Emotion recognition based on deep learning with auto-encoder

Cite as: AIP Conference Proceedings **2217**, 030013 (2020); https://doi.org/10.1063/5.0000679
Published Online: 14 April 2020

I. Made Nomo Wiranata, Pranowo, and Albertus Joko Santoso

Emotion Recognition Based on Deep Learning with Auto-encoder

I Made Nomo Wiranata [1,b)], Pranowo [1,a)] and Albertus Joko Santoso [1, c)]

[1]*Magister Teknik Informatika, Universitas Atma Jaya Yogyakarta, Indonesia*

[a)] Corresponding author: pranowo@uajy.ac.id
[b)]made.wiranata23@gmail.com
[c)]albjoko@staff.uajy.ac.id

Abstract. Facial expression is one way of expressing emotions. Face emotion recognition is one of the important and major fields of research in the field of computer vision. Face emotion recognition is still one of the unique and challenging areas of research because it can be combined with various methods, one of which is deep learning. Deep learning is popular in the research area because it has the advantage of processing large amounts of data and automatically learning features on raw data, such as face emotion. Deep learning consists of several methods, one of which is the convolutional neural network method that will be used in this study. This study also uses the convolutional auto-encoder (CAE) method to explore the advantages that can arise compared to previous studies. CAE has advantages for image reconstruction and image de-noising, but we will explore CAE to do classification with CNN. Input data will be processed using CAE, then proceed with the classification process using CNN. Face emotion recognition model will use the Karolinska Directed Emotional Faces (KDEF) dataset of 4900 images divided into 2 groups, 80% for training and 20% for testing. The KDEF data consists of 7 emotional models with 5 angles from 70 different people. The test results showed an accuracy of 81.77%.

INTRODUCTION

Emotion is an internal condition originating from inside or outside which is indicated externally or the existence of behavior. Emotion is a natural and powerful way of communication between living things. Expressions of emotions can be seen by others even though expressed verbally and non-verbally. Verbal expressions for example, in the form of words by talking about emotions being felt. Non-verbal expressions are facial expressions, physical movements, pronunciation, body cues, and emotional actions. Facial expressions can be shown by various behaviors and facial movements. Facial expressions are one of the most commonly used ways to show how you feel. Facial expressions can represent a person's emotions that can be expressed by facial movements, such as opening lips and raised eyebrows. These facial behaviors are all shaped by facial muscle movements. Facial expressions are very important in communication or interpersonal interactions because it is one of the typical indicators of emotions [1] [2] [3].

In recent years, there has been increasing growth in interest in developing technology to recognize individual emotional states [4]. Face emotion recognition is one of the important and main research fields and one of the most severe in the computer vision field. Facial emotion recognition is still one of the unique and challenging fields of research among the computer vision research community [5] [4] [6]. Through recognition of facial expressions, it is possible to improve, and better understand human interactions, actions, and even feelings [7].

In recent years there have been many advances in artificial neural networks, one of which is convolutional neural network (CNN). CNN learned the representation of new data for the first time using the convolution layer as feature extractors and the fully-connected layer for classification. The process of learning new representations on CNN can be optimized, one of which uses Convolutional Auto-Encoder (CAE). CAE is included in an unsupervised method that makes it possible to train the convolutional layer independently of the classification task to learn new data

The 5th International Conference on Industrial, Mechanical, Electrical, and Chemical Engineering 2019 (ICIMECE 2019)
AIP Conf. Proc. 2217, 030013-1–030013-5; https://doi.org/10.1063/5.0000679
Published by AIP Publishing. 978-0-7354-1971-1/$30.00

representations. The weight learned in this first step can then be used as a starting point for initializing the convolution layer of the neural network [8] [3].

LITERATURE REVIEW

There have been several previous studies in facial expression recognition, such as research conducted by [9] using the Hierarchical Bayesian Theme Models method to recognize expressions from various angles for 7 types of expressions, namely afraid, , surprise angry, sad, disgust, happy and neutral, in this research gets 90.24% precision. In a study conducted by [10] using Facial Movement Features by extracting patch-based 3D Gabor features, selecting 'prominent' patches, and performing patch matching operations, which in this study obtained 92.93% accuracy for the Japanese Female Facial Dataset Expression (JAFFE) and 94.48% for the Cohn-Kanade (CK) dataset.

Auto-encoder is one method that is widely used in research, such as research conducted by [11] to do face recognition using the sparse auto-encoder method by using 3 identification algorithms, namely softmax algorithm, depth-learning top algorithm, and overall fine-tuning algorithm. The data used are ORL, YALE, YALE B, FERET face dataset and get different results for each dataset type. In this research, identification with softmax algorithm gets the best results on ORL dataset with an accuracy of 89.38%, whereas identification with depth-learning top algorithm gets the best results on YALE face dataset with an accuracy of 89.33% and identification with overall fine-tuning algorithm gets the best results on the FERET face dataset with an accuracy of 94.33%.

The use of CNN in various research fields has been widely carried out. Some of these include research conducted by [12] CNN used to identify facial recognition on low-resolution images, in that study CNN was successfully applied and from each of the network architectures used had a 96.02% precision for the CMU PIE dataset and 94, 63% for Extended Yale B dataset. In a study conducted by [13], CNN was used to identify Thermal Face Recognition, using the RGB-D-T Face dataset with a resolution of 640x480 and a resolution of 384x288 for RGB images with a total of 15300 images. In that study, CNN was successfully implemented and from each network architecture used it got better results from traditional recognition such as LBP, HOG and invariant moments with an accuracy of 99.40%. There are also studies of CNN combined with other methods such as the research conducted by [14] to recognize facial patterns using the CNN method and Support Vector Machine (SVM), using a FERET dataset consisting of 1400 images with different light and posture conditions. In that study, CNN and SVM were successfully implemented and produced an accuracy of 97.50%.

METHODOLOGY

In this research, we are introducing facial emotions recognition using CAE as pre-training on CNN. To achieve this, in the first step we train CAE on the selected dataset and then, in the second step, we use the weights of the convolution layer to initialize the convolution layer on CNN. After the training, we conducted a test knowing the accuracy obtained. Visualization of weight transfer can be seen in the **FIGURE** 1.

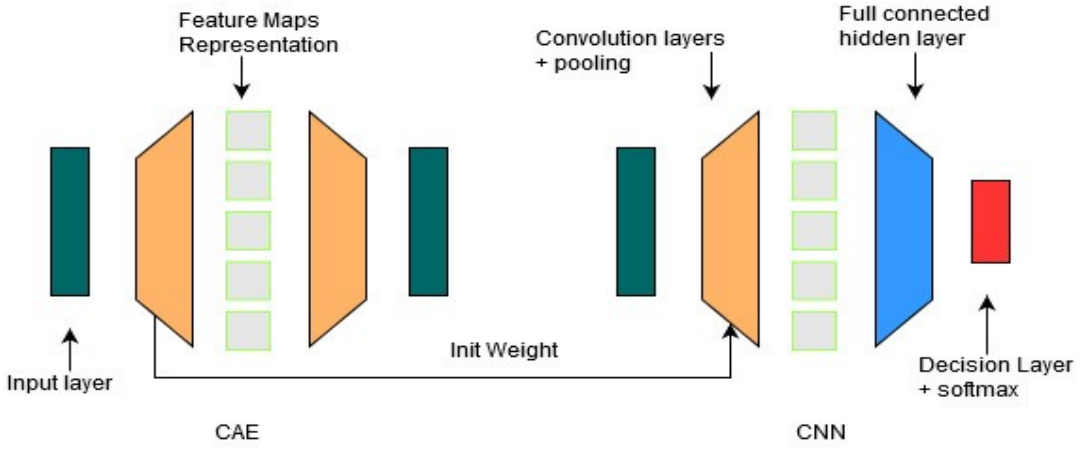

FIGURE 1. Experiment Setup

Autoencoder

CAE first uses several convolution and fusion layers to extract features from inputs into high dimensional feature map representations and then reconstruct inputs using strided transposed convolution [8]. The auto-encoder architecture used in our experiment is described in [15]. The encoding section consists of 3 times the convolution layer followed by 2x2 max-pooling and the ReLU activation layer. More detailed layer information is given in table 1.

TABLE 1. Layer Information

Layer Type	Filter Size	Channels	Activation
Input	None	1(gray)	None
Convolution	3x3	32	ReLU
Max-Pooling	2x2	32	None
Convolution	3x3	64	ReLU
Max-Pooling	2x2	64	None
Convolution	3x3	128	ReLU

For reconstruction, we use one stride transposed convolution layer per convolution layer and Up-sampling for each max-pool layer in the encoding layers. Up-sampling is used to reverse the Down-sampling effect of the max-pooling layers. The weights of the transformed convolution results are the same as those learned in the convolution layer.

Convolutional Neural Network

The architecture used on CNN is the same as the architecture in CAE encoding for feature extraction (convolution and max-pooling) layers and then uses three fully-connected layer layers (sizes 1024, 512 and 7) for decision making (classification), first and second has ReLU activation and the third uses Softmax activation. For the dataset, we divide the available data into training and test sets with a ratio of 80:20 to the total training data of 3920 images and test data of 560. In the test process an evaluation of network accuracy is carried out with the results obtained in the form of accuracy.

The Karolinska Directed Emotional Faces (KDEF) Dataset

The KDEF dataset [16] is an image of human facial expressions consisting of 4900 images, made at the Karolinska Institutet, Stockholm. The images used in this set come from 70 individuals (consisting of 35 women and 35 men) who display seven different types of expressions (fear, anger, disgust, happiness, neutrality, sadness and surprise). Each expression is viewed from five different angles (full left profile, full right profile, half left profile, right half profile and straight) and photographed in two sessions. The KDEF dataset is color photographs. The model used is Caucasian amateur actors (average age, 25 years; range, 20-30 years). They were ordered to pose with a variety of different expressions.

RESULT AND DISCUSSION

Of the total dataset divided into two, 80% is used for the training process and 20% is used for the testing process. In the training process using a dataset of 3920 images with 1500 epochs. The plot loss of the training process can be seen in FIGURE 2. For the plot of training process accuracy can be seen in FIGURE 3.

In FIGURE 2 shows a graph of loss in the training process, in which the value of loss has decreased significantly from the first epoch to the 500th epoch. After the 500th epoch the loss value has decreased slowly but is not yet stable because there is still an increase in the loss value even though it is not too significant. At the 1500th epoch the loss value was 0.01289 while the lowest loss was obtained at the 1411 epoch with a loss value of 0.0047456.

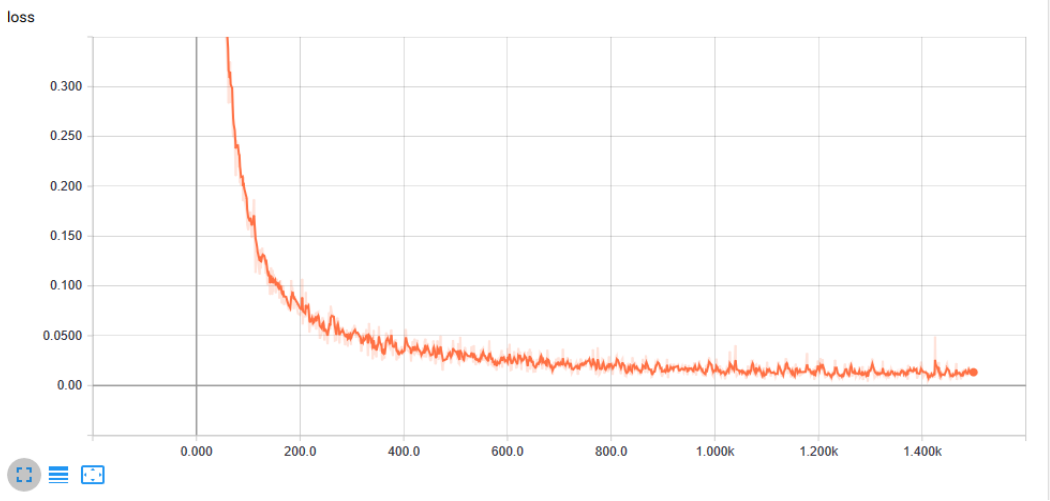

FIGURE 2. Plot Loss In Training Process

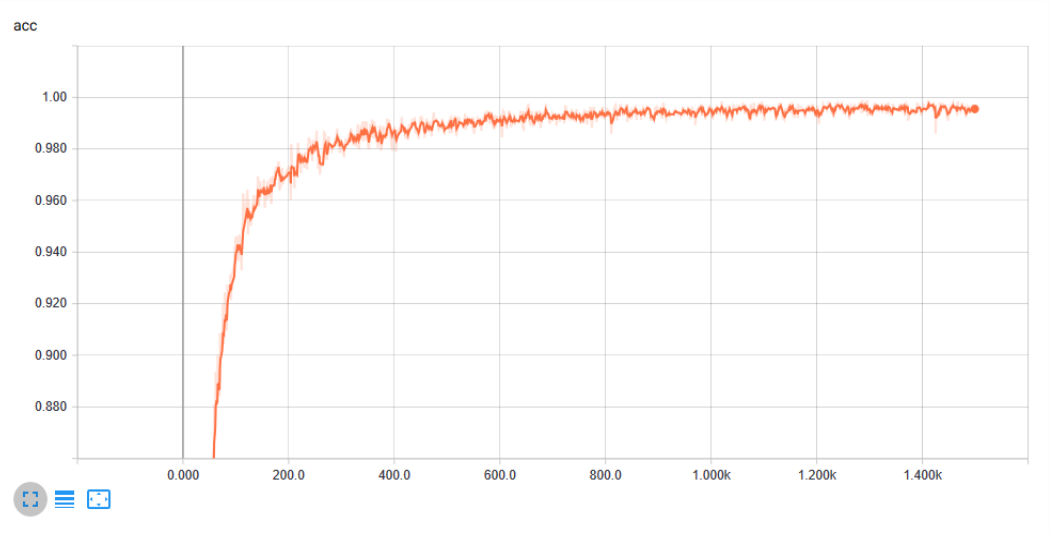

FIGURE 3. Plot Accuracy In Training Process

In FIGURE 3 shows the accuracy graph in the training process, where the value of training accuracy has increased significantly from the first epoch to the 346th epoch. After the 346th epoch the training accuracy value has increased slowly but has not been stable because there is still a decrease in the value of training accuracy even though it is not too significant. At the 1500th epoch the accuracy value was 99.55% while the highest accuracy value was obtained at the 1411 epoch with an accuracy value of 99.90%. In the testing process using a dataset of 560 images with 1500 epochs. The accuracy value in the testing process is far below that of the training process which is 81.77%. while the training process can produce an accuracy of 99.55%.

CONCLUSION

We present research on introducing emotions using convolutional auto-encoder as a pre-train on CNN on the KDEF dataset. CNN training uses weights obtained from convolutional auto-encoder in the pre-train process. The results show the training process gets an accuracy of 99.55% whereas In the process of testing the system can learn to recognize facial emotions with an accuracy of 81.77%.

REFERENCES

1. J. Zhao, X. Mao and J. Zhang, Learning deep facial expression features from image and optical flow sequences using 3D CNN (Springer, Heidelberg, 2018), pp. 1461–1475.
2. B. Hasani and M. H. Mahoor, "Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields,"in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (IEEE, Washington, DC, 2017), pp. 790 – 795.
3. K. Li, Y. Jin, M. W. Akram, R. Han and J. Chen, Facial Expression Recognition With Convolutional Neural Networks Via A New Face Cropping And Rotation Strategy (Springer Berlin Heidelberg, Heidelberg, 2019), pp. 1-14.
4. H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (IEEE, New York, 2016), pp. 1-9 .
5. D. Das and A. Chakrabarty†, "Emotion Recognition from Face Dataset Using Deep Neural Nets," in 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)(IEEE, Sinaia, Romania, 2016), pp. 1-6.
6. J. Siswantoro, Application of Color and Size Measurement in Food Products Inspection, (Indonesian Journal of Information Systems (IJIS), Yogyakarta , 2019), pp. 90 -107.
7. D. Hamester, P. Barros and S. Wermter, "Face expression recognition with a 2-channel Convolutional Neural Network," 2015 International Joint Conference on Neural Networks (IJCNN)(IEEE, Killarney, Ireland, 2015), pp. 1-8.
8. M. Kohlbrenner, R. Hofmann, S. Ahmmed and Y. Kashef, "Pre-Training CNNs Using Convolutional Autoencoders," 2017.
9. Q. Mao, Q. Rao, Y. Yu and M. Dong, Hierarchical Bayesian Theme Models for Multipose Facial Expression Recognition, (IEEE, Piscataway, 2017), pp. 861 - 873.
10. L. Zhang and D. W. Tjondronegoro, Facial Expression Recognition Using Facial Movement Features, (IEEE, Piscataway, 2011), pp. 219 - 229.
11. Z. Zhang, J. Li and R. Zhu, "Deep Neural Network for Face Recognition Based on Sparse Autoencoder," in 2015 8th International Congress on Image and Signal Processing (CISP)(IEEE, Shenyang, China, 2015), pp. 594 - 598.
12. C. Ding, T. Bao, S. Karmoshi and M. Zhu, "Low-resolution face recognition via convolutional neural network," in 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)(IEEE, Guangzhou, China, 2017), pp. 1157 - 1161.
13. Z. Wu, M. Peng and T. Chen, "Thermal Face Recognition Using Convolutional Neural Network," in 2016 International Conference on Optoelectronics and Image Processing (ICOIP) (IEEE, Warsaw, Poland, 2016) , pp. 6 - 9.
14. S. Guo, S. Chen and Y. Li, "Face Recognition Based On Convolutional Neural Network And Support Vector Machine," in 2016 IEEE International Conference on Information and Automation (ICIA) (IEEE, Ningbo, China, 2016) , pp. 1787 - 1792.
15. J. Masci, U. Meier, D. Ciresan and J. Schmidhuber, "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction," in ICANN 2011: 21st International Conference on Artificial Neural Networks (Springer, Espoo, Finland, 2011) , pp. 52-59.
16. D. Lundqvist, A. Flykt and A. Öhman, "The Karolinska directed emotional faces - KDEF," in CD ROM from Department of Clinical Neuroscience, Psychology section (Karolinska Institutet, Stockholm, 1998) , pp. 1157 - 1161.