CONVOLUTIONAL NEURAL NETWORKS FOR FACIAL EMOTION RECOGNITION

Hoang-Anh Le
VNU-HCM University of Science
1612013@student.hcmus.edu.vn

Anh-Quoc Pham
VNU-HCM University of Science
1612543@student.hcmus.edu.vn

Thien-Nu Hoang
VNU-HCM University of Science
1612880@student.hcmus.edu.vn

Abstract

Facial emotion recognition plays an important role in helping human-machine interaction become more intelligent and natural, and automating many surveys and researches in human behavior, health-care and robotics. Knowing its significant, we implement Convolution Neural Networks (CNNs) for this problem. We have two CNNs models, consist of shallow CNNs, deep CNNs, which have different number of layers and are trained by using Kaggle dataset. In result, the highest accuracy is 65.55% belong to deep CNNs. Our motivation is understanding clearly about deep learning, particularly CNNs, and put in on real life. Therefore, we also tunned the hyper parameter of each models such as learning rate, batch size, and number of epochs. In addition, we also used techniques to optimize networks, acting as activation function, dropout and max pooling. Finally, we analyzed the result from two models to observe the relationship between number of layer and accuracy. We believe our result will be valuable to make decision how structure of network before construct it.

1 Introduction

With the need of human–computer interaction, the emotion recognition plays an important role in computer science. There are a lot of ways to recognize emotion included through voice, body gesture, specially facial expression what is the most important way for human to display emotions. In fact, facial emotion recognition has been applied for a lot of applications of variety fields such as customer – attentive marketing, health monitoring and emotionally intelligent robotic interface. Therefore, researching about facial emotion recognition has increasingly attracted many scientists in computer vision.

In 1971 paper titled "Constants Across Culture in the Face and Emotion", Ekman et al. identified six facial expressions that are universal across all cultures: anger, disgust, fear, happiness, sadness and surprise. In recent years, research challenge such as Emotion Recognition in the Wild (Emotion) and Kaggle's Facial Expression Recognition Challenge added the seventh emotion, neutral emotion, into this list for classification.

The first successful applications of CNNs were developed by Yann LeCun in 1990's. Of these, the best known is the LeNet architecture that was used to read zip codes, digits, etc.[9]Day by day, CNNs has been developing by the scientist community. Specially, in computer vision, there are a lot of momentous works which use CNNs approach such as AlexNet[1], VGGNet[7], GoogleNet[5], and ResNet[6]. Besides, there are some contribution from public challenges, typically Facial Emotion Recognition challenge in Kaggle (2013) and Emotion Recognition in the Wild challenge (2015).

In this paper, we executed CNNs approach for facial emotion recognition. The input in to out system is image from Kaggle dataset, then we use CNNs to train and predict the label facial expression, consist of angry, disgust, fear, happy, sad, surprise, and neutral. We tried to build distinct CNNs systems with various layers to find out the best performance. We reached 65.55% of highest accuracy. It can be accepted because the winner of FER2013 challenge achieve 71.162% accuracy. Not only accuracy achieved, we also found some interesting things about CNNs. Although our result and method is not the best, it made us understand deep learning obviously and easier to implement it.This experiment gave us a basic knowledge for out future work.

We mention some related work in Section 2. Dataset is remarked in Section 3. Section 4 show what exactly we did, how our CNNs models in detail. Content in Section 5 is our result. In ending, we also present some conclusion and future work in Section 6

2 Related Work

From 2000 to 2018, many researchers have developed the facial emotion recognition systems. There have been a lot of approaches to solve this problem, from traditional approaches using handcrafted features to deep-learning-based approaches. [8] However, CNNs method was used for most of public challenge and resulted high accuracy. In fact, in FER2013 challenge, the winner, Yichuan Tang, used an ensemble of CNNs trained to minimize the squared hinge loss, and achieved 71.162% accuracy on test set. [2].

In more recent work, Bo-Kyeong Kim et al won the third Emotion Recognition in the Wild (EmotiW2015) challenge with test accuracy of 61.6%. They used a large committee of CNNs with two strategies: varying network architecture (e.g. input preprocessing and receptive field size) in order to obtain more diverse models, and constructing a hierarchical architecture of the committee with exponentially-weighted decision fusion in order to form a better committee. [3]

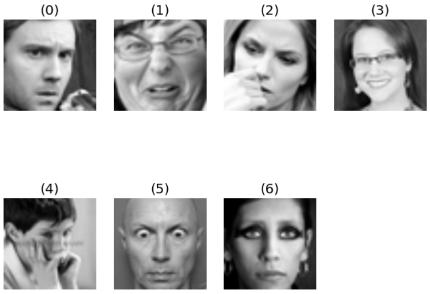

Figure 1: Example of seven emotions in FER2013 dataset: (0) angry, (1) disgusted, (2) fearful, (3) happy, (4) sad, (5) surprised, (6) neutral

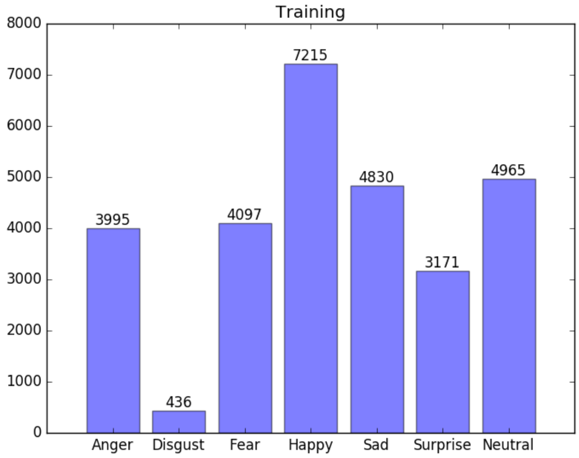

Figure 2: Overview FER2013 data

3 Dataset

We trained and tested our model on Kaggle dataset from Facial Expression Recognition Challenge, which consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. We use training set of 28,709 example, validation test set of 3,589 examples, and test set of another 3,589 examples. [4] In Figure 1, we show seven emotional images included: angry, disgust, fear, happy, sad, surprised, neutral. Although all images are preprocessing, we can see that there are various individuals across the entire spectrum of: ethnicity, race, gender and race, with all these images being taken at various angles. The plot in Figure 2 show the number images of each emotion. Absolutely, disgust is the smallest data (436 images). We predict it will affect to result of experiment.

4 Method

4.1 Overview

We use Convolution Neural Networks (CNNs) to recognize seven facial emotion expression. In CNNs approach, the input image is convolved through a filter collection in the convolution layers to produce a feature map. Each feature map is then combined to fully connected network, and the face expression is recognized as belonging to a particular class-based the output of softmax algorithm. There are two main reasons why we chose CNNs approach:

- CNNs is the most popular network model among the several deep-learning model available. Understanding CNNs helps us develop researching deep-learning career in the future. [8]
- Using deep-learning for facial emotion recognition highly reduce the dependence on face-physics-based model and other pre-processing techniques by enabling "end-to-end" learning to occur in the pipeline directly from the input images. [8]

To reach the motivation of this paper, we implement two classifiers from scratch: (1) shallow CNN with 2 layers, (2) deep CNN with 4 layers. For each of these model, we tunned parameters including learning rate, regularization, and dropout. We also tried using batch normalization and fractional pooling for optimizing time training.
According to the results of three models, we compare these with loss and accuracy to understand exactly how CNNs works.

4.2 Shallow CNN

This network has two convolution layers and one fully connected (FC) layer. In the first convolutional layer, we had 64 filters with kernel size is 3x3, border mode is 'same' and value of input shape is (48,48,1) since input image is 48 x 48 pixels grayscale. The second convolutional is a bit different from first one. In this layer, we have 128 5x5 filters. These convolutional layers also along with batch normalization, max-pooling layer and dropout. Pooling setup is 2x2 with a stride of 2 to reduce the size of the receptive field and avoid overfitting. In dropout layer, a fraction of 0.25 is used.
After 2 convolutional layers, network is added a FC layer after being flattened. FC layer has a hidden layer with 256 neurons and loss function is binary cross entropy (softmax).
Also in all the layers, Rectified Linear Unit (ReLU) is used as the activation function to model non-linearity. ReLU is simply and make high performance.

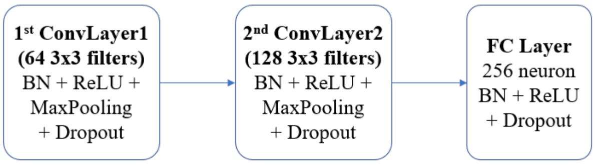

Figure 3: Architecture of Shallow CNN

4.3 Deep CNN

To observe the effect of adding convolutional layers and FC layers to the network, we build a deep CNN with 4 convolutional layers and 2 FC layers. The first and second convolutional layers and the first FC layer in this network is the same with layers in Shallow CNN. The third and fourth convolutional layer is the same, they have 512 3x3 filters, along with Batch-Normalization, Max-pooling layer, Dropout layer and ReLU as activation function. The hidden layer in the second FC layer has more neuron than first FC layer, 512 neuron.

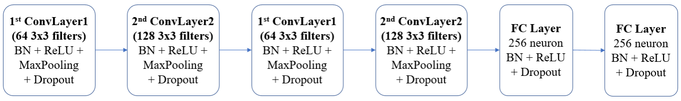

Figure 4: Architecture of Deep CNN

5 Result

In order to compare the results of shallow and deep networks, we computed the confusion matrices for the these models, shown in Figure 5 and Figure 6.

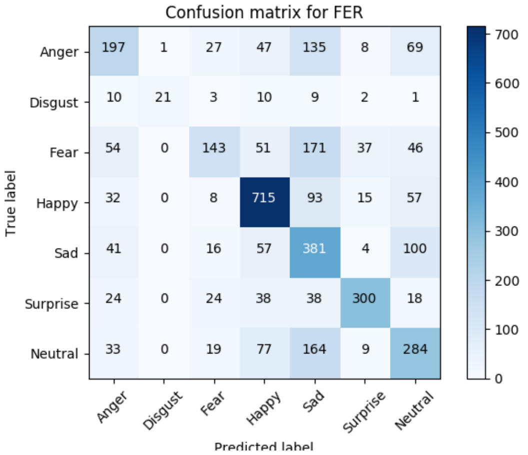

Figure 5: Confusion Matrix of Shallow CNN

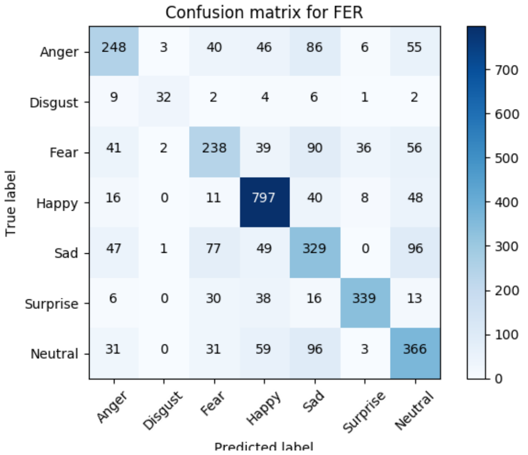

Figure 6: Confusion Matrix of Deep CNN

We can see through the figures that the deep network has a more accurate result than the shallow network's, as most of the cells in the primary diagonal (number of correct recognitions) have higher values , and most of the cells not in the primary diagonal (number of incorrect recognitions) have lower values. Moreover, we can also know which labels are easily to be incorrect recognized, and be confused with other labels. For example, in the shallow network, the Anger, Fear and Neutral often be recognized as label Sad. The label Disgust have a small amount of data, so the accuracy is quite low and not stable.
In addition, we computed the table of accuracy (the percentage of correctly recognitions case of each label).

Table 1: Recognition Accuracy of each Label

Label	Shallow Network	Deep Network
Anger	40.70%	51.24%
Disgust	37.5%	57.14%
Fear	28.49%	47.41%
Happy	77.71%	86.63%
Sad	63.61%	54.92%
Surprise	67.87%	76.7%
Neutral	48.46%	62.45%
Overral	56.31%	65.55%

In the table, we can see the label Happy has the highest correct recognition rate in both network. And the percentage of deep network higher than in shallow network in most of labels.

6 Conclusion

In this paper, we have explored CNNs for recognition facial expression. Firstly, we implemented a shallow CNN, which have two convolution layers, and got a low accuracy (56.31%). In oder to improve this networks and expect a higher accuracy, we demonstrated a deep CNN by adding two convolution networks into a shallow CNN. The highest accuracy we achieve is 65.55%. We also found some emotion that is not be recognized well because of shortage of dataset.

Through this experiment, we learn how to implement a CNNs model to solve a real life problem. In future, we would like to implement a deeper CNN with a paramaterizable number of convolutional layers, and check that if more number of convolutional layers, the higher accuracy is. Moreover, we would like to extend our model for color images with investigating pretrained models such as VGG-Net [7] and AlexNet[1]

References

[1] G. E. H. Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. 2012.

[2] I. J. G. E. L. C. C. M. H. C. T. T.-H. L. Z. R. F. L. W. A. S.-T. M. P. I. P. G. B. X. R. X. C. Bengio. *Challenges in Representation Learning: A Report on Three Machine Learning Contests*. Springer, Berlin, Heidelberg, 2013. Lecture Notes in Computer Science.

[3] J. R.-Y. D.-Y. L. Bo-Kyeong KimEmail. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Multimodal User Interfaces*, 2016.

[4] P.-L. Carrier and A. Courville. Kaggle facial expression recognition challenge. URL `https://www.kaggle.com/c/challenges-in-representation-learning-facial-expr` `data`.

[5] Y. J. P. S. S. R. D. A.-D. E. V. V. A. R. Christian Szegedy, Wei Liu. Going deeper with convolutions. 2014.

[6] S. R. J. S. Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. 2015.

[7] A. Z. Karen Simonyan. Very deep convolutional networks for large-scale visual recognition. 2014.

[8] B. C. Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, page 2, 1 2018.

[9] P. H. Y. B. Yann Lecun, Leon Bottou. Gradient based learning applied to document recognition. *The IEEE*, 1998.