Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230836

Comparing the effect of CNN and linear regression on facial expression recognition

Bohan Fang[1,3], Tongrui Gan[2]

[1]Jinling High School , Nanjing, China, zelangjian@whut.edu.cn
[2]Westsyde Secondary School, 855 Bebek Rd, Kamloops, BC V2B 6P1，Canada

[3]zelangjian@whut.edu.cn

Abstract: In Modern computer vision technology, facial expression recognition plays an important role. Many cases leverages this technology. This paper studies the Convolutional Neural Networks(CNN) in recognising the facial expression. We first show our research method to learn the model. Then we use the dataset to illustrate the efficiency CNN and Linear Regression. By showing the result of the two methods, we can clearly see the advantages of the CNN. We hope our research can help researchers have a better understanding about the CNN model.

Keywords: CNN, Linear Regression, computer vision.

1. Introduction

1.1. History of Machine Learning

Machine learning as actually been around for decades or, arguably, centuries [1]. Dating back to the 17th century, Bayesian and Laplace's derivation of least squares and Markov chains formed the tools and foundations of widely used machine learning. From 1950 (Alan Turing proposed to build a learning machine) to early 2000 (with practical applications of deep learning and more recent advances, such as AlexNet in 2012), machine learning has made great progress. Since the study of machine learning in the 1950s, machine learning has become a new discipline, integrating various learning methods, and a unified view of various basic problems of machine learning and artificial intelligence is also being formed. Machine learning is a common research hotspot in the field of artificial intelligence and pattern recognition, and its theories and methods have been widely used to solve complex problems in engineering applications and science. The winner of the Turing Award in 2010 was Professor Leslie Valliant of Harvard University. One of his winning works was to establish a probability approximate correct learning theory; the winner of the Turing Award in 2011 was Professor Judea Pearll of the University of California, Los Angeles, whose main work was Contribution to the establishment of artificial intelligence methods based on probability and statistics. These research results have contributed to the development and prosperity of machine learning.

Machine learning is a science that studies how to use computers to simulate or realize human learning activities[2]. It is one of the most intelligent and cutting-edge research fields in artificial intelligence.

DOI: 10.54254/2755-2721/6/20230836

Especially in the past ten years, the research work in the field of machine learning has developed rapidly, and it has become one of the important topics of artificial intelligence. Machine learning is used not only in knowledge-based systems, but also in many fields such as natural language understanding, non-monotonic reasoning, machine vision, pattern recognition, etc. Whether a system has the ability to learn has become a sign of "intelligence"[3]. The research of machine learning is mainly divided into two research directions: the first is the research of traditional machine learning, which mainly studies the learning mechanism, focusing on the exploration of the learning mechanism of simulated human; the second is the research of machine learning in the environment of big data. Research, this type of research mainly studies how to effectively use information, focusing on obtaining hidden, effective and understandable knowledge from huge amounts of data[4,5]. After 70 years of development, machine learning, represented by deep learning, draws on the multi-layered structure of the human brain, the layer-by-layer analysis and processing mechanism of neuron connection and interaction information, and the powerful parallel information processing capability of self-adaptation and self-learning. In many aspects Breakthrough progress has been achieved, the most representative of which is the field of image recognition.

1.2. History of Face Recognition

In the past, face recognition was a field that irrelevant to computer science. This field was developed by a group of psychologists to cure psychological diseases through watching patients' facial movements to discover their mental issues. However, with the development in this field, facial expressions became harder for human eyes to detect. As a result, people began to focus on building machines to help them to do face recognition. They use Euclidean Distance to capture people's face features. However, during this age, machines couldn't do face recognition by their own which means people are on the stage that called supervised learning. After a giant development in computer from 1990s, machines can do face recognition without interference from people, this was called unsupervised learning. A lot of convenient equipments were based on this such as face unlocking. With these advanced technology, people begin to research deeper in face recognition. For example, classification of human facial expression. There are various ways to increase the efficiency in interpersonal communication. For example, voice intonation, body language, and more complex methods such as electroencephalograph [6]. Among these, classifying facial expression is much more easier. Detecting facial expressions can help people spread human emotional information and deal with interpersonal relationships since we can understand the true emotion through people's face instead of only through words. According to research by psychologist A. Mehrabia, in the daily communication of human beings, language accounts for only 7% of the information transmission. In contrast, facial expression accounts for almost 55%! According to P. Ekman, there are seven kinds of emotions can be universally recognizable[7]: anger, disgust, fear, happiness, sadness, surprise, contempt. In addition, even people are from different culture, they can reach an agreement on this[8]. This means even if we can't understand some languages, we can still communicate with people from different countries at some aspects. But some people can't detect others emotions quickly through facial expressions. As a result, they will lose a lot of information. By using machine on facial expression classification, we can get more information from others. In some countries, research on micro-expression has been applied to national security, judicial system, medical clinical and political elections. In the field of national security, some dangerous people such as trained terrorists may easily pass the test of lie detectors, but through micro-expressions, we can easily find their true thoughts by analyzing their expressions, and because of this characteristic of micro-expressions, it also has a good clinical application in the judicial system and medicine. Filmmakers, directors or advertising producers can also predict the effect of promotional videos or advertisements through people's facial expressions. This is just as Human-Computer Systems Interaction: Backgrounds and Applications claims, machine facial expression recognition can be used to help medicine, marketing and entertainment[9].

1.3. Facial Expression Recognition

DOI: 10.54254/2755-2721/6/20230836

This paper uses the facial expression data set of the kaggle website(www.kaggle.com). After removing the pictures without facial expressions in the data set, there are 28661 expression pictures in the training set, 7171 expression pictures in the test set, and there are 7 categories of expression pictures, namely angry, disgusted, fearful, happy, neutral, sad, surprised[10]. This paper's goal is to test a machine learning model using convolutional neural networks and another using the linear regression technique and use the training set expression pictures for supervised learning, and use the test set expression pictures for training. Then, we will compare the two models and give predictions.

2. Method

In this section, we show Linear Regression and CNN in detail.

2.1. Linear Regression

The linear model essentially obtains the mapping f from the attribute x to the label y, such that y=f(x,w) (where w is the model parameter) that uses optimization theory to get the ideal value. The equation below may be used to express linear regression when there is just one variable: y = w*x+e, when there are multiple variables: for example, in MNIST, each pixel of the picture is a feature attribute, and each picture has 28(width)*28(height))*1(channel)=784 feature attributes, then y=w1*x1+w2*x2+...+w784*x784+e, another example is the facial expression image of this paper, each image has 48*48*1 =2304 feature attributes, then y= w1*x1+w2*x2+...+w2304*x2304+e, in multi-classification tasks, the output y is generally a vector, called the output vector, and the output vector's dimension is equal to the categorization The number of classifications for, and its element value is a real number, indicating the score of the corresponding category. The closer the sample is to the respective category, the higher the score. The final projected category is the one that corresponds to the output vector's element with the highest value. There are 7 expression classifications in this paper. In order to obtain 7 scores, 7 linear functions are required, which are:

$$\begin{aligned} y1&=w11*x1+w12*x2+...+w1i*xi+e1 \\ y2&=w21*x1+w22*x2+...+w2i*xi+e2 \\ y7&=w71*x1+w72*x2+...+w7i*xi+e7 \end{aligned} \quad (1)$$

If the above variables are represented by a matrix, the linear function can be expressed as y=w*x+e, where the matrix dimension of w is [2304*7], the matrix dimensions of x and e are [1*2304] and [1*7], respectively, and the matrix dimension of the output matrix y is [1*7],so the linear model now becomes the most basic linear model.

The least squares function known as a linear regression equation is used in linear regression to describe the connection between one or more independent and dependent variables. A linear combination of one or more model parameters known as regression coefficients makes up such a function. Simple regression refers to the situation when there is only one independent variable, while multiple regression refers to the situation where there are numerous independent variables.

The first regression analysis technique that underwent in-depth research and had widespread usage in actual applications was linear regression. This is because models that rely linearly rather than nonlinearly on their unknown parameters are simpler to fit, and the statistical features of the resultant estimates are simpler to ascertain.

When we run the program, we find that the running speed of linear regression is very slow, and it takes a long time to make judgments. This also prolongs the duration of the whole treatment. What makes us even more incredible is that the learning performance of linear regression is very low, only 13% accuracy.

2.2. Convolutional Neural Network

Only the linearly separable training set can the linear model successfully classify objects. For the linearly inseparable training set, a neural network can be used to achieve a better classification effect. The main goal is to incorporate a nonlinear transformation, a multi-layer structure, and a nonlinear activation function, the formula is:

DOI: 10.54254/2755-2721/6/20230836

$$\begin{aligned} a &= xW1 \\ h &= f(a) \\ y &= hW2 \end{aligned} \tag{2}$$

Among them, f is the nonlinear function acting on the vector a, which is operated element by element, not the overall operation of the vector a. The components of the vector an are known as neurons, the nonlinear function f is known as the activation function, and h is known as the hidden layer.

Each neuron in a typical neural network is linked to every neuron in the layer before it (fully connected).Consider the images in this article as an example, if using the full connection mode, each neuron in the first hidden layer has 48*48*1=2304 weights, and the hidden layer contains a large number of neurons, a high number of parameters may result in network overfitting, therefore the hidden layer's weights will be enormous, large amounts of memory are needed to store a lot of weights, and the calculation will be slow, these all limit the application of regular neural networks in image classification and recognition. To more effectively address the classification and image recognition problems, it is necessary to implement local feature extraction, local connection, parameter sharing and deep network on the regular neural network. Additionally, these are the crucial components of a convolutional neural network.

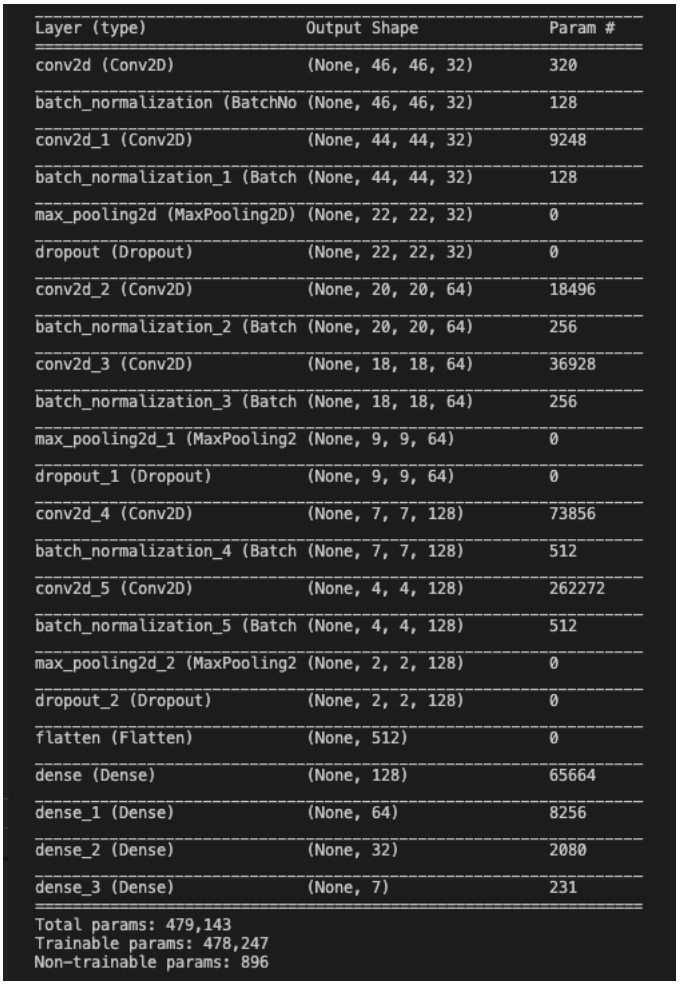
One of the typical deep learning algorithms is the convolutional neural network (CNN), a kind of feedforward neural network with a deep structure that incorporates convolutional computing. Convolutional neural networks are also known as "translation invariant artificial neural networks" since they have the capacity for representation learning and can categorize the input data based on its hierarchical structure.

Time delay network and LeNet-5 were the first convolutional neural networks developed as a result of convolutional neural network research in the 1980s and 1990s. Deep learning theory and numerical computation were introduced in the 21st century. Convolutional neural networks have progressed quickly as a result of equipment advancement and have been used in industries like computer vision and natural language processing.

The convolutional neural network can do both supervised and unsupervised learning since it is built by replicating the biological visual perception process. The convolutional neural network may employ a modest computation since the hidden layer's convolution kernel parameter sharing and the connections between the layers are sparse. Lattice features, like pixels and sounds, may be learned quantitatively with consistent effects and no further feature engineering needs for the data.

The input layers, hidden layers, and output layers make up the three types of layers that make up the convolutional neural network as a whole. In this paper, six convolutional layers are used, with the number of filters being 32, 32, 64, 64, and 128. The first five layers' convolution kernels are 3X3 in size, while the last layer's convolution kernel is 4X4 in size, a BN layer follows each convolutional layer, and after every two convolutional layers, a pooling layer and a dropout layer are added., and finally a Flatten layer and four Dense layers are used for connection output, and the activation function is Relu, the model structure is shown in the following Figure 2.

DOI: 10.54254/2755-2721/6/20230836



Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 46, 46, 32)	320
batch_normalization (BatchNo	(None, 46, 46, 32)	128
conv2d_1 (Conv2D)	(None, 44, 44, 32)	9248
batch_normalization_1 (Batch	(None, 44, 44, 32)	128
max_pooling2d (MaxPooling2D)	(None, 22, 22, 32)	0
dropout (Dropout)	(None, 22, 22, 32)	0
conv2d_2 (Conv2D)	(None, 20, 20, 64)	18496
batch_normalization_2 (Batch	(None, 20, 20, 64)	256
conv2d_3 (Conv2D)	(None, 18, 18, 64)	36928
batch_normalization_3 (Batch	(None, 18, 18, 64)	256
max_pooling2d_1 (MaxPooling2	(None, 9, 9, 64)	0
dropout_1 (Dropout)	(None, 9, 9, 64)	0
conv2d_4 (Conv2D)	(None, 7, 7, 128)	73856
batch_normalization_4 (Batch	(None, 7, 7, 128)	512
conv2d_5 (Conv2D)	(None, 4, 4, 128)	262272
batch_normalization_5 (Batch	(None, 4, 4, 128)	512
max_pooling2d_2 (MaxPooling2	(None, 2, 2, 128)	0
dropout_2 (Dropout)	(None, 2, 2, 128)	0
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 128)	65664
dense_1 (Dense)	(None, 64)	8256
dense_2 (Dense)	(None, 32)	2080
dense_3 (Dense)	(None, 7)	231

Total params: 479,143
Trainable params: 478,247
Non-trainable params: 896

Figure 1. Figure with CNN module

2.2.1. Convolutional Layer. The purpose of the convolution layer, which includes many convolution kernels, is to conduct feature extraction on the input data. The size of the region, known as the "receptive field" in the literature, depends on the size of the convolution kernel, with each neuron in the convolutional layer connecting to several neurons in the neighboring area in the preceding layer, and its meaning can be Analogy to the receptive field of visual cortex cells. When the convolution kernel is working, it will regularly scan the input features, perform matrix element multiplication and summation on the input features in the receptive field, and superimpose the deviation. This paper uses 6 convolutional layers,the convolution kernel sizes for the first five convolutional layers are 3X3, the final convolutional layer is 4X4, and there are 32, 32, 64, 64, 128, 128 filters in total.

2.2.2. Pooling Layer. The pooling layer is often used to minimize the input's spatial size following a convolution layer. Each depth slice of the input volume receives an individual application of it. Volume depth is always preserved in pooling operation. In this paper, a pooling layer is added after every two convolutional layers to purify neurons, and some neurons are removed. The local window size of the pooling layer is 2X2.

2.2.3. Dropout Layer. Overfitting of the model may be avoided by adding dropout layers, dropout works by randomly setting the outgoing edges of hidden units (neurons that make up hidden layers) to 0 at each update of the training phase. In this paper, each pair of convolutional layers is followed by a dropout layer.

DOI: 10.54254/2755-2721/6/20230836

2.2.4. BN Layer. After the multi-layer transformation of the network, the data is no longer normalized, that is, the mean is not 0, and the variance is not 1,this makes network's hidden layers' ability to learn more challenging, adding the BN layer can normalize the input data of the hidden layer and improve the convergence effect and learning speed, this has the effect of stabilizing the learning process and dramatically reducing the number of training epochs required to train deep networks.In this paper, each convolutional layer is followed by a BN layer so that the output data of each convolutional layer is re-distributed normally.

2.2.5. Flatten Layer. Flatten layer is used to flatten the input. For example, if flatten is applied to layer having input shape as (batch_size, 2,2), then the output shape of the layer will be (batch_size, 4),the flatten layer is usually used to flatten the data before making a full connection. In this paper, a flatten layer is also added before the fully connected layer.

2.2.6. Dense Layer. A dense layer is a typical highly connected neural network layer in which every neuron in the layer is linked to every neuron in the layer above it. It is most common and frequently used layer.

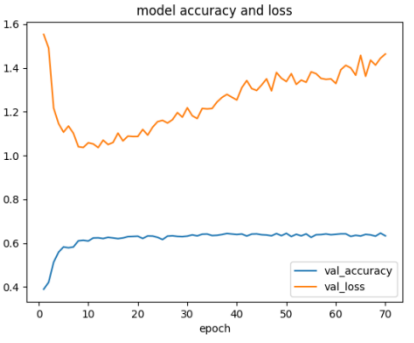
2.2.7. Activation Function. The neural network's activation function, which is utilized to compute the input vector element by element, is its important component. If the output of each neuron passes through a nonlinear function, then the model of the entire neural network is no longer linear,Commonly used activation functions are sigmoid, softmax, Relu, tanh, etc,We use sigmoid, softmax, Relu functions for learning and testing respectively,according to the expression recognition learning results in this paper, the accuracy obtained by using the Relu function is significantly higher than that of sigmoid and softmax.

3. Results

Figure 2 shows our result. From the results, it can be seen that at the level of accuracy, our own CNN score is 64%, while the results of linear regression is 14%. The score of our favourite model is almost five times that of linear regression and 2.5 times that of AlexNet. At the same time, the number of errors in our own CNN is less than the correct number, especially in the expression of joy. However, the cost of higher accuracy is also very high. The training time for linear regression is only about 30 seconds, and the test time is only three seconds. Compared with linear regression, AlexNet's training time has reached 10 to 20 minutes with a 10% increase in accuracy. With the great improvement of accuracy, the training time of our own CNN has also increased, which has been about two hours. From this, we can conclude that although CNN is significantly better than linear regression in facial expression recognition, the time it consumes is also multiplied. We hope to continue to optimize this self-built CNN in the future, so that it can achieve higher accuracy in a shorter time.

	precision	recall	f1-score	support
fearful	0.00	0.00	0.00	1024
disgusted	0.02	1.00	0.03	111
angry	0.00	0.00	0.00	958
neutral	0.00	0.00	0.00	1233
sad	0.00	0.00	0.00	1247
surprised	0.00	0.00	0.00	831
happy	0.00	0.00	0.00	1774
accuracy			0.02	7178
macro avg	0.00	0.14	0.00	7178
weighted avg	0.00	0.02	0.00	7178

Linear Regression



CNN

Figure 2 The comparsion result

DOI: 10.54254/2755-2721/6/20230836

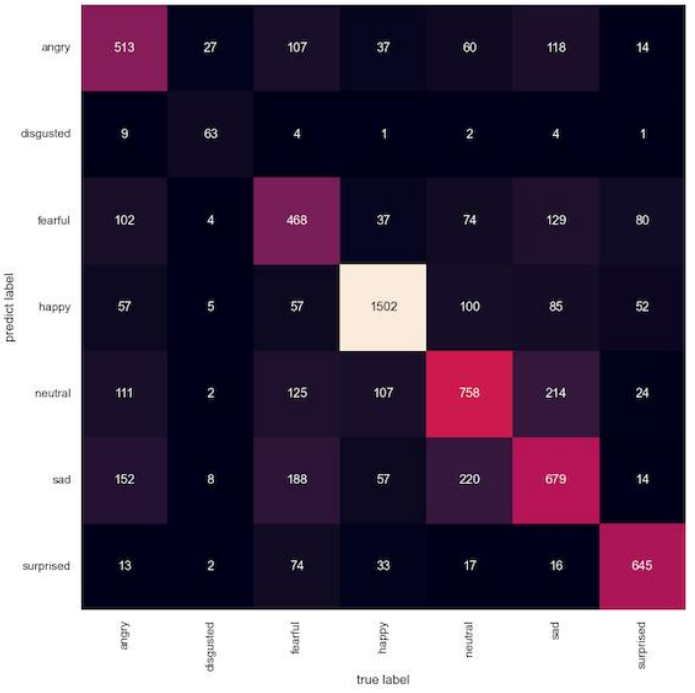


Figure 3. Confusion matrix of CNN model

4. Conclusion

Our work is to test which model can work better in facial expression recognition. As the work progresses, a better model can be gradually selected in this field to realize the routinization of expression recognition. Teachers can judge whether students are confused about concepts with the help of machine expression recognition; bosses can know whether employees are enthusiastic about their work; interrogators can use micro-expression recognition by machines to help judge the authenticity of criminals' languages; and psychologists can use machine assistance to capture the imperceptibility of patients. The expression can better help the treatment.

References

[1] Yan W J, Wu Q, Liu Y J, et al. CASME database: A dataset of spontaneous micro-collected from neutralized faces[C]. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. IEEE, 2013:1-7.

[2] Yan W J, Li X, Wang S J, et al. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation[J]. Plos One, 2014, 9(1):e86041.

[3] Xiang J, Zhu G. Joint Face Detection and Facial Expression Recognition with MTCNN[C]. International Conference on Information Science and Control Engineering. IEEE Computer Society, 2017:424-427.

[4] Cootes T F, Edwards G J, Taylor C J. Active Appearance Models[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2001, 23(6):681-685.

[5] P. Abhang, S. Rao, B. W. Gawali, and P. Rokade, "Article: Emotion recognition using speech and eeg signal a review," International Journal of Computer Applications, vol. 15, pp. 37–40, February 2011. Full text available.

[6] P. Ekman, Universals and cultural differences in facial expressions of emotion. Nebraska, USA: Lincoln University of Nebraska Press, 1971.

[7] P. Ekman and W. V. Friesen, "Universals and cultural differences in the judgements of facial expressions of emotion," Journal of Personality and Social Psychology, vol. 53, 1987.

DOI: 10.54254/2755-2721/6/20230836

[8] A.Kołakowska, A.Landowska, M.Szwoch, W.Szwoch, and M.R.Wro´bel,Human-Computer Systems Interaction: Backgrounds and Applications 3, ch. Emotion Recognition and Its Applications, pp. 51–62. Cham: Springer International Publishing, 2014.

[9] Wang Y, See J, Oh Y H, et al. Effective recognition of facial micro-expressions with video motion magnification[J]. Multimedia Tools & Applications, 2016:1-26.

[10] Xia Z, Feng X, Peng J, et al. Spontaneous micro-expression spotting via geometric deformation modeling[J]. Computer Vision & Image Understanding, 2016, 147(C):87-94.