DOI: https://doi.org/10.15276/hait.06.2023.8
UDC 004.8

An adaptive convolutional neural network model for human facial expression recognition

Olena O. Arsirii[1)]
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com. Scopus Author ID 54419480900
Denys V. Petrosiuk [1)]
ORCID: https://orcid.org/0000-0003-4644-3678; d.petrosyuk1994@gmail.com. Scopus Author ID 54419479400
[1)] Odessa Polytechnic National University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ABSTRACT

The relevance of solving the problem of recognizing facial expressions in the image of a person's face for the formation of a model of social interactions in the development of intelligent systems for computer vision, human-machine interaction, online learning, emotional marketing, and game intelligence is shown. The aim of the work is to reduce the training time and computational resources without losing the reliability of the multivalued classification of motor units for solving the problem of facial expression recognition in a human face image by developing an adaptive model of a convolution neural network and a method for its training with "fine tuning" of parameters. To achieve the goal, several tasks were solved in the work. Models of specialized convolution neural networks and pre-trained on the ImageNet set were investigated. The stages of transfer learning of convolution neural networks were shown. A model of a convolutional neural network and a method for its training were developed to solve the problems of facial expression recognition on a human face image. The reliability of recognition of motor units was analyzed based on the developed adaptive model of a convolution neural network and the method of its transfer learning. It is shown that, on average, the use of the proposed loss function in a fully connected layer of a multi-valued motor unit classifier within the framework of the developed adaptive model of a convolution neural network based on the publicly available MobileNet-v1 and its transfer learning method made it possible to increase the reliability of solving the problem of facial expression recognition in a human face image by 6 % by F1 value estimation.

Keywords: Convolution neural networks; facial expression recognition; deep learning; transfer learning; multilabel classification

***For citation*: Arsirii O.O., Petrosiuk D.V. "An adaptive convolutional neural network model for human facial expression recognition". *Herald of Advanced Information Technology.* 2023; Vol. 6 No. 2. 128–138. DOI: https://doi.org/10.15276/hait.06.2023.8**

INTRODUCTION

Automated of facial expressions recognition facial expression recognition (FER), on depicted human faces is aimed at obtaining information about a person's mental state and is important for forming a model of social interaction. Such a model is the basis for building intelligent systems of computer vision, human-machine interaction, online learning, emotional marketing, machine graphics and game intelligence, etc. [1, 2]. In practice, the deep learning (DL) technology of convolutional neural networks (CNN) is used to solve the problem of facial expressions recognition [3, 4], [5]. However, in order to implement technological solutions of DL CNN with the necessary accuracy, in addition to high-performance graphics processors, a rather large set of pre-marked data such as ImageNet [6] is required. But the use of such popular datasets as DISFA, CelebA and Bosphorus [7, 8] for CNN training for FER does not give good results in terms of accuracy because they require solving the problem of classification by several labels (multi-value classification, or classification of overlapping classes, *multilabel classification*) [9, 10]. At the same time, it should be taken into account that intelligent systems are required for automated FER, which are developed not only for stationary, but also for mobile platforms, which imposes additional restrictions on the resource intensity of the CNN architectures used and their learning speed.

Therefore, the development of an adaptive CNN deep learning model for solving the FER problem with the necessary accuracy and resource intensity is an urgent scientific and practical task.

ANALYSIS OF EXISTING RESEARCH AND PUBLICATIONS

The basis for FER is the facial movement coding system (FACS), proposed by P. Ekman in collaboration with W. V. Friesen [11, 12], [13], which describes the movement of facial muscles using different actions units (AU). In the FACS system, to describe all possible and visually observable changes in the human face, 12 motor units are defined for the upper part of the face and 18 motor units for the lower part, which are associated with a contraction of a certain set of muscles. At the same time, actions can occur separately or in combination.

Two types of state coding are used to describe actions units. The first, simpler, is the coding of the presence or absence of an action unit on the face. In the second case, in addition to the first, the intensity or strength of the AU action is also indicated, while 5 levels of intensity coding are possible (neutral < A < B < C < D < E), where A is the least intense action, and E is the action of maximum strength [12].

Researches show that in recent years, the best results in the field of recognition of faces and objects on the stage show CNN, which is a logical development of the ideas of the cognitions and neocognitron. The undoubted advantage of using CNN for recognition tasks is taking into account the two-dimensional topology of the image, as well as invariance to scaling, rotation, shift and spatial distortion of the image. This is evidenced by the first places of CNN in the annual competition for pattern recognition from the ImageNet dataset – ImageNet large scale visual recognition challenge (ILSVRC) [6]. ImageNet is known to be a huge collectively assembled (crowdsourced) using the Amazon tool Mechanical Turk , a dataset of over 15 million high-resolution images from 22,000 categories . ImageNet large scale visual recognition challenge uses a subset of ImageNet with approximately 1000 images in each of 1000 categories. There are roughly 1.2 million training images, 50,000 validation images, and 150,000 test images.

As researches on the ImageNet dataset show, the undoubted advantage of using CNN for human face recognition is taking into account the two-dimensional topology of the image, which is very relevant when solving human FER problems based on AU. In addition, recent efforts to solve general object classification problems based on the application of CNN [14, 15] have allowed the development of more complex models that are able to generate more reliable feature representations based on the original image without explicitly considering and modeling the local characteristics of various parts of the face and relationships between facial markers. The success of this approach for human FER is evidenced by the work of other authors based on CNN models that currently show the best results on the DISFAset [7]. Thus, in [16], an EAC-Net approach for AU detection was proposed, based on adding two new networks to the previously trained network, trained to recognize AUs by features extracted from the entire image and by pre-cut individual areas of face images representing areas of interest from the point of view of recognition of certain types of AU. The authors use the CNN VGG-19 model in their approach [17, 18]. The estimated area of each AU in the image has a fixed size and a fixed location, which is determined by feature labels in the face image. Based on the structure of E-Net [16], a method of adversarial learning between AU recognition and face recognition is proposed. In [19], a JAA-Net approach with an adaptive learning module was proposed to improve the initially defined areas of each AU in the image. The results of AU classification, obtained in the considered works, will be used in this work as a basis for comparing the recognition reliability using the developed model.

As you can see, the development of CNN models for solving human FER problems based on AU requires significant costs at the stage of designing and setting up the network architecture, as well as a large set of labeled data and long and resource-intensive training. For these reasons, the *transfer learning* approach, which consists in transferring the feature description functions obtained by the CNN model with multiple layers in the process of solving the original recognition problem to the target recognition problem [17, 18], [20]. The following article shows the benefits of using a CNN trained on ImageNet data and adapted to solve the FER problem using data from the DISFA set (Denver Intensity of Spontaneous facial action). It contains videos from 27 subjects – 12 women and 15 men, with each of which a video was recorded with 4845 frames [7, 8]. Each frame is annotated with AU intensity on a 6-point ordinal scale from 0 to 5, where 0 indicates no AU, while 5 indicates maximum AU intensity. The frequency of occurrence of each AU among the 130,814 frames of the DISFA dataset is shown in Table 1. Note the severe data imbalance problem in which most AUs have a very low frequency of occurrence, while only a few other AUs have a higher frequency of occurrence.

Table1. **Number of different types of actions units in the DISFA dataset**

AU	1	2	4	6	9	12	25	26
Number of examples	6506	5644	19933	10327	5473	16851	36247	11533

Source: **compiled by the [4]**

Therefore, when developing an adaptive CNN model using the transfer learning method for solving the FER problem, it is necessary to take into account not only the imbalance of the set data, but also the presence of several AU intensity labels for each record from the set. In this formulation, one has to solve the problem of *multi-label classification*. The papers [21, 22], [23, 24] present an overview of multi-label classification algorithms. There are two main methods for solving multi-label classification problems: problem transformation methods and adaptation method. The problem transformation method transforms the problem into a set of binary classification problems. Adaptation methods perform the classification of a set of class labels; solve the problem in its entirety

All of the above made it possible to formulate the requirements for the adaptive CNN model and the method of its training.

THE AIM AND OBJECTIVES OF THE RESEARCH

The aim of the work is to reduce the training time and computational resources without losing the reliability of the multivalued classification of motor units for solving the problem of facial expression recognition in a human face image by developing an adaptive model of a convolutional neural network and a method for its training with “fine tuning” of parameters.

To achieve the goal, it is necessary to solve the following tasks:

1. Research CNN models pretrained on the ImageNet set and the possibility of their transfer learning.

2. Develop a CNN model and its training method for solving FER problems.

3. Adjust the CNN model to improve the reliability of multivalued AU classification.

4. Evaluate the reliability of AU recognition based on the developed CNN model and its training method.

PRESENTATION OF THE MAIN RESEARCH MATERIAL

Exploring convolution neural networks мodels pretrained on the ImageNet set and the possibility of their transfer learning

The ImageNet data set has been studied according to the following criteria [14]: resource intensity, classification accuracy, and performance on the NVIDIA Titan X Pascal GPU platform.

On Fig. 1 shows a diagram of the dependence of classification accuracy on the number of floating point computations for a fairly wide list of publicly available DL CNNs. The diameter of the circle in the figure corresponds to the number of adjustable parameters (resource intensity) of the CNN.

In the presented diagram, the horizontal red line indicates DL CNNs, which were selected by the authors for further research. These are DenseNet-121 [5], DenseNet-201 [5], MobileNet-v1 [25], MobileNet-v2 [26]. The choice was made taking into account the requirements for learning speed and resource intensity. The last requirement is very important when creating mobile applications for the FER task. We give the following explanations. As can be seen in the diagram, the MobileNet family of networks is superior in compactness to the DenseNet and VGG-19 networks, but inferior to them in classification accuracy on the ImageNet dataset.

At the same time, it is noted in [20] that the MobileNet family of networks is more than three times faster than the selected networks of the DenseNet family and exceeds the VGG-19 network by more than 1.5 times.

In addition, briefly note that VGG-19 is one of the first truly deep networks to achieve great accuracy on the ImageNet dataset, has 19 layers and reads more than 140M parameters. The DenseNet family of networks (Dense Convolutional Network) is designed for stationary devices and is implemented by forming a sequence of dense blocks – each block contains a set of convolution layers and transition layers that resize the feature map.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

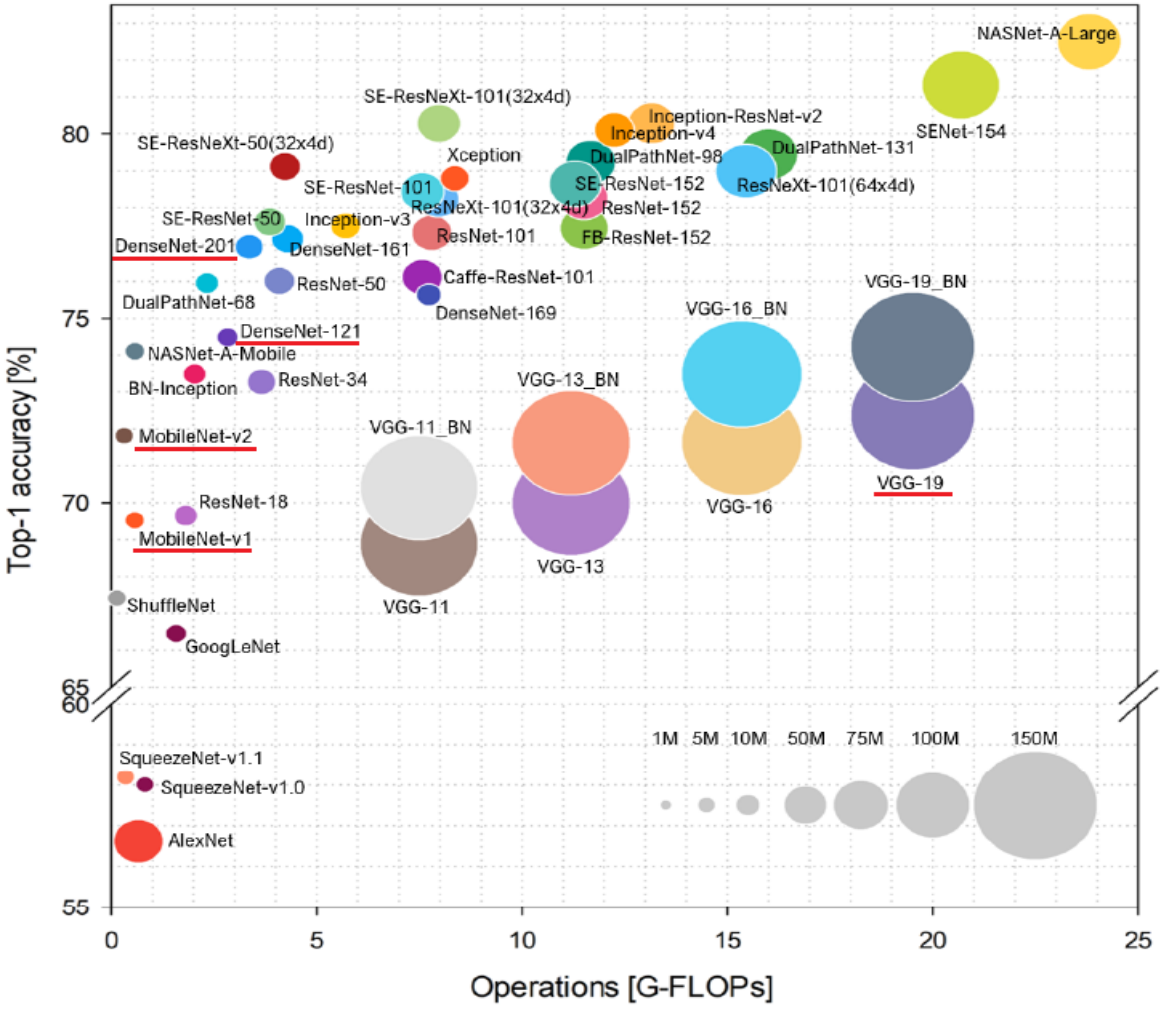

Fig. 1. **Diagram of the dependence of the accuracy of convolution neural networks classification on the ImageNet set on the number of computational operations with floating point (G-10^9FLOPs)**

Source: **compiled by the** [14]

DenseNet is a network with a much smaller number of trainable parameters (about 7M), which is two dozen times less than that of VGG-19, but the classification accuracy is comparable.

MobileNet family of networks [25, 26] due to its lightness (4-3M parameters) made a revolution in computer vision on mobile platforms. The MobileNet model is based on the deep separable convolution framework, which can transform standard convolution into deep convolution and point convolution with a 1 * 1 convolution kernel.

After choosing the so-called pre-train models, we will briefly summarize the stages of their transfer learning, which consists in transferring the feature description functions obtained by the CNN model with numerous layers in the process of solving the initial recognition problem to the target recognition problem [17, 18], [20].

Stage 1. Convolution layers are extracted from the previously trained model (pre-train).

Stage 2. The convolutional layers are frozen to avoid destroying any information they contain during future training epochs (train).

Step 3. Add some new trainable layers on top of the frozen layers. They will learn how to turn old feature maps into predictions for a new dataset.

Step 4. Train new layers on the target (target) dataset.

Step 5. The last step is fine-tune, which consists in deblocking the entire model obtained above (or part of it) and retraining on the target dataset with a very low learning rate. This can potentially lead to

significant improvements by gradually adapting pretrained networks to new data.

Thus, to develop a modelDL CNN based on transfer learning, the following DLs were chosen as pre-train models CNN: DenseNet-121, DenseNet-201, MobileNet-v1, MobileNet-v2. CNNs of the selected architectures have both high learning ability and significant speed, which meets the requirements for network training parameters when solving FER problems on stationary and mobile devices.

Development of a convolution neural networks model and its training method for solving facial expression recognition problems

Convolution neural networks model for solving the human FER problem (Fig. 2) was built on pre-trained public DenseNet numbers 121 and 201 and MobileNet versions v 1 and v 2.The CNNs of the selected architectures have both a high learning rate and a significant performance speed. For each network, fully connected layers were removed at the output, instead of which, after the subsampling layer (Global Average Pooling), an AU prediction block has been added consisting of: a BatchNormalization layer with β and γ as two summary variables for each feature [3, 4], a ReLU activation function layer, and a new fully connected output layer with a sigmoid activation function as the AU classifier.

Transfer learning method of the proposed CNN model consists of two stages: training the classifier on the target data set and fine-tuning the pre-trained network – pre-trained model CNN.

Stage 1. Batch training of the classifier by backpropagation on the target dataset consists of the following steps:

1. (Preparatory). Random values are assigned to weight coefficients of the classifier. At the entrance pre-trained model CNN, the target set of color images is fed – a tensor with dimensions m × m × 3 × N, where (m × m is the image size, N is the batch size).

2. (Forward). As a result of passing pre-trained model CNN, feature maps of certain sizes are formed, (for example, 7 × 7) in the amount L, which determines the size of the Global layer Average Pooling (for example, L= 1024). Exit Global Average Pooling corresponds to the average value of each input feature map and has the form of an L×N matrix. Batch Normalization (BN) is performed for each row of the resulting matrix. At the output of the ReLU layer, only positive values of the coefficients remain (negative ones are set to zero), which are fed to the fully connected classifier layer (Facial classifier – FC, sigmoid, see Fig. 2). Using the target values, an average learning error is calculated and averaged over the entire batch of size N.

3. (Backward). The weight coefficients of the classifier are adjusted using the backpropagation method, taking into account the Dropout operation with a thinning factor of 0.2. For the BatchNormalization layer, the coefficients β and γ are adjusted [3, 4].When the retraining of the classifier is reached (errors of the test and validation samples are tracked), the reverse pass is completed.

Stage 2. Fine tuning is performed to improve the quality of the classification. In this case, all or part of the pre-trained coefficients are unfrozen model CNN, which are also corrected by backpropagation with a low learning rate. Those all forward and backward steps of Stage 1 are performed.

The CNN model constructed in this way and the method for training it form the basis of the deep learning technology of convolutional neural networks, which makes it possible to retrain the last CNN layer using the DISFA image set in a reasonable time without changing the weight of other layers, providing the necessary reliability of AU recognition.

For the binary classification loss function for the fully connected (FC) layer (see Fig. 2), Log-Sum-Exp Pairwise (LSEP) was chosen[27], which gives better results than weighted binary cross entropy. FunctionLSEP is formalized as follows:

$$l_{lsep} = \log(1 + \sum_{v \notin Y_i} \sum_{u \in Y_i} (f_v(x_i) - f_u(x_i)), \quad (1)$$

where $f(x)$ is a label prediction function that maps the feature vector $x$into a K-dimensional label space representing the confidence scores of each label, K being equal to the number of unique labels.

One of the main properties of the function (x) is that it must produce a vector whose values for true labels Y are greater than those for false labels

$$f_u(x) > f_v(x), \forall u \in Y, v \notin Y,$$

where fu (x) is the u-element of the confidence scores for the v-instance in the dataset, respectively. Yi is the corresponding set of labels for the i-instance in the dataset.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

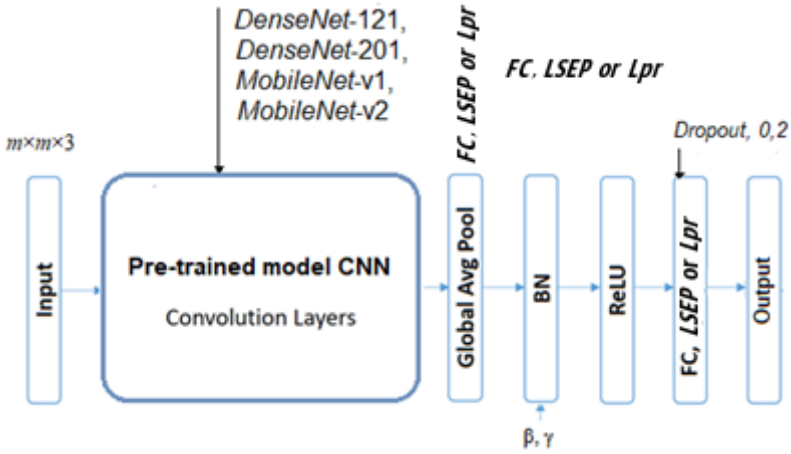

Fig. 2. **Convolutional neural networks model for solving the human facial expression recognition problem based on transfer learning technology**
Source: **compiled by the [4]**

Due to the large imbalance of data in the DISFA set, the recognition reliability was estimated by the value of the F 1-measure (the harmonic mean of the Precision and Recall indicators) as an average value – Avg.F 1 for all AUs:

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall}, \quad (2)$$

$$Precision = \frac{TP}{TP+FP}, \quad (3)$$

$$Recall = \frac{TP}{TP+FN}, \quad (4)$$

where TP are true positive examples, FP are false positive examples, FN are false negative examples.

Adjusting the convolution neural networks model to improve the confidence of multivalve actions units' classification

To solve the multivalued classification problem instead of the *LSEP function* (1) the binary function is used cross-entropies (Binary Cross Entropy, or log loss) is presented as follows:

$$L_{bce} = -ylog(p) - (1-y)log(1-p), \quad (5)$$

where y is the true label; $p$is the predicted value of the probability of the post sigmoid functions activation.

To increase the reliability of the multi-label classification for the fully connected *FC* layer (see Fig 2) was proposed a loss function, which looks as follows:

$$L_{pr} = 1 - Precision \cdot Recall. \quad (6)$$

At the same time, *TP*, *FP* and *FN* is calculated as follows:

$$TP = \sum yp,$$

$$FP = \sum (1-y)p,$$

$$FN = \sum y(1-p),$$

where y is the truth label; p is predicted value probability after sigmoid functions activation .

The view of the loss function (5) was developed taking into account research [3, 4] and the following considerations:

1) the loss function is calculated on the basis of the product of smoothed functions $Precision \cdot Recall$, which is aimed at maximizing the value of the area under the Precision – Recall curve , since the higher the value of the product, the higher and to the right is a point on its value graph;

2) the rectangle, which is formed by perpendiculars from the point of the product value $Precision \cdot Recall$ to the corresponding axes of Precision and Recall, occupies a larger area.

Actions units recognition reliability based on the developed convolution neural networks model and its training method

Testing developed adaptive CNN model and its transfer learning method was carried out on a set of DISFA images with binary classification for a fully connected layer FC using the *LSEP function* on eight AU , which determine the following states of motoractivity of the muscles of the human face: AU1 – the inner parts of the eyebrows are raised; AU2 – the outer parts of the eyebrows are raised; AU4 – lowered eyebrows; AU6 – cheeks are raised; AU9 – wrinkled nose; AU12 – the corners of the lips are

raised; AU25 – lips parted; AU26 – jaw dropped, AU27 – extended mouth.

The highest values of the measure F1 according to (2) were shown by the network models DenseNet-201 and MobileNet-v1 (Fig. 3). At the same time, the MobileNet-v2 – 2×10^6 and MobileNet-v1 – 3×10^6 networks have the smallest number of trained parameters , the DenseNet-201 – 18×10^6 and DenseNet -121 – 7×10^6 networks are more resource-intensive .

A comparative analysis of the reliability results of binary classification AU on a set of DISFA images using specialized CNNs such as EAC-Net , LP-Net and JAA-Net showed that using the *LSEP loss function* [27] and taking into account the number of training parameters, MobileNet convolutional networks-v1 and DenseNet-201 give better results from 1 to 16 % respectively (see Fig. 3)

To check the reliability of the multi-valued classification, the models of MobileNet-v1 networks pre-trained on the ImageNet set were compared with a fully connected layer *FC* built using loss functions L_{bce}(5) and L_{pr}(6) (see Fig. 4a and Fig.4b respectively).

An experiment using the proposed transfer learning method was carried out on images from the DISFA, CelebA and Bosphorus sets, learning rate10^{-5}, and batch size 64, optimizer Adam. The size of the input RGB image is 224×224. As we can see, for almost all AUs, there is a noticeable increase in the reliability of classification; there is a slight decrease in the value of F1 according to AU25. On average, the use of the proposed loss function L_{pr}in the fully connected layer of the multi-valued classifier AU of the developed robust CNN model using MobileNet-v1 and its transfer learning method allowed us to increase the reliability of the solution of the FER problem by 6 % by estimating the value of F_1.

CONCLUSION

The paper substantiates the relevance of solving the FER problem for the formation of a model of social interactions in the development of intelligent computer vision, human-machine interaction, online learning, emotional marketing, game intelligence, etc.

It is shown that CNN deep learning technology is successfully used in practice to solve the FER problem. However, to implement such solutions with the required accuracy, in addition to high-performance graphics processors, a sufficiently large set of prelabeled data of the ImageNet type is required. Using smaller datasets such as DISFA, CelebA, and Bosphorus to train CNNs for FER does not give good accuracy results because it requires solving the problem of multi-valued AU classification by several labels.

Taking into account the above, the goal of the work is formulated: to reduce the training time and computational resources without losing the reliability of the multi-valued classification of motor units for solving the problem of facial expression recognition on a human face image by developing a robust model of a convolutional neural network and a method for its training with “fine tuning” of parameters.

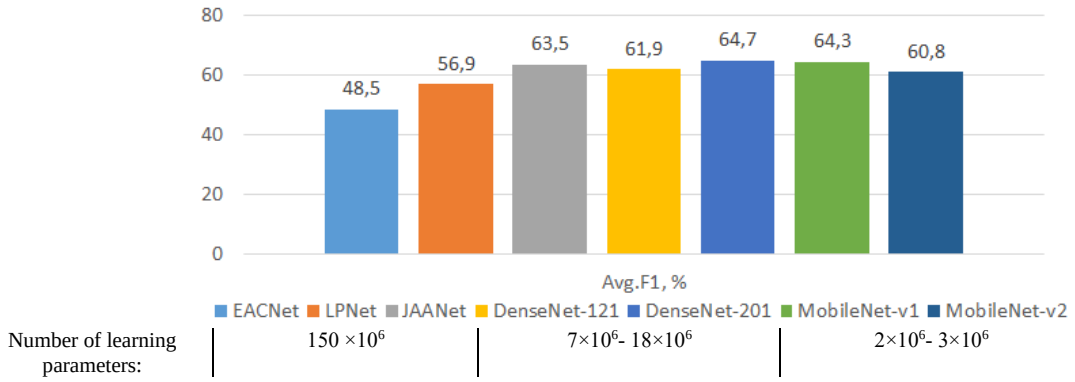

Fig. 3. **Diagram of comparative estimation of the actions units’ recognition quality when solving the facial expression recognition problem using Cevolutional neural networks for the DISFA dataset**

Source: **compiled by the [4]**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

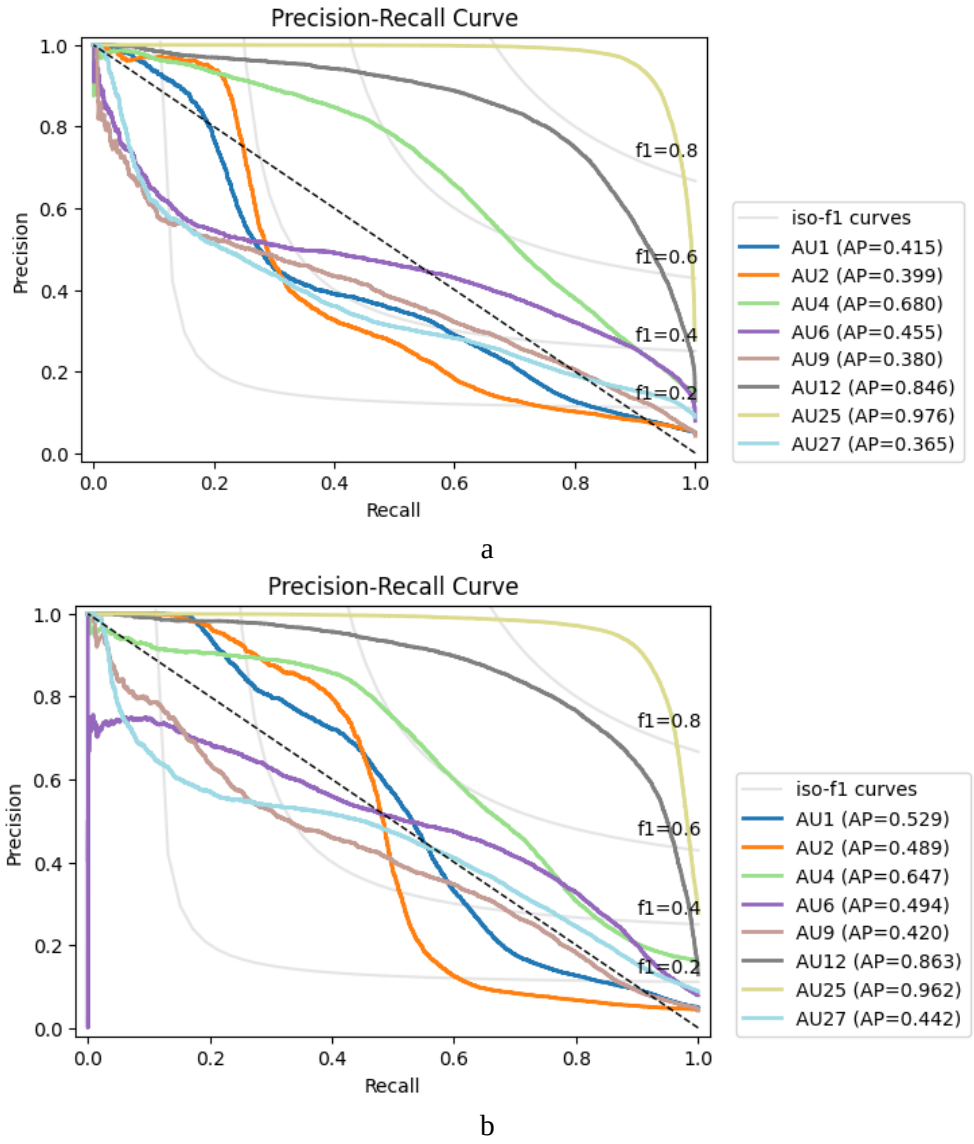

*Fig. 4.***The shape of the Precision-Recall curves and the values of the F_1 measure when classifying eight actions units when solving the facial expression recognition problem with loss functions for a fully connected layer (a - Lbce, b - Lpr)**

Source: **compiled by the authors**

To achieve the goal, the following tasks were solved in the work: models of specialized convolutional neural networks and pretrained on the ImageNet set, the stages of transfer learning of convolutional neural networks are shown; a model of a convolutional neural network and a method for its training were developed to solve the problems of facial expression recognition on a human face image, a loss function was developed for a fully connected layer of pretrained convolutional neural network to increase the reliability of the multivalued classification of motor units of human facial expressions; the reliability of recognition of motor units of facial expressions was analyzed on the basis of the developed robust model of a convolutional neural network and the method of its transfer learning.

It is shown that, on average, the use of the proposed loss function in a fully connected layer of a multivalued motor unit classifier within the framework of the developed robust model of a convolutional neural network based on the publicly available MobileNet-v1 and its transfer learning method made it possible to increase the reliability of solving the problem of facial expression recognition in a human face image by 6 % by evaluation of the value of F1.

In the following works, the authors propose to consider issues related to the use of data pseudo-labeling for machine learning of deep convolutional neural networks when solving the FER problem.

REFERENCES

1. Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C., Xiang, Y. & He, J. "A review on automatic facial expression recognition systems assisted by multimodal sensor data". *Sensors*. 2019; 19 (8): 1863. DOI: https://doi.org/10.3390/s19081863.

2. Li, S. & Deng, W. "Deep facial expression recognition: A survey". *IEEE Transactions on Affective Computing*. 2018. DOI: https://doi.org/10.1109/TAFFC.2020.2981446.

3. Arsirii, O., Petrosiuk, D., Babilunha, O. & Nikolenko, A. "Method of transfer deap learning convolutional neural networks for automated recognition facial expression systems". In: *Babichev S., Lytvynenko V. (eds). Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2021. Lecture Notes on Data Engineering and Communications Technology*. Springer, Cham. 2022; 77: 744–761. DOI: https://doi.org/10.1007/978-3-030-82014-5_51.

4. Petrosiuk, D. V., Arsirii, O. O., Babilunha, O. Ju. & Nikolenko, A. O. "Deep learning technology of convolutional neural networks for facial expression recognition". *Applied Aspects of Information Technology*. 2021; 4 (2): 192–201. DOI: https://doi.org/10.15276/aait.02.2021.6.

5. Valstar, M. & Pantic, M. "Fully automatic facial action unit detection and temporal analysis". In: *IEEE.Conference on Computer Vision and Pattern Recognition Workshop*. 2006. p. 149–149. DOI: https://doi.org/10.1109/CVPRW.2006.85.

6. "ImageNet: ImageNet overview". – Available from: https://image-net.org. – [Accessed: December, 2020].

7. Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. & Cohn, J. F. "DISFA: A Spontaneous facial action intensity database". *In: IEEE Transactions on Affective Computing*. 2013; 4 (2): 151–160. DOI: https://doi.org/10.1109/T-AFFC.2013.4.

8. "CelebFaces attributes (CelebA) Dataset". – Available from: https://www.kaggle.com/datasets/jessicali9530/celeba-dataset. – [Accessed: December, 2022].

9. Zhao, K., Chu, W. S., De la Torre, F., Cohn, J. F. & Zhang, H. "Joint patch and multi-label learning for facial action unit detection". In: *IEEE Conference on Computer Vision and Pattern Recognition.* 2015. p. 2207–2216. DOI: https://doi.org/10.1109/CVPR.2015.7298833.

10. Jiang, B., Valstar, M. F. & Pantic, M. "Action unit detection using sparse appearance descriptors in space-time video volumes". *In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition.* 2011. p. 314–321. DOI: https://doi.org/10.1109/FG.2011.5771416.

11. Ekman, P. & Friesen, W. "Facial action coding system: A technique for the measurement of facial movement". *Consulting Psychologists Press*.1978.

12. Ekman, P., Friesen, W.V. & Hager, J.C. "Facial action coding system". *A Human Face*. 2002. Available from: – https://web.archive.org/web/20080607095042/http://www.face-and-emotion.com/dataface/facs/manual/TitlePage.html. – [Accessed: December, 2022].

13. Ekman, P. "Facial expression and emotion". *American Psychologist.* 1993; 48 (4): 384–392. DOI: https://doi.org/10.1037/0003-066X.48.4.384

14. Bianco, S., Cadene, R., Celona, L. & Napoletano, P. "Benchmark analysis of representative deep neural network architectures". *IEEE Access*. 2018; 6: 64270–64277. DOI: https://doi.org/10.1109/ACCESS.2018.2877890.

15. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. "Densely connected convolutional networks". *IEEE Conference on Computer Vision and Pattern Recognition.* 2017. p. 2261–2269. DOI: https://doi.org/10.1109/CVPR.2017.243.

16. Li, W., Abtahi, F., Zhu, Z. & Yin, L. "EAC-Net: Deep nets with enhancing and cropping for facial action unit detection". *In IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2018; 40 (11): 2583–2596. DOI: https://doi.org/10.1109/tpami.2018.2791608.

17. Lim, Y., Liao, Z., Petridis, S. & Pantic, M. "Transfer learning for action unit recognition*". ArXiv.* 2018. – Available from: https://arxiv.org/abs/1807.07556v1. – [Accessed: December, 2022].

18. Almaev, T., Martinez, B. & Valstar, M. "Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection". In: *IEEE International Conference on Computer Vision.* 2015. p. 3774–3782. DOI: https://doi.org/10.1109/ICCV.2015.430.

19. Shao, Z., Liu, Z., Cai, J. & Ma, L. “JÂA-Net: Joint facial action unit detection and face alignment via adaptive attention”. *Int. J. Comput. Vis*. 2021; 129 (2): 321–340. DOI: https://doi.org/10.1007/s11263-020-01378-z.

20. Ntinou, I., Sanchez, E., Bulat, A., Valstar, M. & Tzimiropoulos, G. “A transfer learning approach to heatmap regression for action unit intensity estimation”. *ArXiv*. 2020. – Available from: – https://arxiv.org/abs/2004.06657. – [Accessed: December, 2022].

21. Baltrusaitis, T., Mahmoud, M. & Robinson, P. “Cross-dataset learning and person-specific normalisation for automatic action unit detection”. *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. 2015; 06: 1–6.

22. Zhou, Y., Pi, J. & Shi, B. E. “Pose-independent facial action unit intensity regression based on multi-task deep transfer learning”. In: *12th IEEE International Conference on Automatic Face and Gesture Recognition*. 2017. p. 872–877. DOI: https://doi.org/10.1109/FG.2017.112

23 . Li, Y., Song, Y. & Luo, J. “Improving pairwise ranking for multi-label image classification”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017. p. 1837–1845. DOI: https://doi.org/10.1109/CVPR.2017.199.

24 . Zhao, K., Chu,W-S., De la Torre, F., Cohn J. F. & Zhang, H. “Joint patch and multi-label learning for facial action unit and holistic expression recognition”. *IEEE Trans Image Process*. 2016; 25(8): 3931–3946. DOI: https://doi.org/10.1109/TIP.2016.2570550.

25 . Howard, G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. DOI: https://doi.org/10.48550/arXiv.1704.04861.

26 . Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.C. “Mobilenetv2: Inverted residuals and linear bottlenecks”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. p. 4510–4520. DOI: https://doi.org/10.1109/CVPR.2018.00474.

27 . Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C. & Zhiyong Lu, J. “ML-Net: multi-label classification of biomedical texts with deep neural networks”. *Journal of the American Medical Informatics Association*. November 2019; 26 (11): 1279–1285. DOI: https://doi.org/10.1093/jamia/ocz085.

Conflicts of Interest: the authors declare no conflict of interest

Received 05.04.2023
Received after revision 08.06.2023
Accepted 19.06.2023

DOI: https://doi.org/10.15276/hait.06.2023.8
УДК 004.8

Адаптивна модель згорткової нейронної мережі для розпізнавання міміки людини за зображенням обличчя

Арсірій Олена Олександрівна[1)]
ORCID: https://orcid.org/0000-0001-8130-9613. e.arsiriy@gmail.com Scopus Author ID 54419480900
Петросюк Денис Валерійович[1)]
ORCID: https://orcid.org/0000-0003-4644-3678; d.petrosyuk1994@gmail.com. Scopus Author ID 54419479400
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

Розпізнавання міміки на зображенні людини для формування моделі соціальних взаємодій при розробці інтелектуальних систем комп'ютерного зору, людино-машинної взаємодії, онлайн навчання, емоційного маркетингу, ігрового інтелекту, є актуальною задачею. Метою роботи є скорочення часу навчання та обчислювальних ресурсів без втрати достовірності багатозначної класифікації рухових одиниць для вирішення задачі розпізнавання міміки на зображенні людини за рахунок розробки адаптивної моделі згорткової нейронної мережі та методу її навчання з «тонким налаштуванням» параметрів. Для досягнення мети в роботі вирішено наступні завдання: досліджено моделі спеціалізованих згорткових нейронних мереж та переднавчених на наборі ImageNet, показано етапи трансферного навчання згорткових нейронних мереж; розроблено модель згорткової нейронної мережі та метод її навчання для розв'язання задач розпізнавання міміки на зображенні людини, розроблено функцію втрат для повнозв'язкового шару попередньо навченої згорткової

нейронної мережі для підвищення достовірності багатозначної класифікації рухових одиниць міміки людини; проаналізовано достовірність розпізнавання рухових одиниць на основі розроблених адаптивної моделі згорткової нейронної мережі та методу її трансферного навчання. Показано, що в середньому використання пропонованої функції втрат у повнозв'язному шарі багатозначного класифікатора рухових одиниць у рамках розробленої адаптивної моделі згорткової нейронної мережі на основі загальнодоступної MobileNet-v1 та методу її трансферного навчання дозволило підвищити достовірність розв'язання задачі розпізнавання міміки на зображенні особи людини на 6% за оцінкою значення F1.

Ключові слова: Згорткова нейрона мережа; розпізнавання міміки люди; рухові одиниці міміки; глибоке навчання; трансферне навчання; багатозначна класифікація

ABOUTTHEAUTHORS

OlenaO. Arsirii - Doctor of Engineering Sciences, Professor, Head of the Department of Information Systems, Odessa National Polytechnic University, 1, Shevchenko Ave. Odessa, 65044, Ukraine.
ORCID:https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com. Scopus Author ID 54419480900
Research field: Information technology; artificialint elligence; decision support systems; machinelearning; neuralnetworks

Арсірій Олена Олександрівна - доктор технічних наук. професор. завідувач кафедри Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Denys Valeriiovych Petrosiuk – PhD Student of the Department of Information Systems. Odessa National Polytechnic University, 1, Shevchenko Ave. Odessa, 65044, Ukraine
ORCID: https://orcid.org/0000-0003-4644-3678; d.petrosyuk1994@gmail.com. Scopus Author ID 54419479400
***Research field*:** Convolutional neural networks; facial expression recognition

Петросюк Денис Валерійович – аспірант кафедри Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна