Facial emotion recognition based on SFEW dataset using Casper neutral network with convolutional neural network and transfer learning

Yuan Ma
Research School of Computer Science, Australian National University
U6712879@anu.edu.au

Abstract: static facial emotion recognition(FER) is a research field interested in detecting emotions from face images. SFEW is a FER dataset which contains faces close to the real world. Facial expressions under real world are different with facial expressions under lab environment, they always provide more ambiguity and most of current algorithms fail to achieve a high accuracy. Early in 2000s, an improved constructive neural network called Casper was created and it shows a strong power of automatically finding the best fitted structure. On the other hand, convolutional neuron network is the most popular algorithm in computer vision field while it shows a strong power of extracting features from images. My aim is to combine Casper network with convolutional neuron network to achieve a high accuracy on detecting facial expressions from SFEW dataset. My model contains a face detection algorithm based on Viola-Jones Face Detector, a feature extraction method based on convolutional neuron network and a classification network based on Casper neural network. Moreover, regarding to the small amount of data in the dataset, I implemented transfer learning by pretraining the model on a similar dataset FER2013 and fine tuning on SFEW. The final result achieves around 39% average accuracy and the implementation of transfer learning achieves a 4% improvement on it.

Keywords: facial emotion recognition, SFEW,Cascade neural network, Casper neural network, deep learning, convolutional neural network, Viola-Jones Face Detector, transfer learning, FER2013

1 Introduction

This paper implemented a method for doing static facial expression recognition based on SFEW dataset. Traditional static facial expression recognition datasets could be classified into two categories - lab-controlled or not. Some of the most popular lab-controlled datasets are JAFFE[1] dataset,CMU Pose Illumination and Expression (PIE) dataset[2] and B+ dataset. All of these datasets require volunteers to perform certain facial expressions in front of a camera with fixed poses and angles. Machine learning methods based on these datasets always give a high accuracy. For instance, current method of combining 2D-LDA(Linear Discriminate Analysis) with SVM(Support Vector Machine) achieves 94.12% accuracy based on JAFFE dataset through cross validation[3]. However, one of the main limitation of these datasets is that they can not represent facial expressions under real conditions. The other type of dataset including SFEW[4] on the other hand attempts to approximate real facial expressions. SFEW did that by extracting 400 images from AFEW[5]. AFEW is a dataset of movie clips containing emotions. SFEW dataset relabeled these clips into 7 classes including neutral, happy, angry, sad, disgust and fear. As movie actors always try to mimic real situations and they normally do it better than others, these clips are closing to facial expressions under real conditions. Because real facial expressions are much more complex than the one under lab environment, algorithm based on the second category of dataset is always more difficult to get a higher accuracy. According to the experiment of original SFEW paper[4], model based on JAFFE has an accuracy 30% higher than the model based on SFEW under the same algorithm.

Recently, both deep learning and constructive neural network show great performance in the field of machine learning. Deep learning methods such as convolutional neural network yield plenty of achievements[6][7] in image classification and proves its superiority of extracting features from images. Constructive neural network including Casper neural network[8] is appreciated with its ability of automatically generating hidden neurons to find the best fitted structure.

In this paper, I explored doing static facial expression through combining convolutional neural network with Casper neural network, as both techniques get multiple achievements and no one has tried combining them in this field. However, convolutional neural network's power is based on the sufficient amount of training data and SFEW only provides 700 images. Thus, I utilized transfer learning as it is accepted as a good way dealing with lack of labeled data[9]. I did that by pretraining the convolutional neural network on FER2013 dataset[10].

The rest of the paper is organized as follows: Section 2 introduces techniques been used including Viola-Jones Face Detector[11], transfer learning, Casper and convolutional neutral network. Section 3 displays and analyzes results of my experiment, Section 4 concludes the results and discusses future work.

2 Method

2.1 Data Preprocessing

One of the biggest difference between lab-controlled facial expression images and real facial expression images is the environment. Among lab-controlled face images, all faces will roughly be in the same position with the same angle. But for real images, face could appear in any position with any angles in the image. Though it is still possible directly extracting information from images, locating faces will greatly benefit our model as all information about emotions is on faces. To do that, I implemented Viola-Jones face detector[11]. Viola-Jones face detector combines ideas of integral image, classifier learning with AdaBoost and the attentional cascade structure. It shows an efficient performance on detecting faces in different scales and it is easy to implement.

In my model, I firstly passed images into Viola-Jones face detector. The output was the position of a rectangular box containing faces inside current image. Then I cropped the image according to that box and resized it to a resolution of 48*48. After that, images were transformed from RGB images to gray scale images. Because most of information is contained in intensities, changing images into gray scale form will increase the efficiency of the model. Moreover, as images may have different lighting conditions and many images in SFEW have an overall low intensity, I also implemented histogram equalization before classification. This technique enhanced the global contrast of images through making intensities better distributed on the histogram. This method has shown to slightly improve the accuracy of FER models in previous research.[12] All of these procedures are implemented through opencv library in python.

The overall data preprocessing procedure is illustrated in figure 1.

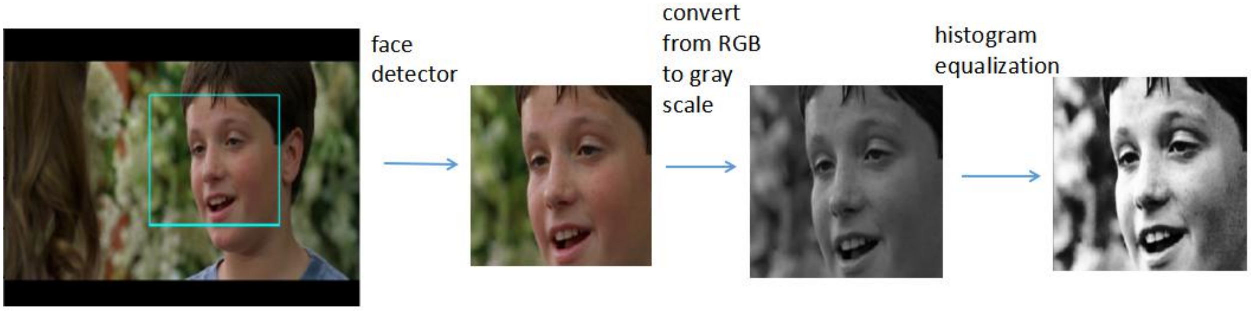

Fig. 1. Data Preprocessing

2.2 Convolutional neural network architecture

.

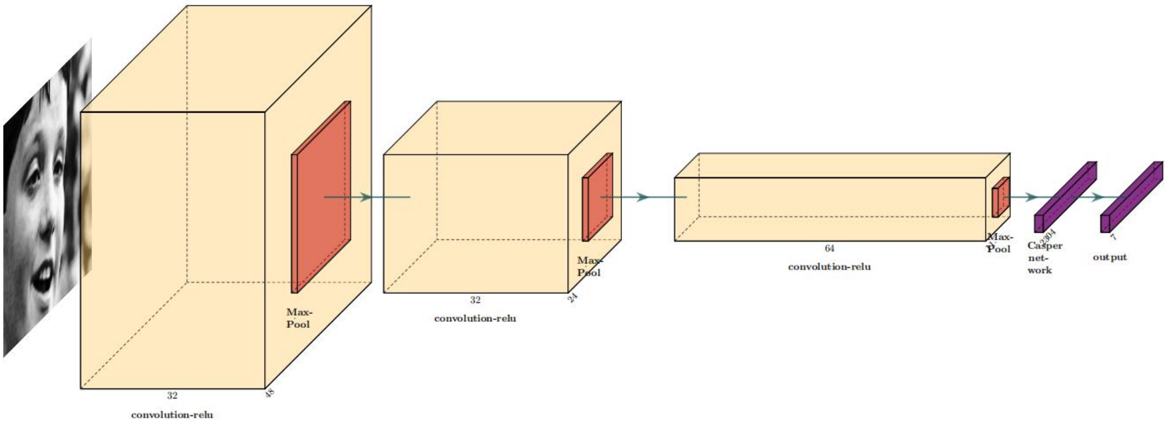

Fig. 2. Convolutional neural netwrok architecture

The architecture of implemented convolutional neural network is shown in figure 2. Convolutional neural network is used for extracting features from images through feature maps and multiple layers with different functionalities. In my method, there are three convolutional layers and three max-pooling layers. The input is a preprocessed 48*48 face image. All convolutional layers use a 5*5 convolutional filter with strides set to 1 and padding set to 2. All max-pooling layers use a 2*2 filter with strides set to 2. Relu(Rectified linear units)[13] is used as an non-linear mapping functions for all hidden layers. Batch normalization is also implemented before non-linear mapping function to make the network more stable[14]. The momentum for batch normalization is set to 0.5. The output of last max-pooling layer is flatten to a 1*2304 vector and connects to a Casper neural network.

2.3 Casper neural network

Casper algorithm is an extension of conventional Cascor algorithm (Cascade correlation method)[15]. They are constructive neural network that could automatically add hidden neurons to themselves. New added neurons will be connected to all the previous hidden neurons and input neurons. In Cascor algorithm, after adding a new neuron, weights of all previous neurons will be frozen (stop changing) and training will only change the weights of new added neurons. This brings an issue that early frozen hidden neurons might capture bad features of dataset and require future neurons to fix it. Casper algorithm solves this issue by dividing neurons into 3 categories L1,L2,L3 and let them keep learning instead of freezing them. Neurons in the L1 category have the highest learning rate and neurons in the L3 category have the lowest learning rate. When a new hidden neuron is added, new hidden neuron will initially be labeled as L1 category and previous neurons will be moved from L1 to L3 category. By doing that, we allow all hidden neurons slightly adapt themselves during the whole training without changing the precaptured features.

In my implemented Casper algorithm, I used Relu activation function and cross entropy loss function. I choose Adam as my optimizer. The learning rate for training the initial network is 0.2. The learning rate for L1, L2, L3 are 0.2, 0.005, 0,001. For each new added neuron, its weight will be initialized in the range of (-0.7,0.7) to prevent from introducing too much noise. A new neuron is added after 15+P*N times of training. N is the number of existed hidden neurons and P is the hyper parameter. Training will stop when loss decreases less than 1% within this time period. These parameters are suggested by the technique paper as they show a good performance independent to the task. P is set to 10 as it is tested give an average good result under this task.

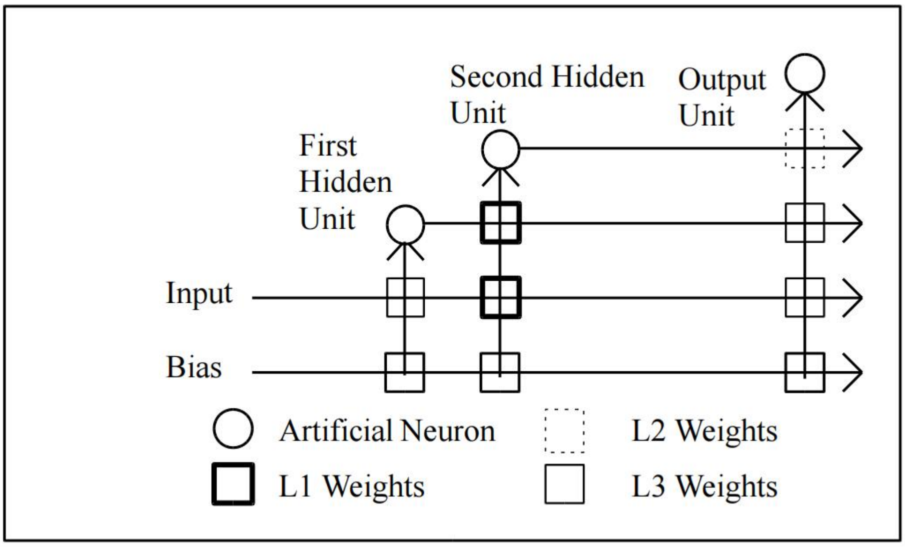

Fig. 3. Casper neuron network structure[5]

2.4 Transfer Learning

One of the biggest necessity for convolutional neural network is the big amount of independent data. In SFEW dataset, there are only 700 images which will greatly limits the performance of our model. One of the best approaches is using transfer learning.[16]. Transfer learning is the method of pretraining our model on other similar dataset and copying its weights as the initial weights for new model. This method provides a good starting point for our model and also increases the efficiency of training. Here, I pretrained my model on FER2013 dataset[10]. FER2013 is a FER dataset with over 3000 facial expression images and it also uses a 7-classes facial expression classification method. As shown in figure 4, images in FER2013 are pretty similar with preprocessed images in SFEW. Thus, I assumed pretraining on FER2013 will benefit out model.

In my method, the architecture of pretraining model is a convolutional neural network connecting with three fully connected layers. Each fully connected layer uses Relu as non-linear activation function. To avoid overfitting, dropout layers are inserted between the first and the second fully connected layer, second and third fully connected layers.The architecture of convolutional neural network is the same as our original model. The learning rate is set to be 0.01 and weight decay is implemented to prevent overfitting. The architecture is shown in figure 5. Here, I did not use Casper neuron network while our goal is to pretrain convolutional neuron network. Moreover, convolutional neuron network connecting with multiple fully connected linear layers has shown to be effective in may classification experiment[17][6]. After pretraining on FER2013, I only kept the convolutional neuron network part, connected it with a Casper neuron network and fine tuned on SFEW dataset. The learning rate of convolutional neuron network is set to be as small as 0.001.

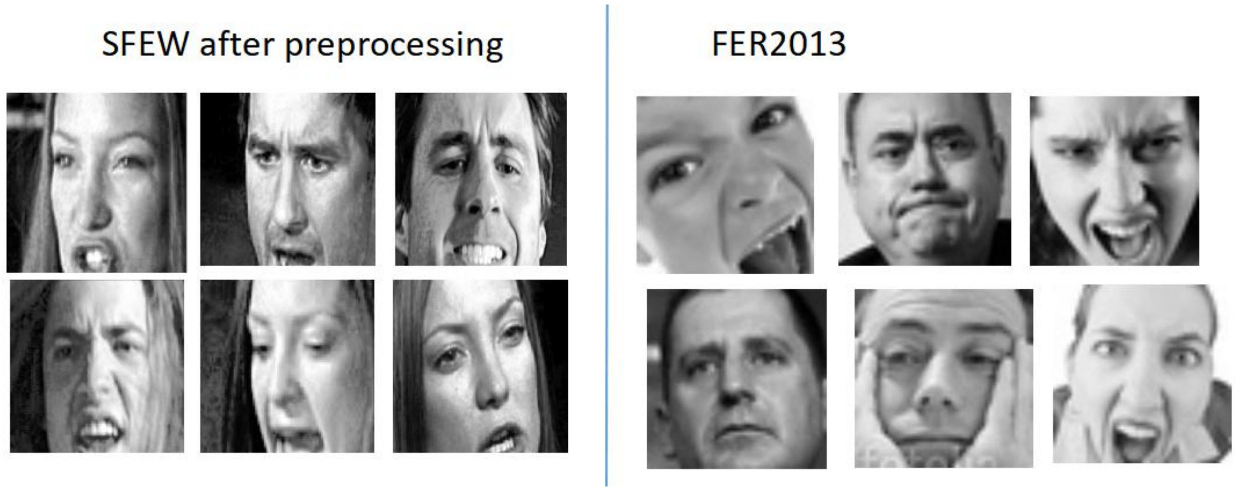

Fig. 4. Comparison between SFEW and FER2013

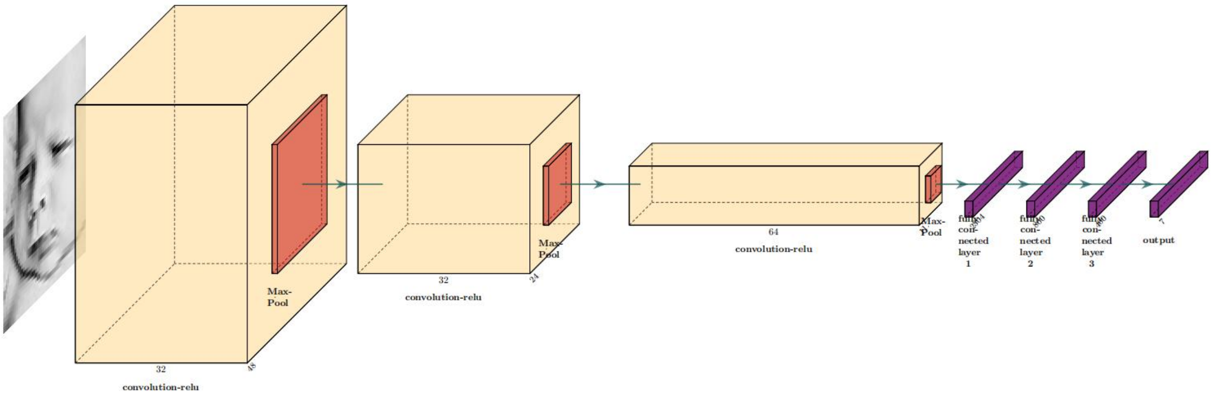

Fig. 5. architecture of pretrained model

3 Result and discussion

In this section, I compared the performance of four model structures: convolutional neural network connecting to Casper neuron network with transfer learning, covolutional neuron network connecting to Casper neuron network without transfer learning, convolutional neuron network connecting to three fully connected linear layers with tranfer learning, convolutional neuron network connecting to three fully connected layers without transfer learning. The structure of convolutional neuron network connecting to fully connected linear layers are the same as architecture shown in figure 5. Models are evaluated through 5-fold cross-validation by separating dataset into five subsets. Four folds are used for training and one fold is used for validation each time. The evaluation is based on the curve of training loss and validation accuracy. The final output is also compared with the SPI baseline mentioned in the dataset paper[4].

3.1 comparing experiment result with baseline model

Figure 6 shows the curve of training loss and figure 7 shows the curve of testing accuracy. The average validation accuracy for Casper model with transfer learning, Casper model without transfer learning, linear model with transfer learning and linear model without transfer learning are 39.209% , 33.514% ,31.09% and 34.04% respectively, meaning that our final model surpasses the accuracy of linear model over 5% and transfer learning also improves the accuracy over 5%.

By observing the training loss curve, we could see that model using transfer learning always has a smaller initial loss than other model with same algorithm, meaning that transfer learning provides a better starting point for the model. Moreover, models using Casper algorithm have a slower loss decreasing, because Casper algorithm needs time to build its structure starting from zero hidden neurons and it also need to deal with the rapid loss rises caused by the noise introduced by new added hidden neurons.

By observing the accuracy curve, we could see model using convolutional neuron network connecting to Casper neuron network with transfer learning has the average highest accuracy. It is also the one converge the fastest. Comparing this model with the one using Casper neuron network without transfer learning, we could see transfer learning not only makes the accuracy improve but also reduces the time of training. In addition, model using Casper algorithm without transfer learning does not outperform model using linear fully connected layers. One of the possible reason is: Casper algorithm adds neurons depending on the reduction of loss. As long as the loss is influenced by convolutional neuron network and Casper neural network at the same time, Casper algorithm may fail to correctly evaluate if it should add more neurons on itself.

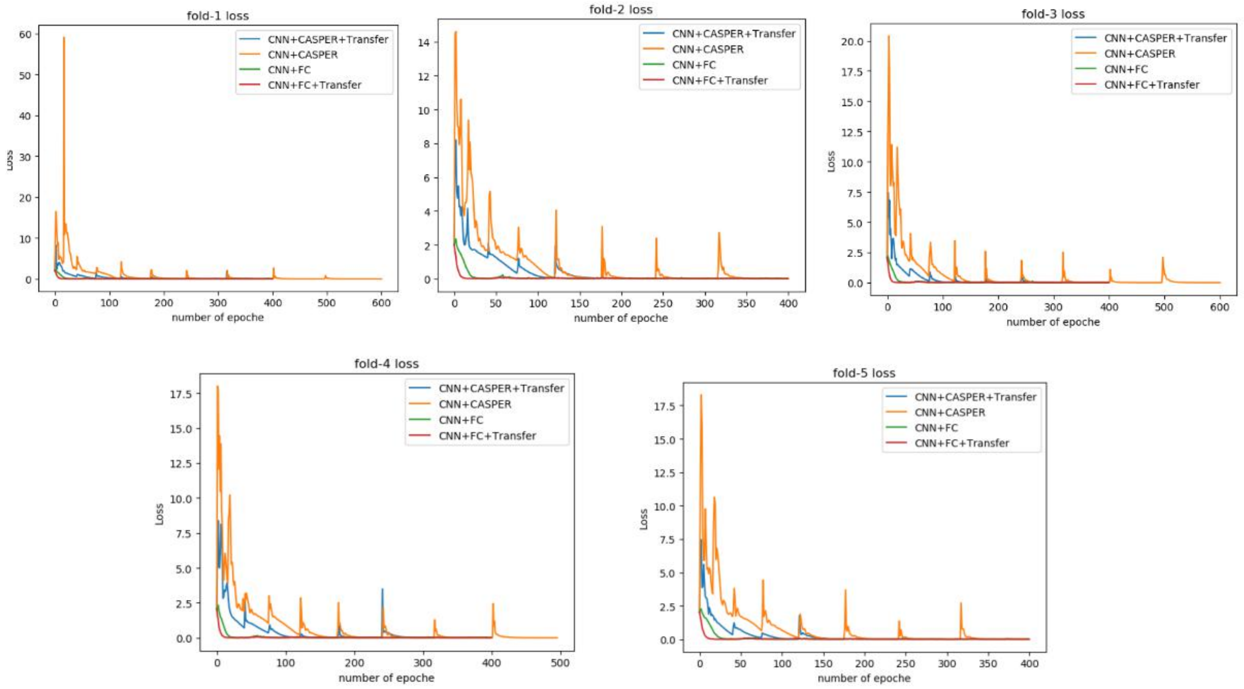

Fig. 6. Training loss

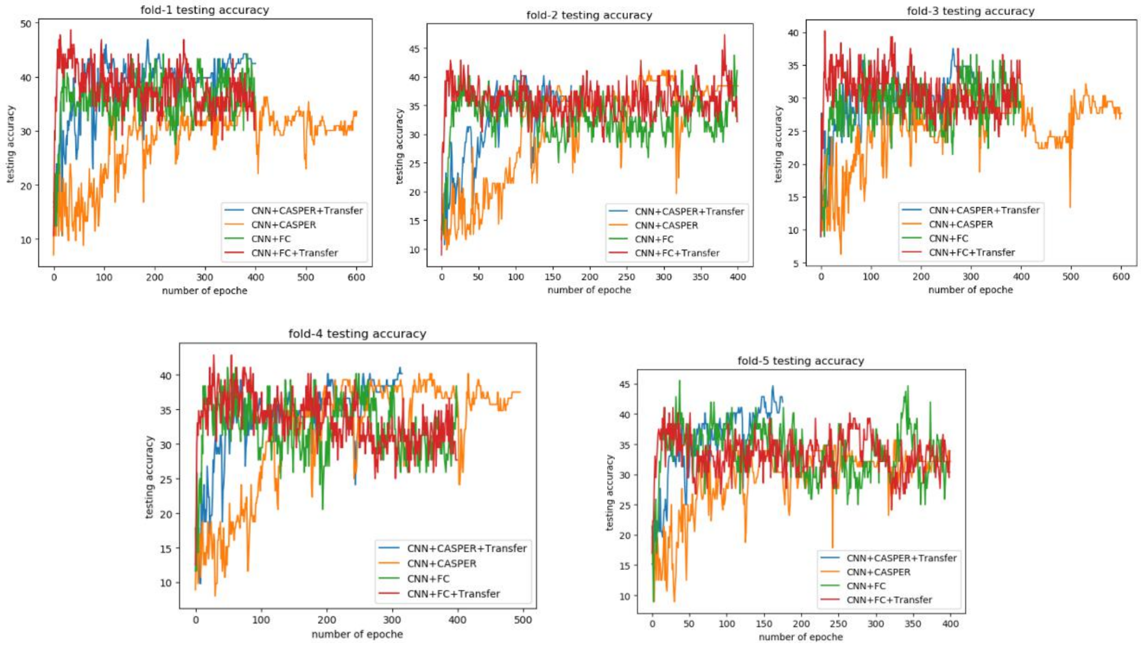

Fig. 7. validation accuracy

3.2 evaluate model based on SPI protocol

In this experiment, I also evaluated model based on SPI protocol mentioned in the original paper. The evaluation is based on the recall, precison, specificity and accuracy of the model. The formulae for evaluation is shown below. The average precision, recall and specificity is shown in table 1. The final average accuracy of my model using convolutional neuron network and Casper with transfer learning greatly surpasses the baseline classification accuracy proposed in the original SFEW paper and the data shown in Table 1 shows an overall better performance of detecting facial expressions. This result shows my model do achieve detecting emotions with SFEW dataset.

$$\text{(1) recall} = \frac{\text{tp}}{\text{tp}+\text{fn}}$$

$$\text{(2)specificity} = \frac{\text{tn}}{\text{fp}+\text{tn}}$$

$$\text{(3)accuracy} = \frac{\text{tn}}{\text{fp}+\text{tn}}$$

tp = true positive, fp = false positive, fn = false negative, and tn = true negative.

Table 1. recall, precision, specificity result

Facial expression	angry	disgust	fear	happy	neutral	sad	surprise
recall	0.417	0.364	0.419	0.447	0.499	0.35	0.399
precision	0.397	0.455	0.405	0.456	0.292	0.369	0.395
specificity	0.909	0957	0.914	0.912	0.897	0.920	0.906

4 Conclusion and Future Work

In this paper, I implemented a deep learning method of combining convolutional neural network with Casper neural network. The method aims to do facial expression recognition based on SFEW dataset. Transfer learning is also used by pretraining the convolutional neural network on the FER2013 dataset. Transfer provides a 5% accuracy improvement on the model and also increases the efficiency of training. The final model achieves an average accuracy of 39%. This result outperforms the baseline classification accuracy as well as the performance of simple convolution neuron network connecting with three fully connected linear layers. However, this result is still 10% lower than the state of the art[6]. A better structure of convolutional neuron network could be explored to improve the accuracy of the model.

One of the biggest limitation of SFEW dataset is the short number of data. Though I used transfer learning to deal with that, it will still influence the accuracy and generality of the model. One of the good extension on this paper is implementing the proposed method on SFEW 2.0 dataset[18]. SFEW 2.0 dataset is an extension of SFEW and it contains over 1500 images which is twice more than the SFEW dataset. Another way to improve this method is using a different face detector. Here, I implemented Viola-Jones face detector. Though it has a strong ability of detecting frontal faces, it is not good at detecting profiles. Other powerful face detector such as HOG[20] and DCNN[19] might generate better result. Moreover, other data preprocessing methods such as image alignment could also be implemented to improve the accuracy of the model.

Reference

1. Lyons, M., Akamatsu, S., Kamachi, M. and Gyoba, J., 1998, April. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition* (pp. 200-205). IEEE.
2. Sim, T., Baker, S. and Bsat, M., 2001. The CMU pose, illumination and expression database of human faces. *Carnegie Mellon University Technical Report CMU-RI-TR-OI-02*.
3. Shih, F.Y., Chuang, C.F. and Wang, P.S., 2008. Performance comparisons of facial expression recognition in JAFFE database. *International Journal of Pattern Recognition and Artificial Intelligence*, *22*(03), pp.445-459.
4. Dhall, A., Goecke, R., Lucey, S. and Gedeon, T., 2011, November. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 2106-2112). IEEE.
5. Dhall, A., Goecke, R., Lucey, S. and Gedeon, T., 2011. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, *2*, p.1.

6. Yu, Z. and Zhang, C., 2015, November. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 435-442).

7. Ko, B.C., 2018. A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), p.401.

8. Treadgold, N.K. and Gedeon, T.D., 1997, June. A cascade network algorithm employing progressive RPROP. In *International Work-Conference on Artificial Neural Networks* (pp. 733-742). Springer, Berlin, Heidelberg.

9. Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), pp.1345-1359.

10. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H. and Zhou, Y., 2013, November. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117-124). Springer, Berlin, Heidelberg.

11. Viola, P. and Jones, M., 2001. Fast and robust classification using asymmetric adaboost and a detector cascade. *Advances in Neural Information Processing System*, *14*.

12. Mungra, D., Agrawal, A., Sharma, P., Tanwar, S. and Obaidat, M.S., 2020. PRATIT: a CNN-based emotion recognition system using histogram equalization and data augmentation. *Multimedia Tools and Applications*, *79*(3), pp.2285-2307.

13. Nair, V. and Hinton, G.E., 2010, January. Rectified linear units improve restricted boltzmann machines. In *Icml*.

14. .Ioffe, S. and Szegedy, C., 2015, June. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

15. Fahlman, C.L., 1990. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems*, *2*.

16. Ponzio, F., Urgese, G., Ficarra, E. and Di Cataldo, S., 2019. Dealing with lack of training data for convolutional neural networks: the case of digital pathology. *Electronics*, *8*(3), p.256.

17. Coşkun, M., Uçar, A., Yildirim, Ö. and Demir, Y., 2017, November. Face recognition based on convolutional neural network. In *2017 International Conference on Modern Electrical and Energy Systems (MEES)* (pp. 376-379). IEEE.

18. Dhall, A., Ramana Murthy, O.V., Goecke, R., Joshi, J. and Gedeon, T., 2015, November. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 423-426).

19. Zhang, C. and Zhang, Z., 2014, March. Improving multiview face detection with multi-task deep convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision* (pp. 1036-1041). IEEE.

20. Pang, Y., Yuan, Y., Li, X. and Pan, J., 2011. Efficient HOG human detection. *Signal Processing*, *91*(4), pp.773-781.