Comparative Analysis of CNN Architecture for Emotion Classification on Human Faces

Oleh Pitsun[1)], Hanna Poperechna [1)], Liudmyla Savanets [1)], Grygoriy Melnyk [1)], Bohdan Halunka[1)]

West Ukrainian National University, 11 Lvivska st., Ternopil, 46001, Ukraine

Abstract

There are many approaches to determining the emotional states of a person in an image or video. Recognizing emotions on a human face is characterized by high complexity compared to voice analysis or measurement of other indicators. There are many factors on a person's face that describe one or another emotional state of a person. Convolutional Neural Networks (CNNs) are widely used for image recognition and classification, and are well suited to the task of classifying emotions on human faces. However, there is no universal approach that perfectly classifies all types of images. Therefore, it is necessary to conduct a comparative analysis of different CNN architectures to determine the best of them for the given task. This study compares AlexNet, LeNet, and MobileNet architectures for the classification of emotional states in human face images. A structure of a convolutional neural network was also developed, which showed better results in comparison with analogues. The results of the analysis will make it possible to draw conclusions about the choice of optimal CNN architectures and develop improved proprietary architectures to improve the quality of emotion classification. Analysis of existing architectures allows to develop a new one, with better results according to the criterion of accuracy.

Keywords

Machine learning, convolutional neural networks, human face, emotions.

1. Introduction

Emotions are an integral part of human life, which allows to improve interaction between people and to understand the attitude of a specific person to a certain situation. Being able to recognize emotions such as happiness, sadness, anger, fear, surprise, and disgustmakes it possible to better understand a person or group of people. Nowadays, artificial intelligence plays a big role in people's lives, allowing to improve the quality of life in various areas. Therefore, the use of artificial intelligence, in particular convolutional neural networks (CNN), to solve the problems of assessing a person's emotional state based on the analysis of facial images is an urgent task. Convolutional neural networks are using to recognize graphic images, which have proven to be the best approach for image classification compared to other neural network approaches. There are a large number of architectures of convolutional neural networks, but they are characterized by high classification quality only for a specific type of images. Unfortunately, there is no universal approach to qualitative classification of all types of images. In addition, such approaches will avoid a subjective vision and assessment of a person's condition. In addition, determining a person's emotional state will allow to automatically identify fake states, which is relevant in connection with the significant development of technologies that allow generating fake information. An important aspect of the development of a qualitative classification system is the sample on the basis of which research is conducted.

We used the CK+48 dataset [2], which contains a sufficient number of classes and the images themselves for the study. The use of such a dataset allows to ensure the adequacy and representativeness of the analysis results.

The First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development; May 10-11, 2024, Ternopil, Ukraine

EMAIL: o.pitsun@wunu.edu.ua (O. Pitsun); hstakhyra@wunu.edu.ua (H. Poperechna); l.savanets@wunu.edu.ua (L. Savanets); mgm@wunu.edu.ua (G.Melnyk);; thisishalunka@gmail.com (B. Halunka)

ORCID: 0000-0003-0280-8786 (O. Pitsun); 0000-0001-6978-4892 (H. Poperechna); 0000-0002-0051-8905 (L. Savanets); 0000-0003-0646-7448 (G.Melnyk) ; 0009-0006-2196-6594 (B. Halunka)

The analysis of architectures of convolutional neural networks made it possible to develop a new architecture for the specific task of recognizing human emotions, which showed better results according to the criterion of accuracy.

The object of research is graphic images of human faces in various emotional states

The subject of research is the architecture of convolutional neural networks for solving classification problems.

The purpose of the work is to develop a new convolutional neural network architecture for recognizing human emotions in images based on the analysis of existing architectures.

2. Literature review

Significant attention of scientists is paid to the approaches of classification and recognition of human facial emotions Several popular datasets have been developed for these tasks, and one of them is CK+48 [2]. In [3], the authors analyzed emotions on the human face based on a dataset of student images. The authors suggest using additional sensors to increase the accuracy of emotional state determination.

The article [4] presents the results of studies of various approaches to the determination of facial emotions based on 3 different datasets. The description, structure, and formation process of the CK+48 dataset are given

in [5]. The work [6] is devoted to the development of approaches to the determination of emotions on people's faces in real time based on 6 emotional states.

A generalized analysis of approaches to recognition of human emotions and classification using neural networks is given in the articles [7,8].

In the article [9], the authors conducted a study of emotion recognition using the FER2013 dataset. The authors used convolutional networks, in particular the VGG architecture, and obtained accuracy rates within 70%.

In the article [10], the authors proposed a convolutional neural network architecture for real-time emotion recognition. In the context of this article, speed is an important factor, approaches to parallelization are given in [11].

The analysis of approaches to the application of elements of artificial intelligence and computer vision systems in the tasks of processing graphic images is considered in articles [12-14]. In particular, the analysis of approaches that allow to speed up image processing processes was carried out and new architectures of convolutional neural networks were proposed for the classification of specific types of images.

In the work, [15] authors consider an approach to searching for similar images.

3. Problem statement

To conduct the research, the following tasks should be implemented:

- Analyze existing approaches to image classification;
- Conduct research on the classification of architectures of convolutional neural networks based on the CK+48 dataset;
- Develop a new architecture for the task of image emotion recognition
- Analyze the results of the obtained studies.

4. Dataset

There are several datasets that are used to conduct research on the tasks of classifying and identifying human emotions. CK+48, which includes up to 1000 images divided into 7 classes: (angry, happy, fear, neutral, disgust, sad, surprise) [1], was chosen as the dataset in this work.

This dataset consists of images with a size of 48 x 48 pixels. To divide the dataset into test and training samples, the images were separated into separate "test" and "training" directories.

5. Generalized image classification algorithm

The generalized algorithm for the classification of images of human facial emotions is shown in Figure 1.

This stage includes both the use of existing architectures and those developed in this work.

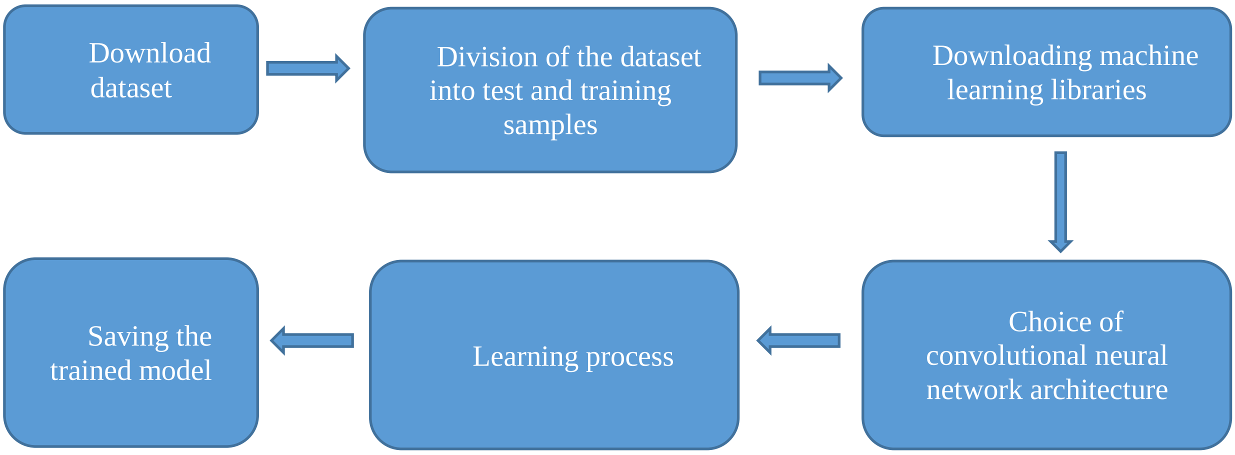
Download dataset
Division of the dataset into test and training samples
Downloading machine learning libraries
Choice of convolutional neural network architecture
Learning process
Saving the trained model

Figure 1: A generalized algorithm for the classification of images of human facial emotions

The generalized emotion classification algorithm consists of two main components: dataset formation and direct neural network training. The quality and time of training will depend on the quality of the dataset distribution and the correct selection of the convolutional neural network architecture.

6. Analysis of architectures of convolutional neural networks

Convolutional neural networks are actively used for the classification of various types of images. The use of images as input data without the need for additional conversion into other formats is the main advantage of their use. CNN consist of the following elements:

- Convolutional layers are one of the main elements of the neural network, which performs the operation of convolution of the input image with a mask. As a result, smaller images are created.
- Pooling layer – allows ones to reduce the spatial size of the collapsed function. This step also makes it possible to highlight features in the image more precisely.
- Fully Connected Layers – this layer is used at the final stage of neural network operation, which allows the model to interact with all functions of the input data.

The number of filters is a crucial parameter when working with a neural network as it directly impacts the network's quality, specifically the allocation of parameters. However, including too many parameters can lead to overtraining and worsen the results.

The size of the filters affects what patterns or features in the images are detected by the network. Larger filters are able to detect broader patterns, while smaller filters are able to detect finer details.

Strides - A larger stride will result in a smaller spatial dimension in the next layer.

Activation function - the following activation functions are the most used - ReLU, tanh, sigmoid.

The depth of the network is the number of layers in it. Greater depth allows the model to learn more complex features, but may lead to overtraining.

This convolutional neural network architecture was one of the first and proved to be effective for classifying small images. A generalized view of the convolutional neural network architecture is shown in Figure 2.

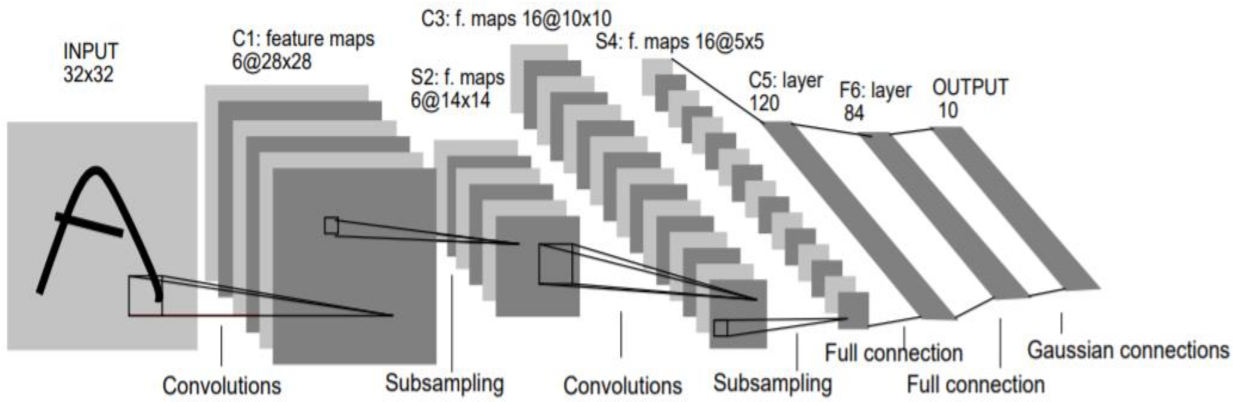

Figure 2: Generalized structure of a neural network.

This architecture is actively used for recognizing road signs or classifying human faces.

AlexNet

The AlexNet architecture is one of the most popular and was first proposed in 2012. In comparison with LeNet, this neural network is characterized by a deeper architecture, which allows to distinguish more detailed objects. ReLU is used as the activation function.

MobileNets

The process of training and operation of neural networks is characterized by the need for significant computing resources and memory.

The MobileNets architecture is designed specifically to work on mobile devices with low latency.

This architecture is designed to work with real-time systems. In many cases, this architecture shows better results compared to VGG16 and others.

A graphical representation of the blocks of the proposed convolutional network structure is shown in Figure 3. This architecture consists of convolutional layers, max-pooling layers, and Dense layers.

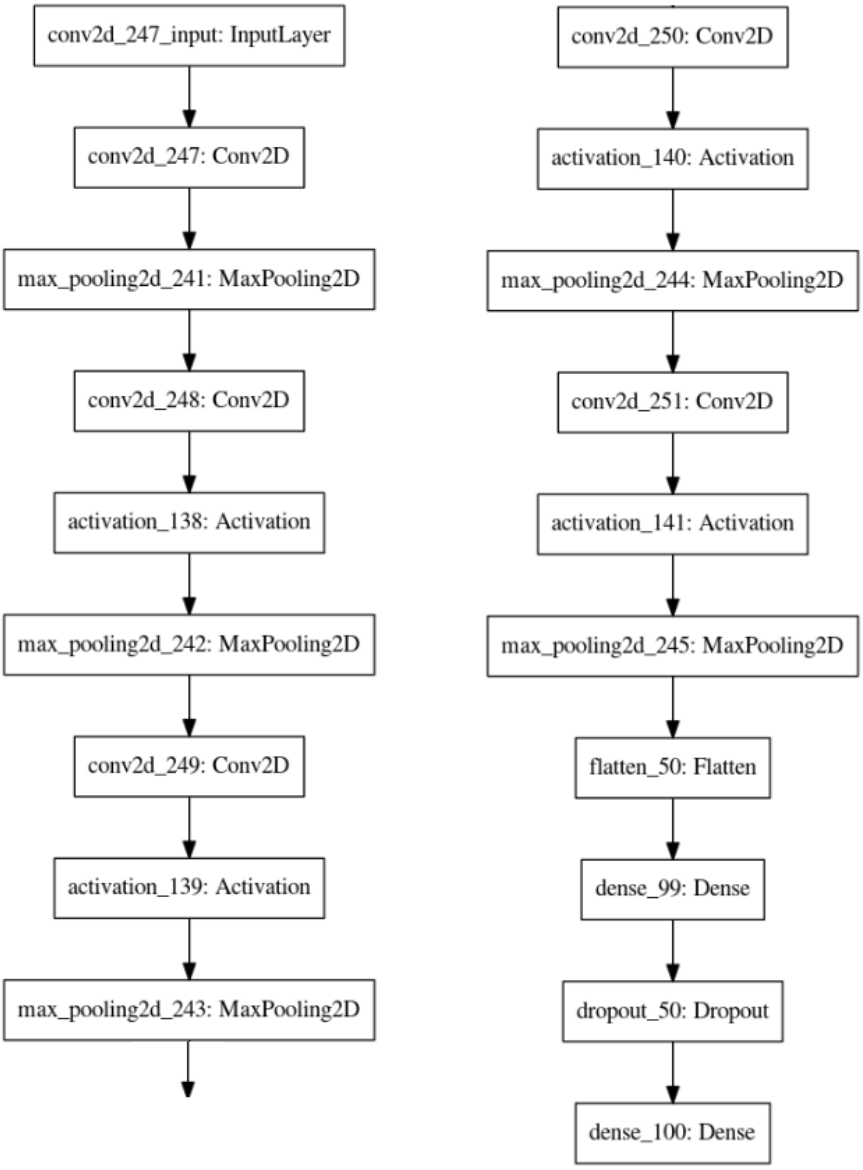
conv2d_247_input: InputLayer
conv2d_247: Conv2D
max_pooling2d_241: MaxPooling2D
conv2d_248: Conv2D
activation_138: Activation
max_pooling2d_242: MaxPooling2D
conv2d_249: Conv2D
activation_139: Activation
max_pooling2d_243: MaxPooling2D
conv2d_250: Conv2D
activation_140: Activation
max_pooling2d_244: MaxPooling2D
conv2d_251: Conv2D
activation_141: Activation
max_pooling2d_245: MaxPooling2D
flatten_50: Flatten
dense_99: Dense
dropout_50: Dropout
dense_100: Dense

Figure 3: Graphic representation of the blocks of the proposed convolutional network structure

Increasing the number of convolutional and pooling layers increases the training time, however, allows for higher research accuracy, which is a higher priority task.

This architecture was selected based on an experimental approach.

7. Computer experiments

Experiments were conducted on the same sample using different architectures with the same ratio of the training sample to the test sample.

AlexNet

The results of using the alexnet architecture are shown in Figure 4

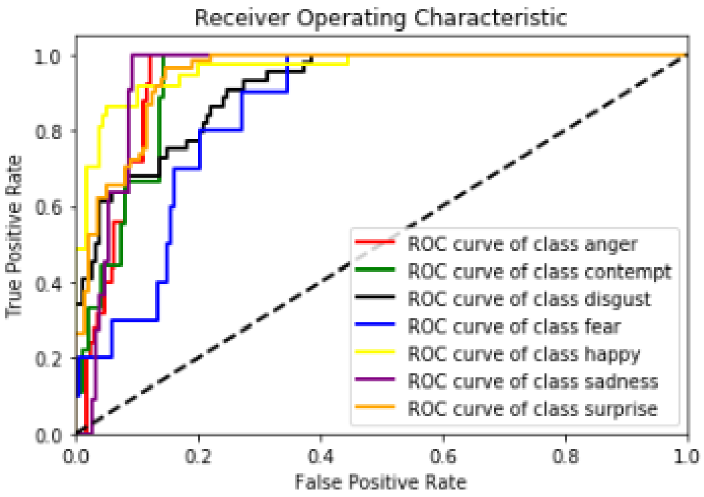

Figure 4: Results of using the alexnet architecture

This graph shows the indicators of ROC curves for 7 classes. Taking into account the given data, it can be concluded that the classification accuracy is high.

Developed architecture

The matrix of classification results is shown in Figure 5.

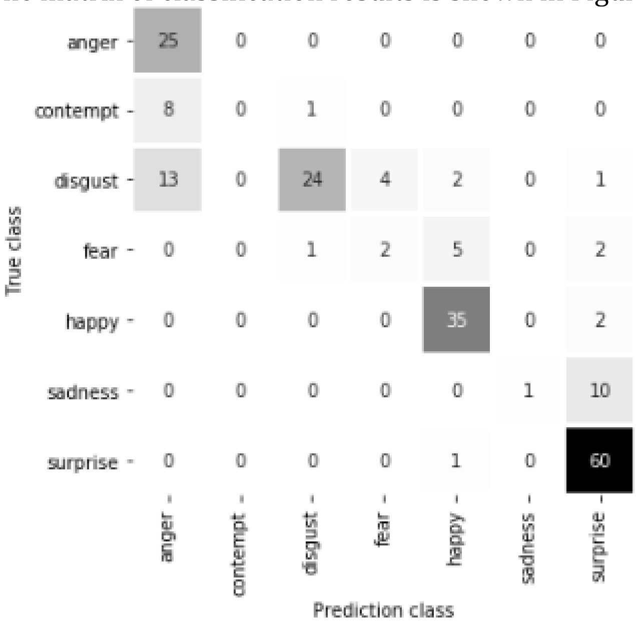

Figure 5: Confusion matrix

Layer parameters for the developed architecture for images from the CK+48 dataset:

```
Layer (type)             Output Shape          Param #
=================================================================
conv2d_73 (Conv2D)           (None, 48, 48, 16)       448
_________________________________________________________________
max_pooling2d_73 (MaxPooling (None, 24, 24, 16)        0
_________________________________________________________________
conv2d_74 (Conv2D)           (None, 24, 24, 16)       2320
_________________________________________________________________
```

```
activation_37 (Activation)   (None, 24, 24, 16)       0

max_pooling2d_74 (MaxPooling (None, 12, 12, 16)       0

conv2d_75 (Conv2D)          (None, 12, 12, 32)       12832

activation_38 (Activation)   (None, 12, 12, 32)       0

max_pooling2d_75 (MaxPooling (None, 6, 6, 32)         0

conv2d_76 (Conv2D)          (None, 4, 4, 64)         18496

max_pooling2d_76 (MaxPooling (None, 2, 2, 64)         0

conv2d_77 (Conv2D)          (None, 2, 2, 64)         102464

activation_39 (Activation)   (None, 2, 2, 64)         0

max_pooling2d_77 (MaxPooling (None, 1, 1, 64)         0

flatten_19 (Flatten)        (None, 64)               0

dense_37 (Dense)            (None, 128)              8320

dropout_19 (Dropout)        (None, 128)              0

dense_38 (Dense)            (None, 7)                903
=================================================================
Total params: 145,783
Trainable params: 145,783
Non-trainable params: 0
```

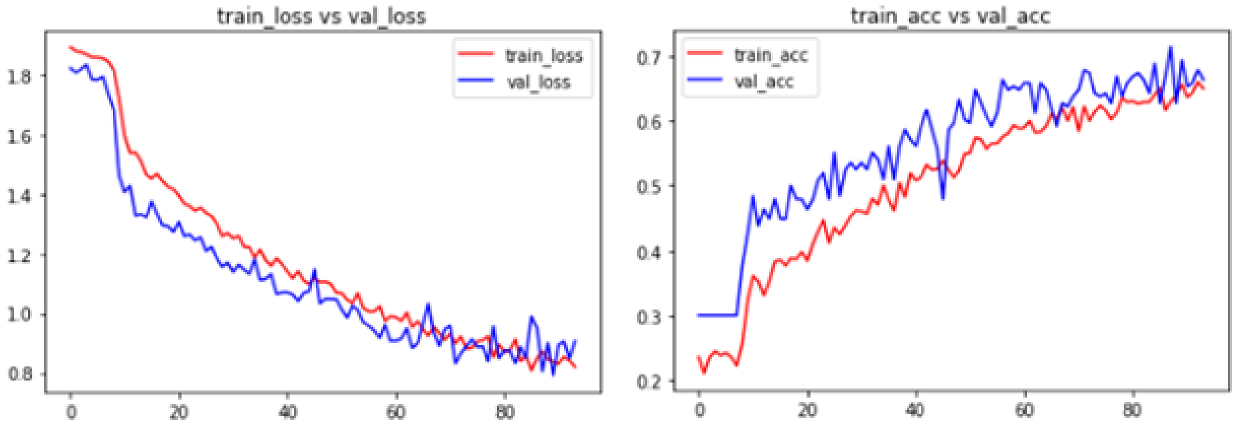

Figure 5: Results of the classification of neural networks based on the developed architecture according to the Accuracy parameter

The training results are shown in Figure 6..

This image shows training and validation metrics over training epochs for a given model. It can be observed that both losses start relatively high but gradually decrease as the training progresses. The validation loss closely follows the training loss, indicating that the model is not overfitting significantly.

The ROC curve for the developed CNN architecture is shown in Figure 6.

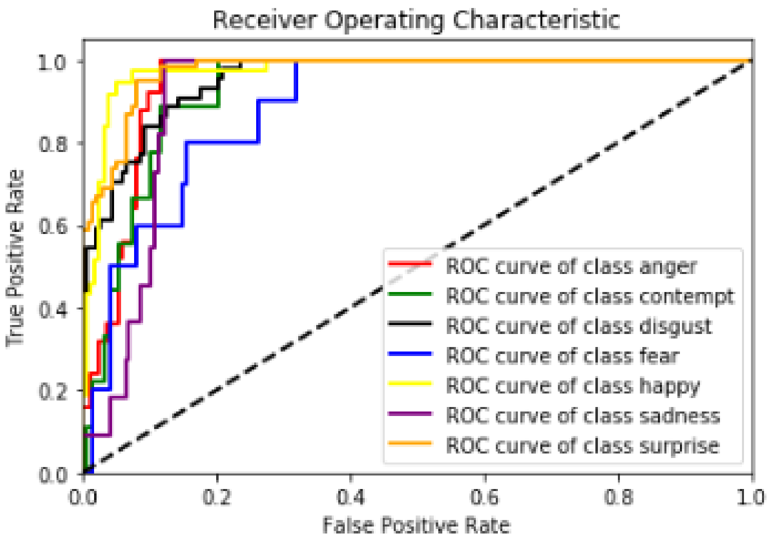

Figure 6: ROC is the curve for the designed architecture

Analyzing the above results, it can be concluded that the developed network showed better results compared to ALexNet and LeNet

The classification results using 4 neural network architectures are shown in Figure 7.

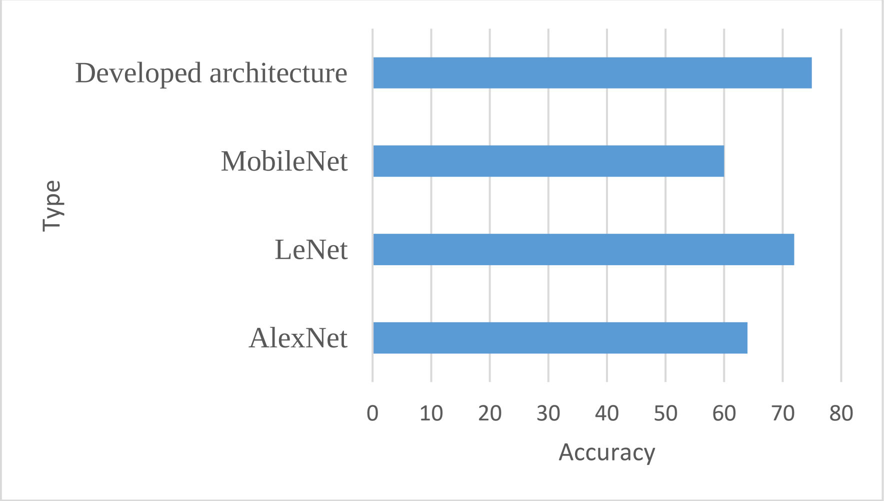

Figure 7: Classification results

As the number of classes increases, the complexity of developing a universal convolutional neural network architecture increases. Therefore, the process of designing a neural network architecture that would show the best results is an important and difficult task. The results of the comparison of the developed architecture with well-known ones show better results, which allows to define the developed architecture as possible to be used for the classification of emotions on the human face.

8. Conclusions

In this work, existing architectures of convolutional neural networks were analyzed and our own was proposed, which showed better results.

As a result of the research conducted on the basis of the CK+48 dataset, it can be concluded that convolutional network architectures show approximately similar results, however, the LeNet architecture showed better results according to the accuracy criterion - 72%, unlike ALexNet or MobileNet - 64 and 60%, respectively. Instead, the developed architecture showed even better results, namely - 75%. The number of training parameters is 145,783.

Further research needs to broaden the network of architectures for research and not to be limited only to the best known.

9. References

[1] Chaudhari, Aayushi, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. 2022. "ViTFER: Facial Emotion Recognition with Vision Transformers" Applied System Innovation 5, no. 4: 80. https://doi.org/10.3390/asi5040080

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262

[3] Magdin, Martin, Ľubomír Benko, and Štefan Koprda. 2019. "A Case Study of Facial Emotion Classification Using Affdex" Sensors 19, no. 9: 2140. https://doi.org/10.3390/s19092140

[4] Minaee, Shervin, Mehdi Minaei, and Amirali Abdolrashidi. 2021. "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network" *Sensors* 21, no. 9: 3046. https://doi.org/10.3390/s21093046

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262 .

[6] E. Dandıl and R. Özdemir, "Real-time Facial Emotion Classification Using Deep Learning", DataSCI, vol. 2, no. 1, pp. 13-17, Jul. 2019.

[7] M. Moolchandani, S. Dwivedi, S. Nigam and K. Gupta, "A survey on: Facial Emotion Recognition and Classification," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1677-1686, doi: 10.1109/ICCMC51019.2021.9418349.

[8] Hu, Tianming, Liyanage C. De Silva, and Kuntal Sengupta. "A hybrid approach of NN and HMM for facial emotion classification." Pattern Recognition Letters 23, no. 11 (2002): 1303-1310.

[9] Khaireddin, Yousif, and Zhuofa Chen. "Facial emotion recognition: State of the art performance on FER2013." arXiv preprint arXiv:2105.03588 (2021).

[10] Duncan, Dan, Gautam Shine, and Chris English. "Facial emotion recognition in real time." Computer Science (2016): 1-7.

[11] M. Dyvak, P. Stakhiv and A. Pukas, "Algorithms of parallel calculations in task of tolerance ellipsoidal estimation of interval model parameters", Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 60, no. 1, pp. 159-164, 2012.

[12] Berezsky, O., Pitsun, O., Melnyk, G., Koval, V., Batko, Y. (2023). Multi-threaded Parallelization of Automatic Immunohistochemical Image Segmentation. In: Hu, Z., Wang, Y., He, M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics IV. CSDEIS 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 158. Springer, Cham. https://doi.org/10.1007/978-3-031-24475-9_23

[13] Berezsky, O., Liashchynskyi, P., Pitsun, O., Liashchynskyi, P., Berezkyy, M. Comparison of Deep Neural Network Learning Algorithms for Biomedical Image Processing. CEUR Workshop Proceeding, 2022, 3302, pp. 135–145. https://ceur-ws.org/Vol-3302/paper7.pdf

[14] Pitsun, O. MLOps Approach for Automatic Segmentation of Biomedical Images / Oleh Berezsky , Oleh Pitsun , Grygoriy Melnyk , Yuriy Batko , Petro Liashchynskyi , Mykola Berezkyi // CEUR Workshop Proceeding, 2023, pp. 241–248. https://ceur-ws.org/Vol-3302/paper7.pdf

[15] O. Veres, B. Rusyn, A. Sachenko and I. Rishnyak, "Choosing the method of finding similar images in the reverse search system", CEUR Workshop Proceedings, vol. 2136, pp. 99-107, 2018.