FSelectMix: Regularization Strategy of Convolutional Neural Networks with Attention

Wanyong Tian[1], Shengyan Li[2], Jianjun Li[3], Yin Ye[4], Chin-Chen Chang[5]*

[1] CETC Key Laboratory of Data Link Technology, the 20th Institute, Xi'an, China
[2] School of Computer Science and Technology, Hangzhou Dianzi University, China
[3] School of Information Science and Technology, Hangzhou Normal University, China
[4] CEC Huada Electronic Design Co., Ltd., China
[5] Department of Information Engineering and Computer Science, Feng Chia University, Taiwan
frogustc@163.com, 960098070@qq.com, lijjcan@gmail.com, yeyin@hed.com.cn, alan3c@gmail.com

Abstract

Network regularization is a valuable approach to enhance network generalization. Unlike previous regularization methods that discard or mix regions at the image level, this paper proposes a regularization method called Feature Selection and Mixing (FSelectMix) based on attention. First, the FSelectMix method utilizes a dual-attention mechanism to select informative features crucial for task reconstruction. It generates adaptive confidence labels for the re-recognition of these features, enhancing the neural network's learning potential. FSelectMix significantly enhances convolutional neural networks' robustness and overall performance by operating at the feature level. Second, the proposed method introduces a multi-objective prediction task with a knowledge distillation network and an adaptive confidence dynamic adjustment strategy to leverage the reconstructed feature samples. This dual strategy not only refines the learning process but also ensures that the network adapts dynamically to varying levels of feature confidence, resulting in more reliable and accurate predictions. Finally, FSelectMix can be seamlessly integrated with existing data augmentation techniques, further boosting the model's performance across different levels. We implemented FSelectMix on the CIFAR10, CIFAR100, and Tiny-ImageNet datasets, resulting in significant performance improvements in all three. The experimental results validate the effectiveness and demonstrate its potential for broad application. Codes are available at https://github.com/ZhugeKongan/FSelectMix.

Keywords: Network regularization, Feature selection, Attention mechanism

1 Introduction

In recent years, as deep neural networks have continued to evolve, the complexity and diversity of deep learning tasks in various fields, such as object recognition [1-2], semantic segmentation [3], and image captioning [4], have increased significantly. To tackle these complex tasks, wider model width and deeper model depth tend to yield better performance [5]. However, larger models also mean more parameters, increasing resource requirements such as memory usage, parameters, operation counts, reasoning time, and power consumption [6]. Moreover, larger models tend to overfit more easily.

Also, depending on the feature extraction method, features can generally be divided into two categories: Traditional hand-crafted descriptors that include global hand-crafted features and locally handcrafted descriptors and deep-learning-based methods [7]. Even though feature extraction methods in deep learning encompass a wide range of architectures and techniques, evolving significantly over time, starting from VGG, Inception, and ResNet to Transformers, and so on. Therefore, network regularization is an excellent choice to improve model robustness and overall performance in these scenarios.

Regularization techniques play a crucial role in improving the generalization ability of deep neural networks, particularly convolutional neural networks (CNNs). It is especially important in many internet-based systems; for example, biometrics typically include fingerprint, iris, vein, and face recognition. When these existing models are applied to a real situation, they may show low accuracy depending on the light reflection and angle of the face. This is because various environmental factors and situations have not been applied to commonly used public face databases [8]. However, regularization techniques help to prevent overfitting and improve the model's robustness to noise and unseen data. There are two types of network regularization: image-level data enhancement [9-11] and feature-level network regularization [12-14]. The former involves purposeful occlusion, clipping, rotation, and adding noise to the image during training, with the goal of simulating possible real-world scenarios through image-level data interference, forcing the network to extract information from obscure features. Feature-level-based network regularization acts directly on the intermediate feature layer and expects all neurons to participate in representing the input data, reducing redundancy, and involving as many neurons as possible in the final prediction.

Moreover, when most convolutional neural networks are computationally and storage-intensive, and it is difficult

*Corresponding Author: Jianjun Li; Email: lijjcan@gmail.com
DOI: https://doi.org/10.70003/160792642026032702002

to deploy on mobile platforms and other micro-devices. Therefore, it is particularly important to compress and accelerate the network model and reduce the computational load and storage space of the convolutional neural network [15]. Regularization techniques based on features are suitable for CNN compression networks.

One of the earliest and most widely used regularization techniques is dropout [16], which randomly drops out neurons during training, forcing the network to learn redundant representations and become less reliant on any single neuron. This helps to reduce overfitting and improve the model's ability to generalize to unseen data. Although current regularization methods have alleviated the overfitting problem, they still have some disadvantages. Image-level-based regularization methods act on the source image, potentially altering the original information due to introduced interference [17]. Feature-level-based methods often have limited applicability, such as Shake-Shake [18] only being applicable to ResNext and ShakeDrop [14], requiring extensive hyperparameter tuning when applied to different models.

Regularization is a fundamental aspect of training convolutional neural networks. It addresses overfitting, improves generalization, enhances robustness, and facilitates the deployment of models in real-world applications. By incorporating these techniques, we can build more reliable, efficient, and effective CNNs that perform well across a variety of tasks and environments. However, these traditional regularization techniques often treat all features equally, which can be suboptimal for complex tasks where certain features may be more informative than others. To address this limitation, attention mechanisms have emerged as a powerful tool for selectively focusing on the most relevant features in the input data.

Besides, attention mechanisms have been successfully applied to various tasks, including machine translation, image captioning, and natural language processing. By allowing the model to focus on the most informative parts of the input, attention mechanisms can significantly improve the model's performance and interpretability.

In this manuscript, we propose an efficient network regularization strategy called FSelectMix, which incorporates an attention mechanism, as shown in Figure 1. FSelectMix randomly selects features from two images within the same batch as input, then uses the attention mechanism to select representative features that are difficult to learn. Our main contributions can be summarized as follows:

1) We identified limitations in existing regularization and data augmentation methods and proposed a novel approach called FSelectMix. This method selects representative features from two different images and mixes them using a dual-attention mechanism. This method improves the robustness and overall performance of convolutional neural networks (CNNs) while avoiding some of the issues of other approaches.

2) We introduced a multi-objective prediction task and a knowledge distillation network to leverage the reconstructed feature samples generated by FSelectMix, which improves the original model's classification accuracy and generalization ability. Additionally, we developed an adaptive confidence dynamic adjustment strategy to address the mismatch problem between images and labels.

3) FSelectMix is a versatile and efficient module that can be incorporated into various models at different training stages. It maximizes a network's learning potential and enhances its ability to handle complex tasks.

4) We demonstrated that FSelectMix, when combined with existing data augmentation methods, effectively mitigates overfitting caused by large models and limited data samples. Our experiments on CIFAR10, CIFAR100, and Tiny-ImageNet datasets show that FSelectMix outperforms previous approaches in terms of classification accuracy.

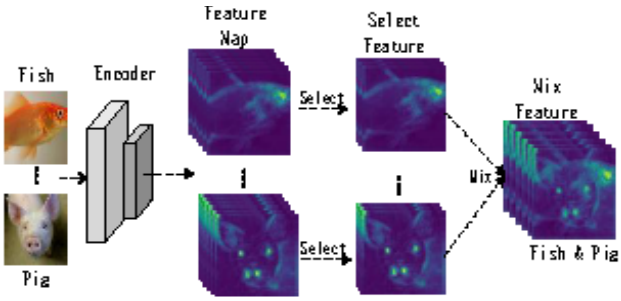


Figure 1. The proposed framework involves selecting two representative feature maps from different images as input and then mixing them by channels to generate new feature samples

The remainder of this paper is structured as follows. Section 2 describes related works. Section 3 presents the proposed approach, database, the structure of the neural network models used in this study. Section 4 gives the experimental results and detailed analysis. Finally, Section 5 concludes the study and discusses the future research direction in Section 6.

2 Related Works

In the context of CNNs, data aggregation and regularization aim to improve the model's generalization. While data aggregation can increase the variety of data the model sees, regularization ensures that the model does not become too specialized for this data.

Firstly, as a well-known technique for improving the completeness of training data and the robustness of models. Data aggregation involves techniques like data augmentation, where the original dataset is artificially expanded by applying transformations like rotation, scaling, or cropping to the images. This increases the diversity of the training data and potentially improves the model's robustness. DeVries et al. [9] proposed a Cutout method to simulate occlusions in training data, enabling the model to better learn how to use context information to predict or replace occluded information. Zhang et al. [10] mix different images, which expands the training dataset and improves the model's generalization ability on unknown samples. Yun et al. [11], DropConnect [20], DropPath [21], shake-shake regularization [18], and ShakeDrop regular-

ization [14]. The basic principle behind these methods is to inject noise into the neural network to avoid overfitting to the training data, thereby reducing redundancy and allowing as many neurons as possible to participate in the final prediction. For instance, Dropout sets some neurons to zero during training, while DropPath sets the entire layer to zero, introducing a lot of non-pixel information that can affect the accuracy of image classification. However, some of these methods are only suitable for small models. For example, Shake-Shake is only suitable for ResNet [22], DropPath is only suitable for multiple branches, and ShakeDrop requires a larger number of training sessions.

Secondly, data aggregation provides a more diverse dataset, which can indirectly benefit regularization by reducing the likelihood of the model overfitting to specific features or noise in the training data. Regularization in CNNs is a technique used to prevent overfitting by adding a penalty term to the loss function, which discourages overly complex models. Common regularization techniques include L1 and L2 regularization (also known as weight decay), dropout, and batch normalization. Regularization helps CNNs generalize better to new, unseen data by discouraging the model from relying too heavily on any single feature or training example.

Thirdly, CNN regularization improves the model's generalization and reduces the influence of data aggregation. However, the data is still huge and can not focus on the important information. To deal with this problem, as an effective feature extraction and enhancement method, the attention mechanism has been widely applied in many fields of deep learning [23-25]. The attention mechanism can record the positional relationship between information and measure the importance of different features based on the weight of information. Dynamic weight parameters are established through relevant and irrelevant choices of information features, and these attention weights are attached to the source features to strengthen the key information and weaken the useless information.

Inspired by the above, we aim to use attention to guide feature recombination in network regularization. We thus use the efficient SENet [23] to score the importance of channel feature maps. Different image features are combined and reconstructed based on the attention score to generate new samples for regularization training. The feature selection and mixing (FSelectMix) method improves the model's robustness and overall performance. We further use multi-objective prediction as the auxiliary supervision task and introduce a knowledge distillation network to improve the model's classification performance and generalization ability.

Finally, we propose an adaptive confidence dynamic adjustment strategy to generate confidence in the mixed feature samples, significantly alleviating the mismatching problem between image and label. The FSelectMix is a convenient and pluggable module that can be deployed in multiple locations of various models, providing high efficiency and model universality and allowing for the full exploration of the network's learning potential.

3 The Proposed Approach

In this section, we will provide a detailed description of the proposed feature selection and mixing architecture called FSelectMix, as well as the self-distillation and adaptive confidence techniques. The pair of images is first input to the feature extractor and then passed by the FSelectMix module, followed by a self-distillation module. The adaptive confidence dynamic adjustment strategy incorporates more sophisticated uncertainty estimation techniques to improve the accuracy and reliability of predictions.

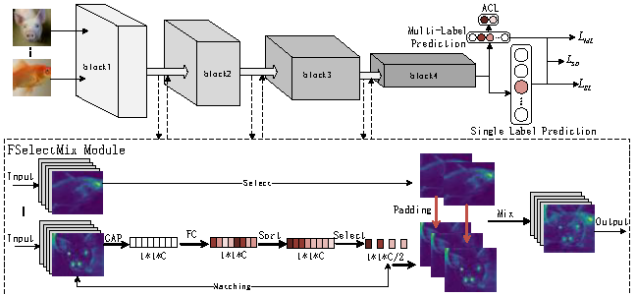


Figure 2. The main framework of our approach

(A multi-label prediction auxiliary task has been incorporated after the backbone to produce an adaptive confidence label (ACL). The proposed FSelectMix is explained in detail below, and it is inserted between the two blocks of the backbone by default.)

3.1 FSelectMix

As depicted in Figure 2, a collection of feature vectors $F = [F^0, F^1, \ldots, F^B]$, obtained from the middle layer of the network, serves as the input. Here, $F^{\{i\}} \in \mathcal{R}^{C\times H\times W}$, and B corresponds to the batch size. Subsequently, the channel attention of each feature vector is computed to facilitate feature combination. Our attention mechanism is inspired by SENet [24]. However, to emphasize the guiding function of attention as much as possible, we replaced the 1×1 convolution with a fully connected (FC) layer. As only a small number of attention layers are needed for guidance, the resulting mechanism is lightweight. We refer to this modified mechanism as efficient SENet (ESE), and its mathematical expression is given by:

$$A_{ESE} = sigmoid\left(F_{GAP} W_{FC}\right) \tag{1}$$

Where F_{GAP} is the global average pooling, and W_{FC} is the fully connected layer after GAP. Then, we sorted each group of feature vectors according to the attention score and extracted the corresponding channel feature maps in order with the stride of 2:

$$F^i_{selection} = F^i_{argsort(A_{ESE})[::2]} \tag{2}$$

The remaining feature maps we choose to fill from another image:

$$F_{padding}^{i} = F_{\{R^c \mid R^c \notin argsort(A_{ESE})[::2]\}}^{j} \quad (3)$$

Where, i_{th} and j_{th} are two groups of feature maps randomly selected in the same batch. Next, under the guidance of attention score, we will purposefully recombine the selected two groups of channel feature maps to generate reconstructed feature samples for regularization training:

$$F_{mix}^{i} = \left[F_{selection}^{i}; F_{padding}^{i} \right] \quad (4)$$

3.2 Self-Distillation

Our approach mixes img-features (F_A, y_A) and (F_B, y_B) generates new feature samples (F_{AB}, y_{AB}), as follows:

$$F_{AB} = \left[selection(F_A) : padding(F_B) \right] \quad (5)$$

$$y_{AB} = \left[y_A : y_B \right] \quad (6)$$

We approach it as a multi-objective classification problem. As depicted in the upper right of Figure 2, we introduced a multi-label classification layer after the backbone network. The output of this layer is denoted as $\hat{y}_M$ and is optimized using binary cross-entropy loss:

$$L_{ML} = L_{BCE}\left(\hat{y}_M, y_{AB}\right) \quad (7)$$

The original single-label prediction $\hat{y}_S$ serves as the standard training and testing output. The proposed FSelectMix is only utilized with a certain probability during training. Specifically, besides using multi-label classification as an auxiliary self-supervised task, we also proposed a knowledge distillation network from multi-label classification to single-label classification. To this end, we formulated a self-distillation loss L_{SD} as follows:

$$L_{SD} = KL\left(Softmax\left(\frac{1}{T} f_M\right) \middle| Softmax\left(\frac{1}{T} f_S\right) \right) \quad (8)$$

where f_M and f_S represent the logits of the multi-label and single-label classifiers, respectively, and T is the temperature hyper-parameter for the softmax function. The Kullback-Leibler divergence (KL) is used to measure the difference between the predicted probability distributions.

$$L_{SD} = L_{KL}(\hat{y}_M, \hat{y}_S) \quad (9)$$

Where L_{KL} represents the Kullback–Leibler divergence loss, $\hat{y}_M$ and $\hat{y}_S$ represent the prediction of multi-objective and single-objective layer, respectively.

3.3 Adaptive Confidence

For single-objective prediction, a cross-entropy loss is used for regular supervised training. However, when training a re-constructed feature sample (F_{AB}, y_{AB}) generated by FSelectMix, it contains both class A and class B features. Therefore, it is necessary to predict the probability of whether the sample belongs to class A and class B simultaneously as follows:

$$L_{SL} = \lambda_A L_{CE}\left(\hat{y}_S, y_A\right) + \lambda_B L_{CE}\left(\hat{y}_S, y_B\right) \quad (10)$$

Where, λ_A and λ_B are respective confidence of class A and class B in single-objective prediction $\hat{y}_S$. In normal training and testing stages, they are $\{\lambda_A = 1-\lambda_B | \lambda_B \in [0,1]\}$. For the confidence label of the proposed FSelectMix, different from previous methods based on the number of features or area proportion, we design an adaptive confidence generation strategy based on a multi-objective prediction layer to achieve the secondary distillation of multi-objective prediction knowledge. Specifically, the output of multi-objective prediction can be expressed as $\hat{y}_M = \{P_1, P_2, \ldots, P_N\}$ and we use this prediction to design a secondary knowledge distillation. That is, the prediction probability $[P_A, P_B] \in \hat{y}_M$ of class A or class B is taken as the confidence and applied to the loss of L_{SL}. The adaptive confidence can be expressed as:

$$\left[\lambda'_A, \lambda'_B\right] = softmax\left(\left[P_A, P_B\right]\right) \quad (11)$$

Substituting the confidence into L_{SL} get L'_{SL}. Finally, the training loss L can be expressed as:

$$L = L'_{SL} + \alpha L_{ML} + \beta L_{SD} \quad (12)$$

Where α and β are the hyperparameters of multi-label prediction and self-distillation loss, respectively. We set $\alpha = \beta = 0$ in the regular training and testing while $\alpha = \beta = 0.5$ during regularizing training where further improvement may be feasible using different ratios of losses.

4 Experiments

In this section, we will evaluate the effectiveness and advantages of the FSelectMix across multiple tasks.

The first performance evaluation for image classification is carried out on both CIFAR-10 and CIFAR-100 benchmark [19]. Three widely used CNN models have been adopted as the backbone, including ResNet [22], wide-Resnet [26] and PyramidNet [27]. Note that during training, we compared with some regularization methods such as Cutmix [11]. Instead of adding the proposed FSelectMix regularization at all stages of training, we add it with a certain amount of weights at each training epoch. The default rule for where and when to add FSelectMix is: For a backbone network with four blocks, such as Res-

Net50, we deploy FSelectMix in the first three-block layers with weights of 0.1, 0.2 and 0.1, respectively, and just one FSelectMix module will be triggered in each training as shown in Figure 2. For the backbone network with only three blocks, such as ResNet110, we deploy FSelectMix in the first two block layers with weights of 0.1 and 0.2, respectively. Also, FSelectMix can be combined with other image-level regularization methods, such as Cutmix. The triggered image-level regularization weight is set to 0.2 by default and different weights can be set in experiments.

This weighted deployment strategy allows for a nuanced integration of the regularization technique, promoting effective learning at different network depths. The following comparisons demonstrate the effectiveness in classic image classification tasks.

4.1 Image Classification

4.1.1 CIFAR-10 and CIFAR-100 Classification

Table 1 and Table 2 present a comparison between the baseline and the proposed FSelectMix on CIFAR-10 and CIFAR-100 datasets, respectively. There are 45000 training images, 5000 validation images, and 10000 test images in both CIFAR10 and CIFAR100 datasets. The overall experimental design is consistent with baseline methods.

Our proposed FSelectMix outperforms the baseline in all models. To provide a main reference for comparison, we also compare FSelectMix with CutMix. Since FSelectMix adapts the feature layer reconstruction through attention guidance, it can better explore the network's learning potential. It is noteworthy that our approach can be combined with various regularization methods, such as CutMix. The combination of FSelectMix and CutMix has achieved the highest Top-1 classification accuracy of 81.54%, 83.40%, and 84.38% on CIFAR-100, respectively.

Table 1. Classification Top-1 accuracies (%) on CIFAR-10

Model	#Params	Top-1 (%)
ResNet18	11.18M	92.36
ResNet18+FSelectMix	11.27M	**94.70 (+2.34)**
Wide-ResNet 40-2	2.25M	93.77
Wide-ResNet 40-2+FSelectMix	2.25M	**95.14 (+1.37)**
PyramidNet-100	3.90M	94.46
PyramidNet-100+ FSelectMix	3.91M	95.47 (+1.01)

Table 2. Classification Top-1 and Top-5 accuracies (%) on CIFAR-100

Model	#Params	Top-1 (%)	Top-5 (%)
ResNet18	11.27M	77.70	93.89
+cutmix	11.27M	80.04 (+2.34)	95.16
+FSelectMix(ours)	11.36M	81.09 (+3.39)	95.60
+FSelectMix+cutmix	11.36M	**81.54 (+3.84)**	**96.27**
ResNet50	23.71M	80.02	95.19
+cutmix	23.71M	81.70 (+1.68)	96.11
+FSelectMix(ours)	25.29M	83.09 (+3.07)	96.72
+FSelectMix+cutmix	25.29M	**83.40 (+3.38)**	**96.60**
PyramidNet-110	28.51M	81.02	95.82
+cutmix	28.51M	83.97 (+2.95)	96.71
+FSelectMix(ours)	28.59M	83.25 (+2.23)	96.72
+FSelectMix+cutmix	28.59M	**84.38 (+3.36)**	**97.25**

4.1.2 Tiny-ImageNet Classification

We further evaluate our method on Tiny-ImageNet [28] using ResNet18, ResNet50 and ResNeXt in Table 3. We follow the same experimental setup and training strategy as CIFAR100. The results show that the model with the Fselectmix regularization strategy is always better than the baseline and Cutmix. When both FSelectMix and Cutmix regularization strategies are used for network training, the network performance is further improved. Top-1 classification accuracy is improved by 2.19%, 3.21% and 4.35% in ResNet18, ResNet50 and ResNeXt, respectively. The results from these experiments underscore the effectiveness of FSelectMix in improving the performance and generalization of CNN models. By selectively focusing on informative features and dynamically adjusting regularization weights, FSelectMix provides a robust and efficient regularization strategy that can be seamlessly integrated with existing methods. Also, the performance evaluation on CIFAR-10 and CIFAR-100 demonstrates that FSelectMix significantly enhances the robustness and accuracy of CNN models. By deploying FSelectMix with strategic weighting and combining it with other regularization methods, our approach offers a powerful solution for improving model performance in various image classification tasks.

4.2 Ablation Study

In this section, we conduct ablation experiments for many factors in FSelectMix to measure their contributions toward our outperforming results. We conducted an ablation study in CIFAR-100 dataset using the same experimental settings in Section 4.1.1.

4.2.1 Comparison Against State-of-the-art Regularization Methods

As shown in Table 4, Cutmix achieves the highest 80.72% Top-1 classification accuracy in the image level, and Fselectmix achieves the highest 81.09% Top-1 classification accuracy in the feature level. In particular, the combination of FSelectMix and CutMix achieves the highest classification accuracy of 81.54%. Our FSelectMix combined with Cutout, Mixup, and CutMix is better than all other regularization methods.

Table 3. Classification Top-1 and Top-5 accuracies (%) on Tiny-ImageNet

Model	#Params	Top-1 (%)	Top-5 (%)
ResNet18	11.37M	62.36	83.72
+cutmix	11.37M	62.86 (+0.50)	84.44
+FSelectMix (ours)	11.46M	63.20 (+0.84)	83.94
+FSelectMix+cutmix	11.46M	**64.55 (+2.19)**	**84.84**
ResNet50	23.91M	63.39	84.27
+cutmix	23.91M	65.18 (+1.79)	85.99
+FSelectMix(ours)	25.70M	65.99 (+2.60)	86.38
+FSelectMix+cutmix	25.70M	**66.60 (+3.21)**	**86.76**
ResNeXt50	23.38M	63.00	83.50
+cutmix	23.38M	66.09 (+3.09)	86.42

+FSelectMix(ours)	25.17M	66.31 (+3.31)	86.53
+FSelectMix+cutmix	25.17M	**67.35 (+4.35)**	**87.05**

Table 4. Comparison of state-of-the-art regularization methods on CIFAR-100

Model	Top-1 (%)	Top-5 (%)
ResNet18	77.70	93.89
+ESE(Baseline)	78.34	94.59
+Cutout	78.22	94.41
+Mixup	79.63	94.78
+Cutmix	**80.72**	**95.86**
+StochDepth	77.85	94.93
+DropBlock	78.12	94.85
+ShakeDrop	78.98	95.00
+Manifold Mixup	79.85	94.56
+FSelectMix(ours)	**81.09**	**95.60**
+ShakeDrop+Mixup	78.75	94.52
+ShakeDrop+Cutmix	80.63	95.83
+Manifold Mixup+Cutout	79.98	94.88
+Manifold Mixup+CutMix	80.55	95.34
+FSelectMix+Cutout	80.99	95.64
+FSelectMix+Mixup	81.17	95.55
+FSelectMix+CutMix	**81.54**	**96.27**

The comparison against state-of-the-art regularization methods, including a dual-attention mechanism, adaptive confidence labels, weighting at different stages, and combination with other regularization methods, clearly shows that FSelectMix offers significant improvements in accuracy and robustness. By focusing on feature-level regularization and leveraging a dual-attention mechanism with adaptive confidence labels, FSelectMix provides a powerful and dynamic regularization strategy. The consistent outperformance across different architectures and datasets solidifies FSelectMix's position as a highly effective regularization technique for convolutional neural networks.

4.2.2 Effectiveness of Adaptive Confidence and Self-Distillation

We investigate the effect of the combination of different losses in Table 5, where the adaptive confidence improves the model by 0.35% and the self-distillation improves it by 0.19%. The overall design of two knowledge distillations and adaptive confidence improves the network by 0.69%. The experiment is based on the ResNet50 backbone network on the CIFAR-100 dataset, which proves the effectiveness of FSelectMix. Among them, Row1 is the result of non-regularized training, and the Top-1 classification accuracy of the standalone feature reconstruction module is 82.50%; Row2 using standalone multi-label classification did not achieve the expected effect; Row3 proves the role of self-supervised tasks by assisting single-target prediction through multi-target prediction; Row4 proves the role of adaptive confidence; Row5 shows that the highest Top-1 classification accuracy of 83.09% is achieved under the joint action of self-supervision, knowledge distillation, and adaptive confidence.

Table 5. Results of FSelectMix with different combinations of losses for ResNet50 on CIFAR-100

L_{SL}	L'_{SL}	L_{ML}	L_{SD}	Top-1 (%)
✓				82.50
		✓		78.32
✓		✓		82.55
	✓	✓		82.90
	✓	✓	✓	**83.09**

The use of adaptive confidence labels resulted in a noticeable performance boost. The ability to dynamically adjust feature weights based on confidence levels contributed to better generalization and accuracy.

4.3 Visualization and Interpretation

In order to understand how FSelectMix works, we obtained some image instances from CIFAR100, and analyzed attention guiding the feature reconstruction process at different scales. Our feature visualization refers to the method [29]. Figure 3 shows the visualized results of FSelectMix in different blocks. Column 7 refers to the confidences of Input A, "insect," and Input B, "flower."

The results of adaptive confidence of Block 1 show that the reconstructed image has a higher probability of belonging to "flower". This is because the guidance effect of low-layer attention is limited, and the features of the "flower" are more obvious. In the deeper layer, with the increase of the number of channels, the ability to guide attention becomes stronger, and the feature map of "insect" in the reconstructed feature map is selected by attention. The confidence of reconstructed images will be more inclined to select the categories. Therefore, the category to which the reconstructed image belongs will become more biased towards the selected category as the guidance capability of attention improves. The changes in confidence from Block 2 to Block 3 for the same two images prove this point. The adaptive confidence is adaptively adjusted as the reconstructed image changes, achieving the expected effect.

The ablation study highlights the critical contributions of each component within the FSelectMix framework and illustrates the visual results. The dual-attention mechanism, adaptive confidence labels, dynamic weighting strategy, and combination with other regularization methods like CutMix all play pivotal roles in enhancing the model's performance. By understanding the impact of these factors, we can better optimize FSelectMix and apply it more effectively to various CNN architectures and tasks.

5 Conclusion

We comprehensively analyzed some existing data augmentation and regularization methods and identified some shortcomings. Based on this analysis, we proposed FSelectMix, which effectively leverages the attention mechanism to address these limitations. FSelectMix introduces a dual-attention mechanism that focuses on selecting the most informative features crucial for task reconstruction,

thus enhancing the neural network's learning potential. By regularizing and guiding CNNs with FSelectMix, we can enhance the robustness of the model. In addition, the reconstructed regularized samples can be used to generate an adaptive confidence label and facilitate knowledge distillation, forming part of a multi-objective prediction task. This dual strategy not only refines the learning process but also ensures that the network adapts dynamically to varying levels of feature confidence, leading to more accurate and reliable predictions. FSelectMix is a pluggable module that can be easily deployed in various tasks and models. This adaptability makes it a valuable addition to the toolkit of machine learning practitioners, as it can be seamlessly integrated with existing data augmentation techniques and other regularization methods. Moreover, it can be used in conjunction with other regularization methods to jointly exploit the learning potential of neural networks, leading to compounded benefits in model performance.

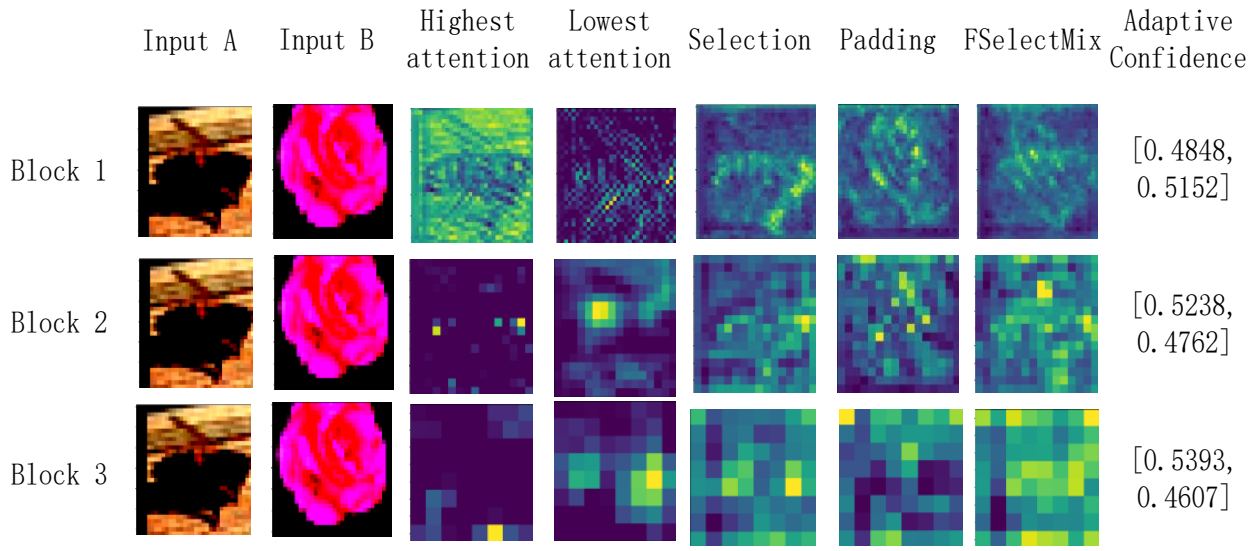


Figure 3. Visualization of FSelectMix results in different blocks

6 Future Work

The effectiveness of the proposed method has been verified in several scenarios. However, in the next step, several promising directions exist for expanding and enhancing the FSelectMix framework. One primary area of focus will be the application of FSelectMix to larger and more diverse datasets. While our current experiments on CIFAR10, CIFAR100, and Tiny-ImageNet have demonstrated significant improvements, extending our approach to datasets like ImageNet could provide a more comprehensive validation of its scalability and robustness. This would involve addressing the higher complexity and variety of images in such extensive datasets, which could further test the limits of our method's generalization capabilities.

We also aim to explore several key areas to enhance the FSelectMix framework further. First, we plan to extend our experiments to more complex and diverse datasets, such as ImageNet, to validate the scalability and robustness of our method in large-scale scenarios. Second, we intend to investigate the integration of FSelectMix with other advanced regularization techniques and data augmentation methods to boost performance further.

Additionally, exploring the application of our approach in different neural network architectures beyond CNNs, such as transformers and graph neural networks, could reveal broader applicability and benefits. We also aim to refine the adaptive confidence dynamic adjustment strategy, potentially incorporating more sophisticated uncertainty estimation techniques to improve the accuracy and reliability of predictions. Lastly, conducting ablation studies to better understand the contribution of each component within FSelectMix will provide deeper insights and guide further optimizations.

Acknowledgments

This work was supported in part by the National Science Fund of China, No. 62471170; Opening Fund of CETC Key Laboratory of Data Link Technology: CLDL-20202207.

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, 2016, pp. 770-778.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, F.-F. Li, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252, December, 2015.

[3] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, USA, 2015, pp. 3431-3440.

[4] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, USA, 2015, pp. 3156-3164.

[5] F. Seide, G. Li, D. Yu, Conversational speech transcription using context-dependent deep neural networks, *Interspeech*, Florence, Italy, 2011, pp. 437-440.

[6] A. Canziani, A. Paszke, E. Culurciello, An Analysis of Deep Neural Network Models for Practical Applications, *arXiv preprint*, arXiv: 1605.07678, May, 2016. https://arxiv.org/abs/1605.07678

[7] J. Xin, F. Ye, Y. Xia, Y. Luo, X. Chen, A new remote sensing image retrieval method based on CNN and YOLO, *Journal of Internet Technology*, Vol. 24, No. 2, pp. 233-242, March, 2023.

[8] S. G. Yu, S. E. Kim, K. H. Suh, E. C. Lee, Effect of Facial Shape Information Reflected on Learned Features in Face Spoofing Detection, *Journal of Internet Technology*, Vol. 23, No. 3, pp. 517-525, May, 2022.

[9] T. Devries, G. W. Taylor, Improved Regularization of Convolutional Neural Networks with Cutout, *arXiv preprint*, arXiv: 1708.04552, August, 2017. https://arxiv.org/abs/1708.04552

[10] H. Zhang, M. Cissé, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk Minimization, *arXiv preprint*, arXiv: 1710.09412, October, 2017. https://arxiv.org/abs/1710.09412

[11] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, J. Choe, CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features, *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019, pp. 6022-6031.

[12] G. Ghiasi, T.-Y. Lin, Q. V. Le, DropBlock: A regularization method for convolutional networks, *Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 10727-10737.

[13] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold Mixup: Better Representations by Interpolating Hidden States, *International Conference on Machine Learning*, Long Beach, California, USA, 2019, pp. 6438-6447.

[14] Y. Yamada, M. Iwamura, T. Akiba, K. Kise, Shakedrop Regularization for Deep Residual Learning, *IEEE Access*, Vol. 7, pp. 186126-186136, December, 2019.

[15] Z. Wen, C. Ye, M. Zhao, F. C. Ou Yang, A Compact Depth Separable Convolutional Image Filter for Clinical Color Perception Test, *Journal of Internet Technology*, Vol. 25, No. 2, pp. 331-340, March, 2024.

[16] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929-1958, June, 2014.

[17] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data Augmentation, *arXiv preprint*, arXiv: 1708.04896, August, 2017. https://arxiv.org/abs/1708.04896

[18] X. Gastaldi, Shake-Shake regularization, *arXiv preprint*, arXiv: 1705.07485, May, 2017. https://arxiv.org/abs/1705.07485

[19] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Technical Report, University of Toronto, April, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[20] L. Wan, M. Zeiler, S. Zhang, Y. LeCun, R. Fergus, Regularization of Neural Networks using DropConnect, *International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013, pp. 1058-1066.

[21] G. Larsson, M. Maire, G. Shakhnarovich, FractalNet: Ultra-Deep Neural Networks without Residuals, *arXiv preprint*, arXiv: 1605.07648, May, 2016. https://arxiv.org/abs/1605.07648

[22] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017, pp. 5987-5995.

[23] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, 2018, pp. 7132-7141.

[24] X. Li, W. Wang, X. Hu, J. Yang, Selective Kernel Networks, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, 2019, pp. 510-519.

[25] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional Block Attention Module, *arXiv preprint*, arXiv: 1807.06521, July, 2018. https://arxiv.org/abs/1807.06521

[26] S. Zagoruyko, N. Komodakis, Wide Residual Networks, *arXiv preprint*, arXiv: 1605.07146, May, 2016. https://arxiv.org/abs/1605.07146

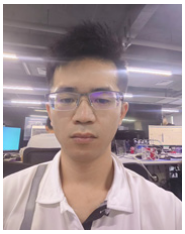
[27] D. Han, J. Kim, J. Kim, Deep Pyramidal Residual Networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017, pp. 6307-6315.

[28] Y. Le, X. S. Yang, Tiny ImageNet Visual Recognition Challenge, *Technical Report*, Stanford University, 2015.

[29] H. Fukui, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, 2019, pp. 10697-10706.

Biographies

Wanyong Tian received the B.Sc. degree in software engineering from Northwest University, Xi'an, China, and the Ph.D degree in computer science and technology from University of Science and Technology, Hefei, China. His research interests include task scheduling algorithms and DEVS modeling.



Shengyan Li is a graduate student at Hangzhou Dianzi University. He received a B.Sc. degree in computer science from Hangzhou Dianzi University in 2023. His research interests include image processing and video understanding algorithms.

Jianjun Li received a B.Sc. degree in information engineering from XiDian University, Xi'an, China, and the M.Sc. and Ph.D degrees in electrical and computer from The University of Western Ontario and the University of Windsor, Canada, separately. He is currently working at HangZhou Dianzi University as a professor. His research interests include micro-electronics, audio, video and image processing algorithms and implementation.

Yin Ye received a B.Sc. and M.Sc. degree in information engineering from XiDian University, Xi'an, China. She is currently working at Huada Electronic Design Corp., Ltd. as the Chief Engineer. Her work areas include information security technology, tinyML system design and IC design.

Chin-Chen Chang (Fellow, IEEE) received a Ph.D. degree in computer engineering from National Chiao Tung University, Hsinchu City, Taiwan. On numerous occasions, he was invited to serve as a Visiting Professor, a Chair Professor, an Honorary Professor, an Honorary Director, the Honorary Chairman, a Distinguished Alumnus, a Distinguished Researcher, and a Research Fellow by universities and research institutes. His current title is the Chair Professor with the Department of Information Engineering and Computer Science, Feng Chia University, since 2005. His current research interests include database design, computer cryptography, image compression, and data structures.