Multi-attribute Recognition,the Key to Universal Neural Network

Jinxin Wei, Qunying Ren

***Abstract*—To achieve the recognition of multi-attribute of object, I redesign the mnist dataset, change the color, size, location of the number. Meanwhile, I change the label accordingly. The deep neural network I use is the most common convolution neural network. Through test,we can conclude that we can use one neural network to recognize multi-attribute so long as the attribute difference of objects can be represented by functions. The concrete network(generation network) can generate the output which the input rarely contained from the attributes the network learned. Its generalization ability is good because the network is a continuous function. Through one more test, We can conclude that one neural network can do image recognition, speech recognition,and nature language processing and other things so long as the output node and the input node and more parameters add into the network. The network is universal so long as the network can process different inputs. I guess that the phenomenon of synaesthesia is the result of multi-input and multi-output. I guess that connection in mind can realize through the universal network and sending the output into input.**

***Index Terms*—Computer vision, multi-attribute, deep neural network, multi-dimension, data processing, universal neural network, parallel processing, speech recognition, nature language processing**

THERE are many multi-label learning examples which contain many labels and networks. I design a single label and a single network to solve the multi-attribute problem.

Because we don't have the dataset fit for my task, so I redesign the mnist dataset. Because the visual attributes of object recognized by mankind are color, size, location, shape, texture, quantity, pattern, so we choose the color, size, location, and shape attributes as example. Because the mnist dataset already has the shape attribute, so we only need to add color, size, location. First, we change the color. Because the color represented by computer is mixed by red, green and blue, so we change the number's color to red, green and blue. We assign the gray's pixel data to red channel, green channel, blue channel separately, the other two channels are zero. The background is all 255. Secondly, we change the size. Shrinking the image to size 18×18, then put the pixel to the white background of size 28×28. When we put the pixel to up part of the background, we change the location. Next we change the label. We use the label form similar to the label form used in classification. Class 0-9 is one hot encoding. For example, 0100000000 is 1. Because there are 3 colors, so the red color is 100, green is 010,blue is001. The index of color label is 10-12. Because there are two sizes which are big and small, so the code is 01, 10 and the index is 13-14. Because there are two locations which are up and middle, so the code is 10, 01 and the index is 15-16.So far,we finish the dataset's processing work. The order of label is number, color, size, location. For example, 01000000001000101 represents big middle red 1. Why we use one hot encoding? Because each class has one output, we can generate the multi-output by regression, and the output is from 0 to 1 which is similar to data normalization.

Now we design the network. The experiment is done by tensorflow framework. The regression network have 3 convolution layers[1] and 2 fully connected layers, no maxpooling and padding, activation function is leaky relu[2]. Because they will cause generation loss. The generation network is the inverse function[4]of regression network. You can read my another paper named "A Functionally Separate Autoencoder", which describes the detail of generation from label to concrete information. When training the regression network,loss function is mean squared error,optimizer is adam[3], metrics is accuracy[3]. I take np.argmax()[3] of prediction[:,0:10], [:,10:13], [:,13:15], [:,15:17] separately, and the real label processes the same way. The regression values become index type through this, and the accuracy can be calculated by comparing the data. When training the generation network, loss function is mean squared error, optimizer is adam, metrics are mean absolute error and cosine similarity, input is multi-attribute label, output is image. So, let's see the result which is shown in Fig.1 (a) and Fig. 1(b).

From Fig. 1(a) and Fig. 1(b), we can see that the labels of the regression are all right, the generation is ok although it is a little blurry. Through test, we can conclude that we can use one neural network to recognize multi-attribute so long as the attributes can be represented by functions.

Now let's modify the label to some degree, then see the generation. First, I change the color to [0.5,0.5,0], [0.5,0,0.5], [0,0.5,0.5], [0.3,0.3,0.3], [0.5,0.2,0.3], [0.2,0.3,0.5], [0.3,0.5,0.2].The result is shown in Fig.2. Second, I change the size to[0.5,0.5], [0.3,0.7], the result is shown in Fig.3. Last, I change the location to[0,1.05], [1.05,0], [0.5,0.5], [0.55,0.45], the result is shown in Fig.4. Why the sum is equal to1? Because the label is one hot encoding and the attribute is exclusive, we can't make one thing big and small simultaneously. The

F. A. Author,Jinxin Wei, Vocational School at Juancheng, 274600, China(e-mail: wjxabai@163.com).

S. B. Author,Qunying Ren, Bureau of Emergency Management of Juancheng, 274600, China(e-mail:renqy1987@163.com)

connection can't make the output to 1 simultaneously.

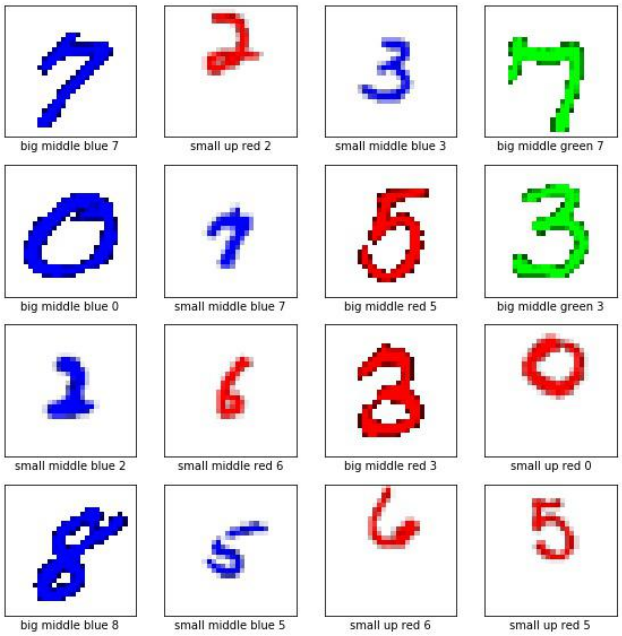

Fig. 1(a) Regression Effect on Test Data (image is input,label is regression result)

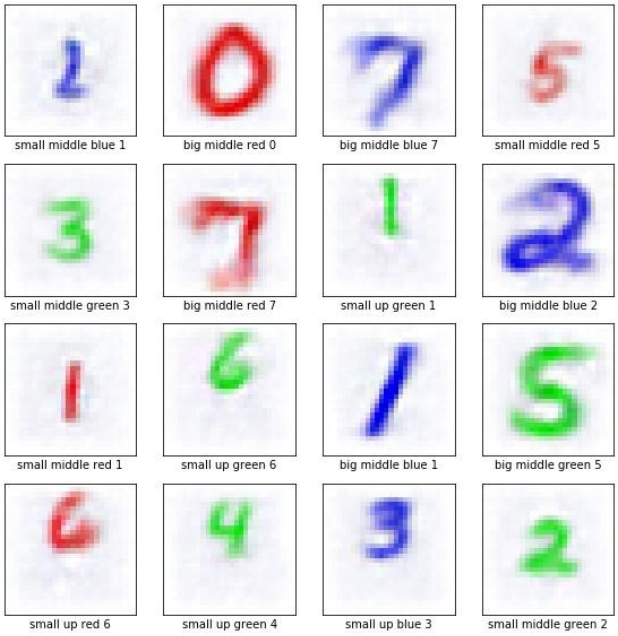

Fig. 1(b) Generation Effect on Test Data (image is generation result,label is input)

From the pictures we can see that we can generalize the rare situation through the attribute we learned.Although we just use three-primary colors, but we can generate other colors. We can generate the medium state of big and small,up and middle.Its generalization ability is good because the network is a continuous function.

Inspired by multi-attribute and computer input/output principle and the difference between speech recognition and image recognition can be represented by function, I design a network to process speech recognition and image recognition simultaneously.It is shown in Fig.5.The multi-attribute network is one input, several outputs, but this network is several inputs and several outputs.

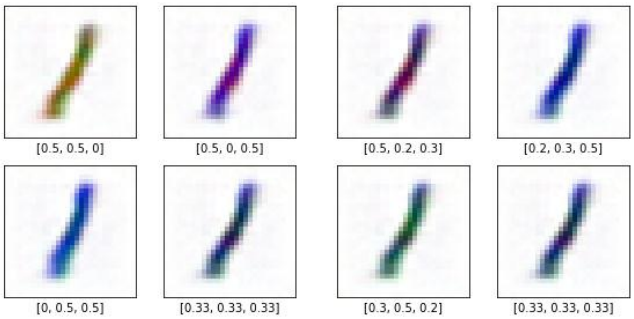

Fig. 2. Effect of Color Change

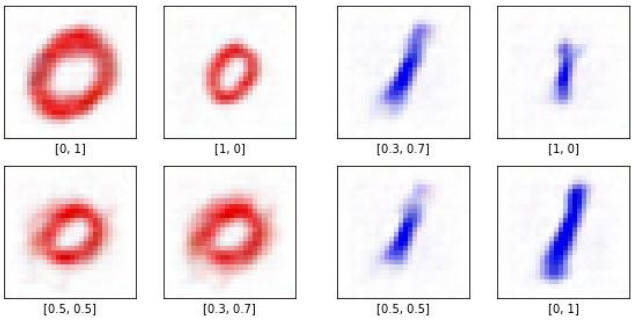

Fig. 3. Effect of Size Change

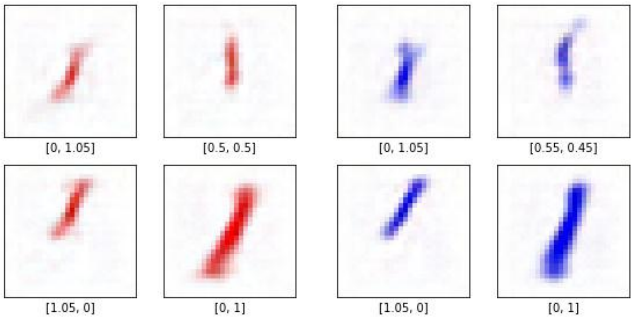

Fig. 4. Effect of Location Change

The dataset is mnist and some downloaded speech data.The amount of speech data is small because I can't find larger dataset. The two task are all classification task, I change them to regression. I concatenate the mnist and the speech data as the input,the labels of the two datasets are also concatenated. For simple, I choose the fully connected network which has 5 layers for my task(When using convolution neural network,the input must be processed to fit the CNN). The first four layers have 60 neuron units each, the last layer has 20 because there are twenty classes(mnist:10 classes, speech dataset:10 classes). The label is one hot encoding. When training it, loss function is mean squared error, optimizer is adam, metrics is accuracy(calculate the two type tasks separately). Because the shape of mnist image is (28,28) and the shape of audio is (20,11), and there are 20477 audio samples which split into train data and test data by the ratio of 0.7. So the shape of train data is (14333,1004) and the shape of test data is (6144,1004). The shape of train label is (14333,20) and the shape of test data is (6144,20). Through test, the accuracy of image regression on test dataset is 86.5% and the the accuracy of speech regression on test dataset is 79.8%. The accuracy will be higher when the dataset is larger. Because the samples of mnist are larger than the audio data, so we can

input images only and make the audio input to -1. If it is 0, this will affect the image data which is 0-1, then this will affect the accuracy of image regression. We can add one output which indicates no input of audio when output is 1 and there is input of audio when output is 0. By adding these changes, the accuracy of image regression is also 86.6%. So the universal network can also process single input. It is important to note that the accuracy of image and audio on test data is 10% less than the accuracy of image and audio on training data. Through test, we can conclude that one network can do two(I don't test more than two) tasks simultaneously, the parameters are shared by the two tasks, so it is parallel processing. I guess it can do more than two tasks simultaneously, so it is universal neural network. When it does several tasks simultaneously, the understanding ability increases. Because we can make decision base on muti-dimension information, so it is more intelligent. I guess that the phenomenon of synaesthesia is the result of multi-input and multi-output, because I don't have other sensory data. Because the parameters of network shared by the multi-input and multi-output, so one input may affect other output when the two outputs are the attributes of some similar object, such as fire and sun which have red color and hot attributes. For example, When we see red(vision input), we may feel warm(warm sensation output), however When we see blue, we may feel cool.

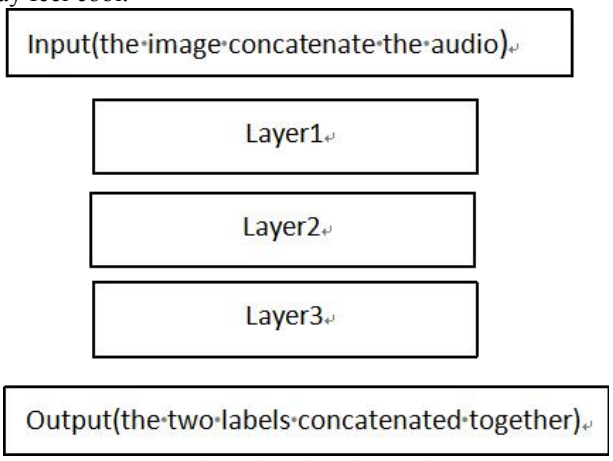

Fig.5 The universal network

Because we can also use the convolution neural network to process sequence data and nature language processing is also represented by functions, so nature language processing can use the same network as image recognition and speech recognition so long as the output node and the input node and more parameters add into the network. The network is universal so long as the network can process the different inputs. We can first send image or speech to the network and the output is language or other information, then send the language or information to the network's input, process it or predict it. I guess that connection in mind can realize through the universal network and sending the output into input. For example, it's cloudy now. This information is sent to the network's input,the predict output is there will be a rain. This output is sent to network's input, the predict output is that I need an umbrella. Because the information about it's cloudy and there will be a rain is represented by image or audio or language, so only universal network can do it. The most likelyhood event can be seen as the effect of the cause. It's causal reasoning.

Conclusion

1.The abstract network(prediction network) achieves the intended functionality of multi-attribute recognition through muti-dimension regression. The concrete network can generate the output which the input rarely contained from the attributes the network learned. Its generalization ability is good because the network is a continuous function.

2.We can use one neural network to do image recognition, speech recognition, nature language processing and other things simultaneously so long as the output node and the input node and more parameters add into the network. The network is universal so long as it can process different inputs. I guess that the phenomenon of synaesthesia is the result of multi-input and multi-output. I guess that connection in mind can realize through the universal network and sending the output into input.

REFERENCES

[1]Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning," USA:MIT Press, 2016, pp. 330-370.

[2]Andrew L. Maas, Awni Y. Hannum and Andrew Y. Ng, "Rectified Nonlinearities Improve Neural Network Acoustic Models," ICML, 2013

[3]Tensorflow Tutorials and Apis, google.inc

[4]Jin-xin Wei, Qun-ying Ren, "A Functionally Separate Autoencoder," unpublished