EasyChair Preprint

№ 10663

A Self-Distillation Assisted ResNet-KL Image Classification Network

Yuanyuan Wang, Haiyang Tian and Yu Shen

August 3, 2023

A Self-Distillation Assisted ResNet-KL Image Classification Network

1st Yuan Yuan Wang
School of Computer Science and Software Engineering, Huaiyin Institute of Technolog
Jiangsu Province IoT Mobile Internet Technology Engineering Laboratory.
Huai'an, China.
zhfwyy@hyit.edu.cn

2nd Hai Yang Tian
School of Computer Science and Software Engineering, Huaiyin Institute of Technolog
Huai'an, China.
2287486860@qq.com

3nd Yu Shen
School of Computer Science and Software Engineering, Huaiyin Institute of Technolog
Huai'an, China.
1101202190@qq.com

***Abstract*—Traditional ResNet models suffer from large model size and high computational complexity. In this study, we propose a self-distillation assisted ResNet-KL image classification method to address the low accuracy and efficiency issues in image classification tasks. Firstly, we introduce depthwise separable convolutions to the ResNet network and enhance the model's classification performance by improving the design of activation functions, using T-ReLU instead of traditional ReLU. Secondly, we enhance the model's perception of features at different scales by incorporating multi-scale convolutions for the fusion of residual layers and attention mechanism layers. To reduce the model's parameter count, we combine feature distillation with logic distillation and optimize the model layer by layer through self-distillation, while applying pruning techniques multiple times to reduce its size. Finally, To assess the efficacy of our methodology, we conduct experimental evaluations on public datasets CIFAR-10, CIFAR-100, and STL-10. The results show that the improved ResNet-KL network achieves an accuracy improvement of 1.65%, 2.72%, and 0.36% compared to traditional ResNet models on these datasets, respectively. Our method obtains better classification performance with the same computational resources, making it promising for applications in tasks such as object classification.**

Keywords—image classification, ResNet, self-distillation, pruning

I. INTRODUCTION

A. Research Background and Objectives

Classification is one of the important tasks in the field of information processing and data analysis. In academia, classification techniques are widely applied in various research areas, such as image recognition, object detection, and facial recognition [1]. Classification techniques can be used for disease diagnosis and image analysis in medical imaging[2].

Despite significant progress in classification techniques, there are still challenges and issues that need to be addressed. Firstly, data imbalance is a prominent issue characterized by a significant disparity in the number of samples across various categories. This can lead to poorer classification performance for minority classes. Secondly,

dealing with noisy data and feature selection is crucial in classification. Real-world data often contains noise or redundant information, which can negatively impact model performance. Continuous improvement and research in classification techniques are therefore crucial for addressing the aforementioned challenges and issues and enhancing accuracy, efficiency, and robustness. This not only holds theoretical importance but also practical significance.

The main contributions of this paper are manifested in three aspects. Firstly, we propose a novel self-distillation assisted ResNet-KL algorithm that leverages feature distillation, knowledge distillation, and pruning mechanisms to enhance model performance. Secondly, we introduce T-ReLU, a fusion of Tanh and ReLU, as a replacement for ReLU activation function. Lastly, we incorporate a multi-scale approach to fuse residual layers and attention mechanism for improved perception of features at different scales.

B. .*Research Status*

In recent years, with the rapid development of computer vision and the widespread application of deep learning techniques, image classification has been a research area of great interest [3]. In China, Yunpeng et al[4] proposed Octave Convolution to reduce redundant information between feature maps, thereby reducing computational complexity and improving the efficiency of image classification. Yao et al[5] proposed a deep mixed multi-graph neural network, which effectively extracts spectral features of nodes and exhibits good noise suppression performance for graphs.

In foreign countries, research on image classification has also made significant progress. Many research institutions and scholars have conducted in-depth studies in the field of image classification using deep learning techniques. For example, Arco et al[6] proposed a multilevel ensemble classification system based on Bayesian deep learning, maximizing performance while providing uncertainty estimation for each classification decision. Chen et al[7] proposed dual-path networks, improving information flow and feature extraction capabilities in image classification through parallel dense connections and depthwise separable convolution layers. Despite the significant contributions made by previous studies in the field of image classification, challenges remain in complex scene recognition, few-shot learning, and zero-shot learning. Therefore, in this paper, we employ self-distillation as an auxiliary method to address these issues during network training.

II. RESNET-KL MODELS

The proposed ResNet-KL model for image classification with self-distillation assistance consists of residual modules, attention mechanism modules, mixed depth separable modules, and multi-scale modules. The innovation of this network lies in having two outputs: one is the logical output after passing through the fully connected layer (FC), and the other is the feature output without passing through any additional layers. These feature maps are then resized to a fixed size [batch_size, 2048, 7, 7] using the ST layer, and finally passed through the FC layer for output. The simplified diagram of the model is shown in Figure 2.1.

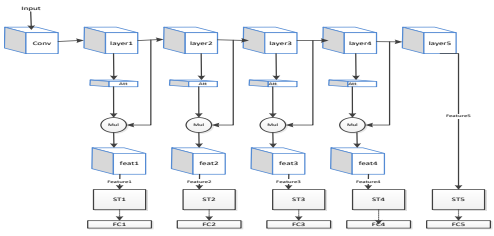

Figure 2.1 ResNet-KL model (dashed arrows indicate that it can go into the fully connected layer below, or it can bypass it).

Huai'an Municipal Science and Technology Plan Project(HABL202101) ;Postgraduate Research & Practice In novation Program of Huaiyin Institute of Technology (HGYK202317).

A. *Mul Block*

In this paper, the information from the residual module, after passing through an attention mechanism, is fused with the information from the residual module itself. Moreover, different sizes of convolution kernels are used at each layer to extract information at multiple scales, capturing different perspectives and obtaining more information. Additionally, the residual module has a simple structure and can be easily combined with other types of neural network modules to form more complex and powerful network architectures. The structural diagram for the multi-scale fusion of the residual module is shown in Figure 3.1.

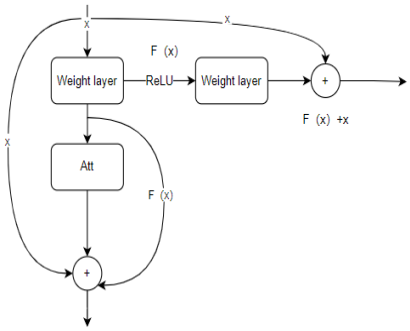

Figure 3.1 Multi-scale residual structure diagram.

B. *Pruning and Self-Distillation*

In this paper, pruning and self-distillation techniques are used to reduce the model parameters, combining feature distillation with logic distillation. The proposed self-distillation pruning method consists of a data input part that involves CutMix data augmentation and feature distillation channels and logic distillation channels for knowledge distillation, which are fused together. The original images undergo training using the CutMix data augmentation technique in the network's input section. The following (1) represents the CutMix data augmentation formula:

$$L = \frac{(x * s)}{N} + (1 - x) * M \qquad (1)$$

Here, L represents the new sample, x represents the pixel value of the cropped area, s represents the area of the cropped region, N represents the total number of pixels, and M represents the pixel value of the remaining area.

The main feature of this network structure is to start from the logic outputs of the upper layers of the network and the feature output of the last layer. These outputs are saved in a list and then obtained sequentially during training. The first 4 layers are logic outputs, while the last layer is the feature output, which is the deepest feature output. Logic distillation and feature distillation are performed during network training, as shown in Figure 3.2. Pruning operation is performed during specific training rounds by calculating the L1 norms of the convolutional layers and the Batch Normalization (BN) layers. The relevance of these layers is evaluated, and the least relevant ones are pruned. The pruned convolutional layers and BN layers are then replaced with the newly obtained ones for further training. The pruning process is conducted in the initial rounds due to the initially chaotic network structure. This allows for the removal of a greater number of irrelevant weights, thereby reducing the overall number of parameters.

The utilization of the self-distillation technique results in a marginal increase in the overall loss function. This is attributed to the combination of logic distillation loss, network loss, and feature distillation loss in the final loss calculation. The calculation formula for the logic distillation loss function is shown in (2):

$$loss = Cross(out\_i, last) * x + Cross(out\_i, labels) * (1 - x) \qquad (2)$$

In this context, "*loss*" refers to the loss function, "Cross(·)" represents the cross-entropy loss, "*out_i*" represents the logic output of each layer, "*last*" represents the output of the deepest layer, and "labels" represent the correct image labels. x represents the weight of the cross-entropy loss function.

In this paper, self-distillation is employed to

align the output features of the network with the predictions generated by previous layers during the model training phase. This approach enhances the network's detection capability in diverse scenarios, albeit at the cost of an increased loss function. Finally, the network utilizes the output results from the prediction channel for object classification.

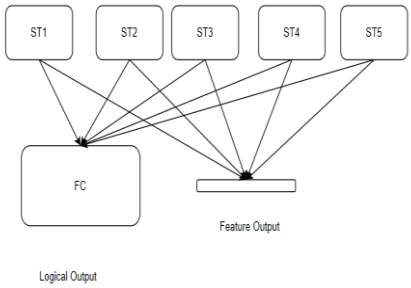

Figure 3.2 Dual output network structure.

C. ST Layer

The ST (Separable Tool) layer is mainly used for feature map scaling in feature distillation. ST layer is composed of a varying number of SepConv convolutional layers and average pooling layers. SepConv convolution includes two steps: depthwise convolution and pointwise convolution.

Depthwise Convolution: Depthwise convolution is a method that performs convolution separately on each input channel. It uses a convolution kernel of the same size as the number of input channels and performs convolution operations on each channel individually. Assuming the input feature map has C channels and the kernel size of depthwise convolution is k×k, the output feature map of depthwise convolution (denoted as *DF*) can be represented by the following (3):

$$DF = DepthConv(F, Wd) \quad (3)$$

Here, *F* represents the input feature map, *Wd* represents the weights of depthwise convolution kernel, and *DepthConv(·)* denotes the depthwise convolution operation.

Pointwise Convolution: Pointwise convolution is a method that performs convolution operations across the entire feature map. It uses a 1×1 convolution kernel and performs a linear combination of channels at each position. Assuming the output feature map of depthwise convolution is *DF* and the weights of pointwise convolution kernel are *Wp*, the output feature map of pointwise convolution (denoted as *PF*) can be represented by the following (4):

$$PF = PointConv(DF, Wp) \quad (4)$$

Here, *Wp* represents the weights of pointwise convolution kernel, and *PointConv(·)* denotes the pointwise convolution operation. By integrating depthwise convolution with pointwise convolution, the formula for depthwise separable convolutional layer (denoted as *SF*) is represented by the following (5):

$$SF = PointConv(DepthConv(F, Wd), Wp) \quad (5)$$

The convolutional structure is illustrated in Figure 3.3.

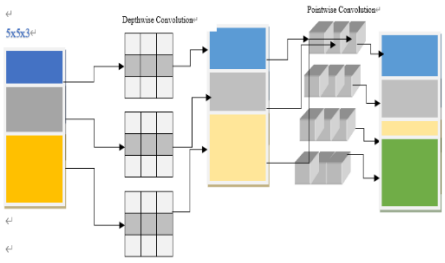

Figure 3.3 SepConv convolution structure

In the domain of image classification, certain images may exhibit minimal disparities between classes, and the absence of negative values in ReLU outputs can lead to the accumulation of biases between activation layers, which can affect the classification performance. To address this issue, the left part of the Tanh function is taken for values less than zero, and the right half of the ReLU function is used for values greater than zero. This combination is referred to as the T-ReLU function. The formula for the T-ReLU function is represented as (6):

$$f(x) = \begin{cases} x & x > 0, \\ \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} & x \le 0. \end{cases} \quad (6)$$

III. Experimental Results and Analysis

A. Experimental Setup

The experimental parameters mainly include learning rate, batch size, epoch, regularization, optimizer, and loss function. The specific parameter settings are shown in Table I. The learning rate is gradually decayed, with a reduction to one-tenth after every one-third of the total epochs.

Table I. Model Parameter Settings

Parameter Name	Parameter Value
learning rate	0.01
batch_size	256
epoch	50
regularization	L1
optimizer	SGD
loss function	CrossEntropy

B. Experimental Results

To verify the effectiveness of the proposed model, experiments were conducted using models on three different datasets: cifar100, cifar10, and STL10. The experimental results are shown in the following tables. Table II presents the experimental results on the cifar100 dataset, Table III shows the results on the cifar10 dataset, and Table IV displays the results on the STL10 dataset.

From Table II, it is apparent that ResNet-KL achieves better accuracy compared to other networks with the same computational resources. However, it should be noted that the loss function in ResNet-KL includes the logic distillation loss, network loss, and feature distillation loss, which contribute to an increase in the loss function. During training, the proposed method exhibits a training time approximately 2 minutes faster than the standard ResNet training approach. This suggests a slightly lower computational complexity compared to the original method.

Table IV shows that most of the performance metrics have decreased on the STL10 dataset. This is because the majority of the images in the STL10 dataset lack annotations. Therefore, the information obtained during network training differs significantly from the previous two datasets, resulting in noticeable differences in accuracy and other metrics.

Table II. Accuracy and other performance metrics of the cifar100 dataset under

model	Recall (%)	ACC (%)	F1	Loss
ResNet18	97.217	82.03	0.947	1.741
AlexNet	92.41	77.58	0.954	1.439
VGG	94.796	80.47	0.953	1.646
DenseNet	90.063	83.009	0.9544	2.15
GooLeNet	93.94	76.701	0.9589	1.822
MobileNet	94.874	79.607	0.963	1.954
Xception	96.15	82.294	0.961	1.534
RepLKNet	97.279	83.511	0.984	1.227
ResNet-KL	96.416	84.75	0.870	8.2472

Table III. Accuracy and other performance metrics of the cifar10 dataset under

model	Recall (%)	ACC (%)	F1	Loss
VGG	89.738	85.13	0.8942	0.4413
ResNet18	93.66	86.06	0.9151	0.6631
ShuffleNet	93.02	83.02	0.9724	0.4349
Xception	91.844	87.06	0.9366	0.3995
RestNet50	93.892	87.28	0.9652	0.3949
ResNet-KL	94.616	87.71	0.8434	6.999

Table IV. Accuracy and other performance metrics of the STP10 dataset under

model	Recall (%)	ACC (%)	F1	Loss
VGG	68.70	57.32	0.6863	0.928
ResNet18	78.40	63.21	0.7852	0.822
ShuffleNet	73.60	62.71	0.7340	0.889
Xception	83.51	61.52	0.8346	0.824
ResNet-KL	78.616	63.47	0.7924	7.231

To validate the effectiveness of self-distillation, this study specifically conducted ablation experiments to demonstrate the improvement. The results of the ablation experiments are shown in Table V.

Table V. Results of ablation experiments.

model	Recall (%)	ACC (%)	F1	Loss
ResNet	93.66	86.06	0.915	0.66
ResNet+Feature distillation	94.063	83.97	0.812	4.7
ResNet+ Logical distillation	94.34	86.34	0.839	15.4
ResNet+ Logical Feature	93.91	87.19	0.849	7.99
ResNet-KL	94.616	87.71	0.843	6.99

The ablation experiments presented in the

above table demonstrate that the accuracy of ResNet generally improves when subjected to various self-distillation methods. Nevertheless, performing individual feature distillation yields a decrease in accuracy. This is attributable to the fact that feature distillation mainly occurs in the deepest layer, leading to a restricted amount of network information in comparison to the gradual accumulation of information that is facilitated by logic distillation across multiple layers. Additionally, across multiple experiments presented in the preceding tables, the performance of F1 measure is not very good. This may be due to data imbalance, as F1 considers both precision and recall, and both can be problematic when dealing with imbalanced data, leading to a lower F1 value.

In this study, the ResNet-KL network merges feature distillation and logic distillation techniques, followed by iterative pruning to progressively decrease the number of parameters. As a result, the accuracy of ResNet-KL improves by 1% compared to the original ResNet, and the recall rate increases by 3%. The effectiveness chart of the ablation experiments is shown in Figure 4.1, where the x-axis represents the experimental steps and the y-axis represents the test accuracy.

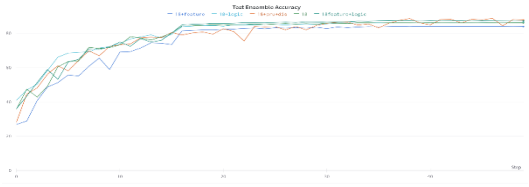

Figure 4.1: The test accuracy in the ablation experiments.

IV. SUMMARY

Although the proposed method demonstrates good performance in handling image classification tasks, it still has some limitations. For example, the training process of this network is relatively complex due to the use of self-distillation strategies and pruning operations, requiring more computational resources and time. Operating in resource-constrained environments can present significant challenges. Additionally, this study mainly focuses on small-scale image classification tasks and may face challenges when dealing with large-scale datasets. Future research can explore solutions to address the issue of class imbalance, such as data resampling, class weighting, or generative methods, to balance the dataset and improve overall classification performance. Furthermore, more efficient and lightweight network structures will be explored to reduce computational resource consumption and meet the requirements of handling large-scale datasets. The next step will also involve integrating ResNet-KL network with other advanced techniques, such as federated learning and semi-supervised learning, to further enhance the performance of image classification tasks.

REFERENCES

[1] Luo H, Zhai W, Zhang J, Cao Y, Tao D. Learning Affordance Grounding from Exocentric Images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2252-2261.

[2] Chowdhury P N, Bhunia A K, Sain A. What Can Human Sketches Do for Object Detection?[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 15083-15094.

[3] Wang Z, Li Y, Chen X. Detecting Everything in The Open World: Towards Universal Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11433-11443.

[4] Chen Y, Fan H, Xu B, Yan Z, Kalantidis Y. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3435-3444.

[5] Yao D, Zhi-li Z, Xiao-feng Z. Deep Hybrid: Multi-graph Neural Network Collaboration for Hyperspectral Image Classification[J]. Defence Technology, 2023, 23: 164-176.

[6] Arco J E, Ortiz A, Ramirez J. Uncertainty-driven Ensembles of Multi-scale Deep Architectures for Image Classification[J]. Information Fusion, 2023, 89: 53-65.

[7] Chen Y, Li J, Xiao H. Dual Path Networks[J]. Advances in neural information processing systems, 2017, 30.