2021 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA) | 978-1-6654-9969-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAMIMIA54022.2021.9809908

Deep Learning Based Facial Emotion Recognition using Multiple Layers Model

Lidia Sandra
Psychology Department
Krida Wacana Christian University
Jakarta, Indonesia

Yaya Heryadi
Computer Science Department,
BINUS Graduate Program-Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480

Lukas
Cognitive Engineering Research Group (CERG),
Faculty of Engineering
Universitas Katolik Indonesia Atma Jaya
Jakarta, Indonesia 12930

Wayan Suparta
Computer Science Department, BINUS Graduate Program-Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480

Antoni Wibowo
Computer Science Department, BINUS Graduate Program-Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480

***Abstract*— A Facial Emotion Recognition (FER) system is an important tool to be implemented in any psychology academic field and beyond. This paper aims to show a system of Facial Emotion Recognition that can be done using the multiple layers model of ResNet50 which is an ANN. The dataset is an array that consists of 7 different emotions represented by the numbers 0-6. The emotion included in the dataset is "anger", "disgust", "fear", "happiness", "sadness", "shock", and "neutrality". This research found that the highest accuracy in the application of the model was recorded at 65% and the test was 60% to recognize facial emotion from a graphical input**.

Keywords*— *Facial Emotion Recognition; Deep Learning; Machine Learning.

I. INTRODUCTION

Humans have multiple ways of communicating, one of the means to send information as a human being is through emotion. Albert Mehrabian famously emerge with a rule that emotion consists of 55% visual information, 38% vocal information, and 7% verbal information [1]. Verbal information is the content of the speech itself, while the vocal is the tone and volume of the speech, visuals comprise of body language and facial emotional expressions [2]. Building a model with deep learning to recognise facial emotion will be beneficial to understand better how human convey information through emotion and cross platform studies of Facial Emotion Recognition (FER) will be thrived.

Facial emotion can be found not only in humans but in also mammals and other species of animals As a human themself, they can show a facial emotion voluntarily or even involuntarily. Voluntarily means that the emotion shown is an outcome of a cognitive decision that the human made, while involuntarily is an action decided by the human from certain inputs. Understanding facial emotion as a feature of communication and information gathering is what we call facial recognition. Creating a system that can do facial emotion recognition is the aim of this paper.

Deep learning itself has a different way of implementation, one of them being supervised, unsupervised, and the other one being semi-supervised [5]. As a part of a bigger dome of machine learning methods which is based on artificial neural networks (ANN), deep learning which is also called deep structured learning works with representation learning [6]. Since deep learning uses layers to have an outcome of a feature that is higher-level from the original input such as the works in image processing, it is well recommended in tackling the task of facial emotion recognition [7].

In the context of the present, a facial recognition system is widely used all around the world due to the nature

978-1-6654-9969-9/21/$31.00 ©2021 IEEE

of it is contactless process [8]. In the practice of facial recognition, the system has been used in many fields that requires the interaction of human and computer [9].

II. METHODS

The method of the research is to use a multi layered deep learning techniques to be able to create a system of facial emotion recognition. In order to do so, a dataset needs to be taken. The dataset itself consist of pictures that later the value will be taken from the pixels of the picture. The category itself is taken into the first column of the tabular which later connected to the dataset. These categories are valued by numbers starting from zero to six. Then, the data pre-processing steps will begin. These steps consist of turning the raw data into pixel value of one-dimensional array. After that, the data will be changed into a two-dimensional array and scaling the value of each pixel by dividing them with RGB value 255. Then the augmentation of the data such as, rotation, zoom, shear, flip, and shift are happening. After all of the pre-processing process works, the data is ready to be taken into the system to train the model, and results will follow.

A. Dataset

The dataset used is taken from Kaggle in the form of tabular which has 35,887 rows and 3 columns with data files that have the extension of ".csv". The first column is filled with *emotion* columns represented by the numbers 0 - 6. Each represents categories of anger, disgust, fear, happiness, sadness, shock, and neutrality. The second column contains *pixels* from the image that will be processed first which was previously a 1-D Array into a 2-D Array. The last column contains *usage* of each of these data

An array is a structure of a data that consist of group of values or variables [10]. Similar to list, the data are represented by one or more array index which can be called a key [11]. Since the simplest type of data is the linear array which we called one-dimensional array [12], we use this as the base of the dataset and turned into a multi-dimensional array for the next step.

TABLE 1. DATASET USED

	emotion	pixels	Usage
0	0	70 80 82 72 58 58 60 63 54 58 60 48 89 115 121...	Training
1	0	151 150 147 155 148 133 111 140 170 174 182 15...	Training
2	2	231 212 156 164 174 138 161 173 182 200 106 38...	Training
3	4	24 32 36 30 32 23 19 20 30 41 21 22 32 34 21 1...	Training
4	6	4 0 0 0 0 0 0 0 0 0 0 0 0 3 15 23 28 48 50 58 84...	Training
...	...	...	...
35882	6	50 36 17 22 23 29 33 39 34 37 37 37 39 43 48 5...	PrivateTest
35883	3	178 174 172 173 181 188 191 194 196 199 200 20...	PrivateTest
35884	0	17 17 16 23 28 22 19 17 25 26 20 24 31 19 27 9...	PrivateTest
35885	3	30 28 28 29 31 30 42 68 79 81 77 67 67 71 63 6...	PrivateTest
35886	2	19 13 14 12 13 16 21 33 50 57 71 84 97 108 122...	PrivateTest

The most emotions found in the images in the dataset were happy categories. The category of happy emotions is the category of emotions that are most easily recognized by models, while the category of emotions disgust is the category of emotions that are the most difficult for models to recognize. This calculation is calculated Figure 1.

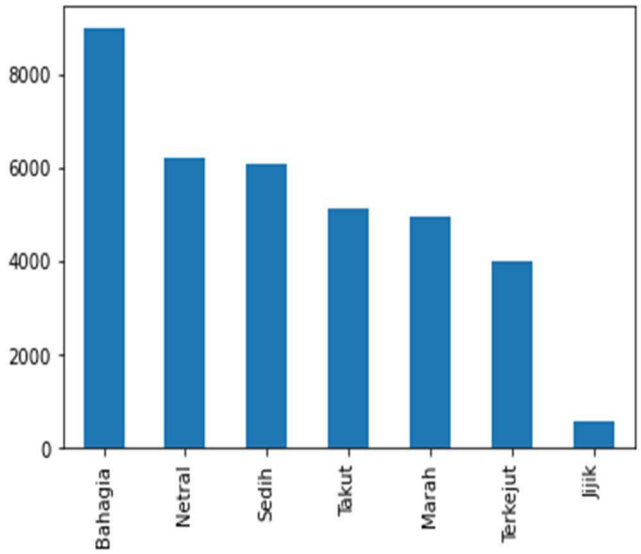

Fig. 1. Model Analysis results based on the Number of Recognized Categories.

B. Data Preprocessing

The preprocessing data flow begins with raw data that is still in the form of 1-D arrays changed using *reshape* to a size of 48x48x1 with 1 channel / *gray*. Then the gray image was converted back to RGB so that the image size changed also to 48x48x3. The next process in preprocessing data is scaling the value of each pixel in the image by dividing its pixel value by 255 so that each pixel value will be worth 0-1. Next is the process of data augmentation in the form of

rotation, zoom, shear, flip, and shift carried out on the image. Then the label format was changed which was previously a 1-D array with 6 emotion columns to a 2-D array with 7 columns so that the model recognized it as multilabel. This label format change is done using hot encoding. The next step is to separate the training data and test data with the amount of test data 30% of the total amount of data. The preprocessing flowchart of the data can be seen on Figure 2.

Fig.2. Preprocessing Data Flow.

Layer (type)	Output Shape	Param #
resnet50 (Functional)	(None, 2, 2, 2048)	23587712
flatten_14 (Flatten)	(None, 8192)	0
dense_25 (Dense)	(None, 2048)	16779264
dense_26 (Dense)	(None, 1024)	2098176
dense_27 (Dense)	(None, 512)	524800
dense_28 (Dense)	(None, 256)	131328
dense_29 (Dense)	(None, 128)	32896
dense_30 (Dense)	(None, 64)	8256
dense_31 (Dense)	(None, 7)	455

Total params: 43,162,887
Trainable params: 43,109,767
Non-trainable params: 53,120

Fig. 1. Dataset with the model used

The model used is a deep learning model with a number of layers applied in the dataset. The first layer applied is to detect facial emotions using the *ResNet50* which is an artificial neural network (ANN) that works by building construct [13]. It is used in this model as it is implemented with multiple layers and normalization of batch in between [14]. The pre-trained models that is already available by hard TensorFlow and has an additional layer of 1 layer flatten and 7 layer dense is also used. ResNet50 pre-trained model is a deep convolutional neural network model with 48 LAYER CNN and 2-layer Carpooling. This model works by reading the input shape (N_rows, 48, 48, 3) and producing the output with the shape (N_rows,7). The dense layer in this model is as hidden layer in the set using the *ReLU* Which is a rectifier activation function [14], while the output layer (the last layer uses *the SoftMax* activation function because it will be accommodated classification cases.

After doing the definition of the model layer, it will then be compiled with a loss function – the one used in training this model is binary cross entropy even though the label has more than two values. Then the optimization process is done using *Adam* with a learning rate of 0.0001 and the metric used as an evaluation material is the level of accuracy. *Epoch* in the model is set with a value of 50 and applied 3 callbacks so that the model can stop doing the training process. These 3 callbacks are *Early Stopping*, *Model Checkpoint*, and *ReduceLRonPlateu*.

The model shown on Figure 3 can be explained as follows. We are working with deep learning methods in this model, resulting in layers of models being used. The first one is a functional pre-trained ResNet50 with an addition layer of a flatten_14 and seven different dense layers. This pre-trained ResNet50 is a deep convolutional neural network with 48 layers CNN and 2 layers of MaxPooling.

III. EVALUATION AND RESULTS

The training results stopped at epoch 15 due to a predetermined set of callbacks. In epoch 8, it can be seen that accuracy and loss training and validation have the same value, which means the results of model analysis are very fit. Nonetheless, in epoch 15 has a tendency to overfitting but can still be tolerated with a difference in loss and accuracy of about 0.05. Shown in the Figure 4, the highest accuracy in the application of the model was recorded at 65% and the test was 60%.

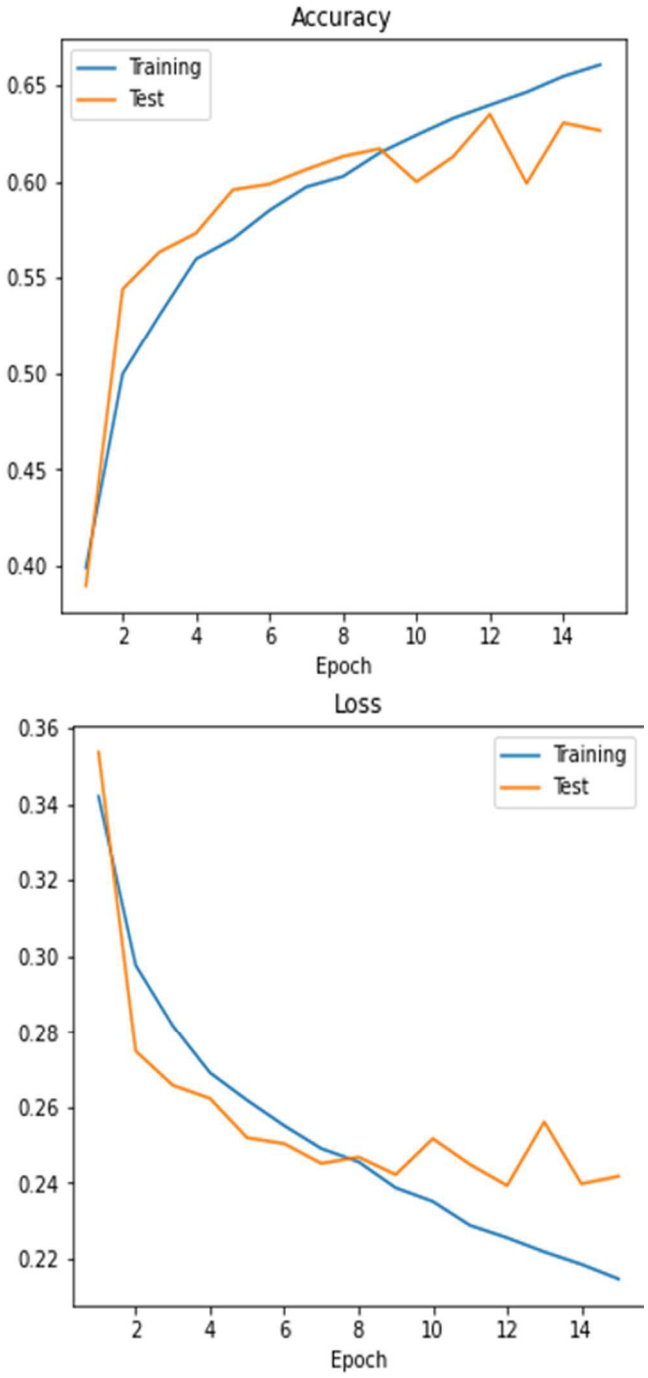

Fig. 2. Model Evaluation Results (Accuracy and Lost

When viewed from the classification report on Figure 5, the accuracy value is 60% which means there is a possibility of influence from the category "disgust" or label 1 whose amount of data is up to 10 times less than other categories. It is necessary to do extra step such as handling imbalance for the category "disgust" so that the effect on accuracy can be eliminated.

	precision	recall	f1-score	support
0	0.67	0.55	0.60	4953
1	0.88	0.37	0.52	547
2	0.65	0.37	0.47	5121
3	0.85	0.88	0.87	8989
4	0.68	0.42	0.52	6077
5	0.77	0.77	0.77	4002
6	0.64	0.65	0.64	6198
micro avg	0.73	0.62	0.67	35887
macro avg	0.73	0.57	0.63	35887
weighted avg	0.72	0.62	0.66	35887
samples avg	0.62	0.62	0.62	35887

Fig.5. Classification Report.

Fig.6. Example of model evaluation result

Fig. 3. Example of model evaluation result

Shown both on Figure 6 and Figure 7 is the performance of the model proposed. The square box on the faces acts as a barrier on where the model works and each of the facial emotion category are directly written above the box.

IV. CONCLUSION AND FURTHER RESEARCH

This research shows that a facial emotion recognition can be done using deep learning methods with multiple layered models used on the dataset. The accuracy value is also considerably high with the model proposed in this paper. There are flaws on the research that we can found, such as the imbalance of the data will results in the decreasing of the overall accuracy. The “disgust” category which does not have as much data as the other category resulting in the fl-score of its category low. Which also resulting in the overall category to be low as well.

With this research we can also see that this kind of problem can be solved by ignoring the imbalance data itself not to be processed by the system, or to fix the amount of the data for that exact category to be able to have the same amount as the other category. From the research as well, we can see that the training results stopped at epoch 15 due to a predetermined set of callbacks. In epoch 8, it can be seen that accuracy and loss training and validation have the same value, which means the results of model analysis are very fit. Nonetheless, in epoch 15 has a tendency to overfitting but can still be tolerated with a difference in loss and accuracy of about 0.05. Shown in the Figure 4, the highest accuracy in the application of the model was recorded at 65% and the test was 60%.

This paper could give a base of what our future research will be. More research that will look into the specific topics can be done with the model of deep learning proposed in this paper. Academics are already looking into these matters with an aim to review facial emotion recognition based on visual information on this paper [15]. Recently we have also one that is specific for interaction between a human and a computer through an application [16]. Another paper is looking into deep learning neural networks in order to have a system of not only emotion recognition but also face recognition [17]. Along with these papers, this paper aims to clearly show the potential of deep learning in handling datasets of a graphical input to be able to create a system that is not only worth looking into as research but also beneficial in other fields[18].

[1] Mehrabian, A. (1971). Silent messages (Vol. 8, No. 152, p. 30). Belmont, CA: Wadsworth.

[2] Mehrabian, A. (2017). Nonverbal communication. Routledge.

[3] Alan J. Fridlund (1994). Human facial expression (1 ed.). San Diego: Academic Press. ISBN 978-0-12-267630-7.

[4] J.A. Russell; J.M. Fernandez Dols (1997). The psychology of facial expression (1 ed.). Cambridge University Press. ISBN 978-0-521-58796-9.

[5] Bengio, Yoshua; LeCun, Yann; Hinton, Geoffrey (2015). "Deep Learning". Nature. 521 (7553): 436–444. Bibcode:2015Natur.521..436L. doi:10.1038/nature14539. PMID 26017442. S2CID 3074096.

[6] Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85–117. arXiv:1404.7828. doi:10.1016/j.neunet.2014.09.003. PMID 25462637. S2CID 11715509.

[7] Deng, L.; Yu, D. (2014). "Deep Learning: Methods and Applications" (PDF). Foundations and Trends in Signal Processing. 7 (3–4): 1–199. doi:10.1561/2000000039.

[8] Chen, S.K; Chang, Y.H (2014). 2014 International Conference on Artificial Intelligence and Software Engineering (AISE2014). DEStech Publications, Inc. p. 21. ISBN 9781605951508.

[9] Bramer, Max (2006). Artificial Intelligence in Theory and Practice: IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream, August 21–24, 2006, Santiago, Chile. Berlin: Springer Science+Business Media. p. 395. ISBN 9780387346540.

[10] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2015-12-10). "Deep Residual Learning for Image Recognition". arXiv:1512.03385

[11] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2016). "Deep Residual Learning for Image Recognition". Proc. Computer Vision and Pattern Recognition (CVPR), IEEE.

[12] Brownlee, Jason (8 January 2019). "A Gentle Introduction to the Rectified Linear Unit (ReLU)". Machine Learning Mastery.

[13] Black, Paul E. (13 November 2008). "array". Dictionary of Algorithms and Data Structures. National Institute of Standards and Technology.

[14] Bjoern Andres; Ullrich Koethe; Thorben Kroeger; Hamprecht (2010). "Runtime-Flexible Multi-dimensional Arrays and Views for C++98 and C++0x". arXiv:1008.2909 [cs.DS].

[15] Garcia, Ronald; Lumsdaine, Andrew (2005). "MultiArray: a C++ library for generic programming with arrays". Software: Practice and Experience. 35 (2): 159–188. doi:10.1002/spe.630. ISSN 0038-0644. S2CID 10890293.

[16] Ko, B. (2018). A brief review of facial emotion recognition based on visual information. Sensors (Basel, Switzerland), 18(2), 401.
[17] Chowdary, M., Nguyen, T., & Hemanth, D. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing & Applications, Neural computing & applications, 2021.
[18] Hussien Mary, A., Bilal Kadhim, Z., & Saad Sharqi, Z. (2020). Face Recognition and Emotion Recognition from Facial Expression Using Deep Learning Neural Network. IOP Conference Series. Materials Science and Engineering, 928(3), IOP conference series. Materials Science and Engineering, 2020-11-01, Vol.928 (3).