Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231122

Comparison and analysis of deep neural networks in facial expression recognition

Jiahang Li

School of Software Engineering, Huazhong University of Science and Technology,Wuhan, 430074, China

Lijiahang226@163.com

Abstract. Recognizing facial expressions automatically is of importance to interactive computer systems since facial expression is an efficient means of conveying emotions. Over the past few years, many researchers have attempted to use deep learning for expression recognition. The advantage of deep learning lies in its ability to learn features from datasets automatically, without relying on hand-crafted features. The paper analyzes what mechanism is useful for expression recognition in deep learning by comparing the performance of different popular models and algorithms from recent research. The paper trains the models on the Facial Expression Recognition Dataset (FER-2013), which is a relatively small and imbalanced dataset. After that, model performance is assessed on the private test dataset of FER-2013. Specifically, Residual Network (ResNet), Visual Geometry Group Network (VGGNet), and MobileNet are evaluated in the experiment. The evaluation is based on running time, the number of parameters, and the accuracy. Squeeze and Excitation (SE) block is utilized in the ResNet, which enables the models to learn useful features from global information. In the paper, ResNet34 inserted with SE block (SE-ResNet34) get the highest private test accuracy of 70.80%. Experimental results show that the residual learning enables models to go deeper without degradation and the SE block is beneficial for the model to learn global information.

Keywords: facial expression recognition, deep learning, ResNet, squeeze and excitation.

1. Introduction

Facial expression recognition is a traditional computer vision task aimed at recognizing and categorizing facial expressions from images of human faces. Identifying the expression of human beings is important in communication since the expression is a powerful signal conveying human emotion. As a result, many previous researchers have tried to put facial expression recognition into many interactive computer systems such as social robotics, medical treatment, and driver fatigue surveillance [1]. Up to now, many algorithms demonstrate excellent performance in the frontal expression recognition task. Traditional facial expression recognition mainly utilized hand-crafted features. However, the problem with the methods based on hand-crafted features is that it is difficult to choose the optimal feature since the face images are manifold in natural situations. In natural conditions, facial expression recognition is still challenging because of subjective identity bias, variations in head pose, illumination, and occlusions [1]. Under such situations, the performance of traditional facial expression recognition methods may suffer great losses. Therefore, more effort needs to be put into the facial expression recognition task to improve the performance of algorithms under naturalistic conditions.

However, in the past few years, Convolutional Neural Network (CNN) has been a popular choice to solve image analyzing problems. According to the previous work, CNN can extract and learn high-level features of human facial images. And the features CNN learns were proven to correspond to the Action Units (AUs), which are used in the Facial Action Coding System (FACS) [2] to recognize facial expressions of humans [3]. This allows facial expression recognition based on CNN to be supported by previous theories. Many previous deep learning-based methods conducted experiments on the Facial Expression Recognition Dataset (FER-2013) [4]. FER-2013 was introduced by Goodfellow et al. as a contest to compare learning methods and methods based on hand-crafted features. In the competition, convolutional neural networks were used by all the top three teams. The winner achieved 71.162% accuracy by combing Support Vector Machine (SVM) and CNN [5]. It can be said that in 2013, deep learning outperformed humans in facial expression recognition. Besides the research mentioned above, Minaee et al. used the attention mechanism in the convolutional network, achieving 70.02% test accuracy [6]. Vulpe-Grigoraşi et al. achieved 72.16% test accuracy by utilizing the hyperparameter optimization technique [7].

The paper aims to compare and analyze the performance of several advanced deep-learning models for recognizing facial expressions including Residual Network (ResNet), MobileNet, and Visual Geometry Group Network (VGGNet) [8-11]. In addition, the characteristics of the various models are analyzed in terms of how they affect the final model performance. Specifically, first, pre-processing of the image data is performed. Since FER-2013 is a relatively small and unbalanced dataset. Data augmentation methods are used to enrich the dataset. In addition to this, class weighting and initial bias are applied to reduce the negative effects of the unbalanced dataset. Second, the hyperparameters are identical to ensure fairness in the training process. Third, Adam is uniformly used as an optimizer to optimize cross-entropy losses. Finally, the paper evaluated the performances of all the models on the FER-2013 private test dataset. The results of the experiment demonstrate that the model with joint residual structure and Squeeze and Excitation (SE) blocks outperforms the other models, achieving 70.80% accuracy. It is further shown that the residual structure allows information to span directly between different layers, thus avoiding the model degradation problems that occur in traditional deep neural networks, resulting in better performance. Besides, it is also demonstrated that the SE mechanism enables models to learn the features more efficiently.

2. Methodology

2.1. Dataset description and preprocessing

There are quantities of available datasets used in the facial expression recognition task. The existing facial expression datasets can be simply divided into two categories, one is full of the complete frontal face images, and the other is taken under natural conditions with various problems like variation of nodes, illumination, and occlusion. In general, it is more difficult to perform expression recognition tasks in the latter datasets. This paper trains the model on FER-2013 [4]. FER-2013 is one of the most challenging datasets because all its images are taken under uncontrolled situations. Even human accuracy on FER-2013 was only 65±5%. All images in FER-2013 have been resized to a size of 48 × 48. It contains 35887 images, including 28,709 training images and 3564 public test images, and 3564 private test images. To validate the generalization capacity of the models, the paper uses the public test dataset as the validation dataset, and the private test dataset is used as the test dataset in the experiment.

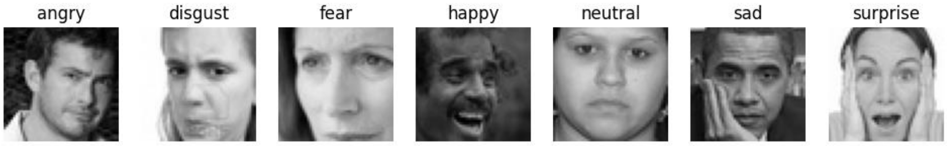

Figure 1. Example images in FER-2013.

As Figure 1 demonstrates, the images in FER-2013 vary a lot in illumination and the angle of the head pose, and the faces in the images may be occluded by hands or some other things, which makes it more difficult for models to extract and learn the features from the images. Besides what is mentioned above, this paper also finds that the dataset is very imbalanced as Figure 2 shows.

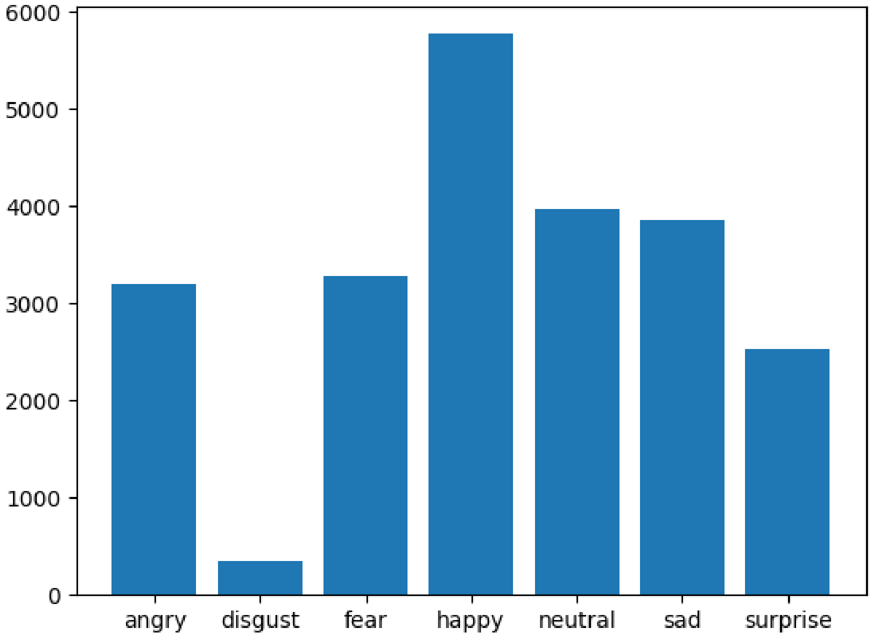

Figure 2. The imbalanced distribution of FER-2013 in the training dataset.

For preprocessing, all the images are normalized to have a pixel value between 0 and 1. Meanwhile, one-hot encoding is utilized to deal with categorical labels.

2.2. Proposed approach

Recognition models are built using VGGNet, ResNet, and MobileNet of different depths. Improvements are made to these models to adapt them to the FER-2013 dataset. Details of the modifications are shown in this section. Moreover, all models utilize softmax in the output layer. To accelerate convergence on the unbalanced dataset, the bias of the last layer is initialized before training in this paper. In addition to this, category weighting is also used. The weights are set to the inverse of the frequencies of classes in the training dataset.

2.2.1. VGGNet. VGGNet was proposed by Simonyan and Zisserman [11]. How the depth of the convolutional neural network affects its performance is evaluated and they found that as the model gets deeper the model tends to perform better. All the VGGNets in the paper utilize filters with size 3 × 3. Padding is used before all the convolutional layers for spatial resolution preservation. The stride of all the convolutions is fixed at 1. Max pool is implemented with pool size 2 and stride 2. In this paper, to fit the images with the size of 48 × 48, the architecture of VGGNet is modified. To reduce the size of the model, the paper utilizes a global average pooling layer to replace the final max pooling layer. Meanwhile, the fully connected layers with nodes of 4096 are removed.

2.2.2. ResNet. ResNet was brought forward by He et al. [8]. To solve the degradation problem, a deep residual learning framework was introduced by them. To make optimization easier, they utilized residual mapping, rather than an underlying mapping. The identity shortcut connection is utilized to help inputs skip layers. The architecture of ResNet resembles the architecture of VGG. Stacked convolutional layers with 3 × 3 filters can also be seen in ResNet. To fit the FER-2013 dataset, the first convolutional layer and max-pooling layer are replaced by a single convolutional layer with 64 3 × 3 filters. Besides, to preserve the image resolution, the stride of the first convolution is set to 1. Two types of building blocks are used in ResNet. The first one, say it is the basic block, consists of two convolutional layers just like the block in VGG, and an identity shortcut connection which is used to skip layers, the outputs of them are added up as the final outputs. The second one, bottleneck, consists of a stack of 3 layers with the kernel size of 1, 3, and 1. ResNet18 and ResNet34 utilize the basic block, while ResNet50 utilizes the bottleneck.

Based on the ResNet, the paper utilizes the SE mechanism. SE is a kind of channel attention mechanism. The squeeze operation aggerates the information from different channels, and the excitation operation tries to capture the channel-wise dependencies. A SE block consists of 1 global average pooling layer, 2 FC layers, and a scale activation. The activation function of the 2 FC layers is ReLU and Sigmoid [12]. SE blocks are placed after the residual blocks in the experiment.

2.2.3. MobileNet. MobileNet was proposed by Howard et al. Depth-wise separable convolutions are used by MobileNet to reduce parameters, which makes it possible to apply low-latency models in mobile applications. Compared to MobileNet, MobileNetV2 [10] utilizes inverted residuals and linear bottlenecks. The bottleneck with inverted residuals is similar to the original residual block in the ResNet. However, the original 3 ×3 convolution is replaced by a depth-wise convolution, reducing many parameters. Besides, the inverted residual block expands channels at first and then reduces channels. Meanwhile, it removes the ReLU at the end of the bottleneck to break the non-linearity, which was proven to be effective to preserve information in low-dimension. In addition, MobileNetV3 [11] redesigns the last stage of MobileNetv2 and introduces hard-swish as an activation function to reduce latency and improve accuracy. Meanwhile, the squeeze-and-excitation block is inserted into the model. To preserve the resolution and prevent the model from losing too much information early, the paper changes the stride of the first convolution layers to 1 in the MobileNets. The last convolution with stride 2 is also replaced by a convolution with stride 1.

2.2.4. Implement detail. All models use Adam as the optimizer to optimize the categorical cross-entropy. The learning rate is 3e-4 at first and decreased to one-tenth of the previous one if there has been no improvement in the validation accuracy for 4 epochs. Besides that, only the weights which achieve the highest validation accuracy are saved and used in the evaluation. The paper trains all the models for 50 epochs, and the mini-batch size is 64. From the experiment, it could be said that all of the models have become overfitting in the 50 epochs, so there is no need to continue training. To address the issue of FER-2013 being a relatively small dataset, this paper attempts various data augmentation methods, including rotation, horizontal flip, crop and resized, and erasing. All methods have a 50 percent probability of being used. Meanwhile, this paper implements several experiments based on the ResNet18 to test the effects of different data augmentation methods.

3. Result and discussion

The performance of the models is demonstrated and compared in this section. Based on the results, the factors contributing to performance improvement are discussed.

3.1. The model performance of the data augmentation

Without any other alternation, several data augmentation methods are introduced to the ResNet18 one by one to evaluate the effects of these methods. As Table 1 demonstrates, with the appliance of methods of data augmentation, the performance of the ResNet18 on the validation dataset is enhanced. After applying the random horizontal flip, the accuracy improves by about 3%, and around 6% accuracy is improved with the help of the random crop and resized. However, using more methods of data augmentation like random rotation and random erasing seems to make the performance worse. It may be caused by the conflict between different data augmentation methods, which leads to the problem that the patterns in the images are even harder for the models to capture.

Table 1. Validation accuracy of the ResNet18 with different data augmentation methods.

Index	Method	Validation Accuracy
1	None	60.33%
2	1+Random Horizontal Flip	63.97%
3	2+ Random Crop and Resized	70.10%
4	3+Random Rotation	67.04%
5	3+Random Erasing	69.32%

In the latter part of the paper, only random horizontal flip and random crop and resized are used in data augmentation.

3.2. The performance on test dataset

As Table 2 demonstrates, SE-ResNet34 achieves the highest private test accuracy of 70.80%. The results demonstrate that residual learning enables the models to go deeper without degradation since the private test accuracies of ResNets are higher than the private test accuracies of VGGNets. Meanwhile, the introduction of the SE block improves the performance of ResNet34 and ResNet50, improving their private test accuracies by 1.34% and 0.47%. The lightest neural network, MobileNetv3Small, gets the lowest private test accuracy of 61.66% and the lowest time cost of 751.9ms.

Table 2. Models and their Performances on FER-2013.

Model	Parameters	Train Accuracy	Validation Accuracy	Private Test Accuracy	Running time on Private Test
VGG11	9.2m	96.51%	68.21%	68.54%	786.7ms
VGG13	9.4m	98.62%	68.34%	69.23%	946.4ms
VGG16	14.7m	90.81%	68.40%	69.49%	1069.8ms
VGG19	20.0m	94.58%	68.40%	69.35%	1229.4ms
ResNet18	11.1m	95.48%	70.10%	70.10%	1528.9ms
ResNet34	21.2m	94.16%	68.43%	69.46%	2574.0ms
ResNet50	23.5m	85.21%	67.85%	69.55%	5227.1ms
Se-ResNet18	11.3m	97.84%	69.07%	70.08%	1655.6ms
Se-ResNet34	21.4m	97.60%	69.41%	70.80%	2844.3ms
Se-ResNet50	26.0m	96.44%	68.88%	70.02%	6620.7ms
MobileNetv2	2.3m	91.59%	65.31%	64.84%	1874.2ms
MobileNetv3Large	4.2m	93.17%	63.47%	63.56%	2167.0ms
MobileNetv3Small	1.8m	90.13%	60.37%	61.66%	751.9ms

In summary, according to Table 1, it can be learned that data augmentation methods including random horizontal flip and random crop and resized could effectively improve the performance of models. This is because that data augmentation helps to enlarge the small data set which is beneficial for the models to learn the features better. Meanwhile, from

Table **2**, Se-ResNet34 performs best in the experiment with 70.80% private test. The paper thinks that the residual learning enables the model to go deeper without degradation and the SE block is of benefit for the model to learn the global information. MobileNets uses the least parameters compared with other models, but their accuracies suffer from the simplicity of the model parameters.

4. Conclusion

This experiment is conducted to analyze the performance of multiple deep neural networks and algorithms on the FER-2013 dataset to understand which mechanism matters in the facial recognition task. Data augmentation is utilized to tackle the problem caused by the relatively small and imbalanced dataset. VGGNet, ResNet, and MobileNet are evaluated in the paper to analyze the effects of residual learning and the SE mechanism in facial expression recognition. In the paper, the models are evaluated based on experiments on the FER-2013 dataset. The SE-ResNet34 achieves the highest private test accuracy of 70.80% among all the models on the FER-2013 dataset according to the experimental results. It can be inferred from the experiment that the SE block does improve the model performance in the facial recognition task by enabling models to learn global information and selectively emphasize the important features. Although the accuracies of MobileNets are not as good as other complex models, the lightweight advantage may enable them to be applied to mobile devices.

References

[1] Li S and Deng W 2020 Deep facial expression recognition: A survey IEEE transactions on affective computing 13(3): pp 1195-215

[2] Ekman P and Friesen W 1978 Facial action coding system Environmental Psychology & Nonverbal Behavior

[3] Khorrami P Paine T and Huang T 2015 Do deep neural networks learn facial action units when doing expression recognition? IEEE International Conference on computer vision workshops (CVPR) IEEE pp 19-27

[4] Goodfellow I et al 2013 Challenges in representation learning: A report on three machine learning contest Neural Information Processing: 20th International Conference (ICONIP) Springer berlin heidelberg pp 117-124

[5] Tang Y 2013 Deep learning using linear support vector machines arXiv Preprint arXiv:1306.0239

[6] Minaee S Minaei M and Abdolrashidi A 2021 Deep-emotion: Facial expression recognition using attentional convolutional network Sensors 21(9): pp 3046

[7] Vulpe-Grigoraşi A and Grigore O 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE) IEEE pp 1-5

[8] He K Zhang X Ren S and Sun J 2016 Deep residual learning for image recognition IEEE Conference on computer vision and pattern recognition (CVPR) IEEE pp 770-778

[9] Sandler M Howard A Zhu M Zhmoginov A and Chen L 2018 Mobilenetv2: Inverted residuals and linear bottlenecks IEEE Conference on computer vision and pattern recognition (CVPR) IEEE pp 4510-4520

[10] Howard A et al 2019 Searching for mobilenetv3 IEEE International Conference on computer vision (CVPR) IEEE pp 1314-1324

[11] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv Preprint arXiv:1409.1556

[12] Hu J Shen L and Sun G 2018 Squeeze-and-excitation networks IEEE Conference on computer vision and pattern recognition (CVPR) IEEE pp 7132-7141