RESEARCH ARTICLE

Attention Enhanced Siamese Neural Network for Face Validation

Hong Qing Yu¹,*
1 School of Computing and Engineering, University of Derby, United Kingdom

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 intelligent

*Corresponding author: Hong Qing Yu, School of Computing and Engineering, University of Derby, United Kingdom. Email: h.yu@derby.ac.uk

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 Convolution Neural Network (CNN) architectures such as VGG-19 [1], ResNet [2] and Inception 3 [3] provided good performances on face validation [4]. However, researchers in this domain realize there is a crucial difference between deep CNN and human learning on face validation and other similar Artificial Intelligent (AI) which is the usage of data volume. A human can learn image concepts from a very small size of examples but deep CNN needs a huge dataset to capture the features but still can make mistakes on novel images [5]. In the meantime, many kinds of research focused on one or few-shot learning algorithms since 2006 [6, 7, 8, 9, 10, 11]. In this paper, we discussed two important issues from the current state-of-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 the paper is organised as:

Different types of Siamese neural networks will be discussed in section 2. Our proposed clustering-based attention enhancement algorithm will be explained in section 3. The evaluation and comparison will be illustrated in section 4. Finally, we will draw a conclusion and future work at the end.

2. Related Work
The Siamese neural network was introduced by [6] 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:

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

Here, the $Y$ can be 0 (same concept) or 1 (different concept) and $D_w$ 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 [12]. The feature models are normally created from Artificial Neural Networks (ANN). If it applies Deep Neural Network (DNN) e.g. multi-layer CNN, then it can be defined as Deep Siamese neural network [13].

2.1. Deep Siamese neural network

The Deep Siamese neural network applies a long sequence of convolution feature filters and each filter is consist of feature 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 [9] in 2015, which contains seven layers of convolution filters.

In the meantime, DeepFace [13]-A Deep Siamese neural network was pro-posed to do a human-level face validation task with a highly promised result of about 97% accuracy applying to the LFW image dataset. The Deepface’s DNN has two connected blocks of CNNs. The first block contains 32 filters of 11×11×3 before the first max-pooling layer. The second one has 16 filters of 9×9×16 followed by the second max-pooling layer that has three subsequent convolution filters. For training such deep neural networks, DeepFace still re-quires a big dataset to train before applying the Siamese function. Thus, the other pathway is to apply transfer learning (Transformers) [14, 15].

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 [16], which transfers a developed Joint Bayesian Method (JBM)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 current the state of the art, Facenet [17] 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 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 [7] is developed that separates the learning process into two phases: base class weight learning (base learning) and meta-learning. The base learning phase is apply 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 work on the novel images that the first phase never learnt before. As the result, attention weights are collected only focused on the novel image. Finally, both weights are integrated through an attractor-regulariser gate to complete the classification task. This work works well on the novel dataset with predefined classes such as dog, cat and fish. For much more similar feature comparing (e.g. different types of fish), it does not work well but this is not the aim of their research anyway. 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) [10] 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) [18]. 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.

![Figure 2: SimSiam and SwAV networks](image)

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 3: Facenet Siamese neural network with the Yale Face Dataset overfitting analysis (x is the epoch, y is the rate of accuracy or loss)](image)
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 [19], 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 Labelled Faces in the Wild Dataset (FLW) [20], the accuracy rate is dramatically dropped compared to the smaller dataset which is clearly displayed in Figure 4. Here the smaller dataset refers to the number of distinct face classification labels is smaller than 20 and the dataset only contains a few hundred of images all together.

![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)](image)

![Figure 5: CFVSiam networks](image)

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

\[
    \text{Loss} = (1 - Y) \frac{1}{2} D(f_1, f_2) + Y \frac{1}{2} (\max(0, \alpha - D(f_1, f_2)))^2
\]  

(1)
Artificial Intelligence and Applications  Vol. XX Iss. XX yyyy

Where,
\[ D(f_1, f_2) = \{(PP \times NPS \times NPD) \mid \sum_{i} \alpha_i^{f_1} \mid f_1[i] - f_2[i] \} \]  

\( \alpha_i \) is the PCA optimised trained parameters of contrastive twin FaceNet DNNs \( f_1 \) and \( f_2 \) 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 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 Algorithm 1 and Algorithm 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 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.

Figure 7 and Figure 8 demonstrate examples of negative pairs from the same clusters of Yale face dataset and LFW dataset. With KMeans cluster-ing (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 7: Different people from the same cluster Yale face dataset example

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 epochs rounds for both datasets. In addition, the accuracy of the performances is both higher than without attention (see Figure 3 and Figure 4) by training on Yale face dataset (train=96, valid=99) and LFW dataset (train=98, valid=99) with the clustering attentions. The, 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 shows the comparing results to the other DNN models tested on the same LFW dataset which are not few-shot machine learning from the survey literature [21, 22, 23]. The proposed attention enhanced Siamese model’s performance is strong as same as the most state-of-art DNN models which request much more computation resources and time.

Figure 8: Different people from the same cluster LFW dataset example
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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Validation Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaceNet+VGGFace2+CNN [24]</td>
<td>0.9965</td>
</tr>
<tr>
<td><strong>Attention+ Siamese (ours)</strong></td>
<td><strong>0.9898</strong></td>
</tr>
<tr>
<td>PSL-CNN [25]</td>
<td>0.9887</td>
</tr>
<tr>
<td>Light CNN-9 [26]</td>
<td>0.9880</td>
</tr>
<tr>
<td>ResNet [27]</td>
<td>98.35</td>
</tr>
<tr>
<td>VGG16+SVM [21]</td>
<td>0.9747</td>
</tr>
<tr>
<td>Deepface [13]</td>
<td>0.9735</td>
</tr>
<tr>
<td>Joint Bayesian [28]</td>
<td>0.9720</td>
</tr>
<tr>
<td>CNN+RBM [29]</td>
<td>0.9252</td>
</tr>
<tr>
<td>VGGnet [30]</td>
<td>0.921</td>
</tr>
</tbody>
</table>
5. Conclusion and Future Work

Table 1 shows the comparing results to the other DNN models tested on the same LFW dataset which are not few-shot machine learning from the survey literature [21, 22, 23]. The proposed attention enhanced Siamese model’s performance is strong as same as the most state-of-art DNN models which request much more computation resources and time.

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 on 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

The author declare that he has no conflicts of interest to this work.

References


Appendices

Algorithm 1 CFVSiam Pairing Algorithm

Input: A face dataset $D\{F_1, F_2, F_3 ... F_n\}$ and each set of $F_i$ presents a person with multiple face image examples $\{f_1, f_2, f_3 ... f_n\}$, 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: $F_i ! = \text{null}$

1: Make a dictionary $FD$ that the key presents the person id and a list of same person’s images as the value
2: Put all images into a clustering image pool $ImageD$ with identifiers of $FD[\text{key}][i]$, where key i is the position in the value list.
3: while $n < \text{size}(FD)/2$ do
4: CLS = KmeanAlgorithm($ImageD$), $K = n$ and CLS is the clusters
5: Score = \(\sum\text{distance}^2\)
6: Return the highest scored CLS and K
7: while $k > 0 \text{ and } k < K$ do create negative pairs from the same cluster
8: for $f_i, f_j$ in CLS(k) do
9: if $f_i . \text{key} \neq f_j . \text{key}$ then
10: NPairs.add($f_i, f_j$)
11: while $k > 0 \text{ and } k < K$ do #create positive pairs from the different cluster
12: for $f_i, f_j$ in CLS(k), CLS($k + 1$), where $k + 1 < K$ do
13: if $f_i . \text{key} \neq f_j . \text{key}$ then
14: PPairs.add($f_i, f_j$)
15: # start normal pairing without considering clusters
16: while $\text{key}_j > 0 \text{ and } \text{key}_i < \text{size(key)}$ do
17: while ( $\text{dokey}_j > 0 \text{ and } \text{key}_y < \text{size(key)}$ ) do
18: if $\text{key}_j == \text{key}_y$ then
19: $n = \text{size}(FD[\text{key}_j]/2))$
20: for $f_x, f_y$ in FD[\text{key}_j], $0 < x =< n, n < y < \text{size}(FD[\text{key}_y])$ do
21: PPairs.add($f_i, f_j$)
22: else
23: $n = \text{Random}(0, \text{size}(FD[\text{key}_j]))$
24: for $f_x$ in FD[\text{key}_j] do
25: NPairs.add($f_i, FD[\text{key}_j][n]$)
Algorithm 2 CFVSiam Network Generating Algorithm

\textbf{Input:} NPairs, PPairs and FaceNet model

\textbf{Output:} Siamese face validation contrastive model

\textbf{Output:} \( NPairs = null \) and \( PPairs = null \)

\begin{align*}
NPV, \ PPV &= \text{Image.load}(NPairs, \ PPairs).\text{Vector()} \\
2: \ NPVP, \ PPVP &= \text{PCA}(NPV, \ PPV) \\
4: \ NPEncode &= \text{Facenet.encoding}(NPVP) \\
4: \ PPEncode &= \text{Facenet.encoding}(PPVP) \\
6: \ \text{Create ANN model of CFVSiam with Relu-based Euclidean distance} \\
& \text{Contrastive loss function and sigmoid-based activation function} \\
8: \ \text{Return the model}
\end{align*}