CNNs without modifications, emphasizes significant data through PCA, and does not require class-specific information. This tool aids in model debugging, offers clear explanations in sensitive areas, and enhances interpretability, making AI decisions more transparent and understandable.

From Fig. 7(a) to (d), they are respectively 32bitResNet14, 1bitResNet14, 1bitSE-ResNet14, and 1bitSE-ResNet14 w/FMB.

We can see that the 1bitResNet14 model, after applying the SE and FMB methods, not only reduces the model size in the channel direction but also has a certain impact in the spatial direction. This is clearly visible from the CAM image in Fig. 7(d), where the outline of the car and the areas focused on by the model can be clearly seen.

V. Conclusion

This paper introduces the FMB (Feature Map Binarization) channel binarization technique and explores the optimal combination with activation functions to enhance model accuracy, as shown in Fig. 3. We also attempt to further streamline the model with more aggressive activation functions such as HardTanh and FMB without losing accuracy or with minimal accuracy loss, as detailed in Table 1 and Table 2. In Table 3, we tested the detection performance of different models for each classification task, which can serve as a basis for optimizing different classification tasks. In the discussion section, we explore various decisions including quantization, channel attention, and FMB, and analyze their EigenCAM results, allowing us to intuitively understand the advantages and disadvantages of different decisions.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors conducted the research. Wu mainly analyzed the data. Wu mainly wrote the paper. All authors had approved the final version.

References

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[2] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[3] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. on Machine Learning (ICML)*, vol. 97, 2019, pp. 6105–6114.

[4] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856.

[5] S. Chakraborty, A. Amrita, T. Choudhury, R. Sille, C. Dutta, and B. K. Dewangan, "Multi-view deep cnn for automated target recognition and classification of synthetic aperture radar image," *J. Adv. Inf. Technol.*, vol. 13, no. 5, pp. 413–422, Oct. 2022.

[6] M. Ashrafuzzaman, S. Saha, and K. Nur, "Prediction of stroke disease using deep cnn based approach," *J. Adv. Inf. Technol.*, vol. 13, no. 6, pp. 604–613, Dec. 2022.

[7] S. N. Kumar and C. Sumanth Kumar, "Fusion of cnn-qcso for content based image retrieval," *J. Adv. Inf. Technol.*, vol. 14, no. 4, pp. 668–673, 2023.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] M. S. Puchaicela-Lozano, L. Zhinin-Vera, A. J. Andrade-Reyes, D. M. Baque-Arteaga, C. Cadena-Morejón, A. Tirado-Espín, L. Ramírez-Cando, D. Almeida-Galárraga, J. Cruz-Varela, and F. Villalba Meneses, "Deep learning for glaucoma detection: r-cnn resnet-50 and image segmentation," *J. Adv. Inf. Technol.*, vol. 14, no. 6, pp. 1186–1197, 2023.

[10] S. Bunrit, N. Kerdprasop, and K. Kerdprasop, "Improving the representation of cnn based features by autoencoder for a task of construction material image classification," *J. Adv. Inf. Technol.*, vol. 11, no. 4, pp. 192–199, Nov. 2020, doi: 10.12720/jait.11.4.192–199.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[12] A. Krizhevsky, "Learning multiple layers of features from tiny images," Master's thesis, University of Toronto, 2009.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[14] H. Zhu, Z. Li, Y. Zhang, Z. Li, Y. Wang, and J. Liu, "MS-HNN: Multi-scale hierarchical neural network with squeeze and excitation block for neonatal sleep staging using a single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2195–2204, 2023, doi: 10.1109/TNSRE.2023.3266876.

[15] X. Jin, Y. Li, J. Wan, X. Lyu, P. Ren, and J. Shang, "MODIS green-tide detection with a squeeze and excitation oriented generative adversarial network," *IEEE Access*, vol. 10, pp. 60294–60305, 2022, doi: 10.1109/ACCESS.2022.3180331.

[16] J. Ai, S. Hou, M. Wu, B. Chen, and H. Yan, "MPGSE-D-LinkNet: Multiple-parameters-guided squeeze-and-excitation integrated D-LinkNet for road extraction in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Art no. 5508205, 2023, doi: 10.1109/LGRS.2023.3306725.

[17] L. Li, T. Tian, H. Li, and L. Wang, "SE-HRNet: A deep high-resolution network with attention for remote sensing scene classification," in *Proc. IGARSS 2020 - 2020 IEEE Int. Geosci. Remote Sens. Symp.*, Waikoloa, HI, USA, 2020, pp. 533-536, doi: 10.1109/IGARSS39084.2020.9324633.

[18] X. Zhang, J. Li, and Z. Hua, "MRSE-Net: Multiscale residuals and SE-attention network for water body segmentation from satellite images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5049-5064, 2022, doi: 10.1109/JSTARS.2022.3185245.

[19] M. Bany Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207366.

[20] S. Woo, J. Park, J. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, 2018, pp. 3–19.

[21] H. Bai, Y. Yang, Y. Liu, Y. Jiang, and W. Wen, "BinaryBERT: Pushing the limit of BERT quantization," in *Proc. Ann. Meeting of the Assoc. for Comput. Linguistics (ACL)*, 2020, pp. 1–12.

[22] A. Abdolrashidi, E. Real, Q. V. Le, and A. K. Hanay, "Pareto-optimal quantized resnet is mostly 4-bit," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, pp. 3085-3093, doi: 10.1109/CVPRW53098.2021.00345.

[23] K. Abdelouahab, M. Pelcat, and F. Berry, "Why tanH is a hardware friendly activation function for CNNs," in *Proc. 11th Int. Conf. on Distributed Smart Cameras (ICDSC)*, 2017, pp. 199–201, doi: 10.1145/3131885.3131937.

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).