DarkSiL Detector for Facial Emotion Recognition

Tarim Dar and Ali Javed

Abstract Facial Emotion recognition (FER) is a significant research domain in computer vision. FER is considered a challenging task due to emotion-related differences such as heterogeneity of human faces, differences in images due to lighting conditions, angled faces, head poses, different background settings, etc. Moreover, there is also a need for a generalized and efficient model for emotion identification. So, this paper presents a novel, efficient, and generalized DarkSiL (DS) detector for FER that is robust to variation in illumination conditions, face orientation, gender, different ethnicities, and varied background settings. We have introduced a low-cost, smooth, bounded below, and unbounded above Sigmoid-weighted linear unit function in our model to improve efficiency as well as accuracy. The performance of the proposed model is evaluated on four diverse datasets including CK + , FER-2013, JAFFE, and KDEF datasets and achieved an accuracy of 99.6%, 64.9%, 92.9%, and 91%, respectively. We also performed a cross-dataset evaluation to show the generalizability of our DS detector. Experimental results prove the effectiveness of the proposed framework for the reliable identification of seven different classes of emotions.

Keywords DarkSIL (DS) emotion detector · Deep learning · Facial emotion recognition · SiLU activation

1 Introduction

Automatic facial emotion recognition is an important research area in the field of artificial intelligence (AI) and human psychological emotion analysis. Facial emotion recognition (FER) is described as the technology of analysing the facial expression of a person from images and videos to get information about the emotional state of that individual. FER is a challenging research domain because everyone expresses their

T. Dar · A. Javed (✉)
University of Engineering and Technology-Taxila, Department of Software Engineering, Taxila 47050, Pakistan
e-mail: ali.javed@uettaxila.edu.pk

S. Anwar et al. (eds.), *Proceedings of International Conference on Information Technology and Applications*, Lecture Notes in Networks and Systems 614,
https://doi.org/10.1007/978-981-19-9331-2_7

emotions differently. Furthermore, several challenges and obstacles exist in this area which makes emotion analysis quite difficult. Nowadays, researchers are focusing to improve the interaction between humans and computers. One way of doing that is to make computers intelligent so they can understand the emotions of humans and interact with them in a better way. Automatic FER systems have the ability to improve our life quality. FER systems can help in the rehabilitation of patients with facial paralysis diseases, they aid in getting customers' feedback on products [1], and robotic teachers having an understanding of students' feelings can offer an improved learning experience. In short, FER systems have extensive applications in various domains, i.e., medical, deep fakes detection, e-learning, identification of emotions of drivers while driving, entertainment, cyber security, image processing, virtual reality applications [2], face authentication systems, etc.

Early research in the field of facial emotion identification is focused on appearance and geometric-based feature extraction methods. For example, the local binary pattern (LBP)-based model presented in [3] introduced the concept of the adaptive window for feature extraction. The approach [3] was validated on Cohn-Kanade (CK) and Japanese Female Facial Expression (JAFFE) datasets against six and seven emotions. Also, Niu et al. [4] proposed a fused feature extraction method from LBP and oriented FAST and Rotated BRIEF (ORB) descriptors. After that, the support vector machine (SVM) classifier was used to identify the emotions. This method [4] was evaluated on three datasets, i.e., CK +, MMI, and JAFFE. The LBP approaches have the limitations of producing long histograms which slows down model performance on large datasets.

Many Convolutional Neural Networks (CNNs)-based methods are developed in the past few decades that have achieved good classification results for FER. For instance, Liu et al. [5] developed CNN-based approach by concatenation of three different subnets. Each subnet was a CNN model which was trained separately. A fully connected layer was used to concatenate extracted features from these subnets and after that softmax layer was used to classify the emotion. The approach [5] was only validated on one dataset which is the Facial Expression Recognition (FER-2013) dataset and obtained an overall accuracy of 65.03%. Similarly, Ramdhani et al. [1] presented a facial emotion recognition system based on CNN. The purpose of this approach [1] was to gather customer satisfaction with the product. This approach was tested on the custom and the FER-2013 datasets. This method [1] has limited evaluation against four emotions on these datasets. Moreover, Jain et al. [6] proposed a deep network (DNN) consisting of convolution layers and deep residual modules for emotion identification and tested the method on JAFFE and Extended Cohn-Kanade (CK +) datasets.

However, there still exist many limitations of these methods such as existing models are not generalized or outperform certain conditions, i.e., variation in face angles, people belonging to different ethnic groups, high computational complexity, variations in lighting conditions and background setting, gender, skin diseases, heterogeneity in faces, and difference in expression of emotion which vary from person to person. In this paper, we presented a robust and effective deep learning model that can automatically detect and classify seven types of facial emotions

(happy, surprise, disgust, fear, sad, anger, and neutral) from frontal and face-oriented static images more accurately. In the proposed work, we customize the basic block of Darknet-53 architecture and introduce the Sigmoid-weighted Linear Unit (SiLU) activation function (a special form of swish function) for the classification of facial emotions. SiLU is a simple multiplication function of input value with a sigmoid function. This activation function allows a narrow range of negative values which facilitates it to recognize the patterns in data easily. As a result of this activation function, a smooth curve is obtained, which aids in optimizing the model in terms of convergence with minimum loss. Furthermore, using SiLU activation in the Darknet-53 architecture optimizes the model performance and makes it computationally efficient. The main contributions of this research work are as follows:

- We propose an effective and efficient DarkSil (DS) emotion detector with SiLU activation function to automatically detect seven diverse facial emotions.
- The proposed model is robust to variations in gender and race, lighting conditions, background settings, and orientation of the face at five different angles.
- We also performed extensive experimentation on four diverse datasets containing images of spontaneous as well as non-spontaneous facial emotions and performed a cross-corpora evaluation to show the generalizability of the proposed model.

2 Proposed Methodology

CNN is a network that contains a different number of layers which assists feature extraction from images better than other feature extraction methods [7]. Deep convolutional neural networks are being developed to improve image recognition accuracy. In this study, we present a customized Darknet-53 model which is the improved and deeper version of Darknet-19 architecture. The input size requirement of Darknet-53 is $256 \times 256 \times 3$. The overall architecture of our customized proposed model is shown in Fig. 1.

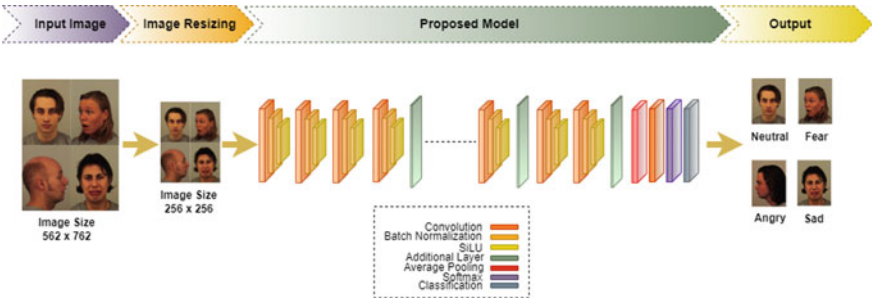

Fig. 1 Architecture of the proposed method

2.1 Datasets for Emotion Detection

To evaluate the performance of our model, we have selected four diverse datasets, i.e., Extended Cohn-Kanade (CK +) [8], Japanese Female Facial Expression (JAFFE) [9], Karolinska Directed Emotional Faces (KDEF) [11], and Facial Expression Recognition 2013 (FER-2013) [10]. JAFFE [9] consists of 213 posed images of ten Japanese models with 256 × 256 resolution. All of the facial images were taken under strictly controlled conditions with similar lighting and no occlusions like hair or glasses. The CK + [8] database is generally considered to be the most frequently used laboratory-controlled face expression classification dataset. Both non-spontaneous (posed) and spontaneous (non-posed) expressions of individuals belonging to different ethnicities (Asians or Latinos, African Americans, etc.) were captured under various lighting conditions in this dataset. The resolution of images in the CK + dataset is 640 × 490. KDEF [11] is a publicly accessible dataset of 4900 images of resolution 562 × 762 taken from five different angles: straight, half left, full left, half right, and full right. This dataset is difficult to analyze because one eye and one ear of the face are visible in full right and full left profile views, making the FER more challenging. FER-2013 [10] contains 35,685 real-world grayscale images of 48 × 48 resolution. As this dataset contains occultation, images with text, non-face, very low contrast, and half-faced images, so, the FER-2013 dataset is more diversified and complex than other existing datasets. A few sample images of all four datasets are presented in Fig. 2.

Fig. 2 Sample images of datasets

2.2 Data Processing

In the pre-processing step, images of each dataset are resized to our model requirement of 256 × 256 resolution with three channels. After pre-processing, images are sent to our customized proposed model to extract the reliable features and later classify the emotions of seven different categories as shown in Fig. 1.

2.3 DarkSiL Architecture

The smallest component of our customized DarkSiL architecture is composed of the convolutional layer, the Batch Normalization (BN) layer, and the SiLU activation layer which are described as follows:

(1) Convolutional layers are the main components of convolutional neural networks. CNN uses a filter or kernel of varied sizes on input to generate a feature map that summarizes the presence of detected features. Darknet-53 architecture contains 53 convolution layers.
(2) Batch Normalization Layer—The use of BN is to normalize the output to the same distribution based on the eigenvalues of the same batch. It can accelerate network convergence and prevent over-fitting after the convolutional layer.
(3) SiLU activation layer—SiLU is a special case of the Swish activation function which occurs at $\beta = 1$. Unlike the ReLU (and other commonly used activation units such as sigmoid and tanh units), the SiLU's activation does not increase monotonically. The property of non-monotonicity improves gradient flow and provides robustness to varying learning rates. One excellent property of the SiLU is its ability to self-stabilize [19]. Moreover, SiLU is a smooth, unbounded above and below activation function. Unboundedness aids in avoiding saturation, and the bounded below property produces strong regularization effects. Furthermore, smoothing helps in obtaining a generalized and optimal model. SiLU activation can be computed as

$$\mathrm{f}(x) = x \times \mathrm{sigmoid}(\beta x) \tag{1}$$

where x is the input value and $\beta = 1$. The smallest component of the Darknet model is repeated 53 times which means its architecture contains 53 convolutions and 53 batch normalization layers. So, 53 SiLU layers are introduced in our customized architecture. We also used the transfer learning approach to train our model on seven output classes of emotions. Feature extraction layers are initialized by using pre-trained Darknet-53 architecture whereas the last three layers after global average pooling, i.e., fc8 (convolution layer with output size 1000), softmax layer, and classification layer are replaced to improve the model.

In the Darknet-53 model, the global average pooling (GAP) layer is presented instead of a fully connected layer. The GAP layer computes the average of all feature maps and feeds the obtained vector into the next convolution layer. The GAP layer has numerous advantages over the convolution layer. One of them is that it imposes a connection between extracted features and categorizations which helps in better interpretation of feature maps as the confidence maps for classes. Second, over-fitting can be prevented in this layer as there is no parameter optimization required in the GAP layer. Moreover, the GAP adds up the spatial information and makes it more robust to spatial translation.

In the softmax layer, numbers in the input vector are converted into values in the range of 0 and 1 which are further perceived as probabilities by the model. The mathematical softmax function in this layer is a generalized case of logistic regression and is applied for the classification of multiple classes.

A classification layer calculates the cross-entropy loss for classification purposes with exclusive categories. The output size in the preceding layer determines the number of categories. In our case, the output size is seven different classes of emotions and the input image is classified into one of these categories.

3 Experimental Setup and Results

For all experiments, the dataset is split into training (60%), validation (20%), and testing (20%) sets. The parameters used for model training on each experiment are Epoch:20, Shuffle: Every epoch, Learning rate: 4×10^{-4}, Batch size: 32, Validation frequency: Every epoch, and Optimizer: Adam. All experiments are carried out on MATLAB 2021a on the machine with the following specifications: AMD Ryzen 9 5900 × 12-core 3.70 GHz processor, 32 GB RAM, 4.5 TB hard disk, and Windows 10 Pro. We employed the standard metrics of accuracy, precision, and recall for the evaluation of our model as these metrics are also used by the contemporary FER methods.

3.1 Performance Evaluation of the Proposed Method

We designed four-stage experiments to show the effectiveness of the proposed model on KDEF [11], JAFFE [9], FER-2013 [10], and CK + [8] datasets.

In the first stage, we performed an experiment on the JAFFE dataset to investigate the performance of the proposed model on a small posed dataset. After training and validation, the proposed model is tested on the test set and the results are mentioned in Table 1. It is worth noticing that our model has achieved an accuracy of 92.9% on the JAFFE dataset, a mean precision of 93.5%, and a mean recall of 92.8%. Results above 90% on the biased JAFFE dataset with mislabeled class problems show the effectiveness of the proposed model for FER.

Table 1 Results of the proposed model on different datasets

Dataset	Accuracy (%)	Mean precision (%)	Mean recall (%)
JAFFE	92.9	93.5	92.8
CK +	99.6	99.1	99.2
KDEF	91.0	93.4	93.0
FER-2013	64.9	65.3	61.1

In the second stage, we conducted an experiment to show the efficacy of the proposed model on a dataset having individuals who belong to different regions, races, and genders. For this purpose, we choose a lab-controlled CK + dataset that contains spontaneous and non-spontaneous facial expressions of people with varying lighting conditions. Table 1 demonstrates the remarkable performance of the proposed model on the CK + dataset. Results of accuracy, precision, and recall close to 100% show that our model can accurately distinguish seven different types of facial expressions in frontal face images of people belonging to different geographical regions of the world.

In the third stage, to check the robustness of the proposed model on varied angular facial images, we designed an experiment on the KDEF dataset as it comprises facial images taken from five different viewpoints. Our proposed model obtained an overall accuracy of 91%, mean precision, and mean recall of 93.4% and 93%, respectively, as shown in Table 1. Obtained results demonstrate that the proposed model not only identifies emotions from frontal face images with higher accuracy but also performs well in the predictions of the facial emotions in images with faces tilted at some angle.

In the fourth stage, we implemented an experiment to examine the effectiveness of the proposed method on a real-world FER-2013 dataset that covers challenging scenarios of intra-class variations and class imbalance. This dataset is originally split into training, validation or public test, and private test sets. Furthermore, the FER-2013 dataset has non-face, low contrast, occlusion, different illumination conditions, variation in face pose, images with text, half-rotated, tilted, and varied ages and gender images which make the classification process more difficult. As reported in Table 1, our model achieved an accuracy of 64.9% which is good in presence of such variation on this challenging dataset. Moreover, the accuracy achieved on this dataset, i.e., 64.9% $\approx$ 65% is very close to the human-level accuracy of 65 $\pm$ 5% on this dataset [10].

3.2 Comparison with Contemporary Methods

To show the effectiveness of our model for facial emotion recognition on multiple diverse datasets, we compared the performance of our method against the existing

state-of-the-art (SOTA) FER methods. In the first stage, we compared the performance of our method with these contemporary methods [12–14] on the JAFFE dataset and the results are provided in Table 2. From Table 2, it is clearly observed that our model achieved an average gain of 12.2% over the existing SOTA. Our proposed model also has a higher discriminative ability than existing works. In the second stage, we compared the results of our method on the CK + dataset with existing methods [6], and [17]. The results in Table 2 depict that our model has a 9–10% better recognition rate in the classification of FER and performs well than comparative methods on the CK + dataset. In the third stage, we compared the performance of our method with state-of-the-art methods [12, 15], and [16] on the KDEF dataset. As shown in Table 2, the accuracy of our model is higher than all of the existing works [12, 15, 16] on the KDEF dataset. The second best-performing method [15] obtained an accuracy of 88% which is 3% lesser than our proposed model. The results state that the proposed method can detect images taken from five angles (0°, -45°, 45°, -90°, and 90°) more accurately than SOTA methods. In the last stage, we compared our model's performance with contemporary approaches of [1, 5], and [18] for the FER-2013 dataset, and results in terms of accuracy are provided in Table 2. It can be seen that the accuracy of the proposed model on the FER-2013 dataset is higher or very close to the best-performing model [5] with a slight difference of 0.13%. It means that our proposed model can detect facial emotions with more accuracy in challenging scenarios of the real world.

3.3 Cross-Corpora Evaluation

The previous works on FER gave less attention to the aspect of model generalizability for seven classes of emotions. So, to overcome this limitation, we conducted a cross-corpora evaluation in which four different datasets are used to demonstrate the generalizability of our model. Previous studies have used one or two datasets for training and performed testing on other datasets and also used a few types of emotions when performing cross-corpora experiments. In this study, we include a wide range of datasets from small posed and lab-controlled ones to real-world and spontaneous expression datasets and straight face to varied angled face image datasets in our cross-dataset experiments. The results of the cross-corpora evaluation are displayed in Table 3.

Despite the very good performance of the proposed model on the individual datasets, it could not perform as well on cross-dataset experiments. A possible reason for the degradation of the accuracy of these experiments is that there exist many dissimilarities among these datasets. These datasets are collected under distinct illumination conditions, with varying background settings in different environments. Types of equipment used in capturing images are different and images are taken from varying distances from the camera. Furthermore, subjects involved in the preparation of these datasets do not belong to the same geographical regions and are of

Table 2 Comparison of DS detector (proposed model) with SOTA

Model		Dataset	Accuracy (%)
Sun et al. [12]		JAFFE	61.68
Kola et al. [3]		JAFFE	88.3
LBP + ORB [4]		JAFFE	92.4
Proposed Model		**JAFFE**	**92.9**
DTAN [17]		CK +	91.44
DTGN [17]		CK +	92.35
DTAGN (Weighted Sum) [17]		CK +	96.94
DTAGN(Joint) [17]		CK +	97.25
Jain et al. [6]		CK +	93.24
Proposed Model		**CK +**	**99.6**
Williams et al. [16]		KDEF	76.5
Sun et al. [12]		KDEF	77.9
VGG-16 Face [15]		KDEF	88.0
Proposed Model		**KDEF**	**91.0**
Talegaonkar et al. [18]		FER-2013	60.12
Ramdhani et al. [1]	With batch size 8	FER-2013	58.20
	With batch size 128	FER-2013	62.33
Liu et al. [5]		FER-2013	65.03
Proposed Model		**FER-2013**	**64.9**

Table 3 Results of the cross-corpora evaluation

Training dataset	Testing dataset	Accuracy (%)
Fer-2013	JAFFE	**31.0**
	KDEF	25.8
	CK +	**67.0**
CK +	JAFFE	21.4
	KDEF	12.2
	FER-2013	28.7
KDEF	JAFFE	**35.7**
	CK +	**40.2**
JAFFE	KDEF	15.9
	FER-2013	14.3
	CK +	24.9

different gender, ages, and races. There also exists a dissimilarity among morphological characteristics of individuals involved in the making of these datasets. Moreover, people belonging to different ethnicities have differences in expressing their emotions. Eastern in contrast to Western shows low arousal emotions. Japanese (eastern) in contrast to European and American (western) tends to show fewer physiological emotions [13]. Datasets available in the domain of FER are also biased like KDEF is ethnicity biased (only European people) and JAFFE is a lab-controlled and highly biased dataset concerning gender (only females) and ethnicity (only Japanese models) and ambiguous expression annotations [14]. Images present in the original datasets are also different from each other in terms of resolution (FER-2013: 48 × 48, JAFFE: 256 × 256, etc.) and image type (grayscale and RGB). Although we upscale and downscale them into the same resolution according to our customized model requirement. But this reason may also affect the results of the cross-corpora evaluation. Despite all these reasons, it can be observed from the results in Table 3 that our proposed model, when trained on the FER-2013 dataset and tested on the JAFFE dataset, obtained an accuracy of 67% which is good in presence of such diversity. Also, the model trained on the KDEF dataset is able to achieve an accuracy of 40.2%. In Table 3, results above 30% are shown in bold.

4 Discussion

In this study, we conducted different experiments on four diverse datasets covering scenarios of straight and varied angled face images, people belonging to different cultures having different skin tones and gender (males, females, and children), variations in lighting conditions, different background settings, races, and a real-world challenging dataset. Our proposed model obtained accuracies greater than 90% except for the FER-2013 dataset. By closely observing the FER-2013 dataset, we found these possible reasons for the degradation of accuracy on this dataset. There exists a similarity in the face morphology of anger, surprise, and disgust classes of emotions in this dataset. Additionally, there exist more images of happy emotions as compared to other classes of emotions, which leads to insufficient learning of traits for these classes. Moreover, the FER-2013 dataset contains images with non-faces, occlusions, half-rotated and tilted faces, and variations in facial pose, age, and gender, which affect the recognition rate of the model. However, in presence of such challenges, our proposed model is still able to achieve human-level accuracy of approximately 65% for this dataset [10]. Table 1 shows the summarized performance of the proposed model on all these datasets. The outperforming results of our model on varied and diverse datasets including challenging scenarios show that our model is effective and robust in recognizing facial emotions. Moreover, the addition of SiLU activation in Darknet architecture not only increases the model's efficiency but also improves accuracy. We also performed cross-corpora experiments to show the generalizability of our approach. From the results, we can say that our model

has covered most of the limitations of existing methods and performed well than comparative approaches.

5 Conclusion

In this research, we have introduced a novel model for facial emotion recognition that is efficient, cost-effective, and robust to variations in gender, people belonging to different races, lighting conditions, background settings, and orientation of the face at five different angles. The presented model was tested on four different datasets and achieved remarkable performance on all of them. The proposed model not only effectively classified emotions from frontal face pictures but also outperformed existing methods on face images with five distinct orientations. We also performed a cross-corpora evaluation of the proposed model to demonstrate its generalizability. In the future study, we plan to create a custom FER dataset to test the performance of our method in real time and further improve the performance of cross-corpora evaluation.

Acknowledgements This work was supported by the Multimedia Signal Processing Research Lab at the University of Engineering and Technology, Taxila, Pakistan.

References

1. Ramdhani B, Djamal EC, Ilyas R (2018, August). Convolutional neural networks models for facial expression recognition. In 2018 International Symposium on Advanced Intelligent Informatics (SAIN). IEEE, pp 96–101
2. Mehta D, Siddiqui MFH, Javaid AY (2018) Facial emotion recognition: A survey and real-world user experiences in mixed reality. Sensors 18(2):416
3. Kola DGR, Samayamantula SK (2021) A novel approach for facial expression recognition using local binary pattern with adaptive window. Multimed Tools Appl 80(2):2243–2262
4. Niu B, Gao Z, Guo B (2021). Facial expression recognition with LBP and ORB features. Comput Intell Neurosci
5. Liu K, Zhang M, Pan Z (2016, September). Facial expression recognition with CNN ensemble. In 2016 International Conference on Cyberworlds (CW), IEEE. pp 163–166
6. Jain DK, Shamsolmoali P, Sehdev P (2019) Extended deep neural network for facial emotion recognition. Pattern Recogn Lett 120:69–74
7. Wang H, Zhang F, Wang L (2020, January) Fruit classification model based on improved Darknet53 convolutional neural network. In 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE. pp 881–884
8. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision And Pattern Recognition-Workshops, IEEE. pp 94–101
9. Lyons MJ, Kamachi M, Gyoba J (2020) Coding facial expressions with Gabor wavelets (IVC special issue). arXiv preprint arXiv:2009.05938
10. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Bengio Y (2013, November) Challenges in representation learning: A report on three machine learning contests.

In International Conference on Neural Information Processing, pp. 117–124. Springer, Berlin, Heidelberg
11. Lundqvist D, Flykt A, Öhman A (1998) Karolinska directed emotional faces. Cogn Emot
12. Sun Z, Hu ZP, Wang M, Zhao SH (2017) Individual-free representation-based classification for facial expression recognition. SIViP 11(4):597–604
13. Lim N (2016) Cultural differences in emotion: differences in emotional arousal level between the East and the West. Integr Med Res 5(2):105–109
14. Liew CF, Yairi T (2015) Facial expression recognition and analysis: a comparison study of feature descriptors. IPSJ transactions on computer vision and applications 7:104–120
15. Hussain SA, Al Balushi ASA (2020). A real time face emotion classification and recognition using deep learning model. In Journal of physics: Conference Series 1432(1), p 012087. IOP Publishing
16. Williams T, Li R (2018, February) Wavelet pooling for convolutional neural networks. In International Conference on Learning Representations
17. Jung H, Lee S, Yim J, Park S, Kim J (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp 2983–2991
18. Talegaonkar I, Joshi K, Valunj S, Kohok R, Kulkarni A (2019, May) Real time facial expression recognition using deep learning. In Proceedings of International Conference on Communication and Information Processing (ICCIP)
19. Elfwing S, Uchibe E, Doya K (2018) Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw 107:3–11