# Max-Feature-Map Based Light Convolutional Embedding Networks for Face Verification

Zhou Yang[1], Meng Jian[1], Bingkun Bao[2,3(✉)], and Lifang Wu[1]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing, China
yangzhoufrank@yeah.net, jianmeng648@163.com, lfwu@bjut.edu.cn

[2] Nanjing Jingjunhai Network Ltd., Nanjing, Jiangsu, China
bingkunbao@gmail.com

[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

**Abstract.** The powerful image feature extraction ability of convolutional neural network makes it possible to achieve great success in the field of face recognition. However, this category of models tend to be deep and paralleled which is not capable to be applied in real-time face recognition tasks. In order to improve its feasibility, we propose a max-feature-map activation based fully convolutional structure to extract face features with higher speed and less computational cost. The learned model has a great potential on embedding in the hardware devices due to its high recognition performance and small storage space. Experimental results demonstrate that the proposed model is ***63*** times smaller in comparison with the famous VGG model. At the same time, ***96.80%*** verification accuracy is achieved for a single network on LFW benchmark.

**Keywords:** Face verification · Convolutional neural network · Feature extraction

## 1 Introduction

With the development of computer vision techniques and advanced hardware support, convolutional neural network (CNN) gradually becomes one of the most widely used and successful face recognition and verification algorithms in recent years. CNN construction is initially employed to solve the multiclass classification problem [1] and achieves great success. With further investigation on this structure, other functions are also developed, for instance, object detection [2], natural language processing [3, 4], face recognition [5–7] and so on.

For the face verification task, since the working principle of convnet is more similar to the characteristics of biological neural network, numerous breakthroughs are made on various challenging recognition tasks. On LFW [19] benchmark, the results of CNN based algorithms [7] outperform human performance [8, 9] on the face verification tasks. CNN is capable of learning robust feature of human faces even the images are filled with noise or disturbance like pose variation, background chaos, illumination and so on.


J. Zhou et al. (Eds.): CCBR 2017, LNCS 10568, pp. 58–65, 2017.
https://doi.org/10.1007/978-3-319-69923-3_7

The deep model is able to capture the tiny but discriminative characteristic of human face which is tough for human eyes to recognize. However, many existing CNN based recognition algorithms depend on complex data pre-processing method [5] or deep network [10] with numerous connection parameters and parallel computing. Thus, for the real-time supervision system, current CNN based algorithms require to transmit captured data to the server which cannot meet the need of real-time recognition tasks. Also, some methods [6, 11] simultaneously utilize both verification and identification supervisory signals to train the deep model. However, the trade-off between identification and verification loss is hard to set.

Light CNN [12–14] utilized Max-Feature-Map (MFM) activation to get a compact and informative feature expression from competitive relationship between two feature maps. This ingenious design reduces the parameters to some degree in the network. The idea of MFM operation solves the problem of key message lost during the first layers caused by ReLU activation. However, the information in the fully connected layer still exists data redundancy which may reduce the availability of real-time deployment of the algorithm. Fully Convolutional Networks (FCN) [15] discards non-convolution portion in CNN and trains a network with only stacked convolutional layers. As an extension of traditional CNN structure, the idea of fully convolutional arrangement also has a great potential in classification field. However, to our knowledge, few of people and organizations try to transplant this idea in face verification and identification tasks.

In this work, we deeply analyze the data distribution and propagation in the network and then combine MFM operation with fully convolutional layer structure and then build a deep learning framework, denoted as Improved Light CNN. The contributions of the proposed Improved Light CNN for face verification task are summarized as follows.

- The MFM operation, which plays the role of activation in the proposed network, produces compact representation and sparse connection. The proposed model employs convolutional structure instead of fully-connected structure to alleviate data redundancy in feature representation. Both MFM and convolutional embedding make the model occupy relatively small disk space with low computational burden.
- Due to the powerful feature selection ability of our network, a low dimension vector (256-d) is capable enough to represent discriminative face characteristics. Such low dimensional data advances the efficiency of information processing and analysis in the later steps. Hence the superiority of less computational burden gives our model a great potential and probability for embedding in real-time applications.

The paper is organized as follows. In Sect. 2, we describe the architecture of Improved Light CNN network and detailed information of each layer. The performance on LFW benchmark and the computational efficiency of the proposed model are presented in Sect. 3. Finally, we briefly conclude this work in Sect. 4.

## 2 Network Architecture

In this section, we firstly introduce the overall architecture of Improved Light CNN framework and then discuss the principle and feasibility of feature extraction from a convolutional layer.

### 2.1 The Improved Light CNN Framework

The configuration and detailed parameters of Improved Light CNN network are outlined in Table 1. The network consists of 6 convolutional layers, 4 max-pooling layers and MFM activations. Also, we utilize 1 × 1 kernel size convolutional layers as network in network architecture [10] which can introduce nonlinear to the network and improve feature extraction ability of the model. MFM function, adopted as activation in our network, is an efficient and convenient feature selection tool. The working principle of MFM operation can be shown via Fig. 1. Since MFM operation brings the competition relationship into network, informative nodes are able to be kept during forward

**Table 1.** Architecture of the proposed Improved Light CNN. In the left column of the table, the layer names refer to the layers in the network built in Caffe framework. Conv refers to convolutional layer, MFM denotes as MFM activation and Pool refers to pooling layer.

| Layer name | Filter size | Output size | Parameters |
|---|---|---|---|
| Conv1 | 6 × 6/2, 2 | 64 × 64 × 96 | 3.4K |
| MFM1 | - | 64 × 64 × 48 | - |
| Pool1 | 2 × 2/2 | 32 × 32 × 48 | - |
| Conv1a | 1 × 1/1 | 32 × 32 × 96 | 4.6K |
| MFM1a | - | 32 × 32 × 48 | - |
| Conv2 | 3 × 3/1, 1 | 32 × 32 × 192 | 82K |
| MFM2 | - | 32 × 32 × 96 | - |
| Pool2 | 2 × 2/2 | 16 × 16 × 96 | - |
| Conv2a | 1 × 1/1 | 16 × 16 × 192 | 18K |
| MFM2a | - | 16 × 16 × 96 | - |
| Conv3 | 3 × 3/1, 1 | 16 × 16 × 384 | 331K |
| MFM3 | - | 16 × 16 × 192 | - |
| Pool3 | 2 × 2/2 | 8 × 8 × 192 | - |
| Conv3a | 1 × 1/1 | 8 × 8 × 384 | 73K |
| MFM3a | - | 8 × 8 × 192 | - |
| Conv4 | 3 × 3/1, 1 | 8 × 8 × 256 | 442K |
| MFM4 | - | 8 × 8 × 128 | - |
| Conv4a | 1 × 1/1 | 8 × 8 × 256 | 32K |
| MFM4a | - | 8 × 8 × 128 | - |
| Conv5 | 3 × 3/1 | 8 × 8 × 256 | 131K |
| MFM5 | - | 8 × 8 × 128 | - |
| Pool4 | 2 × 2/2 | 4 × 4 × 128 | - |
| Conv-fc | 4 × 4/1 | 512 | 1048K |
| MFM-Conv | - | 256 | - |
| Total | - | - | 2165K |

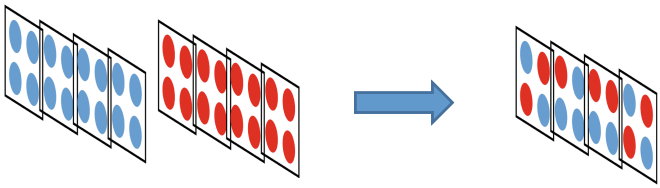

**Fig. 1.** Illustration of MFM operation. The result values from MFM operation are the max ones between the neural nodes in the corresponding positions. The calculated feature map after MFM operation contains the activated neuron from previous feature maps.

propagation. At the same time, convolutional layer can do the similar job as the fully connected layer as long as expanding the kernel size to cover the whole feature map.

Given a convolutional layer which outputs 2 N feature maps. Denoting by W and H as the width and the height of the feature and by $z_{ij}^{l+N}$ as the neural nodes in the feature maps, where N = {1, …, 2 N}, $1 \le l \le N,\ 1 \le i \le H,\ 1 \le j \le W$. The MFM operation can be represented as

$$\hat{z}_{ij}^{l} = max\left(z_{ij}^{l}, z_{ij}^{l+N}\right) \tag{1}$$

## 2.2 Face Feature Extraction from Convolutional Layer

Traditional CNN based face verification models often extract face feature vectors from the fully-connected layer. However, in this case, data redundancy may be caused to some extent because the majority of learned parameters are in the fully-connected layer.

When the verification algorithm is embedded on mobile devices, too much computational pressure may lead to unexpected error which may degrade the user experience.

In order to alleviate the computational burden and enhance the practicability in real-time applications, we propose a new method of applying MFM operation as network activation and replace the fully-connected layer in traditional CNN by a convolutional layer. The core of convolutional layer is local connection while the fully-connected layer is global connection. When we expand the kernel size in a convolutional layer until covering the whole input image, it can be approximately regarded as fully-connected construction. Both of the above two structures utilize dot product to compute neuron output. A 512-dimensional column feature vector will be generated after each global kernel convolution as shown in Fig. 2. The feature representations will be further compressed and simplified by MFM activation to a 256-dimensional face representation.

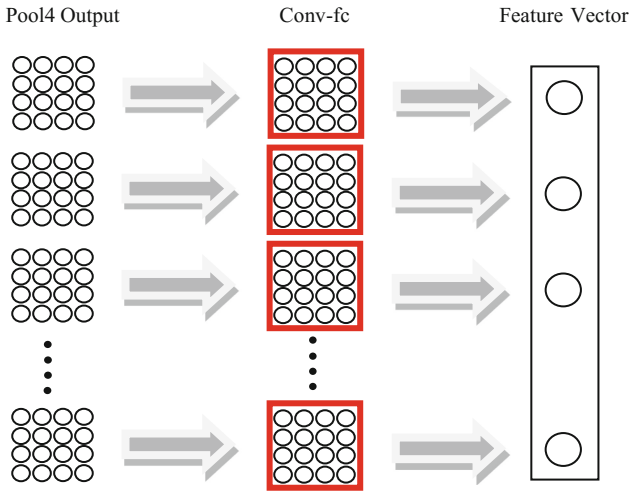


**Fig. 2.** Conv-fc layer process the feature map generated from Pool4 layer by convolutional layer with global information kernel. The output is 512-D column feature vectors.

## 3 Experiments

In this section, we firstly introduce the CASIA-Webface database [11] and the training methodology of Improved Light CNN network. Then we briefly present the face image pre-processing method. Finally, we evaluate our model on LFW standard testing benchmark and compare the results as well as compute efficiency with other deep learning model.

### 3.1 Face Image Pre-processing

CASIA-Webface, one of the biggest public face image databases, is chosen to train our Improved Light CNN network. The database contains 493,456 images from 10,575 various identities. The images are all randomly collected from Internet which can guarantee the generalization ability of learned model. The size of the original images in the database are all $250 \times 250$, which contains some noises including complex backgrounds and face pose variations. To ensure efficiently feature learning of the model, we pre-process the original image to eliminate the interference caused by pose variation to a great extent. We follow the pre-process idea in [14]. After image alignment, the distance between eyes center and mouth center and the distance of eyes center to the top of the image are both fixed into 48 pixels. This operation makes it possible to effectively solve the problem of angled face recognition tasks.

### 3.2 Training Methodology

We utilize the open source deep learning framework Caffe [16] to train the Improved Light CNN model. For the partition of training set and validation set, we randomly

select one image from every identity in the database as the validation part and treat the remaining as training resources. Moreover, the TITAN X GPU is equipped to train the model.

In data layer, we use crop operation to resize the input aligned image from 144 × 144 to 128 × 128. Also, we apply the mirror function to expand the training data which can alleviate the phenomenon of network overfitting. For the specific settings of hyper parameters, we set the initial learning rate to 1e-3 and gradually reduce to 1e-6 by step policy. The dropout ratio after MFM-Conv layer is set to 0.6. The momentum parameter is set to 0.9 and we use Xavier initialization policy in convolutional layer. Besides, because the network may face the problem of loss convergence the gradient explosion, we firstly set the learning rate in all the convolutional layer except the last one to 0 and the back propagation will only make effects on the last convolutional layer to train the classifier. When observing the network loss converges steadily, we recover the learning rate policy in each convolutional layer.

### 3.3 Results on the LFW Benchmark

We followed the standard unsupervised testing protocol on challenging LFW benchmark. There is no identities overlap between LFW dataset and CASIA-Webface dataset, which means all the images in the LFW is unacquainted to our model. Totally 6,000 face pairs which contains 3,000 positive pairs and 3,000 negative pairs are officially given for verification test. The testing images follows the same pre-processing method and are cropped into 128 × 128. Then we straightly extract the 256-D feature vectors from two images in one pair and compute their cosine similarity. The equal error rate (EER) obtained by ROC curve is applied to represent the performance of our model.

As shown in Table 2, our model achieves 96.80% accuracy on LFW database under unsupervised protocol. We do not try to build and test multi network models, because our model aims at real-time applications. The proposed Improved Light CNN drops slightly on accuracy compared with the original version of Light CNN which is mainly because we try to optimize the verification speed and storage place of the model thus it is acceptable to sacrifice part of the verification accuracy for further practicability of the model.

**Table 2.** Single network performance comparison with other method on LFW under unsupervised verification protocol.

| Method | Networks | Accuracy | Protocol |
|---|---|---|---|
| LBPNet [17] | 1 | 94.04% | Unsupervised |
| DeepFace [5] | 1 | 95.92% | Unsupervised |
| DeepID2 [6] | 1 | 95.43% | Unsupervised |
| Webface [11] | 1 | 96.13% | Unsupervised |
| Webface + PCA [11] | 1 | 96.30% | Unsupervised |
| Light CNN A [14] | 1 | 97.77% | Unsupervised |
| Light CNN B [14] | 1 | 98.13% | Unsupervised |
| **Improved Light CNN** | **1** | **96.80%** | **Unsupervised** |

### 3.4 Performance on Computational Cost and Storage Space

The success of a feature extractor model cannot be reflected purely on its recognition performance. The computational efficiency and running burden ought to be considered at the same time. We make a specific comparison on model working efficiency between our model and other released deep learning frameworks including DeepFace, VGG, CenterLoss, Light CNN A and Light CNN B.

To prove the ability of fast face verification, we compare our model with DeepFace, VGG, Light CNN A and Light CNN B on computing efficiency. From Table 3, it is obvious that our model has a superiority on data processing speed and model storage cost. The results also demonstrate that the substitutability of fully connected layer by global kernel convolutional layer in classification tasks and feature extractions. Due to the small size, fast verification ability and less learning parameter, our model is able to make contribution to real-time identification and verification tasks.

**Table 3.** The model working cost compared with other method on parameters, feature dimension, operation time per image and storage space. The data is tested on a single Intel Xeon e3 CPU.

| Model | Parameters | Feature dimension | Times (per image) | Storage space |
|---|---|---|---|---|
| DeepFace | 120,000K | 4096 | 198 ms | - |
| VGG [10] | 27,749K | 4096 | 455 ms | 553 MB |
| CenterLoss [18] | 19,596K | 1024 | 140 ms | - |
| Light CNN A | 3,961K | 256 | 68 ms | 26 MB |
| Light CNN B | 5,556K | 256 | 65 ms | 32.8 MB |
| **Improved Light CNN** | 2,165K | 256 | 29 ms | 8.7 MB |

## 4 Conclusion

In this paper, we proposed a deep learning framework Improved Light CNN for efficiently face verification tasks. The network applies MFM activation to screen informative neural node and provide a compact data distribution for forward propagation. Fully convolutional structure is utilized to further reduce the data redundancy. A low dimensional feature representation is extracted from the model and achieves 96.60% on LFW verification task. Both small occupied space and optimized parameters make it more promising for embedding on smart devices and real-time supervision systems.

**Acknowledgment.** This work was supported in part by the National Natural Science Foundation of China under Grant 61572503, Beijing Natural Science Foundation Grant 4152053.

## References

1. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. CoRR, pp. 1026–1034 (2015)

2. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. CoRR, pp. 779–788 (2016)
3. Nguyen, T.H., Grishman, R.: Relation extraction: perspective from convolutional neural networks. In: The Workshop on Vector Space Modeling for Natural Language Processing, pp. 39–48 (2015)
4. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Comput. Sci. (2015)
5. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: IEEE Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
6. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. Adv. Neural. Inf. Process. Syst. **27**, 1988–1996 (2015). MIT Press
7. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
8. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE International Conference on Computer Vision, pp. 365–372 (2010)
9. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1962–1977 (2011)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014)
11. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. Comput. Sci. (2014)
12. Wu, X.: Learning robust deep face representation. Comput. Sci. (2015)
13. Wu, X., He, R., Sun, Z.: A lightened CNN for deep face representation. Comput. Sci. (2015)
14. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. Comput. Sci. (2016)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
16. Jia, Y., Shelhamer, E., Donnahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia, pp. 675–678 (2014)
17. Xi, M., Chen, L., Polajnar, D., Tong, W.: Local binary pattern network: a deep learning approach for face recognition. In: IEEE International Conference on Image Processing (2016)
18. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). doi:10.1007/978-3-319-46478-7_31
19. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Month (2008)