APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS FOR THE CLASSIFICATION OF HUMAN EMOTIONS

Nikola Matijašević*,
Andreja Samčović,
Marko Đogatović

Faculty of Transport and
Traffic Engineering,
Belgrade, Serbia

Abstract:

Computer vision training is implemented using deep learning. Deep learning can be applied to solve a variety of classification problems. Accordingly, the problem of recognizing human emotions based on facial images can be singled out. For this problem, the use of convolutional neural networks proved to be the best solution. Specifically, the convolutional neural network models used in this paper are ResNet50, VGG16, and VGG19. The CNN model was implemented using the Keras library and the Python programming language. As input to these models, the RAF-DB database containing images of human faces with emotions was used. Based on the results obtained from the mentioned three CNN models, VGG16 proved to be the best, achieving a precision of 83.61%. VGG19 was on the second place with an accuracy of 82.37%, and the worst was ResNet50, whose accuracy was 75.31%.

Keywords:

Convolutional Neural Networks, classification problems, RAF-DB, CNN models.

INTRODUCTION

People rely on the sense of sight and use it to get to know and analyze the environment. The eye captures images of the environment and sends them to the brain for processing. The eye can be treated as a sensor that is responsible for collecting information in the form of an image and further sending it to the brain, which can be treated as a computer. In this way, we arrive at the field of artificial intelligence called computer vision. This field allows computers to extract meaningful information from digital records (images, videos, etc.) and take specific actions or provide recommendations based on them. Artificial intelligence provides computers with the ability to think (form conclusions, make decisions, etc.), while computer vision enables the sense of sight (observation, understanding of the environment, etc.). To compensate for the lack of years of experience that the human eye has, computers use cameras, large data sets, various demanding algorithms, etc. Computer vision training is implemented through deep learning.

Correspondence:
Nikola Matijašević

e-mail:
nikola.matijasevic@sf.bg.ac.rs

DOI: 10.15308/Sinteza-2023-52-59

Some enviable results have been achieved [1][2] when solving problems such as the recognition of various objects such as faces, diseases, etc. The problem of recognizing human emotions can also be singled out. This problem boils down to viewing images of human faces, detecting facial expressions, and classifying them according to the corresponding emotions. Deep learning uses convolutional neural networks (CNNs) that are suitable for solving this type of problem. The focus will be on the application of three existing CNN models on a set of images of human facial expressions and the analysis of the obtained results.

2. DEEP NEURAL NETWORKS

Deep learning is constantly changing. In recent years, significant advances have been made that have established artificial intelligence (and therefore deep learning) as one of the most promising areas of study. At the time of writing, one of the main types is convolutional neural networks [3].

2.1. CONVOLUTIONAL NEURAL NETWORK (CNN)

They are inspired by the biological nervous system and consist of different layers, where each layer has a certain number of neurons. Nodes are interconnected so that the output of a node from a layer represents the input of a node in the next layer. One of the earliest works related to the use of CNNs was published in 1998 [4], where a simple LeNet-5 model was presented that enables the extraction of simpler features of data and their aggregation, thus obtaining more complex features of data. The training process of the used CNN model was carried out using the MNIST database [5].

2.2. STRUCTURE

CNN consists of a part for extracting features (learning properties) of the input data and a part for determining the appropriate class to which the input data belongs (solving the classification problem). Both parts are built from special elements of CNNs called layers. The structure of the CNN model is given in Figure 1.

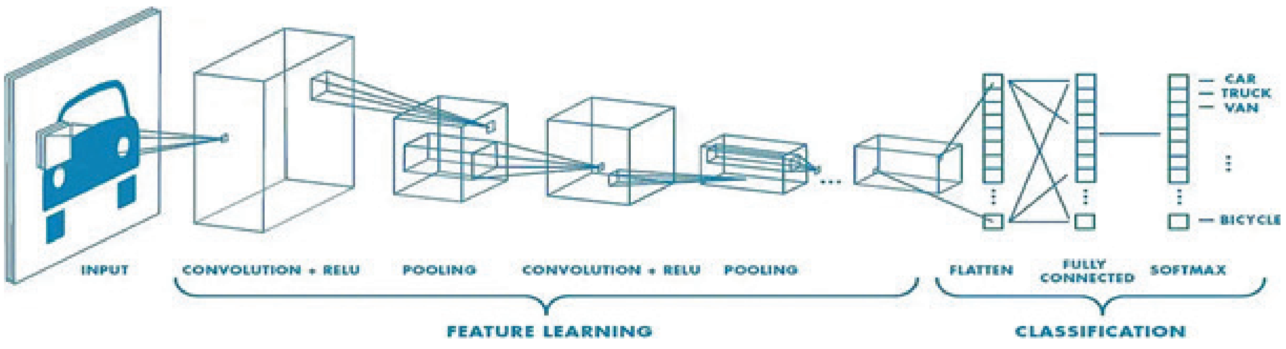


Figure 1 - The structure of the CNN model [6].

By setting the vehicle image is input to the CNN model, the learning of the features occurs which includes a combination of convolutional and pooling layers. One of the most common activation functions used is the Rectified Linear Unit (ReLU) [7]. The pooling layer allows the combining of several values into one. The second part of the CNN model is related to classification and consists of flattening and dense layers and an activation function. When it comes to solving a problem where more than two classes occur, the softmax activation function is used [8]. The focus will be on the following layer groups [9]: main, convolutional, pooling, regularization, and reshaping.

From the group of main layers, the dense layer was used to connect neurons between different layers. The output of one neuron from the previous layer represents the input of all neurons from the next layer. Convolutional layers refer to the extraction of the input data features and are used in combination with pooling. Convolution represents the use of filters of certain dimensions, with the desired step, to perform a more detailed analysis of the input data. There are various ways of pooling, the most used approach being maximum pooling where a 2x2 filter is used which passes the obtained data and returns the maximum value covered by the filter.

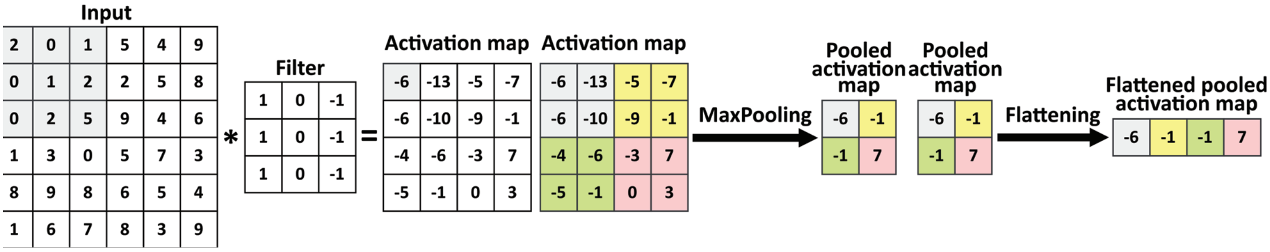
Among the regularization layers, the dropout cancels out the contribution of some randomly chosen neurons. Randomly switching off neurons is used to reduce the possibility of overfitting the CNN model. The most used layer from the reshaping group is the flattening which performs the conversion of all resulting two-dimensional arrays into a single linear vector.

2.3. THE WORKING PRINCIPLE

Input data must be provided for the CNN model which is further sent into a feature learning section. Filters in convolutional layers represent a small matrix or an activation function that serves to detect features. A filter is placed on the input image and the shaded elements are multiplied with the corresponding filter elements and summed up. By passing the filter through the entire image with a given step, an activation map is obtained. In the pooling process, a 2x2 filter is most often applied. The maximum pooling approach looks at the values covered by the filter size and selects the largest one. This approach is repeated until all activation map values for each image have been passed. The part for learning the features of the input data ends here if there are no more convolutional and pooling layers. This whole process is demonstrated through an example in Figure 2.

Furthermore, the flattened data enters fully connected layers. The first fully connected layer accepts this data and must consist of as many neurons as there are flattened data. After that, the model is trained through these layers, i.e. adjustment of weight and bias coefficients to minimize the loss function. Based on these coefficients and using activation functions f (most often ReLU), the value of each neuron can be calculated as shown in Figure 3:

For model prediction, the class with the highest neuron value is taken. However, the training of the model is done through epochs where the model should give the best possible results through each iteration. This is done by modifying the weight and bias coefficients through a process called backpropagation. The main idea behind this is to go back and analyze the obtained neuron values and coefficients. If the obtained prediction is wrong, the neurons and coefficients that influence that prediction are reduced or increased to minimize the loss function. The training stops at the moment when the coefficients can no longer be corrected. If the model did not achieve satisfactory accuracy, it is necessary to review the model, as well as the input data.



Input

2	0	1	5	4	9
0	1	2	2	5	8
0	2	5	9	4	6
1	3	0	5	7	3
8	9	8	6	5	4
1	6	7	8	3	9

Filter

1	0	-1
1	0	-1
1	0	-1

Activation map

-6	-13	-5	-7
-6	-10	-9	-1
-4	-6	-3	7
-5	-1	0	3

Activation map

-6	-13	-5	-7
-6	-10	-9	-1
-4	-6	-3	7
-5	-1	0	3

Pooled activation map

-6	-1
-1	7

Pooled activation map

-6	-1
-1	7

Flattened pooled activation map

-6	-1	-1	7

Figure 2 - Convolution, pooling, and flattening process.

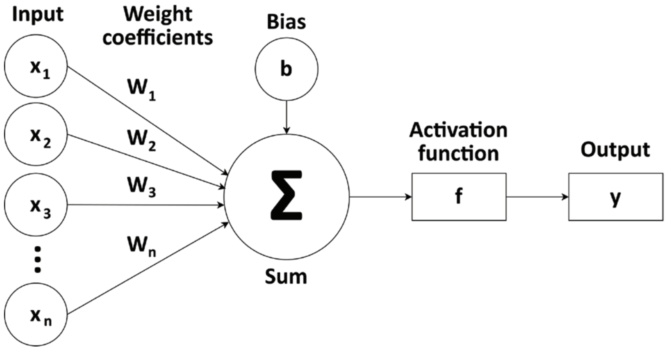


Figure 3 - Calculating the value of neurons.

3. CNN MODELS FOR THE CLASSIFICATION OF HUMAN EMOTIONS BASED ON FACIAL EXPRESSION IMAGES

In this chapter, the CNN models that were used to solve the mentioned classification problem are presented. The practical part requires the implementation of these models which was achieved by using Python libraries.

3.1. CNN MODELS

ResNet50, VGG16, and VGG19 were chosen because they are established as reliable and widely used, but also because of their availability within the Keras library. Residual Network (ResNet) is a CNN model developed by Microsoft [10]. This model can vary with the number of residual blocks they consist of. A residual block is a set of multiple layers in a model. The most used variation is the ResNet50, which consists of a total of 50 layers. ResNet152 is the first model of its kind to receive much attention. The reason for this is their victory at the ILSVRC competition, which took place in 2015, where it achieved an impressive error in image classification, which was only 3.6% [11]. In a network with residual blocks, the output from each layer is sent as input to the next layer and directly to layers separated by usually 2 or 3 layers (skip connections). This approach made it possible to solve existing problems of vanishing gradient and degradation problems [12] [13]. A new group of models was developed by the Visual Geometry Group (VGG) at the University of Oxford [14]. This group was first presented to the public at the ILSVRC competition in 2014, where it finished second [15]. This is certainly an enviable result if we take into account that they achieved a classification error of 7.3%, while in the first place, that error was slightly smaller and amounted to 6.7%. VGG models, although based on the AlexNet model [16], have several differences that set them apart from other models. Instead of using large receptive fields like AlexNet (11x11 with step 4), VGG uses very small receptive fields (3x3 with step 1). Stacking several smaller receptive fields gives better results than using a larger receptive field [14]. They consist of fewer parameters compared to the AlexNet model, which results in saving resources during training. The main two representatives of this group are the VGG16 and VGG19. The numbers 16 and 19 refer to the number of weighted layers. This means that VGG19 contains three more weighted layers compared to VGG16.

3.2. IMPLEMENTATION OF THE CNN MODELS

CNN models can be implemented using different programming languages and various libraries. However, within Python, two libraries have stood out and become the most used in this domain: TensorFlow [17] and PyTorch [18]. For easier work with CNNs within TensorFlow, the Keras [19] library appears. TensorFlow is used to solve a variety of machine learning problems, while the Keras library specializes in deep neural networks. From Keras version 2.4 onwards only TensorFlow is officially chosen as the default library for backend work. Keras library provides rapid experimentation with deep neural networks, ease of use, modularity, and extensibility.

4. PREPARATION OF THE CNN MODELS AND DATASET

The focus will be on explaining a key concept called transfer learning. This concept allows faster training of CNN models, which further provides easier experimentation with the model. Also, it is necessary to review the images and consider whether the set is adequate for use and whether any improvement can be made.

4.1. TRANSFER LEARNING IN CNN MODELS

Transfer learning has become widely represented in CNNs. The idea is to transfer the knowledge of an already trained CNN model. Mostly known CNN models are trained using more than one million images from the ImageNet [20]. That knowledge can be used as a basis for further training and solving some other classification problems. Transfer learning significantly reduces the training time because it is not necessary to train the model from the beginning. Within the Keras library, 38 CNN models can be used [21]. It is possible to set initial weights for these CNN models. These weights are important because they are treated as the knowledge of the model. The weights can be adjusted by loading the weights that were formed during the training on the ImageNet database. This means that CNN models are pretrained and ready for use. There are two strategies with which it is possible to train CNN models and apply the concept of transfer learning: feature extraction and fine-tuning [22].

4.2. DATASET OF FACIAL EXPRESSION IMAGES

The main difficulty that can arise is finding an adequate database which means that it is large enough and that the data is correct. A large number of databases are available on the Kaggle [23], although for some it is necessary to contact the creators of those databases. In this paper, RAF-DB (Real-world Affective Faces Database) [24] was used. It presents a large set of images with various facial expressions collected from the Internet including the following emotions: anger, fear, happiness, neutrality, sadness, surprise, and disgust. Due to the very small number of images within the disgust emotion, this class was discarded. The classes and the number of images are given in Figure 4.

Data augmentation can be applied to this set of images. Augmentation is performing various transformations on data. By applying augmentation, a larger dataset is obtained. Due to the already edited images in the dataset (cropped images of facial expressions, centered facial expressions, solid brightness, etc.), the only transformation that is used is the horizontal rotation of the image.

5. COMPARISON AND ANALYSIS OF THE RESULTS FOR THE CLASSIFICATION OF HUMAN EMOTIONS USING CNN MODELS

In this classification problem, three CNN models were considered, namely: ResNet50, VGG16, and VGG19. The RAF-DB dataset is divided into a set of images for training (80%) and a set of images for validation (20%). During the training and validation of the CNN model, the accuracy and losses are recorded for each of the epochs. The accuracy of the CNN model represents the ratio between the number of correct predictions and the total number of predictions. Loss measures the errors that occur in predictions and the goal is to minimize it. Various functions can be used to measure losses, and here categorical cross-entropy was used.

5.1. RESNET50 MODEL RESULTS

The ResNet50 model reached saturation after eight epochs, and after that, the training and validation of the model were completed. The accuracy and losses of the model during training and validation are shown in Figure 5.

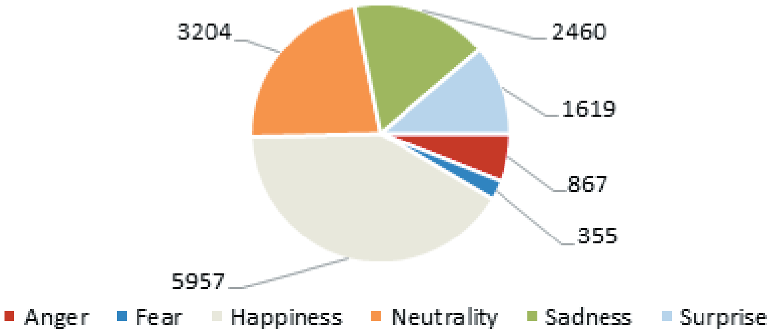


Figure 4 - Structure of the dataset.

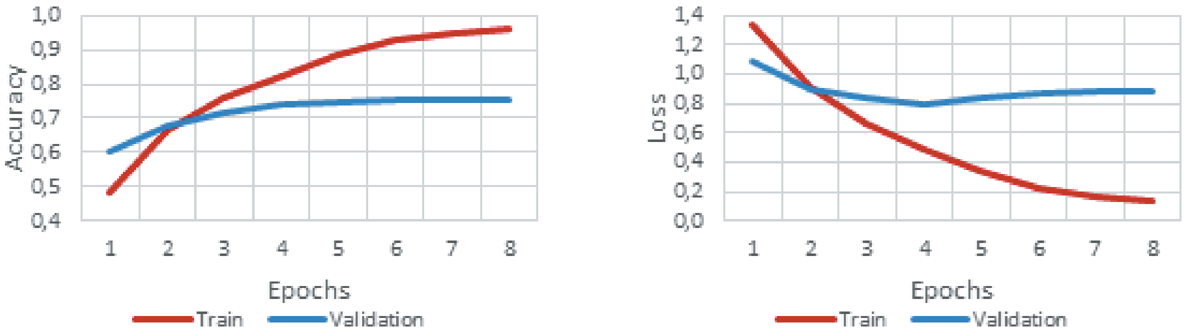


Figure 5 - Accuracy and losses of the ResNet50 model.

Model training and validation are stopped at the end of the eighth epoch because model saturation occurs (accuracy and losses converge). The final values are given in Table 1.

Table 1 - Results for the accuracy and losses of the ResNet50 model in the last epoch.

Last epoch	Training (Loss)	Training (Accuracy)	Validation (Loss)	Validation (Accuracy)
8	0.1378	0.9584	0.8893	0.7531

5.2. VGG16 MODEL RESULTS

The VGG16 model reached saturation after eleven epochs, and then the training and validation of the model were completed. The accuracy and losses of the model during training and validation are shown in Figure 6.

Model training and validation are stopped at the end of the eighth epoch because model saturation occurs (accuracy and losses converge). The final values are given in Table 2.

Table 2 - Results for the accuracy and losses of the VGG16 model in the last epoch.

Last epoch	Training (Loss)	Training (Accuracy)	Validation (Loss)	Validation (Accuracy)
11	0.1785	0.9388	0.5602	0.8361

5.3. VGG19 MODEL RESULTS

The VGG19 model reached saturation after ten epochs, so the training and validation of the model were completed after that. The accuracy and losses of the model during training and validation are shown in Figure 7.

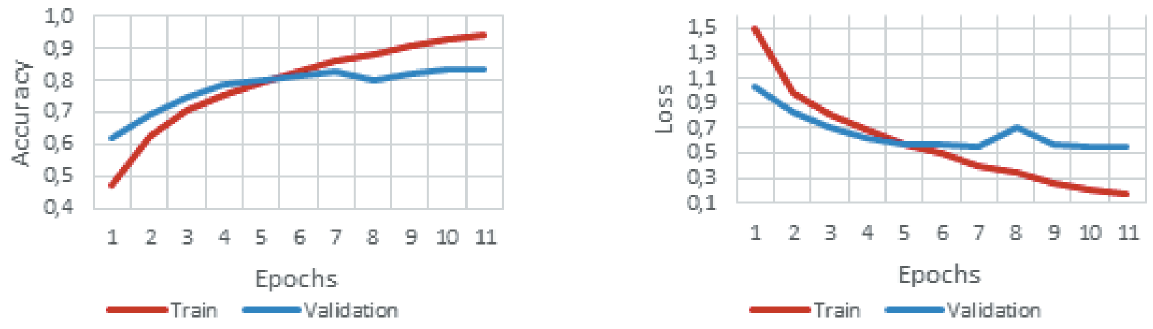


Figure 6 - Accuracy and losses of the VGG16 model.

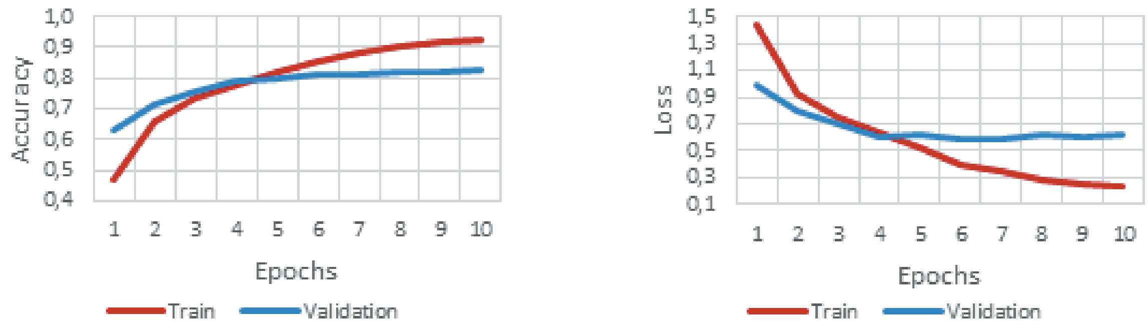


Figure 7 - Accuracy and losses of the VGG19 model.

Model training and validation are stopped at the end of the tenth epoch because model saturation occurs (accuracy and losses converge). The final values are given in Table 3.

Table 3 - Results for the accuracy and losses of the VGG19 model in the last epoch.

Last epoch	Training (Loss)	Training (Accuracy)	Validation (Loss)	Validation (Accuracy)
10	0.2295	0.9213	0.6191	0.8237

5.4. COMPARING MODEL RESULTS

For the final determination of the best model, the accuracy achieved during the validation of the CNN model is looked at. The ranking of the three CNN models considered in this paper is given in Table 4.

Table 4 - Ranking of selected CNN models.

CNN model	Validation (Accuracy)	Rank
ResNet50	0.7531	3
VGG16	0.8361	2
VGG19	0.8237	1

VGG16 turned out to be the best CNN model among those selected with an accuracy of 83.61%. On the second place is VGG19, which achieved a precision of 82.37%. ResNet50 proved to be the worst with an accuracy of 75.31%. Although VGG16 is the best of the selected models, its accuracy of 83.61% is far from the ideal 100%. Two important problems can significantly affect the accuracy. The first problem is related to the wrong choice of CNN model to solve a certain type of problem, errors in the configuration of parameters, and errors in the selection of metrics for evaluation. However, it is possible to use already trained CNN models with certain modifications if necessary which reduces implementation errors. The second problem is related to the database that is used because its size can have a big impact on training a CNN model.

RAF-DB database had the greatest impact on the obtained accuracy results. The class happiness represents 41.2% (5957 images) of the total number of images while the smallest class (fear) has only 355 images (2.45%). It is not possible to ignore this and expect the CNN model to successfully make predictions for all images that should belong to the fear class. Furthermore, some image that is associated with one of the emotions (e.g. fear) can act as an error because it can be interpreted in two ways i.e. according to human perception, that image may be more appropriate for some other emotion (e.g. surprise). Also, understanding emotions based on facial images can be misinterpreted due to cultural differences between people [25]. The last factor refers to the noise in images. RAF-DB contains images that are damaged and distorted i.e. the appearance of blank (completely white or black) images or blurred images in which the face is not visible. By going through the database, these kinds of images were largely moved out. Taking all these factors into account, it is not possible to get close to 100% accuracy without using some additional tools, methods, techniques, etc. There are several successful applications of the CNN model over the RAF-DB database. At the time of writing this paper the highest accuracy result achieved is 90.35% using the EAC model [26]. It is a combination of the ResNet50 and the EAC (Erasing Attention Consistency) method. The EAC method is used to remove the negative influence of noise in the image labels [26].

6. CONCLUSION

People's opinions are divided regarding the emergence and use of artificial intelligence. Some optimists think that artificial intelligence will solve all of the problems that exist, while there are pessimists who think that this is the beginning of the end. Although one should not be at the end of one side or the other, it is necessary to look at the real possibilities of artificial intelligence to gain new knowledge, improve the daily life of mankind, overcome limitations, etc. This paper dealt with the use of CNN models for solving the problem of classification of images of human faces and their emotions. ResNet50, VGG16, and VGG19 models were used. The input data belongs to the RAF-DB dataset. Based on the concept of transfer learning and modification of the RAF-DB dataset, accuracy results were obtained. The VGG16 model stood out as the best, achieving a precision of 83.61%. On the second place was the VGG19 model with an accuracy of 82.37%, and on the third place was the ResNet50 model with 75.31%. These model accuracy results are not ideal, but there is still room for improvement. One of those improvements refers to the combined application of existing CNN models and additional tools, techniques, methods, etc. The direction of future research is the application of additional tools, techniques, and methods to surpass the currently best-achieved result (EAC model). Also, some other ideas include comparing the results of other CNN models or more of those models over the same dataset. Other datasets contain images of facial expressions that can be used to solve the problem of emotion classification.

7. REFERENCES

[1] M. Bakator, D. Radosav, "Deep Learning and Medical Diagnosis: A Review of Literature", *Multimodal Technologies and Interaction,* 2018.

[2] G. Hu, et. al., "When Face Recognition Meets With Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[3] A. Shrestha, A. Mahmood, "Review of Deep Learning Algorithms and Architectures", *IEEE Access,* vol. 7, 2019.

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

[5] MNIST database, [Online], https://yann.lecun.com/exdb/mnist/

[6] MATHWORKS, [Online], https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

[7] F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)", *Computing Research Repository (CoRR)*, 2018.

[8] C. Nwankpa, W. Ijomah, A. Gachagan, S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning", *Computing Research Repository (CoRR)*, 2018.

[9] Keras Layers, [Online], https://keras.io/api/layers/

[10] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[11] ILSVRC 2015, [Online], https://image-net.org/challenges/LSVRC/2015/results

[12] A. Arohan, A. Koustav, S. Abhishek, "A Review of Convolutional Neural Networks", *International Conference on Emerging Trends in Information Technology and Engineering*, 2020.

[13] S. Genevieve, M. Wasfy, "An Overview of Recent Convolutional Neural Network Algorithms for Image Recognition", *IEEE Xplore,* 2018.

[14] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *International Conference on Learning Representations,* 2015.

[15] ILSVRC 2014, [Online], https://image-net.org/challenges/LSVRC/2014/results

[16] A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[17] TensorFlow, [Online], https://www.tensorflow.org/

[18] PyTorch, [Online], https://pytorch.org/

[19] Keras, [Online], https://keras.io/

[20] ImageNet, [Online], https://image-net.org/

[21] Keras Applications, [Online], https://keras.io/api/applications/

[22] R. Kaur, R. Kumar, M. Gupta, "Review on Transfer Learning for Convolutional Neural Network", International Conference on Advances in Computing, Communication Control, and Networking, pp. 922-926, 2021.

[23] Kaggle, [Online], https://www.kaggle.com/

[24] S. Li, W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition", *IEEE Transactions on Image Processing*, vol. 28, pp. 356-370, 2019.

[25] R. Jack, O. Garrod, H. Yu, P. Schyns, "Facial expressions of emotion are not culturally universal", *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[26] Y. Zhang, C. Wang, X. Ling, W. Deng, "Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition", 2022.