Integrate Receptive Field Block into Large-margin Softmax Loss for Face Recognition

Yi Wei$^{1,2}$, Haibo Pu$^{1,2,*}$, Yu Zhu$^{1,2}$ and XiaoFan Li$^{1,2}$

$^1$College of Information Engineering, Sichuan Agricultural University
$^2$Key Laboratory of Agricultural Information Engineering of Sichuan Province, Sichuan Agricultural University, Ya’an, China

*Corresponding author: puhb@sicau.edu.cn

Abstract. Large-margin softmax (L-Softmax) loss for deep neural networks in face recognition tasks. Different from the softmax cross-entropy loss, L-Softmax explicit encourages intra-class compactness and inter-class separability between learned features. RFB Net proposed RFB module to simulate Receptive Fields (RFs) in human visual systems and gain higher accuracy. In this paper, we proposed Integrate Receptive Field Block into L-Softmax Loss for Face Recognition, an enhanced L-Softmax loss with a RFB module, which not only enhance the feature discriminability, but also enhance the feature robustness. Extensive experiments on recognition benchmarks like MNIST, LFW, Experiments with two benchmark datasets show that our proposed approach makes deep learning features more discriminating and thus significantly improves the performance of various visual classification and verification tasks.

1. Introduction

CNN has undoubtedly achieved great success, such as objects [1, 2, 3, 4, 5], scenes [6, 7], and actions [8, 9, 10]. The research status of classification problems has been greatly improved. To be so successful, it is inseparable from large-scale training data [11, 12] and an end-to-end learning framework. CNN is also applied into performing tag prediction and feature learning, then input data is reflected into the output of the last hidden layer (deep features), and then reflected into the predicted tags, such this process is shown as Figure 1. While learning is strong, CNNs are also facing key issues of over-instilling. Considerable effort such as large-scale training data [12], dropout[3], data augmentation [3,13], regularization [14] and stochastic pooling [15] has been put to address the issue. But these methods are not very effective.

The trend of learning more powerful recently is to enhance CNN through more discriminating information. Intuitively speaking, if their intra-class compactness is maximized, as well as inter-class separability, at the same time, the learning characteristics are good. In spite of its difficulties, because there are many large intra-class variations inherent in many tasks, CNN has powerful representation ability, which can make learning invariant features enabled in this direction. Because such thought can be encouraged, [16] large-Margin Softmax Loss (L-softmax) Function is proposed.

The learning objective of L-Softmax loss is very flexible, and the inter class angle boundary constraint can be adjusted. L-Softmax losses generally have three advent-ages. First, it emphasizes that the margin is determined by the angle between classes, which can create more distinguishing features.
Secondly, in order to avoid over-fitting, it will define more difficult learning objectives, so classification can propose different solutions to over-fitting. Thirdly, L-Softmax not only solves the classification problem, but also helps to verify the problem. The ideal learning feature is that the minimum distance between classes is better than the most distance within classes. Under the circumstances, learning well separated functions can obviously promote performance. [16]. However, the network used by L-Softmax cannot fully utilize the multi-layer feature map.

![Figure 1](image.png)

**Figure 1.** The typical framework of convolutional neural networks.

In this paper, in order to solve the problem that the multi-layer feature map cannot be fully utilized, we integrate RFB into L-Softmax to obtain more abundant feature representation and make L-Softmax learning characteristics more compact and better separated. Our idea is verified by Figure 2.

We verified our thoughts at MNIST dataset and LFW dataset. The results indicate that integrate RFB into L-Softmax can improve face recognition accuracy.

Our major contributions can be summarized as follows:

1. We introduce a easy way of Face Recognition by integrating RFB module into L-Softmax to obtain more abundant feature representation and make L-Softmax learning characteristics more compact and better separated.
2. We propose that integrating the RFB into the L-Softmax method is easy to implement in CNN.
3. We evaluate our way on MNIST dataset and LFW dataset with quantitative and qualitative experiments, and prove that our way can improve face recognition accuracy over the original L-Softmax function.

2. Related work

2.1. Deep face recognition

In recent years, face recognition technology is becoming more and more popular. [17,18] CNN supervised by softmax loss is used to solve open-set FR, which basically regards open-set FR as a multi-class classification problem. [19] combined with contrastive loss and softmax loss jointly monitor CNN training, which deeply improves the performance. [20] uses triplet loss of conjoint to learn unified facial embeddedness. Training for nearly 200 million facial images, they achieved the most advanced FR veracity. Linear discriminant analysis [21] proposed the center loss of CNN and obtained promising performance. Generally speaking, the FR CNN [22,23] currently performs well mainly on the basis of contrastive loss or triplet loss. It can be noted that the most advanced FR methods usually use ideas from metric learning (e.g. comparison loss, triplet loss), which shows that open FR can be well solved by discriminatory metric learning. L-Softmax Loss is promoted based on Softmax to satisfy distance constraints that encourage intra-class compactness, and selectivity between classes can improve feature discrimination deeply.
2.2. Spatial RFs With Deep ConvNet

In fact, many studies have analyzed CNN and its related RFs, including GoogleNet and its variants [24], [25], [26], Dilated Convolution [27] and Deformable CNN [28]. Recently, RFB Net [29] proposed the RFB module, which simulates the RF mechanism in the human visual system to improve the lightweight model based on feature representation. It’s a multi-branch convolutional block similar to the Inception [26] block. Together with RFB, all these methods capture information through multi-scale RFs implemented in different manners, and share the properties of end-to-end learning, convenient training, and easy integrating into any CNN architectures. Nevertheless, in contrast to the three counterparts, RFB is biological vision inspired in a daisy-shape configuration, which emphasizes the relation between RFs size and eccentricity.

3. The proposed approach

3.1. Revisiting the Large-Margin Softmax Loss

Following shows the L-Softmax loss function.

$$L_i = -log\left(\frac{e^{\|W_{yi}\|^2\|x_i\|^2}\varphi(\theta_{yi})}{e^{\|W_{yi}\|^2\|x_i\|^2}\sum_j e^{\|W_j\|^2\|x_i\|^2}\cos(\theta_j)}\right)$$  \hspace{1cm} (1)

In general, we need

$$\varphi(\theta) = \begin{cases} \cos(m\theta) & 0 \leq \theta \leq \frac{\pi}{m} \\ D(\theta) & 0 \leq \theta \leq \frac{\pi}{m} \end{cases}$$  \hspace{1cm} (2)

In Eq.1, $x_i \in R^d$ denotes the i-th deep feature, belonging to the $y_i$th class. $W_j \in R^d$ denotes the jth column of the weights $W \in R^d$ in the last fully connected layer. $\theta_j (0 \leq \theta_j \leq \pi)$ is the angle between the vector $W_j$ and $x_i$. where $m$ is an integer that is closely related to the classification margin. With larger $m (m \geq 4)$, the classification margin becomes larger and the learning objective also becomes harder. In this paper, $M$ takes 4, which in [16] has proved that when $m$ is 4, the classification is best. Meanwhile, $D(\theta)$ is required to be a monotonically decreasing function and $D(\frac{\pi}{m})$ should equal $\cos(\frac{\pi}{m})$.

In Figure 2, the 2-D deepening features from results are marked to show the distribution. From Figure 2 we can find that under the supervision of L-softmax loss, although the characteristics of deep learning are separable, the deepening characters are not sufficiently discriminating because they present important intra-class changes. Therefore, it is improper to apply these features directly for identification.
3.2. Receptive Field Block

How can we improve the resolving ability of deep learning performance through effective loss functions? To be precise, the key is to minimize intra-class variations while maintaining the characteristics of separating different classes. Therefore, we propose to integrate the perceptual field block into L-Softmax Loss.

The RFB belongs to a multi-branch convolution block. There are two parts consisted its internal parts: a multi-branch volume layer with various kernels and a trailing expansion pool or convolution layer. The previous part operates to simulate multiple sizes of RF, which is the same as the beginning, and the latter part shows the relationship between RF size and eccentricity in the human visual system. [29]. The structure of RFB module we used is given in Figure 3, and more detail of the two parts follows.

**Multi-branch convolution layer:** Just as CNN defines RF, the way to implement multi-size RF through different cores is very simple and natural. Compared with fixed-size RF, it has advantages. In general, first of all, in each branch, we make use of the bottleneck structure, a 1×1 conversion layer is included, and an n×n conversion layer is added to reduce the number of channels in the feature map. Secondly, in order to reduce the parameters and the deeper nonlinear layer, the 5×5 conversion layer is replaced by two stacked 3×3 conversion layers, and the original n×n conversion layer is replaced by the 1×n plus N×1 conversion layer. Finally, the shortcut design for ResNet [1] and Inception-ResNet V2 [27] will be put into use.

**Dilated pooling or convolution layer:** This concept was originally proposed by Deeplab[30]. this concept was also commonly referred to as the astrous convolutional layer. Creating a higher resolution feature map is the basic intent of the structure, capturing information in larger areas while maintaining the same number of parameters. This design was quickly demonstrated to be semantically separable [27] and subsequently applied to many object detectors such as SSD [31] and R-FCN [32] can greatly improve s improve rate and/or Precision [29].

Distensible convolution is used to irritate the effect of the ellipse of RFs in the human visual cortex in this paper. Figure 3 shows the combination of multi branch volume layer with extensible pool or volume layer. As for every branch, a core size is abided by a connection or convolution layer with a related expansion. Both kernel size and expansion have an analogous active function [29] with the size and centrifugal rate of RF in the visual cortex. Finally, all the branches with feature maps are connected into a convolution array or space pool.

Because of RFB property, integrating into CNNs is very easy and we can need it to improve existing architectures. Our main modification is to integrate the RFB module into the network structure of the original L-softmax loss function.


Table 1. Our CNN architectures for different benchmark datasets. Conv1.x, Conv2.x and Conv3.x denote convolution units that may contain multiple convolution layers. E.g., [3×3, 64]×4 denotes 4 cascaded convolution layers with 64 filters of size 3×3

| Layer       | MNIST                  | LFW                  |
|-------------|------------------------|----------------------|
| Conv0.x     | [3×3,64]×1             | [3×3,64]×1,Stride 2  |
| Conv1.x     | [3×3,64]×3             | [3×3,64]×4           |
| Pool1       | 2×2Max, Stride 2       | 2×2Max, Stride 2     |
| Conv2.x     | [3×3,64]×3             | [3×3,256]×4         |
| Pool2       | 2×2Max, Stride 2       | 2×2Max, Stride 2     |
| Conv3.x     | [3×3,64]×3             | [3×3,256]×4         |
| RFB Block   | RFB                    | RFB                 |
| Conv4.x     | N/A                    | [3×3,256]×4         |
| Fully Connected | 256          | 512                 |

4. Experiments

4.1. Experimental Settings
In this section, the master conducted two sets of experiments to verify the method proposed in this paper. In the first group, we perform image classification; in the second group, we perform face verification. In the visual classification, we use the standard benchmarking dataset: MNIST [33]; in face verification, we evaluate our approach on the widely used LFW dataset [34]. We only use a single model in all baseline CNNs to compare our performance. For convenience, we use L-Softmax to denote the L-Softmax loss. L-Softmax, RFB and L-Softmax in the experiments use the same CNN shown in Table 1. All experiments are carried out using the pytorch framework on NVIDIA GTX2080 GPUs.

4.2. Image Classification
Table 2 shows the previous best results and those for our proposed RFB and L-Softmax loss. From the results, the RFB and L-Softmax losses are not only superior to the original L-softmax loss using the same network, but also achieve the most advanced performance compared to other deep CNN architectures. In Figure 2, we also visualize the learned features by the RFB and L-Softmax loss and compare them to the original L-softmax.

4.3. Face Verification
The proposed method was tested in the field marker surface (LFW) dataset [34]. The LFW dataset has 13,233 facial images. The 13233 facial image is mainly composed of 5749 facial images of different identities. , expressions and lighting have changed a lot. The official 6,000 pairs are used for face verification testing. We followed the standard unlimited external data protocol labeled LFW, and we only trained on the CASIA-WebFace dataset [35]. CASIA-WebFace dataset contains 494 414 facial images, which are composed of 10,575 human facial images. Before training, we used MTCNN [36] to adjust them to 128 x128 pixels.

We trained on the network structure of Table 1. We trained a batch size of 256 and trained a total of 20 batches. The learning rate was initially set to 0.1 and multiplied by 0.1 in the 10th, 14th and 16th epochs. The networks are trained using stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 5×10−4. In Table 3, the accuracies for the Deepid2+ (contrastive loss) [18] and the FaceNet (triplet loss) [20] are reported in the original papers. The FaceNet achieves the highest accuracy of 99.65% by using a very large training set of 200M images. In [16], a higher accuracy of 98.71% for
the L-softmax loss by using both the CASIA-Webface for training, with 0.7M training images in total. When using the CASIA-Webface training dataset only, the RFB and L-softmax loss outperforms the other loss functions.

**Table 2.** Recognition error rate (%) on MNIST dataset

| Method                        | Error Rate |
|-------------------------------|------------|
| Hinge Loss                    | 0.47       |
| Softmax                       | 0.40       |
| L-Softmax (m=2)               | 0.32       |
| L-Softmax (m=3)               | 0.31       |
| L-Softmax (m=4)               | 0.31       |
| RFB + L-Softmax (m=4)         | 0.22       |

**Table 3.** Verification performance (%) on LFW dataset. * denotes the outside data is private (not publicly available).

| Method                        | Outside Data | Accuracy |
|-------------------------------|--------------|----------|
| FaceNet                       | 200M*        | 99.65    |
| Deepid2+                      | 300K*        | 98.70    |
| Softmax                       | WebFace      | 96.53    |
| Softmax + Contrastive         | WebFace      | 97.31    |
| L-Softmax (m=2)               | WebFace      | 97.81    |
| L-Softmax (m=3)               | WebFace      | 98.27    |
| L-Softmax (m=4)               | WebFace      | 98.71    |
| RFB + L-Softmax (m=4)         | WebFace      | 98.82    |

5. Conclusions

In this paper, we propose a way to Integrate Receptive Field Block into Large-margin Softmax (L-Softmax) Loss. By combining the RFB with the L-softmax loss to jointly supervise the learning of the CNN, the discernment of the features of deep learning can be highly enhanced for robust facial recognition. Extensive experiments on several large-scale facial benchmarks have convincingly demonstrated the effectiveness of the proposed method.

Acknowledgement

This paper was supported by Key Laboratory of Agricultural Information Engineering of Sichuan Province Sichuan Agricultural University and funded by the key project of the Sichuan Provincial Department of Education (NO.17ZA0297).

References

[1] He K, Zhang X, Ren S, et al. *Deep Residual Learning for Image Recognition* [J]. 2015:770-778.
[2] He, K, Zhang, X, Ren, S, Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
[3] Krizhevsky A, Sutskever I, Hinton G E. *ImageNet classification with deep convolutional neural networks*[C]/ International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
[4] Simonyan, K, Zisserman, A.: *Very deep convolutional networks for large-scale image recognition*. arXiv preprint (2014).
[5] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. 2014:1-9. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9(2015)

[6] Zhou, Bolei, Khosla, Aditya, Lapedriza, Agata, et al. Object Detectors Emerge in Deep Scene CNNs[J]. Computer Science, 2014.

[7] Zhou B, Lapedriza A, Xiao J, et al. Learning deep features

[8] Baccouche M, Mamalet F, Wolf C, et al. Sequential Deep Learning for Human Action Recognition[C]// International Conference on Human Behavior Understanding. Springer-Verlag, 2011:29-39.

[9] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE

[10] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]// Computer Vision and Pattern Recognition. IEEE, 2015:4305-4314.

[11] Fukunage K, Narendra P M. A Branch and Bound Algorithm for Computing k-Nearest Neighbors[J]. IEEE Trans Comput, 1975, 24(7):750-753.

[12] Olga Russakovsky, Jia Deng, Hao Su, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[13] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. 2014:1-9.

[14] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust[J]. 2014:2892-2900.

[15] Chen Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]// International Conference on Neural Information Processing Systems. MIT Press, 2014:1988-1996.

[16] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering[J]. 2015:815-823.

[17] Sun Y, Wang X, Tang X. Sparsifying Neural Network Connections for Face Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:4856-4864.

[18] Liu J, Deng Y, Bai T, et al. Targeting Ultimate Accuracy: Face Recognition via Deep Embedding[J]. 2015.

[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. 2014:1-9.

[20] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]// Computer Vision and Pattern Recognition. IEEE, 2016:2818-2826.

[21] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.

[22] Chen L C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. 2017.

[23] Dai J, Qi H, Xiong Y, et al. Deformable Convolutional Networks[J]. 2017:764-773.
[29] S. Liu, D. Huang, and Y. Wang, “Receptive field block net for accurate and fast object detection,” 2017.
[30] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, PP(99):834-848.
[31] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Springer International Publishing, 2016:21-37.
[32] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. 2016.
[33] LeCun, Yann, Cortes, Corinna, and Burges, Christopher JC. The mnist database of handwritten digits, 1998.
[34] Huang, Gary B, Ramesh, Manu, Berg, Tamara, and Learned-Miller, Erik. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report, 2007.
[35] Yi D, Lei Z, Liao S, et al. Learning Face Representation from Scratch[J]. Computer Science, 2014.
[36] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.