Facial Expression Recognition with Convolutional Neural Networks via a Data Augmentation Strategy

Binbin Yan; Zhengyu Xiao; Pengxuan Yuan; Kanjia Cai; Qian Chen

School of Informatics Xiamen University
No. 422, Siming South Road, Siming District
Xiamen, Fujian 361000

Abstract

It is a common sense that human express emotions and tend to convey emotions through facial expression. Accordingly, facial expression recognition (FER) has long been a popular field attracts countless scholars. However, there exits over-fitting problem caused by insufficient training data, which is an obstacle resulting in low accuracy. Recently, deep learning technology has shown excellent performance in various fields, based on this fact, deep neural networks have increasingly been leveraged to learn discriminative representations for FER task. We build an available and efficient deep FER system on the strength of convolutional neural network (CNN). By using the generative adversarial networks (GANs) to augment dataset to address the over-fitting problem.

Introduction

Humans express emotions through facial expressions. This is a very important nonverbal communication method that directly reflects inner emotions. There are six common facial expressions like happiness, surprise, sadness, anger, disgust and fear, as well as many compound expressions and micro-expressions.

Facial expression recognition aims to classify facial expressions with given types. Actually, the complexity and subtlety of human facial expressions makes it difficult to classify. However, one thing that cannot be ignored is that the technology of facial expression recognition is an important component of human-computer interaction. It is used in social interaction research, intelligent control, Medical, communication and other fields which are promising, like psychological research in the field of social sciences, user sentiment analysis in marketing, online social games, as well as other features that include human-computer interaction, which can all benefit from the automatic recognition of facial expressions.

Recent years, deep learning methods have been widely used in this recognition task, as for the difficulty of technology and the current accuracy, facial expression recognition can only be used for scenes that do not require high accuracy, and do some aided analysis work. A recent survey paper (Li and Deng 2020) has shown the success of CNN on the FER-2013 dataset.

Used in the ICML Workshop "Representational Learning Challenge" for the Facial Expressions Recognition Challenge, the FER-2013 dataset are created by Google Image Search API. It contains 35,887 images with seven emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutrality. James Bergstra also identified an "empty" model consisting of CNNs. There is no learning process except the final classifier layer. Human accuracy on FER-2013 is 65.5%±5% (Goodfellow et al. 2015). The best performing model in the challenge competition has reached an accuracy of 71.162%. The model is trained with a support vector machine (SVM) loss function, and also uses the L2-SVM loss function, which achieves great results on the FER-2013 dataset and other datasets.

The performance of the algorithm for extracting handcrafted features is up to 68%, though the algorithm for automatic feature learning using deep networks is ahead of handcrafted features, it is unclear whether the performance of the optimal deep network reaches the observable probability in the distribution of this task. A method that combines automatic features learned by CNN with handcrafted features computed by bag-of-visual-words (BOVW) model and take a local learning framework as prediction way (Georgescu, Ionescu, and Popescu 2019) has got a maximum accuracy of 75.42% on FER-2013. In order to achieve better performance on this task, to address the problem of insufficient training data and to remove the effects of expression-independent changes (e.g., head pose, lighting conditions), in this paper, we propose to use GANs for data augmentation, filter irrelevant factors through cascading networks, and attempt network integration to provide the diversity of feature learning with the combination of multiple networks.

Related Work

There have been some studies using CNNs to accomplish recognition task, and some solutions have been proposed for overfitting problems caused by lack of sufficient training data and irrelevant changes in expression (such as lighting, head posture).

Diverse network input

Traditional methods usually use RGB images as the entire input of the network, but these raw data lack some invariances such as image scaling, rotation, occlusion, and light-

ing. Therefore, some methods use handcrafted features as network input to enhance robustness and force the network to pay more attention to facial regions with expression information. Aiming at the invariance of lighting, Levi (Levi and Hassner 2015) and Hassner proposed a novel mapped local binary pattern(LBP) feature. For multi-view FER tasks, scale-invariant feature transform (SIFT), which is robust to image scaling and rotation, can be used.

Generative adversarial networks

Generative adversarial networks have been successfully used in image synthesis. J Zhao (Zhao et al. 2018) et al. proposed a GAN-based face frontalization framework, and Zhang (Zhang et al. 2018) et al. proposed to generate images of different expressions in arbitrary poses for multi-view FER. Yang et al. (Yang, Zhang, and Yin 2018) proposed an identity-adaptive generation (IA-gen) model with two parts, which can well mitigate identity changes.

Cascaded networks

The cascaded network of different structure combinations can learn the hierarchical structure of features. Using this hierarchical structure, you can gradually filter out the variation factors that have nothing to do with the expression. Rifai (Rifai et al. 2012) et al. proposed a multi-scale compressed convolutional network (CCNET) and a retractable autoencoder to separate emotion-related factors from identity and posture. M.Liu et al. (Liu et al. 2013) used the CNN structure to learn incomplete representations, and then used multiple layers of restricted boltzmann machines(RBM) to learn the higher-level features of FER. Instead of simply connecting the outputs of different networks in series, P.Liu et al. (Liu et al. 2014) used a boosted DBN (BDBN), which can iteratively perform feature representation, feature selection and classifier construction in a single loop state, so the ability to discriminate FER is substantially improved.

Approach

Data augmentation by GANs

At present, GANs have made great progress in image generation, and have been able to generate realistic face images. Considering the insufficiency of data in FER-2013, we can use a proper GAN model to generate more expression images and add them into the training set to reduce the overfitting problem.

Since the FER-2013 dataset is not all front faces, there are many pictures with different face angles. Although ck+ has only 981 pictures, the face images are basically front faces, and the facial poses are more uniform. So the quality of facial expression pictures generated by training GAN will be better.

The discriminator and generator of Deep convolutional GAN (DCGAN) use CNN to replace the multi-layer perceptron in GAN, which can generate high-quality images and can learn the semantic information of the image. We use DCGAN to generate facial expression images. However, DCGAN has problems such as training difficulty and the loss of generators and discriminators cannot indicate the training process. Therefore, we also use the improved model of DCGAN called Wasserstein GAN with gradient penalty (WGAN-GP) to solve the collapse mode problem and ensure the variety of generated samples. The model has a value to indicate the progress of the training. The smaller the value is, the better the GAN training, so higher quality images can be obtained.

Since DCGAN and WGAN-GP are both unsupervised learning, their input cannot control the output. Considering that we need to generate different labels, such as angry, frustrated and other facial expressions, we can only do it by training data of each label separately, which means we need to build seven GANs and train them. In order to achieve controlling the output, we use conditional GAN (CGAN), which adds label restrictions to the input on the basis of DCGAN. In addition, we also introduce the gradient penalty of WGAN-GP into CGAN.

Learning model

The first idea we adopt is applying classic machine learning model, support vector machine (SVM) , which have achieved great success in image classification to FER problem.

In the field of deep learning for image classification such as face recognition, the VGG model is used more often. Its characteristics are deep network and small convolution kernel. The depth of the layer makes the feature map wider, which is suitable for large datasets and solving multi-classification problems. Therefore, we use VGG-16 for facial expression recognition tasks. In addition, considering that the FER-2013 dataset is not very large, we also implement a residual network (ResNET) to train a deeper network.

Experiment

A. Image generation

The model structure of DCGAN is as mentioned in Radford's article (Radford, Metz, and Chintala 2015). In training process, we set the batch-size to 16, and trains a total of 80 epochs. WGAN-GP is in the same way as DCGAN generates images. But the batch-size is set to 8, and trained 200 epochs. In the whole training process we set learning rate to 10^{-4}, use ADAM as optimizer. The implementations of CGAN are same with the other two GANs, only the batch-size is set to 64.

We use images of different expressions in the ck+ dataset as training data, that is, one expression corresponds to one GAN, and seven GANs are used to generate seven facial expression images. We generated a total of 7000 pictures, of which each GAN generated 1000. The generated face images shown in Figure 1.

B. Facial expression recognition

DF-SVM We employ VGG-13, VGG-16 and VGG-19 as pre-trained models to extract deep features for SVM (DF-SVM) , which means we didn't train the whole model from scratch. In order to extract effective deep features and prevent overfitting for the latter use, we observe the loss variation and accuracy variation of each model on the test dataset

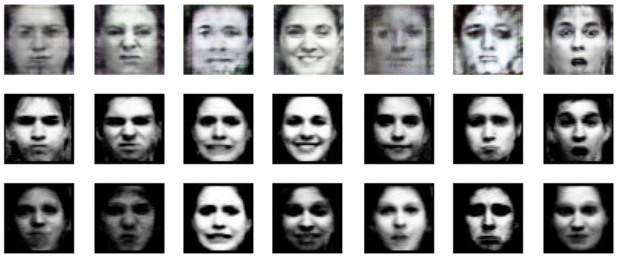

Figure 1: Face images generated by three GANs. DCGAN, WGAN-GP and CGAN from top line to bottom. The images quality of WGAN-GP is the best, but CGAN is the worst because its generated images lack of diversity, they almost look like same one.

and fine-tune the models to improve performance and capability of generalization. Adopt the idea of ensemble learning, we additionally combine all the deep feature we get from pre-trained models and use them as input of another classifier. To extract deep features, we remove the softmax layer of each model, take activation map of last remaining fully connected layers of three models and combine them together into a deep feature vector as input to linear SVM. In order to avoid overfitting problem and maintain the consistency of three deep features, before feeding the vector into the classifier, we normalized the vector using L2-norm. In classifier model, we use default parameters of SVM model from sklearn liabrary.

VGG-16 We use the VGG-16 model proposed by Simonyan and Zisserman in their paper for expression recognition (Simonyan and Zisserman 2014), and use data augmentation methods such as rotation and translation. The batchsize is 32, and train 200 epochs. In the whole training process we set learning rate to 10^{-4}, use SGD as optimizer.

ResNET-50 The ResNET-50 model is proposed by K He et al (He et al. 2016). The other parameters setting are same as VGG-16.

C. Implementation details

Three VGGs for features extraction The VGG-13 architecture is composed of 10 convolution layers and 3 fully connected layers and is pre-trained on ImageNet data set image size of which is different from the FER-2013 dataset we use in this FER experiment. Therefore, we scale all the images in FER-2013 dataset into the same size as ImageNet. Noting that VGG-13 is not particularly designed for FER problem, we keep the weights frozn in first 8 convolution layers because we think the first several layers of the model can still extract effective facial feature which we can use for expression recognition. From the result of VGG-13, we found that freezing weights contributes to the performance of model finitely and VGG-16 is a much deeper network than VGG-13 which consumes a lot of time and resource to fine-tune all the parameters in model. So we fixed the weights of all convolution layers in VGG-16. For VGG-19, We directly employ the model from keras library.

Model	**Dataset**	**Accuracy**
VGG13+VGG16+VGG19+SVM	FER-2013	52.40%
VGG-16	FER-2013	54.92%
ResNET-50	FER-2013	64.28%
VGG-16+DCGAN	FER-2013	66.37%
VGG-16+WGAN	FER-2013	64.61%
ResNET-50+DCGAN	FER-2013	**71.75%**
ResNET-50+WGAN	FER-2013	69.16%

Table 1: Accuracy on test set. ResNET classifier and data augmentation by DCGAN is the best one. Generating face images to augmente data is of help to improve the performance of VGG and ResNET on FER task.

For the rest layers, the weights are randomly normalized by drawing them from a Gaussian distribution with zero mean and 0.1 standard deviation. Then we replace the final softmax layer of 1000 units with a softmax layer of 7 units which represents 7 types of expressions. Besides, we add several dropout layers into the models attempting to improve the performance of the model in FER problem. In the whole re-training process we set learning rate to 10^{-3}, use SGD as optimizer and stop training when accuracy on validation set drops.

D. Results

Before using GANs for data augmentation, ResNET-50 can achieve the best accuracy of 64.28%. When generating face images, WGAN and DCGAN can generate images of better quality than CGAN, but we discarded CGAN's result because its effect is too poor. Among them, the image quality of WGAN is better than that of DCGAN, but the accuracy of the ResNET-50+DCGAN model in the results is better than WGAN, which is increased to 71.75%. Other experiments' results are shown in Table 1.

Conclusion

According to the purpose of the task, firstly, we use different neural networks to extract the features of the face images, fuse the features extracted from different networks, and finally classify with SVM; then we design two end-to-end network structures, based on VGG and ResNET. Finally, considering the irregular face poses of the FER-2013 dataset and partial label errors, we used GANs to generate 1000 frontal images for each label to reduce the impact of irregular images on the results.

After comparing and reflecting on the experimental results, we discussed the most worthy part of analysis and discussion in this experiment, that is, the reason for using poorer quality generated images by DCGAN for data augmentation but with better test accuracy. But we did not arrive at agree view. Because this experiment focuses on the task of facial expression recognition, the experiment on that issue has not been carried out, so we will not discuss too much here, but it is a topic worth continuing to study in the future.

References

Georgescu, M.-I.; Ionescu, R. T.; and Popescu, M. 2019. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access* 7: 64827–64836.

Goodfellow, I. J.; Erhan, D.; Luc Carrier, P.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D. H.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; Ionescu, R.; Popescu, M.; Grozea, C.; Bergstra, J.; Xie, J.; Romaszko, L.; Xu, B.; Chuang, Z.; and Bengio, Y. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks* 64(December 2017): 59–63. ISSN 18792782. doi: 10.1016/j.neunet.2014.09.005.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision Pattern Recognition*.

Levi, G.; and Hassner, T. 2015. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. In *the 2015 ACM*.

Li, S.; and Deng, W. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* .

Liu, M.; Li, S.; Shan, S.; and Chen, X. 2013. AU-aware Deep Networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*.

Liu, P.; Han, S.; Meng, Z.; and Tong, Y. 2014. Facial expression recognition via a boosted deep belief network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (January): 1805–1812. ISSN 10636919. doi:10.1109/CVPR.2014.233.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *Computer ence* .

Rifai, S.; Bengio, Y.; Courville, A.; Vincent, P.; and Mirza, M. 2012. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, 808–822. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer ence* .

Yang, H.; Zhang, Z.; and Yin, L. 2018. Identity-Adaptive Facial Expression Recognition through Expression Regeneration Using Conditional Generative Adversarial Networks. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*.

Zhang, F.; Zhang, T.; Mao, Q.; and Xu, C. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3359–3368.

Zhao, J.; Cheng, Y.; Xu, Y.; Xiong, L.; Li, J.; Zhao, F.; Jayashree, K.; Pranata, S.; Shen, S.; Xing, J.; et al. 2018. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2207–2216.