Choosing the structure of convolutional neural networks for face recognition

Kabul Khudaybergenov
National University of Uzbekistan, kabul85@mail.ru

Follow this and additional works at: https://uzjournals.edu.uz/mns_nuu

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation
Available at: https://uzjournals.edu.uz/mns_nuu/vol3/iss1/2
CHOOSING THE STRUCTURE OF CONVOLUTIONAL NEURAL NETWORKS FOR FACE RECOGNITION

KHUDAYBERGENOV K. K.
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: kabal85@mail.ru

Abstract

Evaluating the number of hidden neurons and hidden layers necessary for solving of face recognition, pattern recognition and classification tasks is one of the key problems in artificial neural networks. In this note, we show that artificial neural network with a two hidden layer feed forward neural network with \( d \) inputs, \( d \) neurons in the first hidden layer, \( 2d + 2 \) neurons in the second hidden layer, \( k \) outputs and with a sigmoidal infinitely differentiable function can solve face recognition tasks. This result can be applied to design pattern recognition and classification models with optimal structure in the number of hidden neurons and hidden layers. In addition, we propose a new type of convolutional neural network, which is capable to extract most powerful features. The experimental results over well-known benchmark datasets shows that the convergence and the accuracy of the proposed model of artificial neural network is acceptable. Findings in this paper are experimentally analyzed on five different face datasets from machine learning repository.

**Keywords:** neural networks, face recognition, hidden layer, convolutional neural networks, deep learning.

**Mathematics Subject Classification (2010):** 97R40, 92B20.

1 Introduction

Face recognition is one of the most research direction in computer vision and image processing. Face recognition is to find regions in a given image (in each frame in real time video) that contain faces and make a prediction of the faces. It is used in almost all of the applications, from surveillance up to entertaining software in mobile applications. Demand for high accuracy and speed are exist and new algorithms and methods are still in development. Some of these algorithms have reached the precision comparable with a human ability of recognition [1-4].

Multilayer neural network is one of the most useful artificial neural network to estimate the functional structure in pattern recognition and classification tasks. Researchers of artificial neural networks have proposed and experimented various kinds of neural network models exploiting multi layered architectures, for instance, convolutional neural networks [5], cascade neural networks [6], deep neural networks [7, 8], graph neural networks [9], recurrent neural networks [10], fuzzy MLP [13, 27] and etc.

In recent years, face recognition has become one of the very popular areas of research. The precision of face identification algorithms and methods has achieved the
human ability of recognition. In general, face recognition methods can be divided into two classes: statistical methods and neural network based methods. Statistical methods for face recognition are based on feature extractors which is made by human and has some level of accuracy.

On the other hand, face recognition based on neural networks proved to be superior over statistical methods. Increasing the number of hidden layers, so called convolutional neural network, the new type of neural network is built. Learning deep neural networks usually requires extremely large learning sets to ensure convergence, as well as to avoid the over-fitting issues. In the case of a lack of training sets, there are also researchers trying to find the reliable solutions.

While solving applied problems, one can use various kinds of architectures of neural network with some appropriately chosen activation functions. If a constructed artificial neural network model is not capable to solve a specific problem or to give acceptable precision, they increase hidden units (neurons) in the layers and/or increase hidden layers on the network model. However, in applications it is necessary to define or indicate how many neurons and how many layers one should take in the neural network architecture. Of course, the larger the number of neurons and hidden layers, the larger the probability of the network to give more precise results. Unfortunately, practicality decreases with the increase of the number of neurons and the number of hidden layers in the neural network model.

In this way, the feed-forward neural networks with a one hidden layer are not always effective if the number of neurons in the hidden layer is prescribed without any base. We show that this idea is no longer true for neural networks with two hidden layers.

The contribution of the paper is as follows:

- We propose to choose a structure for feed-forward neural network in the number of layers and hidden neurons, which is enough to solve the face recognition task.

- We present a method to extract features with a cascaded classifier based on deep learning.

The rest of this paper is organized as follows. In Section 2, we give a brief review of some related works on neural networks. In Section 3, we introduce the detailed implementation of our model for neural network and CNN architecture. Section 4 is related for experiments setup explanation. In Section 5, we show results of experiments conducted on two public benchmark datasets, like the Labeled Face in the Wild dataset (LFW), to verify our algorithms. A conclusion is given in Section 6.

2 Background

To begin with, we give some definitions of neural networks, which plays an important role.
**Definition 1.** A univariate function $\sigma$ defined on $\mathbb{R}$ is called a sigmoidal activation function if the following conditions are met:

$$\lim_{x \to +\infty} \sigma(x) = 1, \quad \lim_{x \to -\infty} \sigma(x) = 0.$$ 

The sigmoidal activation function [14] is a special type of function, which plays a significant role in the research of artificial neural networks, especially in classification and pattern recognition problems.

In this current paper, we use the sigmoidal activation function, which is defined of the following form

$$\sigma(x) = \frac{1}{1 + e^{-\alpha x}}, \quad x \in \mathbb{R}.$$ (1)

where $\alpha \in \mathbb{R}$ is a constant value.

Generally, feed-forward neural networks [12] with a one single hidden layer has the following form

$$f(x) = \sum_{i=1}^{r} c_i \sigma(w^i \cdot x - \theta_i),$$

where the weights $w^i$ are vectors in $\mathbb{R}^d$, $\theta_i$ - threshold values and $c_i$ - the coefficients in real numbers and $\sigma$ - is a univariate function, which is called an activation function in neural networks.

A two hidden layer feed-forward neural network architecture is defined by adding of another hidden layer in neural network into one hidden layer. In general, artificial neural network with two hidden layer of feed-forward type, input $x = (x_1, \ldots, x_d)$, $r$ neurons in the first hidden layer, $s$ neurons in the second hidden layer and $y = (y_1, \ldots, y_k)$ output is defined as follows

$$F(x) = \sum_{p=1}^{s} d_p \sigma \left( \sum_{q=1}^{r} c_{pq} \sigma(w_{pq} \cdot x - \theta_{pq}) - \gamma_p^{(i)} \right).$$

Here $d_p = \left( d_p^{(s)} \right)$, $\lambda_p = \left( \gamma_p^{(s)} \right)$, $c_{pq}$, $\theta_p$ are real numbers, $w_{pq}$ is a real matrix and $\sigma$ is a fixed univariate function.

## 3 Main result

### 3.1 Neural network model

In the current study, we are going to show that there exists a neural network which is a fully connected neural network part of CNN and this neural network has two hidden layers with infinitely differentiable sigmoidal activation function $\sigma$, $d$ inputs, $d$ neurons in the first hidden layer, $2d + 2$ neurons in the second hidden layer, $k$ outputs and having the ability to solve face recognition tasks within a given arbitrary accuracy.

We use this structure of feed-forward neural network with 2 hidden layers as a
fully connected neural network part of the face recognition model which is shown in Figure 1.

Below we present a theorem for feed-forward neural network to choose a structure in the number of hidden layers and hidden neurons.

**Theorem 1.** [11] Let \( Q \) be a compact set in \( \mathbb{R}^d \). For any positive real numbers \( \alpha \) and \( \lambda \) there exists a \( C^\infty (\mathbb{R}) \), sigmoidal activation function \( \sigma : \mathbb{R} \to \mathbb{R} \) which is strictly increasing on \(( -\infty, \alpha )\), \( \lambda \)-strictly increasing on \([\alpha, +\infty)\), and satisfies the following properties: For any \( f \in C(\overline{Q}, \mathbb{R}^k) \) and \( \varepsilon > 0 \), and there exists \( k \)-dimensional real vectors \( d_p = \left( d_p^{(s)} \right) \), \( \lambda_p = \left( \gamma_p^{(s)} \right) \), \( c_{pq}, \theta_p \) and real matrix \( w_{pq} \) for which

\[
\left\| f(x) - \sum_{i=1}^{2d+2} \sum_{p=1}^{k} d_p^{(i)} \sigma \left( \sum_{q=1}^{d} c_{pq} \sigma (w_{pq} \cdot x - \theta_p) - \gamma_p^{(i)} \right) \right\| < \varepsilon
\]

for all \( x = (x_1, \ldots, x_d) \in Q \).

**Remark 3.1.** In the case of recognition, for an input \( x \in \mathbb{R}^d \), its output of the trained network can be computed as follows

\[
F_k(x) = \sum_{p=1}^{2d+2} d_p^{(i)} \sigma \left( \sum_{q=1}^{d} c_{pq} \sigma (w_{pq} \cdot x - \theta_p) - \gamma_p^{(i)} \right) \in \mathbb{R}^k.
\]

From feed-forward neural network theory, all the weights and other parameters \( d_p = \left( d_p^{(s)} \right), \lambda_p = \left( \gamma_p^{(s)} \right), c_{pq}, \theta_p, w_{pq} \) require adjustment via training and tuning using training data set. There exist many optimization algorithms for neural networks, such as gradient descent, newton, conjugate gradient, Levenberg-Marquardt and etc. We use the gradient descent algorithm to optimize the weights and free parameters of the network.

Since, the proposed structure of feed-forward neural network in the model has \( k \) outputs, such as \( f_1, \ldots, f_k \), and a training dataset, \( (x_l, y_l), \ (l = 1, \ldots, M) \), then the overall output of the network is as follows

\[
F(x) = \sum_{i=1}^{k} \sum_{p=1}^{2d+2} d_p^{(i)} \sigma \left( \sum_{q=1}^{d} c_{pq} \sigma (w_{pq} \cdot x - \theta_p) - \gamma_p^{(i)} \right).
\]

We use the learning error of the proposed model is defined as follows

\[
E = \frac{1}{2} \sum_{i=1}^{M} (F_i(x_p) - t_p)^2, \quad i = 1, \ldots, M.
\]

To train the network, the standard back-propagation method was performed using a batch learning.
3.2 Convolutional neural networks

Recent studies show that CNNs have demonstrated a high-performance result on face recognition applications. CNNs are widely used in applications of computer vision, speech recognition, natural language processing, and etc. CNN can efficiently solve image processing problems due to its unique characteristics, such as shared weights, local receptive fields and subsampling. In general, CNN model contain of such blocks, like an input layer, a convolution layer, a combining layer, a fully related layer, and an output layer. The input layer accepts the raw pixel values of the input image; the convolutional layer calculates the output of units that are connected to local domains in the input layer; the pooling layer performs the down-sampling operation along the spatial dimensions; the last part which is called fully connected layer, calculates the scores for each class; and the output layer provides the final results. Thereby, we can make a deep learning models by alternately stacking the convolutional layers and pooling layers to get full the CNN architecture. A simplistic CNN model is shown in Figure 1 below. In order to get the desired accuracy, it is necessary to find most useful features to represent human faces. The feature descriptors, like Local Binary Patterns (LBP), Local Phase Quantization (LPQ) and Binarized Statistical Image Features (BSIF) are not enough powerful, thus in practice, applying these methods we might miss some useful information. In the current note, we use of deep learning to extract the most useful features. The excellence of deep learning models has been shown in a number of recent papers [18, 19, 20]. Moreover, we show that information extracted from deep learning is enough to build face recognition models with feed-forward neural networks, thus we construct a cascade structure of CNN with neural network.

The proposed CNN framework is combined with feed-forward neural network, which its hidden layer size and hidden units’ size are dependent to CNN’s output size. The neural network part, which is fully connected network of the CNN can be called as a classifier part. In order to achieve precise verification, we design the CNN structure as follows.

Figure 1: A CNN model with a fully connected neural network
• detect and resize the face image to 224×224 dimension;
• to get the landmarks, we use method which is given in [15];
• 5 landmarks are chosen as the center of regions (left eye center, right eye center, nose apex, left and right corner of the mouth); Region patch dimension have 2 scales (40×40 pixels);
• In training process, all the patches in the same region of different persons are used to train the proposed model;
• In testing process, different regions in face will be identified by the corresponding network;

We design a convolutional neural network structure to extract the features as given next subsection. We claim that the proposed CNN model is much faster than others are.

3.3 Structure of CNN framework

As illustrated in Figure 2, we propose a distinct layer type of feed-forward which is combined to CNN. The configuration of the proposed model is formed in the following form:

1. Input layer of CNN is set to 224×224×3 (cropped face image size);
2. The second layer is 40×40 local patches (overall three-channel of RGB);
3. The third layer is a convolution layer; kernel size 5×5, stride size 2×2 and feature number is 64, followed by a max-pooling layer Max-Pool-1 with kernel size 3×3 and stride size 2;
4. The fourth layer is also a convolution layer; kernel size 3×3, stride size 2×2 and feature number 256, followed by a max-pooling layer Max-Pool-2 with kernel size 5×5. So the output of the Max-Pool-2 is a 256-dimension vector;
5. The last part of CNN is fully connected neural network which is constructed according to Theorem 1 in Section 3.1. The first hidden layer has 256 hidden neurons and the second hidden layer size equal to 514 (2×256+2) neurons;

We claim that the feed-forward neural network layer in face recognition can have a distinct structure. Namely, we show that artificial neural network with a two hidden layer neural network with \(d\) inputs, \(d\) neurons in the first hidden layer, \(2d+2\) neurons in the second hidden layer, outputs and with a sigmoidal and infinitely differentiable activation function can solve face recognition problems with a given arbitrary accuracy. Here \(d\) is the output size from CNN.
4 Performance evaluation

4.1 Databases for face recognition

The ORL Face Database: consists of 40 subjects, 10 different face images for each subject. Images are of the same size of $92 \times 112$ and vary (slightly) in terms of lighting conditions, facial expressions or facial details.

The LFW database: LFW stands for Labeled Faces in the Wild, a database of face photographs designed for studying the problem of unconstrained face recognition. The data set contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set.

The FERET database [16]: This database was collected in 15 sessions between 1993 and 1996. The database contains 1564 sets of face images for a total of 14126 images that includes 1199 individuals and 365 duplicate sets of face images. A duplicate set is a second set of face images of a person already in the database and was usually taken on a different day.

The SCface database [17]: This is a database of static images of human faces. Images were taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. Database contains 4160 static images (in visible and infrared spectrum) of 130 subjects. Images from different quality cameras mimic the real-world conditions and enable robust face recognition algorithms testing, emphasizing different law enforcement and surveillance use case scenarios.

The AR Face Database: consists of over 4000 color images of 126 subjects, each having 26 facial images taken in two different sessions separated by two weeks. Each session has 13 images with multiple variations in expression, illumination and occlusion (sun glasses and/or scarf). A subset of cropped faces (of size $165 \times 120$) of 50 male and 50 female subjects was used. Eight faces (including neutral expression, smile, angry and scream) of each subject were used in the experiment.

The YALE database: This database was created in the laboratory of Computational Vision and Control Centre of Yale University. The face dataset contains 165
face images of 15 individuals with different facial expressions.

4.2 Experiments setup

For classification with neural networks, the objective value of each sample is a label vector. That is, if there are \( c \) classes to be classified, then we set \( t_p = [0, \ldots, 0, 1, 0, \ldots, 0] \in R^k \), where 1 only appears in the \( i \)th position, and represents the \( p \)th pattern belongs to the \( i \)th class. In the case of recognition, we can judge the class that \( x \) belongs to by finding the position of the maximum component of its real output.

In the fully connected neural network part of proposed model, we used the sigmoidal activation function which is defined in (1).

All the results are the average of 10 runs with independent initialization. The performance evaluations were carried out in Python environment running on a multi core CPU (3.5 GHz) computer with 16 GB of RAM.

4.3 Results and discussion

We compare the accuracy of our proposed CNN structure on face identification task with other models on insufficient training datasets to evaluate the performance. Figure 3 shows the comparison results with different algorithms using different dataset sizes of training sample in ORL. In Figure 3, the x axis indicates the size of the training samples of dataset (image/subject), and the y axis indicates the accuracy of identification. It could be find out proposed CNN structure with a special constructed FNN works better than the established state-of-the-art. The methods that we compared are taken from \([21, 22, 23]\). We applied cross-validation for training on face datasets. On each epoch, we use half of the samples in classes of face dataset.

![Figure 3: Accuracy of models with different training sizes on ORL dataset.](https://uzjournals.edu.uz/mns_nuu/vol3/iss1/2)
with more than one training sample. The average accuracy of proposed model and SVM method is estimated to be 98.50% and 97.57%.

Meantime, we test our proposed model on LFW face dataset. The comparison results are given below in Table 1.

Table 1: Comparison results on LFW dataset with different classifiers.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFace-ensemble [24]</td>
<td>0.9724 ± 0.12</td>
</tr>
<tr>
<td>DeepID [25]</td>
<td>0.9735 ± 0.25</td>
</tr>
<tr>
<td>ConvNet-RBM [26]</td>
<td>0.9227 ± 0.20</td>
</tr>
<tr>
<td>Proposed CNN model</td>
<td>0.9910 ± 0.11</td>
</tr>
</tbody>
</table>

Table 2: Recognition rates results of the different datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>SVM</th>
<th>Proposed CNN model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORL Face</td>
<td>92.50</td>
<td>94.00</td>
<td>98.25</td>
</tr>
<tr>
<td>LFW</td>
<td>95.25</td>
<td>95.50</td>
<td>99.10</td>
</tr>
<tr>
<td>FERET</td>
<td>86.11</td>
<td>74.07</td>
<td>97.80</td>
</tr>
<tr>
<td>SCface</td>
<td>91.50</td>
<td>93.40</td>
<td>97.30</td>
</tr>
<tr>
<td>AR Face</td>
<td>89.20</td>
<td>79.50</td>
<td>98.10</td>
</tr>
<tr>
<td>YALE</td>
<td>90.45</td>
<td>91.70</td>
<td>95.20</td>
</tr>
</tbody>
</table>

From Table 2, we can see that the proposed CNN architecture always has the highest recognition rate than other methods. Therefore, we can claim that our proposed CNN structure is much more effective compared to other algorithms.

Figures 4-9 show training and testing accuracies for selected benchmark face datasets, and the proposed number of hidden neurons and hidden layers using criteria which is given in Section 3. The "Number of neurons" axis in Figures 4-9 depicts the number of the first hidden layer (d-neurons) and similarly, the number of the second hidden layer neurons is calculate according to the first layer neurons amount. Fully connected neural network constructed according to Theorem 3.1 in Section 3: input size is 256 (output from convolution layer), 1-hidden layer size is 256 neurons, 2-hidden layer size is 2*256+2=514 neurons and output size is equal to classification task. From the simulation results, it can be seen, that increasing the number of hidden neurons in the first hidden layer and in the second hidden layer more than d and 2d+2 respectively, we can see that there is no need for that.

From the experimental results over well-known benchmark face datasets, the convergence and the accuracy of the proposed model of neural network is acceptable and even better than we have expected.
Figure 4. The ORL face dataset.

Figure 5. The LWF dataset.

Figure 6. The FERET face dataset.

Figure 7. The SCFace face dataset.

Figure 8. The AR face dataset.

Figure 9. The YALE face dataset.
Conclusion

In this paper, we propose a new CNN cascade framework and specially constructed feedforward neural network inside CNN. This proposed model performs faster and extracts more powerful feature than the other algorithms. It shows high performance results in all face datasets which have been tested in experiments. This deep learning can extract more powerful features for face recognition tasks. Meantime, two hidden layer feedforward network, $d$ neurons in the first hidden layer, $2d + 2$ neurons in the second hidden layer is capable to solve classification tasks.

References


