International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2022, Vol. 6, No. 1, IJSCAI640

An Ensemble Learning Method of Adaptive Structural Deep Belief Network for AffectNet

Takumi Ichimura *, Shin Kamada *

Abstract

Deep Learning is a hierarchical network architecture to express complex abstractions of input patterns of images. A Deep Belief Network (DBN) that builds hierarchical structure of Restricted Boltzmann Machine (RBM) is a well known unsupervised learning method as one of deep learning methods. The adaptive structural learning method of RBM (Adaptive RBM) was developed to find a suitable network structure for the input data set by neuron generation / annihilation algorithm during training. The Adaptive DBN can construct to pile an appropriate number of RBMs up to realize higher classification task. In this paper, our developed model was applied to AffectNet as the facial image data set and showed the better performance of classification rate than the State-of-The-Art CNN models. However, the model outputs incorrect wrong emotion category for some test cases, because the output labels for data set were annotated by two or more human annotators. For the problem, this paper proposes an ensemble learning model of Adaptive DBN, where the ensemble model consists of a parent DBN and some child DBNs. KL divergence is a measure of similarity for the parent and the child to each case. The new neurons are generated at the child to improve the classification according to KL divergence. Moreover, the generated neuron at the child is transferred to the parent to integrate better knowledge. In this paper, the proposed method improved the classification accuracy from 87.4% to 92.5%.

Keywords: Deep Belief Network, Restricted Boltzmann Machine, Adaptive Structural Learning, KL divergence, Ensemble, AffectNet

1 Introduction

The current research in Deep Learning is more influence than we expected both on theoretical of artificial intelligence and the practical use to the learning of hierarchical features from various type of data such as numerical, image, text and audio. The State-of-the-Art deep learning models such as AlexNet [1], GoogLeNet [2], VGG [3] and ResNet [4] have produced grate faculties exceed a human ability of recognition. The models are subordinate to convolutional neural network (CNN) type. The model with the multiple layers trains various features of input data and their learned trait from it. The various complex features

* Prefectural University of Hiroshima, Hiroshima, Japan

of image can be expressed with high classification accuracy by stacking multiple layers hierarchically.

Separately from the CNN model, a Deep Belief Network (DBN) [5] is a popular learning method of deep learning that has high classification performance. The DBN is piling up the Restricted Boltzmann Machine (RBM) [6] as a generative stochastic artificial neural network. It was difficult for DBN to realize the higher classification than CNN model. The way to solve this problem is the automatically determined method of the network structure of DBN and RBM called as the adaptive structural learning method of DBN (Adaptive DBN) and Adaptive RBM. The method can find the optimal number of hidden neurons in RBM and hidden layers in DBN to search for the fitted data space during learning [7,8]. The method employs the neuron generation / annihilation and the layer generation algorithm [9]. Adaptive DBN shows higher classification accuracy than traditional CNN models for some image benchmark data [10] and the real world applications [11, 12].

Adaptive DBN was applied to the facial expression database AffectNet [13] in this paper. There are more than one million facial expressions from the websites and annotating eight kinds of emotion for their facial images manually according to the valence and the arousal. The developed model by Adaptive DBN can reach almost correctly classified for the training data set against the precision of the traditional CNN model. Especially, the accuracy for the specified emotion category was not high (e.g. about 78.4% for the 'anger'). An intensive investigation indicates that the cause of mis-classification is not over-fitting or over-training. Some conflict patterns in the overlapped categories of the test data are found and the cases are arisen by two or more annotator's labeling. The trained model for no cleansed data set may encounter the situation with inconsistency of input-output as shown in case of the use of collected raw data from the Internet. The effective method to solve the problem is the clustering the ambiguous data and then train the cases by multiple models as ensemble learning.

Although an ensemble learning method using parallel two or more DBN models is known to be effective to reach the higher classification, it will consume the wasting computation resources because the construction of DBNs will take huge iterative training with trial and error. Moreover, the learning of the different labeled data by two or more annotators will be required for the ensemble learning of multiple network models. In order to the additional information, the likelihood for the trained models to each input data in the AffectNet will be calculated by using the other additional information such as Kullback-Leibler (KL) divergence.

In this paper, we propose an ensemble Adaptive DBN learning method which consists of three phrases. First, Adaptive DBN model is automatically trained for the database. The model is called as a parent DBN. Second, if the category with not high classification accuracy is found, a new child DBN will be constructed and trained for the confusion patterns in mis-classified cases. Third, the calculated KL divergence is monitored to observe the difference among the parent DBN and the child DBNs [14, 15]. If the KL divergence is large, the generated neuron at the child DBN is transferred to the parent DBN and then the parameters such as weights and parameters are trained by the fine tuning method in [12]. In the experiment results, nine child models were constructed to train several confusion cases in the overlapped region of emotion categories for test data, then the classification accuracy of the original parent model was improved from 87.4% to 92.5% by embedding the child models' neurons into the parent one.

The remainder of this paper is organized as the following four sections. Section 2 explains the adaptive structural learning of RBM and DBN shortly. Section 3 explains the

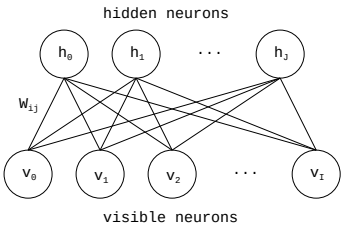


Figure 1: The overview of RBM structure

experiments for the facial database: AffectNet. Section 4 discussed the classification rate of the ensemble learning model of Adaptive DBN. Section 5 concludes this paper.

2 Adaptive Structural Learning Method of DBN

This section gives the comprehension of the fundamental behavior of self-organizing structure of Adaptive DBN briefly. An RBM is a probabilistic unsupervised model. Fig. 1 shows the network structure of RBM which consists of two types of binary layers. A DBN is widely used to build a deep architecture that stacking hierarchical several networks such as the trained RBMs.

2.1 Restricted Boltzmann Machine

An RBM is an unsupervised generative model in the architecture with two types of layers: $v \in \{0,1\}^I$ is visible neurons for input and $h \in \{0,1\}^J$ is hidden neurons for feature vector. An RBM has the three learning parameters $\theta = \{b,c,W\}$ where $b \in \mathbb{R}^I$ for visible neurons, $c \in \mathbb{R}^J$ for hidden neurons, and W_{ij} for weights between them, respectively. The important properties for RBM are the binary valued neurons and no connection among neurons in the same hidden layer. The properties allow each hidden neuron to train independent feature of the specified input space. An RBM aims to minimize the energy function $E(v,h)$ by Eq.(1)-Eq.(3).

$$E(v,h) = -\sum_i b_i v_i - \sum_j c_j h_j - \sum_i \sum_j v_i W_{ij} h_j, \tag{1}$$

$$p(v,h) = \frac{1}{Z}\exp(-E(v,h)), \tag{2}$$

$$Z = \sum_v \sum_h \exp(-E(v,h)), \tag{3}$$

where b_i and c_j are the parameters for visible neuron v_i and hidden neuron h_j, respectively. W_{ij} is a weight among v_i and h_j. The optimal parameters $\theta = \{b,c,W\}$ for the input are calculated by partial derivative of $p(v)$ using maximum likelihood estimation. Contrastive Divergence (CD-k) [16] is used to speed up the Gibbs sampling.

2.2 Deep Belief Network

A DBN [5] is a good generative model which is constructed by hierarchically stacking two or more the trained RBMs. The activated hidden neurons at the $(l-1)$-th RBM are set to the input signals at the l-th RBM by Eq. (4).

$$p(h_j^l = 1|h^{l-1}) = sigmoid(c_j^l + \sum_i W_{ij}^l h_i^{l-1}), \tag{4}$$

where c_j^l and W_{ij}^l are parameters at the l-th RBM. If the visible neuron is set, $h^0 = v$. After a DBN is trained for a classification task, the final output layer is added and the output probability y_k for a category k is calculated by Softmax in Eq. (5).

$$y_k = \frac{\exp(z_k)}{\sum_j^M \exp(z_j)}, \tag{5}$$

where z_j is an output value of the neuron j. M means the number of output neurons. The learning is conducted to minimize the error among the output y_k and the given ground-truth.

2.3 Neuron Generation/Annihilation Algorithm of RBM

Deep learning has been shown to be high classification power for the Big Data, however it is a difficult problem for AI system designers to determine the optimal network size and its appropriate parameters simultaneously during training. For the problem, the adaptive structural learning method in RBM (Adaptive RBM) and DBN (Adaptive DBN) [11] has been developed based on the neuron generation / annihilation algorithm. The method was developed to monitor WD as the variance of various parameters such as weights of BP learning during training. If the hidden neuron is in short supply and the network does not classify the input data sufficiently, the large fluctuation related to the WD will be occurred after a certain period of training. Such fluctuation is caused due to lack of the number of hidden neurons for representing an ambiguous input pattern. In order to represent the ambiguous patterns, the two neurons are required to split the patterns separately. A generated neuron inherits the attributes of the original neuron as shown in in Fig. 2(a).

For the neuron generation in an RBM, the inner product of the variance for two kinds of parameters is monitored, c and W except b in Eq.(6), because the oscillation for the variance of b was observed according to input space where the signals include some noise data.

$$dc_j \cdot dw_j > \theta_G, \tag{6}$$

$$dc_j = \gamma_c dc_j + (1-\gamma_c)(|c_j[\tau] - c_j[\tau-1]|), \tag{7}$$

$$dw_j = \gamma_W dw_j + (1-\gamma_W) Met(W_j[\tau], W_j[\tau-1]), \tag{8}$$

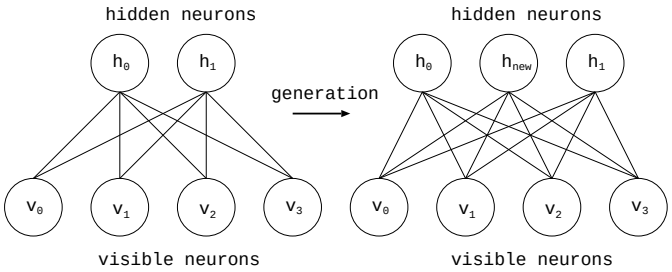
where $dc_j(\geq 0)$ and $dw_j(\geq 0)$ are the variances of the hidden neuron j. θ_G is a threshold with a small value. If θ_G is a smaller value, the neuron generation is easy to occur.

A new hidden neuron h_j^{new} will be generated by Eq.(6)-Eq.(8), if the condition is satisfied. The parameters c_j^{new} and W_{ij}^{new} are c and W for the generated neuron.

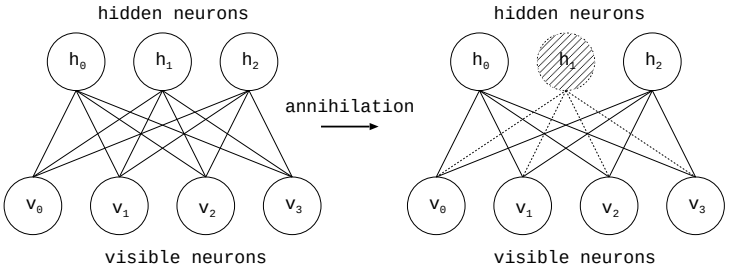
$$c_j^{new} = c_j + N(\mu, \sigma^2),\ W_{ij}^{new} = W_{ij} + N(\mu, \sigma^2), \tag{9}$$

where $N(\mu, \sigma^2)$ is a normal distribution with μ as the mean and σ as the standard deviation. Eq.(9) gives some oscillation, because the assign of new neuron leads to learn a slightly different feature by adding the noise. From the preliminary examination result, we define $\mu = 0$ and $\sigma = 0.1$ in this paper.

Moreover, the neuron annihilation algorithm is developed to remove the redundant neurons after the certain number of neurons are generated. In the other words, it is processed to kill the over increased neuron after generation. The annihilation condition for the neuron



(a) Neuron generation



(b) Neuron annihilation

Figure 2: Adaptive RBM

activities is observed by Eq.(10). If this equation is satisfied after generation, the weights of corresponding neuron are set to zero. Fig. 2(b) shows the annihilation process.

$$\frac{1}{N}\sum_{n=1}^{N} p(h_j = 1|v_n) < \theta_A, \tag{10}$$

where $p(h_j = 1|v_n)$ means a conditional probability of $h_j \in \{0,1\}$ for given input v_n. θ_A is a threshold in $[0,1]$. In case of the large θ_A, more annihilations have been occurred.

2.4 Layer Generation Algorithm of DBN

Generally, a deep layered DBN will be higher classification accuracy compared with a few layers DBN. The deep model is able to represent complex and various features from the abstract concept to the concrete images in the direction of input to output. However, a DBN with more RBMs requires a long time calculation. Thus the discover of DBN with an optimal number of RBMs will be demanded, where it depends on the input space during training.

Based on the neuron generation algorithm, we have developed a hierarchical model of Adaptive DBN which can automatically generate a new Adaptive RBM in the training process as shown in Fig. 3. Since a DBN is a stacking box of pre-trained RBMs, each RBMs' energy value E^l and WD at the l-th layer WD^l can be used as a criterion to the layer generation. If the criterion is larger than the pre-determined threshold, a new RBM layer will be generated to complement lack of representation for given input data [11].

$$\sum_{l=1}^{k} WD^l > \theta_{L1}, \tag{11}$$

$$\sum_{l=1}^{k} E^l > \theta_{L2}, \tag{12}$$

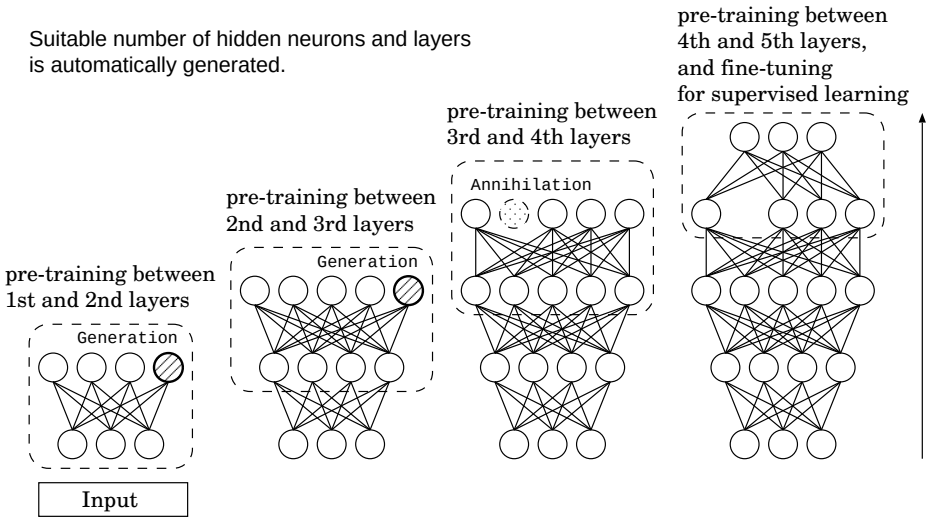


Figure 3: An overview of Adaptive DBN

Table 1: Emotion category in AffectNet

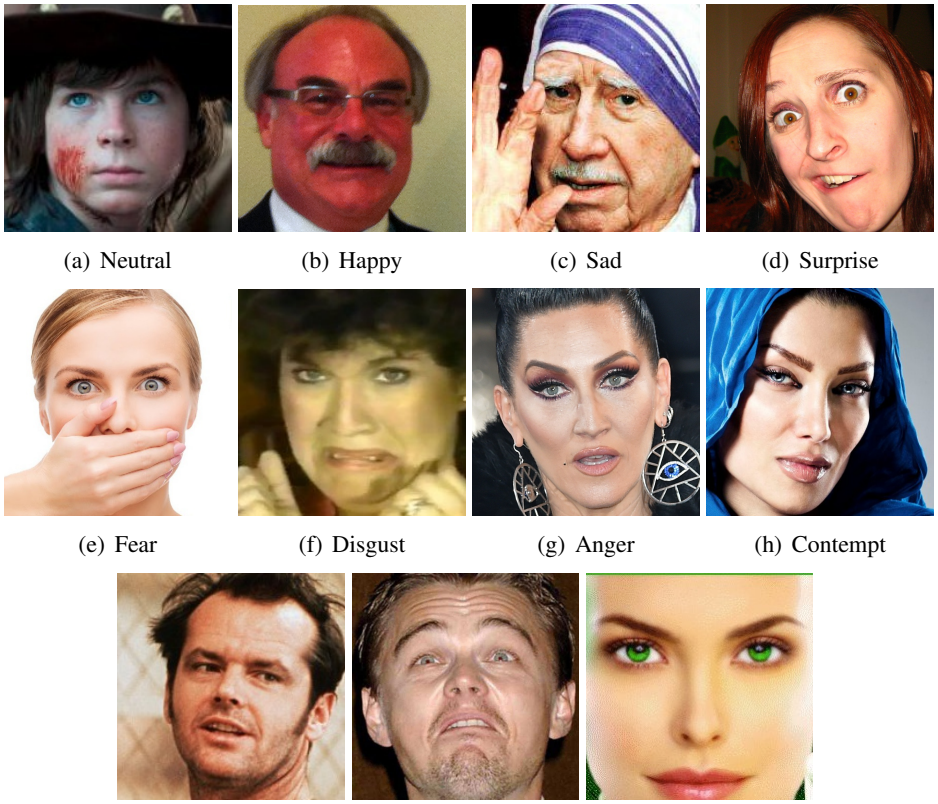
Category	Train Set	Test Set
Neutral	74,874	500
Happy	134,415	500
Sad	25,459	500
Surprise	14,090	500
Fear	6,378	500
Disgust	3,803	500
Anger	24,882	500
Contempt	3,750	500
None	33,088	500
Uncertain	11,645	500
Non-Face	82,414	500
Total	414,798	5,500

where $WD^l = \sum_{j=1}^{J}(WD^l_{c_j} \cdot WD^l_{W^j})$. $WD^l_{c_j}$ and $WD^l_{W^j}$ are WD for c_j and W_j in the l-th RBM, respectively. θ_{L1} and θ_{L2} are the threshold values.

3 Adaptive DBN for AffectNet

AffectNet is the large data set of facial image, valence, and arousal collected in intact from the Internet [13]. The valence and arousal mean the degree of pleasure/displeasure and the degree of emotional strength, respectively. A.Mollahosseini et al. collected 1,000,000 facial images, annotated the emotion category, and prepared for public distribution. As shown in Table 1, the database has eleven emotion categories as shown in Fig. 4 and includes training data set and test data set. The category is labeled by human annotators which includes his/her subjectively according to 'valence' and 'arousal' in Fig. 5. The valence and the arousal takes in $[-1,1]$, respectively. From the figure, we can see the overlap regions of each category for eight emotions.

An Adaptive DBN worked to train only for the eight categories in the training data set. The three categories 'None,' 'Uncertain,' and 'Non-Face' are excluded in the same condition of [13]. The classification result by Adaptive DBN for AffectNet as shown in Table 2



(a) Neutral (b) Happy (c) Sad (d) Surprise

(e) Fear (f) Disgust (g) Anger (h) Contempt

(i) None (j) Uncertain (k) Non-Face

Figure 4: Samples in AffectNet

was better performance than the AlexNet in [13]. Unfortunately, the result for training data set was not appearance in [13]. The classification accuracy by Adaptive DBN is 98.2% on an average for the training case. On the contrary, for the test data set, AlexNet and Adaptive DBN was 56.5% and 87.4% on an average. The category with the best performance was 'Happy'. On the contrary, the worst category was 'Anger'. The categories that did not reach 90% classification accuracy were 'Neutral,' 'Sad,' 'Surprise,' 'Anger,' and 'Contempt.' Further investigation was required for the relation between the facial image and the annotated label for the mis-classified categories.

The mis-classification for the test was not the cause of over-fitting or over-training while training. As a result of the investigation, conflict patterns were found in the data set due to the category subjectively assigned by multiple annotator's answer. In [13], there are some cases that two operators labeled the category for same facial expressions by his/her feelings, but their labels are not same. The agreement matrix between two categories labeled by the two operators is given as shown in Table 3. From the table, the agreement degree is not high. For example, the highest being 'Happy' is 79.6%, while that of 'Anger', 'Contempt,' and 'Neutral is 62.3%, 66.9%, and 50.8%, respectively. Moreover, the degree are moderate based on the Cohen's kappa coefficient [17] calculated from Table 3, but the data set includes two or more operators feelings.

Table 4 shows the confusion matrix for classification accuracy by Adaptive DBN for Table 2. The table element in the matrix is the matched number of the prediction label and the answer label for eight types. Some examples of the mis-classified cases are depicted as shown in Fig. 6. We can find various mis-classified cases from Table 4. For example, 'Disgust $\rightarrow$ Anger' indicates the disgust cases mis-classified into Anger.

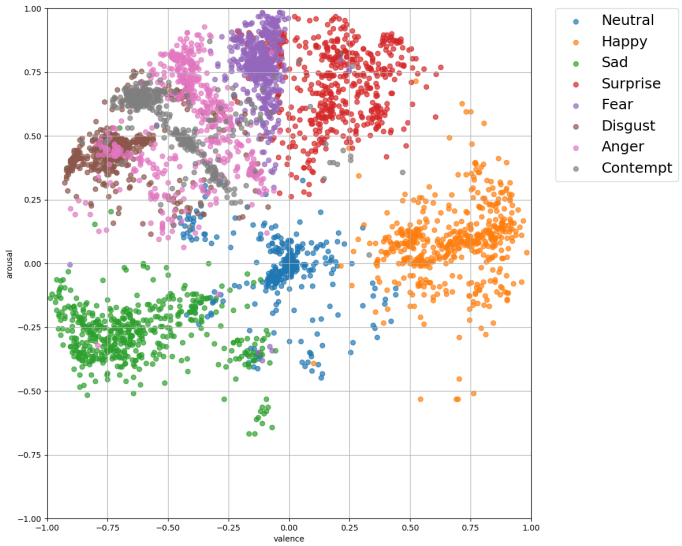


Figure 5: Valence and Arousal

Table 2: The classification results for AffectNet(%)

Category	CNN([13])	Adaptive DBN	
	Test data	Train data	Test data
Neutral	63.0	99.3	87.8
Happy	88.0	99.9	92.4
Sad	63.0	99.2	84.2
Surprise	61.0	99.4	85.8
Fear	52.0	99.5	90.4
Disgust	52.0	99.3	92.4
Anger	65.0	98.2	78.4
Contempt	8.0	98.8	87.6
Average	56.5	99.2	87.4

Since human emotion contains many vague features, the answer by the annotator will be different. For instance, the decision for labeling to the celebrities as shown in Fig. 6 may include the operator's preference to the images during the labeling process. The classification automatically for such cases requires an ensemble learning method of deep learning. An ensemble learning method using a parent DBN and two or more child DBNs will be useful to be higher classification accuracy. The proposed method in this paper can realize it without the wasting computation resources such the huge iterative training for the construction of multiple DBNs. The correction for the mis-labeled data needs the other further information to judge them. The proposed ensemble learning method described in Section 4 was worked to improve the classification capability.

(a) Disgust→Anger (b) Disgust→Anger (c) Disgust→Anger (d) Anger→Disgust (e) Anger→Disgust (f) Anger→Disgust

Figure 6: Examples of mis-classified categories

Table 3: The category of agreement between two annotators (%) [13]

	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	None	Uncertain	Non-Face
Neutral	50.8	7	9.1	2.8	1.1	1	4.8	5.3	11.1	1.9	5.1
Happy	6.3	79.6	0.6	1.7	0.3	0.4	0.5	3	4.6	1	2.2
Sad	11.8	0.9	69.7	1.2	3.4	1.3	4	0.3	3.5	1.2	2.6
Surprise	2	3.8	1.6	66.5	14	0.8	1.9	0.6	4.2	1.9	2.7
Fear	3.1	1.5	3.8	15.3	61.1	2.5	7.2	0	1.9	0.4	3.3
Disgust	1.5	0.8	3.6	1.2	3.5	67.6	13.1	1.7	2.7	2.3	2.1
Anger	8.1	1.2	7.5	1.7	2.9	4.4	62.3	1.3	5.5	1.9	3.3
Contempt	10.2	7.5	2.1	0.5	0.5	4.4	2.1	66.9	3.7	1.5	0.6
None	22.6	12	14.5	8	6	2.3	16.9	1.3	9.6	4.3	2.6
Uncertain	13.5	12.1	7.8	7.3	4	4.5	6.2	2.6	12.3	20.6	8.9
Non-Face	3.7	3.8	1.7	1.1	0.9	0.4	1.7	0.4	1.2	1.4	83.9

Table 4: The confusion matrix for classification results by Adaptive DBN (Num.)

		Predicted Category							
		Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
True Category	Neutral	439	2	7	5	8	16	4	19
	Happy	7	462	2	0	4	12	1	12
	Sad	12	3	421	13	11	20	5	15
	Surprise	15	4	10	429	11	22	0	9
	Fear	10	2	10	10	452	8	3	5
	Disgust	8	2	3	5	8	462	5	7
	Anger	14	4	8	10	9	47	392	16
	Contempt	17	8	6	3	2	21	5	438

4 KL Divergence Based Ensemble Learning Model

This section explains the KL divergence of the ensemble learning model of Adaptive DBN. Fig. 8 shows an overview of our proposed ensemble learning model. One of reasons of misclassification cases for test data was the annotation by two or more humans labeling. A new child DBN is trained only for the overlapped cases that the input patterns are same but the outputs are different as a child model on the basis of the trained model. The KL divergence between the trained network as a parent model and a child model was computed. For the KL divergence with a larger than the threshold value, the parent network requires some new neurons to classify the conflict cases [14, 15]. The neuron generation process was applied to mend the wrong path in the network by using fine-tuning method [12]. The ensemble method took a long calculation time because multiple DBNs work to infer the classification result to the case [14]. The ensemble method in [15] was realized by embedding the features of multiple child models to the parent model and overwrite the difference between their models. In this paper, the ensemble model works to be selected the model with the highest classification ratio as a winner and then to transfer the neurons and their paths from the child network to parent network.

4.1 KL Divergence of Adaptive DBNs

The KL divergence for the original parent DBN T and the child DBN S is calculated by Eq. (13).

$$D_{KL}(T,S) = \sum_i P_T(x_i) \log P_T(x_i)/P_S(x_i), \tag{13}$$

Table 5: Training cases for child network [14, 15]

data set	DBN	Description	number
$Set0$	T	All image cases for Anger and Disgust	1,000
$Set1$	$S1$	The image cases correctly classified at $Set0$	854
$Set2$	$S2$	The image cases wrong classified $Set0$	146

Table 6: KL divergence [14, 15]

Network	KL
$D_{KL}(T,S1)$	0.188
$D_{KL}(T,S2)$	0.660

where $P_T(x_i)$ and $P_S(x_i)$ are the probability distribution of output for the input x_i at the networks T and S, respectively.

In the prior paper [14,15], we tried to construct a child DBN for mis-classified cases and correct labeled cases for the largest overlapped region of two emotion categories: 'Anger' and 'Disgust'. Table 5 shows three kinds of data sets. For two kinds of facial images, we defined all image data set, the correctly classified cases, and wrong classified case as $Set0$, $Set1$, and $Set2$, respectively. The parent DBN, two child DBNs are T, $S1$ and $S2$ and the DBNs were trained for data sets $Set0$, $Set1$, and $Set2$.

Table 6 is the KL divergence of T, $S1$ and $S2$. The result shows that KL divergence is a certain difference and $D_{KL}(T,S2)$ is larger than $D_{KL}(T,S1)$.

The KL divergence between T and $S2$ was calculated to each case and its value takes in $[0.0000, 0.0025]$ as shown in Fig.7. From such an observation, the three cut-off values of KL divergence were set $\theta_{KL} = \{0.0010, 0.0015, 0.0020\}$ to investigate the retraining network for the sub data set that divided mis-classified cases by θ_{KL}. As a result, the classification accuracy for three kinds of child DBNs were 95.8%, 97.2%, and 95.2%, respectively. The child model for the mis-classified cases of test data divided by $\theta_{KL} = 0.0015$ was the best performance. From the observation on the distribution of KL divergence, we consider that a new neuron is required in the DBN for the cases more than a certain value θ_{KL} in case of mis-classified patterns. In this paper, we set $\theta_{KL} = 0.0015$.

There are nine categories where exist over than 15 confusion cases as shown in the Table 4. The nine categories are '1)Anger → Disgust,' '2)Surprise → Disgust,' '3)Contempt → Disgust,' '4)Sad → Disgust,' '5)Surprise → Neutral,' '6)Neutral → Contempt,' '7)Neutral → Disgust,' '8)Sad → Contempt,' and '9)Anger → Contempt.' In this paper, 'Contempt→Neutral' and 'Neutral→Contempt' belong to the same category. The mis-classified cases for the training were 146, 113, 111, 108, 93, 92, 88, 85, and 79, respectively such as $Set2$ in Table 5 in [14, 15]. For such cases, the ensemble Adaptive DBN learning method was applied to be the better accuracy of the classification in the subsection 4.2.

4.2 Knowledge Transferred from Child to Parent

This section explains the transfer of knowledge related to the neuron and its paths from the child DBN to the parent DBN. A new child network is built for only mis-classified cases. The parent network itself is improved due to the difference in network structure between the parent network and the child network. The parent network is improved by additionally embedding the neurons generated in the child network into itself so that an improved

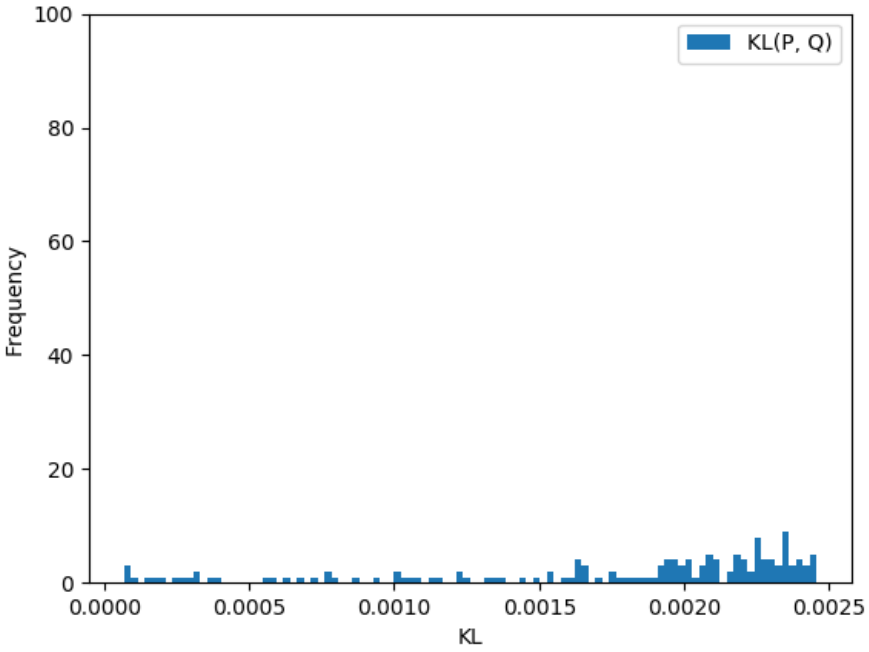


Figure 7: The histogram of $D_{KL}(T, S2)$

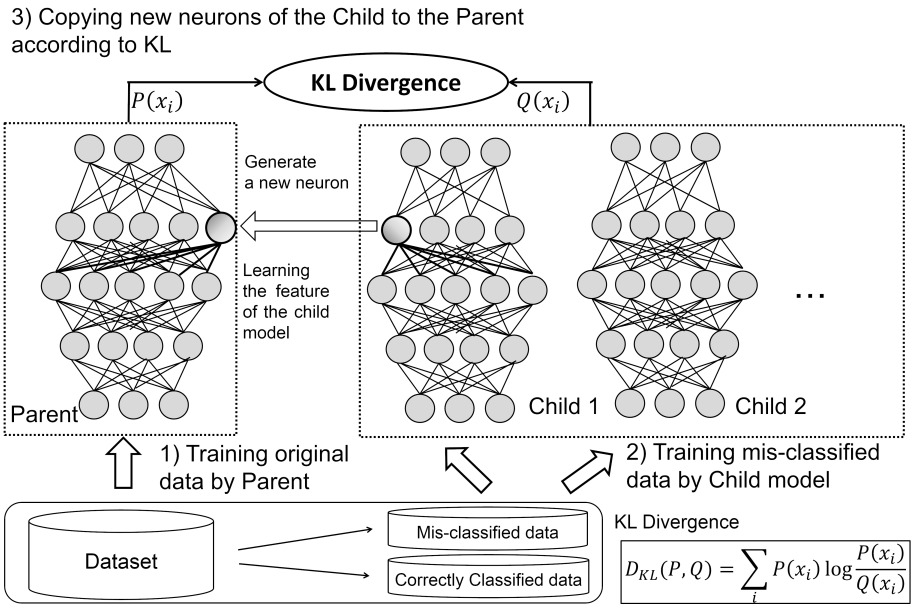
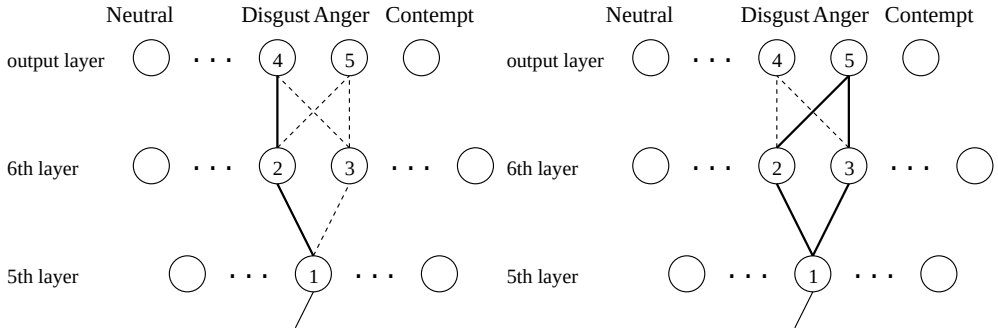


Figure 8: The ensemble learning model of Adaptive DBN



(a) Parent network (b) Child network

Figure 9: Activated neurons and paths in the DBN

Algorithm 1 Fine Tuning Method

1: The DBN structure with the L layers is trained by Adaptive DBN.
2: The fine tuning method in **Algorithm 2** is implemented at the l-th layer ($l \leq L$) from the input layer to the L-th layer.

parent network will calculate a correct output even the incorrectly cases. Moreover, the improvement of the DBN structure at the parent DBN reduces the total calculation time in the ensemble model of Adaptive DBN.

The method initializes a child DBN S with the DBN structure to copy weights from the parent DBN T. S trains the conflict patterns of mis-classified cases such as 'Anger' and 'Disgust' at T. As a result of learning, the classification accuracy at S will be improved outstandingly. The KL divergence between T and S for the given input x_i was given by Eq. (13).

The threshold of KL divergence for the mis-classified cases was varied in $[0.000, 0.0025]$ as described in the subsection 4.1. From the experimental results, the KL divergence D_{KL} between T and S was 0.660, 0.509, 0.367, 0.322, 0.301, 0.312, 0.309, 0.298, and 0.301 for each category of nine confusion patterns, respectively. The threshold value became smaller according to the decrease of confusion cases.

The signal flows on the path of network from the input layer to the output layer at the child DBNs for the input pattern were explored in addition to the KL divergence. If the wrong output pattern is found for the test case, the corresponding neuron and the connected paths are activated incorrectly. The method utilizes that a hidden neuron of RBM is a binary pattern $\{0, 1\}$ and all combinations of binary pattern of neurons are given to the trained DBN network. The flow of signals was analyzed in the direction of input to output and the activated paths in the network were extracted from the framework as inference knowledge. The fine tuning method can find the wrong flow and repair the path partially to improve the classification accuracy. **Algorithm 1-3** are the process of the fine tuning method and the knowledge extraction from the modified network. The method can discover the difference between two networks T and S. (See the details of the method in [12, 18]).

As a result, no significant difference in the path appeared in the lower part closer to the input layer, but characteristic differences were seen in the upper part near the output layer. An example of the activated neurons and paths in case of 'Anger $\rightarrow$ Disgust' is as shown in Fig. 9. A node and a line between nodes are neuron and weight, respectively. The bold line and the dotted line indicate the actually activated path and no activated one. Fig. 9(a)

Algorithm 2 The revise of weight at the l-th layer

1: Let $X(x_1,\cdots,x_p,\cdots,x_N)$ be N input patterns. Let $Y(y_1,\cdots,y_p,\cdots,y_N)$ be target signals. Let l is the output layer for fine tuning at the trained DBN with L layers.
2: Give a X to the trained DBN for feed forward calculation. This process saves the trace and parameters that a neuron is activated from lower layer to upper one. X^T and X^F are correct and wrong set of input patterns for Y, respectively.
3: If a neuron j at the the l-th layer is satisfied with Eq. (14), update weights, $w^{correct}$, connected to the neuron j. Eq. (14) is the ratio of only firing X^T at the neuron j and $w^{correct}$ is a constant value . θ^T is the threshold value. In this paper, $w^{correct}=1$ and $\theta^T=0.3$.

$$|Act_j^T|/(|X^T|+|X^F|)\geq\theta^T, \tag{14}$$

4: If a neuron j at the l-th layer is satisfied with Eq. (15), update weights, w^{wrong}, connected to the neuron j. Eq. (15) is the ratio of only firing X^F at the neuron j. In this paper, $w^{wrong}=0$ and $\theta^T=0.3$.

$$|Act_j^F|/(|X^T|+|X^F|)\geq\theta^F, \tag{15}$$

Algorithm 3 Knowledge extraction

1: x_p in $X(x_1,\cdots,x_p,\cdots,x_N)$ is a I-dimensional vector. The trained DBN has L layers. $Net()$ is the function which calculates output pattern $Y=Y(y_1,\cdots,y_p,\cdots,y_N)$ for given X. y_p indicates a label.
2: Calculate Y by using $Net()$ for X.
3: Extract rules by C4.5 [19].

and Fig. 9(b) are the paths around the output layer at the parent DBN and the child DBN, respectively. In the case, the signal was flowed through 1, 2, and 4 neurons from the fifth layer to the output. But the signal output was the mis-classification to the label 'Disgust' at T in Fig. 9(a). In addition to the same path as T, the child network S activated the path from 1 to 3 neurons where was not activated at T. These new activated paths improved the classification accuracy for the mis-classified output 'Anger' instead of 'Disgust'.

The difference of signal flows between two networks was occurred by the representing ability of features between T and S. If a new activated neuron and their paths in S by neuron generation algorithm is added, T with such neuron and paths can also classify the confusion patterns. For child DBNs with large KL divergence, such neurons and their paths were often seen from the experimental results. In this paper, if the KL divergence is larger than the certain value θ_{KL} and the different path as shown in Fig. 9 is discovered, the parent network is constructed by copying the corresponding neuron in the child network with same weights between neurons. The parent DBN T was trained with small oscillation after the insertion of the transferred neuron at the corresponding path. The automatically assumption of the appropriate θ_{KL} is difficult, but the algorithm with the relation of KL divergence and mis-classified cases will be developing in near future.

The classification accuracy of the proposed ensemble learning was evaluated only for 'Anger' and 'Disgust' with $\theta_{KL}=0.0015$. The ensemble learning model improved the classification accuracy for 'Disgust' and 'Anger' from 92.4% and 78.4% to 94.0% and 94.8%, respectively. Moreover, the training time for the model development was 20.32 hours, while

Table 7: The classification capability by remodeling method

model	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	Total
Parent	87.8%	92.4%	84.2%	85.8%	90.4%	92.4%	78.4%	87.6%	87.4%
Child 1	-	-	-	-	-	94.7%	91.3%	-	-
Child 2	-	-	-	93.4%	-	93.8%	-	-	-
Child 3	-	-	-	-	-	96.4%	-	92.0%	-
Child 4	-	-	94.8%	-	-	-	-	97.2%	-
Child 5	95.0%	-	-	93.8%	-	-	-	-	-
Child 6	94.6%	-	-	-	-	-	-	92.6%	-
Child 7	96.2%	-	-	-	-	91.8%	-	-	-
Child 8	-	-	95.4%	-	-	-	-	93.6%	-
Child 9	-	-	-	-	-	-	94.0%	91.0%	-
Ensemble	93.2%	93.6%	91.4%	92.4%	90.8%	92.4%	91.4%	94.4%	92.5%

Table 8: The classification accuracy of the ensemble model(%)

Category	Adaptive DBN	ensemble method
Neutral	87.8	93.2
Happy	92.4	93.6
Sad	84.2	91.4
Surprise	85.8	92.4
Fear	90.4	90.8
Disgust	92.4	92.4
Anger	78.4	91.4
Contempt	87.6	91.4
Total	87.4	92.5

the reconstruction time by ensemble method was 2.95 hours in the computational environment with two RTX 2080Ti GPUs.

Next, the other categories where exist over than nine confusion cases in the Table 4 were investigated in the same way. Table 7 shows the accuracy of Adaptive DBN and the ensemble method for the original parent DBN and the child DBNs. The ensemble method improved the accuracy ratio from 87.4% to 92.5%. The constructed DBN was remodeling the parent model to the 6 layered child model which consists of 513-441-310-302-321-192 neurons.

Algorithm 4 shows the algorithm of the ensemble learning of TS model. The main procedure is A) learning the original Teacher DBN (Step1), and B) Construct two or more child DBNs to express the case of mis-classification in the Teacher model (Step 2 to 4), C) Neurons and their attributes that output the features of the student DBNs based on the difference of the KL divergence between the Teacher DBN and the student DBN are copied to the original Teacher DBN to improve the classification performance (Step 5 to 6).

Algorithm 4 Algorithm for the ensemble model

1: The adaptive DBN trains the parent DBN T for the training data set.
2: Make some sub data sets $SubSet = \{Set_1, \cdots, Set_i, \cdots, Set_L\}$ for the mis-classified cases in the parent DBN T, where L is the number of sub data sets. In this paper, the sub data set has a certain number of mis-classified cases between categories based on the confusion matrix of training cases.
3: For each sub data set Set_i, build a specific child DBN S_i. The student DBN copies the network structure of the parent DBN T and is initialized at the parameters such as weights of the parent DBN with a little perturbation.
4: Train each student DBN for S_i.
5: Calculate the KL divergence between the parent model T and each child model S_i by Eq.(13). If the KL divergence is larger than a certain threshold θ_{KL} for a given input signal and the difference neuron firing paths are found in the parent DBN and each child DBN as shown in Fig.9, inherits from the attribute of the corresponding neuron in the child DBN to their parent DBN. The fine tuning method described in [18] employs to re-training with a perturbation of parameters at the parent DBN.
6: For all child DBNs, the algorithm ends when no new neurons are copied to the parent DBN. For feed forward calculation, the child DBNs are not used and only the reconstructed parent DBN is applied.

5 Conclusive Discussion

In this paper, we constructed Deep learning model of Adaptive DBN for the AffectNet: Facial Expression Image database. Adaptive DBN is an outstanding function to build an optimal network structure during training. The classification result for training data set was improved, however, the overlapped conflict patterns for the certain emotion categories in the test dataset were found, because human emotion includes many ambiguous and complex features due to the labeling. The classification ratio was decreased for them. The ensemble learning method was proposed to improve the mis-classification using the parent model and multiple child models. The KL divergence between two networks was measured and the fine tuning method can find the difference of them. Since one DBN was used for the reduction of computation resources after training models, the inference knowledge is summarized into the parent model by transferring the generated neuron from child models. In the experiment results, nine child models were constructed to train several confusion cases in mis-classified emotions, then the classification accuracy of the original parent model was improved from 87.4% to 92.5% by embedding the child models' neurons to the parent.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19K12142, 19K24365, and obtained from the commissioned research by National Institute of Information and Communications Technology (NICT, 21405), JAPAN.

References

[1] A.Krizhevsky, I.Sutskever, G.E.Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, Proc. of Advances in Neural Information Processing Systems 25 (NIPS 2012) (2012).

[2] C.Szegedy, W.Liu, et.al., *Going Deeper with Convolutions*, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1-9 (2015).

[3] K.Simonyan, A.Zisserman, *Very deep convolutional networks for large-scale image recognition*, Proc. of International Conference on Learning Representations (ICLR 2015) (2015).

[4] K.He, X.Zhang, S.R en, J.Sun, *Deep residual learning for image recognition*, Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778 (2016).

[5] G.E.Hinton, S.Osindero and Y.Teh, *A fast learning algorithm for deep belief nets*, Neural Computation, vol.18, no.7, pp.1527-1554 (2006).

[6] G.E.Hinton, *A Practical Guide to Training Restricted Boltzmann Machines*, Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science (LNCS, vol.7700), pp.599-619 (2012).

[7] S.Kamada and T.Ichimura, *An Adaptive Learning Method of Restricted Boltzmann Machine by Neuron Generation and Annihilation Algorithm*, Proc. of 2016 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2016), pp.1273-1278 (2016).

[8] S.Kamada and T.Ichimura, *A Structural Learning Method of Restricted Boltzmann Machine by Neuron Generation and Annihilation Algorithm*, Neural Information Processing, vol.9950 of the series Lecture Notes in Computer Science, pp.372-380 (2016).

[9] S.Kamada and T.Ichimura,*An Adaptive Learning Method of Deep Belief Network by Layer Generation Algorithm*, Proc. of IEEE TENCON2016, 2971-2974 (2016).

[10] A.Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Master of thesis, University of Toronto (2009).

[11] S.Kamada, T.Ichimura, A.Hara, and K.J.Mackin, *Adaptive Structure Learning Method of Deep Belief Network using Neuron Generation-Annihilation and Layer Generation*, Neural Computing and Applications, pp.1-15 (2018).

[12] S.Kamada, T.Ichimura, T.Harada, *Knowledge Extraction of Adaptive Structural Learning of Deep Belief Network for Medical Examination Data*, International Journal of Semantic Computing, Vol.13, No.1, pp. 67-86 (2019).

[13] A.Mollahosseini, B.Hasani, M.H.Mahoor: *AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild*, IEEE Transactions on Affective Computing, vol.10, No.1 pp.18-31 (2017).

[14] T.Ichimura, S.Kamada, *Re-learning of Child Model for Misclassified data by using KL Divergence in AffectNet: A Database for Facial Expression*, Proc. of 2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA2019), pp.15-20 (2019).

[15] T.Ichimura, S.Kamada, *A Distillation Learning Model of Adaptive Structural Deep Belief Network for AffectNet: Facial Expression Image Database*, Proc. of the 9th International Congress on Advanced Applied Informatics(IIAI AAI 2020), pp.454-459 (2020).

[16] G.E.Hinton, *Training products of experts by minimizing contrastive divergence*. Neural Computation, vol.14, pp.1771-1800 (2002).

[17] J.Cohen, *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, vol.20, no.1, p.37-46, (1960).

[18] S.Kamada and T.Ichimura, *Fine Tuning of Adaptive Learning of Deep Belief Network for Misclassification and its Knowledge Acquisition*, International Journal Computational Intelligence Studies, Vol.6, No.4, pp.333-348 (2017).

[19] J.R.Quinlan, *Improved use of continuous attributes in c4.5*, Journal of Artificial Intelligence Research, Vol.4, No.1, pp.77-90 (2016).