ISSN 2224-087X. Електроніка та інформаційні технології. 2020. Випуск 13. С. 38–49
Electronics and information technologies. 2020. Issue 13. P. 38–49

УДК 004.93'12 DOI: https://doi.org/10.30970/eli.13.4

FACE EMOTION RECOGNITION WITH CONVOLUTIONAL NEURAL NETWORK

O. Kaskun, R. Shuvar, A. Prodyvus

Ivan Franko National University of Lviv,
50 Drahomanova St., 79005 Lviv, Ukraine
oleh.kaskun@lnu.edu.ua , roman.shuvar@lnu.edu.ua ,
andriy.prodyvus@lnu.edu.ua

Facial expression recognition is the part of facial recognition that is gaining more importance and its need increases enormously. The challenges include face identification and recognition, suitable data representation, appropriate classification scheme, appropriate database, among others.

Although there are methods to identify expressions using machine learning and artificial Intelligence techniques, this work tries to use deep learning and image classification method to recognize expressions and classify expressions according to images.

In this paper, we propose to implement a general convolutional neural network (CNN) building framework for designing real-time CNNs. We validate models by creating a real-time vision system that performs face detection and emotion classification tasks simultaneously in one step using proposed CNN architecture.

Keywords: convolutional neural network, face detection, facial recognition, TensorFlow, deep learning.

Introduction

The recognition of human expressions and emotions has caught the attention of researchers, as the capability of recognizing expressions helps in human-computer interaction [1], to right advertising campaigns, crowning with an augmented and enhanced human communication.

There are many ways to inspect the recognition of human expressions, ranging from facial expressions [2], body position, tone of voice [3], etc. Emotions can be expressed in many different ways that may or may not be seen by the naked eye. Therefore, with the right tools, any indication that precedes or follows them may be subject to detection and recognition. In this article, we have focused on the recognition of facial expressions.

The application must be able to detect information only from the face of its user to identify the emotional state. The correct interpretation of any of these elements using machine learning techniques has turned out to be complicated due to the high variability of the samples within each task. This carries models with millions of parameters trained under miles of samples [4].

Furthermore, human precision to classify an image of a face into one of 7 different emotions is 65% ± 5% [5]. The difficulty of this task can be seen when trying to manually

classify the images from the FER-2013 [5] data set in Fig. 1 into the following classes {"angry", "disgust", "fear", "happy", "sad ", "surprise", "neutral"}.

Fig. 1. Training data of FER2013. Each row consists of faces of the same expression: starting from the first row: Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral.

The most modern imaging-related tasks, such as image classification and object detection, rely on convolutional neural networks (CNN). These tasks require CNN architectures with millions of parameters; therefore, their deployment in low-performance platforms and real-time systems becomes unfeasible. In this paper, we propose to implement a general CNN building framework for designing real-time CNNs. The implementations have been validated in a real-time facial expression system that provides face-detection and that achieves human-level performance when classifying emotions.

Related works

In one of the most iconic works on emotion recognition by Paul Ekman [6], happiness, sadness, anger, surprise, fear, and disgust were identified as the top six emotions (besides neutral). Ekman later developed FACS using this concept, so he established the standard setting for work on emotion recognition. Neutral was also included later, in most human recognition data sets, resulting in seven basic emotions [6].

Earlier developments on emotion recognition rely on the traditional two-step machine learning approach [19]. In the first step, some features are determined from the images, and in

the second step, a classifier (such as SVM, neural network, or random forest) is used to detect emotions. Some of the popular handmade features used for facial expression recognition include Oriented Gradient Histogram (HOG) [7], local binary patterns (LBP) [8], Gabor wavelets and Haar features [9]. A classifier would then assign the best emotion to the image. These approaches seemed to work fine on simpler datasets, but with the advent of more challenging datasets (which have more class variation), they started to show their limitation.

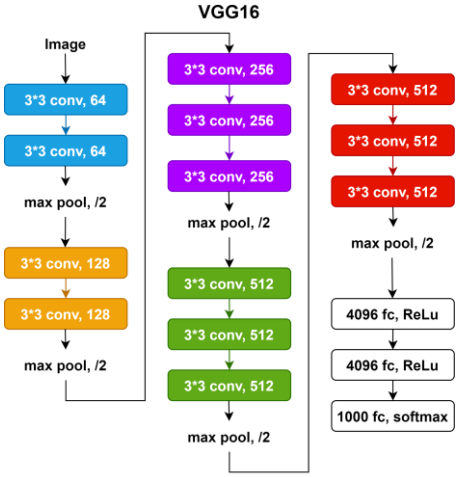

Fig. 2. VGG16 - Convolutional Network for Classification and Detection.

The popular CNNs used for feature extraction include a set of fully connected layers at the end. Fully connected layers often contain most of the parameters in a CNN. In particular, VGG16 [10] contains approximately 90% of all its parameters in its last fully connected layers. Figure 2 illustrates the complete architecture of VGG16 Model.

New architectures like Inception V3 [11], is shown in Fig. 3, reduced the number of parameters in its last layers by including a Global Average Pooling operation. Global Average Pooling reduces each feature map into a scalar value by taking the average of all the items on the feature map. The average operation forces the neural network to select the global features of the input image. Xception [12], a variant of an Inception module, is shown in Fig. 4. This CNN architecture takes advantage of the combination of two of the most successful CNN experimental assumptions: the use of residual modules and depth-wise separable convolutions [13]. Depth-wise separable convolutions further reduce the number of parameters by separating features extraction processes and combining features within a convolutional layer.

In addition, the modern model for the FER-2013 dataset is based on CNN trained with the square hinged loss [14]. This model achieved an accuracy of 71% [15] using about 5 million parameters. In this architecture, 98% of all parameters are located in the last fully connected layers. The method using an ensemble of CNNs achieved an accuracy of 66% [15].

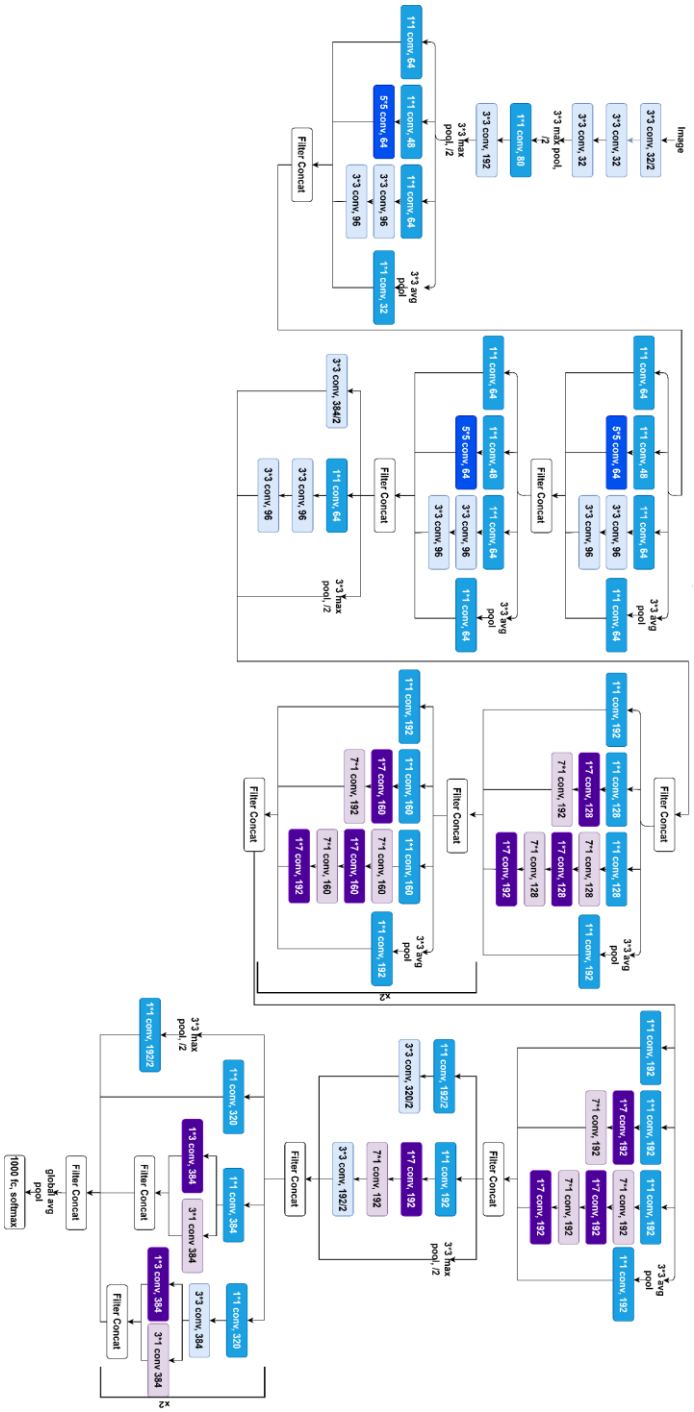

Fig. 3. Architecture of Inception V3

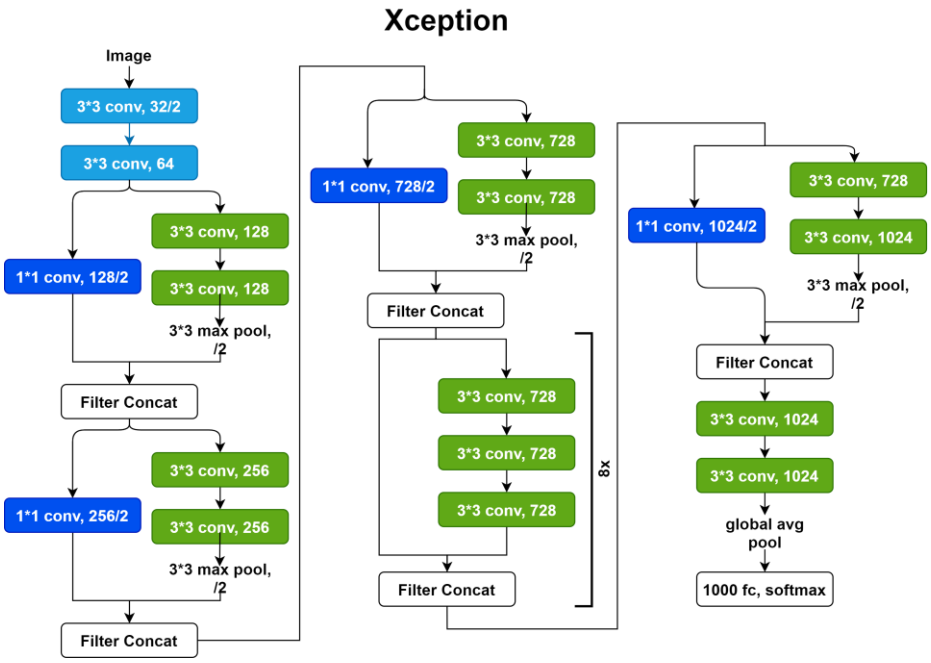

Fig. 4. Architecture of Xception.

All of the above models achieve significant improvements compare to the traditional on emotion recognition, but there seems to be missing a simple piece for attending to the important face regions for emotion recognition. In this work, we try to solve this problem, by proposing a model based on the attention convolutional network, which is able to focus on salient face regions.

Proposed model

During the development, we created two different test versions of the models. Both versions were designed with the idea of creating the best accuracy with less number of parameters. Reducing the number of parameters helps us overcoming major problems. The use of a small CNN model divests us from slow performances and the reduction of parameters provides a better generalization.

Our first model relies on the idea of avoidance completely fully connected layers. The second architecture combines the removal of the fully connected layer and the inclusion of the combined depth-wise separable convolutions and ResNet modules. We use separable convolution blocks to replace the traditional convolution layer. Separable convolution usually used to extract spatial correlation and channel correlation of residuals, to increase the signal to noise ratio, and obviously improve accuracy.

Both architectures were trained with the Adam Optimization Algorithm [16]. We use this popular algorithm in the field of deep learning because it achieves good results fast. Using large models and datasets, we demonstrate Adam can efficiently solve practical deep learning

problems. For example, we found a comparison of Adam to Other Optimization Algorithms Training a Multilayer Perceptron, this can be observed in Fig. 5.

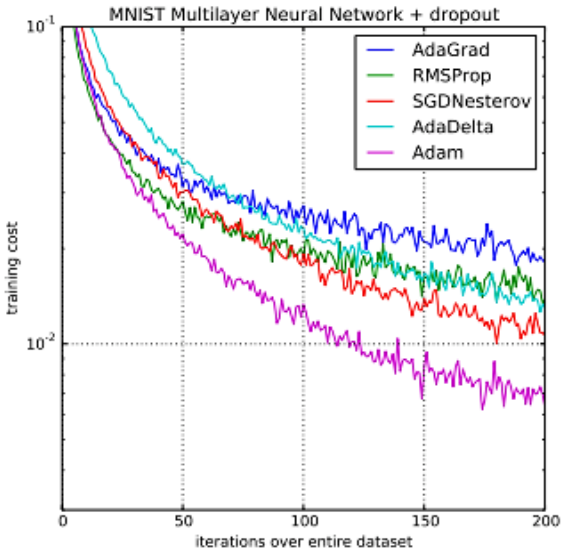

Fig. 5. Comparison of Adam to Other Optimization Algorithms Training a Multilayer Perceptron

According to the previous models, our first model used Global Average Pooling to completely remove any fully connected layers. We reached it by having in the last convolutional layer the same number of feature maps as number of classes, and applying a softmax to every reduced feature map. Our first model is a typical fully convolutional neural network composed of 9 convolution layers, ReLUs, Batch Normalization, and Global Average Pooling. About 600,000 parameters are contained in this model. It was trained on the FER-2013 dataset. Our first model achieved an accuracy of 66%.

Dataset FER2013: The Facial Expression Recognition 2013 (FER2013) database was first introduced in the ICML 2013 Challenges in Representation Learning [17]. This dataset contains 35,887 grayscale images of 48x48 resolution, most of which are taken in wild settings. Originally the training set contained 28,709 images, and validation and test each includes 3,589 images. This database was created using the Google image search API and automatically registered faces. Faces are labeled to one of the following classes {“angry”, “disgust”, “fear”, “happy”, “sad”, “surprise”, “neutral”}. Compared to the other available datasets, FER has more variation in the images, including face obstruction (mostly with hand), partial faces, low-contrast images, and eyeglasses. Sample images from the FER dataset are shown in Fig. 1.

Our second model is inspired by Xception [12]. This model unify the use of residual modules [18] and depth-wise separable convolutions [13]. Residual modules change the desired mapping between two subsequent layers so that the learned features become the difference between the original feature map and the desired features. Therefore, the desired features $G(x)$ are modified in order to solve an easier learning problem $F(x)$ such that:

$$G(x) = F(x) + x$$

When we deleted the last fully connected layer in the first proposed architecture, we reduced the number of parameters by removing them from the convolutional layers. This was done with using of depth-wise separable convolutions. Depth-wise separable convolutions are composed of two different layers: depth-wise convolutions and pointwise convolutions. The main goal of these layers is to separate the spatial cross-correlations from the channel cross-correlation [12].

This was done by applying an *S*x*S* filter on every *M* input channels and after applying *N* 1x1x*M* convolution filters to mix the *M* input channels into *N* output channels. Applying 1x1x*M* convolutions combines every value in the feature map without considering their spatial relation to the channel.

Depth-wise separable convolutions ease the computation with relation to the standard convolutions by a factor of $\frac{1}{N} + \frac{1}{S^2}$ [13]. The visual difference between a normal convolution layer and a depth-wise separable convolution can be observed in Fig. 6.

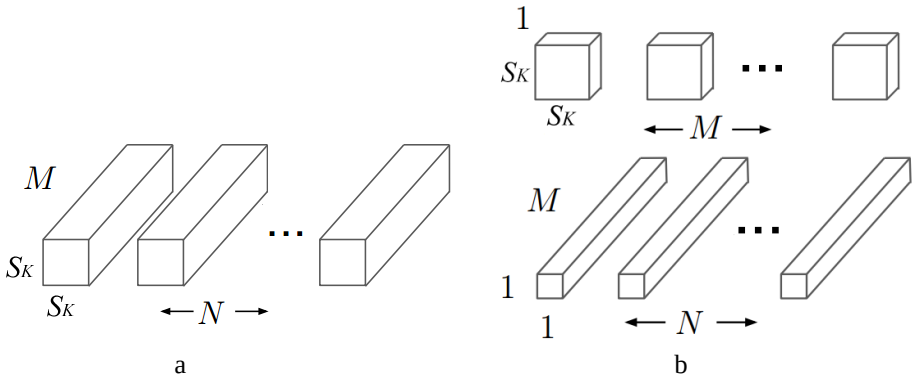

Fig. 6. Comparison (a) standard convolutions and (b) depth-wise separable convolutions.

Our last model is a fully convolutional neural network that contains 4 residual depth-wise separable convolutions where after every convolution follows a batch normalization operation and a ReLU activation function. In the last layer, we apply a global average pooling and a soft-max activation function to make a prediction. In this model, we have approximately 60,000 parameters; which present a decrease of x10 when compared to our first implementation, and ~x80 when compared to the other classic CNN.

Our model is like cropped Xception, which can observe in Fig. 7. This model obtains an accuracy of 66% for the emotion recognition task. Final model weights can be stored in a small file, around 800 kilobytes.

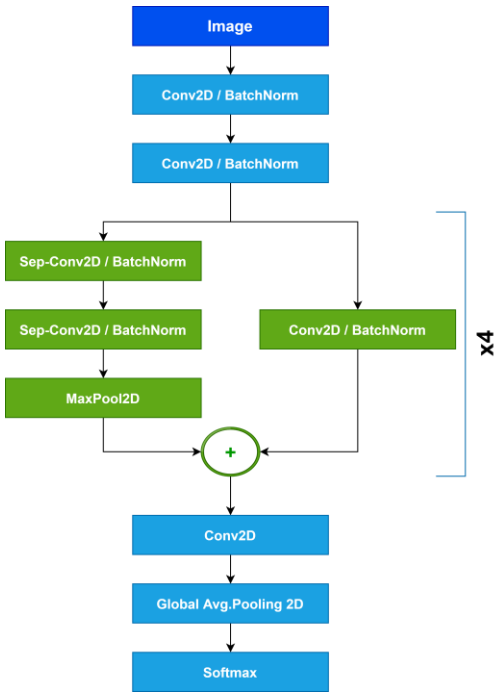

Fig. 7. Our model designed based on Xception.

Results

Results of the emotion recognition task on faces can be observed in Figure 8. In Figure 9 we show the confusion matrix results of our emotion recognition model. We can observe some common misclassifications such as predicting “angry” instead of “disgust” and predicting “sad” instead of “fear”. We compared the learned features between some emotions and the result of both models can be observed in Fig. 10.

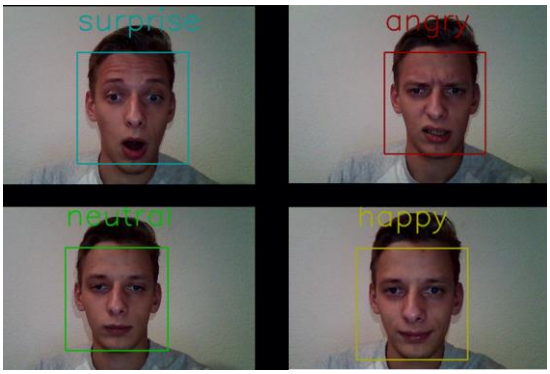

Fig. 8. Results of the emotion classification

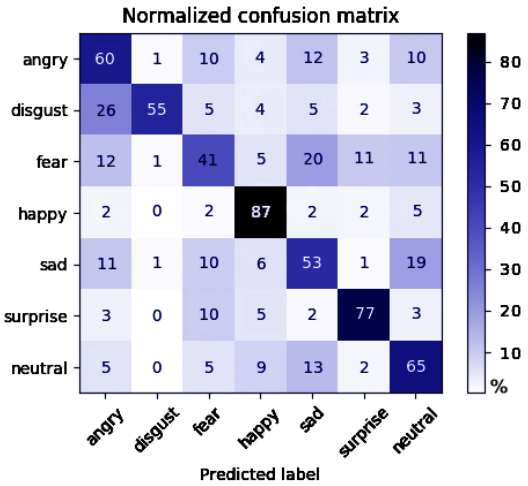

Fig. 9. Normalized confusion matrix of our second proposed model of network.

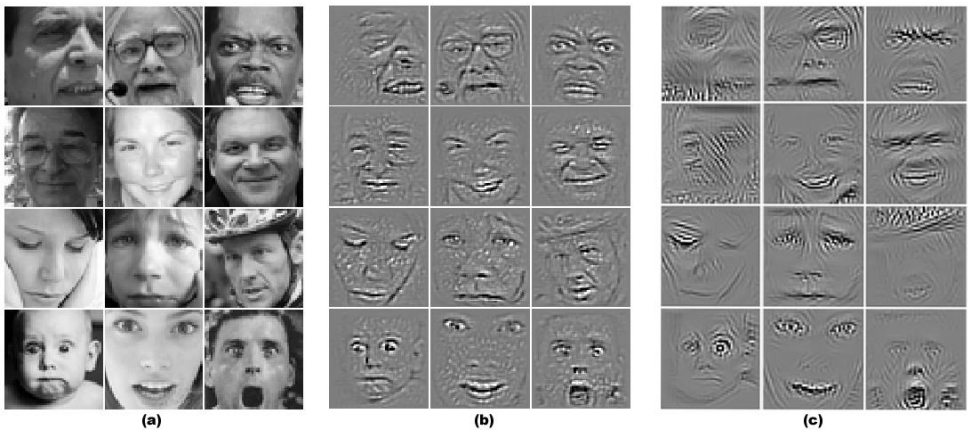

Fig. 10. Every row in all sub-figures starting from the top corresponds in accordance with the emotions "angry", "happy", "sad", "surprise". (a) Samples from the FER-2013; (b) Back-propagation visualization of our final model; (c) Back-propagation visualization of our sequential fully CNN.

Also, we have deployed a visualization method to highlight the salient regions of face images which are the most crucial parts thereof in detecting different facial expressions. The light areas in figure 8b belong to the pixel values that activate a selected neuron in our last convolution layer. The selected neurons were always selected in accordance with the highest level of activation. So, we can observe that the network learned to get activated by considering features such as the eyebrows, the teeth, and the widening of eyes and that every feature remains constant over the same class.

Results show that the neural network learned to interpret understandable human-like features, which provide generalizable elements. Due to results, we understand some common misclassifications. For example, person with glasses being classified as "angry". This happens because features for class "angry" are highly activated when the network considers a person is frowning, and these features get confused with glass frames.

In addition, we can observe that the features learned in our last model based on Xception are more interpretable than the ones learned from sequential fully CNN. So using more parameters in our simple implementations provides less robust features.

Conclusion

This paper proposes the new models for facial expression recognition using an attentional convolutional network. We have proposed models that have been systematically built to reduce the number of parameters, but with saving quality of classification. We started by eliminating completely the fully connected layers and by reducing the number of parameters in the remaining convolutional layers with depth-wise separable convolutions.

Also, we have developed a system that performs face detection and emotion recognition in a single application. We are very close to human-level performance in our classification task.

Finally, we made a visualization of the learned features in our neural network using the guided back-propagation visualization. This visualization shows us the high-level features learned by our models and gives us the opportunity to discuss and compare their interpretability.

REFERENCES

1. *R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor.* ”Emotion recognition in human-computer interaction.” IEEE Signal processing magazine 18, no. 1: 32-80, 2001.
2. *A. Mollahosseini, D. Chan and M. H. Mahoor.* ”Going deeper in facial expression recognition using deep neural networks.” 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). DOI: 10.1109/WACV.2016.7477450
3. *Kun Han, Dong Yu and Ivan Tashev*. ”Speech emotion recognition using deep neural network and extreme learning machine.” 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014), Singapore September 14-18, 2014.
4. *Amodei D, Anubhai R, Battenberg E. et al.* “Deep speech 2: End-to-end speech recognition in English and Mandarin.” CoRR, vol. abs/1512.02595 3, 2015.
5. *Xavier Glorot, Antoine Bordes and Yoshua Bengio*. ”Deep sparse rectifier neural networks”. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 315–323, 2011.
6. *Ekman Paul and Wallace V. Friesen.* ”Constants across cultures in the face and emotion.” Journal of personality and social psychology 17.2: 124, 1971.
7. *Paul V.C. Hough*. ”Method and means for recognizing complex patterns.” U.S. Patent 3,069,654, issued December 18, 1962
8. *Shan Caifeng, Shaogang Gong, and Peter W. McOwan.* ”Facial expression recognition based on local binary patterns: A comprehensive study.” Image and vision Computing 27.6: 803-816, 2009.
9. *M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel and J. Movellan.* ”Recognizing facial expression: machine learning and application to spontaneous behavior.” In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, pp. 568-573. IEEE, 2005.
10. *Karen Simonyan and Andrew Zisserman.* Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

11. *Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens and Zbigniew Wojna*. "Rethinking the inception architecture for computer vision". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016.
12. *François Chollet.* Xception: Deep learning with depthwise separable convolutions. CoRR, abs/1610.02357, 2016.
13. *Andrew G. Howard et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.
14. *Yichuan Tang*. Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239, 2013.
15. *Ian J. Goodfellow et al.* Challenges in Representation Learning: A report on three machine learning contests, 2013.
16. *Diederik Kingma and Jimmy Ba.* Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
17. *P.-L. Carrier and A. Courville.* Facial expression recognition dataset. In ICML 2013 Challenges in Representation Learning, 2013.
18. *Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun.* Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
19. *Shervin Minaee, Amirali Abdolrashidi.* Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. arXiv preprint arXiv:1902.01019, 2019.

РОЗПІЗНАВАННЯ ЕМОЦІЙ ІЗ ВИКОРИСТАННЯМ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

О. Каськун, Р. Шувар, А. Продивус

Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005, м. Львів, Україна
oleh.kaskun@lnu.edu.ua , roman.shuvar@lnu.edu.ua ,
andriy.prodyvus@lnu.edu.ua

Розпізнавання емоцій - це частина розпізнавання обличчя, яка набуває все більшого значення, і потреба в ній значно збільшується. Завдання включає знаходження та розпізнавання облич, відповідне представлення даних, схему класифікації, базу даних для проведення навчання.

На відміну від існуючих методів ідентифікації виразів за допомогою машинного навчання та методів штучного інтелекту, у цій роботі використовується метод глибокого навчання та класифікації зображень, щоб розпізнати та класифікувати вирази обличчя на зображенні.

У цій роботі ми пропонуємо впровадити загальну структуру побудови нейронної мережі (CNN) для проектування CNN, що працюватиме у режимі реального часу. Ми тестуємо моделі, створюючи систему бачення в реальному часі, яка виконує завдання виявлення обличчя та класифікації емоцій за один крок, використовуючи запропоновану архітектуру CNN. Розроблена архітектура була систематично побудована, задля того, щоб зменшити кількість параметрів, для цього використана технологія глибинного розподілення згортки. Дана архітектура зменшила кількість параметрів у 80 разів в порівнянні із класичною архітектурою, але при цьому відбулися втрати для емоції

«відраза». Для навчання нейронної мережі використано базу даних FER2013, яка розповсюджується на платформі Kaggle.

Необхідно зауважити, що нейронна мережа проявила здатність активізуватися, розглядаючи такі особливості, як наявність зубів на зображені, положення брів і очей. Цей результат запевняє, що мережа навчилася розпізнавати зрозумілі людиноподібні особливості, які забезпечують визначення емоції.

В результаті тестування стали зрозумілі кілька поширених помилок, наприклад особи в окулярах класифікуються як «сердиті». Це відбувається через те що темні рамки окуляр, вказують мережі що людина хмуриться, а дана особливість є одною з основних ознак емоції «злість». Успішність визначення емоцій рівна 66%, що є доволі високим показником для задач даного типу.

Ключові слова: згорткова нейронна мережа, знаходження обличчя, розпізнавання обличчя, TensorFlow, глибоке навчання

Стаття: надійшла до редакції 12.05.2020,
доопрацьована 15.05.2020,
прийнята до друку 18.05.2020