Evaluation of A Convolutional Neural Network System on Emotion Recognition

Hao Che
College of Engineering and Computer Science
Australian National University
Canberra ACT 2600 Australia
u6548236@anu.edu.au

Abstract. In this paper, a convolutional neural networks (CNN) network system for an emotion recognition task has been designed and developed. The goal of the designed CNN network system is to classify movie image into one of the seven facial emotion classes. As a static facial expression dataset, SFEW is used as training and testing. The designed CNN network system applied image preprocessing methods include histogram equalization, grayscale, resize and normalize and data augmentation methods include horizontal flipping, image rotation to enhance system performance on SFEW dataset. Some pervious works performed emotion classification on SFEW dataset will be compared with the CNN network system designed for evaluation. Experiments show that when input small size image to the designed CNN network system still can achieve good accuracy performance on SFEW dataset.

Keywords: Convolutional Neural Networks, Casper, Fully connect neural network, SFEW dataset, PHOG, LPQ, Image Preprocessing, Data Augmentation, Leakey ReLU, k-fold cross validation

1 Introduction

Although humans can recognize facial expressions with little effort, people have difference in the accuracy of recognizing the emotional aspects of others. In the past few years, some progress has been made in face detection, feature extraction for facial expression classification. However, it is difficult to develop a system that can accomplish emotion recognition automatically [1]. Most emotion recognition tasks can be divided into static emotion recognition and dynamic emotion recognition. Dynamic emotion recognition as a sequence task in neural network and deep learning can extract information from time series inputs. In contrast, static emotion recognition extract information from static images or features. As a video-based method, dynamic emotion recognition is more robust than static emotion recognition. In this paper, a simple network based on Convolutional Neural Networks (CNN) has been developed to evaluate for facial emotion expression recognition.
Convolutional Neural Network CNN is an improvement of the Artificial Neural Network (ANN), which is an "end-to-end" model. CNN artificial neurons can respond to the surrounding units in a part of the coverage area. It has features including local connection and weight sharing, which reduces network parameters, speeds up training, and improves regularization effects. CNN has excellent performance for large-scale image processing.

1.1 Dataset and features

In deep learning and neural network, emotion recognition is a hot area of research. Static Facial Expressions in the Wild (SFEW) [2] is a static facial expression database which can be used for static emotion recognition. SFEW dataset contains 675 images, which include 75 disgust emotion images and 100 angry, fear, happy, natural, sad and surprise emotion images. Resources of images in SFEW dataset are movies. SFEW dataset covers facial expressions that include unconstrained facial expressions, various head poses, larger age ranges, different facial resolutions, close to real-world lighting and so on. Most of images in SFEW dataset are not face cantered images. Some images are in low brightness and dark background. Comparing to other emotion recognition datasets, although SFEW is a small dataset, the images in SFEW contain a variety of information with different environment. Therefore, SFEW is suitable for evaluating the generalization performance and fit velocity of a system.
Dhall et al. [2] used pyramid of histogram of oriented gradients (PHOG) descriptor and local phase quantisation (LPQ) descriptor to extract features from images in SFEW dataset. Both PHOG and LPQ descriptor can extract the PC of images. PC is Principal Components Analysis. Original image data always as large as expensive to deal with. Principal Components Analysis is a method to reduce dimension of data. Although 5 PC features cannot represent all information of data, main pattern of image can be kept. PHOG [3] is a descriptor extracting features that describing the spatial shape. It has strong anti-noise performance and certain anti-rotation ability. PHOG counts the histogram distribution of the gradient direction of edge image at different levels. The LPQ is a descriptor that extract features on the grid and based on calculating the short-term Fourier transform (STFT). LPQ descriptor shows a better performance than local binary pattern (LBP) [4].

1.2 Previous works

Hao [5] developed an improved Casper network and a fully connected network to perform classification on SFEW dataset. The improved Casper that is a Casper algorithm [6] based algorithm. With the rapid development of neural network and deep learning in 20 years, a variety of techniques can be proved to improve training networks. The techniques used in original Casper algorithm are techniques that have been claimed 20 years ago. Techniques include Z-score transformation, adding noise to data, weight sampler, cross entropy loss, leakey ReLU and RMSprop learing rate algorithm have been applied to improve Casper learning algorithm for the emotion classification task on SFEW database. The input of the two networks are top 5 pca PHOG and top 5 LPQ pca features.
The training for new added hidden neuron in improved Casper is designed no limitation to total epoch number that has been set before training. Based on the original Casper learning algorithm, the time period for training a new hidden neuron for Casper is 15+P*N, where N is the number of existed neurons, and P is set parameter [6]. While the N, P and maximum number of hidden neurons are not variant number, the total epoch cannot be identified as an exact number before training. If the total epoch number set as a fixed data before training, the new added hidden neuron does underfitting the Casper model and lead to the high error cost and low accuracy. Therefore, a flexible total epoch number has been designed to fit the issue. When the remaining epoch number is smaller than the time period 15+P*N while a new hidden neuron added, the total epoch adds the number of 15+P*N. Then, the training process can be extended and fit the new hidden neuron to Casper.
The seven emotion classes classification accuracy performance of the improved Casper network and fully connected network on SFEW dataset are 25.11% and 24.19% respectively. The experiments show that two networks have close accuracy performance, but improved Capser can reduce the cost of space on the redundant number of neurons. Improved Casper also has the good performance on fitting and generalization.

1.3 What next

In this paper, a CNN based network system has been designed. The performance of the designed system on SFEW dataset will be compared with results provided by Hao and Dhall. The rest of the paper is organized as follows: Section 2 presents the methods that used to build CNN network system. Section 3 presents and discuss the experimental results on SFEW dataset for comparing designed CNN network system and previous works. At last, Section 4 makes a conclusion and looking forward to future work.

2 Method

Considering the limited training time cost and memory size cost, the deep learning network is designed as a small CNN network. The considerations of designing method are influenced by the cost limitation. The following parts introduced the inputs and outputs of developed system, applied data augmentation methods, applied methods used for fitting training process, and details of designed CNN network model.
The model and relevant methods was developed in PyTorch [7] and a graphics processing unit (GPU) has been used for speeding up the computation training process. The designed CNN network trained on a desktop with i7-8700 CPU, and GTX 1060 6GB GPU. The batch size is set to 64. The designed network training can achieve the optimal in 30 epochs without using data augmentation and 150 epochs with data augmentation.

2.1 Image Preprocessing

The images in SFEW are close to real-world lighting. Some image preprocessing methods can be applied to improve the performance of extracting features. Four methods that contain histogram equalization, grayscale, resize and normalize have been used in image preprocessing.
Histogram equalization is an image processing method that adjust contrast based on the histogram of an image. The images in SFEW database come from movie scenes, which usually have high contrast [8]. Shin also mentioned that after using histogram equalization method to preprocess SFEW dataset images, the final classification result can be improved. Therefore, histogram equalization has been applied in image preprocessing stage.
As the size of original image in database is color image with 576(width) x 720(length) x 3(channel) size, the data size of each image is 1244160. To reduce the training workload, the images converted from color image to grayscale image which only has 1 channel. The data size of each image decreased to 414720. After that, resize images to 231 x 288 size. Then the data size of each image decreased to 66528 which is about one-nineteen of data size of original image.
The range pixel values in image is 0 to 255. Use the original features value as inputs makes gradient descent to converge slow for deep learning network. In addition, most activation functions are sensitive to values around zero. Z-score transformation method has been used to normalize input features. Z-score transformation method can center data around 0 and adjust the distribution of dataset.

2.3 Data Augmentation

Data Augmentation can improve the generalization of system. Considering the number of samples in SFEW dataset is 675 which is not large, some data augmentation methods such as image horizontal flipping, image rotation, have been applied. Although using data augmentation cannot add new features to dataset, it can improve the ability to extract more features from dataset. Image horizontal flipping method is to flip image horizontally and add them as input to training. Due to the feature of the data set is that images are not face centered and the vertical asymmetry of the images, some data augmentation methods such as flip image vertically and centering cannot be applied.

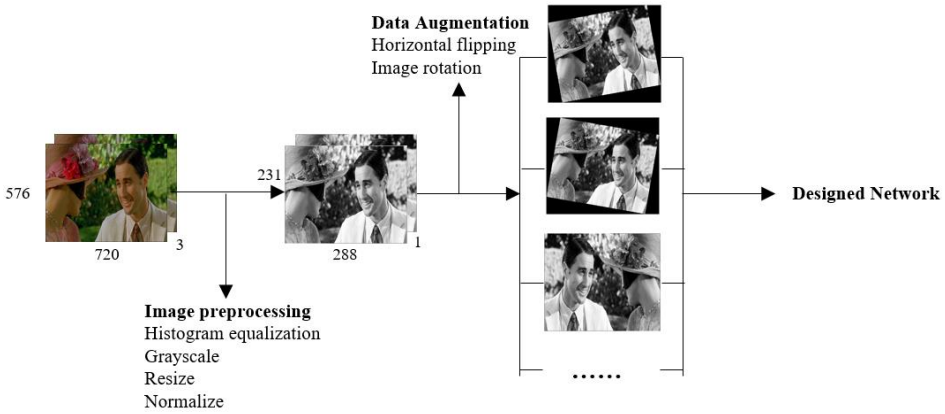

Fig1. Works before inputs data into designed network. Image preprocessing part applied histogram equalization, grayscale, resize and normlization method. Data augmentation part applied horizontal flipping and random image rotation (-10 degree to 10 degree) methods.

2.4 Inputs and outputs and network model

The input of designed network (Figure 2) is 231 x 288 x 1 image. The output of designed network is a number from 0 to 6 that are corresponding to labels angry, disgust, fear, happy, natural, sad and surprise respectively.
A CNN network (Figure 2) has been designed to evaluate SFEW dataset. The CNN network contains two convolution block and two fully connected layers. A convolution block contains one convolution layer, one 2d batch norm layer, one Leaky ReLU layer and one 2d Max Pool layer. Fully connected layers have been used to predict emotion classes from features that extracted by convolution layers.

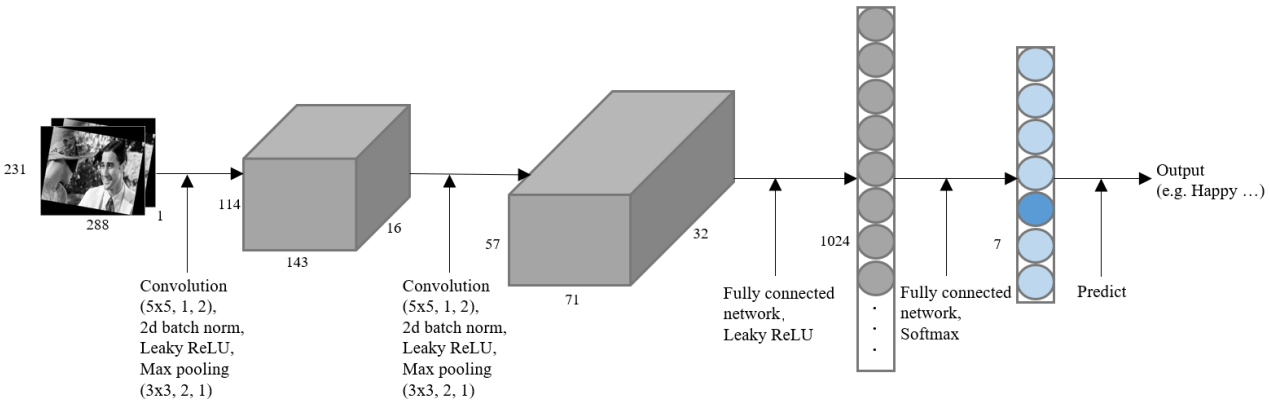

Fig2. Structure of designed CNN network. After data augmentation stage completed, 231 x 288 x 1 size image is import into network. The image data will be compuated by two convlution layers and two fully connected layers. The bule circles are the probability produced by softmax.

In this task, the number of features extracted from CNN network is 57x71x32 that is 129504 which is a large number of inputs for neural network. In model design, fully connected layers have been used to predict emotion classes from features that extracted by convolution layers rather than Casper algorithm. Although improved Casper used a variety of techniques to fit the current neural network environment, build enough hidden neurons from scratch for handling so much input data is high time and computation cost.
Leakey ReLU has been used as the activation function rather than rectied linear unit (ReLU) in the designed CNN network and fully connected network. Since ReLU [9] introduced by Nair and Hinton, ReLU has replaced sigmoid as activation function in a variety of classification tasks. As ReLU setting negative values to zero, some connections between neurons cannot contribute for neural network training. However, leaky rectied linear (Leaky ReLU) that has non-zero gradient over entire domain. It set negative values to a sloping negative not zero. Due to the designed network

is small, the entire network is intended to be fully utilized. Using Leakey ReLU makes all neurons works for network and reduce the vanishing gradient problem. Cross entropy loss function has been used to compute loss and predict labels.

3 Results and discussion

The following three parts introduced the dataset split method, performance comparison, and results discussion.

3.1 Evaluate by Stratified k-fold cross-validation

Stratified k-fold cross-validation, which is an improved version of k-fold cross validation [10] has been applied as an accuracy estimation method to designed CNN network system. In the stratified k-fold cross-validation method which applied for SFEW, the data set is split to ten folds with the almost same number of records. To compare designed CNN network and other works, the split process is randomly but fixed by an unchanging random seed. Nine folds are used as the training set in turn, the remaining fold is used as the verification set, and the average of these k results is used as our Evaluation results of the model. In the emotion classification task on SFEW dataset, k is selected by ten. After ten rounds of training and validation, the final result can be calculated by the average error and accuracy on ten validation sets. The following network evaluation is based on ten folds test values.

3.2 Accuracy Comparison

Dhall et al. (2011) used svm to perform classification on LPQ and PHOG features that extracted from SFEW dataset. The classification accuracy provided by Dhall et al. is 43.71% for LPQ and 46.28% for PHOG respectively. Hao (2020) developed an improved Casper network and a fully connected network to perform classification on top 5 LPQ pca and top 5 PHOG pca image data that extracted from SFEW dataset. The Performance comparison between designed CNN, Dhall's svm and Hao's networks Tables 1.

Table 1. Performance comparison between designed CNN, Dhall's SVM and Hao's networks

	Main technique	Inputs	Standard deviation on 10fold	Average Accuracy on Test Sets %
Dhall et al.	SVM	LPQ features	Not provided	43.71%
Dhall et al.	SVM	PHOG features	Not provided	46.28%
Hao	Improved Casper	Top 5 LPQ pca and Top 5 PHOG pca	0.0399	25.11%
Hao	Neural network	Top 5 LPQ pca and Top 5 PHOG pca	0.0664	24.19%
Designed CNN network system	CNN network	Resized grayscale image	0.0383	45.35%

The results show that the accuracy (45.35%) of designed CNN network system on SFEW is close to the accuracy of Dhall's SVM (46.28%) on PHOG features of SFEW.

And the accuracy performance is higher than other works include Hao's networks on top 5 LPQ pca and top 5 PHOG pca features and Dhall's SVM on LPQ features of SFEW. The results also show that the Standard Deviation of designed CNN network system is not too large and similar with improved Casper network.

3.3 Results Discussion

Serval reasons may influence the performance of designed CNN network system that not exceed the highest accuracy 46.28%. First, while evaluating designed CNN network system, there are 25 images with label 'Disgust' is not in dataset. Second, due to cost limitation, the original image data size decreased from 576 x 720 x 3 to 231 x 288 x 1 in image preprocessing stage. Thus, some of image data information were lost. Third, random seed selection may not be optimal. Different random seed may cause tiny evaluation performance.

The main reason that designed CNN network is better that performance of Hao's network is that the input of Hao's networks is only 10 pca features. The input information is much less than image input. Although the 10 pca features are the top 5 features of PHOH and LPQ, many image information were not included in these features. In addition, the inputs were only the features of original images. No data augmentation was used before extract image data by the PHOG and LPQ descriptor in Hao's work.

4 Conclusion and Future Work

Based on the feature extraction function of the CNN network and some image preprocessing and data augmentation methods applied, the designed CNN network system classification task has achieved good performance. The accuracy performance of designed CNN network is much higher than the Hao's previous work. However, the accuracy performance is close to but not better than the Dhall's SVM based on PHOG features. Some additional works can be added to improve the performance. As the designed CNN network is not as complex as some effective deep learning networks such as Resnet and GoogleNet produced by other researchers. Designing a new network system according to these complex and effective network structure will be the works in the future.

5 References

1. Sebe, N., Lew, M., S., Cohen, I., Sun, Y., Gevers, T., Huang, T., S.: Authentic Facial Expression Analysis. Image and Vision Computing 25.12: 1856-1863. (2007)
2. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static Facial Expression Analysis In Tough Conditions: Data, Evaluation Protocol And Benchmark. Proceedings of the IEEE International Conference on Computer Vision and Workshops BEFIT, pp. 2106–2112. (2011)
3. Bosch, A., Zisserman, A., and Munoz X.: Representing shape with a spatial pyramid kernel. Proceedings of the ACM International Conference on Image and Video Retrieval. (2007)
4. Ojansivu, V., Heikkil, J.: Blur insensitive texture classification using local phase quantization. Image and Signal Processing, ser. Lecture Notes in Computer Science. (2008)
5. Che, H.: Empirical Evaluation of Improved Casper Learning Algorithm on Emotion Recognition. College of Engineering and Computer Science Australian National University. (2020)
6. Treadgold, N.K., Gedeon, T.D.: A Cascade Network Employing Progressive RPROP. Int. Work Conf. on Artificial and Natural Neural Networks, pp. 733--742. (1997)
7. https://github.com/torch
8. Shin, M., Kim, M., Kwon, D.: Baseline CNN structure analysis for facial expression recognition. 25th IEEE International Symposium on Robot and Human Interactive Communication. (2016)
9. Nair, V., Hinton, G.E.: Rectied linear units improve restricted Boltzmann machines. In ICML, pp. 807. (2010)
10. Jung, Y., Hu, J.: A k-fold averaging cross-validation procedure,'' J. Nonparam. Statist., vol. 27, no. 2, pp. 167–-179. (2015)