A Novel Facial Expression Recognition Method for Identifying and Recording Emotion

Ziqi Wei
The Hill School, Pottstown, PA, USA
ZWei20@thehill.org

Abstract. In this study, an accurate, innovative, and real-time emotion recognition and recording system based on novel deep learning techniques is proposed. A nine-layered convolutional neural network (CNN) is built with the addition of backpropagation (BP) to increase accuracy and Haar-Cascade with Open Source Computer Vision (OpenCV) for frontal face detection. In past research, mass datasets for investigations into emotion are often collected through case studies, surveys, or naturalistic observations. Due to their subjective nature, these studies often yield biased results. Through the use of facial recognition software trained with the Adam optimizer and evaluated on Kaggle’s Fer2013 public dataset, this CNN can recognize six universal emotions – happiness, sadness, anger, fearful, surprise, and disgust – with the addition of neutral for other or no emotion. After training, this method has reached a train accuracy of 95% and a test accuracy of 53.55%. In addition, the system uses an unprecedented method that creates a graph demonstrating the short-term fluctuation of emotion on a millisecond basis. The final emotion value of each short-term evaluation will appear in a separate file to document long-term emotional status. Aside from software, a smart mirror including a camera and a LED Electronic Screen is built for real-time feedback of people’s emotional states. Every 100 millisecond, a label and an emoji will appear on the mirror to create an easily accessible and entertaining form of emotion detection.

1. Introduction
In the past year in the United States alone, 6.9% of adults have experienced at least one major depressive episode [1], 9.7% of adults suffer from mood disorders [2], and 18.1% of adults have experienced anxiety disorders [3]. Emotional disorders as such contribute to suicide being the 10th leading cause of death nationwide [4], and the lack of control over one’s emotions is the culprit [5]. With common research methods like case studies, naturalistic observations, and surveys, people are unable to receive immediate feedback on their emotional status and often receive subjective evaluations. Having a unique combination of a deep learning network through the use of a CNN, BP, and OpenCV, this system can give people real-time feedback while giving them both a short-term and long-term record of their emotions.

This research is based on deep learning. As a mathematical modeling tool and algorithm designed for automatic learning, deep learning has strong adaptive learning features useful for implicit representation [7, 8, 9]. Combined with a professional facial expression recognition dataset (Fer2013), this system has achieved the purpose of real-time emotion classification through live video. Past research
has shown 6 emotions that are universally the same – happiness, sadness, anger, fear, surprise, and disgust [6]. After also including neutral for other or no emotion, this system can be used universally. In addition, through the use of python MatPlotLib and python file module, this study is able to record short-term fluctuation of emotion every millisecond and automatically represent it on a line graph. A long-term record of the final emotional level during each short-term recording will be put in a separate text file.

2. Method

2.1. Dataset (Fer2013)
The facial expression database for training this CNN is the Fer2013 public dataset from the Kaggle website\(^1\). Used for the Kaggle facial expression recognition challenge, this dataset consists of 28,709 examples for the training set and 3,589 examples for the test set. The final test set used for determining the winner of the competition contains another 3,589 examples.

| Number | Emotion | Amount in Dataset |
|--------|---------|-------------------|
| 0      | Anger   | 3,995             |
| 1      | Disgust | 436               |
| 2      | Fear    | 4,097             |
| 3      | Happy   | 7,215             |
| 4      | Sad     | 4,830             |
| 5      | Surprise| 3,171             |
| 6      | Neutral | 4,965             |

2.2. Experiment Design

This research is designed with a convolutional network. A convolutional network, also known as a convolutional neural network [10], is a neural network designed for processing data like time series data (a one-dimensional grid regularly sampled on the time axis) and image data (a two-dimensional pixel network) [11]. It differs from other neural networks through its use of a mathematical operation of a special linear operation – convolution. Replacing a general matrix multiplication operation with a convolutional operation, a convolutional network consists of at least one layer and excels at image, text, and audio application.

The basic CNN consists of four layers: a convolutional layer, a pooling layer, a nonlinear activation layer (such as ReLU), and a fully connected layer. The convolutional layer is used to implement the convolution operation through the convolution kernel and has the characteristics of parameter sharing and local sensing. Through the convolution operation, an automatic layer-by-layer extraction of features is realized. Generally following the convolutional, the pooling layer is used for downsampling input features (such as images that through this process will be filled with “SAME” and “Valid”), reducing the number of parameters, and retaining core input features. This can be achieved through either average pooling (pooling window takes Pixel average) or maximum pooling (pooling window takes Pixel maximum). This study utilizes maximum pooling. Nonlinear activation functions – such as ReLU, tanh, etc. – are then used in the convolutional layer to perform a nonlinear transformation on the result. The fully connected layer is last and acts as a classifier, that is, the final model output.

\(^1\) https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
Table 2. CNN Structure Design

| Layer Name          | Parameter Setting                                      |
|---------------------|--------------------------------------------------------|
| Input Layer         | 48 * 48 * 1                                            |
| Convolutional Layer 1| 64 @ 5 * 5; Stride = 1; Activation = ReLU               |
| MaxPool Layer 1     | 3 * 3; Stride = 2; Padding = “Same”                    |
| Normalization Layer 1| Depth_Radius = 2; Bias = 1, Alpha = 0.001/9.0; Beta = 0.75 |
| Convolutional Layer 2| 64 @ 3 * 3; Stride = 1; Activation = ReLU               |
| Normalization Layer 2| Depth_Radius = 2; Bias = 1, Alpha = 0.001/9.0; Beta = 0.75 |
| MaxPool Layer 2     | 3 * 3; Stride = 2; Padding = “Same”                    |
| Flatten Layer 1     | Units = 9,216; 3-D (12 * 12 * 64) convert to 1-D (9,216); Activation = ReLU |
| Fully Connected Layer| Units = 192                                           |
| Output Layer        | Units = 7                                              |

Excluding the input layer, this CNN model has nine layers. After a 48 * 48 frontal face greyscale image is passed through the input layer, it goes through the convolutional layer, the maximum pooling layer, and the normalization layer twice. After, the data maintain the same distribution in the statistical sense as it passes through the flatten layer, the fully connected layer, and finally outputs the probability scores for all seven categories by using the softmax function. The maximum probability is viewed as the final category and will be expressed through the use of an emoji corresponding to its category.

![Output (Happy)](image1.png)  ![Output (Surprise)](image2.png)  ![Output (Sad)](image3.png)

In training this CNN, the backpropagation method is used for updating weights by comparing the actual output with the expected output. The derivative of the weight, \( w \), calculated by the chain rule, is defined as follows:

\[
\frac{\partial L(z,y)}{\partial w} = \frac{\partial L(z,y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w}
\]  

(1)

The cross-entropy loss function \( L(z,y) \) is also defined:

\[
L(z,y) = - \left[ y \log(z) + (1 - y) \log(1 - z) \right]
\]

(2)

In formula (2), \( z \) represents the actual output while \( y \) represents the expected output. As a result, the weight update is as:

\[
w \leftarrow w - \eta \frac{\partial L(z,y)}{\partial w}
\]

(3)

In formula (3), \( \eta \) represents the learning rate.

The process of weight update consists of four steps. First, the training dataset is split for the first batch; then relevant losses is calculated through forward propagation; After, calculate the gradient of the loss function through backpropagation; and finally, use the gradient to update weights of the CNN.

To achieve backpropagation, this study utilizes the Adam optimizer, since it produces the most accuracy. The learning rate is set as 1e-4 while other parameters, implemented by Kingma and Ba [12], are used as default.
After a pre-trained CNN model is created, the haarcascade_frontalface_default.xml file is loaded with OpenCV. As a result, Haar-Cascade, a pre-trained classifier for determining if a certain object exists in an image, is implemented for frontal face detection in the video stream. Subsequently, OpenCV will crop the frontal face image extracted from the video to a size of 48 * 48 * 1 for future steps.

![Figure 4. 48 * 48 * 1 Greyscale Picture](image)

If the human face is detected, a predict algorithm will run to identify facial expression. After being resized to 48 * 48 * 1, the frontal face image is passed as input through the ConvNet. The network will output a list of softmax scores for the 7 categories. Finally, the emotion with the highest score will be displayed as an emoji on the screen, as shown in Fig. 1, 2, and 3.

Finally, a quantitative design is used for recoding the fluctuation of emotion. All emotions are split into three categories – positive, neutral, and negative. Happy is positive, neutral and surprise are neutral, and sadness, fear, anger, and disgust are negative. As the real-time video is streamed, the computer outputs an emotion every millisecond. This emotion will be computed into a graph using the criteria as follows:

![Table 4. Classification and Weights of the 7 Emotions](image)

The accumulative results of each session are then calculated and saved for long-term emotion tracking. The specific formula is shown as:

\[ p = 4 \times C_p + (-1) \times C_n + 0 \times C_t \]  

In formula (4), \( C_p \) represents the number of occurrences of positive expressions, \( C_n \) represents the number of occurrences of negative expressions, \( C_t \) represents the number of occurrences of neutral expressions, and \( p \) represents the final emotion value of each session.

In addition, with an external camera, a light emitting diode (LED) electronic display screen, and an acrylic see-through mirror, a convenient and fun way of detecting emotion in everyday life is created.
3. Data Analysis

3.1. Train Process

In the training process of this CNN model, the parameter is set as follows:

| Layer Name      | Parameter Setting                                |
|-----------------|--------------------------------------------------|
| Input           | Image = 48 * 48 * 1 Greyscale; Label = 7        |
| Loss            | Cross Entropy                                    |
| Optimizer       | Adam; Learning Rate = 10^{-4}                    |
| Batch Size      | 50; Shuffle = True                               |
| Train Steps     | 20,000 – 30,000                                  |
| Accuracy        | Arguments of the Maxima Function + Reduce Mean Function |
| MetPlotLib      | Train Process with Iteration, Loss, and Accuracy |

The specific train process is shown as such:

![Figure 5. Train Steps](image)

After 30,000 steps of training, a dynamic graph of the change in loss and accuracy with respect to the number of steps in the training set is shown.
In Fig. 4, the blue lines represent the change of loss with respect to train steps. The smaller the loss value, the closer the actual output of the represented model is to the expected output. In other words, during its train process, as the loss decreases, the accuracy of the CNN increases. During this train process, the loss value continues to drop from 4.87 to 0.0012. Simultaneously, the red line, representing the change of accuracy with respect to train steps, is increasing. From 0 – 15,000 steps, the accuracy rises linearly, increasing from 0.12 to 0.95. Similarly, the loss value decreases in linear form during the first half of training. As the model increasing receives data, its learning speed is steady as it achieves the update of parameters (weights and biases) through all neural levels. However, through steps 15,000 – 30,000, the growth of accuracy and the decay of loss levels off to a relatively horizontal line. This is due to the use of the Adam optimizer, which is an adaptive learning method. Therefore, the stochastic gradient descent (SGD) converges to a different or sub-optimal minimum value. As a result, the accuracy stabilizes to a 0.95 in the latter stage, indicating that the model achieves convergence during the train process.

3.2. Test Performance
At the same time, the graph for the accuracy of the test set is created.

Through the graph, it can be observed that in general, the accuracy of the test set is also increasing and stabilizes at around 0.5 – 0.55. The accuracy of the best model output is around 0.536. Comparing the train set with the test set, it can be seen that both are increasing in the first 15,000 steps while leveling off to a rather stable point in the second 15,000 steps, which is closely related to the slow update of the
overall parameters of the model. According to the Kaggle Facial Recognition Contest, the accuracy of this model ranks at #22 worldwide, earning it a bronze medal.

Figure 8. Kaggle Official Website Facial Recognition Contest Results

3.3. Recording Emotion
The short-term emotion performance recording method achieved through a unique design creates graphs as followed:

Figure 9. Emotion Fluctuation Graph in 2s
Figure 10. Emotion Fluctuation Graph in 12s

In the figures above, the x-axis represents the length of time in milliseconds, the y-axis represents the emotion value, and the red line represents the emotion alert value. If the emotion value is above the red line, a person’s emotional status is considered positive. Vice versa, if the emotion value is below the red line, a person’s emotional status is considered negative. The graph also depicts emotion fluctuation during each session in great detail.

Each time the software is closed, the final emotion value is recorded into a text file. In order to track long-term emotional status, the method of persistence to the local disk is used. The result is shown below:

---

2 https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/leaderboard
Figure 11. Long-Term Emotion Record Through Text File

From the text file above, it is shown that each line contains the date and time interval, the emotion attribution (positive, negative, or neutral), and the final emotion value. Many lines combine to present long-term emotion status. For a better understanding of people’s emotional status, specific times of the day (morning, evening, or night) can be chosen for recording emotions. With controlled variables (time, lighting, or situation), people’s emotional responses can be recorded to reflect their emotional state for either self-reflection or determining the need for treatment.

4. Discussion and Conclusion
This research implements facial expression recognition through the use of convolutional neural networks. By incorporating BP and the Adam optimizer in this model, the accuracy increased drastically. HaarCascade implemented by OpenCV also created the opportunity for timely feedback through live video. Not only is the software accurate and real-time, but also the hardware is designed to attract interest and presents an easily accessible form of emotion detection. Most importantly, this model is designed to record both short-term and long-term emotional fluctuation in a unique way, providing a method for both a qualitative and quantitative measurement of emotional status.

In future research, through engineering fine-tuning or increasing the number of network structure layers, the accuracy of this model can increase. To better assess emotion, adding a system to detect voice tones, body language, or context can help researchers understand the different variables that contribute to emotion. Aside from adding variables, expanding the model to read more than just the frontal face can increase the practicality of this model since it creates more space for mobility when evaluating emotion. Also, as large as the Fer2013 dataset may be, it does not incorporate different lighting situations. With different datasets including different types of lighting situations – like datasets with images in low lighting or with overlooking lights – the software can be altered to better adapt to different situations. In addition, the hardware of this research can be tweaked in different ways for different populations. For example, for children, the camera can be placed in a teddy bear as its eyes with the LED electronic display screen on the belly.

There is also a wide range of future applications. High-precision models can be deployed in the cloud to connect to real-time cameras, which will achieve commercialization applications such as customer sentiment analysis in malls through store cameras. For educational purposes, this model can also be applied to learning. For example, the response of students to different learning styles can help determine effective ways to achieve higher learning. From a psychological perspective, a large data collection of a person’s emotion can help diagnose certain mental illnesses that stem from unbalanced emotions. It can also help therapists to better understand their patients since it has the capability of catching microexpressions the human eye may not notice that are crucial to understanding patients’ mental states.

More simply, this model can be used for entertainment or merely for people to understand themselves better. Overall, this model can be applied to various fields to create a more convenient and accurate way of understanding emotion.
References
[1] National Institute of Mental Health, “Major Depression,” November 2017.
[2] National Institute of Mental Health, “Any Mood Disorder,” November 2017.
[3] National Institute of Mental Health, “Any Anxiety Disorder,” November 2017.
[4] World Health Organization, “Suicide,” August 2018
[5] D. Barlow, “Disorders of Emotion,” Psychology Inquiry, vol. 2, issue 1, pp. 58–71, 1991.
[6] P. Ekman, “Universal Facial Expressions of Emotion,” California Mental Health Digest, vol.8, issue 4, 1970.
[7] Hinton, G.E., “Deep Belief Nets,” 2017 NIPS Tutorials Notes, Neural Information Processing Systems Conference, Vancouver, BC, December 2007.
[8] Bengio, Y. and Lecun, Y, “Scaling Learning Algorithms towards AI,” In Large Scale Kernel Machines, 2007a.
[9] Delalleau, O. and Bengio, Y., “Shallow vs. Deep Sum-Product Networks,” In NIPS, 2011.
[10] LeCun, Y., Jackel, L.D., Boser, B., Denker, J.S., Graf, H.P., Guyon, I., Henderson, D., Howard, R.E., and Hubbard, W., “Handwritten Digit Recognition: Application of Neural Network Chips and Automatic Learning,” IEEE Communication Magazine, 27 (11), 41-46, 1989.
[11] Chen, B., Ting, J. A., Marlin, B. M., and de Freitas, N., “Deep Learning of Invariant Spatial-Temporal Features from Video,” NIPS*2010 Deep Learning and Unsupervised Feature Learning Workshop, 2010.
[12] Kingma, D. and Ba, J., “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv: 1412. 6980, 2014.