A comparative study of deep convolutional neural network-based facial expression regression

Zhengyuan Sun

School of Information and Engineering, Xi'an Technology and Business College, Xi'an, Shaanxi, 710200, China

bbates61558@student.napavalley.edu

Abstract. Facial expression is an important way for people to express their emotions. Recently, with the advancement of the computer vision, facial expression recognition has become a current research hotspot and has made significant progress, which can be utilized in the interaction between humans and computers, emotional computing and other computer vision fields, the evolution of artificial intelligence and deep learning has better promoted the research of facial expression recognition. Conventional approaches are largely based on machine learning, which leverages artificial way for feature extraction. The extracted facial expression features are interfered by human factors, so that the trained classifier cannot effectively interpret the expression information, which ultimately leads to insufficient model generalization ability and low recognition accuracy. However, deep learning-based facial expression recognition in real scenes is still affected by factors such as human pose, different degrees of facial occlusion, background environment and light interference, and the recognition accuracy still needs to be further improved. This paper focuses on representative convolutional neural network architectures and summarizes the strengths, weaknesses, and innovations of these networks, through which the advancement of neural network architectures shows the great potential of the direction.

Keywords: Convolutional neural network, Facial expression, Deep learning.

1. Introduction

Facial expression is an important visual signal for human communication. Recently, with the advancement of the computer field, facial expression recognition (FER) has become a current research hotspot and has made significant progress, which can be utilized in human-computer interaction, emotional computing and other computer vision fields, the advancement of artificial intelligence and deep learning has better promoted the research of FER [1]. The traditional FER approaches based on machine learning adopts artificial way for feature extraction, and the extracted facial expression features are interfered by human factors, so that the trained classifier cannot effectively interpret the expression information, which ultimately leads to insufficient model generalization ability and low recognition accuracy. However, deep learning-based FER is still affected by factors such as human body posture, different degrees of facial occlusion, background environment and light interference, and the recognition accuracy still needs to be further improved.

This paper focuses on representative neural network models and analyzes the strengths and weaknesses of each network model as well as the innovations developed on top of the original model, to continuously improve the convolutional neural network model and enhance the feature expression ability and recognition accuracy of the network. The nature, advantages and disadvantages of different activation functions and loss functions are investigated in terms of their influence on the training of the network and the feature expression ability of the model.

2. Preliminary Knowledge

2.1. Convolutional Neural Networks (CNNs)

One of the deep learning algorithms that best exemplifies this is CNN, which has rich data representation learning capability. Compared with traditional solutions, it can automatically extract descriptive features. In addition, conventional algorithms generally preprocess data first, and then extract and locate features, CNN generally input the original data and shape of the image as information, thus shortening the algorithm's time. As shown in Figure 1, the structure of CNN, in order to facilitate the operation of the subsequent steps, usually in the input layer of data preprocessing work, and then through the convolutional layer and excitation layer for feature extraction, pooling layer is responsible for the extracted features of data compression, and finally through the fully connected layer to achieve the classification [2].

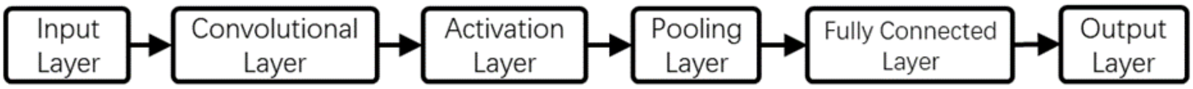

Figure 1. Architecture of CNN (Figure Credits: Original).

2.2. Conventional Machine Learning-Based Facial Expression

In the traditional machine learning facial expression recognition research, feature extraction is mainly the use of computers to calculate and process the digitized facial expression data, and then use mathematical methods are used to extract some features, which is applied to the feature extraction algorithm because of the extraction of some of the features of the facial image, so to a certain extent, also plays a role in reducing the feature dimension, machine learning-based facial expression feature extraction is generally divided into Two methods: overall extraction and local extraction.

Holistic extraction: the core idea of the holistic extraction method is to holistic human facial expression for feature extraction. When a person produces emotional changes, the human facial organs will undergo more obvious deformation, which will have an impact about the face's global information, then the expression features can be extracted from the global perspective.

Localized extraction: Changes in facial expressions are not only reflected in the whole, but there are also local differences, which requires localized extraction of features. The human face is divided into several parts, and the importance of each part is not the same, using the localization method to extract the features of the part with relatively high importance, and the feature analysis of the part with less importance is weakened [3].

2.3. Deep Learning-Based Facial Expression

The term Deep Learning (DL) was first coined by Hinton of Canada, who is also known as the father of neural networks. Hinton and Ruslan solved the trouble of eliminating the gradient vanish in the training of deep networks in a 2006 paper, and since then the wave of Deep Learning has begun to rise, and research based on Deep Learning has rapidly developed in fields such as healthcare, finance, and drones. The research based on deep learning is rapidly developing in the fields of healthcare, finance, and driverless driving.

At present, image classification is inextricably linked to deep learning, and FER is a subtask of image classification. Combining deep learning-based algorithms with facial expression recognition has many benefits compared to traditional machine learning-based expression recognition methods. First, feature extraction and feature classification in machine learning-based facial expression recognition algorithms are two independent steps that need to be studied separately, while feature extraction and feature classification in deep learning are designed and optimized in the same algorithm, which can simplify the algorithm and reduce the complexity of the algorithm; second, feature extraction in the traditional machine learning method relies on manually extracting features, which is not only tedious, but also the extracted features are more complex. Secondly, feature extraction in traditional machine learning methods relies on manual extraction of features, which is not only cumbersome but also the extracted features are easily interfered by human factors, whereas in deep learning, the features are automatically extracted by the neural network that has a better ability to extract the features from the image, which makes the FER methods according to deep learning have a better ability to express the features.

3. FER using a convolutional neural network

3.1. POSTER

The POSTER based network architecture is optimized and improved for the FER task, but the network structure of POSTER is not only complex but also leads to expensive computational costs, so in then based on POSTER, POSTER++ is proposed in order to alleviate the computational pressure. POSTER++ improves on POSTER in three ways: Cross Fusion, Dual Streaming and Multi-scale Feature Extraction [4].

In cross-fertilization this work uses the window-based cross-attention mechanism instead of the original POSTER cross-attention mechanism. The window-based attention mechanism not only provides good linear computational complexity, but also enhances certain modeling capabilities, and removes the dual-stream design's image-to-landmark branch, which contains two main branches in POSTER: image-to-landmark and landmark-to-image. The landmark-to-image branch, as the core of POSTER, is essential for solving inter-class similarity and intra-class differences, and the image-to-landmark branch only provides some information that is not taken into account in the structure of landmarks, so the image-to-landmark branch is deleted from the dual-stream design to reduce the computational rate. branch from the dual-stream design, which reduces the computation rate. For multi-scale feature extraction, POSTER++ will not use the pyramid architecture for feature extraction, but will directly extract multi-scale features from the focus part of the image and the facial landmark detector, through the above design POSTER++ will be more powerful than before, and at the same time, through the testing of multiple datasets POSTER++ achieves 92.21% of the total number of features extracted in RAF-DB, 92.21% in AffectNet (7 cl) and 92.21% in AffectNet (7 cl), and 92.21% in AffectNet (7 cl). on AffectNet (7 cls), 67.49% on RAF-DB, and 63.77% on AffectNet (8 cls). This proves the reliability and effectiveness of the optimization improvements.

3.2. Poker Face Vision Transformer

Representation learning and feature decoding has been widely researched aspects in the domain of FER, the prevailing ambiguity of emotional expression labels is difficult for those methods based on traditional visual learning [5]. Also, the mapping from facial expression images to emotional labels lacks explicit supervised signals of facial details in direct learning.

This essay suggests a new FER model known as Poker Face Vision Transformer (PF-ViT), which is composed of five parts, the first part is the encoder: a complete mapping of facial expressions to an encoder that can represent them; the second part is the separator: a separator that decomposes the representations into emotional components and orthogonal residuals; the third part is the generator: it can reconstruct the expression face and then synthesize the poker face; the fourth part is the discriminator: it can distinguish the fake face produced by the generator and train adversarial with the encoder and the generator; the fifth part is the classification head: it is used to recognize the emotion.

This model separates and recognizes disturbing and difficult to discern agnostic emotions from static facial images by generating specific poker faces, where this work considers an expression face to be the result of a series of facial muscle movements in an expressionless face, inspired by facial action coding systems. The proposed PF-ViT uses the vanilla Vision transformer and starts to work on a large dataset without any large dataset without any expression labels, performed with training and obtained excellent resolution. However, it is now experimentally found that ViTs need more training to obtain better feature representations, since ViTs use much weaker inductive bias compared to CNN and hybrid ViTs. However, it could be found that by separating and recognizing emotions in facial expressions through the auxiliary task of poker face generation without the need for paired images, the method mitigates the effects of interference and achieves significant improvements. By working on unlabeled facial expression data, results confirm the great potential of PF-ViT in FER. Results show that a tiny ViT with only 5M parameters can already achieve competitive performance compared to previous SOTA methods.

This method again innovated new accuracy peaks with accuracies of 92.07%, 67.23%, 64.10%, and 91.16% at the RAF-DB, AffectNet-7, with AffectNet-8, and FERPlus datasets, respectively. Besides, for the first time, PF-ViT successfully synthesized a very realistic poker face with any expression for the first time using the vanilla Transformer backbone.

3.3. Distracted Attention Network (DAN)

Two key observations of biological visual perception are the basis of DAN. It is observed that multiple facial expression categories have essentially similar beginnings in appearance with very small disparities, while facial expressions express themselves through multiple facial feature regions, but in order to discriminate each facial expression, a coding method for high order interaction between local features is needed, and with these issues as the main focus, three components of the DAN are proposed: the feature clustering network (FCN) , Multiple Attention Network (MAN) and Attention Fusion Network (AFN) [6].

FCN: Considering the performance and parameters of the model, a residual network is used as the backbone and a loss function for discrimination, called affinity loss, is proposed; MAN: Joint inference of multiple local regions and contains several parallel heads, which are independent of each other; AFN: AFN uses a log-softmax function to scale the attention feature vectors, which is used to stress the most important parts, on the basis of which a zone loss method is recommend to avoid overlapping of attention and to focus on in different critical regions, and finally the attention feature vectors are merged for the calculation of output probability.

The specific process is that FCN takes face images are input and delivers embeddings representing recognition capabilities. Then, a MAN is used to learn different attention maps that capture several items facial expression regions that are cross-sectional. These attention then seeks to be trained by AFN and to focus on different areas. Finally, The features of all the attentional heads are fused by AFN, which then makes predictions about the representation categories of the input images.

Results demonstrate that the accuracy of the DAN method on AffectNet-8, AffectNet-7, RAF-DB, and SFEW 2.0 is 62.09%, 65.69%, 89.70%, and 53.18%, respectively, which illustrates the feasibility of this network model on the face expression recognition system.

3.4. Hybrid Multi-Task Learning (HMTL)

When investigating the effect of fine-grained FER with ImageNet pre-trained weights, it could be found that restarting the training was better than pre-training on ImageNet if the images could be sufficiently augmented, this led to a method for improving fine-grained and field FER called HMTL, where classically supervised learning (SL) during the Auxiliary tasks are performed by HMTL in the shape of Multi-task Learning (MTL) using Self-supervised Learning (SSL), which is utilized during training to append information from the images that are derived from the main fine-grained SL task [7]. This work investigates how two custom versions (puzzles and fixes) can be used by the proposed HMTL design in the FER domain, achieving advanced results in the AffectNet benchmarks through both pattern of HMTL and no prior training with extra data. Text results with HMTL shows the differences and advantages of HTML that are pre-trained and presented by SSL, while experiments on gender recognition and head pose estimation taskes illustrate the potential for fine-grained facial representation is improved by HMTL.

The results show that by using HMTL approach on the AffectNet dataset, both types of sentiment recognition (i.e., dimensionality and categorization) are significantly improved in all performances even with a severely imbalanced training set of 20%.

3.5. Audio Visual Emotion Recognition (AVER)

The fusion of two deep neural networks becomes AVER, which is the depth CNN model extracted by FER architecture and the second modified FER VGGish model are pre-trained separate audio and visual depth CNN recognition modules respectively. A very standard, very formal method of fusion is used to merge audio and video feature representations, and the spatial and temporal representations are processed using RNN to model temporal dynamics on the basis of advance training flat and independent audio and CNN modules [8].

The Deep CNN structure in multiple modes consists of three parts: a visual facial expression embedding network, an audio embedding network for feeling identification, and an audio-visual fusion model; its main role is to fuse using a model-level fusion method, Multiple feature representations of audio and visual direction are created., and then use a RNN to capture changes in time. This method greatly outperforms others in Predict valence states on the RECOLA data set. In addition, the performance of the proposed facial feature extraction network is better than that of other networks about When we compared Google's facial expression database with the AffectNet Facial expression dataset, we found that the results were very good. For AffectNet with eight discrete facial expression categories, a logistic regression model can be trained based on features extracted by the student network for the entire AffectNet training set, with an accuracy of 61.6% using this dimension.

3.6. Pyramid with a Super-Resolution (PSR)

The FER is a very interesting task, although it can greatly improve the ability of human-computer interaction but is always constrained for some reasons when applied out of the lab in the real world, to solve this problem, Individual field (ITW) images can be automatically recognized by FER was proposed, but ITW photos are harmed from practical troubles with respect to form, orientation and image emotion degree [9]. In order to put together an optimized model, the research team proposed a pyramid with a super-resolution (PSR) network architecture for the ITW FER task, along with the introduction of A loss function, which apply some knowledge to each expression in the FER task, along with a loss function on three of the most popular Texts are conducted on Three of the most popular ITW data sets on the market show that PSR is superior to other algorithms at this stage.

Despite the success of data in the lab, the rise of ITW in these years has posed significant throw down the gauntlet to researchers. One of the biggest reasons for this is that data in the lab is data created under controlled conditions, in comparison to ITW datasets which loud, not accurate enough uncontrollable. The research team found about both observations are FER ITW data sets task, respectively, that ITW datasets have varying image sizes; CNNs are usually sensitized based on the input size of the image, so the researchers proposed a PSR ITW FER tasks of different image sizes of the network architecture are processed.

The PSR has six modules which are a network architecture with respect to spatial transformer, a high-performance processor for processing facial features, a connection block that receives the bit flow from the physical layer into a frame stream for use by the network layer and a non-exclusive two processors, a face scaling processor. Through experimental findings comparing Seven and eight facial expression structures are included in dataset, this model has an accuracy of 60.68% in the classification of eight emotions, exceeding the state-of-the-art of 59.58% currently realized by Georgescuetal.

3.7. EfficientFace

A team proposed an efficient and robust FER network called Efficient Face, which not only has few parameters but also has higher accuracy and robustness for FER [10]. so as to maintain the lightweight system's durability, a deep convolutional based local feature extractor and null-channel spatial modulator were designed, and the use of the deep convolution so that the network is able to recognize both locally and globally significant facial features.

Due to the limited capability of lightweight networks in feature learning and the challenges of occlusion and pose changes in field FER, direct use of lightweight networks for FER may lead to poor performance in terms of accuracy and robustness, Experiments conducted on datasets led the researcher to introduce the Labeled Distribution Learning (LDL) method as a training strategy. the occlusion and pose are lifelike changes have shown that effentface is also stable when the character is obscured or when the body moves and changes.

Experiments on real occlusion and pose change datasets show the robustness of the proposed method for occlusion and pose change problems. With a small number of parameters and FLOPs, this method achieve the most advanced results in the RAF-DB, CAER-S, AffectNet7 datasets, resulting in significant results on the AffectNet-8 dataset.

3.8. Aff-Wild2

With the rise of deep learning models, various deep learning models are used for default selection of computer vision characters, but when wild-FER emerged, many existing databases of small wild have many problems such as the size of the database is too small, contains few topics, the content is not audio-visual, and there is no annotation of behavioral values. To address these problems, researchers expanded one of the largest field-available databases (Aff-Wild) to study ongoing emotions such as value and novelty. In addition, the researchers labeled the database sections [11].

Specifically the researchers conducted extensive experiments on CNN and CNN-RNN architectures using visual and audio patterns while training them on Aff-Wild2 and then evaluating their performance on 10 publicly available sentiment databases, in the study it was found that at the time of the study there was no database containing annotations for all the main behavioral tasks (VA estimation, AU detection, Expr classification) were annotated, so the researchers created a new dataset containing 260 videos with about 1.4 million frames annotated for VA and merged this with Aff-Wild finally generating what is known as the Aff-Wild2 database then annotated portions of Aff-Wild2 with AU and the seven basic expression labels to create about 398K and 403K AU and Expr annotations, respectively, and finally demonstrated the strong performance advantage of this network in FER through experimental data.

4. Results

AffectNet is a dataset created with the goal of providing a rich source of facial images for emotional analysis research. The dataset was constructed by collecting over 1 billion images from the internet and carefully screening and annotating them using computer vision techniques and machine learning algorithms. The AffectNet dataset is publicly available for research purposes. The results are demonstrated in Table 1.

Table 1. Performance comparison of CNN.

	Facial Expression Recognition (FER) on AffectNet	
name (of a thing)	Accuracy (8 emotion)	Accuracy (7 emotion)
POSTER++	63.77%	67.49%
PF-ViT	64.10%	67.23%
DAN	62.09%	65.69%
HMTL	61.72%	
AVER	61.60%	65.4%
ITW	60.68%	
EfficientFace	59.89%	60.68%
Aff-Wild	63%	

5. Conclusion

Times are progressing the level of science and technology is also progressing, while the algorithm is constantly improving, at the same time, the deep convolutional network is still the mainstream model in deep learning and machine learning, other recognition compared to the face expression has been developed for a long time, with a large number of theoretical support, at the same time in the industrial world has been used on a large scale, but due to the environment in the laboratory and the real environment is still a certain difference that However, due to the difference between the environment in the laboratory and the real environment, the accuracy rate is not high in the noisy environment in reality, but the expression recognition has an irreplaceable position in clinical medicine, human-computer interaction, and asset security, which has a broad application prospect, and at the same time, it needs to increase the research efforts in the recognition technology in the outdoor area. In conclusion, the powerful feature extraction capability of convolutional neural networks greatly promotes the development of expression recognition, and expression recognition based on CNN has great potential for development and application.

References

[1] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing, 13(3), 1195-1215.

[2] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.

[3] Tian, Y., Kanade, T., & Cohn, J. F. (2011). Facial expression recognition. Handbook of face recognition, 487-519.

[4] Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., & Huang, A. (2023). POSTER V2: A simpler and stronger facial expression recognition network. arXiv preprint arXiv:2301.12149.

[5] Li, J., Nie, J., Guo, D., Hong, R., Wang, M., (2023). Emotion Separation and Recognition from a Facial Expression by Generating the Poker Face with Vision Transformers. arXiv preprint arXiv:2207.11081

[6] Wen, Z., Lin, W., Wang, T., & Xu, G. (2023). Distract your attention: Multi-head cross attention network for facial expression recognition. Biomimetics, 8(2), 199.

[7] Pourmirzaei, M., Montazer, G. A., & Esmaili, F. (2021). Using self-supervised auxiliary tasks to improve fine-grained facial representation. arXiv preprint arXiv:2105.06421.

[8] Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognition Letters, 146, 1-7.

[9] Vo, T. H., Lee, G. S., Yang, H. J., & Kim, S. H. (2020). Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access, 8, 131988-132001.

[10] Wang, G., Li, J., Wu, Z., Xu, J., Shen, J., & Yang, W. (2023). EfficientFace: An Efficient Deep Network with Feature Enhancement for Accurate Face Detection. arXiv preprint arXiv:2302.11816.

[11] Kollias, D., & Zafeiriou, S. (2019). Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855.