Journal of Physics:
Conference Series

PAPER • **OPEN ACCESS**

Real Time Emotion Classification Based on Convolution Neural Network and Facial Feature

To cite this article: A Nasuha *et al* 2021 *J. Phys.: Conf. Ser.* **1737** 012008

Real Time Emotion Classification Based on Convolution Neural Network and Facial Feature

A Nasuha[1], F Arifin[2], A S Priambodo[3], N Setiawan[4] and N Ahwan[5]

[1,2,3,4,5] [1]Electronics and Informatics Engineering Education Departement, Universitas Negeri Yogyakarta, Indonesia

Email: [1]arisnasuha@uny.ac.id

Abstract. Health problems due to emotion disorder should not be taken lightly because they have worse effects on health. Emotion disorder leads to prolonged stress and causes mental fatigue. Therefore, emotions need to be classified as early as possible. This classification result can be utilized to determine a person's emotion and treatments required. In this paper, we proposed emotion classifier based on facial features. Here, we used Convolution Neural Network (CNN) to extract facial features from input images and classify them into 7 basic emotions: angry, sad, happy, neutral, fear, disgust, and surprise. Dept-wise separable convolution is applied, instead of the ordinary convolution in CNN, to reduce the number of trainable parameters so that the overall architecture of CNN can be made as simple as possible without compromising the accuracy. The simple architecture of the CNN allows us to make it work in real time. Our proposed method achieves an accuracy of 66% on 3.589 input images of FER2013 data set.

1. Introduction

Emotion is a form of a person's response to environmental conditions. This is an important factor that needs to be considered in interacting with the others. Someone emotions cannot be known easily due to its nature as an unvoiced state of person [1]. However, along with the development of artificial intelligence and image processing,we can read a person's emotion through a facial instrument.

Research in Facial Emotion Recognition (FER) has been developed since 1970. Research was initiated by Paul Ekman [2] who developed the Facial Action Coding System (FACS) to classify 7 basic human emotions: anger, sadness, happiness, neutral, fear, disgust, and surprise. Later, Research [3] started to apply the machine learning algorithm in FER system to improve its accuracy. This system uses an ideal images such as ideal colors, ideal contrast, and ideal backgrounds. So this system has a drawback for the face images taken naturally [4].

Many applications such as security systems, interactive Human-Computer Interaction, job interview, business and marketing purposes used FER to classify person's emotion. This output classification can be used to support a decision making. For example, person's emotion detected by FER can be used as consideration for accepting or rejecting candidates in a job interview.

In recent year, all method used to solve image-related tasks such as images detection and image classification [5] are developed based on Convolution Neural Network (CNN). Convolution Neural Network is a subfield of artificial intelligence that has outstanding performance in computer vision task. Unlike ordinary neural network method, the convolution layer in CNN allows us to extract the

feature of input image automatically. This layer eliminates the computation process of feature extraction in ordinary Neural Network [4].

In this paper we proposed real-time Facial Emotion Recognition (FER) based on Convolution Neural Network to classify 7 basic emotions : “angry”, “sad”, “happy”, “neutral”, “fear”, “disgust”, and “surprise”. Dept-wise separable convolution is applied in CNN to reduce the trainable parameter. The performances of proposed FER have been validated in a real-time Emotional level Detector which provides face detection and emotional classification.

2. Related Work

Researches of Facial Emotion Recognition (FER) are initialized by adopting facial action coding system (FACS) to design a framework of basic human emotions. FACS system provides a way for detecting a facial emotion based on the movement of face muscle that arranges the person’s emotion. Along with the development of artificial intelligent and computer vision, classification of facial emotion is carried out using the machine learning methods such as K-Nearest Neighbors (KNN) and Support vector Machine (SVM) [6]-[9].

As a reinforcement learning method, CNN grow rapidly in recent year. CNN has outstanding performance in computer vision tasks, including in facial-based emotion classification. Convolution layer in the CNN allow us to extract the image features automatically. So that only simple pre-processing is required before the features are fed to the fully-connected layer for training process [10]. The number of trainable parameters feeded to fully connected layer needs to be considered. The fewer parameter, the simpler neural network structure. For a FER system, the number of parameter is minimized by capturing the face area firstly and extracts its features later, so that the size of input image can be reduced [11].

Even though the feature extraction has been carried out, fully-connected layer is still considered to contain most of the parameter in CNN. For example, VGG [9] still uses 90% of all input parameters for their last fully connected layer. Modern architectures such as Inception V3 [12] reduce the number of trainable parameters by implementing a global average pooling operation in the last layer instead the fully-connected layer. Global average pooling convert each feature map into a scalar value by taking average over all features map component. Recent architecture, Xception [5], use Depth-wise separable convolutions to reduce further the amount of trainable parameters by separating the feature extraction process.

3. Theory

3.1. Convolutional Neural Network

Convolutional Neural Network or commonly abbreviated as CNN is a machine learning algorithm that has reliable performance in processing two-dimensional data such as image. CNN can be used to detect and recognize certain objects based on the image input. CNN is developed from artificial neural network and it is a kind of supervised learning method. Generally, CNN consist of three type of layer as shown in figure 1. First layer is convolution layer, utilized to extract the feature from input images [13]. Some filters (kernels) are convolved with the input image to extract its features. The second layer is pooling layer. Pooling layer is used to reduce the dimension of the input image. There are 2 types of pooling layer: maximum pooling and average pooling. And the last layer is fully-connected layer. This layer is responsible for classification at the end of the network architecture.

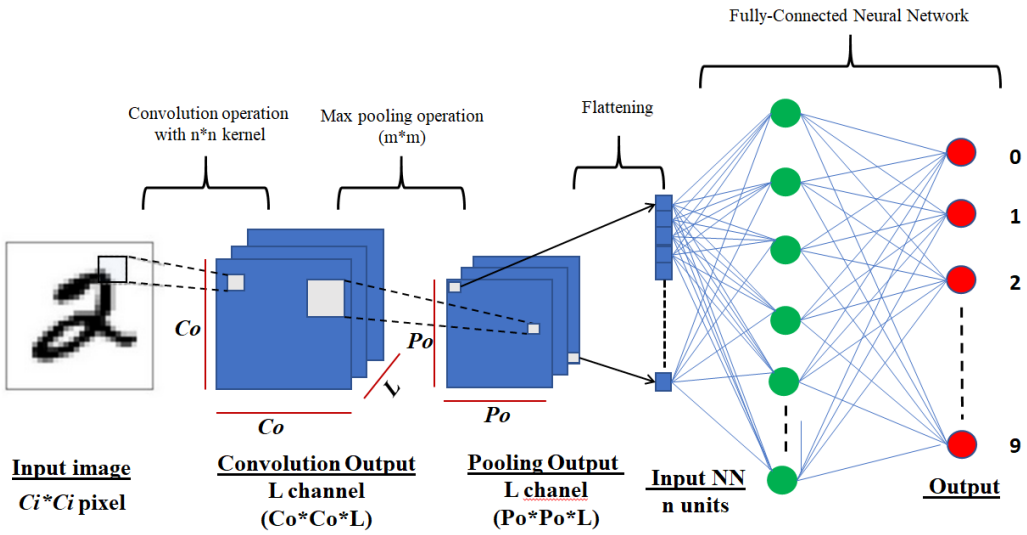

Figure 1. Architecture of CNN

3.2. Depth-wise Separable Convolutions

Ordinary CNN convolves a number of filters across the channels input as shown in figure 2. For simple networks architectures, ordinary CNN will work properly. But, for deep networks, convolution generates a large number of trainable parameter and is computationally costly. For example, an image consist of L channel input and $Ci \times Ci$ pixel size for each channel. Each input channel is convoluted to **P** kernels of size n × n. The number of multiplications for this convolution process are :

$$Convolution = (n \times n) \times (Co \times Co) \times L \times P$$

where $Co = Ci - n$, is the dimension of the output channel.

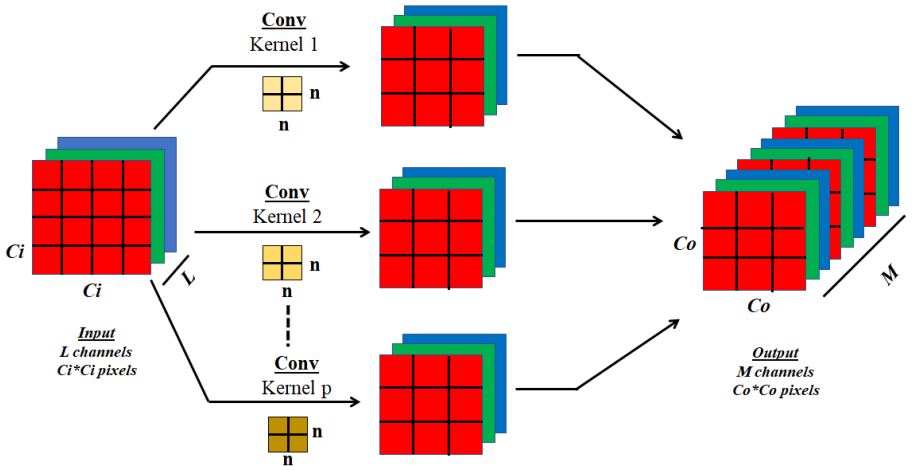

Figure 2. Ordinary Convolutions

Depth-wise Separable convolutions [14], offer a method to reduce the number of computation and parameter in convolution process. The main idea of a this method is separate the spatial cross-correlation from the channel cross-correlation [5]. Depth-wise Separable convolutions consist of 2 convolution step: depth-wise convolutions and point-wise convolution as shown in Figure 3. In the

first step, every L input channels is convoluted to single kernel n×n size separately compared to all channels in ordinary convolution. The number of multiplication for this convolution process is

$$Conv1 = (Co \times Co) \times (n \times n) \times L$$

Meanwhile for the second step, K filters $1 \times 1 \times L$ size are convoluted to all channels in the output of previous step. This convolution process requires the number of multiplications as follows

$$Conv2 = (Co \times Co) \times L \times P$$

The number of multiplications required for overall convolutions are the sum of the multiplications in the depth-wise convolutions and point-wise convolution. The number of multiplication and trainable parameter is reduced by $\frac{1}{P} + \frac{1}{n^2}$ compared to ordinary convolution [15].

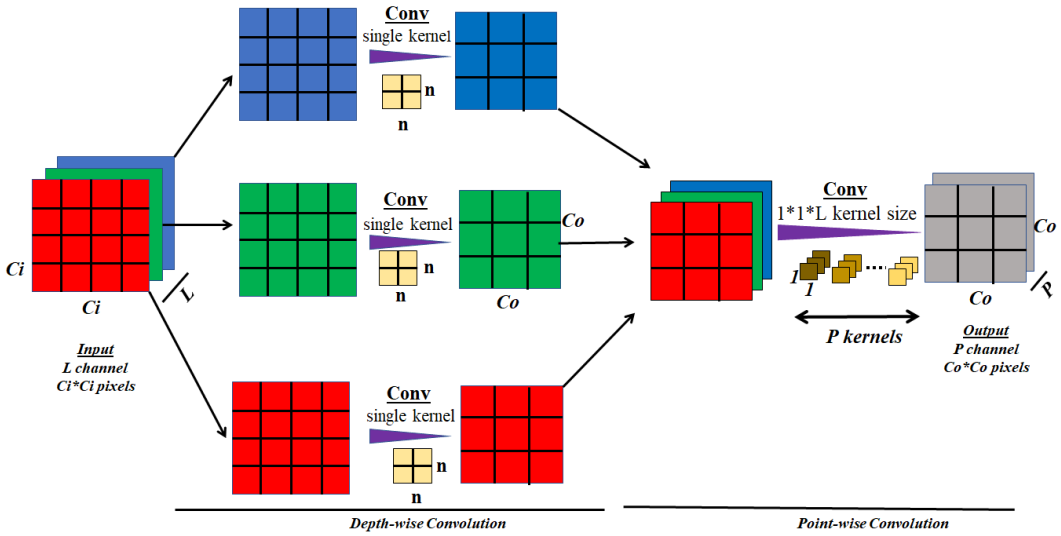

Figure 3. Depth-wise separable convolution

4. Proposed Model

We proposed real time Facial Emotional Detection based on convolution neural network with depth-wise separable convolution layer. The proposed model contains 4 depth-wise separable convolution layer where each layer is followed by Batch normalization operation and a Rectifier Linear Unit (ReLU) activation function. Since we use depth-wise separable convolution layer, the trainable parameter can be reduced significantly compared to ordinary CNN. In this proposed model, the number of parameter is reduced from 600.000 parameters to 60.000 parameter. Softmax activation function and global average pooling in the last layer are utilized to generate a prediction. This architecture is inspired by mini Xception model in [16] and [5]. The final architecture of proposed model is shown in figure 4.

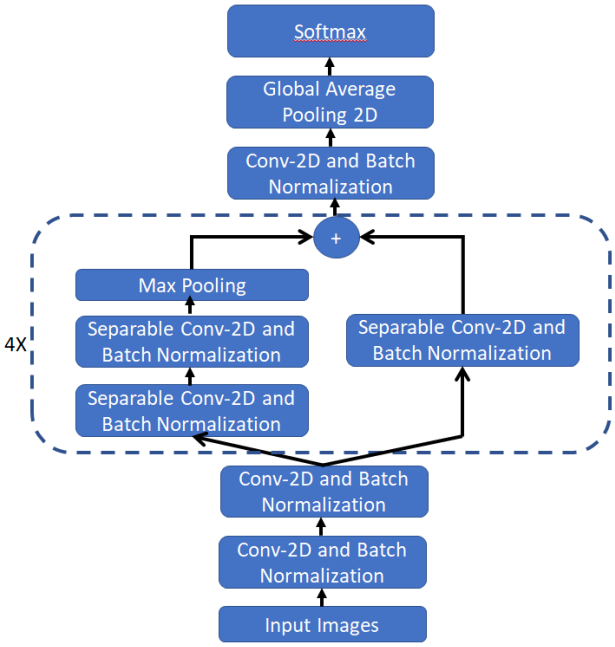

Figure 4. Architecture of proposed model

5. Experiment Result

The proposed model above is trained in the FER-2013 dataset by Kaggle [17]. FER-2013 consist of 35.887 grayscale images. Each image contains person's facial expression with size 48*48 pixels. This data set is split into 3 types of data: 28.709 for data train; 3.589 for validation, and 3.589 for testing. Figure 5 show the example images in FER-2013 data set.

Figure 5. Facial expression images in FER-2013 data set

As previously stated, 80% of the images in the FER-2013 dataset are used to train the proposed CNN model. The training process is done for 102 epoch in a computer with specifications shown in

table 1. Overall, our proposed model get accuracy of 66% for emotion classification in FER-2013 data set. The result classify the person's facial emotion into 7 category : "angry", "disgust" , "scarred", "happy", "sad", "surprised" and "neutral".

Table 1. Computer specification

Operating system	Windows 7
Processor	Intel Core i7 – 16 GB RAM
Python version	3.6
GPU	NVIDIA Quadro K1000M

An example for our emotion classifier system can be seen in figure 6 which is containing face detection and emotion classification. In table 2, we provide the accuracy of our proposed emotion classifier. Here, happy classification gets the highest accuracy. This indicates that happy emotion is easy to be recognized. But, sme misclassification is occurred for all emotion categories. For example, our model classifies "scarred" instead of "sad", or classify "scared" instead of "angry". This is because in some perspective, the scarred emotion can be interpreted as sad emotion.

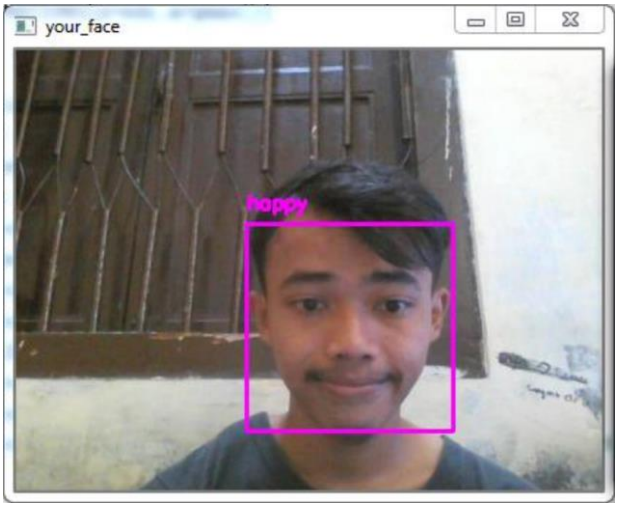

Figure 6. Result of real-time emotion classifier system

Table 2. Classification result

Classification Output \ Accuracy (%)	Angry	Disgust	Scared	Happy	Sad	Surprised	Neutral
Angry	**42.02**	29.20	7.56	0.34	19.65	11.78	17.7
Disgust	0.13	**32.10**	0.72	0.04	0.32	0.12	0.55
Scared	29.27	7.93	**34.63**	1.8	17.03	29.70	15.18
Happy	0.05	0.40	4.38	**72.42**	2.66	1.68	1.38
Sad	13.05	13.33	24.73	0.58	**36.79**	6.58	14.51
Surprised	4.08	1.72	0.28	0.33	3.02	**40.30**	4.89
Neutral	11.40	13.32	27.70	24.48	20.54	9.84	**49.79**
	Angry	Disgust	Scared	Happy	Sad	Surprised	Neutral

Truth Emotion

6. Conclution

The model of CNN for emotion classification was presented in this paper. This model, inspired by Xception model, uses the dept-wise separable convolution layer to reduce the number of parameter from 101K to 33K parameters. Our proposed model get an accuracy 66% in classifying 7 basic emotion : “angry”, “disgust” , “scarred”, “happy”, “sad”, “surprised” and “neutral”. Happy emotion get the highest accuracy. But our model get the lowest accuracy of 31.10 % for disgust classification. Overall accuracy is still low. For future work, the performance of classifier must be improved by redesigning the model architectures.

References

[1] R. Beale and C. Peter, “The Role of Affect and Emotion in HCI,” in Affect and Emotion in Human-Computer Interaction: From Theory to Applications, C. Peter and R. Beale, Eds. Berlin, Heidelberg: Springer, 2008, pp. 1–11
[2] “Facial Action Coding System,” Paul Ekman Group. [Online]. Available: https://www.paulekman.com/facial-action-coding-system/. [Accessed: 26-Sept-2020]
[3] J. Hamm, C. Kohler, R. Gur and R. Verma, ”Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders”, Journal of Neuroscience Methods, vol. 200, no. 2, pp. 237-256, 2011. Available: 10.1016/j.jneumeth.2011.06.023
[4] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” arXiv:1804.08348 [cs], Apr. 2018
[5] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. CoRR, abs/1610.02357, 2016
[6] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[7] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision (IJCV), July 2016.
[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[9] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014
[10] 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.
[11] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, vol. 1, pp. I–I
[12] 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
[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial Transformer Networks,” CoRR, vol. abs/1506.02025, 2015
[14] P. Lucey, J. 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,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, 2010, pp. 94–101.
[15] Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.
[16] O. Arriaga, M. Valdenegro-Toro, and P. Ploger, “Real-time Convolu- ¨ tional Neural Networks for Emotion and Gender Classification,” CoRR, vol. abs/1710.07557, 2017.

[17] I. J. Goodfellow et al., Challenges in Representation Learning: A report on three machine learning contests. 2013