Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220657

Human Facial Expression Recognition based on Attention Mechanism

Zezhong Yang[1]

[1]University of Washington, Seattle WA 98195, USA

yangz63@uw.edu

Abstract. In the current study of facial expression recognition, CNNs have been extensively used and proven to be useful models. In this work, we analyze and improve VGG-16 network by embedding SENet between the convolutional layers and three fully connected layers of the original network to form the SE-VGG-16 network. The proposed network architecture is able to extract additional spatial information regarding interdependencies among feature channels, which is complementary to facial expression recognition. The experiment is performed on dataset FER2013 and improve the accuracy with top-1 error 33.5.

Keywords: Facial Expression Recognition, VGG, Image processing, visual attention.

1. Introduction

The face is the most expressive and informative of a human being, and the facial expressions have significant meaning in the understanding of human emotions and psychological activities. The recognition task has significant research value and a wide range of practice applications, including robot control, human-computer interaction, and health monitoring system. Some prior research works from the traditional approach, such as SVM [1], but the performance is limited by the representational power of the mathematical model. Due to the advancement in deep learning, CNNs has been extensively used in recognition task and favored for their high accuracy. Today, the study of facial expressions with neural networks has become the main focus. For instance, a sequential learning network is introduced in Lv et al.[2] firstly detects face expression-related components and classifies parsing components by a stacked autoencoder. In Lopes et al.' work [3], a combination of Convolutional Neural Network is used, and a preprocessing operation is proposed to extract specific facial expressive features and explore the presentation order of those features to improve the robustness of the expression recognition.

In this work, we aim to improve the performance of facial expression recognition by capturing the interdependencies among feature channels. Those interdependencies are exploited by introducing the concept of attention mechanism, which intentionally biases the allocation of computational resources toward the most informative components of the signal. The SE-VGG-16 network (as shown in figure 1) is introduced for this task. This network is based on VGG-16 [4] network for facial expression feature extraction and embeds the SENet [5] into the VGG-16 network to model the channel-wise relationships and enhance the representational power of the network. First, a parallel module to embedding after the original convolutional layers of the VGG-16 network to study the channel-wise relationship as weights and use them to rescale to original feature maps. It contains an additional average pooling inserted after the original convolutional layers of the VGG-16 network. This additional computation is used to obtain

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2/20220657

global spatial information of each feature map. Then, the additional module containing two fully connected layers, one ReLU function, and one Sigmoid function at the end, is embedded after the average pooling to study the weights and restrain the values, which ensures the multiple channels are allowed to be emphasized. Last, the original feature channels are rescaled by the weights obtained from this module, allowing specific channels that are conducive to facial expression recognition to be emphasized in the network, thereby improving the performance of the network model.

2. Method

To improve the accuracy of Facial Expression Recognition, the convolutional neural network SE-VGG is introduced based on the original VGG architecture incorporated with attention mechanisms to further improve the prediction accuracy. Also, the experiment is performed on the FER2013 dataset [6].

2.1. VGG Network

Table 1. Structure of VGG[4]. The Network consists of 16 weighted layers with 6 different configurations shown in each column.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

The VGG module was introduced in 2014 by Simonvan and Zisserman [4] after LeNet [7] and AlexNet [8]. This module provides an inspirational module structure for later Network such as GoogleNet [9] and ResNet [10]. Unlike previous modules, VGG is designed based on the concept of blocks. The architecture of the VGG module is shown in Figure 1. Each block is consisted of several convolutional layers and a maximum pooling to reduce the input dimension passed to the next block. The convolutional layers in VGG have a size of 3x3 instead of 7x7 in the AlexNet and pooling has the size of 2x2 instead

DOI: 10.54254/2755-2721/2/20220657

of size 3x3 in the AlexNet in order to increase the overall depth of the neural network while maintaining the same effective receptive field as the AlexNet. This allows the VGG module have better performance over multinomial logistic regression task as described in Simonvan and Zisserman's work[4].

2.2. Attention Mechanism

The core of convolutional neural networks is the convolution operator in which enables networks to extract spatial and channel-related information. To strengthen the representation power of a CNN, Hu et al [5] propose squeeze and excitation network that introduces a compact module to exploit the interdependencies between channels. This module contains three components – squeeze, excitation, scaling (see figure 1). It is a compact module embedded into another CNN Architecture. With input X, the transformation Ftr refers to the original convolutional layers of the original CNN. The squeeze function Fsq takes the output features U and reduces the spatial dimensions H and W to 1 x 1 using global average pooling to extract channel information. The excitation function Fex consists of one fully connected layer of dimension 1x1xC/r, one ReLU activation function, one fully connected layer of dimension 1x1xC, and one Sigmoid activation function (see figure 2). Two fully connected layers are used here to model the gating mechanism [5] that captures the channel-wise dependencies and sigmoid is used to ensure that multiple channels are allowed to be emphasized. The scaling Fscale gives the final output obtained by rescaling features U with the channel weights obtained from the excitation function. This compact module provides additional attention and gating mechanism in which enhances important features extracted from relevant channels and suppresses features coming from other channels.

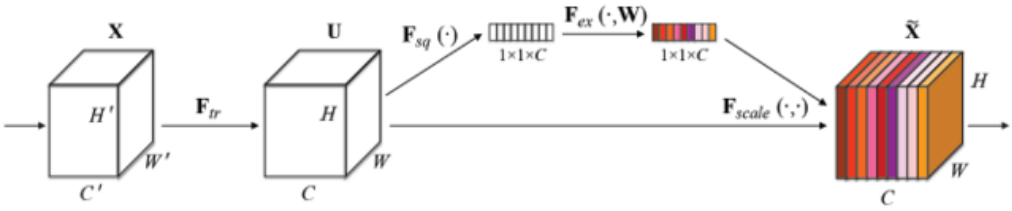

Figure 1. Structure of SENet [4]. The Network consists of 3 components - Fsq, Fex, Fscale.

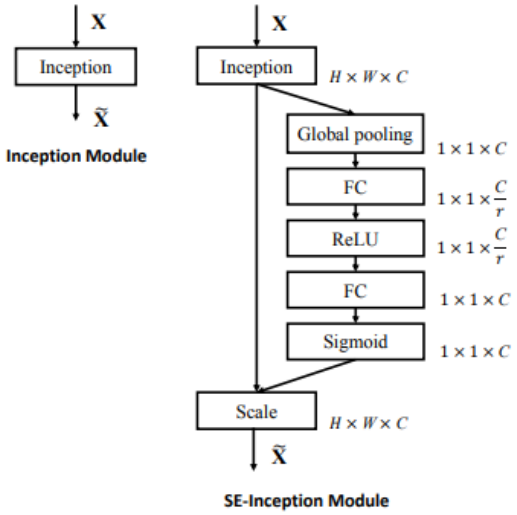

Figure 2. Schema of SENet [5]. The schema of the original Inception module (left) and the SEInception module (right).

DOI: 10.54254/2755-2721/2/20220657

2.3. Attention-based VGG Network

In this work, we propose the SE-VGG-16 Network modified from VGG-16 Network with the attention mechanism introduced. The original 16 weighted layers in VGG-16 Network are preserved to capture the hierarchical spatial information. But the additional parallel SENet module are embedded after 13 convolutional layers in the original VGG-16 Network. This parallel module computes the weights that represents channel-wise interdependencies, rescales the extracted feature outputs from the original convolutional layers in the VGG-16 Network channel-wisely, and passes the modified features to the original fully connected layers. This modification biases the channel information and enhances the representational power of the neural network. In the experiment, we use the Adam optimizer with a learning rate 0.0001 and 60 train epochs. To speed up the experiment, the system is equipped with NVIDIA 2080 TI GPUs with 32 GB DDR4 SDRAM.

3. Results and Discussion

3.1. Dataset description

The experiment is performed on the FER2013 dataset [6] created using Google image search API by Pierre Luc Carrier and Aaron Courville. It contains 35887 48x48 pixel grayscale images of faces, with 4953 “Anger” images, 547 “Disgust” images, 5121 “Fear” images, 8989 “Happiness” images, 6077 “Sadness” images, 4002 “Surprise” images, and 6198 “Neutral” images. Total of 35877 images are split into a training set (28709 examples) and a public test set (3589 examples). Also, seven emotions are also categorized numerically (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). In the experiment, the SE-VGG is trained on the FER2013 training set, and performance is tested on the FER2013 public test set.

3.2. Comparison results

Table 2. Top-1 error on FER2013 validation set and complexity comparison. The left column refers to two network we compared, and the right column refers to the top-1 error measured over the FER2013 validation set.

	Top-1 err.
VGG-16 [3]	**36.1**
SE-VGG-16	**33.5**

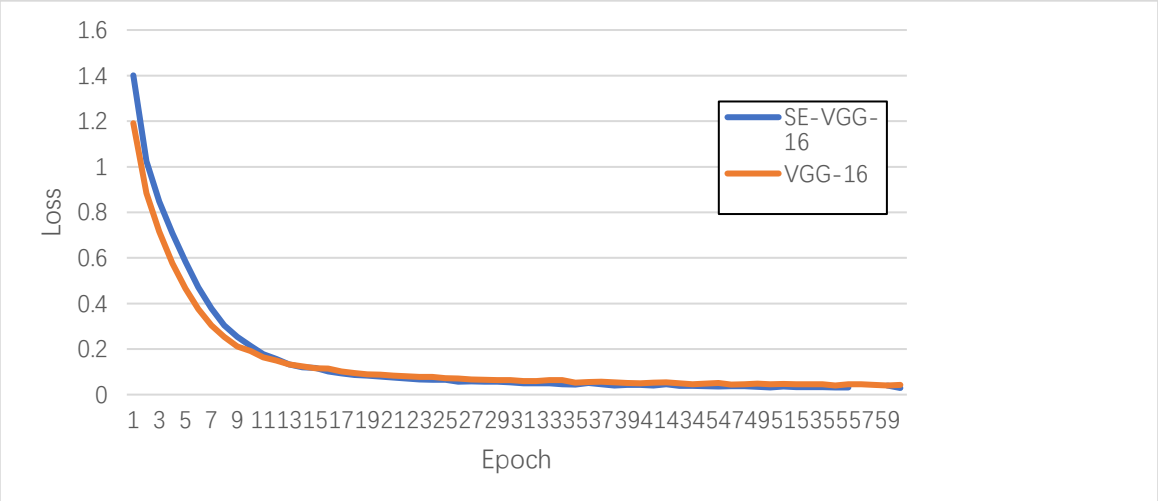

Figure 3. Training loss collected on FER2013 training set. The SE-VGG-16 exhibits an improved performance over the training process.

DOI: 10.54254/2755-2721/2/20220657

To evaluate the SE-VGG-16 Network, we compare the SE-VGG-16 Network against original VGG-16 architectures. We observe an improved performance in SE-VGG-16 Network with a top-1 error of 33.5%, exceeding VGG-16 (36.1%) by 2.6% (shown in table 2). Though the SE-VGG-16 Network with additional SE blocks implemented increases depth in architecture, it costs in an extremely computationally efficient manner and yields faster convergent speed measured over the FER2013 training set (figure 3). This suggests that the additional SE block introduced into VGG-16 enhances the performance on the facial expression recognition task and could be implemented in other convolutional networks to improve their performance in the facial expression recognition task.

4. Conclusion

In this work, we aim to introduce the visual attention mechanism to obtain a better target representation for facial expression recognition. In specific, we embed the SENet into the popular VGGNet, namely SE-VGG-16. And then, we evaluate the performance of the SE-VGG-16 Network for the Facial Expression Recognition task. In FER2013, the VGG network with SENet achieves a 2.5% decrease on top-1 error, which demonstrates the improvement in accuracy of the VGG 16 Network with the additional attention mechanism implemented, and the enhancement in recognition task on the FER2013 dataset. Our work again confirms the representational power of the SE-VGG-16 Network on the Facial Expression Recognition task.

References

[1] M. Abdulrahman and A. Eleyan, "Facial expression recognition using Support Vector Machines," 2015 23nd Signal Processing and Communications Applications Conference (SIU), 2015, pp. 276-279, doi: 10.1109/SIU.2015.7129813.

[2] Y. Lv, Z. Feng and C. Xu, "Facial expression recognition via deep learning," 2014 International Conference on Smart Computing, 2014, pp. 303-308, doi: 10.1109/SMARTCOMP.2014.7043872.

[3] Lopes A T, De Aguiar E, De Souza A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order [J]. Pattern Recognition, 2017, 61: 610-628.

[4] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[5] Hu, J., Shen, L., & Sun, G. Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 2017.

[6] Goodfellow, I.J., Erhan, D., Carrier, P.L., et al. Challenges in Representation Learning: A report on three machine learning contests. arXiv preprint arXiv: 1307.0414, 2013

[7] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[9] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.