School of Informatics

Informatics Research Review
Convolutional neural networks for facial expression recognition

[redacted]

January 2021

Abstract

In this research review, I searched some paper of conventional neural network of facial expression recognition at first. Then I conclude the general structure of a FER CNN method. Then I introduced other CNN methods for dynamic FER, 3D FER and FER with occlusion. For dynamic FER, an objective function is helpful. FRR CNN and DFF CNN is proposed for 3D FER problem. Some variants of CNN for FER are also discussed. In the end, I introduced some researches that work on identify what FER CNN learned form the dataset.

Date: Friday 29th January, 2021

Supervisor: [redacted]

1 Introduction

The communication of human is realised more by gesture, pose and expression rather than language. Therefore, human face expression recognition is a very important problem in human-machine interaction.

With the development of deep learning, the convolutional neural network inspired by the way how human brain process visual signal has shown great effectiveness on image recognition. Usually, a CNN network consists of convolutional layers, pooling layers and fully connected layers.

In this review, I searched for the recent works of CNN for facial expression recognition problem. The FER problem can be divided into 2D FER, 3D FER, 2D+3D FER, FER with occlusion, static FER and dynamic FER. There lots of datasets for training and validating FER methods. FER-2013 is widely used for the development of a FER CNN. The images in FER-2013 are collected by google image search API. It consists of 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels. JAFFE, the Japanese female facial expression dataset contains 213 facial expression samples from 10 Japanese females. CK+, the extended Cohn–Kanade database is the most widely used FER laboratory dataset. Similar to CK+, MMI is also a laboratory-controlled dataset. AFEW, the acted facial expressions in the wild dataset consists of video clips in different movies. Therefore, the images in AFEW are under a complex condition. The illumination, pose is random. There could also be occlusion in the face. It's a dataset in the wild condition. The static facial expressions in the wild (SFEW) is a subset of AFEW. Multi-PIE dataset is developed by CMU. It contains expression images for different viewpoints. It's often used for dynamic FER. The BU-3DFE is a static 3D FER dataset. BU-4DFE is the dynamic 3D FER dataset. AffectNet dataset is an extremely large dataset that contains more than one million images.[1] These images are collected by querying different search engines using emotion-related tags.

Firstly, I introduced some typical CNN structure for FER. I analysed the process of FER CNN. At first, images from datasets should by preprocessed by cropping, rotating and alignment. Then we need to recognise faces from the processed images. The face feature will be the input of the FER CNN network. Conventionally, a CNN network consists of convolutional layer, pooling layer and a fully connected layer. Then I introduced a method for dynamic FER problem. After that I also discussed two methods on the occlusion-aware FER. It includes the patch-gated CNN and CNN with attention mechanism. Next, I organized some innovative variants of FER CNN including feature redundancy-reduced convolutional neural network, unsupervised domain adaptation for FER using generative adversarial networks. Finally, I discussed how CNN learn from dataset through facial action units.

2 Literature Review

2.1 General Methods of Facial Expression Recognition Using CNN

2.1.1 Dataset and Face Preprocessing

Here are some datasets that are frequently used for training CNN model, like FER-2013, SFEW2.0, CK+, KDEF, Jaffe. Images of FER-2013 are gathered by Google image search API. SFEW2.0 provides images collected from movies. Thus faces in SFEW2.0 are presented in a more natural way which means the environment is more complex. It is more difficult to realise

FER on SFEW2.0. Jaffe contains images of Japanese female faces and the faces are perfectly straight towards the viewer. All these datasets contains 7 types of emotion expression. Each image is labelled correctly.[2]

Then we need to detect faces from datasets. At first, to improve the robustness and accuracy, images from these databases need to be cropped, aligned and rotated to generate images of faces that are easier to be recognised.[3] Dlib C+ Library provided by King works well on it. However, this method cannot detect face in every image of the dataset. The success rate of each dataset is different. The main reason for the failure of face detection is that only part of the face is shown in some images. Besides, we should also care about the illumination and contrast in these images. Five methods (raw, histogram equalization, isotropic diffusion-based normalization, DCT-based normalization, difference of Gaussian) are tested on five datasets on four CNN model. The result shows that Hist-eq performs best on FER-2013, SFEW2.0 and CK+. IS performance best on KDEF.

Another way to detect face from datasets is to combine several detectors: the joint cascade detection and alignment (JDA) detector, the Deep-CNN-based (DCNN) detector from and Mixtures of Trees (MoT). The three detector are connected in a sequence. The input images are processed by JDA first, if fail, it will be passed to DCNN. If DCNN is unable to recognise face in the picture, MoT will take over it in the end. By this way, the success rate of this method is extremely high. Only one image could not be recognised in 372 SFEW images.[4]

2.1.2 The Proposed CNN Models

One of the successful CNN models contain five convolutional layers with three fully connected layers in the end. To get a better performance with limited training data, author chose stochastic pooling layers rather than max pooling. Besides, dropout was used in the fully connected layer. It also adds randomized perturbation to the samples to make the network more robust. This model achieves the accuracy of 70% on FER dataset, 52.29% on SFEW validation and test set. [2]

Another CNN model contains three convolutional layers and two fully connected layers. Max pooling layers with ReLu activation function are used to connect these convolutional layers. Behind two connected layers is the softmax layer. This model achieves the recognise rate of 86.54% on CK+.[5]

2.1.3 Conclusion

From above, we can see that if we want to develop a convolutional neural network for facial expression recognition, first we need to choose some datasets. Then, we need to preprocess these images by cropping, rotating and aligning the. After preprocessing, we need to recognise faces from these images.

The convolutional neural network usually consists of several convolutional layers with stochastic/max pooling layers and ReLU. After that is several fully connected layers. The softmax is implemented for classification.

2.2 Dynamic Facial Expression Recognition

It is not enough to realise static facial express recognition because in the application out of laboratory, we should consider other factors like pose and illumination. Wissam proposed a

FER method that is flexible to the variations of the images.[6]

FRE methods can be based on geometric, appearance or mixed. Geometric-based FER method is more limited than appearance-based method because of the complexity of the environment that the photo was taken. The FER method based on appearance focus on the changes of texture. To make sure the system can recognise facial expression successfully of the same expression at different image variations.

This method implements a new objective function that can reduce the error of classification while reduce the difference of features between images with variable changes. Features learned by objective function will be passed to Siamese networks. The objective function consists of two terms. One minimizes the classification error, the other one minimizes the feature difference between images. Features learned by this objective function are more robust. By this way, the network shows robustness to variation of images without additional computational complexity.[6]

The Siamese network consists of two CNN networks with three convolutional layers, each convolutional layer is followed by a max pooling layer with ReLu. Behind them is a fully connected layer of 300 nodes. Softmax is implemented for classification. Compared to baseline CNN, the accuracy of proposed method is higher than baseline CNN by 3.51%. When tested on dataset with pose variations, the proposed method has higher accuracy rate than baseline CNN set across all poses. Furthermore, the performance of proposed method towards variation of illumination is also measured. The result shows in each illumination the proposed method has higher accuracy than baseline CNN. In conclusion, the use of objective function to learn features improves the robustness of CNN without increasing the computational complexity of the network.

2.3 Occlusion-aware Facial Expression Recognition

2.3.1 Patch-Gated CNN

Occlusion is one of the inevitable problems that appears in the application of FER. In many cases, the face of a person is not fully shown in the picture. Many well-performed FER systems are constructed controlled environment. These FER systems may not work so well in wild condition.

To overcome this challenge, the patch-gated convolution neutral network is proposed. It simulates the way how human recognises facial expression. When some part of the face is covered, human may use the symmetric part of face or other highly related region to make judgement. PG-CNN detects the blocked face patch and focus on the unblocked patch that provides information. According to the related face landmark, PG-CNN crops the part that it interests in. For each patch, the weight is recalculated by PG-Unit. PG-Unit finds the blocked face patch and gives them low weight. The unblocked face patch will be given high weight.

The key methods that make PG-CNN work are region decomposition and occlusion perception.[7] At first the input images are represented as features. The PG-CNN decomposes these features to 24 sub-features for 24 patches. Each patch is encoded to a weighted vector by PG-Unit.[7] The weight is computed according to the attention net. In the end, the weighted features are connected to represent the occluded face. Next there are three fully connected layers followed by softmax.

Compared with DLP-CNN, PG-CNN performs better on RAF-DB and AffectNet database because patch-based model can reflect the movement of muscle. On occlusion datasets. PG-

CNN also exceeds DLP-CNN.[7] With the increase of training set, the performance gap between them is decreasing.

The performance of PG-CNN on cross-dataset is also evaluated. The PG-CNN was trained on RAF-DB or AffectNet dataset and evaluated on CK+, MMI, Oulu-CASIA dataset.[7] When trained on CK+ and evaluated on MMI, the accuracy rate of PG-CNN surpassed other existent FER methods.

The comparison between CNN and P-CNN (PG-CNN without PG-Unit) and comparison between P-CNN and PG-CNN are conducted to identify how PG-CNN improves the performance of CNN. On occluded dataset and unoccluded, the performance of P-CNN surpasses VGG-16. On RAF-DB and AffectNet dataset, the accuracy of PG-CNN is higher. This is because PG-Unit enables the system to process related local patches and transfer the attention to other related local part. Apart from MMI, the performance of PG-CNN surpasses P-CNN on almost all datasets. P-CNN relies on the whole face region while PG-CNN can make judgement by regional patches. Besides, PG-CNN will not put too much attention of occluded patches.[7]

2.3.2 CNN With Attention Mechanism

Another FER method for images with occlusion is the convolution neutral network (CNN) with attention mechanism (ACNN). It can perceive the blocked region of the face, and focus on the unblocked region.[8] It reweights the representation of region of interest according to the suggested gate unit. There are two kinds of ACNN: patch-based CNN (pACNN) and global-local-based CNN (gACNN). These ACNNs are tested on two in-the-wild facial expression datasets (RAF-DB and AffectNet).[9] The result shows ACNNs can improve the accuracy of occluded face and un-occluded face. ACNN can transfer attention from blocked patches to related unoccluded patches.

gACNN is very similar to Patch-Gated CNN. Besides focus on part of the facial region, gACNN also consider the global face region. On the on hand, global-local-based attention method helps get local details and the global context.[9] On the other hand, gACNN can learn the diversity of features. Global-Gated Unit is embedded into gACNN to balance global facial expression automatically. The GG-Unit consists of two parts: one that encodes the input features to global vector representation, one that contains an attention network which learns scalar weight to represent the contribution of global representations.[9]

Compared with DLP-CNN on RAC-DB and AffectNet database, the performance of gACNN is much better.[9] This is because model with global attention can capture subtle muscle movement better.[9] With the help of gate unit, gACNN surpasses DLP-CNN on occlusion datasets.

2.4 3D Facial Expression Recognition

2.4.1 Fast and Light Manifold CNN

One novel method of 3D facial expression recognition is the fast and light manifold CNN (FLM-CNN). FLM-CNN can save much memory compared to other manifold CNNs without losing the information in the original data. The manifold CNN will not be affected by the rotation. The FLM-CNN is robust to variation of images.[10]

FLM-CNN adopts a human vision inspired pooling structure and a multi-scale encoding strategy to enhance the representation of image to magnify the facial difference between expressions.

Besides, through preprocessing based on sample tree, the cost for storage pf data reduces greatly. FLM-CNN shows high accuracy compared to other state-of-the-art methods on BU-3DFE.[11]

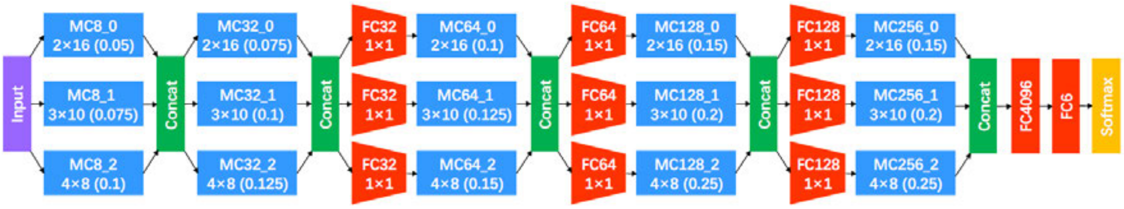

Figure 1: The Framwork of FLL-CNN

2.4.2 Deep Feature Fusion CNN

3D face model will not be affected by illumination or pose variations compared to 2D model.[12] The proposed method combines multiple features of 3D face model.[13] To do FER on 3D datasets, we need to extract 2D attribute information from it. The CNN will learn features from the combination of multiple features of the 2D attribute. What's more, the use of global average pooling can reduce the number of parameters needed in the model so that the problem of overfitting could be avoid. This method is evaluated on Bosphorus database and shows good performance.[14]

Deep feature fusion FER was evaluated on Bosphorus database. The result shows that by combining normal, texture, depth and curvature we can get the highest accuracy. The adding of 2D texture can improve the accuracy because it's more detailed.[15] Besides, the combination of curvature and texture can also reach high accuracy.[16] The features obtained by this method can lead to better performance than handcraft features. Fine-tuning can improve the performance of model and get better results. Compared with other state-of-the-art methods on the Bosphorus Database, the proposed method improves the performance of the model.[15] Furthermore, the combination of 2D and 3D facial information will help the model get the best performance.[17]

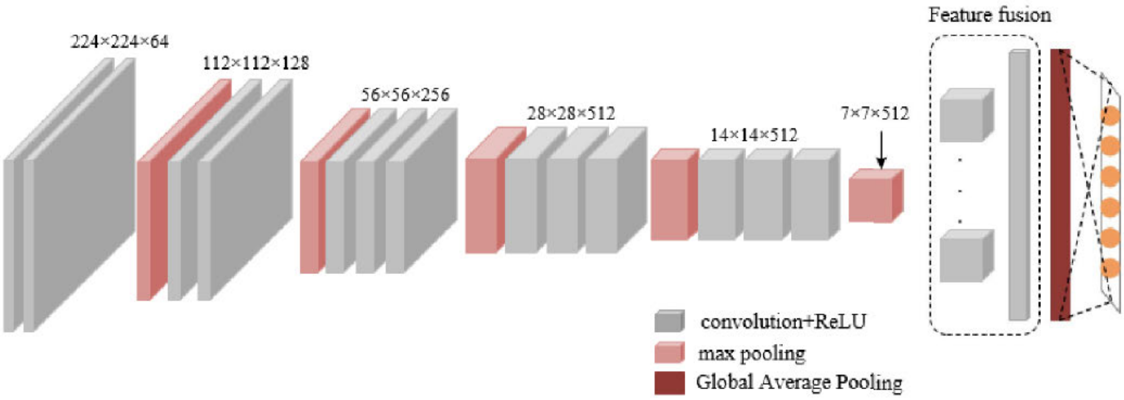

Figure 2: The framwork of DFF model

2.5 Variants of CNN used on FER

2.5.1 Feature redundancy-reduced convolutional neural network

To overcome the redundancy of features in over learning convolutional kernels, the feature redundancy-reduced CNN is proposed. Besides, there are some transformations that are redundant for solving FER problem, parallel Siamese network is implemented in the network. The transformation-invariant pooling strategy will produce high-level representation. Each pair of feature maps in convolutional layer is used to generates more kernel to reduce feature redundancy.

The feature vectors generated by conventional CNN are combined with redundant noise. Therefore, it cannot describe the images well when doing expression analysis. Siamese network with multiple channels are implemented in this system to enhance the system's ability of judging features. In this model, images from the same transformation are set to the input of the same channel. Then when a image passes the convolutional module, it will generate a concatenate feature vector. Weights are shared among all channels. A new convolutional module is also proposed to reduce the feature redundancy. A new regularisation term is added to the loss function.[18]

$$J = \frac{1}{2}\left\|O(X) - Y\right\|^2 + \sum_{c=1}^{m} \frac{\alpha_c}{L_c} \tag{1}$$

This method is evaluated on the extended Cohn-Kanade and The Japanese Female Facial Expression database. In the experiment, FRR-CNN is configured with three parallel channels. Compared with other FER models like conventional CNNs and deep learning CNNs, the proposed method performs best on CK+. FRR-CNN is competitive on the classification of surprise, happiness and disgust which has strong facial movements.

2.5.2 Unsupervised Domain Adaptation for FER using Generative Adversarial Networks

Some FER methods may perform well on one dataset while not so well on other datasets due to the difference of expression feature distribution on different datasets. To improve the performance of the CNN model on crossed dataset, the unsupervised domain adaptation is applied to CNN.[19] The proposed method is suitable for unlabelled small datasets. To get more samples, generative adversarial network is trained on the objective dataset.[19] Through training, GAN will generate fine-tuned samples. These fake samples are labelled to be different with original samples. The proposed method could be applied to any existing CNNs. After testing on multiple datasets, the results show this method can improve the performance of CNN on FER effectively.

Large training samples can help solve the problem of overfitting. However, usually it's costly to get large facial expression image samples. Most of the time there is not so large amount of samples. The proposed method is suitable for small datasets. This method aims to learn knowledge from one dataset and transfer the knowledge to other datasets. Combined with samples generated by GAN, the proposed method can work on small unlabelled samples.

The proposed method is evaluated on several datasets. The validation set is different from training set. Two models are chosen to be improved by the proposed method, AlexNet and VGG 11. For each model, the last fully connected layer is modified. Before training, the images in the datasets are cropped to the same size. The GAN is also trained on cropped databases for 5000 epochs to produce large amounts of samples.

The original datasets with labels are used for training the model. The objective dataset consists of unlabelled samples generated by GAN. The model will be tested on objective dataset. When FER-2013 is used as the original dataset, Alexnet is the CNN model, JAFFE, MMI and CK+ are the objective datasets, the accuracy is improved by 3.76%, 3.72%, and 4.41% on them.[19] When the original dataset is FER-2013, the objective dataset is JAFFE, the CNN model is VGG-11, the accuracy is improved by 15.02%.[19] Then, when smaller dataset is used as the original dataset, like CK+, the accuracy of the model is also improved. The result shows the proposed method can increase the cross-dataset accuracy of CNN on FER effectively. Compared with other published method, the proposed method do not need the objective dataset to be labelled.

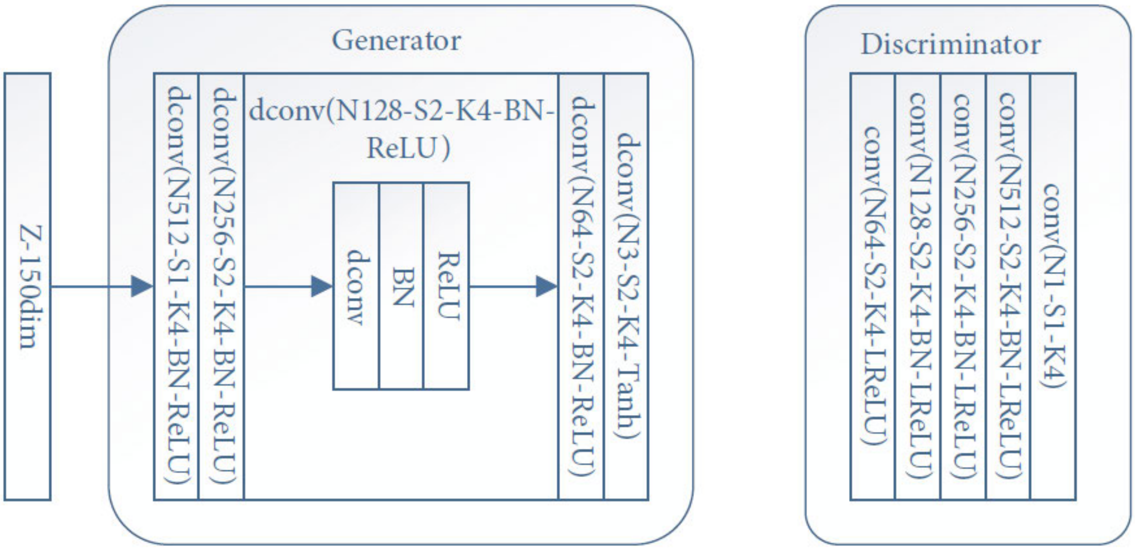
Z-150dim
Generator
dconv(N512-S1-K4-BN-ReLU)
dconv(N256-S2-K4-BN-ReLU)
dconv(N128-S2-K4-BN-ReLU)
dconv
BN
ReLU
dconv(N64-S2-K4-BN-ReLU)
dconv(N3-S2-K4-Tanh)
Discriminator
conv(N64-S2-K4-LReLU)
conv(N128-S2-K4-BN-LReLU)
conv(N256-S2-K4-BN-LReLU)
conv(N512-S2-K4-BN-LReLU)
conv(N1-S1-K4)

Figure 3: The Structure of GAN netowrk

3 Action Units in CNN for FER

3.1 Whether Deep Neural Networks Learns Facial Action Units When Doing Expression Recognition

The expression of human face consists of multiple facial action units. To study how much knowledge CNN learns from the dataset, we need to visualize the neurons in the network using deconvolutional network. The result shows that the visualized neurons correspond to the facial action units.

The network used in the experiment contains three convolutional layers. Consider the third convolutional layer and N images that produce strongest response in training set, then keep the strongest neuron high and remove all other activations, reconstruct pixel data using deconvolutional network. The reconstruction employs Guided Backpropagation. Analysing the patterns learned in Toronto Face Dataset, the result shows that some filters are sensitive to region that aligns with several facial action units.[20] On CK+ dataset, the feature patterns in CNN model are defined clearly and fits well with facial action units.

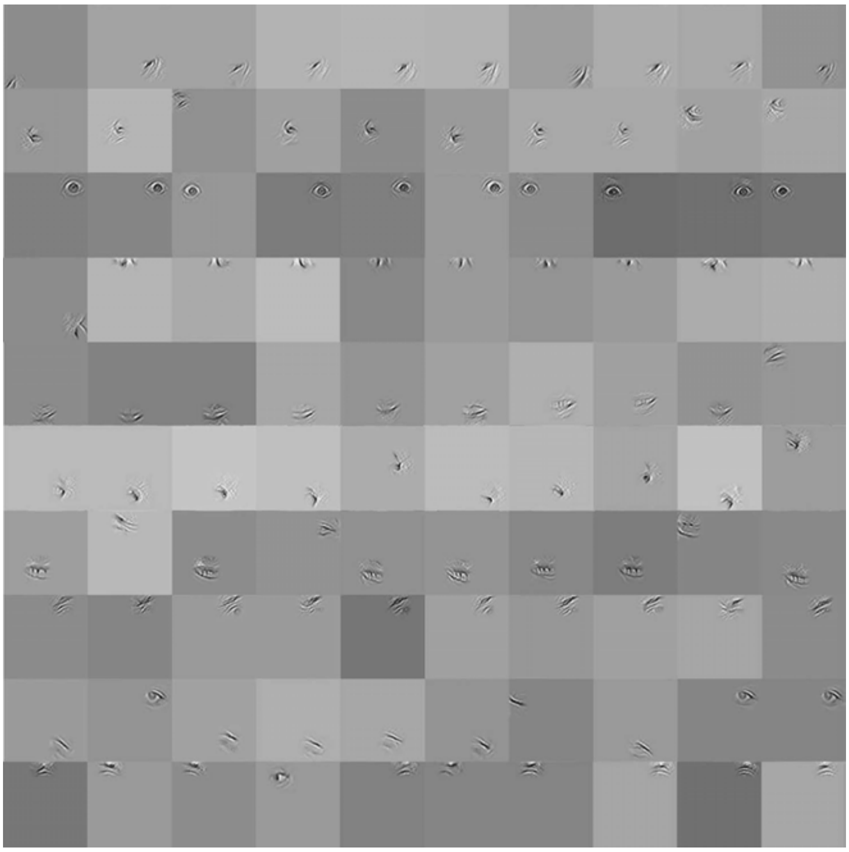

Figure 4: Visualization of spatial patterns

3.2 Deep Explanation Model for Facial Expression Recognition Through Facial Action Coding Unit

The expression is related to the facial action units. In the datasets, AUs have some relationship with the label. Sunbin et al proposed a model to explain the classification results of the CNN. The CNN model is trained on CK+ dataset and classifies images according to the features learned by the CNN model. The explanation model classifies multiple facial action units according to the features. The experiment shows that explanation model will generate AUs well only when the features and emotion category are obtained from CNN model.[21]

The facial expression can be decomposed to several facial action units. AUs generated from the decision process could be the foundation of how model make decisions. We assume we could explain the decision of CNN model by the AUs. The proposed model can explain the decisions of CNN model well, which classify the facial images by different emotion. The explanation model takes features and decision generated by CNN model and outputs AUs to explain the CNN model.[21]

The deep explanation model includes interpretable model and explanation model. Explanation model generates explanation from bounding box or text or generates visualisations with pixel-level precision. The proposed explanation model is based on the explanation model. The proposed model consists of CNN FER classifier and DNN explanation model. The DNN explanation model explains the decision made by CNN model. The explanation model classifies AUs according to the features learned by CNN FER model. Among most of the AUs, the proposed model performs better than other models.[?]

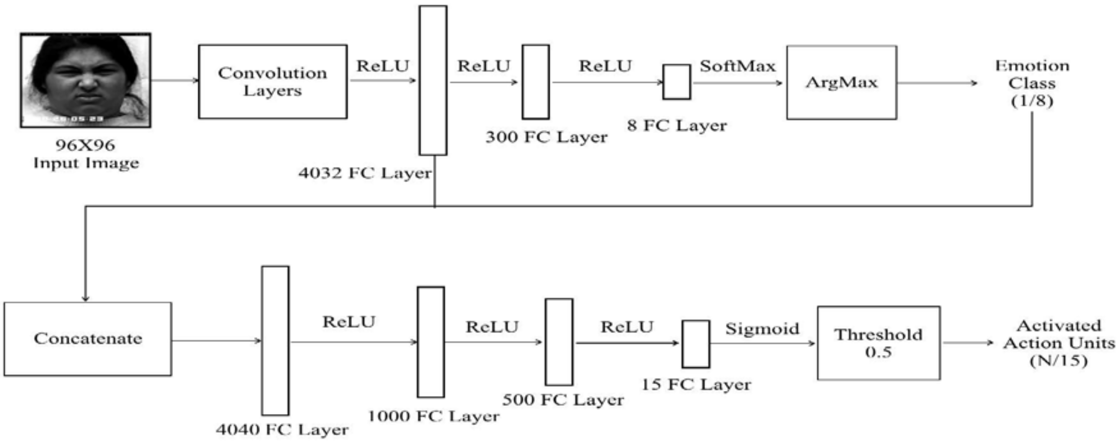

Figure 5: DNN Explanation model

4 Summary & Conclusion

In recent years, there are many works on FER CNN. There are lots of progress in the development of FER methods. The conventional CNN has shown high accuracy on all kinds of datasets. To solve the dynamic FER problem, we can add the objective function that reduces the difference of the same facial expression with variations. When it comes to the FER with occlusion, patch-gated CNN and CNN with attention mechanism is effective. The main idea of them is to recognise the occluded part of the face and give them lower weight so that the attention of the system will be transferred to the un-occlude part of the face.

For 3D FER, the fast and light manifold CNN shows high accuracy while saving lots of memory. Another way is the deep feature fusion CNN method. It combines geometric and texture features. The result shows that by combining multiple features can improve the CNN network on 3D FER datasets.

The FRR CNN can reduce the feature redundancy of CNN on FER problem. To improve the cross-dataset performance, we need a dataset with large number of images. However, it's hard to get so many samples. The generative adversarial network can produce new unlabelled samples by fine-tuning the original images. The generated samples could be used as the testing set.

There are also some researches that explore what the CNN learns on the dataset through analysing the AU generated from the features produced by CNN. The two researches mentioned previously show that according to the action units reproduced by the network, these AU are highly related to the visualized pixel pattern, which verifies that the network did learn from the datasets.

References

[1] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[2] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015.

[3] Ke Shan, Junqi Guo, Wenwan You, Di Lu, and Rongfang Bie. Automatic facial expression recognition based on a deep convolutional-neural-network structure. In *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 123–128. IEEE, 2017.

[4] Heechul Jung, Sihaeng Lee, Sunjeong Park, Byungju Kim, Junmo Kim, Injae Lee, and Chunghyun Ahn. Development of deep learning-based facial expression recognition system. In *2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, pages 1–4. IEEE, 2015.

[5] M Shin, M Kim, and DS Kwon. Baseline cnn structure analysis for facial expression recognition. 25th ieee int. In *Symp. Robot Hum. Interact. Commun. RO-MAN*, volume 2016, 2016.

[6] Wissam J Baddar, Dae Hoe Kim, and Yong Man Ro. Learning features robust to image variations with siamese networks for facial expression recognition. In *International Conference on Multimedia Modeling*, pages 189–200. Springer, 2017.

[7] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated cnn for occlusion-aware facial expression recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2209–2214. IEEE, 2018.

[8] Siyue Xie, Haifeng Hu, and Yongbo Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition*, 92:177–191, 2019.

[9] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[10] Zhixing Chen, Di Huang, Yunhong Wang, and Liming Chen. Fast and light manifold cnn based 3d facial expression recognition across pose variations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 229–238, 2018.

[11] Sa Wang, Zhengxin Cheng, Xiaoming Deng, Liang Chang, Fuqing Duan, and Ke Lu. Leveraging 3d blendshape for facial expression recognition using cnn. *Science China Information Sciences*, 63(2):1–3, 2020.

[12] Kuang Liu, Mingmin Zhang, and Zhigeng Pan. Facial expression recognition with cnn ensemble. In *2016 international conference on cyberworlds (CW)*, pages 163–166. IEEE, 2016.

[13] Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831, 2017.

[14] Chao Li, Ning Ma, and Yalin Deng. Multi-network fusion based on cnn for facial expression recognition. In *2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018)*, pages 166–169. Atlantis Press, 2018.

[15] Kun Tian, Liaoyuan Zeng, Sean McGrath, Qian Yin, and Wenyi Wang. 3d facial expression recognition using deep feature fusion cnn. In *2019 30th Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE, 2019.

[16] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Multi-region ensemble convolutional neural network for facial expression recognition. In *International Conference on Artificial Neural Networks*, pages 84–94. Springer, 2018.

[17] Guihua Wen, Zhi Hou, Huihui Li, Danyang Li, Lijun Jiang, and Eryang Xun. Ensemble of deep neural networks with probability-based fusion for facial expression recognition. *Cognitive Computation*, 9(5):597–610, 2017.

[18] Siyue Xie and Haifeng Hu. Facial expression recognition with frr-cnn. *Electronics Letters*, 53(4):235–237, 2017.

[19] Xiaoqing Wang, Xiangjun Wang, and Yubo Ni. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience*, 2018, 2018.

[20] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE international conference on computer vision workshops*, pages 19–27, 2015.

[21] Sunbin Kim and Hyeoncheol Kim. Deep explanation model for facial expression recognition through facial action coding unit. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4. IEEE, 2019.