# ONE-SHOT LEARNING FOR FACE RECOGNITION USING SIAMESE NEURAL NETWORKS AND TRANSFER LEARNING

Filip Kornata
*School of Computer Science*
*University of Nottingham*

## ABSTRACT

This paper describes a deep learning approach for solving a face recognition problem with a small dataset and only one observation per class in a minimal time. It starts with relevant data augmentation techniques including the use of an external, aligned dataset. The concept of two Siamese networks with shared weights is explained and used along with transfer learning using pre-trained AlexNet. A shared, fully connected layer is added to calculate Cross-entropy loss. The model is compared to a hard-coded template-based algorithm. The introduced method gives a better result but requires a longer execution time to achieve it. However, it performs worse than existing approaches, where the time factor is not considered.

***Index Terms***— Siamese Networks, Transfer Learning, One-shot Learning, Face Recognition, Convolutional Neural Networks

## 1. INTRODUCTION

There are many highly accurate deep learning methods for face recognition trained on large databases. [1] gives a detailed review of most of them comparing different designs, databases, protocols, and application scenes. This paper focuses on classification with a restriction of only one observation per class, one-shot learning. It was initially approached using a variational Bayesian framework in [2], [3]. Then, [4] used a Siamese Neural Networks approach for one-shot learning to classify images of characters from the Omnigolot dataset. In this method, results from two similar CNNs are combined to determine whether the two input images are similar or dissimilar. Meanwhile, transfer learning had been gaining more interest in the field of face recognition [5]–[7]. In 2020, Siamese Networks, along with transfer learning, have been applied to the face recognition problem with underrepresented data in [8]. The results achieved in this study are high; however, performance in one-shot learning and time efficiency are not included, which leaves many questions about this technique unanswered. This paper focuses on exploring the mix of transfer learning and Siamese Networks to solve a one-shot learning face recognition problem in a limited time. The paper goes through each of the utilized methods, which include data preprocessing, network architecture, loss function, and evaluation. Then, the experimental results are presented. The aim of the experiments is to obtain predictions accuracy higher than the one achieved by the template-based approach, which is 25%, in the shortest possible time.

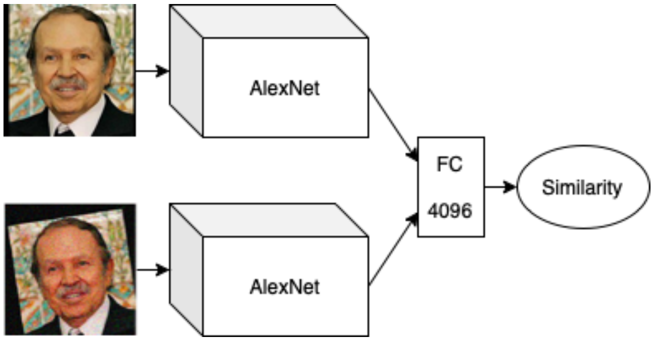

*Figure 1.* Siamese networks' architecture with two similar AlexNet neural networks and a fully connected layer (FC) combining the outputs and producing a similarity score. Two input images, the class representative (top) and an augmented version (bottom) visualize the nature of solving a one-shot learning problem.

## 2. METHODOLOGY

One-shot learning is a challenging problem because of the lack of diversity in training data. Augmentation helps to overcome this obstacle by making geometrical and visual changes to the images. The new images can be combined with the originals, to obtain a bigger and more diverse set of observations. Siamese network is a good choice for this problem, as it can effectively learn based on the dissimilarities of two images. To take advantage of that, an additional set of images, without any class affiliation, can be introduced to the network.

### 2.1. Data augmentation

The dataset of 100 faces of different people has been augmented based on the expectations of the data that the

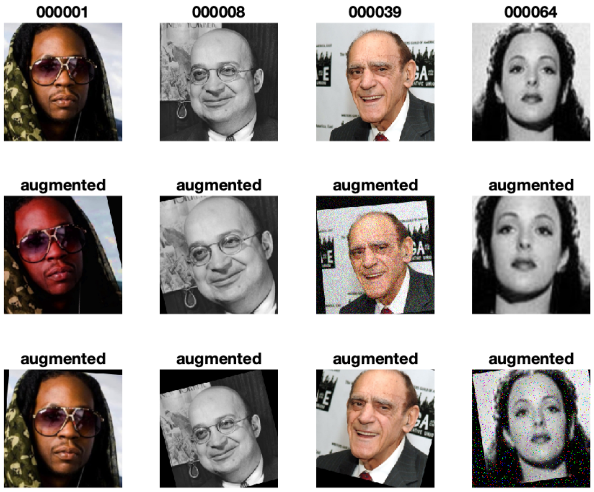

*Figure 2.* Examples of augmentation for four face images. Four original images on the top with two augmented examples of each.

model may be used for in the future. Scale, rotation, and translation are used to simulate different face positions and are applied at a random intensity between values -20 to 20 (in pixels), 0.9 to 1.2, and -20 to 20, respectively. Reflection is used with 0.1 probability. Hue, saturation, brightness, and contrast imitate different cameras and environmental conditions. These are applied with a probability of 0.1 and at random scales between -0.1 and 0.1 for the first three options, and 0.9 to 1.1 for contrast. Blur and noise help to recognize older images better. These modifications are applied at a random scale with a probability of 0.03 and 0.1, respectively. The augmentation process is repeated 10 times such that at the end of it, the size of the entire training and validation set is multiplied by 11. That size will allow the preprocessing to produce a variety of image mutations in each of the classes. Examples of augmented images are presented in Figure 2. I also used an additional dataset, CelebA [9] of unclassified, aligned faces and applied the same augmentation techniques as mentioned before. The size of this set is equal to the size of augmented training and validation sets combined. The pairs of images are grouped into batches. Each of them contains 20 images, which is enough to achieve a sufficient estimate of the gradient in a short time. In the process of grouping the pairs, the images from the augmented external dataset are used with a probability of 0.5, only for the pairs labeled as "dissimilar".

### 2.2. Network architecture

Siamese networks architecture consists of two similar CNNs that share the same weights. For this study, I use AlexNet [10] as a transfer learning base, which is a model trained on 1.2 million ImageNet images for 1000-class classification. Although it is already 10 years old, with its reasonable number of weights, it can give satisfying results in a relatively short time. For the reasons mentioned before, I keep the input sizes unchanged, even though it isn't necessary. The network contains 8 layers in total: 5 convolutional, and 3 fully connected layers. The only structural difference in the network used in this study is the removal of the last 1000-weight layer, as my loss function doesn't require a specific number of outputs at this point. The first three layers are completely frozen. That is, because of the random initialization of the weights in the fully connected layers. As some big weight updates might happen at some point, the trained weights could be affected. This analogy is explained in [8]. The next 2 convolutional layers are fine-tuned with the initial weights set to pre-trained. The last two layers are initialized randomly. The weights are updated using ADAM Optimizer [11] with a learning rate equal to 0.000005, gradient decay factor equal to 0.9, and squared gradient decay factor equal to 0.999.

| **Execution Time** | **Iterations** | **Prediction Accuracy** |
|---|---|---|
| 13 minutes | 500 | 12% |
| 35 minutes | 1000 | 24% |
| 68 minutes | 2000 | 32% |

*Table 1.* The accuracy of Siamese networks using transfer learning from AlexNet on a 100-class dataset with a single observation per class.

### 2.3. Loss function

Because of the unusual network architecture, a distance loss function is required to measure the magnitude of error in continuous probabilistic prediction. [8] used Contrastive loss function for that purpose. However, it didn't perform well for one-shot learning. Cross-entropy function shows better results and, therefore, is used to optimize the weights. To do that, each pair from a batch is passed through two separate subnetworks. Then, a sigmoid function is applied to each of the outputs to convert them into a probabilistic form. A calculated difference between the vectors is then passed through one additional fully connected layer with 4096 weights. The output is the absolute similarity value for each of the pairs in a batch. After using the sigmoid function again, Cross-entropy loss is calculated as follows:

$$L = -y * log(Y) - (1 - y) * log(1 - Y) \qquad (1)$$

where $Y$ is a predicted probability that a pair of images show the same person, and $y$ is the discrete label equal to 1 for the same person or 0, for images of two different people.

### 2.4 Validation and testing

The model is tested based on the similarity between each of the images in the test set and the anchor

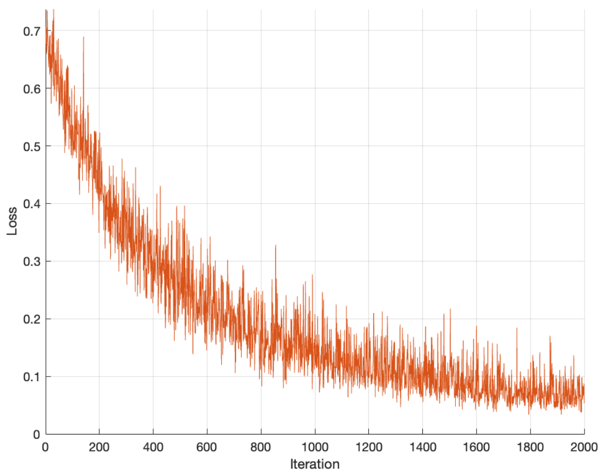

*Figure 3.* Cross-entropy loss in 2000 iterations of Siamese networks using AlexNet transfer learning on a 100-class dataset with a single observation per class.

image of each class. As it is highly inefficient to pass each of the images through the network separately, this is done using the entire test set at once. The same applies to the anchor images. Each of the test outputs is then multiplied 100 times and packed into a batch with 100 different class-representative images. After passing the batch through the fully connected layer mentioned in the previous subsection, the highest argument for each batch determines the predicted class. The same method is used for validation, which is performed every 20 iterations. The batch size used for this part is 30, as it provides a fair trade-off between speed and fidelity. The training/validation split used in the implementation is 70/30. All the images are downsized to 227x227x3 to reduce the training time while keeping their important features extractable.

## 3. METHOD EVALUATION

The aim of this experiment is to achieve the best result in the shortest possible time. There are several benchmarks of one-shot face recognition [8], [12]–[14], but there is no information on the time efficiency of the evaluated approaches. The result achieved by the template-based method is 25% prediction accuracy. The evaluation results of my solution to this problem are shown in Table 1. According to the table, the approach described in this study can achieve that in slightly over 35 minutes. Cross-entropy loss and validation accuracy curves for the training on 2000 iterations are presented in Figures 3 and 4. The approach with the result of 24% required 1000 iterations, and the one with 12% - 500. The experiments were carried out in MATLAB using Apple M1 Pro CPU and 16 GB of main memory.

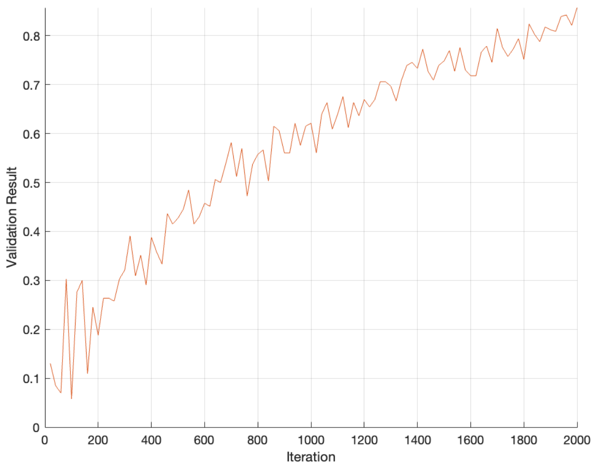

*Figure 4.* Cross-entropy loss in 2000 iterations of Siamese networks using AlexNet transfer learning on a 100-class dataset with a single observation per class.

## 4. DISCUSSION

The studied model managed to match the result to the template-based method after a 25-minute training, which is a success. It is slower than the mentioned method but gives better results with more time. However, the time efficiency of this method is not as good as expected. According to [10], AlexNet is well optimized for training on GPUs. Therefore, the results should be better for training with more powerful units. The validation curve in Figure 4 reaches near-maximum accuracy, which indicates that there is a big gap between validation and testing. Because the validation set is a subset of the underrepresented observations, it is unable to produce realistic accuracy predictions. It tends to overestimate them. This validation method doesn’t take advantage of the external dataset, which could be achieved using a measure counting the “good guesses” of similar/dissimilar pairs. On the other hand, such an approach would lack consistency with the evaluation. Another solution to that problem could be adding further adjustments to the augmentation stage. According to [1], recently, there has been a significant improvement in the field of loss functions. That suggests that a bit outdated Cross-entropy could be replaced with a better technique.

## 5. CONCLUSION

In this paper, I found a solution for a face recognition problem of 100-class classification using one observation per class. The solution achieved by Siamese networks with transfer learning based on AlexNet achieved the goal set in the introduction. This study shows that hard-coded approaches can be successful if a fast solution is necessary, and a medium-quality result is sufficient. Deep learning shows better results in longer periods. This study can be explored further by improving the validation technique, experimenting with other loss functions, or expanding the augmentation.

I would also recommend trying out different pre-trained CNNs for transfer learning.

## 6. REFERENCES

[1] M. Wang and W. Deng, ‘Deep Face Recognition: A Survey’, *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021, doi: 10.1016/j.neucom.2020.10.081.

[2] Li Fe-Fei, Fergus, and Perona, ‘A Bayesian approach to unsupervised one-shot learning of object categories’, in *Proceedings Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1134–1141 vol.2. doi: 10.1109/ICCV.2003.1238476.

[3] Li Fei-Fei, R. Fergus, and P. Perona, ‘One-shot learning of object categories’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006, doi: 10.1109/TPAMI.2006.79.

[4] G. Koch, R. Zemel, and R. Salakhutdinov, ‘Siamese Neural Networks for One-shot Image Recognition’, p. 8.

[5] A. Ar and A. Sethi, ‘Face Recognition and Verification using Transfer Learning’, 2018. doi: 10.13140/RG.2.2.26851.99367.

[6] R. M. Prakash, N. Thenmoezhi, and M. Gayathri, ‘Face Recognition with Convolutional Neural Network and Transfer Learning’, in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Nov. 2019, pp. 861–864. doi: 10.1109/ICSSIT46314.2019.8987899.

[7] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, ‘Feature Transfer Learning for Face Recognition With Under-Represented Data’, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5697–5706. doi: 10.1109/CVPR.2019.00585.

[8] M. Heidari and K. Fouladi-Ghaleh, ‘Using Siamese Networks with Transfer Learning for Face Recognition on Small-Samples Datasets’, in *2020 International Conference on Machine Vision and Image Processing (MVIP)*, Feb. 2020, pp. 1–4. doi: 10.1109/MVIP49855.2020.9116915.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang, ‘Deep Learning Face Attributes in the Wild’, *ArXiv14117766 Cs*, Sep. 2015, Accessed: May 11, 2022. [Online]. Available: http://arxiv.org/abs/1411.7766

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘ImageNet Classification with Deep Convolutional Neural Networks’, in *Advances in Neural Information Processing Systems*, 2012, vol. 25. Accessed: May 11, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[11] W. K. Newey, ‘Adaptive estimation of regression models via moment restrictions’, *J. Econom.*, vol. 38, no. 3, pp. 301–339, Jul. 1988, doi: 10.1016/0304-4076(88)90048-6.

[12] Y. Guo and L. Zhang, ‘One-shot Face Recognition by Promoting Underrepresented Classes’, p. 10.

[13] Z. Ding, Y. Guo, L. Zhang, and Y. Fu, ‘One-Shot Face Recognition via Generative Learning’, in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 1–7. doi: 10.1109/FG.2018.00011.

[14] L. Wang, Y. Li, and S. Wang, ‘Feature Learning for One-Shot Face Recognition’, in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct. 2018, pp. 2386–2390. doi: 10.1109/ICIP.2018.8451464.