Robust Image Classification via Using Multiple Diversity Losses*

Yi Fang 2112112057@zjut.edu.cn
Zhejiang University of Technology, Hangzhou 310023, China

Wen-Hao Zheng zheng_wenhao@163.com
Zhejiang University of Technology, Hangzhou 310023, China

Qihui Wang 20050018@hznu.edu.cn
Qianjiang College, Hangzhou Normal University, Hangzhou 311121, China

Xiao-Xin Li* mordekai@zjut.edu.cn
Zhejiang University of Technology, Hangzhou 310023, China

Editors: Berrin Yancıkoğlu and Wray Buntine

Abstract

Many research works focus on the robustness of convolutional neural networks (CNNs) on image classification. Diversity loss has been demonstrated to be an effective method to boost robustness. However, the existing diversity losses did not fully consider the strong correlation between regional features when filters are locally activated. They focused on improving filter responses constraint with classification loss. However, diversity loss has deeper optimization space. We explore the combinations of different filter diversity losses and feature diversity losses. We enhance the orthogonality between pair-wise filters to make them more diverse and penalize irrelevance between regional response mappings. We make multiple combinations and propose several methods on improving orthogonality, which have different adaptations for different datasets and network models. We evaluate their effectiveness in experiment. Our combinations could improve the efficiency of robust image recognition.

Keywords: diversity loss; filter orthogonality; robust image classification

1. Introduction

Convolutional neural networks (CNNs) have achieved remarkable success on image recognition Dosovitskiy et al. (2020); Sun et al. (2005); Lu and Weng (2007); Petrou and Petrou (2010). However, it remains challenging to build a robust classifier that can handle varying data in new scenarios effectively, due to the inevitable domain shift when the actual data is encountered at test time. As shown in Fig. 1, handwritten digits can have different distributions caused by different factors, such as various writing styles Hull (1994), different application scenarios Neto et al. (2020), and the ray transformation.

A common solution is using transfer learning. That is, using a small well-labeled dataset to predict relevant but unlabeled samples in a target domain, so that the distribution of the

* Corresponding author: Xiao-Xin Li. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grants LGF22F020027, GF22F037921 and LGF20H180002, and in part by the National Natural Science Foundation of China uncer Grant 62271448.

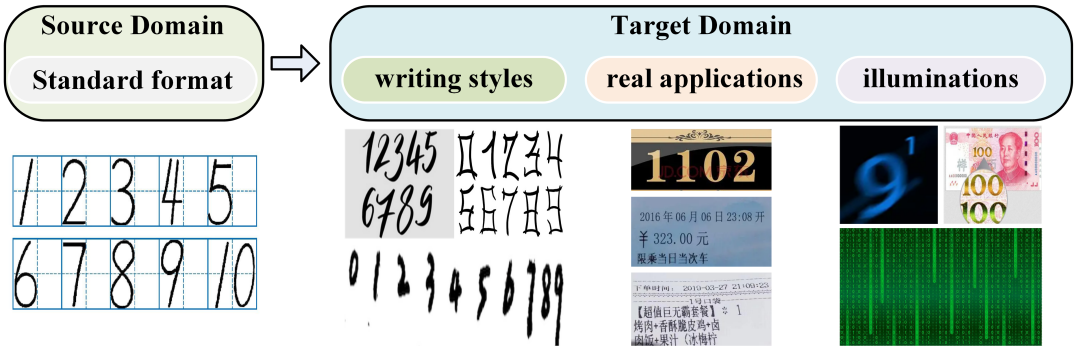

Figure 1: Examples of domain shift in handwritten digit recognition.

target data is as close as possible to that of the input data. However, its actual deployment is difficult to advance in depth, and the two most significant issues are: *i)* the dependence on labeled training data does not allow for manual labeling of all test data and is expensive; *ii)* the feature distribution of input data will produce unpredictable displacement changes after entering the CNN. The two problems are closely related. In convolutional space, the input data is distributed in the form of feature scatter points, and the well trained feature distribution in CNN implicitly reflects the sub distribution of the input image. On the contrary, models with good robustness and generalization ability often exhibit rich feature responses and fully activate important semantics.

We therefore explore the correlation between filter distributions in convolutional space and input distributions. If a specific filter only captures a typical feature, it can naturally weaken its response to noise signals, and thereby achieves a *natural* robustness.

Inspired by this idea, Zhuang *et al.* Zhuang et al. (2021) proposed a simple but effective method to explain the activation of a filter kernel in a CNN model and pointed out that training networks obtained by suppressing other visual features unrelated to the activation of specific kernels are naturally interpretable. However, this natural interpretability is very fragile, due to the strong correlation of feature responses at adjacent spatial positions and multiple filters work together to respond to local image regions.

Then, how to enrich the filter responses as much as possible by weakening their correlations? The spatial activation diversity (SAD) loss in the work of Yin et al. (2019) inspired us greatly. The feature of a region can be strengthened or weakened by punishing different filter weights, and the correlation between the filter responses can be weakened by combining the modified classification losses to enrich the useful features in convolution space.

In this work, we focus on the potential variations of the SAD loss Yin et al. (2019). We find that the number of filters used in the SAD loss is often very large so that the SAD loss is hard to be used in practice. To ensure that diverse filters are well learned without introducing a large computational cost, we, inspired by the method in MobileNets Howard et al. (2017), proposed a new method to calculate the similarity between filters . Specially, the filter kernels can be decomposed into multiple low-dimensional sub-filters along the channel direction, and their similarity can also be measured. We use two modified filter decomposition methods: *Point-Wise* and *Depth-Wise*. They can be used to reduce the convolutional computation and to improve the training speed. We introduce these two

combinations in detail in the method section and verify their effectiveness in the experimental section.

Further, in order to enhance the discriminability between feature responses, we have further improved the orthogonal combination of response constraints. In the work Yin et al. (2019), they had proposed the second term response diversity loss combined with classifier loss could decorrelation filter response and we optimize it. Given the fact, when training a large sample set in a neural network model, a fixed batch (usually not 1) is often set. The size of the batch determines the number of samples to be trained at once. When updating the weight, it is also updated by batch, which usually affects the optimization process and training speed of the model. Based on this discovery, we orthogonal the batch. Specifically, from the perspective of four-dimensional mapping space, N can be used as the number of Batch channels in the sample direction and F as the number of channels in the number of filters. Orthogonal combinations of response graphs with larger than two-dimensional sizes can be made, and their combinations can be more diversified. We believe that the addition of F and N is conducive to weakening the strong correlation between adjacent responses and ensuring certain independence of features.

To summarize, our contributions are in three folds: *1)* We modified diversity loss to make it be more suitable for dealing with robust image recognition problems; *2)* We incorporate the ideas of **Point-Wise** and **Depth-Wise** into the filter constraint, which greatly reduces the convolutional computation and enrich feature diversity; *3)* We modify the filter response constraint to incorporate batches into decomposition combinations in high dimensional space, thereby reducing the correlation between characteristic responses.

2. Related Work

2.1. Understanding Features on Convolution

Deep CNNs have been pushing the frontier of visual recognition over past years. Besides recognition accuracy, the research community's strong need for understanding deep cellular neural networks also prompted the development of tools that dissect pre-trained models to visualize how they make prediction process.

The attention model Rush et al. (2015) borrows from this feature and uses the attention mechanism to make the network model selectively "look" at different parts of the input data and apply adjustable weights so that the subtitles pushed to each prediction can respond effectively in a timely manner. Its idea is basically in line with ours. However, attention mechanisms often require large-scale data labeling, which introduces additional costs compared to traditional models. In addition, it may ignore some insignificant but important semantic information, resulting in information loss. We want to study interpretability from the characteristics of the model itself, without additional self-attention.

One potential way is to provide annotations for learning filters with local activation and building structured representations from the bottom up. However, this is challenging to execute and costly to manually mark in practice, which opposes our initial goal of machine learning for convenience, and is also worse than end-to-end learned filters. In Yin et al. (2019), interpretability is incorporated into the activation of filters and their responses, and spatial activation diversity loss is proposed to encourage filters to be fully activated locally and to capture more discriminating visual cues, especially in blocking face recognition.

2.2. Kernel Orthogonality in Neural Networks

Filters play a significant role in convolutional layer and each filter has an image feature response that it pays attention to, such as vertical edge, horizontal edge, color, texture, etc.

Orthogonal kernels are shown to stabilize the training of CNNs and make more efficient optimization Wisdom et al. (2016); Arjovsky et al. (2016); Lezcano-Casado and Martınez-Rubio (2019). Orthogonal weight initialization is proposed in Mishkin and Matas (2015); Saxe et al. (2013). However, this orthogonality may not last as training progresses. In order to ensure that weights in the whole training process are orthogonal, optimization methods based on Stiefel manifolds are used in Harandi and Fernando (2016); Huang et al. (2018); Miyato et al. (2018), and further extended to convolutional layers in Miyato et al. (2018). However, their work is not directly focusing on how to improve the orthogonal measures. And it is more complicated.

Feature redundancy is another character in deep convolution. Many works use the redundancy to compress or speed up networks He et al. (2017); Howard et al. (2017). However, the highly nonuniform spectrum may contribute to the redundancy in CNNs. To overcome the redundancy by improving feature diversity, multi-attention Zheng et al. (2017), diversity loss Li et al. (2018), and orthogonality regularization Chen et al. (2017) have been proposed.

2.3. Spatial Activation Diversity Constraint

Novotny *et al.* Novotny et al. (2017) proposed a spatial activation diversity loss (SAD-Loss) for semantic matching by penalizing the correlations among filters weights and their responses. The diversity constraint is implemented by two diversity losses $\mathcal{L}_{\text{SAD}}^{\text{filter}}$and $\mathcal{L}_{\text{SAD}}^{\text{response}}$, encouraging the orthogonality of the filters and of their responses, respectively. $\mathcal{L}_{SAD}^{filter}$ penalizes their correlations to make filters orthogonal:

$$\mathcal{L}_{\text{SAD}}^{\text{filter}}(F) = \sum_{i \neq j} \left| \sum_{p} \frac{\left\langle F_i^p, F_j^p \right\rangle}{\|F_i^p\|_F \left\|F_j^p\right\|_F} \right| \tag{1}$$

where F_i^p is the extract hypercolumn of filter F_i at the spatial location p : if p is a point, then $p = (u, v)$. Notly, orthogonal filters are likely to respond to different image structures, but this is not necessarily the case. Thus, the second term $\mathcal{L}_{\text{SAD}}^{\text{response}}$is introduced directly decorrelate the filters' *response maps* $\psi_k^I = \psi(F_k * \Phi(I))$, where the compact response mappings $\Phi(I)$ are compressed from the extract hypercolumns, I is the input image:

$$\mathcal{L}_{\text{SAD}}^{\text{response}}(I; \Phi; F) = \sum_{i \neq j} \left\| \frac{\left\langle \psi_i^I, \psi_j^I \right\rangle}{\left\|\psi_i^I\right\|_F \left\|\psi_j^I\right\|_F} \right\|^2 \tag{2}$$

Further, the regularized term using smoothing response maps $\psi'(I) = g_\sigma * (\psi(I))$ to optimize $\mathcal{L}_{\text{SAD}}^{\text{response}}$loss computing. Here, the Gaussian kernel g_σ plays a role to encourage filter responses to spread further apart by dilating their activations.

In above-mentioned work, they impose orthogonal constraint on convolutional filters and their responses. This measure strength the activation on local region and they called it *diversity constraint*. Eq. (1) and Eq. (2) show the proposed combination separately.

But they ignored that the decompositions of 4-Dimensional filter tensors and response tensors are not unique. Especially in filter diversity constraint, the combined forms could be changeable. We distinguish them and propose our modified filter constraint and response constraint aiming at orthogonal combinations.

3. Proposed Method

In this section, we will specifically introduce our improved filter diversity constraint and filter response diversity constraint, respectively.

For simplicity, we assume a convolutional layer has F filters of size $k \times k$ which are applied to an input consisting of N samples, each with C channels, height H and width W. Then, the layer has three 4D tensors associated with it: input $X \in \mathbb{R}^{N\times C\times H\times W}$, kernel weight $W \in \mathbb{R}^{F\times C\times k\times k}$ and $Y \in \mathbb{R}^{N\times F\times H'\times W'}$.

3.1. Depth-Wise Combination on Filters

Filters play an important role in "interpreting" input semantic information. The information input into the network are firstly captured by rich filters, then split by regional filters and respond locally, and distributed in mapping space in the form of feature scatter points. A model with superior recognition is often efficient for exacting useful information. So, how to make the interest features be fully activated, our main idea is focusing on transform *filter constraint* Novotny et al. (2017) , as shown in Eq. (1). According penalize different kernel weights to strengthen valuable features or weaken useless features, making the network be interpretable.

We further illustrate it by geometric representation in Fig. 2. For easy comparison with our proposed method, we dub it *Point-Wise* Filter Diversity Loss (pFDL) combination here. We first discuss the 4D kernel weight decomposition on convolutional filters. A single filter is described as a cubic with 3-Dimensional size ($C_{in} \times k \times k$) and group number is C_{out}. Considering the orthogonality of filters at the same spatial position p, it would be two understandings: i) orthogonality between 3D filters; ii) further consider the internal 2D sub-filter orthogonality.

If we discuss the first combination, the constraint is extended towards C_{out} depth direction and then there are totally $C_{out}(C_{out}-1)/2$ pairs, we calculate values along 3D filters at spatial position p (ideally point).

Furthermore, we make a deep analysis on 2D sub-filter orthogonality. A 3D filter can be understood as C_{out} numbers of sub-filter planes, which can be unfold to a column or a row. As shown in the *internal Pair-Wise* part of Fig. 2, a 2D plane on C_{in} direction can be stretched to 9 sub-boxes if $k = 3$. We further calculate the orthogonal values between these boxes and this constraint is stronger. Then, the orthogonal cosine similarity of 2D sub-filters f_i can be represent as:

$$\mathcal{L}_{\text{Internal Pair-Wise}}(F) = \left| \sum_{m,n\in k^2 \cup (m\neq n)} \frac{\langle f_i^m, f_i^n \rangle}{\|f_i^m\| \, \|f_i^m\|} \right| \tag{3}$$

in which the size of a single sub-filter is $k \times k$, which point m and n both $\in k^2$.

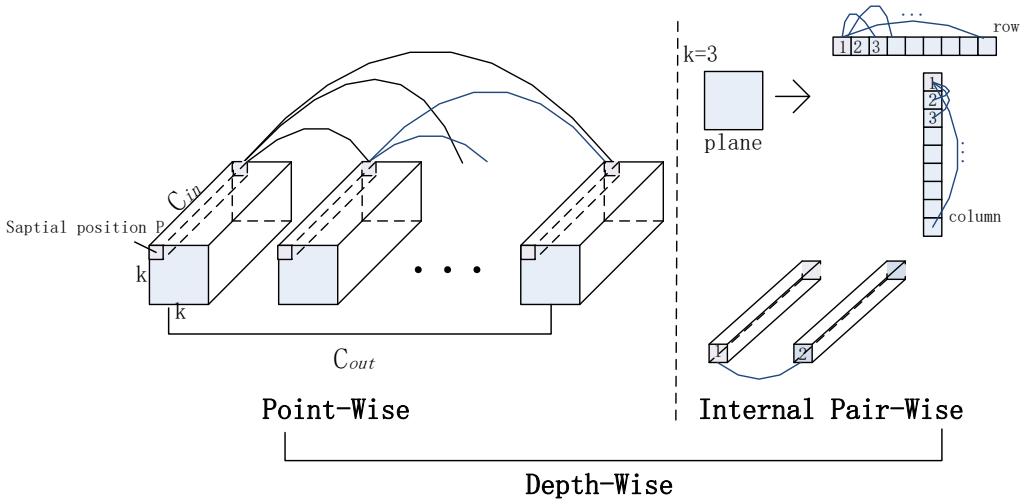

Figure 2: Geometric Representation of pFDL and dFDL.

We combine Eq. (1) and Eq. (3) and dub it *depth-wise* Filter Diversity Loss (dFDL). Its constraint is stronger than the front one. Summarily, our modified filter diversity constraint can be:

$$\mathcal{L}_{p\mathrm{FDL}}(F) = \sum_{i \neq j} \left| \sum_{p} \frac{\left\langle F_i^p, F_j^p \right\rangle}{\left\| F_i^p \right\|_F \left\| F_j^p \right\|_F} \right| \tag{4}$$

$$\mathcal{L}_{d\mathrm{FDL}}(F) = \sum_{i \neq j} \left| \sum_{p} \left| \sum_{m,n \in k^2 \cup (m \neq n)} \frac{\langle f_i^m, f_i^n \rangle}{\| f_i^m \| \, \| f_i^m \|} \right| \right| \tag{5}$$

3.2. Channel-Wise Combinations on Responses

In this section, we aim at another aspect: weakening the strong correlation between adjacent features to make network be more interpretable. This somewhat frees up the design of complex network structures and focus on fully understanding the depth features themselves.

Previous work Novotny et al. (2017); Yin et al. (2019); Zhang et al. (2018) had proposed that the second term $\mathcal{L}_{\mathrm{SAD}}^{\mathrm{response}}$ is introduced to directly decorrelate the filtersâĂŹ response maps $\psi_i(\mathrm{I})$ and further regularized by using channel-wise Gaussian kernel, which is encouraged to expand their activation. The *filter response constraint* is described as Eq. (2). In order to make features be more discriminative, Novotny *et al.* proposed to combine the second term $\mathcal{L}_{\mathrm{SAD}}^{\mathrm{response}}$ with classifier loss.

Inspired by filter-decomposition idea in Section 3.1, we think the constraint strength of combination would have a positive effect on decorrelate feature response. For a simple 2D feature map, there is only one combination to measure the similarity between it and the 2D feature maps mapped by neighboring filters. However, combinations of similar measures in higher dimensions are often not unique, and we focus on achieving different orthogonal combinations of **channel-wise** directions.

Conveniently comparison, we rename Eq. (2) to $\mathcal{L}_{i\text{RDL}}$, in which the filter response mappings (F, H', W') are activated by 3D filters F and sum them. Considering the 2D mappings inside 3D filters, we then decompose (N, F, H', W') and calculate the loss between 2D response mapping (H', W'). The formula is:

$$\mathcal{L}_{o\text{RDL}}(I;\Phi;F) = \sum_{i\neq j}\left|\frac{\left\langle \psi_i^I, \psi_j^I \right\rangle}{\left\|\psi_i^I\right\|_f \left\|\psi_j^I\right\|_f}\right| \tag{6}$$

Further, we put samples N into filter response loss. Sample size N and filter size F could be thought as two channels in different directions of 2D response mapping (H', W'). Then, wise channel N, F and calculate the similarity measure of strong mapping $(N * F, H', W')$. We defined it as ψ_k^I and the number of k is $f(N) = N * F$, which the formula is as follows:

$$\mathcal{L}_{a\text{RDL}}(I;\Phi;F) = \sum_{i\neq j}\left|\frac{\left\langle \psi_i^I, \psi_j^I \right\rangle}{\left\|\psi_i^I\right\|_{f(N)} \left\|\psi_j^I\right\|_{f(N)}}\right| \tag{7}$$

the constraint is stronger than $\mathcal{L}_{i\text{RDL}}$ and $\mathcal{L}_{o\text{RDL}}$. This operation could reduce strong correlation between regional responses and enhance the feature discriminability.

3.3. Discussion

In above, we introduced several orthogonal combinations of filter diversity loss $\mathcal{L}_{\text{FDL}}$ and filter response diversity loss $\mathcal{L}_{\text{RDL}}$. The constraint degree of them are different and the effectiveness would be discussed latter in experiment section.

Further, the regularized hyperparameters could also affect the accuracy. To stabilize the model training, we experimental finding optimal values. Combined with classifier loss $\mathcal{L}_{\text{Soft}}$, the total function could be :

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Soft}} + \alpha\mathcal{L}_{\text{FDL}} + \beta\mathcal{L}_{\text{RDL}} \tag{8}$$

We believe that diversity is transitive, and shallow-layer diversity would enrich deep feature responses to a certain extent and the modified diversity loss is applied to certain deep convolutional layer. Shown in Fig. 3 , the depth of lattice color shows intuitive representation of cosine similarity between pair filters without or with diversity loss. Applying diversity loss could reduce the high correlation between filters in some degree, then enrich the diversity of filter responses.

4. Experiments

We will demonstrate the effectiveness of our proposed method in this section.

4.1. Experimental Setting

In order to make our proposed method be more adaptable, we also conduct similar experiments on VGG16. The setting of batch size is set by considering the size of the utilized

Figure 3: Block diagram of cosine similarity results for pairwise filters in LeNet5 Conv1 trained on MNIST. The size of filter a_i is 3×3.

datasets and the complexity of the adopted network architecture. Experiments were performed on five GPUs. SGD and RMSProp were used as the optimization method. We initialized the learning rate as 0.01, which was divided by 10 after the 16K and 24K iterations.

The effect of our method were evaluated on the following datasets.

MNIST Variation Larochelle et al. (2007)**.** MNIST Variation was generated by perturbing samples from the popular MNIST digits, which leads to several benchmarks: mnist-rot, mnist-back-rand, mnist-back-image, mnist-rot-back-image. The samples of MNIST Variation are shown in Fig. 4.

CIFAR-10. CIFAR-10 Krizhevsky et al. (2009) contains 60,000 color images with size 32×32, which can be categorized into ten classes: planes, cars, birds, cats, deer, dogs, frogs, horses, boats and trucks.

4.2. Digit Recognition on MNIST Variations

We compare the effects of the proposed diversity losses and their combinations by using LeNet5 LeCun et al. (2015) and SphereFace10 Liu et al. (2017) on MNIST-Variations. MNIST Variation Larochelle et al. (2007) was generated by perturbing samples (mnist-basic) from the popular MNIST digits, which leads to several benchmarks: mnist-rot, mnist-back-rand, mnist-back-image, and mnist-rot-back-image. The mnist-basic subset is used for training, while the other variations are used for testing. The training and test samples of MNIST Variation are shown in Fig. 4.

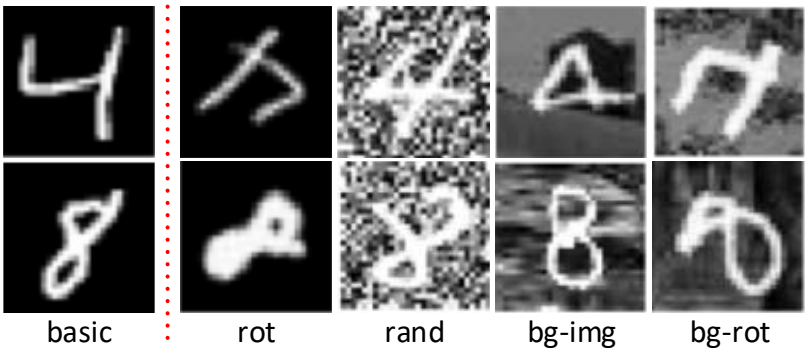

Figure 4: Training (left) and test (right) samples of MNIST Variation.

Table 1: The architecture of SphereFace10. Conv$d$.x denotes convolutional blocks in the d-th layer. $[k \times k, n] \times c$ denotes c cascaded convolution layers with n filters of size $k \times k$. Sd denotes the stride is d in the convolution layers. FC is the full-connection layer.

Layer	10-layer CNN
Conv1.x	$[3 \times 3, 64] \times 1$, S2
Conv2.x	$[3 \times 3, 128] \times 1$, S2; $\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 1$
Conv3.x	$[3 \times 3, 256] \times 1$, S2; $\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix} \times 2$
Conv4.x	$[3 \times 3, 512] \times 1$, S2
FC	512

As shown in Fig. 4, only the basic subset was chosen for training and the other four variation subsets for testing. The architecture of SphereFace10 was shown in Table 1.

We experimentally find that imposing DL on the first or second layer usually leads to superior performance. This is because the diversity can be transited layer by layer: if the first layer contains diverse filters, its output will contain rich features that will further enrich the filter kernels and filter responses of the second layer. However, enforcing DL on more layers can certainly boost diversity, but might not boost performance. We therefore impose the proposed DL on the first layer of LeNet5 and on Conv2.x of SphereFace10.

Table 2 compares the recognition results induced by the four diversity loss and their combinations. The first row in Table 2 shows the recognition results without using DL. As can be seen, for LeNet5, the FDLs (both pFDL and dFDL) lead to performance degradation although pFDL outperforms dFDL. On the other hand, for SphereFace10, the RDLs (both oRDL and aRDL) lead to performance degradation. It means that only applying FDL or RDL cannot guarantee performance boosting and combing them together may be necessary.

Table 2: Recognition results on MNIST Variations. Numbers with **bold faces** indicate performance boosting, while numbers with gray color represent performance degradation.

				LeNet5				SphereFace10				
*p*FDL	*d*FDL	*o*RDL	*a*RDL	rot	rand	bg-img	bg-rot	basic	rot	rand	bg-img	bg-rot
				39.48	68.91	68.28	27.10	96.77	37.60	33.63	60.84	23.17
✓				39.32	71.25	68.48	27.17	**96.80**	37.62	34.07	61.04	23.21
	✓			37.98	47.03	63.09	23.88	96.79	37.67	33.79	61.07	23.17
		✓		39.91	68.92	**69.90**	**28.00**	96.74	37.24	34.40	60.89	22.94
			✓	40.40	74.69	68.73	27.74	**96.80**	37.63	34.00	60.82	23.10
✓		✓		40.03	73.64	69.28	27.78	96.77	37.61	**34.41**	**61.14**	**23.99**
✓			✓	**40.63**	**74.68**	69.87	27.26	**96.80**	**37.68**	34.10	60.89	23.19

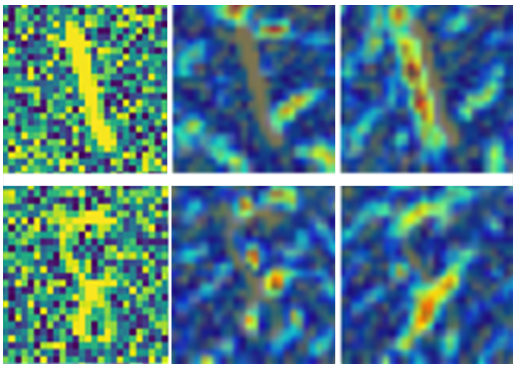

Figure 5: Visualizing the feature maps of the digits "1" and "8" by using Grad-CAM. Left: the input digit images, Middle: without using DL, Right: using *d*FDL and *o*RDL.

By combining FDL and RDL, as shown in the last two rows of Table 2, performance degradation is well resolved. Also, FDL and RDL achieves the optimal performance in most cases for both LeNet5 and SphereFace10.

Clearly, the model capacity could also affect the recognition performance. Generally speaking, larger network capacity could boost performance. However, by comparing the results of LeNet5 and SphereFace10, we observe that LeNet5 outperforms SphereFace10. This is caused by the fact that SphereFace10 produces small feature-maps caused by large stride in the convolutional operation.

In order to more intuitively observe the effect of the proposed diversity loss, we visualize the feature-maps produced by the first layer of LeNet5 by using Grad-CAM. Fig. 5 shows the visualization result. Clearly, DL enhances the significance of the output feature-maps greatly.

4.3. Object Recognition on CIFAR-10

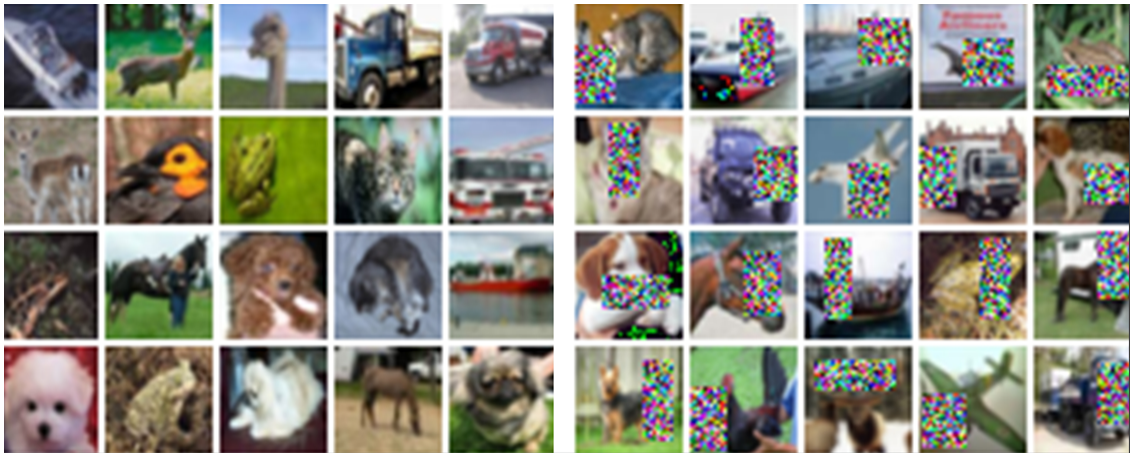

Figure 6: Training (left) and test (right) samples from CIFAR-10.

Table 3: Recognition results on CIFAR-10. "\" represents no diversity loss was used.

	\	pFDL	dFDL
\	56.25	59.18	58.34
oRDL	56.35	59.16	56.12
aRDL	56.59	59.21	**59.52**

We also evaluate our method on CIFAR-10 Krizhevsky et al. (2009). CIFAR-10 contains 60,000 color images with size 32×32 for ten categories: planes, cars, birds, cats, deer, dogs, frogs, horses, boats and trucks. To test the robustness of the proposed diversity losses, we add 20% random noise blocks to the test set of CIFAR-10. Training and test samples are shown in Fig. 6. VGG16 was selected for evaluation. We impose the DL on the first convolutional layer of VGG16.

Table 3 reports the recognition results. The optimal performance is achieved by combing dFDL and aRDL.

4.4. Face Recognition on AR

We further test the proposed diversity loss on the AR face dataset Martínez (1998). The AR dataset is commonly used for evaluating robust face recognition (FR) algorithms. It contains over 4,000 facial images from 126 subjects (70 men and 56 women). For each subject, 26 facial images are taken in two separate sessions. These images suffer different facial variations including various facial expressions (neutral, smile, anger, and scream), illumination variations (left light on, right light on and all side lights on), and occlusions (sunglasses and scarves). In this section, we selected a subset of the database that consists of 119 subjects (65 males and 54 females). The grayscale images were resized to resolution 112×92.

For training, we choose 8 unoccluded frontal view images with varying facial expressions for each of 119 subjects from the first session. For testing, we consider two separate test

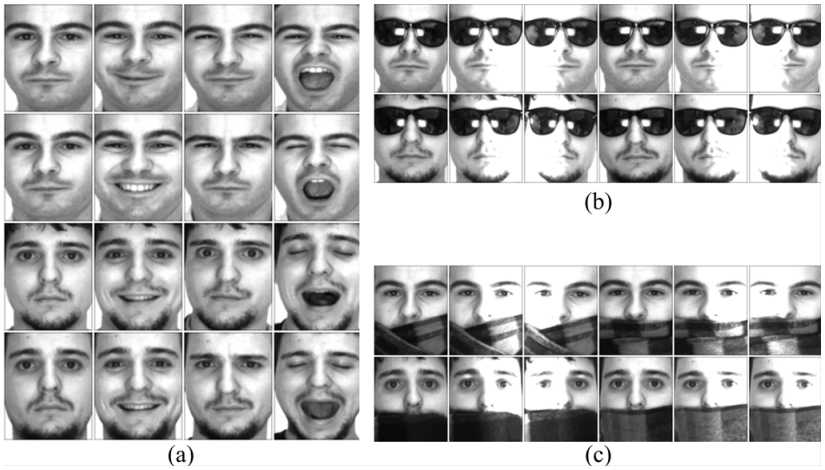

Figure 7: Samples from the Training Set (a), Test Set I (b), and Test Set II (c) of the AR dataset.

Table 4: Recognition results on AR dataset

	Test Set I			Test Set II		
	\	pFDL	dFDL	\	pFDL	dFDL
\	40.20	40.34	**42.02**	34.87	35.29	35.01
oRDL	41.04	41.32	41.60	**36.56**	35.71	34.31
aRDL	41.04	40.48	40.34	36.28	35.15	35.71

sets. The first/second test set contains 119×6 images with sunglass/scarf occlusion and illumination variations from the two sessions. Fig. 7 shows the sample images of the training and test sets.

SphereFace10 was selected for evaluation and the DL was imposed on Conv2.x. Batch size is set to 16. Table 4 reports the recognition results.

For Test Set I, we can see that all DL combinations outperform only using the softmax loss and dFDL achieves the optimal performance. For Test Set II, we can see that all DL combinations, except for the combination dFDL+oRDL, outperform only using the softmax loss and oFDL achieves the optimal performance.

5. Limitations

Our modified diversity loss $\mathcal{L}_{\text{FDL}}$ and $\mathcal{L}_{\text{RDL}}$ have multiple combinations and the constraint degrees then are variable. Based on it, we find the optimal combination is changeable which depends on the complexity of test datasets and which convolutional layer we expect to impose. Thus, our limitation is that we cannot develop a unified framework to regulate our orthogonal method. In addition, limited by lab equipment, we are not testing our approach on a larger datasets and a more open-source environment. We also find the regularized super-parameters in diversity loss would also affect the recognition effect to a great extent, which we think is worth further research and mining.

6. Conclusions

In this paper, we study the adaptive problem in the field of interpretable robust image recognition. We focus on spatially activated diversity loss and strive to explore the inner relationship of high-dimensional variables. Furthermore, motivated by MobileNets, we propose two orthogonal combination methods: Depth-Wise and Point-Wise. To reduce the correlation of regional feature response mappings, we also improve response space and propose two new combinations: out-channel and all-channels. We demonstrated the effectiveness of our proposed method through experiments. In the future, we would further explore and refine our approach against the limitations of our method.

References

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International conference on machine learning*, pages 1120–1128. PMLR, 2016.

Yunpeng Chen, Xiaojie Jin, Jiashi Feng, and Shuicheng Yan. Training group orthogonal neural networks with privileged information. *arXiv preprint arXiv:1701.06772*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mehrtash Harandi and Basura Fernando. Generalized backpropagation,\'{E} tude de cas: Orthogonality. *arXiv preprint arXiv:1611.05927*, 2016.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the international conference on Machine learning*, pages 473–480, 2007.

Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14, 2015.

Mario Lezcano-Casado and David Martınez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, pages 3794–3803. PMLR, 2019.

Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 369–378, 2018.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5): 823–870, 2007.

A.M. Martínez. The AR face database. Technical report,, 1998.

Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Arthur Flor De Sousa Neto, Byron Leite Dantas Bezerra, Estanislau Baptista Lima, and Alejandro Héctor Toselli. Hdsr-flor: a robust end-to-end system to solve the handwritten digit string recognition problem in real complex scenarios. *IEEE Access*, 8:208543–208553, 2020.

David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2017.

Maria MP Petrou and Costas Petrou. *Image processing: the fundamentals*. John Wiley & Sons, 2010.

Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Quan-Sen Sun, Sheng-Gen Zeng, Yan Liu, Pheng-Ann Heng, and De-Shen Xia. A new method of feature fusion and its application in image recognition. *Pattern Recognition*, 38(12):2437–2448, 2005.

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9348–9357, 2019.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836, 2018.

Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

Jia-Xin Zhuang, Wanying Tao, Jianfei Xing, Wei Shi, Ruixuan Wang, and Wei-shi Zheng. Understanding of kernels in cnn models by suppressing irrelevant visual features in images. *arXiv preprint arXiv:2108.11054*, 2021.