PROCEEDINGS

of the

2017 Symposium on Information Theory and Signal Processing in the Benelux

May 11-12, 2017, Delft University of Technology, Delft, the Netherlands
http://cas.tudelft.nl/sitb2017

Richard Heusdens & Jos H. Weber (Editors)

ISBN 978-94-6186-811-4

The symposium is organized under the auspices of

Werkgemeenschap Informatie- en Communicatietheorie (WIC)

& IEEE Benelux Signal Processing Chapter

and supported by

Gauss Foundation (sponsoring best student paper award)

IEEE Benelux Information Theory Chapter

IEEE Benelux Signal Processing Chapter

Werkgemeenschap Informatie- en Communicatietheorie (WIC)

TU Delft
Delft University of Technology

IEEE

Deep Verification Learning

Fieke Hillerström Raymond Veldhuis Luuk Spreeuwers
University of Twente
Dept. Services, Cybersecurity and Safety

Abstract

Deep learning for biometrics has increasingly gained attention over the last years. The expansion of computational power and the increasing dataset sizes, increased verification performances. However, large datasets are not available for every application. We introduce Deep Verification Learning, to reduce network complexity and train on smaller datasets. Deep Verification Learning takes two images to be verified at the input of a network, and trains directly towards a verification score. We applied Deep Verification Learning on the face verification task, also it could be extended to other biometric modalities.

1 Introduction

Deep learning face recognition has been extensively studied during the last years and has obtained impressive results [1, 2, 3, 4]. The increasing availability of computational power and training data allows for the training of deeper networks. We introduce Deep Verification Learning to reduce the network complexity and enable training on smaller datasets (see Figure 1). Most of the state-of-the-art deep learning face recognition systems use convolutional networks. For face verification, commonly a framework based on multi-class classification is used [4] (see Figure 2). We define this type of learning as 'Identification Learning'. One of the challenges in deep learning face recognition is data bias [4]. The availability of training data is limited for applications that do not utilize public web images (see Table 1). For those applications it is interesting to investigate less complex deep learning architectures. We propose a Deep Verification Learning system, directly trained for a verification score (see Section 3). We applied Deep Verification Learning on the task of face recognition and it could be extended towards other biometric modalities. Deep Verification Learning offers several advantages over Identification Learning. Training a network in pairs enables the creation of extra training samples. Providing two images as input of the network enables it to learn face similarities and differences directly at the first layers. Training towards a verification score instead of multi-class classification reduces the number of network parameters drastically. Given these advantages, we hypothesize that our network can train more effectively on small datasets.

We investigate the ability of Deep Verification Learning by comparing Deep Verification Learning with Identification Learning (see Section 4). We explore the benefits of increasing the dataset size in a controlled manner. The research questions addressed in this paper are:

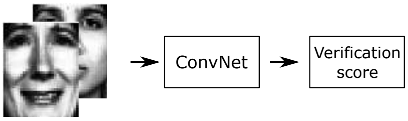

Figure 1: Deep Verification Learning. Two images are presented as input of the network and the system is directly trained towards a verification score. Face images are preprocessed images from the FRGC dataset [5].

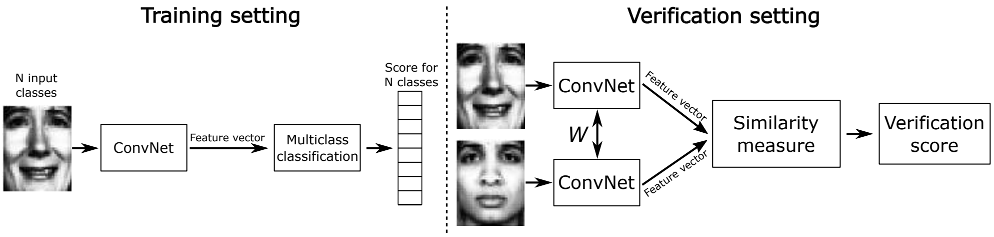

Figure 2: Identification Learning convolutional network. Left: Training for multi-class classification. Right: Two networks are replicated in verification setting. The networks have the same fixed weights W. A new top-layer is trained.

Dataset	**#Images**	**#Subjects**	**Access**	**Source**
LFW [6]	13,233	5,749	Public	Celebrity search
CelebFaces+ [7]	202,599	10,177	Public	Celebrity search
CASIA-WebFace [8]	494,414	10,575	Public	Celebrity search
MS-Celeb-1M [9]	10M	100K	Public	Celebrity search
Social Face Classification [1]	4.4M	4,030	Private	Facebook
Google [3]	100-200M	8M	Private	Undefined
Megvii Face Classification [4]	5M	20K	Private	Celebrity search
FRGC [5]	39,328	568	Public	Photo sessions

Table 1: Datasets used for training deep learning face recognition networks and their characteristics.

1. Can Deep Verification Learning result in similar or better face verification performance then Identification Learning?

2. What is the effect of the number of images in a dataset on the face verification performance of both Deep Verification Learning and Identification Learning?

Related background is described in Section 2. We explain Deep Verification Learning in Section 3, followed by our experiments. Finally we present our conclusions.

2 Related Work

Deep learning face verification systems commonly use a framework based on multi-class classification [4]. The networks are trained for subject identification. The identification layer is removed and a feature vector remains. A similarity measure is added for verification (see Figure 2). Different types of top layers can be used. Examples of untrained methods are the inner product between two normalized feature vectors [1] or the L2 norm [4]. Possible trained methods are the weighted-χ^2 distance, Joint Bayesian [10], and a new-trained neural network [10]. Most of the traditional and deep learning topologies extract the features of two faces separately. Verification signals can enhance training. DeepID2 [11] combines verification and identification signals into a joint cost function. FaceNet [3] uses a triplet loss function, which separates genuine pairs from impostors. Most architectures use the output of a non-final layer as feature vector for verification. Deep Verification Learning trains directly towards a verification score, instead of using an intermediate representation. The datasets used for training these deep networks are commonly large (see Table 1). The trained networks do not generalize well to new applications [4]. In case of controlled applications, only small datasets are available. Deep Verification Learning reduces the number of parameters significantly: for a dataset of N subjects, the last layer of an Identification Learning

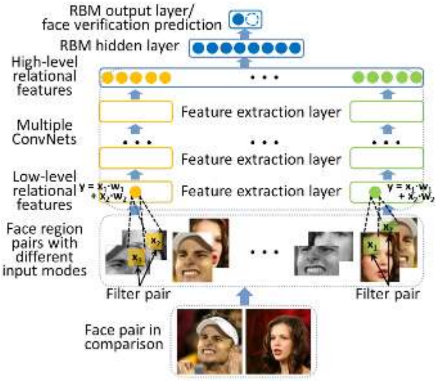

Figure 3: System architecture of the hybrid ConvNet-RBM model, proposed by Sun et al.. Image from [12].

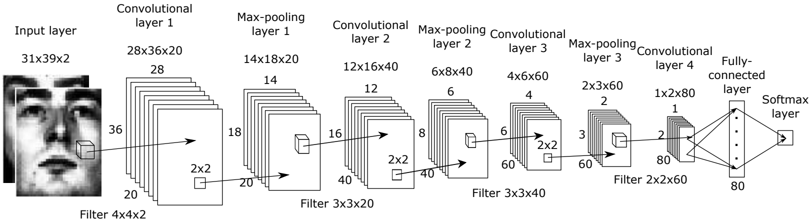

Figure 4: Deep Verification Learning network. At the input two grayscale images are presented. The network is directly trained towards a verification score. Topology based on [12].

network is an N-way softmax-layer. With K the length of the second last layer, $K \cdot N$ parameters need to be trained. For Deep Verification Learning this number is only $2 \cdot K$ parameters.

Sun et al. [12] proposed a similar verification learning architecture (see Figure 3). They train networks for twelve different face regions, each containing five different convolutional networks and average the output of eight different input modi. On top of these networks they train a classification RBM. They train on a large dataset. Our aim is to design a less complex network that can train on smaller datasets. Therefore we propose a simplified version of the architecture proposed by Sun et al., containing a basis of their proposed network.

3 Deep Verification Learning

We introduce a Deep Verification Learning network, based on the architecture proposed by Sun et al. [12]. The network trains directly trained towards a verification score (see Figure 4). Max-pooling layers and the ReLU activation function are used. Two grayscale images (31x39) are presented as input of the network, each in a separate channel. The output layer is a 2-way softmax layer, predicting a verification score. Using grayscale images reduces the number of trainable parameters and does not reduce face recognition performance [13]. The number of parameters in the network, excluding the final softmax-layer, is 13,612. The softmax-layer adds 160 parameters. In the case of Indentification Learning, the final softmax-layer adds $80 \cdot N$ parameters, which increases the trainable parameters in the network immensely.

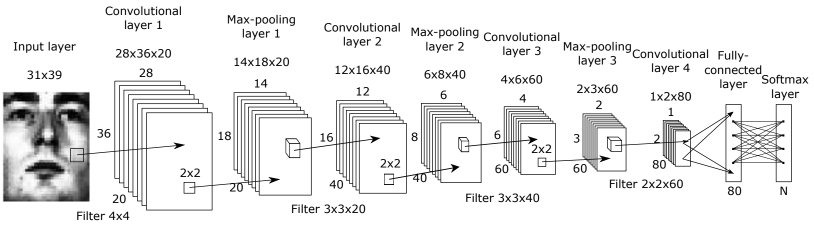

Figure 5: Training the convolutional network for multi-class classification. A N-way softmax-layer calculates a probability score for N subject input classes.

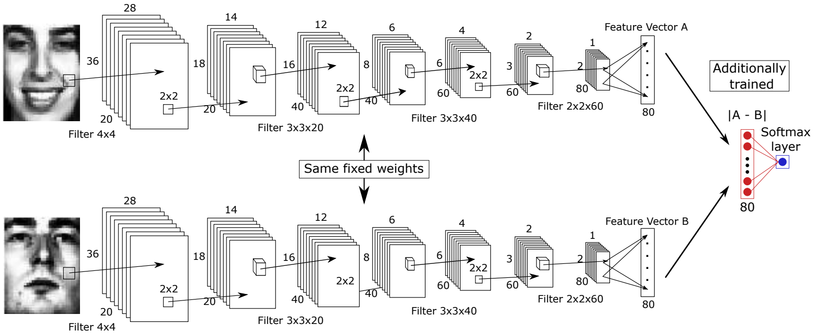

Figure 6: Identification Learned convolutional network for face verification. The feature vectors of both images are taken as input for a two-way softmax-top-layer.

4 Experiments

4.1 Experiment setup

The architectures of both systems to be compared are identical, except for the input and output layer. The Deep Verification Learning network is discussed in Section 3. The Identification Learned network is trained for classification (see Figure 5). After training, the classification layer is removed and a feature vector is obtained. Two feature vectors of the images to be compared are normalized to zero mean and unit variance and taken as input for a new-trained two-way softmax-layer. With x_A and x_B referring to normalized feature vectors, the input for the softmax-layer is $|x_A - x_B|$ (see Figure 6). Normalized input vectors resulted in higher verification performances in our experiments, probably because the ReLU expects the input data to be zero centered.

The networks are trained using SGD and a mini-batch size of 32 and learning rate of 0.005. Before every epoch, the training samples are shuffled. The cost function is the negative log-likelihood and normalized to the number of samples in target class t, to compensate for an unbalanced dataset. Xavier initialization [14] is used. Early stopping is applied, to prevent the networks from overfitting. A validation set is created with pairs of face images. The Identification Learning validation set contains single face images from the N training classes. During an epoch, the validation cost is evaluated after ten mini-batches. The parameters resulting in the lowest validation cost are saved as 'best'. When training does not improve the validation cost, training is stopped and the saved 'best' are taken as final model. For Identification Learning the rank-1 recognition score is used for early stopping.

The first experiment is performed using the controlled images of the FRGC dataset [5], which contains 24,614 controlled images, from 568 subjects. The second and third experiment use a combination of twelve different public datasets, merged and used by Zeng et al [15], containing 438,319 images of 13,671 subjects. Training and validation is split 50-50 on subjects. The validation set is used for both the model evaluation and in the early stopping algorithm. This adds a bias to the results in deciding when the network start to overfit. The validation set is not used for updating the parameters itself and thus the learned features. For a particular training set and network combination we expect that the network starts overfitting around the same number of training iterations when tests are repeated. Therefore we expect the bias to be small and to have no influence on the comparison.

The input images are converted to grayscale values and registered on the eyes, using annotated coordinates. The images are cropped to a fixed box around the eyes' and scaled to a fixed size of 31x39 pixels. Thereafter the images are histogram equalized and converted to have zero mean. Data augmentation is applied in the training set, by adding all possible modi (horizontal flipped and different order).

4.2 Performance comparison

The two architectures are compared on their verification performance using the controlled images of the FRGC dataset. The subjects are randomly split in training and validation sets (284 subjects each). Training set contains 32,136 pairs from 11,920 images. Identification validation set contains 1,192 images and verification validation set 10,000 pairs from 12,694 images. The training set contains genuine pairs for every subject, with a maximum of 66 pairs per subject. In the validation set the genuine pairs are equally distributed. The initialization and training of both networks is repeated five times. The ROC curves of all ten tests are shown in Figure 7. Identification Learning networks show a higher variance in performance. The variance in the Area Under Curve (AUC) is 1.4×10^{-6} and 3.0×10^{-8}, respectively. We re-training the top layer for a fixed network five times, to explore the cause of this variance. The results are shown in Figure 8 and the variance in the AUC is 3.0×10^{-9}. We conclude that the variance in performance is created when training the network for multi-class classification.

The performance of Identification Learning is highly dependent on the multi-class classification training, which we expect is caused by the high number of parameters. The Deep Verification Learning model has only 13,772 parameters to train, as opposed to the 36,316 parameters of the Identification Learning network. With a dataset of only 10,728 training images, it is likely that the multi-class classification network overfits. Deep Verification Learning performs substantially better than Identification Learning in our experiments. However, some remarks have to be made. Only one type of newly-trained top layer for Identification Learning is evaluated. The dataset used for testing has its limitations. The number of images is low compared to the number of parameters in the networks, especially in the case of Identification Learning this can lead to overfitting. Further experiments should be done on datasets with a higher variety.

4.3 Dataset configuration comparison

Increasing the amount of training data typically improves the performance. Datasets could be expanded by adding more subjects to a dataset and/or by adding more images per subject. We attempt to evaluate the effect of both aspects separately. The validation set is kept the same in every test, containing 20,000 pairs from 221,562 images of 6,835 subjects. The networks are trained on the datasets specified in Table 2 and tests are repeated three times. The Area Under Curve values of these tests are shown in Figures 9 & 10. Contrary to the expected increase in performance, the performance in fact declines, which we expect is caused by the network strongly overfitting on the

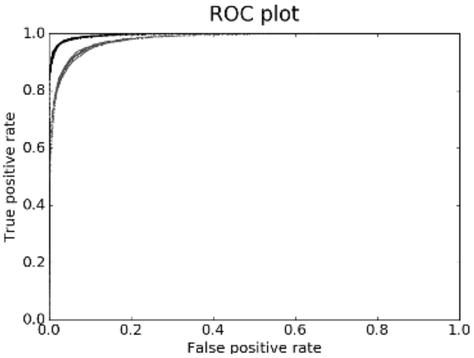

Figure 7: ROC curves for the five Deep Verification Learned networks (black) and the five Identification Learned networks (gray) on the FRGC controlled dataset.

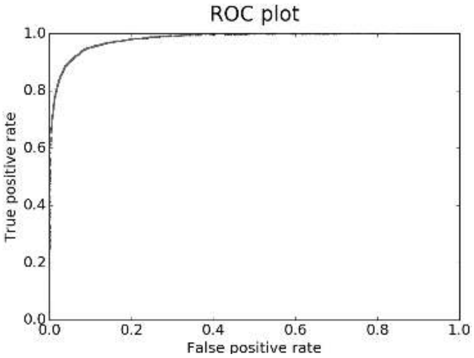

Figure 8: ROC curves for the five re-trained top layers with the same fixed convolutional network.

Increasing number of images per subject				
#Subjects	#Images per subject	#Images total	#Gen. pairs per subject	#Training pairs
211	20	4,220	190	80,180
211	40	8,440	780	329,160
211	80	16,880	3,160	1,333,520
211	160	33,760	12,720	5,367,840

Increasing number of subjects				
#Subjects	#Images per subject	#Images total	#Gen. pairs per subject	#Training pairs
211	20	4,220	190	80,180
422	20	8,440	190	160,360
844	20	16,880	190	320,720
1688	20	33,760	190	641,440

Table 2: Dataset details for testing with increasing number of images per subject, increasing number of subjects in the dataset, and the validation set.

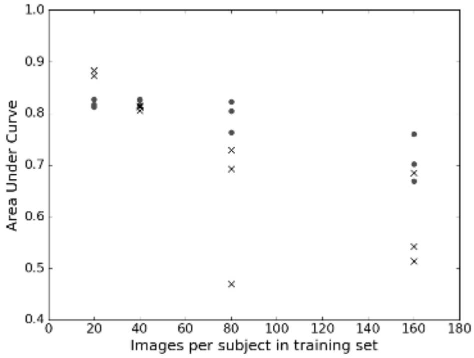

Figure 9: Area Under Curve values for the Deep Verification Learned networks (x) and the Identification Learned networks (o), for different number of images per subjects in the dataset.

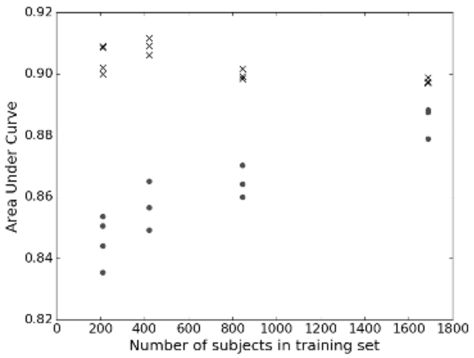

Figure 10: Area Under Curve values for the Deep Verification Learned networks (x) and the Identification Learned networks (o), for different number of subjects in the dataset.

Test	**Imp. different datasets**			**Imp. same datasets**	
80 Images per subject	0.729	0.469	0.693	0.719	0.735
160 Images per subject	0.542	0.514	0.684	0.718	
844 subjects	0.899	0.898	0.902	0.898	0.900
1688 subjects	0.897	0.899	0.897	0.904	

Table 3: AUC values for Deep Verification Learning network for different imposter pair construction in the trainingset. Imposters are formed within all the possible datasets (left) or imposter pair forming is restricted to subjects within the same dataset (right).

training data. The Deep Verification Learning performance is more affected by this than Identification Learning is. Only subjects with 160 images or more are used in the training set, creating a bias. It turned out subjects from only two datasets, met this requirement and were sampled into the training set. This causes overfitting to these types of data. We recommend repeating these experiments on a larger homogeneous dataset.

Another remarkable result is found when increasing the number of subjects in the training set. For Deep Verification Learning the performance declines when too many subjects are added, which may be caused by the way the network learns. When training in pairs it is possible to learn to compare type of images instead of recognizing faces; different type of images are unlikely to form a genuine pair. This could cause overfitting to the type of images represented in the training set. To investigate this hypothesis, with the imposter pairs only formed within a dataset. We found a small increase in performance (see Table 3), but not sufficient enough to draw conclusions. Therefore more experiments should be performed.

5 Conclusions & Recommendations

We introduced Deep Verification Learning and applied it to face verification. We compared Deep Verification Learning with a network trained for multi-class classification on the FRGC dataset and evaluated the effect of different dataset-sizes on verification performance. We found a notable improvement when Deep Verification Learning is used instead of Identification Learning. Increasing the dataset-sizes does not improve the verification performances as expected. Care must be taken in training set selection, to prevent the networks from overfitting.

Despite the promising results, more extensive experiments should be performed, with a separate validation set for the early stopping algorithm. We recommend the use of cross-validation and more different types of datasets. The network should be compared to state-of-the-art face recognition systems. The tests regarding the influence of different dataset sizes, should be performed in a way that dataset bias is not possible. The learning process is important for the final network performance. More insight into the current working of the network should be obtained. We concentrated on the simplicity of our network and advise to investigate advanced techniques that are found to enhance performance and learning. Effort should be made to obtain a training set with the same type of images as is used in the final application. The optimal number of images per subject and number of genuine/imposter pairs compared to the number of subjects in the dataset should be found. Data augmentation could be extended to increase the dataset size.

References

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.

[2] Y. Sun, D. Liang, X. Wang, and X. Tang, “Deepid3: Face recognition with very deep neural networks,” *arXiv preprint arXiv:1502.00873*, 2015.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[4] E. Zhou, Z. Cao, and Q. Yin, “Naive-deep face recognition: Touching the limit of lfw benchmark or not?,” *arXiv preprint arXiv:1501.04690*, 2015.

[5] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the face recognition grand challenge,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 947–954, IEEE, 2005.

[6] E. Learned-Miller, G. Huang, A. RoyChowdhury, H. Li, G. Hua, and G. B. Huang, “Labeled faces in the wild: A survey,”

[7] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, 2014.

[8] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv preprint arXiv:1411.7923*, 2014.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition,” in *European Conference on Computer Vision*, pp. 87–102, Springer, 2016.

[10] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, 2014.

[11] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in *Advances in Neural Information Processing Systems*, pp. 1988–1996, 2014.

[12] Y. Sun, X. Wang, and X. Tang, “Hybrid deep learning for face verification,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1489–1496, 2013.

[13] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Li, and T. Hospedales, “When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 142–150, 2015.

[14] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks.,” in *Aistats*, vol. 9, pp. 249–256, 2010.

[15] D. Zeng, H. Chen, and Q. Zhao, “Towards resolution invariant face recognition in uncontrolled scenarios,” in *Biometrics (ICB), 2016 International Conference on*, pp. 1–8, IEEE, 2016.