Proceedings Paper:
Jalal, M.A., Mihaylova, L. orcid.org/0000-0001-5856-2223 and Moore, R.K. (2020) An end-to-end deep neural network for facial emotion classification. In: 2019 22th International Conference on Information Fusion (FUSION). 22nd International Conference on Information Fusion, 02-05 Jul 2019, Ottawa, Canada. IEEE . ISBN 9781728118406

An End-to-End Deep Neural Network for Facial Emotion Classification

Md Asif Jalal[a], Lyudmila Mihaylova[b], Roger K Moore[a]

[a]Department of Computer Science, University of Sheffield, S1 4DP, UK
[b]Department of Automatic Control and Systems Engineering, University of Sheffield, S1 3JD, UK
{majalal1,r.k.moore,l.s.mihaylova}@sheffield.ac.uk

***Abstract*—Facial emotional expression is a nonverbal communication medium in human-human communication. Facial expression recognition (FER) is a significantly challenging task in computer vision. With the advent of deep neural networks, facial expression recognition has transitioned from lab-controlled settings to more neutral environments. However, deep neural networks (DNNs) suffer from overfitting the data and biases towards specific categorical distribution. The number of samples in each category is heavily imbalanced, and overall the number of samples is much less than the full number of samples representing all emotions. In this paper, we propose an end-to-end convolutional-self attention framework for classifying facial emotions. The convolutional neural network (CNN) layers can capture the spatial features in a given frame. Here we apply a convolutional-self-attention mechanism to obtain the spatio-temporal features and perform context modelling. The AffectNet database is used to validate the framework. The AffectNet database has a large number of image samples in the wild settings, which makes this database very challenging. The result shows a 30% improvement in accuracy from the CNN baseline.**

***Index Terms*—facial emotion, classification, attention networks, convolutional neural networks, deep neural architectures**

I. Introduction

Facial expression is a non-verbal signal for conveying emotions. Emotion carries para-linguistic cues and information about individual intention. The understanding of human emotions is an essential requirement for human-human communications and human-robot communications because it makes the communication more natural and feedback driven. The robot-robot and human-robot interactions are common and we use these robots in different aspects of our social life. The anthropomorphic and the zoomorphic robots are gaining popularity in the education sector, the health care industry and the entertainment industry. Thus, the importance of understanding emotions is rapidly increasing. Every research paper on this topic has different perspectives on the facial expression recognition (FER) problem. This paper presents an approach that can recognise the emotional state from various facial expressions.

FER systems can be either static or dynamic [1]. The static-based models [2], [3] encode the spatial features from a single image and the dynamic-based models encode the spatio-temporal features over a time span [4], [5]. The FER systems were built with hand-crafted feature descriptors (e.g histogram oriented gradient (HOG) [6], [7], the local binary pattern (LBP) approach, the local ternary pattern (LTP) approach [8] and the Gabor Filter [9]). These feature descriptors have their advantages and disadvantages [10]. These features are used with generative or discriminative classifiers (eg., Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) [10], [11]).

However, with the advent of deep neural networks, FER systems changed profoundly. Hybrid models with handcrafted features and neural networks became popular [12], [13]. Furthermore, the increases in cheap processing power of GPU units promotes the use of deep neural networks as end-to-end systems. Convolutional Neural Network (CNN) architectures have shown excellent results on image related problems learning spatial features and having rich representation power [14]. LeNet, VggNet and GoogleNet explore the architectural representations of CNNs emphasising on the depth, width and orientation of CNN blocks [15]–[17]. These CNN structures are used to build SER systems as well [2], [18], [19]. CNNs are good at learning deep representations with increasing size, but they do not perform that way for learning long-term positional and contextual dependencies. Attention architectures can learn the long-term positional and contextual dependencies [20]–[24].

In this paper, we propose a deep neural network structure to understand the positional dependencies between the facial regions across the channels in CNN. The goal of this network is to learn those dependencies to produce the categorical distribution of emotions from static facial expressions.

We have used an attention mechanism based on convolution blocks. The paper is organised as follows, section II presents the related works, section III presents the new approach towards the architecture and relevant models, section IV describes the proposed architecture, section V discusses the dataset and the experimental settings and the results, section VI elaborates the implications of the results and section VII draws the conclusion and propose possible future enhancements.

II. Related Works

Deep learning has been incorporated in FER systems for a long time [25], [26]. The high level of abstraction and non-linear representations have resulted in state-of-the-art results

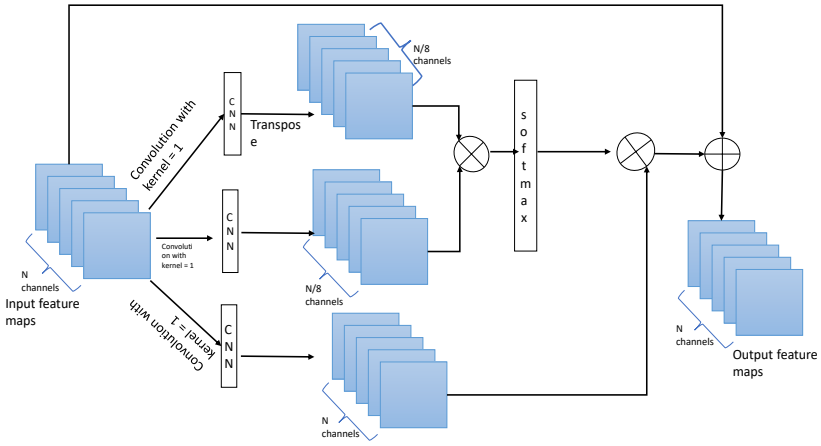

Fig. 1. Attention mechanism with convolution.

in FER systems. The deep networks can be trained for flexible tasks [27].

Initially, the facial muscle expressions recognition was biased towards its subject. The generalisation for correlation between the facial parts is crucial for subject independent emotion classification. The problem with subject independent scenarios in FER can be handled efficiently with deep neural networks. Traditional CNN such as AlexNet has performed very well in these problems [2], [28]. In [29], CNNs are trained across various databases and showed the usefulness of CNNs in FER problems in the cross-corpus situation. The network correlates among cross-corpus data very well. Jung et. al [30] extract temporal image features and temporal geometric features from a series of images using two different CNNs by fine-tuning and fusing them to improve the FER performance. Although this fusion approach improves the performance, CNNs are not very good at learning temporal dependencies in general. Some research has proposed temporal context modelling for learning temporal dependencies by using recurrent neural networks (RNNs) [31], [32]. A dual channel network is used for learning temporal features [33]. Fan et. al used VggNet [16] and RNN in one channel and 3D CNN on the other channel. These hybrid models often result in high performance for FER tasks. These architectures combined the spatial local feature extraction capability of CNNs with RNNs temporal modelling [34], [35]. They propose an average-based aggregation strategy for the features contrary to [19] who employed feature level fusion.

Vaswani et. al [22] propose an attention mechanism based architecture with feedforward neural networks where they show that the sole attention mechanism can learn the global dependencies between the input and output. Attention networks become vastly popular for modelling long term dependencies [20]–[24]. Wang et. al [36] propose a method to measure positional dependency within the same sequence. Parikh et. al [37] also have a similar approach, i.e. to decompose a sequence into a subsequence and compare it with the other subsequences in the same sequence.

III. The Developed Approach

Classifying emotions from facial expressions is based on the relative differences in facial muscles at a given instance. Traditionally, FER systems consider facial muscles (action units), and visually salient facial landmarks for classifying emotion. However, here we consider the problem slightly differently. Rather than looking for particular regions in the face we consider looking for the relational dependencies between each position with the others within the same image by adapting self-attention [36]–[38]. The proposed approach is built based on a stack of CNN with different kernel sizes and this affords the network to learn the different spatial features.

A. Convolutional Neural Network

In the convolution stage, feature maps are convoluted with kernels $g[x_b, y_b]$. The kernels are the local receptive field, and they share the same set of weights while doing convolution on different parts of the input image $f[x_a, y_a]$. While the kernels slide through the image, they extract visual features (edges, corners or more abstract features) and combine the set of outputs to form feature maps. If the kernels of size $[h \times w \times N]$ ([height $\times$ width $\times$ depth], and $n = 1, 2, \cdots, N$) are used, the n^{th} convolutional feature map will be:

$$y_n = f\left(\sum_j g_n * x_j\right), \tag{1}$$

where g_n is the n^{th} kernel and x_j $(j = 1, 2, \cdots, J)$ is the j^{th} input feature map of size $[A \times B]$ and $f(\cdot)$ is a nonlinear activation function. The kernel size is significant for preserving locality in the whole network as well as controlling representations [39].

We opt to use smal kernel size to preserve locality by considering a small neighbourhood at a time. Relatively small kernel sizes can increase non linearity in the network and enable feature fusion [40], [41]. Stacking more than one CNN layer results in increased non-linearity and rich representations [14], [16]. Sometimes, the CNN layer is followed by a pooling layer that focuses on improving the discriminability power of the network and robustness to shift and distortions [42]. However, it is crucial to control the kernel size in pooling to keep it from losing information. The network learns faster if the network parameters are decorrelated and they are linearly transformed to have zero means and unit variances [42]. Loffee et. al [43] proposed batch normalisation for approaching this issue by reducing the internal covariance shift in a batch of samples.

B. Self-Attention Networks

Self-attention networks have the flexibility of modelling long-term inter-sequence dependencies. We use self-attention as non-local networks [36], [38] to model the relationships between the regions in the feature maps from previous layers. The features from previous CNN layers **y** are transformed into three feature spaces **j**, **k** and ***l*** . Where

$$j\left(y\right) = W_j y \qquad k\left(y\right) = W_k y \qquad l\left(y\right) = W_l y \tag{2}$$

where W_j, W_k and W_l are network weights learned through back-propagation. Number of channels in W_j, W_k is less than number of channels in the features. However, W_k has the same number of channels as input feature y. Dot product is used to calculate the relationship between j and k. Then it is normalised using the $softmax$ function.

$$e_{ij} = softmax\left(j\left(y_i\right)^T k\left(y_j\right)\right) \tag{3}$$

The attention map is calculated by a matrix multiplication between e and $l(y)$. A scaling factor is multiplied with the attention map and the result is added with the input feature map.

$$attention\_output = \gamma \cdot \left(e \cdot l(y)\right) + y \tag{4}$$

In this work, γ is randomly initialized. This layer learns the non-local dependencies as well as the local neighbourhood.

IV. The Proposed Framework

The proposed framework is shown in Figure 2. The first CNN block (c1) has one convolutional layer of output channel size 32 and kernel size 3. This is followed by batch normalization and rectified linear unit (ReLU) nonlinearity. The second CNN block (c2) has two convolution layers of kernel size 3 and 5. Maxpooling is used to introduce sparsity in the network. Two maxpooling layers are used with kernel size 2. Throughout the network the same kernel size (2) has been used for maxpooling. The input feature channels and output feature channels in c2 are 32 and 192, respectively. The third CNN block (c3) has three convolutional layers. The first convolutional layer is followed by a maxpooling layer and batch normalisation. The remaining two convolutional layers are stacked together. The input feature channels and output feature channels in c3 are 192 and 128 respectively. Both upsampling and downsampling is performed in c3. Sample outputs feature maps have been visualized in figure 3.

Three CNN blocks are followed by a self-attention layer (a1). The network weights W_j, W_k, W_l, mentioned in Section III-B, are three convolutional layers with kernel size 1. The number of output channels for W_j and W_k is one-eighth of the input channels in the self-attention layer. The number of output channels is decreased to reduce the computation time. The number of output channels for W_l is the same as the input channels in the self-attention layer. The scaling factor γ is randomly initialised. The kernel size in the attention layer is set to 1 to perform feature level fusion.

The output feature maps from a1 are fed into the fourth CNN block (c4). This block has a convolution layer with input channel size 128 and output channel size 64 with kernel size 2. This convolution is followed by an adaptive average pooling layer to produce a fixed length of 3×3 sized feature maps. Throughout the network, after each pair of convolution and maxpooling, batch normalization is performed to reduce the correlation between the parameters in the network.

The final block (l1) is a dense layer. It (d1) has two parts. The first part contains one fully-connected layer followed by a dropout (0.85) layer. The rectified linear unit (ReLU) activation function has been used with this. The second part has one fully connected layer followed by a softmax layer. We have used dropout (0.85) for regularization.

V. Performance Evaluation

A. Dataset

AffectNet [44] is an image database for detecting and recognising facial expressions of emotions in the wild. It consists of both the categorical (joy, anger, sad, happy etc.) and the dimensional labelling (valence, arousal) of emotion in the wild. Having more than 140000 image samples without any controlled environment (e.g., lighting condition, alignment, rotation, head posture etc), makes this database very difficult. The facial landmarks (OpenCV [45]) of the images are also provided with the dataset. The average image resolution is 425×425 with STD of 349×349 pixels. Professional annotators have labelled the images.

There are eight emotion categories in AffectNet database. The emotions are neutral (80276), happy (146198), sad (29487), surprise (16288), fear (8191), disgust (5264), anger (28130), contempt (5135). The numbers with those emotion labels are the numbers of annotated samples for each category.

B. Data Preparation

The number of samples in each category clearly shows that the classes are heavily imbalanced . The standard training split of the database is also imbalanced. To cope with this problem, we adopted up-sampling and down-sampling strategy. For training, we randomly chose 13000 images for each category from the official training split. The under-represented classes are up-sampled by replicating and augmentation. So, the

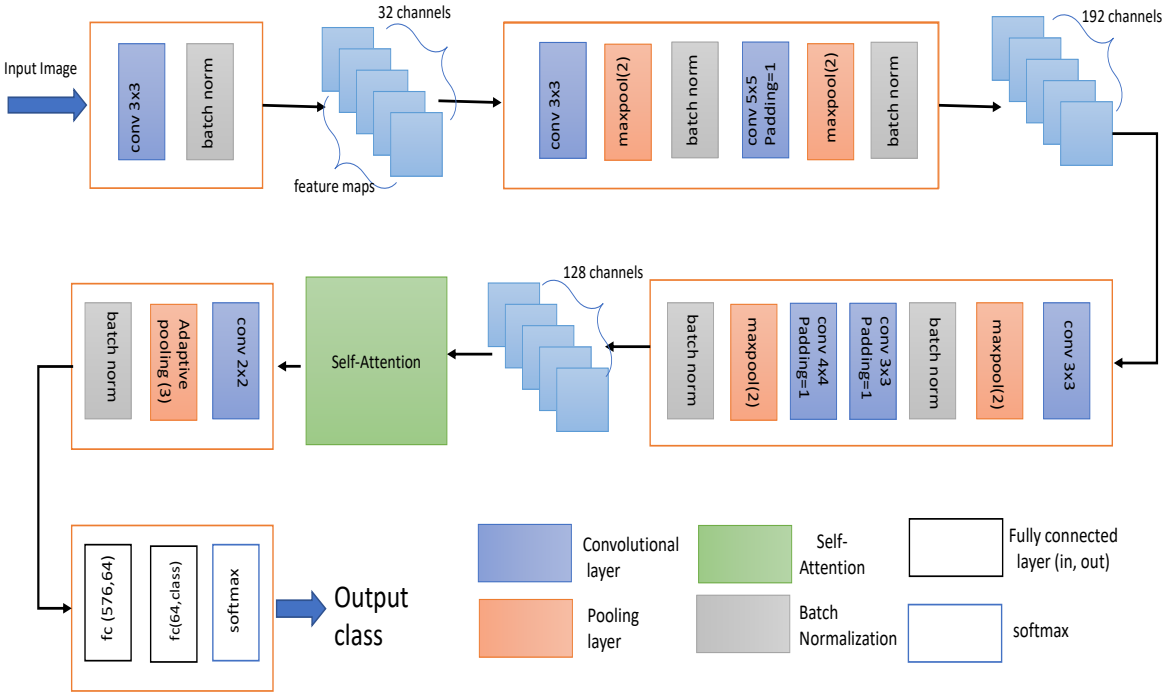

Fig. 2. The proposed self-attention network architecture for emotion classification

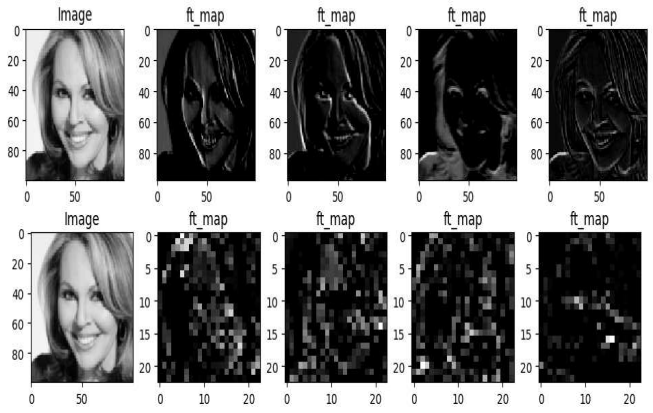

Fig. 3. Original Image and sample feature maps after c1 (top row) and after c2 (bottom row)

number of samples for each class is neutral (13000), happy (13000), sad (13000), surprise (13000), fear (12756), disgust (11409), anger (13000), contempt (11250). However, it can be seen that the data is imbalanced in contrast to the approach taken by [44] where the authors of [44] took 15000 samples and up-sampled heavily for making the data balanced thus making our experimental scenario more challenging.

The official standard validation split is used for testing the performance of the proposed framework as the validation set is published for testing purposes. Each of the classes has 500 sample images for testing.

C. Experiment

The experiments have been conducted using the PyTorch [46] deep learning framework. A Nvidia GTX 1080ti GPU has been used for executing the experiments.

1) Learning: The adam optimiser [47] has been applied to mini-batch of 500 images with categorical cross-entropy loss. The momentum and weight decay are set to 0.9 and 0 respectively. Throughout the network the learning rate is set to 0.0001. To prevent the network from overfitting dropout layers have been used with the fully connected layers. In the proposed model, 2, the dropout rate is set to 0.85 in the classifier section.

2) Data Augmentation: Random horizontal flipping and random cropping have been applied on the frames for data augmentation in order to increase the diversity of the training samples. The frames are normalized and re-sized to $[224 \times 224]$ images.

TABLE I
TEST ACCURACY ON AFFECTNET VALIDATION SET

Method	Accuracy (%)
Baseline (AlexNet)	58.0
Proposed Network	**93.8**

D. Results

The proposed framework has been evaluated with 'AffectNet' [44] database. The creators of 'AffectNet' database haven't released any official test set yet. After contacting the authors [44], they advised us to use the official validation set as the test set. According to Mollahosseini et al. [44], a baseline system has been implemented on AlexNet [14]. The results on validation set are reported officially in [48]. Also, we follow the architectural reference mentioned in the paper [44]. AlexNet is built upon sixty two million parameters, making it a very deep neural network architecture. Our proposed framework has fewer parameters compared to the very deep neural network architectures. However, it can be clearly seen in Table I that the performance gain with the proposed framework is more than 30%. We showed that the performance of the proposed framework is not biased upon any particular emotion category. In Figure 4 the confidence scores are given. We ran the experiment five times and took the mean of the top accuracy scores. The accuracy on AffectNet validation set is reported as 93.8%.

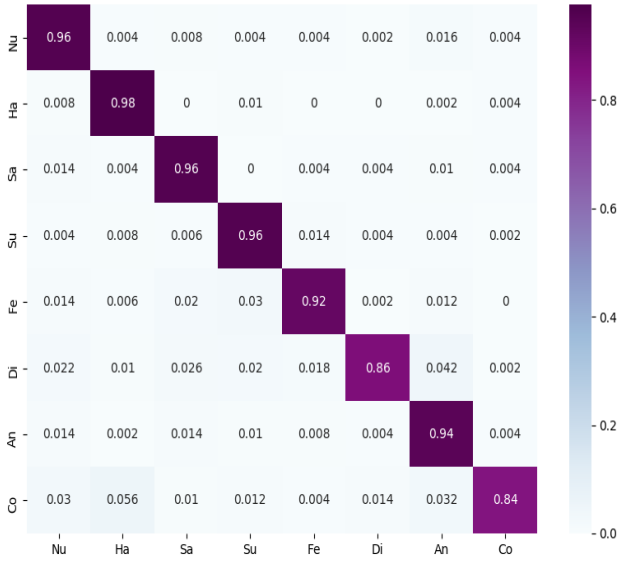

Fig. 4. Confusion Matrix for the eight emotion category labels Nu (neutral), Ha (happy), Sa (sad), Su (surprise), Fe (fear), Di (disgust), An (anger), Co (contempt)

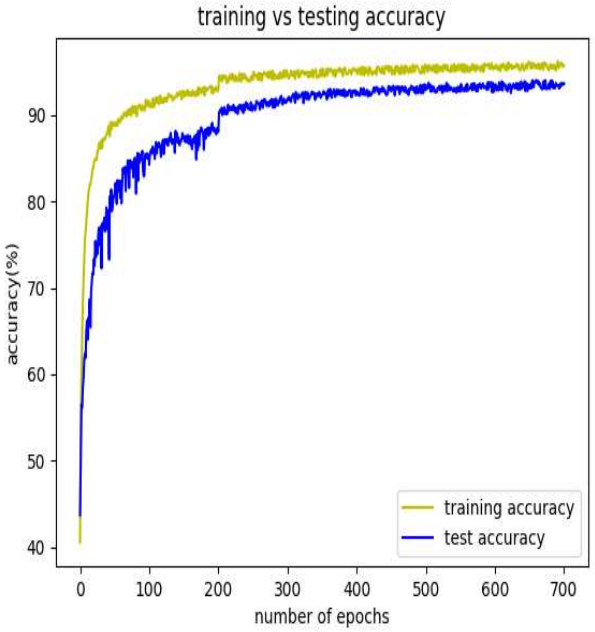

Fig. 5. Training and Test accuracy (%) in 700 epochs

VI. DISCUSSION

The network in the proposed framework consists of 1.69 million parameters, which is considerably less than the parameters in AlexNet and also most of the state-of-the-art deep neural networks. Hence, we managed to reduce the computational cost without compromising with the performance of the framework. Figure 5 clearly shows that the network learns generalized deep representations on training data. Since the initial training phase, the network successfully managed to avoid overfitting.

Figure 4 shows the confidence values among the classes. The images with "happy" labels show the highest confidence, followed by surprise, sad, neutral and anger. There has been conflict among the disgust and contempt emotion labelled images. However, human annotators agree upon only 60.7% of the image labels [44].

We extract the non-local feature dependencies in the image segments and prioritise image segments. Also, small neighbourhoods were considered with smaller convolutional kernels while creating the feature maps. Furthermore, rapid downsampling and upsampling has been avoided throughout the network. These approaches have led to richer representations, that justifies the improved performance gain of the proposed framework.

VII. CONCLUSION

We presented a convolutional self-attention framework in this paper. The self-attention mechanism is for learning the positional dependencies among the facial regions for a given categorical distribution. Also, we achieved 30% gain in accuracy while maintaining a reasonable computational cost. This

research brought the state-of-the-art to a different level with a simpler network as explained in section VI. Finally, this work brought together good practices for designing CNN and deep networks.

REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018. [Online]. Available: http://arxiv.org/abs/1804.08348

[2] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *CoRR*, vol. abs/1511.04110, 2015. [Online]. Available: http://arxiv.org/abs/1511.04110

[3] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1805–1812. [Online]. Available: https://doi.org/10.1109/CVPR.2014.233

[4] X. Zhao, X. Liang, L. Liu, T. Li, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," *CoRR*, vol. abs/1607.06997, 2016. [Online]. Available: http://arxiv.org/abs/1607.06997

[5] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007. [Online]. Available: https://doi.org/10.1109/TPAMI.2007.1110

[6] C. Orrite, A. Gañán, and G. Rogez, "Hog-based decision tree for facial expression classification," in *Pattern Recognition and Image Analysis*, H. Araujo, A. M. Mendonça, A. J. Pinho, and M. I. Torres, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 176–183.

[7] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based hog features," in *Face and Gesture 2011*, March 2011, pp. 884–888.

[8] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, Sep. 2008, pp. 1–8.

[9] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, ser. FG '98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 454–. [Online]. Available: http://dl.acm.org/citation.cfm?id=520809.796139

[10] P. Carcagnì, M. D. Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," in *SpringerPlus*, 2015.

[11] M. U. Tariq, J. Yang, and T. S. Huang, "Maximum margin gmm learning for facial expression recognition," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, 2013.

[12] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid cnn–sift aggregator," in *Multi-disciplinary Trends in Artificial Intelligence*, S. Phon-Amnuaisuk, S.-P. Ang, and S.-Y. Lee, Eds. Cham: Springer International Publishing, 2017, pp. 139–149.

[13] M. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *CoRR*, vol. abs/1804.10892, 2018. [Online]. Available: http://arxiv.org/abs/1804.10892

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[18] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 435–442. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830595

[19] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518301302

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057. [Online]. Available: http://proceedings.mlr.press/v37/xuc15.html

[21] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[23] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *CoRR*, vol. abs/1412.7755, 2015.

[24] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *CoRR*, vol. abs/1807.06521, 2018. [Online]. Available: http://arxiv.org/abs/1807.06521

[25] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Computer Vision – ACCV 2014*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, pp. 143–157.

[26] E. P. Ijjina and C. K. Mohan, "Facial expression recognition using kinect depth sensor and convolutional neural networks," in *2014 13th International Conference on Machine Learning and Applications*, Dec 2014, pp. 392–396.

[27] H. Khalajzadeh, M. Mansouri, and M. Teshnehlab, "Face recognition using convolutional neural network and simple logistic classifier," in *Soft Computing in Industrial Applications*, V. Snášel, P. Krömer, M. Köppen, and G. Schaefer, Eds. Cham: Springer International Publishing, 2014, pp. 197–207.

[28] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.

[29] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *CoRR*, vol. abs/1705.01842, 2017. [Online]. Available: http://arxiv.org/abs/1705.01842

[30] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2983–2991, 2015.

[31] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 49–56. [Online]. Available: http://doi.acm.org/10.1145/2808196.2811638

[32] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 467–474. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830596

[33] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI

'16. New York, NY, USA: ACM, 2016, pp. 445–450. [Online]. Available: http://doi.acm.org/10.1145/2993148.2997632

[34] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550. [Online]. Available: http://doi.acm.org/10.1145/2522848.2531745

[35] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 494–501. [Online]. Available: http://doi.acm.org/10.1145/2663204.2666274

[36] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CoRR*, vol. abs/1711.07971, 2017. [Online]. Available: http://arxiv.org/abs/1711.07971

[37] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," *CoRR*, vol. abs/1606.01933, 2016. [Online]. Available: http://arxiv.org/abs/1606.01933

[38] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-Attention Generative Adversarial Networks," *arXiv e-prints*, p. arXiv:1805.08318, May 2018.

[39] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[40] M. Lin, Q. Chen, and S. Yan, "Network in network," *CoRR*, vol. abs/1312.4400, 2013. [Online]. Available: http://arxiv.org/abs/1312.4400

[41] S. H. HasanPour, M. Rouhani, M. Fayyaz, M. Sabokrou, and E. Adeli, "Towards principled design of deep convolutional networks: Introducing simpnet," *CoRR*, vol. abs/1802.06205, 2018. [Online]. Available: http://arxiv.org/abs/1802.06205

[42] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*. London, UK, UK: Springer-Verlag, 1999, pp. 319–. [Online]. Available: http://dl.acm.org/citation.cfm?id=646469.691875

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[44] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," 2017.

[45] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[48] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet," 2017. [Online]. Available: http://mohammadmahoor.com/affectnet/