Received: 26 April 2023 | Revised: 4 July 2023 | Accepted: 26 July 2023 | Published online: 14 August 2023

Artificial Intelligence and Applications
2024, Vol. 2(1) 21–27
DOI: 10.47852/bonviewAIA32021018

BON VIEW PUBLISHING

RESEARCH ARTICLE

Attention Enhanced Siamese Neural Network for Face Validation

Hong Qing Yu[1,*]

[1]*School of Computing and Engineering, University of Derby, UK*

Abstract: Few-shot computer vision algorithms have enormous potential to produce promised results for innovative applications which only have a small volume of example data for training. Currently, the few-shot algorithm research focuses on applying transfer learning on deep neural networks that are pre-trained on big datasets. However, adapting the transformers requires highly cost computation resources. In addition, the overfitting or underfitting problems and low accuracy on large classes in the face validation domain are identified in our research. Thus, this paper proposed an alternative enhancement solution by adding contrasted attention to the negative face pairs and positive pairs to the training process. Extra attention is created through clustering-based face pair creation algorithms. The evaluation results show that the proposed approach sufficiently addressed the problems without requiring high-cost resources.

Keywords: few-shot machine learning, Siamese neural network, face validation, artificial intelligence

1. Introduction

Face validation is one of the important machine learning research topics for a wide range of smart applications. In the last decade, the development of convolutional neural network (CNN) architectures such as VGG-19 (Simonyan & Zisserman, 2014), ResNet (He et al., 2015), and Inception 3 (Boonyuen et al., 2019) provided good performances on face validation (Gwyn et al., 2021). However, researchers in this domain realize there is a crucial difference between deep CNN and human learning on face validation and other similar artificial intelligence (AI) which is the usage of data volume. Humans can grasp visual concepts from just a few image examples, whereas deep CNNs require extensive datasets to extract features and yet may make mistakes when encountering new images. In the meantime, many kinds of research focused on one or few-shot learning algorithms since 2006 (Chen & He, 2020; Fei-Fei et al., 2006; Koch, 2015; Lake et al., 2011; Müuller et al., 2022; Ren et al., 2018). In this paper, we discussed two important issues from the current state-of-the-art Siamese neural network on face validation, which are overfitting or underfitting (for simplification, we use overfitting as the general term in this paper) and less accuracy on large classes. We propose an enhanced pairing algorithm to address the issues.

The rest of this paper is organized as:

Different types of Siamese neural networks are discussed in Section 2. Our proposed clustering-based attention enhancement algorithm is explained in Section 3. The evaluation and comparison are illustrated in Section 4. Finally, a conclusion and future work are drawn at the end.

*__Corresponding author:__ Hong Qing Yu, School of Computing and Engineering, University of Derby, UK. Email: h.yu@derby.ac.uk

2. Related Work

The Siamese neural network introduced by Fei-Fei et al. (2006) presents a learning structure that has two parallel neural networks. One network is to understand the same concept and the other is to understand the differences between different concepts. However, these two networks are sharing the weights of the features from the learning process on the dataset. Therefore, the dataset needs to be pre-processed into two types of set pairs: pairs of the same concept and pairs of different concepts. In the face validation domain, they are image pairs of the same person (positive pairs) and image pairs of different people (negative pairs). The core mathematics function behind the Siamese neural network is the contrastive loss function:

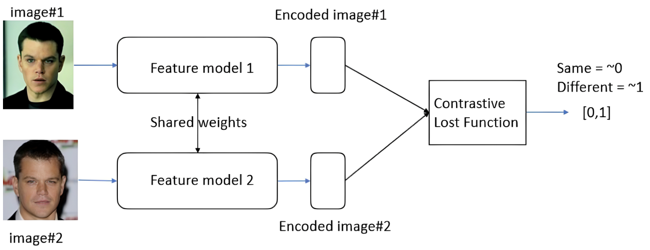
$$\text{Loss} = (1 - Y)\frac{1}{2}D_w^2 + Y\frac{1}{2}(\max(0, \alpha - \mathrm{D_w}))^2$$

Here, the Y can be 0 (same concept) or 1 (different concept) and D_w^2 presents the similarity measurement. The similarity measurement is based on the Bayesian likelihood function. In face validation, the facial features can be projected into a Euclidean space where distance calculations directly correspond to a measure of face similarity. Figure 1 shows the overall working process of the Siamese neural network (Roy et al., 2019). The feature models are normally created from artificial neural networks (ANNs). If it applies deep neural network (DNN), e.g., multi-layer CNN, then it can be defined as deep Siamese neural network (Taigman et al., 2014).

2.1. Deep Siamese neural network

The deep Siamese neural network applies a long sequence of convolution feature filters and each filter consists of feature

Figure 1
Siamese neural network



mapping, convolution activation function, and max-pooling (CNN process). The Siamese neural network will create a pair of deep CNN by joining them at the end with the loss function. The best deep Siamese neural network for image verification was claimed in Koch (2015), which contains seven layers of convolution filters.

In the meantime, DeepFace (Taigman et al., 2014) – a deep Siamese neural network – was proposed to do a human-level face validation task with a highly promised result of about 97% accuracy applying to the Labeled Faces in the Wild (LFW) image dataset. The DeepFace's DNN has two connected blocks of CNNs. The first block contains 32 filters of $11 \times 11 \times 3$ before the first max-pooling layer. The second one has 16 filters of $9 \times 9 \times 16$ followed by the second max-pooling layer that has three subsequent convolution filters. For training such deep neural networks, DeepFace still requires a big dataset to train before applying the Siamese function. Thus, the other pathway is to apply transfer learning (transformers) (Vaswani et al., 2017; Zhuang et al., 2021).

2.2. Transfer learning-based Siamese neural network

Transfer learning means adopting well-build facial feature extraction models (general model) that are trained on a big dataset but with tunes on the last layer using the application-specific (small) dataset. In this way, the general model can be well-trained regardless of cost consumption on computation resources and time.

One of the earliest transfer models was introduced in Cao et al. (2013), which transfers a developed joint Bayesian method learning model from other domains to perform face verification training on the LFW dataset. The accuracy of the work can achieve at 96.33%.

In the current state of the art, FaceNet (Schroff et al., 2015) is the most well-known transfer model in the face validation domain. The unique character of the FaceNet model is to extract face mapping features into a compact Euclidean space. As a result, the similarity of face images can be directly measured as Euclidean distances. Therefore, FaceNet has the suitable character to work with Siamese neural network as the transfer model.

2.3. Incrementation and simplification processes on image classification

Based on the transfer learning idea, there are two opposite directions of research on few-shot learning recently on image classification. In 2019, an incremental few-shot learning algorithm (Ren et al., 2018) is developed that separates the learning process into two phases: base class weight learning (base learning) and meta-learning. The base learning phase applies transfer learning to collect network weights for general classes on pre-trained and classified images. The meta-learning phase is to only extract feature weights through Siamese neural network on the novel images that the first phase never learnt before. As a result, attention weights are collected and only focused on the novel image. Finally, both weights are integrated through an attractor-regularizer gate to complete the classification task. This study effectively handles new datasets featuring distinct, predefined categories like dogs, cats, and fish. However, it struggles to distinguish between entities with closely similar characteristics, such as different types of fish. In contrast to adding an extra layer to Siamese neural network, the Facebook AI research team claimed a surprising exploring research outcome that a simple Siamese neural network (SimSiam) (Chen & He, 2020) can get enough meaningful features to do image classification of images even without having negative sample pairs, large batches, and momentum encoders. The core component that makes this happen is a one-side stop-gradient operation (see Figure 2 left). A combined and cloud-based clustering algorithm (SwAV) was also developed by having feature learning on the same images through two different image-augmented versions (see Figure 2 right) (Caron et al., 2020). The SwAV algorithm first encodes the class features into prototype vector *C* similar to the base class weights collection. Then the cloud online classification applies the swap prediction method to cluster the images into different classes.

However, there is a major difference between these proposed image classification algorithms and the face validation Siamese neural network, which is the contrastive network designed to predict/cluster the classes on the left side and validation on the right side for the same encoded image and face validation in Siamese neural network is to identify if the two different images are the same.

Figure 2
SimSiam and SwAV networks

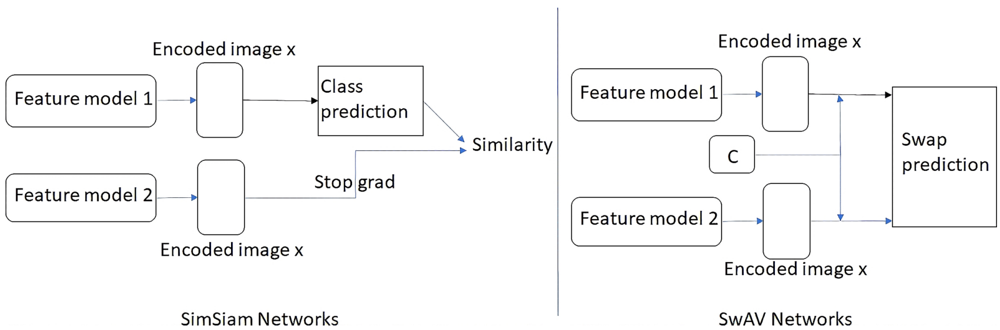


Figure 3
FaceNet Siamese neural network with the Yale face dataset overfitting analysis (x is the epoch, y is the rate of accuracy or loss)

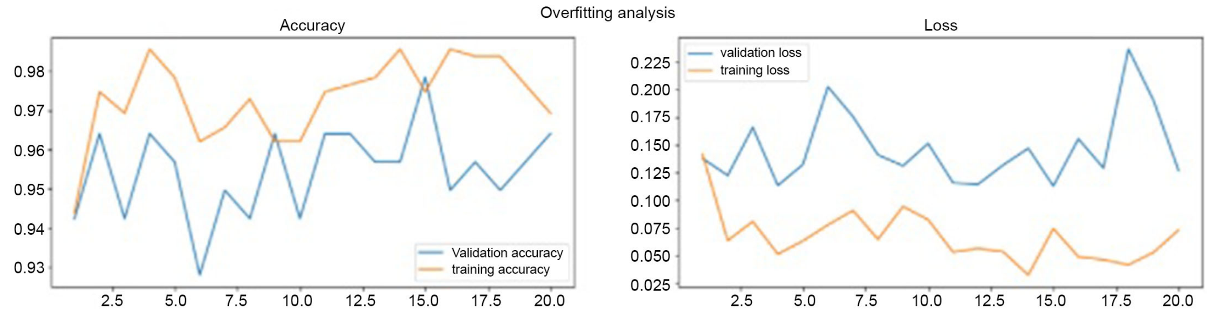
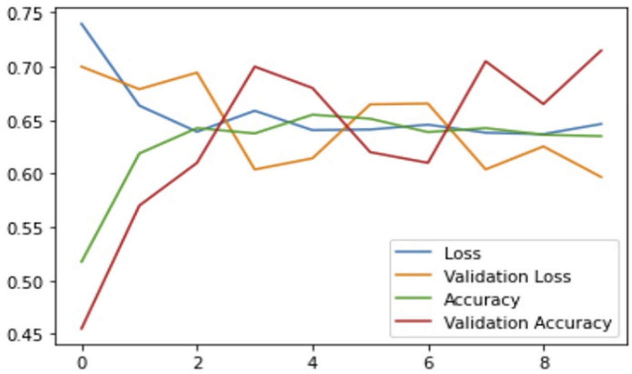


Figure 4
FaceNet Siamese neural network with LFW dataset accuracy analysis (x is the epoch, y is the rate of loss, validation loss, accuracy and validation accuracy)



2.4. Limitations

In general, the current Siamese neural network approaches suffer two problems:

- Significant overfitting problem for a smaller training dataset even with transformer. Applying the FaceNet Siamese neural network to the Yale face dataset (Belhumeur et al., 1997), we can clearly see an accuracy gap between training and validation as shown in Figure 3.

- Poor performance for a large training dataset without incremental computing (time cost) and significant pre-trained deep networks (very cost to transfer). Applying only the FaceNet Siamese neural network to the LFW dataset (Huang et al., 2008), the accuracy rate is dramatically dropped compared to the smaller dataset which is clearly displayed in Figure 4. In this context, the term "smaller dataset" means that there are fewer than 20 unique labels for face classification, and the dataset comprises just a few hundred images in total.

3. Proposed Clustering-Based Attention Siamese Neural Network

In this section, we introduce the clustering-based face validation Siamese neural network (CFVSiam). The hypothesis of CFVSiam is that clustered pairing algorithm can reduce significant numbers of image pairs for training but is more efficient and accurate because the networks are more sensitive to similar faces. More precisely, CFVSiam has three major steps of clustering (unsupervised learning) the few-shot face dataset based on the Kmean algorithm, creating negative pairs from the same cluster and different clusters (the positive pair will be created normally) and applying FaceNet-based deep Siamese neural network to encoding and contrastively compute the face validation. Figure 5 shows the architecture of CFVSiam.

3.1. Definition of CFVSiam

$$\text{Loss} = (1-Y)\tfrac{1}{2}D(f1,f2)_w^2 + Y\tfrac{1}{2}(\max(0,\alpha - D(f1,f2)_w))^2 \quad (1)$$

$$D(f1,f2) = \left\{[PP \times NPS \times NPD]\left|\sum \alpha_i^{pca} f1[i] - f2_2[i]\right|\right\} \quad (2)$$

where α_i is the PCA optimized trained parameters of contrastive twin FaceNet DNNs $f1$ and $f2$ over positive pairs (same person – *PP*), negative pairs from the same cluster (*NPS*), and negative pairs from different clusters (*NPD*). The reason that we can reduce the number of pairs is that the algorithm only takes one image from

Figure 5
CFVSiam networks

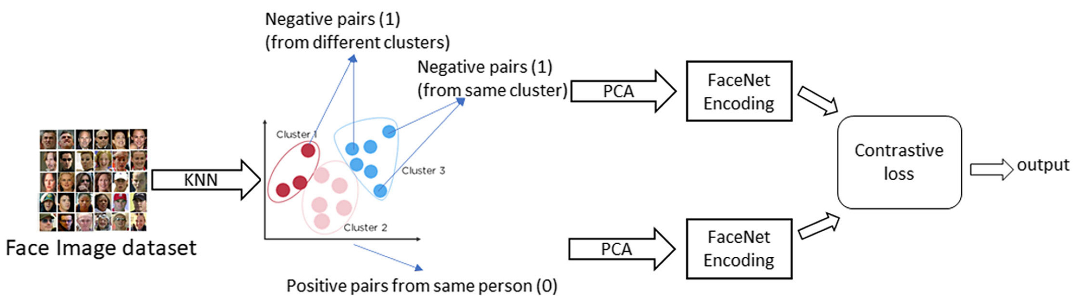
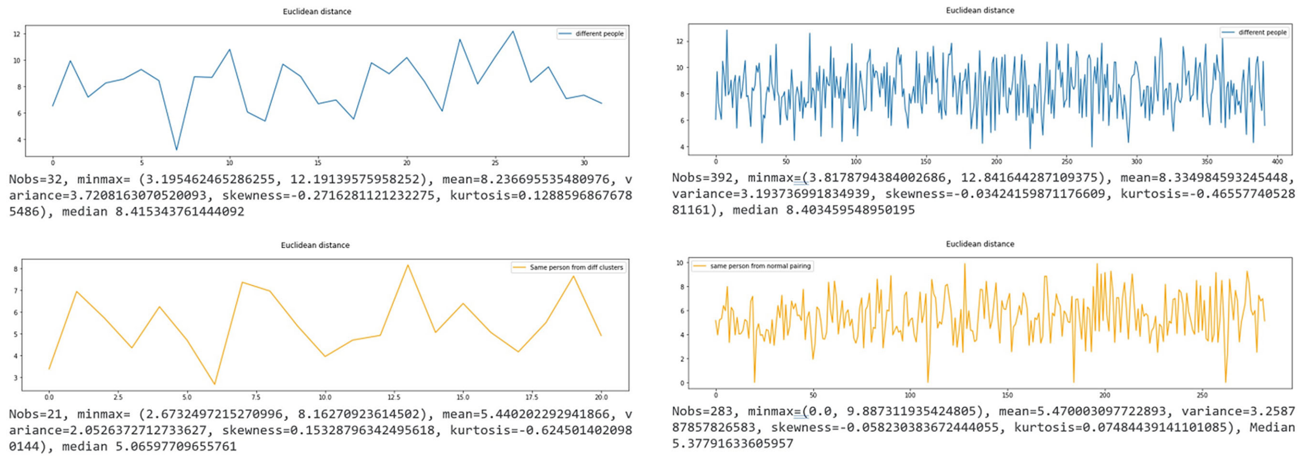


Figure 6
Face distance analysis



different clusters to create negative pairs but focuses on creating more negative pairs in the same clusters. Thus, the Siamese neural network is more sensitive and accurate. The idea behind this is that it is not difficult for Siamese neural networks to learn differences between face images from two different clusters but it is hard to get accurate feature learning in the same cluster.

3.2. Algorithm

The overall creation process of CFVSiam network has two algorithms that are presented in Algorithms 1 and 2 (see Appendices section).

Algorithm 1 presents the process of creating CFVSiam pairs of negative pairs and positive pairs. Attention is made to the negative pairs from the same clusters and positive pairs from the different pairs.

Algorithm 2 presents the creation process of the CFVSiam networks using a pre-trained FaceNet deep neural network. All the python implementations of these two algorithms are available on GitHub to review (https://github.com/semanticmachinelearning/AttentionSiameseNN).

4. Evaluations

The hypothesis is that the negative pair of people in the same cluster should have closer distances between them and the variance should be larger among them than in the random negative pairing generation process. Oppositely, the positive pair of the same person from different clusters should have longer distances and the variance should be smaller among them than in the random positive pairing generation process. Figure 6 shows

Figure 7
Different people from the same cluster Yale face dataset example

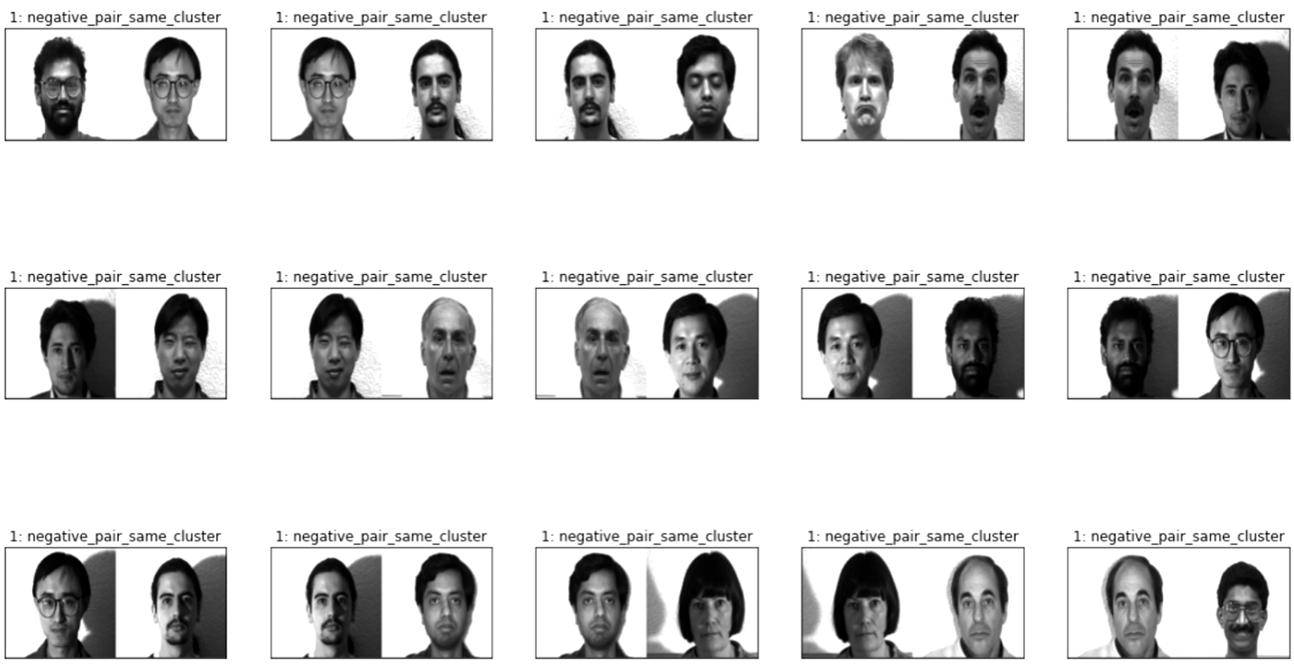
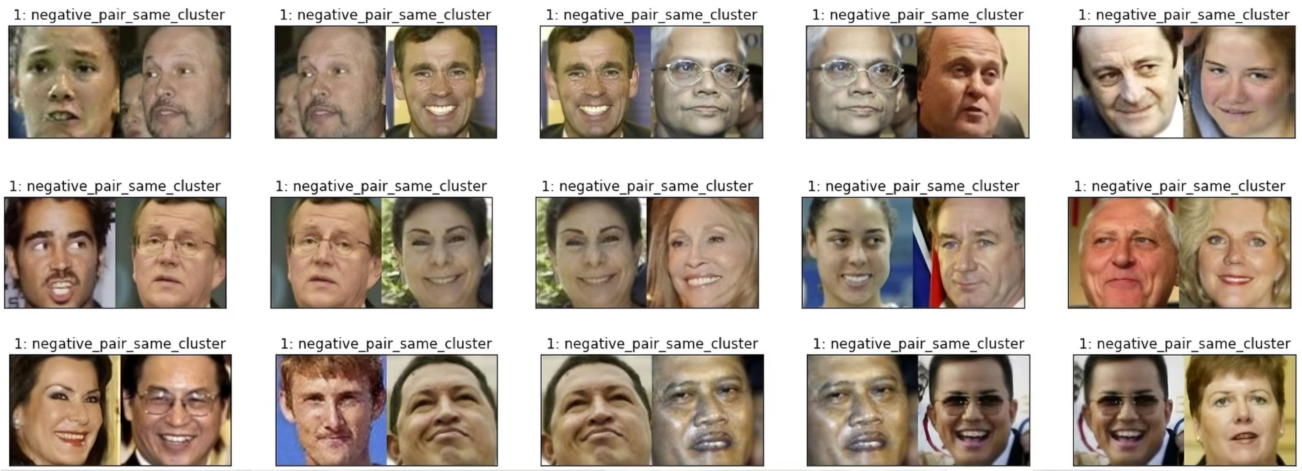


Figure 8
Different people from the same cluster LFW dataset example



that the assumption is correct if we take Yale face dataset as an example. On the top of the figure, the left figure presents different people paired in one of the clusters (mean = 8.237, variance = 3.721) compared to the right figure which includes all the negative pairs (mean = 8.335, variance = 3.194). Therefore, we believe that the negative pairs created through clustering make more attention to different people who have a certain degree of similarity and vice versa for the positive pairs.

Figures 7 and 8 demonstrate examples of negative pairs from the same clusters of Yale face dataset and LFW dataset. With KMeans clustering (sklearn-clustering-MiniBatchKMeans python package), the faces can be grouped with a certain level of similarity that will make the data pay more attention to extract the different features for different peoples' face images that may be difficult to separate from the same cluster and more common features for same person's face images that look very different from different clusters.

Figure 9 shows the accuracy and overfitting analysis of the proposed CFVSiam network. We can prove that the overfitting problem is dramatically addressed. Compared to Figure 3, CFVSiam's validation results are rarely worse than training results in all epos rounds for both datasets. In addition, the accuracy of the performances is both higher than without attention (see Figures 3 and 4) by training on Yale face

Figure 9
CFVSiam accuracy and overfitting analysis (x is the epoch, y is the rate of accuracy or loss for training and validation)

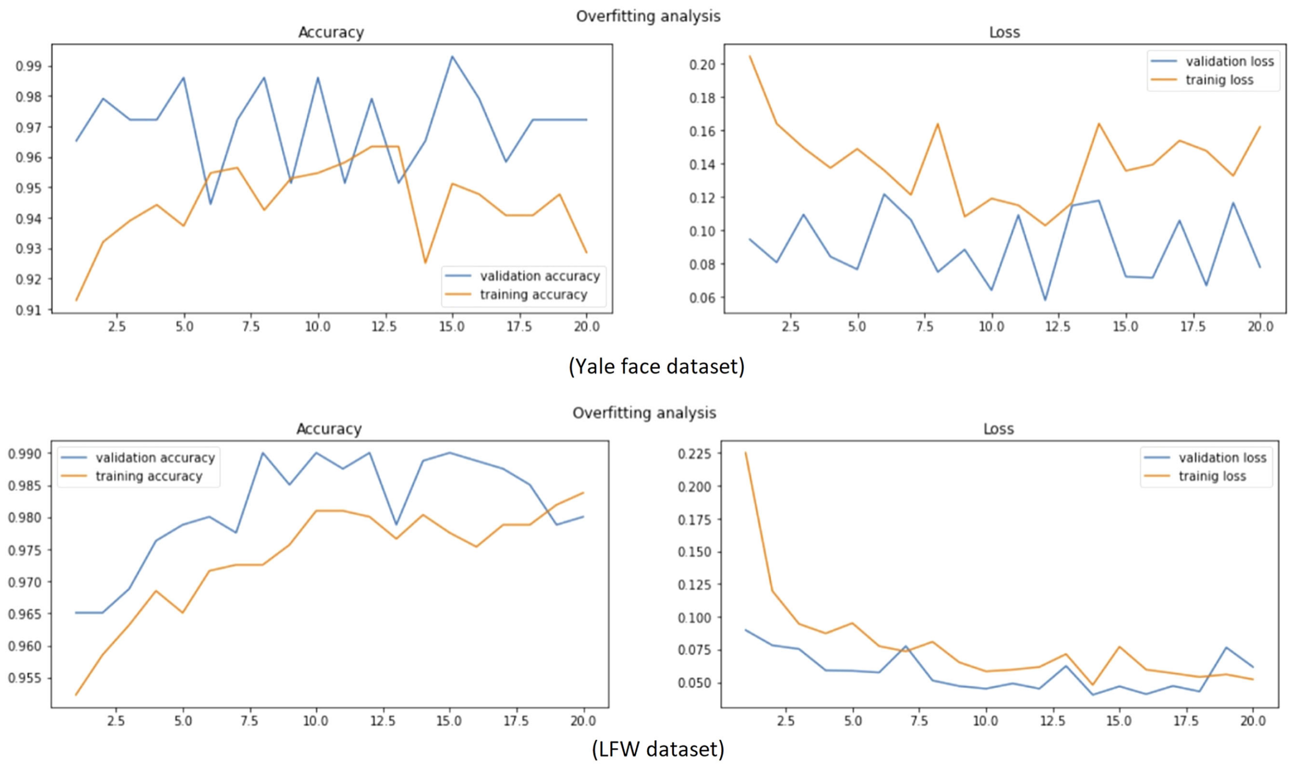


Table 1
Comparison DNN models on LFW dataset

Methods	Validation accuracy
FaceNet + VGGFace2 + CNN (Cao et al., 2017)	0.9965
Attention+Siamese (ours)	**0.9898**
PSI-CNN (Nam et al., 2018)	0.9887
Light CNN-9 (Wu et al., 2015)	0.9880
ResNet (He et al., 2015)	98.35
VGG16 + SVM (Chen & Haoyu, 2019)	0.9747
DeepFace (Taigman et al., 2014)	0.9735
Joint Bayesian (Chen et al., 2012)	0.9720
CNN + RBM (Cheng, 2019)	0.9252
VGGnet (Zhiqi, 2021)	0.921

dataset (train = 96, valid = 99) and LFW dataset (train = 98, valid = 99) with the clustering attentions. So, we can prove that CFVSiam's performance will not be affected by the size of the data (numbers of different people and color or black-white).

Table 1 presents a comparison of the results with those from other DNN models, all tested on the same LFW dataset. These models do not utilize few-shot learning techniques, as referenced in the survey literature (Chen & Haoyu, 2019; Swapna et al., 2020; van Dijk, 2019). The proposed attention enhanced Siamese model's performance is strong as same as the most state-of-the-art DNN models which request much more computation resources and time.

5. Conclusion and Future Work

Few-shot learning algorithms have the advantage of using fewer data examples from each class to address classification or prediction problems. This advantage will enable the algorithms to train faster with lower costs for resource-limited applications. However, we found that the Siamese neural network has problems with overfitting and low accuracy for big size of classes. To address these problems, we proposed CFVSiam network by adding a cluster attention mechanism to the pair data creation process. The evaluation results on two different datasets proved our hypothesis on the proposed enhancement in the face validation domain. Future research will focus on:

- applying CFVSiam network to the real-world applications for further evaluation.
- generalization of the clustering-based attention algorithm to other neural network and application domains.

Conflicts of Interest

Hong Qing Yu is an Associate Editor for Artificial Intelligence and Applications, and was not involved in the editorial review or the decision to publish this article. The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle: https://www.kaggle.com/datasets/olgabelitskaya/yale-face-database and https://www.kaggle.com/datasets/jessicali9530/lfw-dataset

References

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 711–720.

Boonyuen, K., Kaewprapha, P., Weesakul, U., & Srivihok, P. (2019). Convolutional neural network inception-v3: A machine learning approach for leveling short-range rainfall forecast model from satellite image. In *International Conference on Swarm Intelligence 2019: Advances in Swarm Intelligence*, 105–115.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2017). Vggface2: A dataset for recognising faces across pose and age. *arXiv Preprint: 1710.08092*.

Cao, X., Wipf, D., Wen, F., Duan, G., & Sun, J. (2013). A practical transfer learning algorithm for face verification. In *2013 IEEE International Conference on Computer Vision*, 3208–3215. https://doi.org/10.1109/ICCV.2013.398.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *arXiv Preprint:2006.09882*. https://arxiv.org/abs/2006.09882

Cheng, W., Sun, Y., Li, G., Jiang, J., & Liu, H. (2019). Jointly network: A network based on CNN and RBM for gesture recognition. *Neural Computing & Applications*, *31*(1), 309–323. https://doi.org/10.1007/s00521-018-3775-8.

Chen, D., Cao, X., Wang, L., Wen, F., & Sun, J. (2012). Bayesian face revisited: A joint formulation. In *12th Euorpean Conference on Computer Vision*, 566–579.

Chen, H., & Haoyu, C. (2019). Face recognition algorithm based on VGG network model and SVM. *Journal of Physics: Conference Series*, *1229*(1), 012015. https://doi.org/10.1088/1742-6596/1229/1/012015

Chen, X., & He, K. (2020). Exploring simple Siamese representation learning. *arXiv Preprint: 2011.10566*.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611. https://doi.org/10.1109/TPAMI.2006.79.

Gwyn, T., Roy, K., & Atay, M. (2021). Face recognition using popular deep net architectures: A brief comparative study. *Future Internet*, *13*(7), 164. https://doi.org/10.3390/fi13070164.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv Preprint: 1512.03385*.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

Jadon, S., & Jadon, A. (2020). An overview of deep learning architectures in few-shot learning domain. *arXiv Preprint:2008.06365*.

Koch, G. R. (2015). *Siamese neural networks for one-shot image recognition*. Master's Thesis, University of Toronto.

Lake, B. M., Salakhutdinov, R., Gross, J., & Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33), 2568–2573.

Müuller, T., Pérez-Torró, G., & Franco-Salvador, M. (2022). Few-shot learning with Siamese networks and label tuning. *arXiv Preprint: 2203.14655*.

Nam, G. P., Choi, H., Cho, J., & Kim, I.-J. (2018). PSI-CNN: A pyramid-based scale-invariant CNN architecture for face recognition robust to various image resolutions. *Applied Sciences*, *8*(9), 1561. https://doi.org/10.3390/app8091561.

Ren, M., Liao, R., Fetaya, E., & Zemel, R. S. (2018). Incremental few-shot learning with attention attractor networks. *arXiv Preprint: 1810.07218*.

Roy, S., Harandi, M., Nock, R., & Hartley, R. (2019). Siamese networks: The tale of two manifolds. In *2019 IEEE/CVF International Conference on Computer Vision*, 3046–3055. https://doi.org/10.1109/ICCV.2019.00314.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for largescale image recognition. *arXiv Preprint: 1409.1556*.

Swapna, M., Sharma, D. Y., & Prasad, B. (2020). A survey on face recognition using convolutional neural network. In *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19*, 649–661. https://doi.org/10.1007/978-981-15-1097-7_54.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. https://doi.org/10.1109/CVPR.2014.220.

van Dijk, J. (2019). *Face verification for poor resolution images*. Bachelor's Thesis, University of Groningen. https://fse.studenttheses.ub.rug.nl/20227/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems.*

Wu, X., He, R., Sun, Z., & Tan, T. (2015). A light CNN for deep face representation with noisy labels. *arXiv Preprint: 1511.02683*.

Zhiqi, Y. (2021). Face recognition based on improved VGGNET convolutional neural network. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference*, 2530–2533. https://doi.org/10.1109/IAEAC50856.2021.9390856.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., . . . , & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555.

How to Cite: Qing Yu, H. (2024). Attention Enhanced Siamese Neural Network for Face Validation. *Artificial Intelligence and Applications,* 2(1), 21–27. https://doi.org/10.47852/bonviewAIA32021018

Appendices

Algorithm 1. CFVSiam pairing algorithm

Input: A face dataset D$\{F_1, F_2, F_3 \ldots F_n\}$ and each set of F_i presents a person with multiple face image examples $\{f_1 f_2 f_3 \ldots f_n\}$, where f_i is the accessing location of the image. NPairs=[] holds negative face image pairs of different people and PPairs=[] holds positive face image pairs of same person.

Output: NPairs, PPairs

Output: $F_i\ ! = null$

Make a dictionary *FD* that the key presents the person id and a list of same person's images as the value

Put all images into a clustering image pool *ImageD* with identifiers of

$FD[key][i]$, where key i is the position in the value list.

while $n < size(FD)/2$ **do**

CLS = KmeanAlagorithm(*ImageD*), K = *n* and *CLS* is the clusters

$$\text{Score} = \sum \text{distance}^2$$

$\text{Score} = \sum \text{distance}^2$

Return the highest scored *CLS* and K

while $k > 0$ **and** $k < K$ **do** create negative pairs from the same cluster

for $f_{i,}\ f_j$ in $CLS(k)$ do

if $f_{i.}key \neq f_{j.}key$ **then**

NPairs.add(f_i,f_j)

while $k > 0k < K$ **do** #create positive pairs from the different cluster

for $f_{i,}\ f_j$ in $CLS(k)$,$CLS(k+1)$, where $k+1 < K$ do

if $f_{i.}\ key \neq f_{j.}\ key$ **then**

PPairs.add(f_i,f_j)

\# start normal pairing without considering clusters

while $key_i > 0\&key_i < size(key)$ **do**

while (**do**$key_j > 0\&key_j < size(key)$) do

if $key_j == key_i$ **then**

$n = size(FD[key_i]/2))$

for $f_x f_y$ in FD[key_i], $0 < x =< n, n < y <$size($FD[key_i]$) do

PPairs.add(*fi*,*fj)*

else

n = *Random*(0,*size*(FD[key_j]))

17

for f_x in $FD[key_i]$ do

NPairs.add(f_i,FD[key_j][n])

Algorithm 2. CFVSiam network generating algorithm

Input: NPairs, PPairs and FaceNet model

Output: Siamese face validation contrastive model

Output: $N\ Pairs! = null$ and $PPairs! = null$

NPV, PPV = Image.load(NPairs, PPairs).Vector()

NPVP, PPVP = PCA(NPV, PPV)

NPEncode = Facenet.encoding(NPVP)

PPEncode = Facenet.encoding(PPVP)

Splitting the NPEncode and PPEncode into training and testing datasets of NPEncode train, PPEncode train, NPEncode test, PPEncode test

Create ANN model of CFVSiam with Relu-based Euclidean distance Contrastive loss function and sigmoid-based activation function

model = CFVSiam.compile.fit(NPEncode train, PPTrainEncode train)

Return the model