Proceedings of Machine Learning Research 95:502-517, 2018 ACML 2018

Collaboratively Weighting Deep and Classic Representation via l_2 Regularization for Image Classification

Shaoning Zeng ZSN@OUTLOOK.COM
1. Department of Computer and Information Science, University of Macau, Avenida da Universidade, Taipa, Macau, China
2. School of Information Science and Technology, Huizhou University, Avenue Yanda, NO. 46, Huizhou, Guangdong, China

Bob Zhang BOBZHANG@UMAC.MO
Department of Computer and Information Science, University of Macau, Avenida da Universidade, Taipa, Macau, China

Yanghao Zhang YZ16N18@SOTON.AC.UK
Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom

Jianping Gou GOUJIANPING@UJS.EDU.CN
School of Computer Science and Communication Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang, Jiangsu, China

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Deep convolutional neural networks provide a powerful feature learning capability for image classification. The deep image features can be utilized to deal with many image understanding tasks like image classification and object recognition. However, the robustness obtained in one dataset can be hardly reproduced in the other domain, which leads to inefficient models far from state-of-the-art. We propose a deep collaborative weight-based classification (DeepCWC) method to resolve this problem, by providing a novel option to fully take advantage of deep features in classic machine learning. It firstly performs the l_2-norm based collaborative representation on the original images, as well as the deep features extracted by deep CNN models. Then, two distance vectors, obtained based on the pair of linear representations, are fused together via a novel collaborative weight. This collaborative weight enables deep and classic representations to weigh each other. We observed the complementarity between two representations in a series of experiments on 10 facial and object datasets. The proposed DeepCWC produces very promising classification results, and outperforms many other benchmark methods, especially the ones claimed for Fashion-MNIST. The code is going to be published in our public repository[1].

Keywords: Sparse representation, Collaborative Representation, Collaborative Weight, L2 Regularization, Image Classification

1. Introduction

Machine learning methods have been applied to deal with various multi-media and computer vision tasks. Traditionally, linear models such as sparse representation (SR) Wright et al.

1. https://github.com/zengsn/research

(2009) and collaborative representation (CR) Zhang et al. (2012) have drawn much attention and gained promising results in image classification. Lately, nonlinear deep learning LeCun et al. (2015) models, e.g., ResNet He et al. (2016a) and VGG Simonyan and Zisserman (2014), have produced state-of-the-art results in many image-based tasks, including face recognition, object detection, video tracking, etc. Linear sparse models can be utilized to improve deep neural networks Xin et al. (2016). On the other hand, more and more conventional methods took deep features as input to gain more promising classification results Cai et al. (2016); Zeng et al. (2017c, 2018). However, recent studies showed that deep features from neural networks are usually designed for SVM-like classifiers Li et al. (2018). Using deep features as sole input in non-SVM classical models could be dubious. For this problem, we believe that the technique of linearly representing images can be applied to enhance nonlinear deep models.

Deep learning shows a very strong capacity to learn discriminative image features. CNN features off-the-shelf Sharif Razavian et al. (2014) were demonstrated to be powerful for recognition tasks. Learning deep features can help to obtain state-of-the-art results for different tasks like face recognition Wen et al. (2016b), scene recognition Zhou et al. (2014), person re-identification Xiao et al. (2016) and general image classification Shao et al. (2017); Valada et al. (2018). The good news is that many conventional machine learning methods can also learn credible features from images. In recent years, Sparse Representation (SR) Wright et al. (2009) via l_1 regularization has shown huge potential in feature extraction and image classification. On the other hand, l_2 regularization-based Collaborative Representation (CR) Zhang et al. (2012) can also build a similarly robust linear model. The l_2 regularization inside CR helps to create an equally discriminative but faster sparse representation Xu et al. (2017b). According to our observation, sparseness plays an important role in both linear and nonlinear models. It is likely for these two paradigms, deep and classic representations, to generate a new representation learning model when collaborating with each other.

In this paper, we propose a Collaborative Weight-based Classification method that brings deep and classic non-deep representation together, to implement a more promising image classification. We name it DeepCWC for short. The contribution of this work includes: 1) proposing a new classifier to integrate features from linear and nonlinear models, 2) giving an analysis on how black-box deep features work in a sparse classification model, 3) conducting image classification experiments on different CNN models and convolutional layers inside them, to demonstrate the performance of DeepCWC in a consistent and comprehensive way. The proposed method produces promising results on face and object recognition. In particular, it ranks first in recognition (97.66%) on the Fashion-MNIST dataset.

2. Related Work

The root inspiration comes from the popular deep residual network (ResNet) He et al. (2016a). ResNet keeps an identity map learned from the last layer, and applies it to next layer of learning. Then, it constructs a new building block $y = \mathscr{F}(x, \{W_i\}) + x$, as shown in Fig. 1(a). This explicitly allows these layers to fit a residual mapping, so as to make it easier for the residual to be zero (sparse) than to fit an identity mapping by a stack of nonlinear

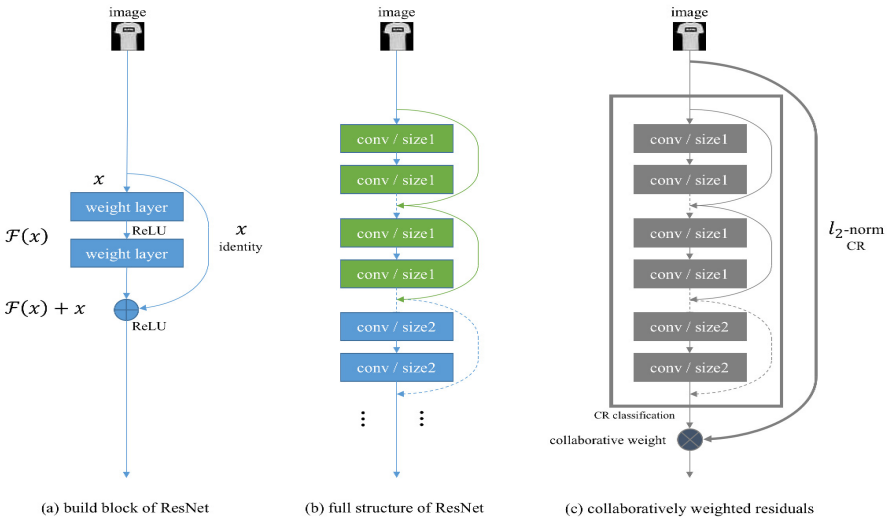

Figure 1: Collaboratively weighted residuals of CR and ResNet models.

layers. In this way, the discrimination learned in the previous layers will be propagated layer-by-layer. Also, there would be a linear transformation between some layers, if two connected blocks have a different dimension, as shown in Fig. 1(b). Denote the linear projection as W_s, where the building blocks become $y = \mathscr{F}(x, \{W_i\}) + W_s x$.

ResNet attracts great attention and progresses observably He et al. (2016b), while our main focus is the way how it utilizes the prior information. It is possible to include sparse learning, as prior information, in deep neural networks as well. For example, grouping multiple sparse regularizations for simultaneously optimizing deep neural networks Scardapane et al. (2017). Wen et al. proposed to learn a structured sparsity in deep neural networks to regularize the inside structures (i.e., filters, channels, filter shapes, and layer depth) Wen et al. (2016a). Afterwards, a fixed linear sparse filter can be cascaded with a thresholding nonlinearity to maximize sparsity in deep neural networks Xin et al. (2016). It becomes an emerging trend to utilize linear sparse models to collaborate with nonlinear neural networks.

As shown in Fig. 1(c), our idea has a similar structure following the building block of ResNet. The key is fusing the identity map learned from l_2-norm collaborative representation Zhang et al. (2012) to the result after deep residual learning. To simplify the implementation structure, the linear model is not injected into the building block of the neural network. Instead, it performs on the classifier. There are several reasons for this structure. Firstly, it helps to avoid overheads in the training process of the neural networks. Furthermore, it creates a more general structure that can be easily extended to other types of neural networks, which are not limited to ResNet. For example, we also implement this in Inception Szegedy et al. (2017) and VGG Simonyan and Zisserman (2014), which will be demonstrated in Sec. 4. The idea behind pairing nonlinear deep learning with an additional linear representation is to make the network more capable in different classification tasks.

However, the usage of the prior learned information is different in our implementation. ResNet adds up the learned x in model training, while the proposed DeepCWC will

introduce a collaborative weight in classification, which is obtained by an element-wise multiplication instead of addition. The next section explains the detailed implementation.

3. Deep collaborative weight-based classification

The key idea in Deep Collaborative Weight-based Classification is straightforward: using the model of Collaborative Representation (CR) Zhang et al. (2012) to learn a classifier from the original images and deep features in pairs. CR is based on l_2 normalization and emphasizes the collaboration among all samples in the representation. However, more and more evidence points to the fact that the collaboration requires help from the sparseness in the representation to maintain a high level of performance Akhtar et al. (2017). We believe that using deep features is one of the possible solutions.

3.1. Pair of residual learning

In CR based classification (CRC), the role of collaboration among classes is stressed, rather than sparsity in the representation, when representing a test sample. Let A denote the training samples selected from all C classes, while y is the test sample. Both A and y will be normalized to have l_2-norm. The representation of y by A can be denoted as an approximate linear problem $y \approx A\alpha$, where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_C]$ is the representation coefficient to be solved.

First of all, a regularized least square method Zeng et al. (2018) is used to solve the problem and perform the collaborative representation of the original image sample as follows

$$(\hat{\alpha}) = \arg\min_{\alpha}\{\left\|y - A \cdot \alpha\right\|_2^2 + \lambda \left\|\alpha\right\|_2^2\}, \tag{1}$$

where λ is the regularization parameter, which introduces a certain number of "sparsity" to the solution. The solution of this linear problem by using regularized least square can be derived as

$$\hat{\alpha} = (A^T A + \lambda \cdot I)^{-1} A^T y. \tag{2}$$

Let $P = (A^T A + \lambda \cdot I)^{-1} A^T$, such that P is a projection matrix that can be pre-solved and independent of the test sample y. The projection makes CR much faster than the conventional SR. It is noted that this operation may not fit in the memory of a large-scale dataset Yu et al. (2012). In our implementation, we used an incremental strategy Xiao et al. (2014); Ristin et al. (2014) to deal with this problem, despite the fact that there are other potential solutions, e.g., dictionary learning Cai et al. (2016); Xu et al. (2017a).

From this, we can obtain the coefficient vector $\hat{\alpha}_i$ related with the ith class. Typically, in SRC the coefficient is utilized to solve the representation residual of a specific class by $\left||y - A_i \cdot \hat{\alpha}_i|\right|_2$. Besides this, in the original CRC implementation, the l_2-norm of the sparse coefficient $\left||\hat{\alpha}_i|\right|_2$ was also added to obtain more discrimination when performing classification Zhang et al. (2012). Finally, the residual is obtained by

$$res_{img} = \left||y - A_i \cdot \hat{\alpha}_i|\right|_2 + \lambda \left\|\alpha_i\right\|_2. \tag{3}$$

At the same time, a side-by-side CR is performed on deep features. The so-called deep features are specific to the layer in the deep model. Theoretically, any layer can be utilized

to extract deep features. For example, the *global_pool* layer He et al. (2016a,b); Szegedy et al. (2017) in the ResNet and Inception models. Here, we denote the deep features from one specific layer (j) of the neural networks for all training samples as

$$B = fea_{cnn}(A, layer_j), \tag{4}$$

and therefore the query sample becomes

$$y_{cnn} = fea_{cnn}(y, layer_j). \tag{5}$$

The size of the feature set is determined by the design of the specific layer, where it is normally mismatched with the size of the original images. It is hard to simply perform integration on the feature level. Therefore, fusion of the feature pairs will be performed on the residuals, which have the size same as the class number.

Then, the representation coefficient obtained from the deep features by a similar CR process is

$$\hat{\beta} = (B^T B + \lambda \cdot I)^{-1} B^T y_{cnn}. \tag{6}$$

After that, the residual between the query feature set y_{cnn} and each class of training feature sets can be solved with the same method as Eq. (3)

$$res_{cnn} = \left|\left|y - B_i \cdot \hat{\beta}_i\right|\right|_2 + \lambda \left\|\beta_i\right\|_2. \tag{7}$$

3.2. Fusion based on collaboratively weighting

Our proposed method manages to retrieve this part of the missed information via a novel fusion operation. The fusion is performed on two residuals, since both have an equal dimension depending on the number of classes. Therefore, fusion on the residuals is not only straightforward, but also faster.

Let us denote the residuals solved from two groups of samples as $res_{img} = [d_{img,1}, d_{img,2}, \ldots, d_{img,C}]$ and $res_{cnn} = [d_{cnn,1}, d_{cnn,2}, \ldots, d_{cnn,C}]$, where C is the number of classes in the dataset. Then, the fusion via the collaborative weight is performed on the residual vector via an element-wise multiplication,

$$res_{fusion} = res_{img} \odot res_{cnn}, \tag{8}$$

where the residual entry related with the ith class is calculated by the collaborative weight. This weight means that each entry in the residual vector is assigned a weight solved by the collaborative representation of the original images. The information carried by this weight compensates the missing part of the abstract higher layers in a neural network. In this way, we obtain the final fusion residual. Although additional or weighted averages are a more common approach to perform fusion in many other methods Zeng et al. (2017a,c); Wen et al. (2018), they require a set of fine-tuned factors to obtain a good result. What is more, we observed a descending accuracy when adding up two residuals.

Finally, we classify the test sample to a class with minimal residual as follows

$$identity(y) = \arg\min_i (res_{fusion,i}). \tag{9}$$

The idea of our collaborative weight is simplistic and intuitive. The collaborative weights are determined by the relative contribution of each class from the original samples. Each residual of the deep features is overlapped by a weight solved using the collaborative representation of the original samples, in order to integrate its contribution.

3.3. Why deep features works in CR

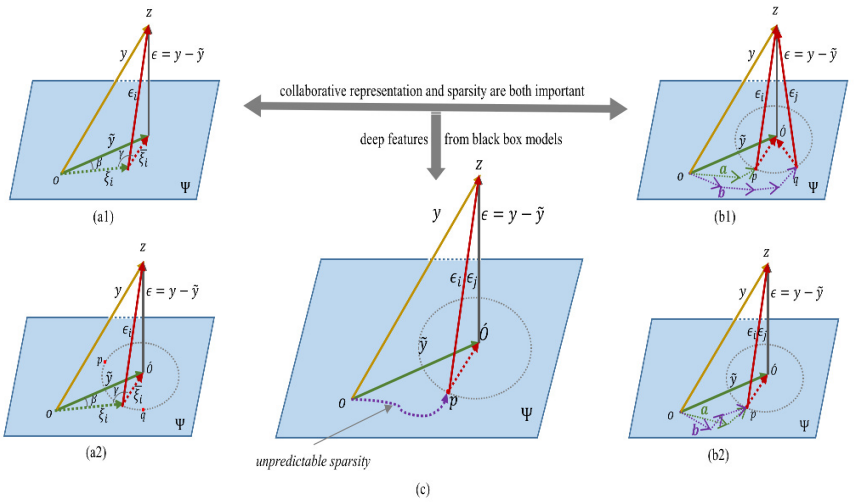

Figure 2: The sparsity from deep features in the CR model.

To answer this question, we first need to answer another question: What is the relationship between collaboration and sparsity? Many had tried to give an answer with some considering sparsity as being more important Wright et al. (2009); Deng et al. (2013). On the other hand, others insist that collaboration matters more Zhang et al. (2012); Peng et al. (2014). However, as for the rest, they treat collaboration and sparsity as equal factors Zeng et al. (2017b,a); Akhtar et al. (2017). So far, the last viewpoint well explains our proposed collaboratively weighting deep and non-deep representation.

The subspace occupied by the columns of a sparse dictionary Φ can be denoted as a set of Ψ. Fig. 2 shows the geometrical illustration of this subspace. A test sample y can be approximately represented by the columns of Φ, and the error is $\epsilon = y - \tilde{y}$. In addition, vector $\tilde{y}$ can be decomposed to ξ_i and $\tilde{\xi}_i$, as depicted in Fig. 2 (a1). According to Zhang et al. (2012), the angles β and γ together decide the robustness of the CR model. However, Fig. 2 (a2) shows that it is likely to have more than one right answer, which is depicted as the circle. The distances of pz and qz are the same. Therefore, CR by itself without considering sparsity may not be robust enough Akhtar et al. (2017).

Fig. 2 (b1) and (b2) show how the sparsity can help in the CR model. In these two cases, where (b1) $\xi_i \neq \tilde{\xi}_i$ and (b2) $\xi_i = \tilde{\xi}_i$, a and b are two paths to points p and q. The distances are also the same $||\epsilon_i||_2 = ||\epsilon_j||_2$ in these two instances, while $a \neq b$ in (b1) and $a = b$ in (b2). This means that the class-specific residuals are equal in both cases, but construction of vectors ξ_i and ξ_j may be different. The components consisting of the path depict the sparsity in the representation, where fewer steps of ξ_i indicates a sparser collaborative representation coefficient α. Using sparser features in CR can help to produce a more robust classification, which is the very reason why DeepCWC works.

The black box deep model provides an unpredictable sparsity in the deep features. Currently, we can only accept this fact according to the largely contracted dimension of

the deep feature set, with feature learning being the most powerful characteristic of deep learning. As shown in Fig. 2 (c), the effort of CR on deep features can be painted as a random and unpredictable curve between o and p. This can be treated as potentially the most efficient path and is also the result observed in the experiments.

3.4. Why the fusion is positive

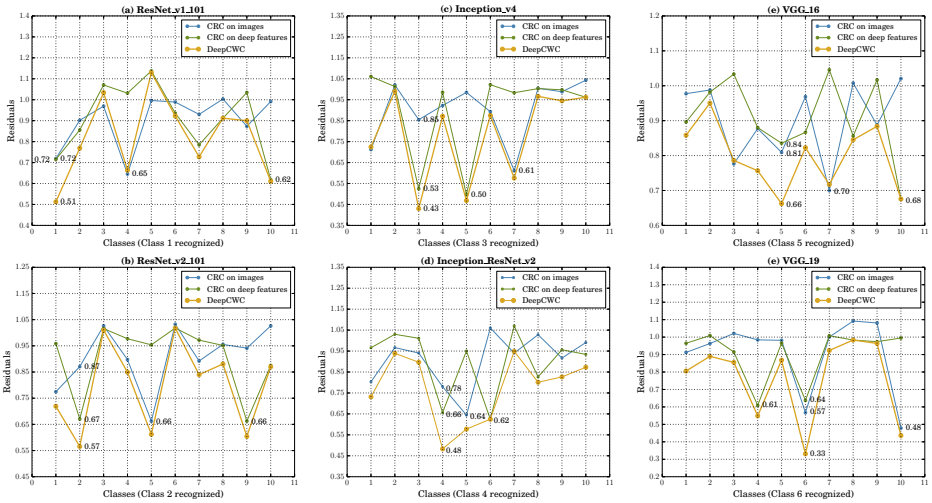

Figure 3: The residuals from the pair of CR representation and the final fusion.

When the deep features are ready and fit well in the CR model, the next problem is how to consolidate both of them into an united set. This is where the collaborative weight works. Previous work showed that well constructing the residuals is helpful to generate a robust sparse model, i.e., the two-phase sparse representation model Xu et al. (2011). To illustrate the impact of the weight, we captured some runtime data from our experiments, which is plotted in Fig. 3.

The purpose of the collaborative weight is to expand the more promising residuals, while restricting the other ones. The target class is selected by a final minimization, hence, we look for the smallest values. For example, the residuals of classes 1-10 are shown on the $ResNet\_v1\_101$ model in Fig. 3 (a). The correct class label is 1, where the distances of CRC on images and deep features are both below 1 (0.72). However, the minimal values of them are 0.65 and 0.62, respectively. This could lead to a wrong classification result (Class 4 and 10). After the fusion, the resultant residual becomes even smaller, resulting in the correct class being chosen (Class 1). This ensures that the classification will not be affected by other nearby classes. The same phenomenon can be found on other models, which are annotated in Fig. 3 (b) - (f). On the other hand, the classes with a larger distance value, e.g., $d_i > 1.0$, the fusion will make it much larger ($m * n > 1.0$, if $m > 1.0$ and $n > 1.0$). This in turn helps to avoid negative results.

Another point to note in the fusion is that the parameters do not need to be tuned. This is not like other conventional fusion schemas, which usually introduces one or two fusion factors Zeng et al. (2017a,c); Wen et al. (2018), where more parameters call for more tuning. The fusion result is determined by the residuals themselves. We barely need to look for an optimal parameter and maintains the effectiveness of DeepCWC.

4. Experimental results

This section describes the experiments to demonstrate the robustness and performance of the proposed method. First of all, six facial datasets, including FERET Phillips et al. (2000), MUCT Milborrow et al. (2010), Yale B Georghiades et al. (2001), Georgia Tech (GT) Chen et al. (2005), AR Martinez (1998), and ORL Cambridge, are selected to evaluate the performance of face recognition. These experiments were conducted due to the fact that CRC is usually applied to face recognition. Secondly, another set of experiments had been performed on some object datasets, including a leaf dataset Flavia [26], and three object datasets CIFAR-10 Krizhevsky and Hinton (2009), Fashion-MNIST Xiao et al. (2017) and COIL-100 Nayar et al. (1996), which are often utilized to evaluate deep learning methods.

Also, in order to evaluate the robustness after introducing the collaborative weight between the original images and the deep features, we extracted the deep features using multiple state-of-the-art deep CNN models, including ResNet_v1_101, ResNet_v2_101, Inception_v4, Inception_ResNet_v2, VGG_16 and VGG_19. All of these are trained previously on the ImageNet dataset Deng et al. (2009) in Google TensorFlow[2]. Our assumption is the proposed DeepCWC works on different deep CNN models. The feature extraction is performed on the TensorFlow-Slim library. Besides this, another goal is to investigate which layer of features in a CNN model are more suitable for collaborative weight. Based on the considerations, we conducted a set of relevant experiments and obtained the following results.

4.1. Experiments for face recognition

We ran experiments on six popular benchmark facial datasets. These datasets are relatively small. The smaller datasets do not contain enough samples to train a robust model by CNN, but we can extract the deep features using pre-trained deep CNN models. Our goal in this group of experiments is to compare the classification result between our proposed method and state-of-the-art methods. The best results are shown in Fig. 4 (a).

It is clear that the proposed DeepCWC method outputs a higher recognition accuracy than normal CRC, CRC using deep features and other state-of-the-art methods, no matter which CNN model is used to extract the deep features. It is uncertain that using deep features would generate a higher recognition accuracy than using the original images. For example, when utilizing the ResNet_v1_101 model to extract features of the AR dataset, CRC performs better on original images than deep features, as shown in Fig. 4 (a). And this is also true in some other experimental cases, which shows one of the limitations of a typical deep learning method. This is the very reason why we proposed the DeepCWC.

2. https://github.com/tensorflow/models/tree/master/research/slim#Pretrained

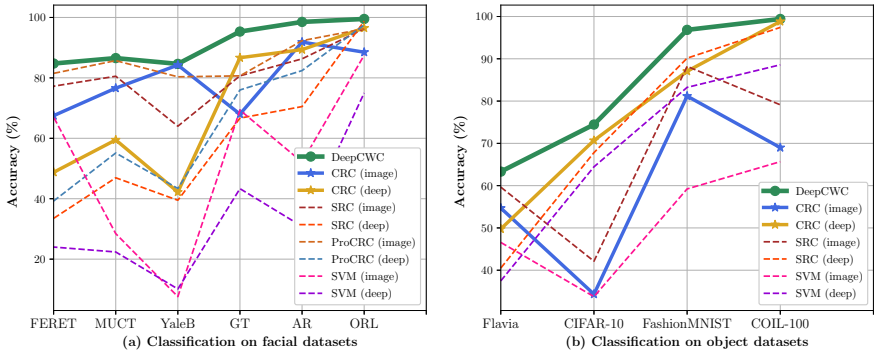

Figure 4: Face and object recognition accuracy comparison.

No matter which dataset, performing fusion of two feature sets based on the collaborative weight generates a higher recognition accuracy. Even in cases that merely use deep features without collaborative weight, e.g., on FERET, MUCT, Yale B, GT, AR and ORL, the proposed DeepCWC helps the recognition by fusing features from the original images and the CNN models.

4.2. Experiments for object recognition

The next set of experiments were performed on some leaf and object datasets, including Flavia (leaf), COIL-100, CIFAR, and Fashion-MNIST. The results are consistent on all datasets, as shown in Fig. 4 (b). Incorporating the deep features learned by ResNet_v1_101, the recognition accuracy (yellow) is much higher than the result on the original images (blue), except the result on the Flavia. However, DeepCWC further pushes the recognition up to an even higher level, and the improvements are stable on all datasets.

The highest accuracy is obtained on the COIL-100 dataset when using the first 60 samples in each class (83%) as the training samples. Deep features are beneficial to classification on this dataset, where DeepCRC (up to 98.83%) outperformed CRC (only 69.0%) by over 30%. That being said, the proposed DeepCWC still produces the highest accuracy of 99.42%, which reached a state-of-the-art level in recognition. On the Flavia leaf dataset, the improvement generated by collaborative weight is remarkable, though the accuracy is relatively lower, as shown in Fig. 4 (b). The results on the CIFAR-10 dataset get an improvement as well. And the improvement (the column Impr) is calculated by the rate of the accuracy from the DeepCWC over the higher one between CRC on images and deep features, and the improvements on the Flavia and Fashion-MNIST are up to 21.01% and 12.41%, respectively.

4.3. Experiments on different layers

Two versions of the ResNet pre-trained models, ResNet_v1_101 He et al. (2016a) and ResNet_v2_101 He et al. (2016b), are tested in this set of experiments. As described above, we borrowed a similar architecture idea from the deep residual network, as shown in Fig. 1. For this reason, we design the first implementation of DeepCWC based on ResNet. There

are 101 layers in the network, and we evaluate the proposed method on the features obtained from three layers, which are *global_pool*, *logits*[3] and *spatial_squeeze*[4] from shallower to deeper. The layer *spatial_squeeze* is the layer before the last convolutional layer, while the *logits* layer is before *spatial_squeeze*. These two layers have the same feature set size (1000). However, the *global_pool* layer is before this layer and has a larger size of 2048. The classification results are demonstrated in Fig. 5 (a) and (b).

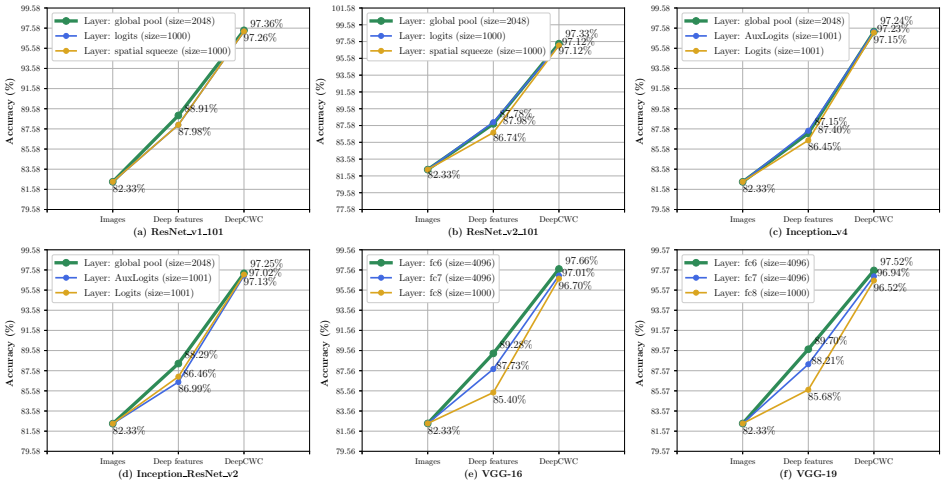

Figure 5: Accuracies with different layers of six deep models.

In both models, performing CRC on deep features obtained from each layer produces a higher accuracy than using the original images. DeepCWC generates an even higher accuracy than both of these. We observe exactly the same result (97.26%) in both *logits* and *spatial_squeeze* layers. For classification, the feature maps of the last convolutional layer are fed into fully connected layers followed by a softmax logistic regression layer Krizhevsky et al. (2012). The global average pooling Lin et al. (2013) is introduced to avoid overfitting in the fully connected layers. Using the features maps captured from the *global_pool* layer produces a slightly higher accuracy (97.36%). Every result reaches a state-of-the-art level, and is higher than all current implementations (See subsection 4.4). It is noted that the computation time increases due to a larger size (double) of the feature maps from layer *global_pool*. Therefore, the *logits* (or *spatial_squeeze*) layer should be a better choice when considering the balance between accuracy and speed.

To investigate the performance when using a different CNN model, two Inception models are evaluated in a similar way. They are Inception_v4 and Inception_ResNet_v2 models pre-trained on ImageNet. Besides the *global_pool* layer, the *Logits* and *AuxLogits* layers are also utilized to extract deep features, before being fed into to the linear CR model. A set of similar results are observed in the experiments, as shown in Fig. 5 (c) and (d).

3. *resnet_v1_101/logits* or *resnet_v2_101/logits*.
4. *resnet_v1_101/spatial_squeeze* or *resnet_v2_101/spatial_squeeze*.

Table 1: State-of-the-art accuracies on the Fashion-MNIST dataset.

Deep Model (∗)	Preprocessing	Highest Accuracy
CapsNet	None	90.6%
VGG16 26M parameters	None	93.5%
GoogleNet with cross-entropy loss	None	93.7%
MobileNet	Yes	95.0%
DenseNet-BC 768K params	Yes	95.4%
Dual path wide resnet 28-10	Yes	95.7%
WRN-28-10 + Random Erasing	Yes	96.3%
WRN40-4 8.9M params	Yes	96.7%
Ours		
DeepCWC with Inception_v4	None (global pool)	97.24%
DeepCWC with Inception_RN_v2	Yes (global pool)	97.25%
DeepCWC with ResNet_v2_101	Yes (global pool)	97.33%
DeepCWC with ResNet_v1_101	None (global pool)	97.36%
DeepCWC with VGG_19	None (fc6)	97.52%
DeepCWC with VGG_16	None (fc6)	**97.66**%

∗ Data claimed in https://github.com/zalandoresearch/fashion-mnist
∗ CapsNet in https://github.com/naturomics/CapsNet-Tensorflow#results

DeepCWC achieves an accuracy over 97% on deep features obtained from three layers. The highest result (97.24%) is the one with the largest size (2048) from the *global_pool* layer. In this case, the *AuxLogits* layer, with a smaller size (1001) than *global_pool*, produced an approximately equal accuracy of 97.23%. The result from *Logits* is close to this. In fact, all of the results in DeepCWC for the three cases are stable and close to each other.

The last set of models are of the VGG implementation. We chose VGG-16 and VGG-19 models, and utilized the feature maps from their $fc6$, $fc7$ and $fc8$ layers. The size of both $fc6$ and $fc7$ is 4096, while the last $fc8$ layer has a smaller size of 1000. The largest feature set in this group of experiments produced the most promising classification results. What is more, the trend is consistent with before. As shown in Fig. 5 (e) and (f), the highest accuracy is up to 97.66%, using the shallower $fc6$ layer in VGG-16, which is also the most promising result we obtained using this dataset and in all cases. The results achieved by VGG-19 are slightly lower than this, but higher than the other cases. The larger feature size helps to produce a more accurate classification.

4.4. Comparison to the state-of-the-arts on Fashion-MNIST

The results obtained on the pre-trained models (over 97%) are all state-of-the-art, as shown in in Tab. 1. According to the description of the current methods, all of them are tuned and trained on the Fashion-MNIST dataset locally. Also, most of them applied one or two preprocessing techniques, and used the same deep neutral network architecture, e.g., VGG, ResNet, etc. Previously, the most promising accuracy was obtained by the Wide Residual Networks (WRN) model Zagoruyko and Komodakis (2016), both of which applied the standard preprocessing (mean/std subtraction/division) and augmentation (random

Table 2: Classification time and speed on different models.

Model	Layer	Feature Size	Time (sec)	Speed (sec/image)
VGG-16	/fc6	4096	16342.347	0.233
VGG-16	/fc8	1000	6826.632	0.098
VGG-19	/fc6	4096	16567.730	0.237
VGG-19	/fc8	1000	6526.632	0.093
Inception_v4	/Logits	1001	7177.612	0.103
Inception_v4	/global_pool	1536	8626.181	0.123
Inception_RN_v2	/Logits	1001	6734.609	0.096
Inception_RN_v2	/global_pool	1536	7884.362	0.113
ResNet_v1_101	/global_pool	2048	9108.256	0.207
ResNet_v1_101	/logits	1000	7302.812	0.104
ResNet_v2_101	/global_pool	2048	8373.421	0.120
ResNet_v2_101	/logits	1000	7254.657	0.104

crops/horizontal flips). For example, one used the random erasing technique Zhong et al. (2017) and produced an accuracy of 96.3% [5], while the other one with 96.7% accuracy had 8.9 M parameters and utilized freezing layers [6].

Compared to current state-of-the-art methods, the proposed DeepCWC produces a higher result using multiple CNN models. The classification accuracy ranges from 97.24% to 97.66%, which are all higher than previous methods. The highest accuracy is generated on VGG-16 from the $fc6$ layer with a size of 4096.

4.5. Discussion

Our experimental machine was configured with the following hardware, including an Intel® Core™ i7-7820X CPU@3.60GHz x 16, 64 GB RAM, 1.3 TB SSD and one NVIDIA TITAN Xp GPU. The code was run on TensorFlow 1.6, MATLAB R2016 and Ubuntu 16.04 OS. The recorded time consumption of each experimental case is shown in Tab. 2.

This time includes the whole training and testing of both the original images and deep features in CR, but does not count the time for feature extraction by the pre-trained models. Therefore, the speed (seconds per sample) is calculated by dividing the total time by the size of dataset (70000). Considering that the running state of the machine may fluctuate, the speed is between 0.1 - 0.2 seconds per sample. Furthermore, the following can be discussed about the proposed method.

Linear representation such as CR can improve deep neural networks based representation learning. Even using a pre-trained model, the proposed DeepCWC achieved a state-of-the-art classification result on the Fashion-MNIST dataset. The accuracy and performance outperformed current popular methods as well. This gives us a clue that linear methods have a new way to cooperate with nonlinear models.

The collaborative weight of two diverse representations help produce an accurate classifier. Currently, more work is focusing on the neural network architecture

5. https://github.com/zhunzhong07/Random-Erasing
6. https://github.com/ajbrock/FreezeOut

and/or parameter tuning. However, the proposed DeepCWC neither pays much attention to deep learning model itself, nor tunes any parameters. Fusing multiple representations creates a robust classifier that works well on multiple deep learning models. The results are all at a state-of-the-art level.

Multiple layers in a deep CNN model show an effective capacity to extract discriminative features, including the global average pooling layer Lin et al. (2013) and the fully connected layer. This is confirmed in the experiments, as shown in Fig. 4.3, and Tab. 1. However, the size of feature map decides the time consumption.

5. Conclusions

We propose a deep collaborative weight-based classification (DeepCWC) method. It first performs the linear representation on original images and deep features, extracted from nonlinear neural networks. Then, both of them collaboratively weight each other to build a strong discriminative classifier. The method is extensively evaluated using multiple popular deep CNN models, like ResNet, Inception, and VGG. The experimental results are promising on more than one layers in these neural networks, with most of the results belonging to a state-of-the-art level.

The l_2-norm based CR model is chosen as the linear constraint in this work to enhance the classification based on pre-trained CNN models. However, there are still some questions, for example, whether there are other linear models (like sparse representation, dictionary learning, etc.), more suitable for the same task, or whether it can bring one more step of break-through when applied on locally trained and tuned CNN models. We will keep working on these open topics in the future.

Acknowledgments

This research was funded by National Natural Science Foundation of China (NSFC) (61602540, 61502208), Science and Technology Development Fund (FDCT) of Macao SAR (124/2014/A3), and Scientific and Technical Program of City of Huizhou (2016X0422037, 2017C0405021). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Efficient classification with sparsity augmented collaborative representation. *Pattern Recognition*, 65:136–145, 2017.

Sijia Cai, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. A probabilistic collaborative representation based approach for pattern classification. In *Computer Vision and Pattern Recognition*, pages 2950–2959, 2016.

AT&T Laboratories Cambridge. The orl database of faces. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. Online; accessed 12-Octobor-2017.

Ling Chen, Hong Man, and Ara V Nefian. Face recognition based on multi-class mapping of fisher scores. *Pattern Recognition*, 38(6):799–811, 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

Weihong Deng, Jiani Hu, and Jun Guo. In defense of sparsity based face recognition. In *Computer vision and pattern recognition (cvpr), 2013 ieee conference on*, pages 399–406. IEEE, 2013.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. IEEE, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016b.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

Yu Li, Peter Richtarik, Lizhong Ding, and Xin Gao. On the decision boundary of deep neural networks. *arXiv preprint arXiv:1808.05385*, 2018.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Aleix M Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

S. Milborrow, J. Morkel, and F. Nicolls. The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*, 2010. http://www.milbo.org/muct.

S Nayar, S Nene, and Hiroshi Murase. Columbia object image library (coil 100). *Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96*, 1996.

Xi Peng, Lei Zhang, Zhang Yi, and Kok Kiong Tan. Learning locality-constrained collaborative representation for robust face recognition. *Pattern Recognition*, 47(9):2794–2806, 2014.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

Marko Ristin, Matthieu Guillaumin, Juergen Gall, and Luc Van Gool. Incremental learning of ncm forests for large-scale image classification. In *Computer Vision and Pattern Recognition*, pages 3654–3661, 2014.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

Ling Shao, Ziyun Cai, Li Liu, and Ke Lu. Performance evaluation of deep feature learning for rgb-d image/video classification. *Information Sciences*, 385:266–283, 2017.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inceptionresnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

Abhinav Valada, Luciano Spinello, and Wolfram Burgard. Deep feature learning for acoustics-based terrain classification. In *Robotics Research*, pages 21–37. Springer, 2018.

Jie Wen, Bob Zhang, Yong Xu, Jian Yang, and Na Han. Adaptive weighted nonnegative low-rank representation. *Pattern Recognition*, 81:326–340, 2018.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2074–2082, 2016a.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016b.

John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, Yuxin Peng, and Zheng Zhang. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *ACM International Conference on Multimedia*, pages 177–186, 2014.

Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

Bo Xin, Yizhou Wang, Wen Gao, David Wipf, and Baoyuan Wang. Maximal sparsity with deep networks? In *Advances in Neural Information Processing Systems*, pages 4340–4348, 2016.

Yong Xu, David Zhang, Jian Yang, and Jing-Yu Yang. A two-phase test sample sparse representation method for use with face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(9):1255–1262, 2011.

Yong Xu, Zhengming Li, Bob Zhang, Jian Yang, and Jane You. Sample diversity, representation effectiveness and robust dictionary learning for face recognition. *Information Sciences*, 375(C):171–182, 2017a.

Yong Xu, Zuofeng Zhong, Jian Yang, Jane You, and David Zhang. A new discriminative sparse representation method for robust face recognition via $l\_\{2\}$ regularization. *IEEE transactions on neural networks and learning systems*, 28(10):2233–2242, 2017b.

Hsiang Fu Yu, Cho Jui Hsieh, Kai Wei Chang, and Chih Jen Lin. Large linear classification when data cannot fit in memory. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):23, 2012.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Shaoning Zeng, Jianping Gou, and Xiong Yang. Improving sparsity of coefficients for robust sparse and collaborative representation-based image classification. *Neural Computing & Applications*, pages 1–14, 2017a.

Shaoning Zeng, Xiong Yang, and Jianping Gou. Multiplication fusion of sparse and collaborative representation for robust face recognition. *Multimedia Tools and Applications*, 76 (20):20889–20907, 2017b.

Shaoning Zeng, Bob Zhang, and Yong Du. Joint distances by sparse representation and locality-constrained dictionary learning for robust leaf recognition. *Computers and Electronics in Agriculture*, 142:563–571, 2017c.

Shaoning Zeng, Bob Zhang, Yuandong Lan, and Jianping Gou. Robust collaborative representation-based classification via regularization of truncated total least squares. *Neural Computing and Applications*, pages 1–9, 2018.

Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision*, pages 471–478, 2012.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.