Facial Expression Recognition Based on Transfer Learning and SVM

Facial Expression Recognition Based on Transfer Learning and SVM

Lei Yang, Haiqing Zhang, Daiwei Li*, Fei Xiao and Shanglin Yang

Department of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China
Email: ldwcuit@cuit.edu.cn

Abstract. The facial expression datasets always have a problem: data with small amount or large amounts of data but also with large noisy. Both problems will affect the facial expression recognition accuracy of the model. A transfer learning method for facial expression recognition is proposed by combining the Convolutional Neural Network (CNN) and Support Vector Machine (SVM). SVM have good performance on small data sets and CNN based on transfer learning have better ability of feature extraction for large noisy data set. This method reduces the training time of model and increase the facial expression recognition accuracy. The experimental results show that the accuracy of the proposed method on the CK+ and FER2013 data sets has reached 99.6% and 68.1%.
Keywords. Convolutional neural networks; facial expression; transfer learning; support vector machine; Inception-ResNet-v1.

1. Introduction

Facial expression recognition technology has a wide range of applications in the fields of driving safety, business and education. However, there are still many difficulties in facial expression recognition in practice. In the real expression dataset, there are usually the following problems: brightness effect (too bright or too dark), various face angles, occluder and more. The facial expression recognition based on deep learning is more effective than traditional method in these issues. In 2016, Zhao et al. [1] proposed peak-piloted deep network (PPDN). PPDN uses samples with peak expression (simple samples) to supervise the intermediate feature responses of the same type of non-peak expression samples (hard samples) from the same subject. PPDN performs well on the Oulu-CASIA and CK+ datasets. In 2021, Cui et al. [2] proposed improved VGGNet and improved Focal Loss which achieved extremely good performance in CK+, JAFFE and FER2013.

To improve the accuracy of facial expression recognition and enhance the generalization of the training model, the model based on the traditional Inception-ResNet-v1 [3] has been modified by replacing softmax with SVM. The new model is training by transfer learning. The improved model has better performance on CK+ [4] and FER2013 [5].

The rest of this paper is organized as follows: Section II details proposed method; Section III introduce experiments preparation; results and analysis of experiments are in Section IV. Finally, we conclude our work and present future challenges in Section V.

2. The Proposed Methods

In our methods, the Inception-ResNet-v1 model derived from FaceNet [6] is adopted, which is trained on the VGGFace2. The dataset including more than three million images of various objects, with an

average of 362.6 images for each subject. Even though Inception-ResNet-v1 model is based on VGGFace2 which is specially organized to face recognition instead of facial expression recognition, the features are highly similar between face recognition and facial expression recognition. Therefore, some knowledge learned by Inception-ResNet-v1 model can be shared with the task of facial expressions recognition.

2.1. The Framework of the Proposed Method

The framework of the proposed method is based on the Inception-ResNet-v1 with replacing the softmax with SVM. Some layer parameters are transferred from the pre-trained model, so those layers will be frozen. Only training the unfrozen layers and SVM with new datasets. The overall architecture of our proposed method is shown in figure 1. And the main steps of our method are Fine-tuning Pre-trained Model and training classifiers.

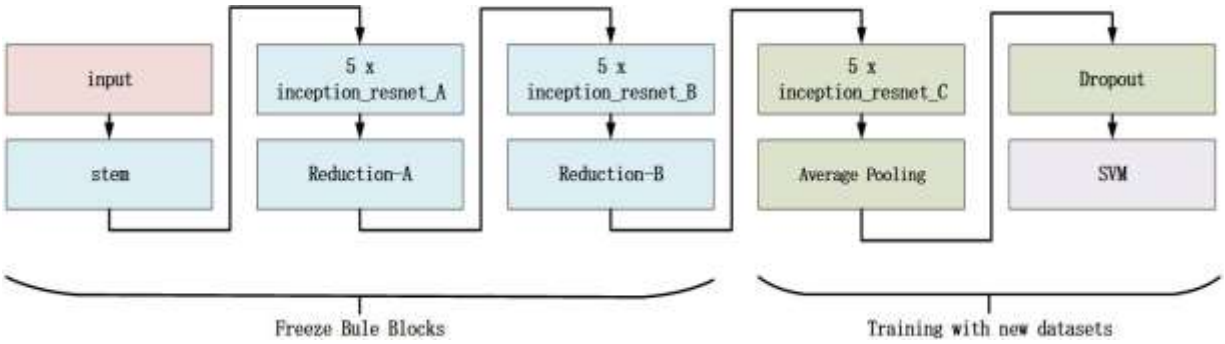

Figure 1. The architecture of the proposed method.

Inception-ResNet-v1. Inception-ResNet-v1 is a convolutional neural network which combined Inception architecture with residual connections in 2017. The residual connections can accelerate the training of the Inception network with a small increase in accuracy. And the computational cost of Inception-ResNet-v1 is similar to Imception_v3 [7]. Also, this network is widely used in image classification.

Transfer Learning. Transfer learning is widely used in the field of image processing for improving learning and training efficiency. Principles of transfer learning is transferring information from the related domains (source domain) to another domains (target domains). In this paper, transfer learning is used to share the learned model's parameters with the proposed model.

Support Vector Machine. Support Vector Machine is a supervised classification algorithm and a linear classifier with the largest interval in the feature space. SVM maps the vector from low-dimensional space to high-dimensional space using kernel function method to solve the nonlinear two-class classification problem [8].

2.2. Fine-Tuning Pre-trained Model

The pre-trained model is chosen in this paper is trained by Google in 2018 which has state-of-art performance in face recognition. The details of the pre-trained model are shown in table 1.

Table 1. Details of pre-trained model.

Model name	LFW accuracy	Training dataset	Architecture
20180402-114759	0.9965	VGGFace2	Inception-ResNet-v1

Firstly, the pre-trained model was loaded, and stem, 5xInception_resnet_A, Reduction_A, 5xInception_resnet_B and Reduction_B were freezed. Secondly, the facial expression datasets were

input to train the rest layers and softmax. In this process, the weight coefficients of frozen layers won't be updated.

2.3. Training Classifiers

The output of the network is embeddings. Therefore, the model is loaded without softmax above all. For calculating embeddings, facial expression datasets in models are input to run forward pass. Then machine learning classifiers are trained with those embeddings.

3. Experiments Preparation

This section includes data preparation and experiment environment. All the experiments were evaluated in a PC computer by running in the CentOS 7 system, with a 2.10 GHz Inter Xeon Gold 6230R CPU, 124.4G RAM and 1.0 TB hard disk. The software environment is TensorFlow-gpu (1.7.0) and python 2.7.

3.1. Data Preparation

In this part, three databases and data enhancement methods in experiments will be introduced. CK+ is lab-controlled data. SFEW [9] and FER2013 are wild environment data. All of them have seven emotion labels, which include six basic expressions. The details of emotion labels are shown in table 2.

Table 2. Details of Emotion Labels in Three Datasets.

Dataset	Six Common Basic Expressions		Plus
CK+	Anger	Disgust	Contempt
SFEW 2.0	Fear	Sadness	Neutral
FER2013	Happiness	Surprise	Normal

3.2. DataSet Introduction

- CK+: The CK+ database is laboratory-controlled database. Sequences in CK+ show a shift from neutral expression to peak expression and shown in figure 2. Only the peak expression images are selected in our experiments.
- FER2013: The FER2013 database is large-scale and unconstrained. The image data of it is collected automatically by Google image search API. There are 28,709 training images, 3,589 teat images and 3,989 validation images in FER2013.
- SFEW 2.0: The Static Facial Expression in the Wild (SFEW) database, which has been selected static frames from movies, is different from the laboratory-controlled database. There are 958 training images, 372 teat images without labels and 436 validation images in SFEW 2.0.

Figure 2. part of the sequences in CK+.

3.3. *Data Enhancement*

In order to match Inception-ResNet-v1 model, it is necessary to preprocess and enhance the data. First, the images are resized to 160*160 to match the input size. Second, the original images are enlarged by flipping, brightening, darkening, adding GaussianNoise and GaussianBlur, rotating 180 degrees and

90 degrees, which also enhanced data. The examples of data enhancement are shown in figure 3. After enhancing data, there are 17,411 training images of CK+ and 33,039 training images of SFEW 2.0.

Figure 3. Data enhancement.

4. Experimental Results and Analysis

In order to illustrate the performance of our method, experiments have been done in this section. Firstly, models training from scratch and models training based on pre-train model are compared in three datasets. And then, combining those two type models with SVM (linear), SVM (rbf) and Random Forest. In addition, SVM (linear) means SVM with linear kernel and SVM (rbf) means SVM with gaussian kernel.

4.1. Model Performance Comparison

For better illustrate the performance of our method, the network architecture is Inception-ResNet-v1 in all models. And the details of the parameters are shown in table 3. Those parameters are choosing through many experiments.

Table 3. Parameters of network.

Parameters	Setting
Learning Rate	0.1
Batch Size	128
Epoch Size	200
Optimizer	ADAM
Embedding Size	128
Embedding Size	512
Epoch Numbers (CK+)	10
Epoch Numbers (SFEW 2.0)	10
Epoch Numbers (FER2013)	50

The embedding size of the pre-trained model is 512. But setting embedding size is 128 for models training from scratch will have better performance than setting embedding size is 512. So, experimented on different embedding size settings. To check the effectiveness of our method quickly, small epoch numbers are chosen for different datasets. And the experimental results shown in table 4. Specifically, the rules of model code are “dataset + digital”. “1” means models training based on pre-train model. “2” means models training from scratch with embedding size is 512. “3” means models training from scratch with embedding size is 128.

In general, on the condition of same embedding size, models training based on pre-train model have significantly better performance for all datasets than models training from scratch and improved accuracy results by 5.571%-11.929%. Particularly, embedding size is a crucial factor in our experiments. In FER2013, FER3 even have subtly better performance than FER1.

4.2. Classifier Performance Comparison

In this part, six models which trained from last part are selected to replace softmax with the traditional machine learning classifiers. The parameters of the classifiers are shown in table 5. Those main parameters are choosing through many experiments.

Table 4. Model performance results.

Model code	Dataset	Pre-trained model	Embedding size	Accuracy
CK1	CK+	✓	512	**0.97714**
CK2	CK+	×	512	0.92143
CK3	CK+	×	128	0.96714
SFEW1	SFEW 2.0	✓	512	**0.57571**
SFEW2	SFEW 2.0	×	512	0.51429
SFEW3	SFEW 2.0	×	128	0.52357
FER1	FER2013	✓	512	**0.68500**
FER2	FER2013	×	512	0.56571
FER3	FER2013	×	128	**0.69429**

Table 5. Parameters in classifiers.

Classifier	Main setting
SVM (linear)	gamma=10, probability=True, C=10
SVM (rbf)	gamma=10, probability=True, C=10
Random Forest	n_estimators=1000, max_depth=None

And the six models are CK1, CK2, CK3, FER1, FER2 and FER3. They are combining with machine learning classifiers and training separately on CK+ and FER2013. The results are shown in table 6.

Table 6. Training on CK+ and FER2013.

CK+				FER2013			
Model code	Classifier	Accuracy	Trained time	Model code	Classifier	Accuracy	Trained time
CK1	SVM (linear)	**0.996**	**11.147s**	FER1	SVM (linear)	**0.681**	**40.019s**
CK1	SVM (rbf)	0.991	456.312s	FER1	SVM (rbf)	0.679	846.796s
CK1	Random Forest	**0.996**	**272.248s**	FER1	Random Forest	**0.681**	**991.081s**
CK2	SVM (linear)	0.995	14.733s	FER2	SVM (linear)	0.490	3638.737s
CK2	SVM (rbf)	**0.998**	**498.703s**	FER2	SVM (rbf)	0.503	4168.842s
CK2	Random Forest	0.993	393.786s	FER2	Random Forest	**0.520**	**1329.447s**
CK3	SVM (linear)	**0.998**	**5.008s**	FER3	SVM (linear)	**0.657**	**131.248s**
CK3	SVM (rbf)	0.991	317.055s	FER3	SVM (rbf)	0.627	513.965s
CK3	Random Forest	**0.998**	**170.966s**	FER3	Random Forest	0.644	877.651s

In CK+, all types of classifiers achieved better performance than softmax, especially SVM (linear) improved accuracy results by 1.886%~7.357%. In some case, Random Forest and SVM (linear) do have same accuracy. However, the training time of Random Forest is more than 20 times that of SVM (linear). In FER2013, SVM (linear) is also the best choice among the comparison classifiers. The classifiers combined with FER1 better than others. And that proved FER1 have the best ability of feature extraction. Comprehensively, SVM (linear) achieved the best performance for most of the

cases. Specially point out, classifiers combined with CK2 or FER2 all have significantly different performance with others.

Compared with classifiers combined with CK3 or FER3, the more suitable embedding size for facial expression recognition must be 128. And transfer learning can improve models' ability of feature extraction. Additionally, the effect of embedding size is particularly obvious. In tables 4 and 7, the "dataset" +2 is always worse than "dataset" +3. It seems 128 is the suitable embedding size for face expression recognition. Far more, we compared our models with others in CK+ and FER2013. The results show that our models all have good performance in accuracy. The result shows in table 7.

Table 7. Comparison with different methods on CK+ and FER2013.

CK+		FER2013	
Method	Accuracy	Method	Accuracy
Zeng [10]	97.35%	Zeng [10]	61.86%
Nwosu L [11]	97.71%	Bag of words [14]	67.40%
M-MobileNet [12]	99.29%	proposed method	68.1%
EM-AlexNet [13]	94.25%	I_FL [2]	72.49%
Proposed method	99.6%	VGG [15]	73.28%

Comparing with excellent models, proposed model still slightly improved the accuracy by 0.31%~5.35%. On FER2013, proposed model not achieved the state-of-art performance but better than 65%. Human accuracy achieved on this dataset is at par with 65% [11].

5. Conclusions

This paper combined Inception-ResNet-v1 with SVM and training model by transfer learning. Also, embedding size verified in this paper is a crucial factor for face expression recognition and 128 is the suitable choice. This work can be further extended to training new face recognition model with 128 embedding size as pre-trained model. Also, studying at improve Loss function is another way.

Acknowledgments

This research is supported by the Building Skills 4.0 through University and Enterprise Collaboration (Erasmus+Shyfte 4.0) Project (Erasmus+Programme: 598649-EPP-1-2018-1FR-EPPKA2-CBHE-JP, http://shyfte.eu), and the National Natural Science Foundation of China (NSFC) (No. 61602064), and the Science and Technology Agency Project of Sichuan Province (Nos. 2021YFH0107).

References

[1] Zhao X, Liang X, Liu L, et al. 2016 Peak-piloted deep network for facial expression recognition. *European Conference on Computer Vision* pp 425-442.
[2] Cui Z, Pi J, Chen Y, et al. 2021 Facial expression recognition combined with improved VGGNet and focal loss *Computer Engineering and Applications* 2007-0492.
[3] Szegedy C, Ioffe S, Vanhoucke V, et al. 2017 Inception-v4, inception-ResNet and the impact of residual connections on learning. *Artificial Intelligence.* 221-231.
[4] Lucey P, Cohn J F, Kanade T, et al. 2010 The extended Cohn-Kanade dataset (CKD): A complete dataset for action unit and emotion-specified expression *Computer Vision and Pattern Recognition Workshops* pp 94-101.
[5] Goodfellow I J, Erhan D, Carrier P L, et al. 2013 Challenges in representation learning: A report on three machine learning contests *Neural Information Processing* 117-124.
[6] Schroff F, Kalenichenko D and Philbin J 2015 FaceNet: A unified embedding for face recognition and clustering *Computer Vision and Pattern Recognition* 815-823.
[7] Szegedy C, Vanhoucke V, Ioffe S, et al. 2016 Rethinking the inception architecture for computer vision *Computer Vision and Pattern Recognition* 2818-2826.

[8] Wu H, Huang Q, Wang D, et al. 2018 A CNN-SVM combined model for pattern recognition of knee motion using mechanomyography signals *Journal of Electromyography and Kinesiology* **42** 136-142.
[9] Dhall A, Goecke R, Lucey S and Gedeon T 2011 Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark *Computer Vision Workshops* pp 2106-2112.
[10] Zeng G, Zhou J, Jia X, et al. 2018 Hand-crafted feature guided deep learning for facial expression recognition *Automatic Face & Gesture Recognition* 423-430.
[11] Nwosu L, Wang H, Lu J, et al. 2017 Deep convolutional neural network for facial expression recognition using facial parts *Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, Big Data Intelligence and Computing and Cyber Science and Technology Congress* pp 1318-1321.
[12] Wang W, Zhou X, He X, et al. 2020 Facial expression recognition based on improved MobileNet *Computer Applications and Software* **37** 137-144.
[13] Yang X and Shang Z 2020 Facial expression recognition based on improved AlexNet *Laser & Optoelectronics Progress* **57** (14) 243-250.
[14] Ionescu R T, Popescu M and Grozea C 2013 Local learning to improve bag of visual words model for facial expression recognition *Workshop on Challenges in Representation Learning.*
[15] Khaireddin Y and Chen Z 2021 Facial emotion recognition: State of the art performance on FER2013 *arXiv preprint arXiv:2105.03588.*