중심손실과 ResNet50를 통한 얼굴 사진으로부터 차별적 감정 특징 학습

(Learning discriminative emotional features from facial images via ResNet50 with center loss)

*** † *** ‡ *** §

(**** ***) (**** ****) (***** *****)

요 약 본 논문은 중심 손실을 활용해 사람의 표정으로부터 차별적인 감정 특징을 학습하는 CL-ResNet50를 제안한다. 먼저 입력된 이미지로부터 분리될 수 있는 (separable) 특징을 추출하기 위하여 대규모 ImageNet 데이터에 사전 학습된 ResNet50를 활용한다. 해당 모델이 감정 특징을 학습하기 위해 표정 데이터에 재학습 된다. 이때 차별적인 (discriminative) 감정 특징을 학습하기 위해서 mini-batch마다 계산된 각 감정에 해당하는 깊은 특징의 중심과 해당하는 깊은 특징들 사이의 거리를 최소화함으로써 각 감정 클래스에 해당하는 깊은 특징들은 각각의 중심으로 모이도록 했다. 제안하는 CL-ResNet50는 RAF-DB, FER2013 데이터에서 softmax loss만 사용한 것 보다 각각 2.44%, 1.32%의 분류 정확도 향상을 보였다.

키워드 : 얼굴 표정 인식, ResNet50, 중심 손실, RAF-DB

Abstract This paper presents CL-ResNet50 that learns discriminative emotional features from facial images via center loss. Firstly, to extract separable features from input images, we adopt ResNet50, which is pre-trained on large ImageNet data. This model is retrained on facial expression dataset to learn emotional features. During this process, the corresponding deep features of each emotion class cluster to their respective centers, which are computed at every mini-batch by minimizing the distance between the center of each emotion and its corresponding deep features for discriminative emotional feature learning. The proposed CL-ResNet50 shows improvements in classification accuracies than using softmax loss alone by 2.44% and 1.32% on RAF-DB and FER2013 datasets.

Key words : facial expression recognition, ResNet50, center loss, RAF-DB

† 비 회 원 : *****
****@*****
‡ 비 회 원 : **** ***
****@*****
§ 종신회원 : **** ****
******@*****

논문접수 : 2023년 04월 05일
심사완료 :

정보과학회논문지: 정보통신 제31권 제6호(2004.12)

1. Introduction

Automatic recognition of facial expression is an important area of affective computing as it helps the computer understand human behaviors, which result in more natural interaction between them [1]. During the past decades, significant progress has been made on facial expression recognition (FER) using large-scale data-driven deep networks [1]-[7]. For FER, large pre-trained models are effective at learning separable feature representations from facial images [8]. However, such networks are not very good at learning discriminative features due to their label predictions using softmax loss [9]. Although nearest neighbor and k-nearest neighbor algorithms can learn discriminative features, they are not made for label predictions [9]. Thus, the intra-class variance poses a greater challenge in FER attributing to noisy characteristics of facial images belonging to the same emotion class. To learn both separable and discriminative emotional features from still images of faces, we train the ResNet50 [10] with center loss [9], which penalizes the distance between the center of each emotion's deep features and the corresponding deep features.

Previous to our work, a similar approach has been taken by Guo et al. [2], in which the center loss is deployed for discriminative loss function to train the inception-v3. Different from their work, we choose ResNet50 because it has shown better accuracies over other large pre-trained models including Inception V3 according to the earlier study of Melinte and Vladareanu for FER [8]. Its higher generalization performance is mainly attributed to its bottleneck design using shortcut connections that effectively address the issues of vanishing gradients [10]. To enhance the model with discriminative powers, we train the ResNet50 with joint supervision of softmax loss and center loss. The center loss has been first introduced by Wen et al. [9] and they defined the deep features as the last hidden layer before the label prediction. The separable features and the discriminative features are characterized by having inter-class variance and intra-class compactness, respectively [9].

When it comes to our proposed model, we replaced the number of units of the output features of the last fully connected (FC) layer of the ResNet50 [10] with seven units, which are the number of categories of facial expressions in our FER task. As the number of units of input features to the last FC layer of the ResNet50 is 2048, the deep features of our proposed model, CL-ResNet50 has 2048 units.

Through our extensive experiments, we have found that the values of the center weight and its balancing factor are instrumental in obtaining improvements from the center loss. By carefully tuning these parameters, we could gain higher classification accuracies over the ResNet50 that only utilizes softmax loss without center loss. Thus, our contributions can be summarized as follows:

1. We have designed the deep learning framework with the pre-trained ResNet50 to learn effective separable features from facial expressions for emotion recognition as well as discriminative features by minimizing the intra-class variance caused by the distance between the deep features and their corresponding centers.
2. The ablation study shows that the center weight and its balancing factor significantly act on the performance of the center loss. Thus, we have adjusted the parameters through experiments on two benchmark datasets called RAF-DB [11] and FER2013[1] resulting in the effective generalization of the proposed model.

To verify the effectiveness of our model, we present the experimental results with classification accuracies as well as the visualization of the predicted deep features in the experimental section.

This paper is organized as follows. In Section 2, we show related works regarding facial expression recognition and center loss. Next, we describe the model architecture and details of the learning method in Section 3. In Section 4, we compare the performance of the CL-ResNet50, which is trained under the joint supervision of softmax loss and center loss with the baseline model solely using softmax loss. Finally, we draw conclusions with the future research direction in Section 5. We publicly release the implementation code, which will be available at:

1 https://www.kaggle.com/datasets/msambare/fer2013

https://github.com/KangHyunWook/Journal-of-KIISE-2023.

2. Related Works

2.1 Facial expression recognition

In the past decades, many researchers have worked on automatic facial expression recognition [1]-[7]. Most of them focused on distinguishing six basic emotions as they are commonly detected in our daily lives [4]. Classical feature extraction methods utilized hand-crafted visual features such as SIFT [12], HOG [13], and LBP [14]. Inspired by the success of deep networks on image classification [15] and object detection [16], following studies on FER used deep learning models. Most works have focused on expression recognition on frontal faces and a few attempted to detect pose-invariant facial expressions [4].

Many different algorithms have been introduced for FER as follows. Zhao et al. [6] attempted to achieve invariance to expression intensity by exploiting the peak-piloted deep network (PPDN) that is fed with facial images paired with a peak and a non-peak expression. To avoid annotation bias from multiple inconsistently labeled facial expression datasets, Zeng et al. [3] proposed the inconsistent pseudo annotations to latent truth (IPA2LT) framework that discovers noise patterns by tagging multiple labels for each sample. Later, Wang et al. [1] proposed Self-Cure Network (SCN) to suppress the uncertainties for large-scale facial expression recognition. Considering that most emotions are combinations of basic emotions, label distribution learning (LDL) method is introduced by Zhao et al. [7] to train their proposed EfficientFace. It consists of local-feature extractor and a channel-spatial modulator that effectively improve the robustness of the lightweight network [7]. Not only six basic emotions but also predictions of compound emotions from still images of faces have been attempted by Guo et al. [2]. Melinte and Vladareanu [8] performed both face recognition (FR) and FER for the interaction between humans and NAO robot by using two optimized CNNs.

2.2 Center loss

The center loss has been originally introduced to learn discriminative features for FR task [9]. For this purpose, the networks are trained under the joint supervision of softmax loss and center loss [9]. The original form of the center loss is defined as:

$$L_C = \frac{1}{2}\sum_{i=1}^{m}||x_i - c_{y_i}||_2^2 \quad (1)$$

where, x_i denotes the i-th deep features of emotion class y_i, and c_{y_i} denotes the center of the deep features corresponding to the respective emotion class, given m number of samples [9]. However, due to the tremendous amount of complexity, such type of Eq. 1 has never been attempted [9]. To address, Wen et al. [9] proposed a new formula, in which the center loss is computed at every mini-batch. This initial attempt triggered huge inspirations on later studies of computer vision, zero-shot learning and speech recognition.

In the field of computer vision, Gune et al. [17] adapted center loss for zero-shot learning. In their work, the distance between the class wise visual centroids and semantic class prototypes are minimized for the alignment of the inter-class structures of visual and semantic space [17]. Later, Zhan et al. used the similar method to project the embedded class semantic features to the mean vector of learned latent adjective-noun pairs (ANP) features for visual sentiment analysis [18]. Center loss was also deployed for speech emotion recognition by Dai et al. [19]. They trained the network with joint supervision of crossentropy softmax loss and the center loss, in which the former loss enables features to be separable from different emotion classes and the latter pulls the features within the same emotion class to the center.

3. Methodology

The main goal of this paper is to predict the correct emotion expressed in human faces. There are seven discrete emotion categories for both RAF-DB and FER2013 as shown in Table 2. We adapt the pre-trained ResNet50 for FER and the details of its architecture is described in Table 1.

Table 1. Modified architecture of ResNet50 for FER task.

layer name	output size	50-layer
conv1	112×112	7×7, 64, stride 2
conv2_x	56×56	3×3 max pool, stride 2 $\begin{pmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{pmatrix}\times3$
conv3_x	28×28	$\begin{pmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{pmatrix}\times4$
conv4_x	14×14	$\begin{pmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{pmatrix}\times6$
conv5_x	7×7	$\begin{pmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{pmatrix}\times3$
	1×1	average pool
	1×1	7-d fc

The output features of average pooling layer with 2048 units are the deep features denoted as d in Eq. (2). At every mini-batch, the respective center of the corresponding deep features of same emotion class are computed as:

$$\frac{\Delta c_j=\sum_{i=1}^{m}\delta(y_i=j)\cdot(c_j-d_i)}{1+\sum_{i=1}^{m}\delta(y_j=j)} \quad (2)$$

where, c_j is the center of deep features corresponding to j-th emotion class. Given m number of samples, the function δ returns 1 if its parameter condition $y_i=j$ is satisfied, that is, the emotion class of the i-th deep features belongs to the emotion category j. Otherwise, it returns 0. The center loss sums up the distances between the deep features and their respective center. Then they are divided by the number of samples within the same emotion class. Note that 1 is added to the number of samples belonging to same emotion class to prevent the division by zero. Let the computed gradients of the center respective to the i-th deep features be denoted as L_C, then the final loss can be defined as:

$$L=L_S+\lambda L_C \quad (3)$$

where, L_S denotes the softmax loss. Aforementioned formulae for the center loss is inspired by [9], in which more details can be found.

4. Experiments

4.1 Datasets

Table 2. Statistics of two benchmark datasets of RAF-DB and FER2013. It shows the number of samples that belong to their respective emotion classes.

	emotion categories						
Dataset	anger	disgust	fear	happiness	neutral	sadness	surprise
RAF-DB	705	717	281	4772	2524	1982	1290
FER2013	3995	436	4097	7215	4965	4830	3171

For the experiments, we use two benchmark facial expression datasets called RAF-DB and FER2013. The statistics of each dataset is shown in Table 2, in which the number of each emotion class in the trainset is shown.

4.1.1 RAF-DB

There are 30,000 facial images with basic or compound expressions [11]. It provides labels both in basic or compound expressions. In our experiment, we only consider the samples with basic emotions: anger, disgust, fear, happiness, neutral, sadness, surprise. Thus, the CL-ResNet50 is trained with 12,271 images and evaluated on the rest 3068 images. It is notable that the benchmark provides split train and test set for fair evaluation.

4.1.2 FER2013

It is comprised of 28709 training images, and 3589 validation set [5]. All images are of shape 48×48 pixels. The images belong to one of the 7 emotion classes, which are anger, disgust, fear, happiness, neutral, sadness, and surprise.

4.2 Experimental settings

The original training data is split into 70% and 30% into trainset and validation set, respectively. The input facial images are resized to 224×224 and they are randomly flipped horizontally with random rotations to avoid overfitting. Additionally, the input images are cropped at a random location. For normalization, the pixel values belonging each channel of each input image is subtracted by the mean values of all training data corresponding to the respective channel. Then they are divided by the standard deviation of all pixel values corresponding to each channel of all training data.

To train the proposed CL-ResNet50, the adam optimizer is set with initial learning rate of 1e-3 for both softmax loss and center loss. The proposed CL-ResNet50 is trained with 50 epochs with the batch size of 62. To prevent the model from overfitting, the patience is set to 6 for early stopping.

4.3 Experimental results

Firstly, we compare the classification accuracies and the F1 score of CL-ResNet50 with a vanilla ResNet50 [10]. As shown in Table 3, the joint supervision of softmax loss and center loss improves the vanilla ResNet50 model that is trained with softmax loss alone. The performance is measured with the alpha and the lambda values set to 1.9 and 0.2, respectively. The CL-ResNet50 outperforms the vanilla ResNet50 model by 2.44% and 1.32% in classification accuracies on RAF-DB and FER2013, respectively. In F1 score, the performance has improved by 2.42% and 0.92% on RAF-DB and FER2013 datasets, respectively.

Table 3. Comparison of performances in classification accuracies and F1-score with ResNet50 and CL-ResNet50.

	ResNet50		CL-ResNet50	
Dataset	Acc	F1	Acc	F1
RAF-DB	76.239	75.580	**78.683**	77.996
FER2013	62.079	61.566	**63.402**	62.487

Next, we reduce the number of units of the output deep features from the CL-ResNet50 to 2 units using principal component analysis (PCA) to observe if they cluster to their respective centers via the proposed learning method. We plot the PCA embeddings, which are the predicted deep features on the train data of RAF-DB and the result of the visualization is shown in Fig. 1. The left figure shows the predicted vectors of the vanilla ResNet50 with softmax loss alone and the right figure shows the predicted vectors of the CL-ResNet50, which is trained under the joint supervision of softmax loss and center loss. As can be seen in Fig. 1, the predicted deep features in the left suffers from large intra-class variance, while the predicted deep features in the right are clustering to their respective centers, which result in the effective intra-class compactness.

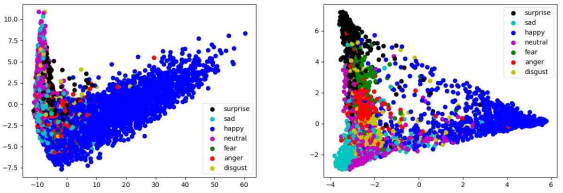

Fig. 1. Visualization of predicted deep features on train data of RAF-DB. The original dimensions of predicted deep features are 2048 units and they are embedded to 2 dimensions using PCA to imitate the original distribution. The X-axis denotes the first element of the 2-dimensional pcaed deep features and the Y-axis denotes the second element.

To further investigate the FER performance of CL-ResNet50, we also present the following confusion matrices as shown in Fig. 2 and Fig. 3.

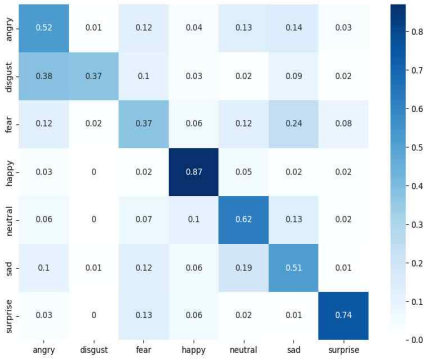

Fig. 2. Confusion matrix showing the performance on FER2013 in classification accuracies. The row denotes the true labels and the right column denotes the predicted emotion labels.

In both figures, we observe that the highest recognition rate is achieved by happy emotion whereas the lowest recognition rate is observed in disgust emotion. Additionally, we find that on FER2013 in Fig. 2, the 24% of fear samples are misclassified as sad. While 30% of fear samples are misclassified as surprise on RAF-DB. The high misclassification rate of fear emotion on RAF-DB may be attributed to the small number of fear samples for training. As a result, the predicted fear emotions may

have had similar activation levels with the predicted surprise emotions.

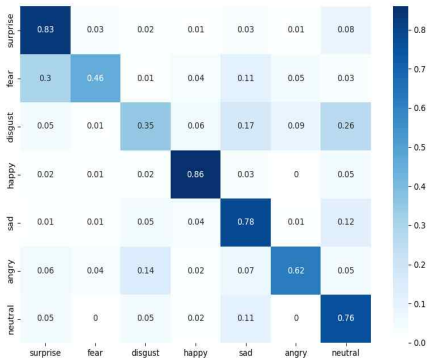

Fig. 3. Confusion matrix showing the performance on RAF-DB in classification accuracies. The row denotes the true labels and the column denotes the predicted emotion labels.

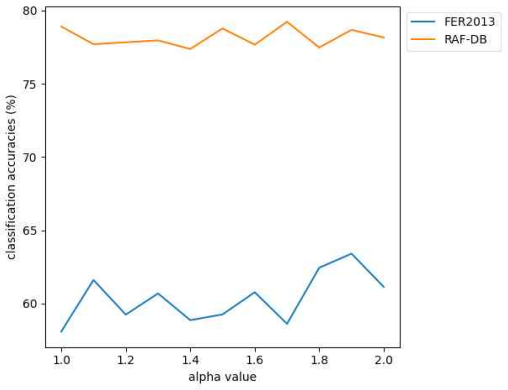

Fig. 4. Classification accuracies according to different balancing factors. The graph shows the result of the center weight fixed to 0.2. The blue line represents the relative performances of the CL-ResNet50 on FER2013 with different alpha values and the orange line shows the result on RAF-DB.

Finally, we observe the role of the balancing factor denoted as alpha for the center loss as shown in Fig. 4. The table shows the relative performances according to different balancing factors with the center weight fixed to 0.2. The higher balancing factor of the center loss drives the gradients of the center loss to be updated with higher step sizes.

According to Fig. 4, the best performance on RAF-DB can be achieved with the balancing factor of 1.7. However, it rather degrades the performance on the FER2013. As an alternative, we have chosen 1.9 for the balancing factor, which achieves best improvements both on RAF-DB and FER2013 datasets.

5. Conclusion

Recently, facial expression recognition has attracted many attentions from researchers due to their various applications. Although a significant performance has been achieved by earlier studies, training a large pre-trained model with softmax loss alone suffers from a large intra-class variance. While the center loss can minimize the intra-class variance, careful selection of center weight and its balancing factor is required to ensure the robustness of the trained model across different datasets. The visualization of the predicted deep features shown in Fig. 1 verify the effectiveness of the proposed CL-ResNet50 in drawing them to their respective centers. The experimental results in Table 3. shows that the improvements using the proposed model on FER2013 is not as good as what's been noticed on RAF-DB. We assume that it is due to the many noises contained in FER2013. As demonstrated by [1], we aim to improve our proposed model by adapting methods, which deal with datasets that contain many noises for a future work.

References

[1] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6897-6906, August, 2020.

[2] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, and B. Knyazev, et al., "Dominant and complementary emotion recognition from still images of faces", IEEE Access, Vol. 6, pp. 26391-26403, Apr. 2018.

[3] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets", In Proceedings of the European conference on computer vision (ECCV), pp. 222-237, 2018.

[4] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition", In Proceedings of the IEEE conference on computer vision and pattern

recognition, pp. 3359-3368, June, 2018.
[5] M. I. Georgescu, R. T. Ionescu, and M. Popescu, “Local learning with deep and handcrafted features for facial expression recognition”, IEEE Access, Vol. 7, pp. 64827-64836, May, 2019.
[6] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, “Peak-piloted deep network for facial expression recognition”, In Computer Vision - ECCV 2016: 14th European Conference, pp. 425-442, October, 2016.
[7] Z. Zhao, Q. Liu, and F. Zhou, “Robust lightweight facial expression recognition network with label distribution training”, In Proceedings of the AAAI conference on artificial intelligence, pp. 3510-3519, May, 2021.
[8] D. O. Melinte and L. Vladareanu, “Facial expressions recognition for human - robot interaction using deep convolutional neural networks with rectified adam optimizer”, Sensors, Vol. 20, No. 8, April, 2020.
[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition”, Computer Vision - ECCV 2016: 14th European Conference, pp. 499-515, October, 2016.
[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
[11] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2852-2861, July, 2017.
[12] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, International journal of computer vision, Vol. 60, pp. 91-110, 2004.
[13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Vol. 1, pp. 886-893, June, 2005.
[14] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, “Learning active facial patches for expression analysis”, 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2562-2569, June, 2012.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, Communications of the ACM, Vol. 60, No. 6, pp. 84-90, June, 2017.
[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.
[17] O. Gune, B. Banerjee, and S. Chaudhuri, “Structure Aligning Discriminative Latent Embedding for Zero-Shot Learning”, BMVC, 2018.
[18] C. Zhan, D. She, S. Zhao, M. M. Cheng, and J. Yang, “Zero-shot emotion recognition via affective structural embedding”, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1151-1160, October, 2019.
[19] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, “Learning discriminative features from spectrograms using center loss for speech emotion recognition”, In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7405-7409, May, 2019.