COMPARISON OF DIFFERENT CONVOLUTION NEURAL NETWORK ARCHITECTURES FOR THE SOLUTION OF THE PROBLEM OF EMOTION RECOGNITION BY FACIAL EXPRESSION

A.O. Vorontsov [1,a], A.N. Averkin [1,2,b]

[1] *Dubna State University, Institute of system analysis and management; 141980, Dubna, Moscow reg., Universitetskaya str., 19*

[2] *FRC "Computer Science and Control" of RAS, 117333, Moscow, Vavilova str., 40*

E-mail: [a] dealwithbotalfred@gmail.com, [b] averkin2003@inbox.ru

In this paper the usage of convolution neural networks has been considered for solving the problem of emotion recognition by images with facial expression. Emotion recognition is a complex task and the result of recognition is highly dependent on the choice of the neural network architecture. In this paper various architectures of convolutional neural networks were reviewed and training experiments were conducted on selected neural networks. The overviewed neural network architectures were trained on FER2013 and AffectNet datasets, both widely used for emotion recognition experiments. A comparison of the selected neural network architectures was made using the accuracy metrics. At the end of this paper the comparative analysis was made and obtained results were overviewed.

Keywords: emotion recognition, deep learning, convolution neural networks

1. Prerequisites and the concept of emotion recognition

Emotion recognition is one of the hottest topics in artificial intelligence today. At first glance, it seems to be poorly applicable in everyday life in fact can be widely useful in various interesting areas. The use of emotion recognition technology is very diverse. You can build behavioral models of people, evaluate the quality of customer service in stores and conduct marketing research, supplement additional analytics into systems of “smart” cities, assess the emotional state of students and much more.

Emotion is a special kind of mental processes that express the human experience of their relationship to the world and themselves. Emotions play a huge role in human life and interpersonal communication. Emotions can be expressed in various ways and through various sources: facial expressions, gestures, posture, motor responses, voice. However, the face of a person possesses the most informative and the reason of it lays in its expressiveness. Each person expresses emotions in different ways, but all of them have a common basis. American psychologist Paul Ekman in one of his studies found that there is a set of emotions that are universal and can be understood by another person regardless of race, culture or gender. Such "basic" emotions are joy, surprise, anger, fear, sadness, disgust, contempt [1].

This categorical model allowed researchers to start working towards constructing first emotion classifiers. After the acceptance of convolutional neural networks as strong technology in 2012 and its subsequent development it became possible to process images in order to detect and classify emotions in images of people's faces [2]. This work formed the basis for a new direction of science - emotions recognition. In addition to the described categorical model, there are some other ones. Emotions can be detected via key points (action units) and their movements. Result depends on movements of these points and the correspondence of these movements to a specific set of action unit rules [3]. In addition, a valence-arousal scale that provides a more flexible two-parameters-way to define emotions [4]. Nevertheless, one way or another in the end it all comes down to “basic” emotion model.

The concept of emotions recognition through images consists of two steps:

1. Search for and detect faces in the image.
2. Classify emotions.

Moreover, if with the first step there are no problems today then the second one causes certain difficulties. Today there are a number of international competitions in emotion recognitions in which participants struggle finding the optimal models of neural networks in order to get greater accuracy values.

2. Our research

The first emotion recognition competition was the “Challenges in Representation Learning: Facial Expression Recognition Challenge” competition (FER2013) launched on the Kaggle platform in 2013. Participants were asked to create a classifier that would classify emotions from a photograph of a person’s face choosing emotions from seven classes. Those classes were taken from the Paul Ekman’s research described earlier in this paper. The dataset consists of 35 thousand grayscale images and each image is 48 by 48 pixels in size. Any Kaggle user is allowed to take a part in this competition even today. The small size of dataset and its general accessibility made it possible for this dataset to become leading and most used at the first research stages in the development of neural networks for recognition of emotions. Today, almost all researchers in this field use the FER2013 dataset to test the workings of the proposed and developed algorithms and models.

For comparison, the largest publicly available dataset for emotion recognition is AffectNet. AffectNet contains about a million color high-resolution images [4]. However, even this amount of data is not enough to achieve high training accuracy values.

The growth of neural network technologies directly connected to the growth of computing power and the development of cloud computing. This helped to create new types of neural networks and develop a huge variety of complicated neural network architectures. To date, there has been tremendous progress in artificial intelligence technologies. The best models of neural networks could

track and detect different objects with an accuracy above of 98%. Despite this, today researchers who engaged in the emotion recognition get mean accuracy results in the range from 60% to 70%. It can be justified by the features of the research field itself. A person's face can express several emotions at once and a neural network is prone to be mistaken choosing between them. Positive emotions are detected quite accurately, but the negative ones can be tangled. Such results are not bad and they allow conducting the research itself clearly showing how the proposed methods and algorithms work. Nevertheless, those accuracy values are not enough for applying emotion recognition technologies in production systems.

There are many competitions being held and not all of them are dedicated only to emotion recognition through people's face, but they are connected with this source of information in one way or another.

EmotioNet competition offers participants to train neural networks using action units dateset [5]. At the 2017 competition the accuracy of the model proposed and trained by the winning team was about 60% [6].

The Multimodal Emotion Recognition Challenge (MERC) challenged participants in defining emotions by a combination of voice, face, and movement. The training dataset consisted of thousands of small videos. The accuracy result of the winner was 67.9%.

In the Emotion Recognition in the Wild Challenge (EmotiW) participants were faced with several tasks one of which is similar to the MERC competition. The winners' results ranged from 59.7% to 60.3%, [7].One of the participating teams suggested using four neural networks in parallel execution with the result averaging [8].

In addition to competitions, comparative characteristics are also being made. Table 1 presents the results of pre-trained neural networks with popular architectures that were retrained with the FER2013 dataset [9].

Table 1. Accuracy results of neural networks retrained with FER2013

Neural network name	**Layers**	**Accuracy %**
GoogLeNet	22	63
CaffeNet	8	68
VGG16	16	68,2
ResNet152	152	69,7
VGG16-FACE	16	70,7

As it can be seen from table 1 there is no dependency between the depth of an neural network and the result. Moreover, deeper neural networks do not always have the best results. Each of the presented networks was pre-trained on ImageNet dataset excluding VGG16-FACE that was trained on the facial detection. Probably that causes that the result is the best among those presented networks.

In this work we attempted to create our own model of a neural network and train it from scratch using the FER2013 dataset that has been already discussed in this paper. The neural network architecture is simple and consists of 11 layers: 4 consecutive blocks of a convolutional layer and a pooling layer and 3 fully connected layers. Neural network training was conducted on virtual machines in the Amazon Web Services cloud for 20 hours on 1 GPU. The result of training was 62% for the validation batch of samples and 60% for the test batch. This is not that bad comparing to the results of neural networks presented in the table 1.

In addition, as part of this work, a second training experiment was conducted on same model with the use of the AffectNet dataset. The result is 71% validation accuracy and 67% on the test sample. This result is still in the range of values and to the further improvement, it requires both the study of the model itself and additional work with the data.

3. Future work

In order to improve the practical results of the research we plan to re-train the proposed neural network on the FER2013 dataset extended by augmentation [10]. An increase of data and its normalization (even distribution of training samples across emotion classes) can help to get a better result.

4. Conclusion

In this paper emotion recognition neural network models and the results of major competitions have been reviewed and a practical study was conducted. Training results that have been achieved does not so far behind training results obtained by most researchers in this field. Thus, now it is easy to assume that everyone can start working on the problem of emotion recognition and it does not require much. Actually, results of such studies can be quite strong (for the first step).

Unfortunately, solution of the problem of emotion recognition and real production systems are still very far away. First, it requires the accumulation of data for training, as well as the construction of new models and the development of new methods, such as combining several neural networks and averaging the result or using various sources taken together, such as detecting emotions by facial expression and voice or face and gestures.

Acknowledgement

The work was supported by the grant of RFBR 17-07-01558.

References

[1] Ekman P. Basic emotions. // In T. Dalgleish and M. Power (Eds.). Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd., 1999, pp 45-60.

[2] Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. // NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, December 2012.

[3] N. Hussain, H. Ujir, I. Hipiny, J-L Minoi. 3D Facial Action Units Recognition for Emotional Expression. // Advanced Science Letters Volume 24 (ICCSE2017), December 2017.

[4] Ali Mollahosseini, Behzad Hasani, Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. // IEEE Transactions on Affective Computing, August 2017.

[5] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, Aleix M. Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), December 2016.

[6] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, Aleix M. Martinez. EmotioNet Challenge: Recognition of facial expressions of emotion in the wild. // March 2017

[7] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi. From Individual to Group-Level Emotion Recognition: EmotiW 5.0. // Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp 524-528, November 2017.

[8] Knyazev B., Shvetsov R., Efremova N., Kuharenko A. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. // November 2017.

[9] Yin Fan, Xiangju Lu, Dian Li, Yuanliu Liu. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. // Conference: International Conference on Multimodal Interaction, At Tokyo, Japan, November 2016.

[10] Luis Perez, Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. // December 2017.