Facing Face Recognition with ResNet: Round One

Ivan Gruber[1,2,3(✉)], Miroslav Hlaváč[1,2,3], Miloš Železný[1], and Alexey Karpov[3,4]

[1] Faculty of Applied Sciences, Department of Cybernetics, UWB, Pilsen, Czech Republic
grubiv@ntis.zcu.cz, {mhlavac,zelezny}@kky.zcu.cz
[2] Faculty of Applied Sciences, NTIS, UWB, Pilsen, Czech Republic
[3] ITMO University, St. Petersburg, Russia
karpov@iias.spb.su
[4] SPIIRAS, St. Petersburg, Russia

Abstract. This paper presents initial experiments of an application of deep residual network to face recognition task. We utilize 50-layer deep neural network ResNet architecture, which was presented last year on CVPR2016. The neural network was modified and then fine-tuned for face recognition purposes. The method was trained and tested on challenging Casia-WebFace database and the results were benchmarked with a simple convolutional neural network. Our experiments of classification of closed and open subset show the great potential of residual learning for face recognition.

Keywords: Face recognition · Classification · Neural networks · Data augmentation · Residual training · Computer vision

1 Introduction

Face recognition has been one of the most intensively studied topics in computer vision for last few decades and received great attention because of its applications in various real-world problems, among which can be counted security and surveillance systems, general identity recognition or person database investigation. The most significant usage of face recognition is in biometrics. Compared with some other biometrics techniques (fingerprints, iris, etc.) face recognition has the potential to non-intrusively recognize subject without any further cooperation of the subject. This and the fact, that camera sensors are much cheaper than various fingerprints or iris scanners, make face recognition most attractive biometric technique. Another important thing about face recognition is, that humans identify other people according to their face too, therefore they are likely to be comfortable with systems that use this approach.

In this article, we presented a method for face classification in unconstrained conditions based on an residual neural network [1]. The method is trained and

A. Ronzhin et al. (Eds.): ICR 2017, LNAI 10459, pp. 67–74, 2017.
DOI: 10.1007/978-3-319-66471-2_8

also tested on the major part of Casia-WebFace database [2]. We performed two types of experiments: (1) face classification of the closed subset; (2) face classification of the open subset; both with very promising results.

This paper is organized as follows: in Sect. 2 we provide quick look on the literature in this area; in Sect. 3 we describe used database and our data augmentations; in Sect. 4 we discuss presented method; in Sect. 5 we show obtained experimental results, while we draw conclusions and discuss future research in Sect. 6.

2 Related Work

In recent years, face recognition is experiencing renaissance through the deep convolutional neural networks (CNNs). After the breakthrough article from Krizhevsky et al. [3] in 2012, many novels approaches using CNNs for face recognition appeared. To name the most important ones, let's mention following works. In 2014, Taigman et al. [4] with their method based on CNN and effective face alignment significantly improved recognition rate on the LFW dataset [5], standard benchmark dataset for face recognition. Later in 2014, Sun et al. [6] presented method using CNN to learn effective features for face recognition. They further improved their features in article [7]. In 2015, Lu and Tang [8] surpassed human-level face verification performance on LFW using novel method based on GaussianFaces. Finally, Schroff et al. [9] reached almost flawless performance on LFW with their deep neural network trained to provide 128-D feature vector using triplet-based loss. It should be noted, that their method doesn't need any face alignment. This is caused by enormous amount of used training data (260 million images).

Despite all of the above-mentioned methods, the obtained state-of-the-art results are unrealistically optimistic from the position of pose-invariant face recognition, because most of the LFW images can be classified as near-frontal. Therefore most of existing algorithms can have problems with pose variances. To address this problem there were created several datasets (for example CASIA WebFace [2] or IJB-A [10]) containing variable external conditions including extreme pose variances. In 2016, Masi et al. [11] improved state-of-the-art results for datasets including large pose variances by presenting method based on Pose-Aware CNN Models.

Later in 2016, He et al. [1] proposed novel DNN architecture containing residual learning blocks to address a problem of degradation during learning very deep networks. This entry, which won the 1st place in ILSVRC-2015 [12], has very high potential for face recognition and it is probably the biggest upgrade of neural networks since Krizhevsky's AlexNet.

3 Dataset

As a training set for tested neural networks was chosen Casia-WebFace database. Casia-WebFace database is the second biggest publicly available database

for face recognition (the biggest one is The Megaface dataset [13]) and contains 494414 RGB images of 10575 subjects with resolution of 250×250 pixels. All images were semi-automatically collected from the Internet, i.e. persons are captured in variable conditions including pose, illumination, occlusion, age variations, haircut changes, sunglasses, etc. Most faces are centered on the images. For exemplary images see Fig. 1.

Fig. 1. Exemplary images from Casia-WebFace database

For the training, we decided to use only identities, which have at least 100 images presented. This leaves us with 181901 images for 925 identities. In order of CNN training, the resolution of all images was decreased to 64×64 pixels. To enrich these data we generated flipped version of each image. To further enriching training set and balancing counts of images per identity (which is very important for neural network training) we used data augmentation. To be more concrete, we modified images with Gaussian blur, noise, and brightness transformations. This leads to 908953 images in total. The database was appropriately split (each unique image with all its augmentations was in the same subset) into three subsets - training, validation and testing set, whereas 70% of images were used for the training set, 15% for the validation (development) set and 15% for the testing set.

4 Method

Due to the neural networks improvements in recent years, most hand-crafted feature descriptors for face recognition, if enough data is available, become obsolete. This results from the fact, that such descriptors use the same, hardly optimal,

operator to all locations of the face. On the other hand, learning-based descriptors can discover optimal operators for each location in the facial image.

Firstly, we train simple CNN on a multi-class (925 classes) face recognition task by minimizing cross-entropy loss to obtain benchmark results. The input is an RGB face image with size 64×64 pixels. The CNN's architecture contains three convolutional layers (32 filters each), each followed by ReLU non-linearity and max-pooling layer (to fight with overfitting and more robustness to local translation). The architecture is ended with a fully-connected layer containing 925 neurons (one for each class). For updating CNN's parameters $\boldsymbol{W}$ we used standard SGD optimization method. CNN was trained with the mini-batch size of 64 images during 850k iterations.

Secondly, we train deep residual network (DRN) with an architecture based on ResNet-50. DRN was proposed by He et al. [1] to address a problem of degradation during learning very deep neural networks. During testing, authors observed counter-intuitive phenomenon - adding more layers to the architecture causes higher training error. Historically this problem occurred, because of vanishing/exploding gradients during backpropagation, however, this phenomenon has been largely addressed by normalized initialization and intermediate normalization layers. Therefore, this degradation of training accuracies indicates that not all systems are similarly easy to optimize.

Authors address the degradation problem by introducing a deep residual learning framework. Instead of learning each group of stacked layers directly fit a desired underlying mapping, they let these layers fit a residual mapping. Let $H(\boldsymbol{x})$ be the desired underlying mapping of a group of stacked layers with $\boldsymbol{x}$ denoting the input to the first of these layers. Based on the hypothesis that multiple nonlinear layers can asymptotically approximate complicated functions, multiple nonlinear layers should be able to asymptotically approximate the residual function, i.e. $F(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}$. The original function thus becomes $F(\boldsymbol{x}) + \boldsymbol{x}$. Although both forms should be able to asymptotically approximate the desired functions, the ease of learning is different. The formulation of $F(\boldsymbol{x}) + \boldsymbol{x}$ we realized by the shortcut connections as element-wise addition (Fig. 2).

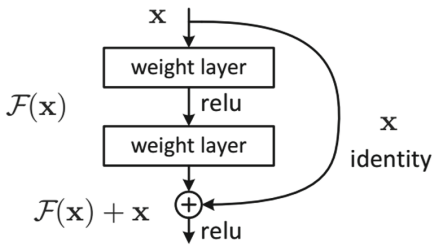

Fig. 2. Residual learning: a building block [1]

Formally building block is defined as:

$$\boldsymbol{y} = F(\boldsymbol{x}, \{\boldsymbol{W}_i\}) + \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{y}$ is the output vector of the layers and $F(\boldsymbol{x}, \{\boldsymbol{W}_i\})$ represents the residual mapping to be learned. The main advantage of this configuration is that the computational complexity of element-wise addition is negligible. The dimensions of $\boldsymbol{x}$ and F must be equal, however, if this is not the case, then linear projection of $\boldsymbol{x}$ can be performed. The function F can represent multiple fully-connected or convolution layers, in the later case the element-wise addition is performed channel by channel. He et al. experimented with F that contains one, two or three layers, however, if F has only a single layer, no advantages were observed.

As was already said, we trained NN with the architecture based on their 50-layer residual network. The network contains in total 16 residual learning building blocks each followed by ReLU non-linearity. Function F represents three convolutional layers, first two are also followed by ReLU non-linearity, in all cases. The only two differences between our and their architecture are the different size of the last (fully-connected) layer and the different size of the last average-pooling layer in ours. This second change was made because our input to the network has size 64×64 instead of original 224×224 pixels. The initial idea was to trained ResNet from the beginning with random initialization of weights $\boldsymbol{W}$, however, we believe, that low-level features extracted in the original classification task (ILSVRC-2015) are for face recognition relevant enough, that we can utilize already pre-trained weights of the original ResNet-50. Therefore, we decided to perform fine-tuning of original weights, whereas weights of convolutional layers in first 14 residual learning blocks were fixed. ResNet-50 was trained with the batch size of 6 during 120k iterations. For updating weight parameters $\boldsymbol{W}$ we used SGD optimization again.

Both neural networks were implemented in Python using Caffe deep learning framework [14].

5 Experiments and Results

In this section, we present the experimental results of our two experiments. Both experiments were performed on our test set described in Sect. 3.

5.1 Classification of the Closed Subset

In our first experiment, we evaluate our networks on the face classification of closed subset task. This type of task is easier than the classification of the open subset, because we don't allow any face image of foreign identity to come to the input of the network, therefore we don't need any thrash class or confidence threshold. NN simply always classifies the input face image into one of 925 trained classes.

Simple CNN reached after 850k training iterations only average results, to be more concrete 72.4% and 71.0% recognition rate (RR) on the development set and the test set respectively. It should be noted, that it can not be expected

Table 1. Comparison of classification recognition rates

Method	Development set	Test set
CNN	72.4%	71.0%
ResNet-50-60	89.3%	87.9%
ResNet-50-120	91.5%	90.7%

to obtain state-of-the-art results with such simple architecture. Nevertheless, we trained this network primarily to get benchmark results.

First, we fine-tune ResNet-50 with only 60k iterations. Thus trained network already reaches very promising result (89.3% and 87.9% RR). Next, we double the number of training iterations to reach even better results - 91.5% and 90.7% RR on development and test set respectively, which are very good rates on such challenging database as Casia-WebFace. We do not reach better results with further increasing of the number of iterations. Nevertheless, we assume, that with bigger training set, we can reach even better results and we definitely will examine this possibility in our future research. The comparison of results of classification is showed in Table 1.

Overall, this experiment showed the great potential of ResNet and residual learning for face recognition.

5.2 Classification of the Open Subset

In our second experiment, we evaluate our method on the face classification of open subset task. As it was already suggested, classification of closed subset can be simply extended by implementing confidence threshold. If the value of probability for the most probable class would be below this threshold, person's identity would be claimed as an unknown. It would be classified without any change otherwise. As the unknown persons, we use facial images of identities discarded during creating training set (all identities with less than 100 images presented). Moreover, we utilize facial images from the MUCT face database [15].

The optimal confidence threshold is found on the basis of false acceptance rate and false rejection rate, see Fig. 3. As you can see, best possible results are reached by choosing threshold value approximately equal to 0.974, however, it should be noted, that for practical usage both rates are not equally problematic. Usually, it is not a problem to repeat login process into your bank account, because of false rejection of the system. But if someone else logs into your bank account, it will be definitely considered as a big security hole in the system. Therefore, we believe, that more strict threshold, 0.984 for example, is actually better.

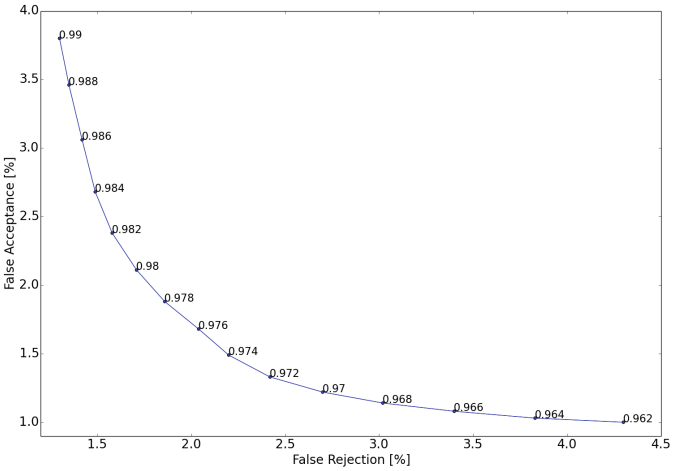

Fig. 3. False acceptance vs false rejection for open subset classification

6 Conclusion and Future Work

Face Recognition is a very challenging problem, which, despite all the effort and popularity in recent years, can be in uncontrolled conditions hardly solved. Our work shows that residual training of deep neural networks has big potential for face recognition tasks. We believe, that with some minor modifications of ResNet architecture and more augmentations of the training set, or with bigger training set, has residual training of neural networks potential to improve state-of-the-art results.

In our future research, we would like to focus on the testing of ResNet on verification task, which has big significance for real-world applications. We are planning to extract features from the penultimate layer of the network and on the basis of their weighted chi-square distance perform face verification. Moreover, we would like to enrich training set by synthesizing new facial images in different poses using a 3D morphable model.

Acknowledgments. This work is supported by grant of the University of West Bohemia, project No. SGS-2016-039, by Ministry of Education, Youth and Sports of Czech Republic, project No. LO1506, by Russian Foundation for Basic Research, projects No. 15-07-04415 and 16-37-60100, and by the Government of Russian, grant No. 074-U01. Moreover, access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
2. Yi, D., Lei, Z., Liao, S., Li, Z.: Learning face representation from scratch. CoRR (2014)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing, pp. 1106–1114 (2012)
4. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
5. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments, pp. 07–49 (2007)
6. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification, pp. 1–9. CoRR (2014)
7. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust, pp. 2892–2900. CoRR (2014)
8. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with Gaussian face. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 3811–3819 (2015)
9. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
10. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1931–1939 (2015)
11. Masi, I., Rawls, S., Medioni, G., Natarajan, P.: Pose-aware face recognition in the wild. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4838–4846 (2016)
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV), pp. 211–252 (2015)
13. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The MegaFace benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding (2014). arXiv preprint arXiv:1408.5093
15. Milborrow, S., Morkel, J., Nicolls, F.: The MUCT landmarked face database. In: Pattern Recognition Association of South Africa (2010)