Rethinking VGG19 and ResNet Architecture For Facial Expression Recognition

Jiayan Kong
University of Toronto
Toronto, ON
jiayan.kong@mail.utoronto.ca

Xiang Chen
University of Toronto
Toronto, ON
legendxiang.chen@mail.utoronto.ca

Keyi Zhang
University of Toronto
Toronto, ON
kya.zhang@mail.utoronto.ca

Abstract

Facial expression recognition has attracted the attention of researchers in computer science and many related disciplines and has been widely used in practical applications. In this article, we reproduce two types of deep neural networks that are widely used in the facial expression recognition field, ResNet and VGG, and how they gain an advantage over the other. We train and test our model on datasets recombined from RAF Database and FER-2013. Then we tune the hyperparameters of the models and achieved accuracies of more than 70%. However, complex models do not generalize as well as simpler models in specific problems given a fixed dataset as we found out for the facial emotion recognition task, ResNet09 achieve higher accuracy than ResNet18 in classifying all 7 emotions. Distribution of labeled training data will affect the bias for test accuracy in 7 classes as our experiment results support.

1 Introduction

Facial expression is a key signal of human emotion (Cui et al., 2020). More and more advanced deep learning models have been introduced since the early 21-st century. Research on facial expression recognition in recent years focuses on deep neural networks, such as VGGNet and ResNet, to obtain powerful representations and classifications. After we have gone through Very Deep Convolutional Networks for Large-Scale Image Recognition (Simonyan and Zisserman, 2014), Deep Residual Learning for Image Recognition (He et al., 2016a) and Facial Expression Recognition Based on VGGNet Convolutional Neural Network (Jun et al., 2018), we find that there have been lots of exploration based on these two models in this area; therefore, we focus on the specific variants of VGGNet and ResNet and reproduce them in this paper.

With work and experiments from previous researchers, we get a rough idea that the growth of complexity in the same series of models usually increases the accuracy, i.e. from VGG16 to VGG19. However, researchers also show that it is not always the case, different tasks require different models with appropriate complexity given the size of training dataset. For the facial expression recognition task, we would like to reproduce ImageNets of VGG19, ResNet09, and ResNet18 and further evaluate these three models in multi-aspects including performance, efficiency, classification bias, and hyper-parameters in this paper. Besides the comparison among models, how dataset and model affect the accuracy on different test label classes is another topic covered in our analysis.

2 Background and Related Works

We collected, analyzed some papers from arXiv.org, NeurIPS and IEEE that released a novel and upgraded implementation of CNN focusing on facial expression recognition. We want to make a comparison between two world-leading technologies in facial expression recognition that have some different points of focus but all improve the FER model in terms of boosting accuracy.

VGG network. The VGG network architecture was first introduced in the paper Very Deep Convolutional Networks for Large-Scale Image Recognition (Simonyan and Zisserman, 2014). The feature of VGGNet is that it uses all 3x3 convolution kernels and 2x2 pooling layers and improves performance by deepening the network structure. It has outstanding performance in applications such as extracting image features, while it takes a long time to train due to its number of fully connected nodes and depth.

ResNet network. In 2016, He et al. first introduced the ResNet network architecture in the paper Deep Residual Learning for Image Recognition. In the paper Identity mappings in deep residual networks (He et al., 2016b) published subsequently, the author analyzed the mathematical principles behind the residual module, then updated the residual module and used identity mappings to make the network easier to train and improve the accuracy.

3 Proposed Method

Data processing. We have used several datasets that are widely used in the facial expression recognition field to train our model. These datasets contain the seven basic emotion categories that we intend to study: surprise, fear, disgust, happy, sad, anger, and neutral. To further improve the performance of our model, we manually process the dataset, mix the data from different datasets to generate a new data set, to reduce the influences caused by the difference between datasets. Besides, we also control the number of images for each type of emotion category in our dataset to ensure that there will be no special bias towards some kinds of emotions during our training process.

Head Pose Estimation. We have investigated several methods to achieve head pose estimation, such as the build-in method, Cascade Classifier, in OpenCV. We found the traditional computer vision method can not provide us with satisfactory accuracy. To further improve the accuracy, we use the technique of combining deep neural nets with the traditional computer vision method, which is Multi-task Cascaded Convolutional Networks(Zhang et al., 2016). It has three stages. In the first stage, it will exploit a fully convolutional network called Proposal Network to obtain the candidate windows and their bounding box regression vectors. In the second stage, it will feed them to another convolutional network called Refine Network to further reject a large number of false candidates. In the third stage, it will extract five facial landmarks.

One Cycle Learning Rate. We use a popular method in the modern industry to set the learning rate during training, which is 1 Cycle Policy. This policy can give us very fast results when training complex models. It sets up based on the Cyclical Learning Rate to get faster training time with regularization effect. Previous work (Smith and Topin, 2019) proposes that a cycle with two steps of equal lengths, first step going from lower learning rate to higher before second step to back to the minimum, and the optimal learning rate should follow the following calculation. We have seen that the loss function could be written as the following.

$$f(\theta_i - \epsilon \nabla_\theta f(\theta_i)) \approx f(\theta_i) + (\theta_{i+1} - \theta_i)^T \nabla_\theta f(\theta_i) + \frac{1}{2}(\theta_{i+1} - \theta_i)^T H(\theta_{i+1} - \theta_i)$$

After the calculation, we can get the optimal learning rate as the following.

$$\epsilon^* = \epsilon \frac{\theta_{i+1} - \theta_i}{2\theta_{i+1} - \theta_i - \theta_{i+2}}$$

, where $\theta \in R^N$ represents the parameters of the neural net.

Bias Comparison. We want to find whether or not each model has a bias performance for one or more specific expression identifications. Therefore, we set seven different test sets, each expression for one set, and we run our models on these sets to test their performance on specific

expressions. We also investigated the bias due to different training sets. Therefore, we use multi data sets to train and test the final performance for each model.

4 Experiments

Databases. We consider two datasets generated from two benchmarks: dataset1 consists of 10374 images from Real-world Affective Faces (RAF) Database (Li et al., 2017), dataset2 has 13000 images from the combination of RAF Database and part of the Facial Expression Recognition 2013 (FER-2013) Dataset (Goodfellow et al., 2013). Both datasets have 7 emotion categories as {0:Surprise, 1:Fear, 2:Disgust, 3:Happy, 4:Sad, 5:Anger, 6:Neutral}. Images from the RAF are cropped manually to the size of 48×48 with the help of MTCNN (Zhang et al., 2016). As our method requires a more equally distributed training data on each emotion category, some extra training images are taken from FER-2013 to form the second dataset. As shown in Table 1, though we have combined all images for emotion fear from the training examples of the two datasets, the total number 1087 is still relatively low compared to other categories. The lack of information for a specific training class is a very common weakness among different datasets of facial expression; however, our experiment results show that other factors like the similarity of images between two categories have a stronger impact on the accuracy.

Table 1: Training Datasets Information

Name	Surprise	Fear	Disgust	Happy	Sad	Anger	Neutral
dataset1	1329	356	651	3452	1781	842	1963
dataset2	2000	2000	1000	2000	2000	2000	2000

Models and hyperparameters. VGG19, ResNet09, ResNet18 models are applied on both datasets. An average pooling layer of kernel and stride size 1 is added at the end of the VGG19 model as well as a linear classifier layer that has 512 input channels and 7 output channels. The ResNet09 and ResNet18 models all have the same linear layer as VGG19 to classify the 7 emotion classes. All models are applied with the cyclical learning rate (Smith, 2017). The training epochs are 25, 30 and 35 for VGG19, ResNet09 and ResNet18 respectively.

4.1 Comparison of Model Accuracy with Different Datasets

The experiment results in Figure 1 and Table 2 show that for both datasets, ResNet09 achieves comparable performance with VGG19 in final validation accuracy and loss while both of them outperform ResNet18 by $6\%\sim8\%$ in accuracy unexpectedly. Although the increased model size and computational cost tend to translate to immediate quality gains for most tasks (Szegedy et al., 2016), it is reasonable to state that the limited number of labeled training data is impeding the performance of ResNet18, causing a relatively low validation accuracy. In terms of computational cost, when the batch size and epoch number are fixed, both ResNets train apparently faster than the VGGNet. Our experiments results demonstrate that for the task of facial emotion recognition, if the given dataset is not large enough, i.e. contains around 10,000 to 15,000 number of training data, ResNet trains faster than VGG19 with significantly less computational cost and ResNet09 outweighs the other two models considering aspects like performance and efficiency.

Table 2: Validation Accuracy and Loss

Name	VGG19	ResNet09	ResNet18
dataset1: Accuracy	0.7921	0.8037	0.7214
dataset1: Loss	0.8069	0.7338	0.9285
dataset2: Accuracy	0.7802	0.7808	0.7380
dataset2: Loss	1.0157	0.8985	1.0802

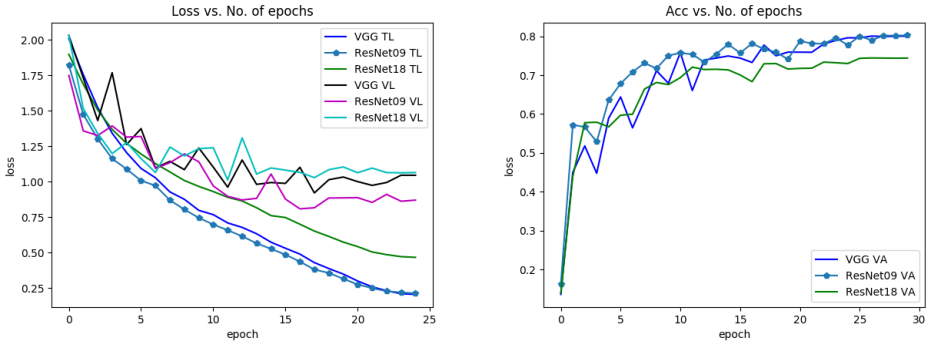

Figure 1: Plot of Loss and Accuracy

4.2 Bias Comparison

To evaluate if equally distributed data would affect the bias, we can see from Table 3 with corresponding columns from dataset1 to the one in dataset2. From dataset1 to dataset2, the dataset of images for all classes are expanded to 2000 images except the Happy category and 1452 images are removed from it. As a result, all models experience a decrease in classifying happy faces by 12%∼13% and the accuracy for this class also leans toward the final average validation accuracy on all 7 categories. Meanwhile, all three models expose an increasing accuracy for 1%∼3% in Surprise, Fear, Anger and Neutral. Nearly 72% of the 7 emotion classes are classified with a smaller bias due to the change from dataset1 to dataset2. This would support our conjecture that an equally distributed dataset would reduce the bias in classifying different labels.

Table 3: Test Accuracy On 7 Categories

	Dataset1			Dataset2		
Category	VGG19	ResNet09	ResNet18	VGG19	ResNet09	ResNet18
Surprise	0.8567	0.8687	0.7542	0.8629	0.8804	0.7829
Fear	0.2931	0.3190	0.2931	0.3621	0.3448	0.3190
Disgust	0.4127	0.4603	0.3757	0.4497	0.4286	0.4021
Happy	0.9033	0.9200	0.8817	0.7650	0.7842	0.7633
Sad	0.7497	0.7460	0.6745	0.7411	0.7643	0.6498
Anger	0.7303	0.7191	0.6236	0.7640	0.7472	0.6685
Neutral	0.9214	0.8996	0.8166	0.9301	0.9127	0.8472

However, given the same dataset, all three models display an analogous accuracy difference in each label class which indicates that the three model does not have any sensitivity towards a specific class. The cause for the low accuracy in classifying Fear and Disgust is that the features for the two emotions are not very distinguishable when we look through the training data. Besides, lack of training images for Disgust increase the difficulty of model's training.

5 Conclusion

In this paper, we have applied three deep learning models, ResNet09, ResNet18 and VGG19, to the task of facial expression recognition. We used the recombined datasets from RAF Database and FER-2013 to train our model and tuned the hyperparameters and the experiments results support our conjecture that an equally distributed dataset would reduce the bias in classifying different labels. Finally, we realized ResNet09, ResNet18, VGG19 models with 79%, 80%, and 73% validation accuracy. And for different emotion categories, the recognition results of the three models for fear and disgust emotions are significantly worse than other categories. For future work, we hope to build datasets with larger sample size and more equally distributed data in each emotion category to improve the bias problem.

Contributions

We believe that the three authors have made many contributions in establishing models, conducting experiments, and writing reports. In more detail, Xiang Chen has taken the main responsibility for model implementation; Keyi Zhang built and trained the dataset and created some experiments; Jiayan Kong implement model prediction and visualizing experimental results.

References

Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. *Advances in Neural Information Processing Systems*, 33, 2020.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

He Jun, Li Shuai, Shen Jinming, Liu Yue, Wang Jingwei, and Jin Peng. Facial expression recognition based on vggnet convolutional neural network. In *2018 Chinese Automation Congress (CAC)*, pages 4146–4151. IEEE, 2018.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.