IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 72-76

Light CNN For Deep Face Representation Using Multilevel Residual Networks

Kiran Singla[1], Prakash Mohod[2]
[1]*4th sem CSE, MTech., Department of Computer Science and Engineering, Rashtrasant tukadoji Maharaj Nagpur University, Nagpur, Maharashtra, India*
[2]*Assistance Professor, Department of Computer Science and Engineering, Rashtrasant tukadoji Maharaj Nagpur University, Nagpur, Maharashtra, India*

Abstract: *Deep convolutional neural networks have shown remarkable performance in the field of face recognition tasks in recent years. The networks are continuously growing larger to better fit large amount training data thus increasing the complexity and vanishing gradient problem. It normally takes a lot of time and more computational power to train these deeper neural networks. This paper presents a Light CNN with multilevel residual network(L-MRN) to learn a compact embedding on the large scale face dataset without fine tuning. This network is designed to obtain better performance from Light CNN without increasing the number of parameters and computational costs.*
Keywords: *Convolutional Neural Network, Face Recognition, Residual network, Light CNN, RoR*

I. Introduction

The emergence of deep convolutional neural networks has greatly contributed to advancements in solving complex tasks [6, 4, 5, 7, 10, 14, 15] in computer vision with significantly improved performance. Various tasks like image classification, face recognition, object detection have benefited from CNNs. To achieve optimal accuracy, the scale of the training dataset for CNN has been consistently increasing. To fit this large amount of training data the CNNs have been growing larger, thus increasing the computational costs of the network. Increasing the network depth is known to improve the model capabilities, which can be seen from AlexNet [6] with 8 layers, VGG [15] with 19 layers, and GoogleNet [14] with 22 layers. However, increasing the depth can be challenging for the learning process because of the vanishing/exploding gradient problem [2], which hamper convergence from the beginning. When deeper networks start converging, a degradation problem has been exposed, with the network depth increasing, accuracy gets saturated and then degrades rapidly. Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error.

Deep residual networks [22] avoid this problem by using identity skip-connections, which help the gradient to flow back into many layers without vanishing. The identity skip-connections facilitate training of very deep networks up to thousands of layers that helped residual networks win five major image recognitions tasks in ILSVRC 2015 [24] and Microsoft COCO 2015 [16] competitions. After the success of residual networks more and more variants of residual networks and architectures have been proposed [8, 9, 11, 12, 13, 18, 19, 20, 21].

However, an obvious drawback of residual networks is that every percentage of improvement requires significantly increasing the number of layers, which linearly increases the computational and memory costs [22].

Very deep residual models will also suffer vanishing gradients and overfitting problems; Thus, the performance of thousands layer ResNets is worse than hundred-layer ResNets. Then the Identity Mapping ResNets (Pre-ResNets) [12] simplified the residual networks training by BN-ReLU-conv order. Pre- ResNets can alleviate the vanishing gradients problem, so that the performance of thousand-layer Pre-ResNets can be further improved.

More and more residual network variants and architectures have been proposed, and they form a residual networks family together. One of them is RoR [8], it adds level-wise shortcut connections upon original residual networks to promote the learning capability of residual networks. To dig the optimization ability of residual networks, RoR substitutes residual mapping of residual mapping for optimizing original residual mapping. RoR achieved better performance than ResNets by using the same number of layers on different data sets. RoR is not only suitable for original ResNets, but also fits in nicely with other residual networks. Any residual network can be improved by RoR. Hence, RoR has a good prospect of successful application on various image recognition tasks.

Deeper and deeper CNNs have been proposed to achieve better performance and the results of these models revealed the importance of network depth, but this increased the difficulty of training such deeper networks as the number of parameters increased and thus increasing the computational cost. To solve this issue LightCNN [1] was proposed, it introduced a Max-Feature-Map (MFM) operation to obtain a compact representation and perform feature filter selection. Light CNN networks are carefully designed to obtain better performance meanwhile reducing the number of parameters and computational costs. MFM operation is a special case of maxout to learn a Light CNN with a small number of parameters. Compared to ReLU whose threshold is learned from training data, MFM adopts a competitive relationship so that it has better generalization ability and is applicable on different data distributions. In the context of CNN, MFM operation plays a similar role to local feature selection in biometrics. MFM selects the optimal feature at each location learned by different filters. It results in binary gradient (1 and 0) to excite or suppress one neuron during back propagation. This model leads to better performance in terms of speed and storage space.

In this paper we propose a network combining the features of Light CNN and RoR, we add level-wise shortcut connection to the Light CNN architecture. Although this approach seems quite simple, it is surprisingly effective in practice and achieves good performance. We hypothesize that the residual mapping of residual mapping will optimize the ability of LightCNN by adding a few identity shortcuts without increasing the layers of LightCNN.

The remainder of the paper is organized as follows. Section II illustrates the proposed L-MRN framework. Experimental results and analysis are discussed in Section III, leading to conclusion in Section IV.

ARCHITECTURE:

In this section, we discuss the architecture of our proposed L-MRN network. Our basic network architecture is same as proposed in LightCNN [1]. The LightCNN is based on MFM operation, which is a special case of maxout to learn a network with small number of parameters. MFM adopts a competitive relationship by choosing more competitive nodes from previous convolution layers by activating the maximum of two feature maps. MFM perform feature selection and facilitate to generate sparse connections.

Residual blocks are added to the 29-layer Light CNN network. The residual blocks contain two 3 x 3 convolutional layers and two MFM operations without batch normalization.Although batch normalization is efficient to accelerate the convergence of training and avoid overfitting, in practice, batch normalization is domain specific which may be failed when test samples come from different domains compared with training data. Besides, batch statistics may diminish when the size of training minibatches are small. Fully connected layer is employed instead of the global average pooling layer on the top, because while training input images are all aligned, so that each node for high-level feature maps contains both semantic and spatial information which may be damaged by the global average pooling.

The basic idea of Resnet block is that residual mapping is easy to optimize. It skips blocks of convolutional layers by using shortcut connections to form shortcut blocks. RoR [8] adds more level-wise shortcut connection upon these residual blocks. We add this level-wise shortcut connection to LightCNN framework. Table I Shows the architecture of Light CNN framework and Fig.1. Shows how we add the level-wise shortcut connections to it.

We used shortcut level number of m = 3 as it gave the best results as compared to others. Level 3 resblock is the basic resblock as given in LightCNN architecture. We don't add shortcuts by dividing each resblock group equally like RoR as LightCNN architecture is different from basic network architecture, we add Level-2 shortcuts after every max-pooling layer. Finally, the root or Level-1 shortcut connection is added after Initial Block as shown in Fig.1. Projection shortcuts (done by 1 x 1 convolutions) are used for increasing dimensions, while other shortcuts are identity connections.

II. Methodology

Table 1: The Architectures Of The Light Cnn-29 Model.

Block Names	Type	Filter Size/Stride, Pad	Output Size
Initial Block	Conv	5 x 5/1,2	128 x 128 x 96
	MFM	-	128 x 128 x 48
POOL1	Pool	2 x 2/2	64 x 64 x 48
RESBLOCK1	Conv	$\begin{bmatrix} 3\ X\ 3/1,1 \\ 3\ X\ 3/1,1 \end{bmatrix}$ x 1	64 x 64 x 48
GROUP1	Conv	1 x 1/1	64 x 64 x 96
	MFM	-	64 x 64 x 48
	Conv	3 x 3/1, 1	64 x 64 x 192
	MFM	-	64 x 64 x 96
POOL2	Pool	2 x 2/2	32 x 32 x 96

RESBLOCK2	Conv	$\begin{bmatrix} 3\ X\ 3/1,1 \\ 3\ X\ 3/1,1 \end{bmatrix}$ x 2	32 x 32 x 96
GROUP2	Conv	1 x 1/1	32 x 32 x 192
	MFM	-	32 x 32 x 96
	Conv	3 x 3/1, 1	32 x 32 x 384
	MFM	-	32 x 32 x 192
POOL3	Pool	2 x 2/2	16 x 16 x 192
RESBLOCK3	Conv	$\begin{bmatrix} 3\ X\ 3/1,1 \\ 3\ X\ 3/1,1 \end{bmatrix}$ x 3	16 x 16 x 192
GROUP3	Conv	1 x 1/1	16 x 16 x 384
	MFM	-	16 x 16 x 192
	Conv	3 x 3/1, 1	16 x 16 x 256
	MFM	-	16 x 16 x 128
RESBLOCK4	Conv	$\begin{bmatrix} 3\ X\ 3/1,1 \\ 3\ X\ 3/1,1 \end{bmatrix}$ x 4	16 x 16 x 128
GROUP4	Conv	1 x 1/1	16 x 16 x 256
	MFM	-	16 x 16 x 128
	Conv	3 x 3/1, 1	16 x 16 x 256
	MFM	-	16 x 16 x 128
POOL4	Pool	2 x 2/2	8 x 8 x 128
Fc1	Fc1	-	512
MFM_Fc1	MFM_fc1	-	256

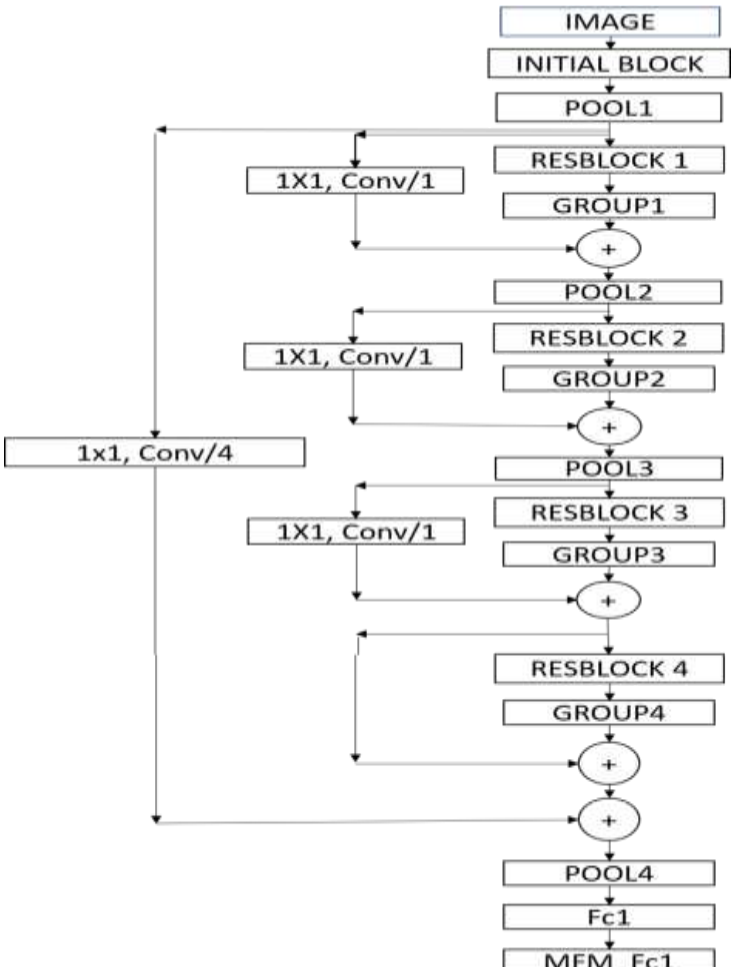

Figure 1: *L-MRN Architecture*

III. Performance Evaluation

Our experiment shows that L-MRN framework converges faster than Light CNN framework and has better accuracy. To train the L-MRN, we randomly select one face image from each identity in the dataset as the validation set and the remaining images as the training set. PyTorch, an open source deep learning framework is employed to implement the model. MS-Celeb-1M dataset is used to train the model. Like LightCNN we also use gray scale face images instead of RGB image for both training and testing purpose. The face images are aligned

to 144 x 144 by the five landmarks [23] and then randomly cropped to 128 x 128 as inputs as shown in fig.2. Besides, each pixel (ranged between [0, 255]) is divided by 255.

Figure 2: *FACE-ALIGNMENT*

It is important to choose a suitable number of shortcut levels for a satisfying performance. The more shortcut levels chosen, the more branches and parameters are added and thus increase the computational cost and storage space. Also, the overfitting problem will be exacerbated, and the performance may decrease. Therefore, for knowing the best number of levels we conducted some experiments by varying the shortcut levels. The improvement will be less obvious if the number of levels is too small or too high. After experimenting with the different shortcut levels, we found that shortcut level m=2 didn't gave much difference than the original network. Level m=3 and m=4 gave nearly same results, but level 3 gave best performance considering that the number of parameters in level 4 increased. After further increasing the levels, the performance started degrading. So, we chose that m=3 will be the best shortcut-level number for L-MRN. Our L-MRN network achieves 99.25% accuracy on MS-CELEB-1M dataset without fine tuning, which is slightly better than Light CNN.

Fig.3. shows how rapidly the accuracy of L-MRN increases than Light CNN, showing that our framework converges faster than the Light CNN framework while training without increase in the number of parameters of the network.

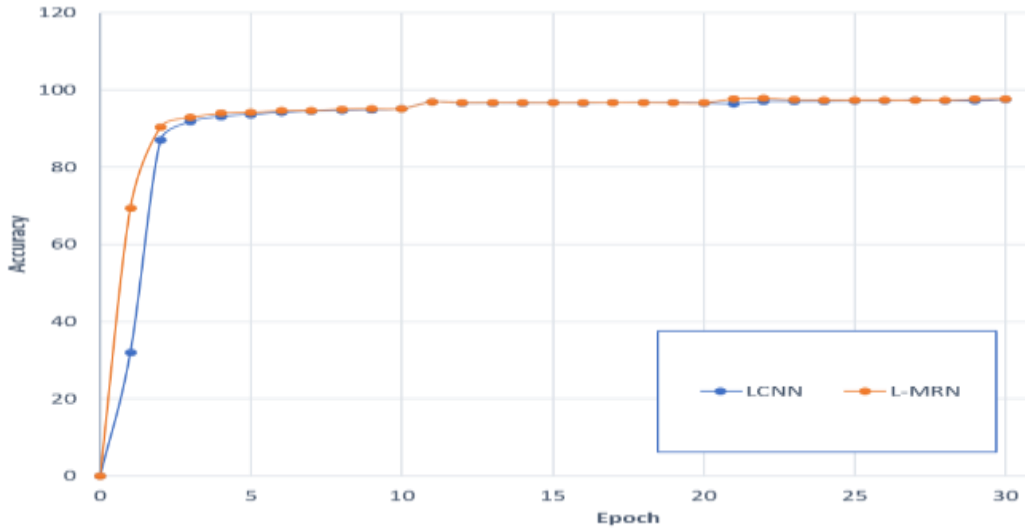

Figure 3: *Accuracy of LCNN and L-MRN while training on MS-CELEB-1M dataset.*

IV. Conclusion

In this paper, we have developed a Light CNN with multilevel residual network (L-MRN) framework to learn a robust face representation. Inspired by RoR we added level wise shortcut connections to the LightCNN framework and increased the efficiency of LightCNN network. Results shows that our L-MRN framework converges faster than LightCNN network without increasing the number of parameters.

References

[1]. X. Wu, R. He, Z. Sun and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," in *IEEE Transactions on Information Forensics and Security*, Nov. 2018.

[2]. Y. Bengio, P. Simard, and P. Frasconi. "Learning long-term dependencies with gradient descent is difficult". *IEEE transactions on neural networks, 5(2):157–166, 1994.*

[3]. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks". *In Advances in Neural Information Processing Systems*, pages 550–558, 2016.

[4]. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition". *Neural computation*, 1(4):541–551, 1989

[5]. J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition". *In ICML*, 2014.

[6]. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks". *In NIPS*, 2012.

[7]. M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional neural networks". *In ECCV*, 2014.

[8]. Ke Zhang, Miao Sun, Tony X. Han, Xingfang Yuan, LiruGuo, Tao Liu; "Residual Networks of Residual Networks: Multilevel Residual Networks". *arXiv:1608.02908v2 [cs.CV]*

[9]. S. Zagoruyko and N. Komodakis, "Wide residual networks". *arXiv preprint arXiv:1605.07146*, 2016.

[10]. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: "Integrated recognition, localization and detection using convolutional networks". *arXiv preprint arXiv:1312.6229*, 2013.

[11]. M. Abdi, S. Nahavandi;" Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks". *arXiv:1609.05672v4*, 2016

[12]. K. He, X. Zhang, S. Ren, and J. Sun, "Identity mapping in deep residual networks,"*arXiv preprint arXiv:1603.05027*, 2016.

[13]. G. Huang, Y. Sun, Z. Liu, and K. Weinberger, "Deep networks with stochastic depth," *arXiv preprint arXiv:1605.09382*, 2016.

[14]. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions". *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[15]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *In ICLR*, 2015

[16]. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," *In European Conference on Computer Vision, pages 740–755. Springer*, 2014.

[17]. Chu, D. Yang, R. Tadinada; "Visualizing Residual Networks", *arXiv:1701.02362v1*

[18]. F. Shen, and G. Zeng, "Weighted residuals for very deep networks," *arXiv preprint arXiv:1605.08831*, 2016.

[19]. Han D, Kim J, Kim J, "Deep pyramidal residual networks," *inProc. CVPR.*, 2017

[20]. Ke Zhang, LiruGuo, Ce Gao, Zhenbing Zhao, "Pyramidal RoR for Image Classification", *arXiv:1710.00307 [cs.CV]*

[21]. S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[22]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition". *arXiv preprint arXiv:1512.03385, 2015*.

[23]. Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," *in IEEE Conference on Computer Vision and Pattern Recognition, 2013.*

[24]. Shah, S. Shinde, E. Kadam, and H. Shah, "Deep residual networks with exponential linear unit," *arXiv preprint arXiv:1604.04112, 2016.*