ECE 520.638 Deep Learning

Final Project Report

Project Title: Facial Expression detector

Team: Chongyu, Daijie Bao, Runtian Tang

Summary:
Mental health and old-age healthcare have been increasingly serious problems in the world; However, both problems are hard to detect and pose a great challenge to the government. Some research has shown that facial expression is an effective way to keep track of an individual's health. We built and trained a model pipeline using deep learning methods that can detect face images in pictures and classify various facial expressions. The model was trained on the AffectNet dataset [1] and achieved an accuracy of 50%.

Related work:
To study facial expression identification, Andrey trained several multi-task learning of lightweight convolutional neural networks on cropped faces without margins. These models are presented based on Mobile-Net, Efficient Net and RexNet architectures. It was experimentally demonstrated that they lead to near state-of-the-art results in age, gender and race recognition on the UTK-Face dataset and emotion classification on the Affect Net dataset.[2]

Kaipeng and Zhanpeng adopt a cascaded structure with three stages of carefully designed deep convolutional networks that predict face and landmark locations in a coarse-to-fine manner. In addition, in the learning process, they proposed a new online hard sample mining strategy that can improve performance automatically without manual sample selection. Their method achieves superior accuracy over the state-of-the-art techniques on the challenging FDDB and WIDER FACE benchmark for face detection, and AFLW benchmark for face alignment, while keeping real-time performance[3]

The classification models we choose are ResNet18[5] and VGG-16[6]. They are pre-trained on ImageNet dataset[7]. These two networks are suitable for lots of classification tasks.

Approach:
We built and trained a model pipeline that can detect face images with bounding boxes and classify eight different expressions (anger, contempt, disgust, fear, happy, neutral, sad, and surprise). The pipeline includes two deep neural networks for face detection and facial expression classification respectively. The first neural network for face detection is Multi-task Cascaded Convolutional Networks which we adopted from Kaipeng and Zhanpeng's paper [3]. Multi-task Cascaded Convolutional Networks (MTCNN) is a framework developed as a solution for both face detection and face alignment. The process consists of three stages of convolutional networks that are able to recognize faces and landmark locations such as eyes, nose, and mouth. MTCNN code was obtained from a public repository on Github which is a pre-trained

framework.[4] The details of MTCNN architecture are shown in figure 1.

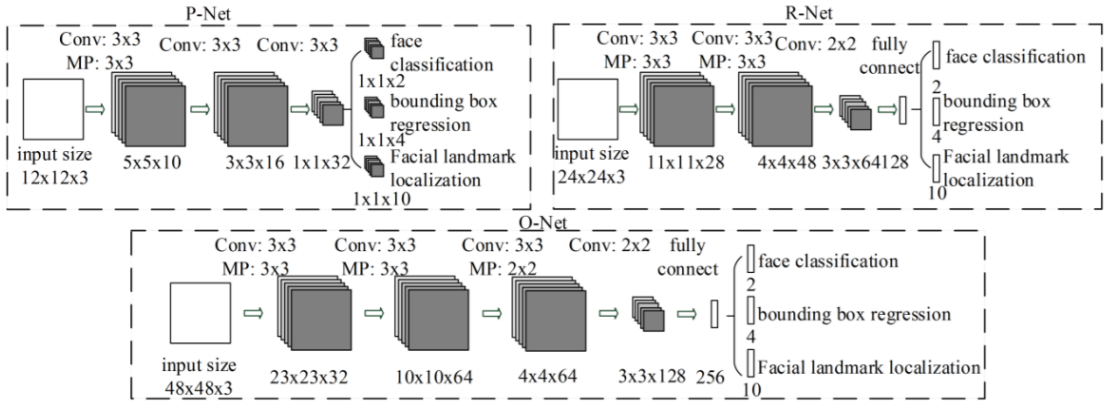

Figure 1: Multi-task Cascaded Convolutional Networks architecture

To detect the face from our input images, we first created a face detector by using the MTCNN framework, then the framework outputs the key coordinates of the face in the image. Furthermore, we used matplotlib libraries to draw a bounding box around the target face based on the framework coordinate result. The sample detection result from the MTCNN framework is shown in figure 2.

Figure 2: Face detection sample result from MTCNN framework

After we obtain the target face from the image, the face inside the bounding box will be extracted, preprocessed, and fed into the second neural network which is used for facial expression recognition. We chose pre-trained VGG-16 and ResNet-18 for facial expression classification. They can classify the extracted human faces into eight expression categories. At last, we save the face images with labels for future health analysis. Figure 3 shows a demo result coming from our project model pipeline.

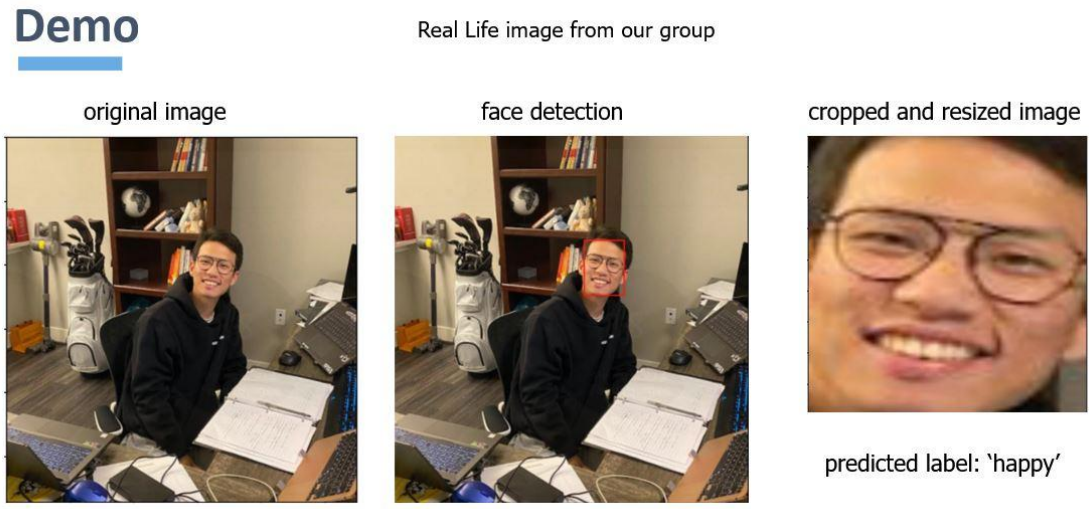

Figure 3: Demo result from our project pipeline

Datasets, Experiments and Results:

We chose the AffectNet dataset to train the facial expression classification network. It contains more than one million labeled facial images collected from the internet with 1250 emotion-related keywords in six different languages. Since this dataset is too large for this project and will take too long to train the model, we randomly select 1600 training images, 400 validation images, and 5000 test images with eight different labels. Before training the classification models, we replaced the last layer of the pre-trained VGG-16 and ResNet-18 with an eight-node fully connected layer and preprocessed the data with resizing and normalization. We used cross-entropy loss function and Adam optimizer to train the models. The performance of VGG-16 is shown in figure 4, the accuracy of training data reaches 98.87% and the accuracy of validation data reaches 44.75%. As shown in figure 5, ResNet-18 achieves 100% accuracy on training data and 50.25% accuracy on validation data. We can see that VGG-16 model is overfitting after 5 epochs, and ResNet-18 model is overfitting after 2 epochs. It may be because the number of training data is too small and the difference between some expressions is too subtle to classify.

Figure4. accuracy and loss of pre-trained VGG-16 network

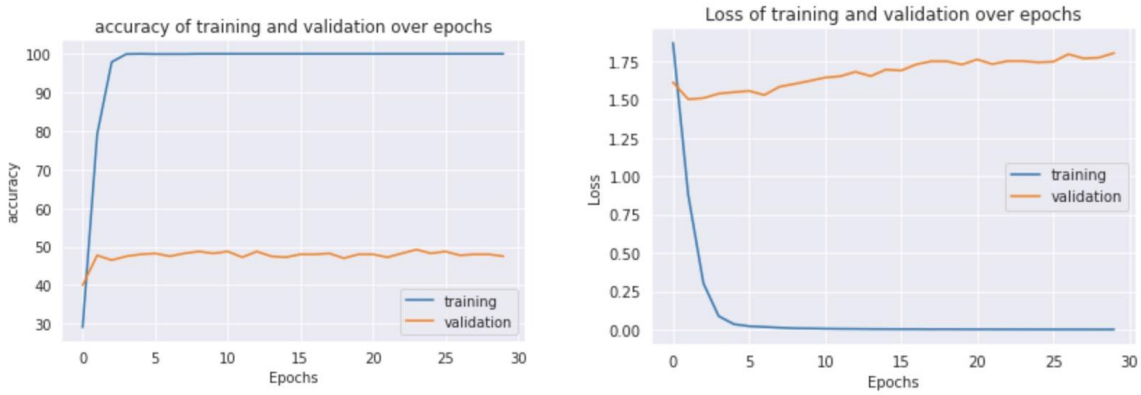

Figure5. accuracy and loss of pre-trained ResNet-18 network

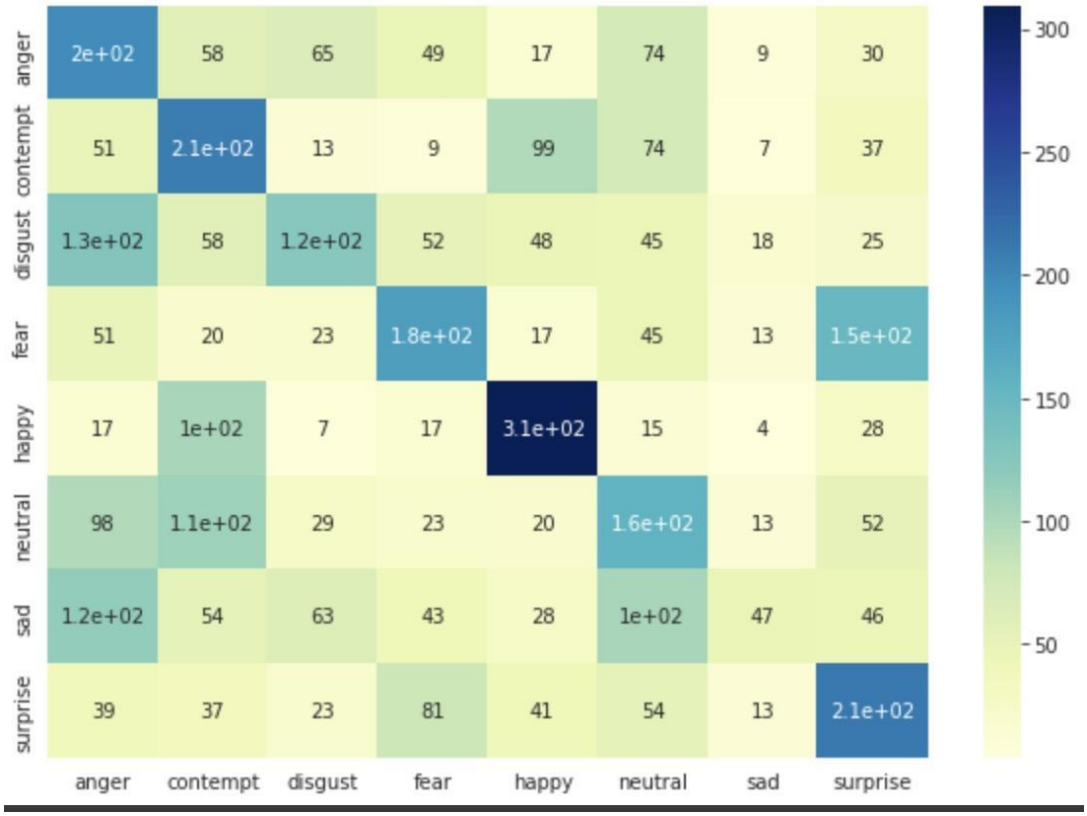

Figure6. Confusion matrix of VGG-16 on 5000 test images

In order to further explore the reason, we plotted a confusion matrix of VGG-16 results on 5000 test images as shown in figure6. It shows that happy expression has the best classification accuracy and some other expressions such as fear, disgust, also perform well; however, some expressions such as sad and anger are indistinguishable from other expressions. By looking at surprise and fear impressions, it makes sense that the model is always confused by these two impressions because human beings may also tend to make mistakes in identifying them.

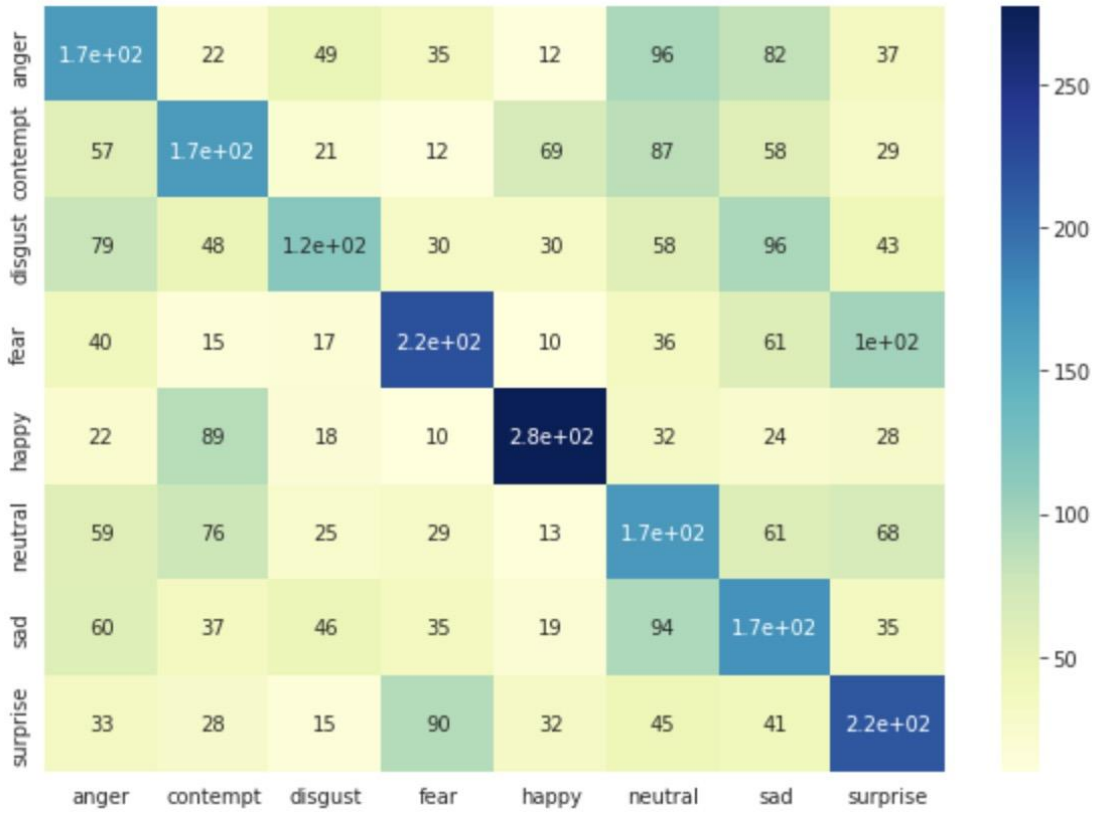

Figure7. Confusion matrix of ResNet-18 on 5000 test images

The confusion matrix of ResNet-18 results is shown in figure 7. Similar to the results of VGG-16, 'happy" perceptron has the highest classification accuracy. However, ResNet-18 has significantly greater performance on classifying 'sad' which is indistinguishable from the VGG-16 model. The classification accuracy of other expressions also shows that ResNet-18 has better performance on our face expression classification task.

Reflection

Our result shows that face detection can be effectively achieved by MTCNN network, however, facial expression classification is not an easy problem for networks such as VGG-16 and Resnet-18. In order to improve the performance of facial expression classification, more training data or data augmentation could be the potential solution, or a more complicated network such as Resnet-128 is recommended.

We hope this technique can be implemented in the school, nursing homes, and private houses, so it can consistently keep track of the health conditions of individuals and send warnings in case of any potential illnesses.

Reference:

[1] A. Mollahosseini, B. Hasani and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, 1 Jan.-March 2019, doi: 10.1109/TAFFC.2017.2740923.

[2] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), 2021, pp. 119-124, doi: 10.1109/SISY52375.2021.9582508.

[3] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[4] https://github.com/ipazc/mtcnn

[5] 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.

[6] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, pp. 730-734, doi: 10.1109/ACPR.2015.7486599.

[7]J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.