Teacher-directed learning in view-independent face recognition with mixture of experts using overlapping eigenspaces

Reza Ebrahimpour a,c,* , Ehsanollah Kabir b , Mohammad Reza Yousefi a,c

a School of Cognitive Sciences, Institute for Studies in Theoretical Physics and Mathematics, Niavaran, Tehran, P.O. Box 19395-5746, Iran
b Department of Electrical Engineering, Tarbiat Modarres University, Tehran, P.O. Box 14115-143, Iran
c Department of Electrical Engineering, Shahid Rajaee University, Tehran, Iran

Received 5 March 2007; accepted 30 October 2007
Available online 22 November 2007

Abstract

A model for view-independent face recognition, based on Mixture of Experts, ME, is presented. In the basic form of ME the problem space is automatically divided into several subspaces for the experts, and the outputs of experts are combined by a gating network. In our proposed model, the ME is directed to adapt to a particular partitioning corresponding to predetermined views. To force an expert towards a particular partitioning corresponding to predetermined views, a new representation scheme, overlapping eigenspaces, is introduced, that provides each expert with an eigenspace computed from the faces in the corresponding neighboring views. Furthermore, we use teacher-directed learning, TDL, in a way that according to the pose of the input training sample, only the weights of the corresponding experts are updated. The experimental results support our claim that directing the experts to a predetermined partitioning of the face space improves the performance of the conventional ME for view-independent face recognition. Comparison with some of the most related methods indicates that the proposed model yields excellent recognition rate in view-independent face recognition.

© 2007 Elsevier Inc. All rights reserved.

Keywords: View-independent face recognition; Mixture of experts; Teacher-directed learning; Single-view eigenspaces; Global eigenspace; Overlapping eigenspaces

1. Introduction

In performing face recognition, the visual system shows a remarkable capacity to distinguish between significant and unimportant image changes, to learn from examples to recognize new views of faces, and to generalize from known to novel views. In this paper, we focus on one aspect of this problem, the ability to recognize faces from different viewing directions. The problem of view-independent recognition is difficult because the image of a face seen from a novel viewing direction can be different from all previously seen images of the same face.

There are different methods for handling pose variations in face recognition. These methods are divided into the following three major groups: (a) the invariant features methods, (b) the 3D model-based methods and (c) the multiview methods [10,52,56].

Invariant features methods attempt to extract features that do not change when faces are seen from novel views, such as geometric invariants [1,7,54]. A drawback of these methods is the unfeasibility of finding sufficient number of invariant features for reliable recognition. In addition, there are many informative features that are intrinsically view-dependent and are not used in these methods.

The 3D model-based methods focus on constructing a prototypical view (frontal view) from a 3D model which is extracted from the input image. A recent survey of approaches to 3D face recognition is provided in [5]. Such methods work well for small rotation angles, but they fail when the angle is large causing some important features to be invisible [56].
Most proposed methods are based on using a number of multiview samples. In multiview methods, an adequate number of different views of a face are used to deal with the pose problem [10]. An example is the work by Beymer [2], which models faces with templates from 15 views, sampling different poses from the viewing sphere. The recognizer consists of two main stages, a geometrical alignment stage where the input is registered with the model views and a correlation stage for matching.

Under the category of multiview methods, there are other works in which the attempt is made to propose representation schemes that are robust to changes in viewpoint. Of such methods, the most famous one is the single-view eigenspaces. The concept of single-view eigenspace was first introduced in [38], based on the Principal Component Analysis, PCA (originally proposed in [28,46] and popularized by Turk and Pentland [50]). They use the face images in five common poses to build five single-view eigenspaces. For a test face, the distance to each single-view eigenspace is calculated and the pose class with the minimum distance is recognized. The single-view eigenspaces has also been used in [42] with three projection spaces of frontal, half profile and profile. The alternative solution for view-independent recognition based on PCA technique is the global eigenspace. The global eigenspace is created from all face images in different poses. This method has been used in [36] to simultaneously perform object pose estimation and recognition.

There are two main strategies in combining classifiers: fusion and selection. In classifier fusion, it is supposed that each ensemble member is trained on the whole feature space, whereas in classifier selection, each member is assigned to learn a part of the feature space. This way, in the former strategy, the final decision is made considering the decisions of all members, while in the latter strategy, the final decision is made by aggregating the decisions of one or a few of experts [31].

Recently, combining classifiers based on the fusion of outputs of a set of different classifiers have been developed in the field of face recognition as a method of improving the recognition performance [57,48,8,49,30]. Ref. [48], proposes a face recognition committee machine, which assembles the outputs of various face recognition algorithms, eigenface, Fisherface, elastic graph matching, support vector machine and neural network, to obtain a unified decision with improved accuracy. In [49], a fusion of classifiers based on a single representation space is employed. Their combined classifier uses the generalization capabilities of both learning vector quantization, and radial basis function networks to build different classifiers from the training patterns. There are however some trainable fusion strategies such as decision templates [32] or the behavior knowledge space [55] method, proposed for person identity verification in [29,23].

Classifier selection has not attracted as much attention as classifier fusion. However, classifier selection is probably the better of the two strategies, if trained well [31]. One of the most popular methods of classifier selection is the mixture of experts, ME, originally proposed by Jacobs et al. [25]. The ME models the conditional probability density of the target output by mixing the outputs from a set of local experts, each of which separately derives a conditional probability density of the target output. The outputs of expert networks are combined by a gating network which is trained to select the expert(s) that is performing the best at solving the problem [20,22,6]. In the basic form of ME [25], the expert and gating networks are linear classifiers, however, for more complex classification tasks, the expert and gating networks could be of more complicated types. For instance, Ref. [13] proposes a face detection model, in which they use multi-layered perceptrons, MLPs [21,40,3] in forming the gating and expert networks to improve the face detection accuracy. In [19] authors use MLPs with one hidden layer as experts and an RBF network as the gating network, for designing compensators for intensity modulated radiation therapy. Refs. [51,20] use ME with MLP experts for medical diagnostic systems and ECG beats classification, respectively.

Utilizing combining classifiers for view-independent face recognition is one of the subjects that have been less studied. A classifier fusion model is proposed in [27] where the experts are different in their nature owing to different sources of information and architectures used. They use four experts, two of which explicitly generate virtual-view face images and exploit them for learning discriminative features, and the two others compute view-robust representations of face input images. As the purpose of their study is to examine the potential for improving the recognition accuracy by fusing different types of pose-invariant face experts, simple fixed fusion rules such as the product and sum are employed as the aggregating operator. Under the classifier selection category, for the first time, Ref. [11] proposed a model, based on ME, for view-independent face recognition. In their model, a global eigenspace is used as the representation layer for both gating and expert networks. Their model reveals promising results in recognizing faces in intermediate unseen views by combining the outputs of experts which are specialized in recognizing specific views of faces. In order to sustain the specialization of experts, Ref. [12] employs a single-view representation scheme for the ME experts. In this work, each expert has its own eigenspace computed from the faces of a single view. According to the input layer eigenspace, each expert obtains specialization in recognizing faces of a single view. This way, combining the outputs of two experts leads to an enhanced performance in recognizing faces in intermediate views; however, generalization to previously unseen views is not yet satisfying, since the representation of the single-view eigenspaces degrades as faces rotate farther from the views that form the eigenspaces.

In this paper, we propose a model for view-independent face recognition which is essentially based on the divide and conquer approach, according to which a complex computational task is solved by dividing it into a number of
The proposed model contains experts which have expertise over different areas of the face space and a gating network which finds the pose of the input face image and combines the experts’ outputs through multiplicative connections (Fig. 1). The gating network essentially learns to decide which expert to choose in identifying a given input image. The selection which is performed by the gating network is based on the pose of the input face image, so we assign a global eigenspace to it, enabling the gating network to well recognize the view of a face. Fig. 2 illustrates the reconstructions of the faces in different views which are projected into a global eigenspace. As shown in the reconstructed faces of Fig. 2, the global eigenspace provides a useful representation for the face view.

To force the experts to become adapted to predetermined views, two different strategies are considered. The first strategy involves the representation layer of each expert, such that the representation layer of each expert is computed from the faces from those views that the expert is specialized in. The details of the representation scheme used in our model is described in Section 2.1.

The second strategy focuses on the training process of experts and the way the training samples are presented to each expert. This way, we attempt to direct each expert towards learning those training samples that correspond to its area of expertise. The details of this way of learning, called teacher-directed learning, is described in Section 2.2.

2.1. Overlapping eigenspaces

Our proposed overlapping-eigenspaces method is a trade-off between the two previously known eigenspace methods for representation. We neither define a global eigenspace for the whole space (Fig. 3a) [35,36] nor use one particular view-based eigenspace for each pose (Fig. 3b) [43,35]; rather, we divide the face space into overlapping subspaces, each containing two neighboring views, and represent each subspace through a distinct eigenspace (Fig. 3c).

In our model, each eigenspace encodes a portion of the face space. Therefore, its representation is more accurate than the global representation through a single eigenspace. On the other hand, as each eigenspace encodes two neighboring views, it can represent faces of intermediate unseen views, whereas single-view representation performs poorly in this case.

2.2. Teacher-directed learning

In the teacher-directed learning method, teacher information is included in the training process [26]. We use TDL to specialize the experts in their corresponding views,
in a way that, according to the pose of the input training sample, only the weights of the corresponding expert(s) are updated.

TDL is a subgroup of supervised learning, in both of which the target is known. There is however a tiny difference between them, mainly arising from the existence of teacher in TDL. In TDL the teacher has a meta-knowledge over the problem solution and directs the learner towards the desired output during the training process, whereas in supervised learning, the input–output map is obtained just with knowing the target. In our work, the meta-knowledge according to which the teacher directs the experts, is the information that it has over the pose of each training sample.

In our model, we use MLPs for expert and gating networks. In the training phase, the weights of MLPs are learned using the error back-propagation, BP, algorithm [21,40,3]. To apply TDL to BP algorithm, updating of the weights in the learning process are controlled, such that only the weights of the expert(s) having the input training sample in its eigenspace are updated and the others are kept unchanged. For instance, for a training sample of frontal view (0°), only the weights of experts responsible for −45°/0° and 0°/+45° views are updated (Figs. 1b and 3).

Fig. 2. The global eigenspace reconstruction of faces in different views. The original and reconstructed images are shown in the upper and lower rows, respectively. All reconstructions are computed using the first 50 eigenvectors.

Fig. 3. (a) Global eigenspace: just one eigenspace is used to represent all views of faces. (b) Single-view eigenspaces: each view of face is represented through a distinct eigenspace. (c) Overlapping eigenspaces: the face space is divided into four overlapping areas and each two neighboring views are represented through a distinct eigenspace. The thick gray lines indicate the poses that are used to build each eigenspace.

Considering the learning rules of ME described in the Appendix, according to Eqs. (4)–(8), the value of \( h \) defines the extent to which the weights of an MLP are updated, such that for \( h = 0 \) there is no weight updating and for \( h = 1 \) the maximum of calculated weight updates are applied. Succinctly stated, \( h \) is an estimate of the posterior probability that expert \( i \) can generate the desired output. Therefore, the most straightforward way of directing experts in the training phase is to control the value of \( h \) for each training sample to define which expert to update and which to leave unchanged.
To implement the above procedure, we should have one or two expert(s) for each training samples, which have a nonzero $h$, so their weights are updated towards learning the corresponding pose. We use the teacher matrix, $T_M$, in which the element-by-element multiplication of the $j$th column on $h$, results in a new $h$ that is nonzero for the $j$th expert and zero for all other experts. For the training samples of $-90^\circ$, $-45^\circ$, $0^\circ$, $+45^\circ$ and $+90^\circ$, $j$ is assigned the value of 1, 2, 3, 4 and 5, respectively.

$$T_M = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}$$ (1)

3. Experimental results

Our experiments are divided into two parts. The first experiment investigates the performance of the proposed model in recognizing faces of intermediate unseen views. In this experiment, the performance of the proposed model is compared against those of MEs with global and single-view eigenspaces. Our second experiment is designed to examine the role of each expert in performing the recognition task. In this experiment, the network is to recognize unseen images of faces in similar views as the training samples. During this experiment, the way that each expert improves the performance of the proposed model is studied.

3.1. Experiment I

In our first experiment, the ME task was to recognize the intermediate unseen views of faces. We used a subset of the PIE dataset [45] which consists of 180 images of nine views from 20 people (Fig. 4). The nine views of each person were evenly spaced from $-90^\circ$ to $+90^\circ$, with $22.5^\circ$ steps, along the horizontal plane. The proposed model was trained on the views of angles $\{ \pm 90^\circ, \pm 45^\circ, 0^\circ \}$ and was tested on the intermediate views of angles $\{ \pm 67.5^\circ, \pm 22.5^\circ \}$ (Fig. 5a). As in the PIE dataset there is just one sample of each pose for each identity, we face the “small sample size” problem. This problem exists in those pattern recognition tasks, where the number of available samples is small compared to the dimensionality of the samples [33]. Many techniques have been developed to attack this problem [47]. We tried to solve it with the basic idea of synthesizing multiple new face images. We produced 14 new
images, 7 of which were made by changing the contrast (Fig. 5b, the upper row) and the other 7 were made by adding Gaussian blur to the original images (Fig. 5b, the lower row). The changes in contrast and brightness of images were made using the built-in Matlab function ‘imadjust’, in which the contrast limits were varied from 0.25 to 0.95. The blurred images were produced using the Gaussian filter of the built-in Matlab functions ‘fspecial’ and ‘gaussian’. A rotationally symmetric Gaussian lowpass filter with a total filter size of 12-by-12 pixels was used. The standard deviation of the filter was varied between 0.2 and 1.4, in steps of 0.2. In this way, we had 1500 images for the training and 1200 images for the test.

Using synthesized images to train the networks is reminiscent of a similar technique used in 3D model-based methods. There are however differences between our work and such methods. In 3D model-based methods a 3D morphable model is used to compute 3D face models from a number of input images of each subject in the training set. The 3D models are rendered under varying pose conditions to build a large set of synthetic images. These images are then used for training a face recognition module [53,4]. In our work, different from the 3D morphable models, we do not synthesize new views of faces, but to tackle the small sample size problem, we generate multiple new images from the existing views.

As shown in Fig. 3, we formed four overlapping eigenspaces. Therefore, the ME used in our model had four experts (Fig. 1b). To form an eigenspace, for instance for the neighboring views of +45° and +90°, 600 training images of these views were used and mapped to the PCA space [50] of dimension 50. For the gating network, which is to mediate between these experts, a global eigenspace, of dimension 50, was used.

To evaluate the performance of our proposed model, we trained and tested two MEs with global [11] and single view.
global and single-view representations. In all experiments of overlapping representation and 5 output nodes for the persons. The MLP of the gating network had 50 input random weights. The MLPs of all experts had 50 input responding models, each time trained with different initial row of Table 1 are the averages of 10 times testing the corresponding models, for different number of hidden neurons, for gating and expert networks, are listed. The results in each experiment, the gating network had a global eigenspace and four experts for the overlapping representation. In all experts for both single view and global representations, spaces, with and without TDL, respectively. We had five experts for both single view and global representations, and four experts for the overlapping representation. In all experiments, the gating network had a global eigenspace in its input layer. In Table 1, the recognition rates of different models, for different number of hidden neurons, for gating and expert networks, are listed. The results in each row of Table 1 are the averages of 10 times testing the corresponding models, each time trained with different initial random weights. The MLPs of all experts had 50 input nodes for PCA components and 20 output nodes for 20 persons. The MLP of the gating network had 50 input nodes and 4 output nodes corresponding to four experts of overlapping representation and 5 output nodes for the global and single-view representations. In all experiments $\eta_e$ and $\eta_g$ of models were considered 0.022 and 0.046, respectively (Eqs. (4)–(7)). As shown in Table 1, the ME using overlapping eigenspaces and teacher-directed learning, outperforms the others.

In the literature of combining classifiers, early stopping has been shown to be useful for generalization [42–44]. Here, we used the results obtained for a validation set as the criterion for stopping the training of networks. The validation set was 25% of the training set described above. In this scheme, when error in the validation set increases, the training is stopped and the network is considered to have the best generalization ability at that point.

To analyze the errors, we considered the best result obtained from overlapping eigenspaces with TDL, with the recognition rate of 94.5% which had 66 misrecognized cases. Examining the results, the errors occurring in the recognition of intermediate unseen views can be categorized into the following three groups:

• First, some of the testing views are more difficult for our network to recognize. Among the misrecognized faces, those with $\pm 67.5^\circ$ rotation seem to cause more mistakes, as the mentioned network had 23, 11, 12 and 20 errors in the recognition of faces with $-67.5^\circ$, $-22.5^\circ$, $+22.5^\circ$ and $+67.5^\circ$ of rotation, respectively. It seems that as the intermediate view deviates farther from the frontal to profile view, it results in a more different image, such that the $0^\circ/\pm 45^\circ$ overlapping eigenspace provides a more helpful representation for the network, than the $\pm 45^\circ/\pm 90^\circ$ overlapping eigenspace. This way, the intermediate views of $\pm 22.5^\circ$ are recognized with less mistakes. One solution might be to form smaller eigenspaces as the views become closer to the profile views.

• Second, the original unseen images were all correctly recognized, but 48 of 66 misrecognized faces were those synthesized by contrast changes and the other 18 by blur changes.

• And third, we found that some people are similar to each other such that the network mistakes one for the other. Table 2 shows the confusion matrix of the recognition results of the mentioned model. For instance, two of the most misrecognized faces belong to persons 11 and 16 (See Fig. 4). As shown in Table 2, the network mistakes 6 images of person 11 for person 16, and it also mistakes 5 images of person 16 for person 11. The same error occurs for persons 14 and 20.

The PIE dataset has a number of faces with rotations in the vertical plane. To examine the performance of the proposed model in recognizing such views, we tested our model on faces with up and down poses in $0^\circ$ and $\pm 67.5^\circ$ rotations, namely e07, e09, e25 and e31 faces [45]. For the 40 faces associated with the frontal view, 87.5% were correctly recognized and for other 40 faces associated with $\pm 67.5^\circ$ rotations, the performance degraded to 77.5%. Comparing with the case that we had no pose change in the vertical plane, the recognition rate has a 6.5% and 15.33% decrease for frontal and $\pm 67.5^\circ$ views, respectively. Note that this results are obtained in a situation that we had no training sample with rotations in the vertical plane. However, considering the better recognition rate of the up and down faces in the frontal view, one solution might be extending the overlapping eigenspaces to cover the vertical plane; that is if we have overlapping eigenspaces for faces in up and down poses of frontal, half profile and profile views, and also the corresponding experts, the model will be able to handle a variety of poses in both vertical and horizontal planes.

Finally, we would like to compare the performance of the proposed model with two of the most related works in the literature for view-independent face recognition, [35] and [27]. These models were implemented and tested on our dataset under the same condition as our experiment with the intermediate unseen views. The results are tabu-
lated in Table 3. Its is clearly shown that the recognition rate of the proposed model is higher than the two existing methods: the recognition rate of the proposed method is 6% higher than the fusion method of \[27\] and 12.14% higher than the view-based eigenspaces method of \[35\]. Note that the overlapping eigenspaces are the best among the eigenspace methods, but when TDL is applied to the weakest eigenspace method, the global eigenspace, it outperforms the overlapping eigenspaces, showing the robustness of the proposed learning method in performing view-independent face recognition. As would be expected, applying TDL to overlapping eigenspaces reveals the highest recognition rate of 91.58%.

### 3.2. Experiment II

In this experiment, we examine the role of specialization of experts in our model. In order to attain a better understanding of the functionality of the experts, we performed this experiment on recognizing unseen face images in similar views as the training samples. These images were synthesized by the technique described in Section 3.1. The experiment was carried out with 1500 images of \(-90^\circ, -45^\circ, 0^\circ, +45^\circ\) and \(+90^\circ\) views. The performance of each expert of the models in the previous experiment, with the best topologies, was observed.

Fig. 6 illustrates the performance of each expert of the MEs with global and single-view eigenspaces and of the proposed model, on the unseen face images of similar views to the training, averaged over 10 runs. The bars denote the average recognition rates of experts, broken down by pose class. The standard deviation is also shown. Note that in Fig. 6b and c the left most bar in each pose corresponds to \(0^\circ\) and \(45^\circ\) experts, respectively. Considering Fig. 6a for the ME with the global eigenspace, for any input image irrespective of its pose, the experts reveal almost the same recognition rate. In Fig. 6b for the ME with single-view eigenspaces, there is comparatively stronger expertise for the experts in their corresponding pose while for other poses the recognition rate, in its best case, falls below 74%. This kind of specialization is not very useful in recognizing the intermediate unseen views of faces. The use of overlapping eigenspaces solves this shortcoming by providing expertise on neighboring views. For instance, as shown in Fig. 6c, the \(-90^\circ/45^\circ\) expert reveals recognition rates of about 90% for both \(-90^\circ\) and \(-45^\circ\) poses. This way, faces of intermediate unseen views are properly represented and recognized by the corresponding expert.

### Table 3

Comparison between the proposed method and two of the most related works in the literature implemented and tested on our dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlapping eigenspaces with TDL</td>
<td>91.58</td>
</tr>
<tr>
<td>Fusion of pose-invariant face-identification experts [27]</td>
<td>85.63</td>
</tr>
<tr>
<td>Overlapping eigenspaces</td>
<td>81.04</td>
</tr>
<tr>
<td>Single-view eigenspaces [12]</td>
<td>80.51</td>
</tr>
<tr>
<td>View-based eigenspaces [35]</td>
<td>79.44</td>
</tr>
</tbody>
</table>

Off-diagonal values represent errors in recognition, whereas diagonal values, typed in bold, signify correct recognitions. The rows show the output classes while the columns correspond to the input faces.

### 4. Conclusion

A model for view-independent face recognition, based on ME, was presented. The basic idea was to specialize the experts on overlapping areas of the face pose space,
Fig. 6. Recognition rates, averaged over ten test runs, each time trained with different random initial weights, on unseen synthesized images of training views broken down by pose class. (a) The experts of ME with the global eigenspace are not biased to prefer one class of pose to another. (b) The experts of ME with single-view eigenspaces are specialized in recognizing faces of a single view, which does not make a noticeable improvement in recognizing faces of unseen intermediate views, whereas experts of the proposed model, (c), demonstrate expertise on neighboring views covering the intermediate views, so that they are able to recognize faces of intermediate unseen views.
such that each expert is enabled to perform a view-independent face recognition in its area of expertise. Two different strategies were employed for this purpose; first, overlapping eigenspaces were used to partition the representation layer of experts to neighboring views; and second, a teacher-directed learning technique was employed in the training process to specialize the experts in their corresponding partitions. Basically, TDL improves the specialization of experts by having knowledge on the pose of each training sample and allowing only the weights of the corresponding expert(s) to be updated. Experiments were carried out with a training set on the views of angles \{±90°, ±45°, 0°\} and a test set on the intermediate views of angles \{±67.5°, ±22.5°\}. Testing our proposed representation scheme, the recognition rate was strongly increased to 91.58% by applying a teacher-directed learning method. Comparison with other related methods in the literature also demonstrated the improved performance of the mixture of experts by overlapping eigenspaces and TDL methods in view-independent face recognition.

Appendix. Mixture of experts

In a revised version of mixture of experts model, to improve the performance of the expert networks, we use MLPs instead of linear networks or experts in Fig. 1a. The application of MLPs in the structure of expert networks calls for a revision in the learning algorithm. In order to match the gating and expert networks the learning algorithm is corrected by using an estimation of the posterior probability of the generation of the desired output by each expert. Using this new learning method, the MLP expert networks’ weights are updated on the basis of those estimations and this procedure is repeated for the training data set. It should be mentioned that we do not use the notation of [25] to formulize the learning rules of the modified ME, but we follow the one which is described in Appendix of [9], since its clear explanation of learning rules makes its extension easier for our purpose (The learning algorithm of the mixture structure with linear classifiers as experts is described in [25]).

Each expert is an MLP network with one hidden layer that computes an output \(O_e\) as a function of the input stimuli vector, \(x\), and a set of weights of hidden and output layers and a sigmoid activation function. We assume that each expert specializes in a different area of the input space. The gating network assigns a weight \(g_i\) to each of the experts’ outputs, \(O_e\). The gating network determines the \(g_i\) as a function of the input vector \(x\) and a set of parameters such as weights of its hidden and output layers and a sigmoid activation function. The \(g_i\) can be interpreted as estimates of the prior probability that expert \(i\) can generate the desired output \(y\). The gating network is composed of two layers: the first layer is an MLP network, and the second layer is a softmax nonlinear operator. Thus the gating network computes \(O_g\), which is the output of the MLP layer of the gating network, then applies the softmax function to get:

\[
g_i = \frac{\exp(O_{gi})}{\sum_{j=1}^{N} \exp(O_{gj})}, \quad i = 1, \ldots, N
\]

where \(N\) is the number of expert networks. So the \(g_i\) are nonnegative and sum to 1. The final mixed output of the entire network is

\[
O_T = \sum_{i=1}^{N} O_g i, \quad i = 1, \ldots, N
\]

The weights of MLPs are learned using the error back-propagation, BP, algorithm. For each expert \(i\) and the gating network, the weights are updated according to the following equations:

\[
\Delta w_y = \eta_e h_i (y - O_i) (O_i (1 - O_i)) O_{hi}^T
\]

\[
\Delta w_h = \eta_e h_i w_y^T (y - O_i) (O_i (1 - O_i)) O_{hi} (1 - O_{hi}) x_i
\]

\[
\Delta w_{yg} = \eta_g (h - g) (O_g (1 - O_g)) O_{hg}^T
\]

\[
\Delta w_{hg} = \eta_g w_{yg}^T (h - g) (O_g (1 - O_g)) O_{hg} (1 - O_{hg}) x_i
\]

where \(\eta_e\) and \(\eta_g\) are learning rates for the expert and the gating networks, respectively. \(w_h\) and \(w_y\) are the weights of output to hidden and hidden to output layer, respectively, for experts and \(w_{hg}\) and \(w_{yg}\) are the weights of input to hidden and hidden to output layer, respectively, for the gating network. \(O_{hi}\) and \(O_{hi}^T\) are the transpose of \(O_{hi}\) and \(O_{hg}\), the outputs of the hidden layer of expert and gating networks, respectively. \(h_i\) is an estimate of the posterior probability that expert \(i\) can generate the desired output \(y\):

\[
h_i = \frac{g_i \exp\left(-\frac{1}{2}(y - O_i)^T(y - O_i)\right)}{\sum g_j \exp\left(-\frac{1}{2}(y - O_j)^T(y - O_j)\right)}
\]

As pointed out by [9], in the network’s learning process, “the expert networks “compete” for each input pattern, while the gate network rewards the winner of each competition with stronger error feedback signals. Thus, over time, the gate partitions the input space in response to the expert’s performance”.

References


