Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 1527–1534

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Deep self-attention network for facial emotion recognition

Arpita Gupta[a], Subrahmanyam Arunachalam[a], Ramadoss Balakrishnan[a,*]

[a] *National Institute of Technology, Tiruchirappalli,620015, India*

Abstract

Human emotion detection is one of the major problems in computer vision. Human emotions consist of several sub-emotions which are difficult to classify into a specific class. In this proposed work, we have tried to classify the emotions into 6 basic categories (happy, sad, disgust, fear, surprise, anger) and neutral human emotions. We have proposed deep-learning framework which consist of CNN, ResNet and attention block which gives visual perceptibility to the network. The proposed model has greater applicability in real life facial emotion detection. The proposed model has achieved satisfactory result and has shown effective results on FER dataset.

Keywords: *Deep Learning, Attention, ResNet, Facial Emotion Recognition;*

1. Introduction

In the past decade with the growing computational power of the machine's, scientists/researchers are trying to replicate the working of human brain in many tasks [1]. One of the areas of research in human emotion detection. In the recent years the data collection has grown exponentially with cameras placed in the public network and the specialty datasets are being collected and maintained for better results for specific tasks.

Human facial expression e.g. happy, disgust, sad, fear, surprise, and anger [2][3][4][5] are one of the methods for non-verbal communication. Computer vision and its techniques are trying to automate and solve the challenging problem of identifying the human's emotion accurately

* Corresponding author. Tel.:+91-989-460-2642.
E-mail address: arpitagupta2993@gmail.com

10.1016/j.procs.2020.04.163

Identifying the emotion of human is one of the most interesting and useful applications in computer vision, it could be used for identifying the mental state of the human [6], detecting the lies, understanding the mood[7], treating mental patients, to make the search system better, counselling etc[8]. Detecting of the expression is done by human in daily life without any problem and effortlessly [9], but it's hard to make machines understand this. There are six basic expressions of human anger, sadness, surprise, happy, disgust and fear. Most of the current automated expression detecting methods are trying to recognize these expressions. Some of the systems can also recognize neutral expressions. There are two types of methods in automated facial expression detection (FER)[10]; static [11][12][13]and dynamic[14][15]. Static expression detection is the one in which only single frame of human face is being used while in dynamic multiple frames of human expression is used. The proposed model uses static method expression detection and can detect 6 basic and neutral emotion.

Detecting the facial emotion and analysing consist of three steps [9]: acquisition of face [16][17][18], extraction and representation of facial data and recognition of facial emotions. Once the face shape is detected the changes are used to detect the emotion. These facial changes are the expressions or features which are used to classify as an emotion. To detect the changes there are two methods geometric feature based [19][20][21][22] and appearance based [13][18]. Geometric feature-based methods are the one in which the distance between the different components of the face is calculated and expression is detected while in appearance-based feature vector is extracted from the face and these features are used to detect expression of emotion of the face [9].

Nomenclature

I	Extracted image features through prior convolution
Q	Fully connected layer of dimension 1 applied on the extracted image features I
T(u,v)	2D filter
y_i	image
C_{yi}	Category of image
I(r,c)	Image on which convolution is applied
N	Number of datapoints
CNN	Convolution Neural Network
ResNet	Residual Network
FER	Facial Emotion Recognition

2. Related Work

Deep learning has the great power of learning features which is being used in FER, but some of the characteristics create problem in applying deep learning as the current available datasets are not sufficient to train and achieve the accuracy like humans [23]. The databases should consist of all the variable like age, gender, ethnic background, level of expression [24]. When the facial expression is recognized in the controlled environment the variation in pose, occlusion or lighting could affect the results [23].

2.1 CNN

CNN is widely used in various application of computer vision and facial emotion recognition is one them [23]. CNN has proved to be vigorous to different variations in the facial characteristics and has outperformed the other techniques in the case of an unseen pose variation of the face [25][26]. CNN could be used to recognize the facial expression problems such as rotation, translation and subject independence.

CNN comprises of three sorts of heterogeneous layers: convolutional layers, pooling layers, and completely associated layers. The convolutional layer features a set of learnable channels to convolve through the full input picture and deliver different types of actuation highlight maps. The benefits of convolution layer are: neighborhood network, which learns correlations among neighboring pixels; weight sharing within the same feature outline, which significantly diminishes the number of the parameters to be learned; and shift-invariance to the area of the question. The convolution layer is followed by pooling layer which reduces computation cost of the network and the spatial size of the feature maps. Two most commonly utilized nonlinear down-sampling procedures for translation invariance are average and max pooling. To ensure that all neurons in the layer are fully connected to activations in the previous layer the fully connected layer is used [23].

CNN models are being used in FER since a long time now and have shown promising results. There are CNN which are region based [27][28][29] which is employed in feature learning of FER. The high-quality region could also be concentrated on for feature learning by CNN models [30][31]. CNN could also be used to learn the spatio-temporal features of the datasets [32] or to find the actions in the images by considering the sequence of images [33]. CNN is being in different scenarios of recognition in FER [34][35].

2.2 ResNet and Attention

Attention is a mechanism relating to different pixels in the same channel to efficiently understand the picture. Research from the process of human perception shows the importance of attention which uses top knowledge to direct the process of bottom-up feedforward [36]. Attention is been used in many deep learning networks recently. In the training phase Deep Boltzman Machine also uses attention to rebuild [37]. In the field of NLP attention is being used very often to find the sequence in the text by applying it with RNN and LSTM [38][39][40][36][41]. Attention is used to take in account the past information or the top information and then find the next most relevant feature in the future or next sequence of information.

Residual Networks help in reducing the degradation problem and increase the learning power by making the use of unnecessary layers and making them do the identity mapping. It helps in using the depth of the deep networks for better performance. Residual networks are being used in many fields of computer vison and facial emotion recognition is one of them. The facial affect estimation uses combination of CNN and ResNet [42] to find the level of valence and arousal in the wild. In recent research different models are being combined like CNN and ResNet to find the emotion of the face also [43] which has shown great results. Our research has combined CNN, ResNet and attention mechanism to find the emotion of the face and has shown promising results as one networks helps other to concentrate on the better feature further in the next layer computation.

3. Proposed Work

The framework of our proposed method consists of convolutional layers with residual connections. Attention block is introduced in between the convolutional and dense layers in order to allow the model to greater visual perceptibility as already shown in [44]. Attention is implemented using the following equation 1:

$$\text{Attention}\ (Q, I) = \text{softmax}(\text{Flatten}(\tanh(Q)) * I \qquad (1)$$

ResNet is the network which consists of residual connections which overcomes the problem of vanishing gradients which usually occurs in very deep networks. The convolutional layers used are of 3*3 kernel size with 32 or 64 filters used accordingly. Dropout of 25% and 50% are introduced in the network as shown in the fig.1. The convolutional as well as the dense layers are sandwiched by batch normalisation and ReLU layers. The final layer of the network is a fully connected layer which gives 7 outputs between 0-1 which is maintained by the softmax function. Each of the values indicate the probability of the image belonging to that class.

Two other frameworks are also proposed here which does not achieve as high accuracy as the above one but is computationally less expensive. The second framework is a standard convolutional layer architecture with residual connections while the third framework consists of only convolutional and max pooling layers with no residual connections. The simple 2D convolution is given by the following equation 2:

$$J(r,c) = \sum_{u=-h}^{h} \sum_{v=-h}^{h} I(r+u, c+v) * T(u,v) \tag{2}$$

Window size is given by 2h+1. The training accuracy, validation accuracy and training time per epoch has been listed in table 1.

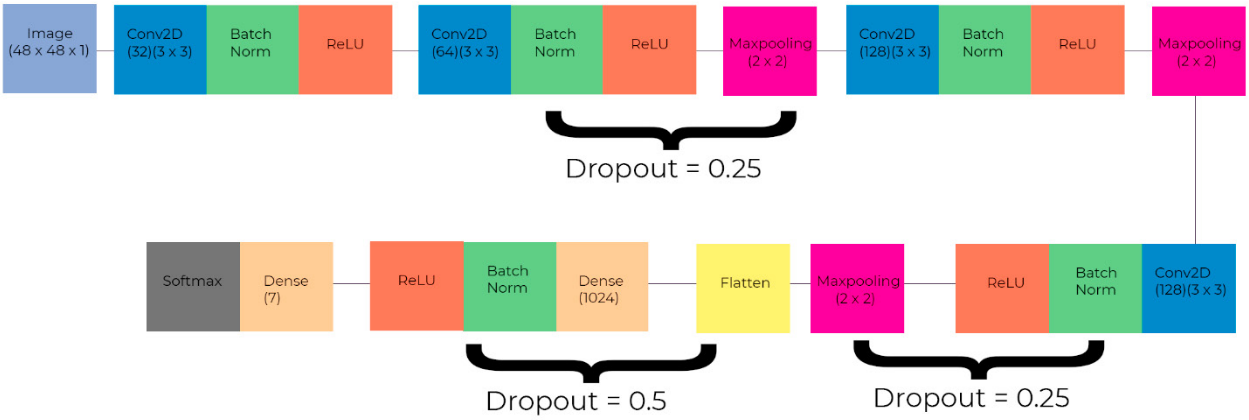


*Fig.1. Convolutional layers of 3*3 kernel size with 32, 64 and 128 filters used in succession. ReLU is used as the activation function. Dropout is used to avoid overfitting.*

There are totally 28,669 images in training part of FER dataset with 3955 images belonging to angry category, 436 images belonging to disgusted category, 4097 images of fearful category, 7215 images of happy category, 4965 images of neutral category, 4830 images of sad category and 3171 images of surprised category. All these images are resized to a shape of 48*48*1 (taking only the Y-channel of the YCbCr) before feeding into the network.

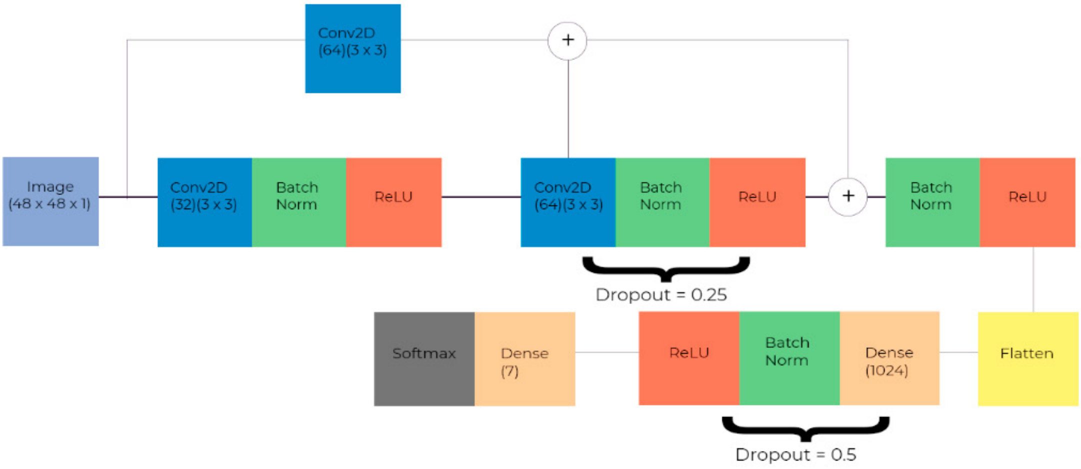


Fig. 2. Residual connections used to avoid the problem of vanishing gradients.

The convolutional network along with residual connections and attention mechanism performs exceedingly well. The residual connection layer is made sure to be of same shape as the main network by inserting a Conv2D operation. The attention model gives out values between 0 and 1 which is then multiplied with a 1936 * 64 sized matrix to give useful features. Useless feature vectors are multiplied by values close to zero whereas useful feature vectors are multiplied by values close to one.

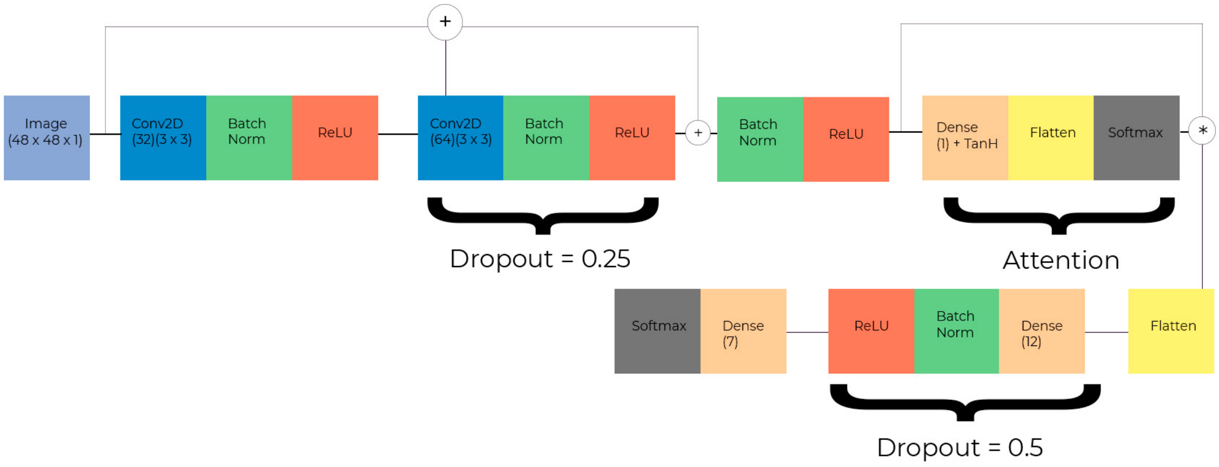


Fig. 3. Attention on extracted features giving better results.

4. Experimental Results

The network is trained on Keras library in Python for 50 epochs. Categorical cross-entropy is used as the loss function which is given by the following formula:

$$\frac{-1}{N}\sum_{i=1}^{N} log\, p_{model}\left(y_i \in C_{yi}\right)$$

$P_{model}(y_i \in C_{yi})$ is probability that y_i image belongs to category $C_{yi..}$

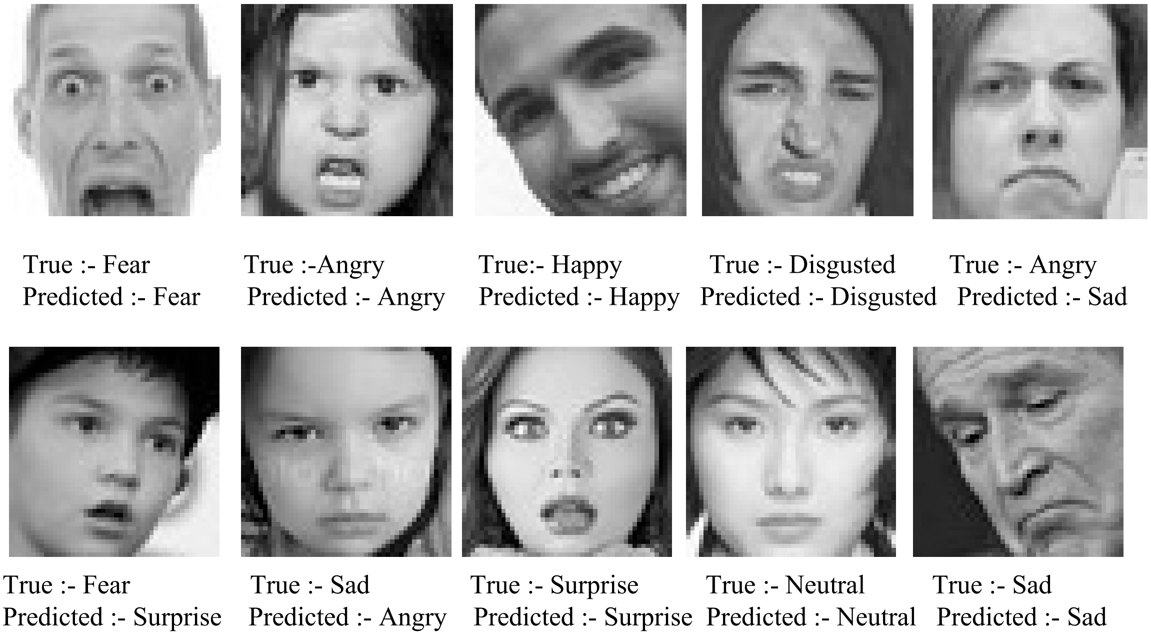
It is optimised using Adam optimiser with initial learning rate of 10^{-4} and decay of 10^{-6}. All the three networks were trained in NVIDIA P100 GPU and the results are tabulated as shown below in table 1.

Table 1. Details of the result obtained

Model	Training accuracy (%)	Testing accuracy (%)	Training time per epoch (sec)
Convolutional Network – A [1]	62.12	63.00	5
Convolutional Network – B [1]	60.48	53.00	7
Convolutional Network – C [1]	52.34	63.00	6
Convolutional Network (ours)	85.76	62.74	9
Convolutional + Residual connections (ours)	82.55	63.83	1880
Convolutional + Residual + Attention (ours)	77.78	64.40	2640

In [1] they have proposed a model with three different networks based on CNN named A, B and C. The network A is based on Krizhevsky and Hinton[45] and consisted of three convolution layer, two fully connected layers, max pooling and dropout. The network B is based on Alexnet consisting of three convolution layers, three connected layers, normalization and dropout. The network C is based on Gudi [46] consisting of one convolution layer, normalization, fully connected, dropout and max pooling.

In the proposed model with convolution networks only the training accuracy achieved is 85.76%, when convolution networks are added with residual networks accuracy achieved is 82.55% while when both attention and residual connections are added accuracy achieved is 77.78%. While the testing accuracy of the models are 62.74%, 63.83% and 64.40% respectively. The best accuracy achieved in training phase by existing networks is 62.12 while the best accuracy achieved by our proposed network is 85.76. The best accuracy achieved in testing phase by existing networks is 63 while the best accuracy achieved by our proposed network is 64.40. As can be seen in table 1. The proposed networks have outperformed the existing networks. The tradeoff between training time per epoch and testing accuracy could be seen in table 1 clearly.



True :- Fear
Predicted :- Fear

True :-Angry
Predicted :- Angry

True:- Happy
Predicted :- Happy

True :- Disgusted
Predicted :- Disgusted

True :- Angry
Predicted :- Sad

True :- Fear
Predicted :- Surprise

True :- Sad
Predicted :- Angry

True :- Surprise
Predicted :- Surprise

True :- Neutral
Predicted :- Neutral

True :- Sad
Predicted :- Sad

Fig. 4. Above images show the true emotion and the prediction by our proposed model.

5. Conclusion

In this paper we have proposed a deep self-attention network for facial emotion recognition which consists of attention block, residual connections and convolution networks. Every block of the network has a distinct function to perform as residual connections are helping in removing the vanishing gradients problem, convolution network is extracting the features and attention has proved effective as it gives better visual perceptibility to the network. The proposed model has outperformed the existing CNN based networks by achieving the higher accuracy of 85.76 in training phase and 64.40 in testing phase. The proposed model has shown better results on the FER dataset and could be used in real-time applications. Moreover, this network could be employed and extended in other computer vision tasks.

References

[1] Correa, E., A. Jonker, M. Ozo, and R. Stolk. "Emotion Recognition using deep convolutional neural networks." *Tech. Report IN4015* (2016).

[2] Filippini, Massimo, and Lester C. Hunt. "US residential energy demand and energy efficiency: A stochastic demand frontier approach." *Energy economics* 34, no. 5 (2012): 1484-1491.

[3] Darwin, Charles, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[4] Mehrabian, Albert. "Communication without words." *Psychology today* 2, no. 4 (1968).

[5] Ekman, Paul, and Wallace V. Friesen. "Constants across cultures in the face and emotion." *Journal of personality and social psychology* 17, no. 2 (1971): 124.

[6] Mandal, Manas K., Rakesh Pandey, and Akhouri B. Prasad. "Facial expressions of emotions and schizophrenia: a review." *Schizophrenia bulletin* 24, no. 3 (1998): 399-412.

[7] Dureha, Anukriti. "An accurate algorithm for generating a music playlist based on facial expressions." *International Journal of Computer Applications* 100, no. 9 (2014): 33-39.

[8] Kumari, Jyoti, R. Rajesh, and K. M. Pooja. "Facial expression recognition: A survey." *Procedia Computer Science* 58 (2015): 486-491.

[9] Jain, Anil K., and Stan Z. Li. *Handbook of face recognition*. New York: springer, 2011.

[10] Fasel, Beat, and Juergen Luettin. "Automatic facial expression analysis: a survey." *Pattern recognition* 36, no. 1 (2003): 259-275.

[11] Liu, Ping, Shizhong Han, Zibo Meng, and Yan Tong. "Facial expression recognition via a boosted deep belief network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805-1812. 2014.

[12] Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. "Facial expression recognition based on local binary patterns: A comprehensive study." *Image and vision Computing* 27, no. 6 (2009): 803-816.

[13] Liu, Weifeng, Caifeng Song, and Yanjiang Wang. "Facial expression recognition based on discriminative dictionary learning." In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 1839-1842. IEEE, 2012.

[14] Byeon, Young-Hyen, and Keun-Chang Kwak. "Facial expression recognition using 3d convolutional neural network." *International journal of advanced computer science and applications* 5, no. 12 (2014).

[15] Lien, James Jenn-Jier, Takeo Kanade, Jeffrey F. Cohn, and Ching-Chung Li. "Detection, tracking, and classification of action units in facial expression." *Robotics and Autonomous Systems* 31, no. 3 (2000): 131-146.

[16] Chen, Chih-Rung, Wei-Su Wong, and Ching-Te Chiu. "A 0.64 mm^2 Real-Time Cascade Face Detection Design Based on Reduced Two-Field Extraction." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 19, no. 11 (2010): 1937-1948.

[17] Garcia, Christophe, and Manolis Delakis. "Convolutional face finder: A neural architecture for fast and robust face detection." *IEEE Transactions on pattern analysis and machine intelligence* 26, no. 11 (2004): 1408-1423.

[18] Zhang, Zhiwei, Dong Yi, Zhen Lei, and Stan Z. Li. "Regularized transfer boosting for face detection across spectrum." *IEEE signal processing letters* 19, no. 3 (2011): 131-134.

[19] Bartlett, Marian Stewart, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. "Recognizing facial expression: machine learning and application to spontaneous behavior." In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 568-573. IEEE, 2005.

[20] Zhang, Zhengyou, Michael Lyons, Michael Schuster, and Shigeru Akamatsu. "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron." In *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*, pp. 454-459. IEEE, 1998.

[21] Yang, Peng, Qingshan Liu, and Dimitris N. Metaxas. "Boosting coded dynamic features for facial action units and facial expression recognition." In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-6. IEEE, 2007.

[22] Jain, Suyog, Changbo Hu, and Jake K. Aggarwal. "Facial expression recognition with temporal modeling of shapes." In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pp. 1642-1649. IEEE, 2011.

[23] Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." *arXiv preprint arXiv:1804.08348* (2018).

[24] Valstar, Michel F., Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. "Meta-analysis of the first facial expression recognition challenge." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, no. 4 (2012): 966-979.

[25] Fasel, Beat. "Robust face analysis using convolutional neural networks." In *Object recognition supported by user interaction for service robots*, vol. 2, pp. 40-43. IEEE, 2002.

[26] Fasel, Beat. "Head-pose invariant facial expression recognition using convolutional neural networks." In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, p. 529. IEEE Computer Society, 2002.

[27] Sun, Bo, Liandong Li, Guoyan Zhou, Xuewen Wu, Jun He, Lejun Yu, Dongxue Li, and Qinglan Wei. "Combining multimodal features within a fusion network for emotion recognition in the wild." In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 497-502. ACM, 2015.

[28] Sun, Bo, Liandong Li, Guoyan Zhou, and Jun He. "Facial expression recognition in the wild based on multimodal texture features." *Journal of Electronic Imaging* 25, no. 6 (2016): 061407.

[29] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587. 2014.

[30] Li, Jiaxing, Dexiang Zhang, Jingjing Zhang, Jun Zhang, Teng Li, Yi Xia, Qing Yan, and Lina Xun. "Facial expression recognition with faster R-CNN." *Procedia Computer Science* 107 (2017): 135-140.

[31] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In *Advances in neural information processing systems*, pp. 91-99. 2015.

[32] Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 4489-4497. 2015.

[33] Ji, Shuiwang, Wei Xu, Ming Yang, and Kai Yu. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 1 (2012): 221-231.

[34] Fan, Yin, Xiangju Lu, Dian Li, and Yuanliu Liu. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks." In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445-450. ACM, 2016.

[35] Nguyen, Dung, Kien Nguyen, Sridha Sridharan, Afsane Ghasemi, David Dean, and Clinton Fookes. "Deep spatio-temporal features for multimodal emotion recognition." In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1215-1223. IEEE, 2017.

[36] Wang, Fei, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. "Residual attention network for image classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156-3164. 2017.

[37] Larochelle, Hugo, and Geoffrey E. Hinton. "Learning to combine foveal glimpses with a third-order Boltzmann machine." In *Advances in neural information processing systems*, pp. 1243-1251. 2010.

[38] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9, no. 8 (1997): 1735-1780.

[39] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." In *Proceedings of the IEEE international conference on computer vision*, pp. 1520-1528. 2015.

[40] Srivastava, Rupesh K., Klaus Greff, and Jürgen Schmidhuber. "Training very deep networks." In *Advances in neural information processing systems*, pp. 2377-2385. 2015.

[41] Kim, Jin-Hwa, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. "Multimodal residual learning for visual qa." In *Advances in neural information processing systems*, pp. 361-369. 2016.

[42] Hasani, Behzad, and Mohammad H. Mahoor. "Facial affect estimation in the wild using deep residual and convolutional networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 9-16. 2017.

[43] Huang, Christina. "Combining convolutional neural networks for emotion recognition." In *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1-4. IEEE, 2017.

[44] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

[45] Krizhevsky, Alex, and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Vol. 1, no. 4. Technical report, University of Toronto, 2009.

[46] Gudi, Amogh. "Recognizing semantic features in faces using deep learning." *arXiv preprint arXiv:1512.00743* (2015).