1-17

Deep Learning Imitating Relationship between Ears and Brain for Temporal Data

Shinsaburo KITTAKA Yoko UWATE Yoshifumi NISHIO
(Tokushima University)

1. Introduction

Artificial intelligence (AI) is hot topic among not only researchers but also the public. Because AI is hoped as new labor force. Japanese cabinet publicly disclosed the prediction that japanese population is decreasing and drops to below 100 million by the year 2050. From this prediction, we expect labor force to become decreasing. However, computer capability is improving on Moore's Law without population and labor force decreasing. So AI will serve as labor force in foreseeable future. And now, most of AI discriminates object from images. Because current AI is focused on image processing, and the structure is similar to relationship between eyes and brain.

In this study, we set up the structure of deep learning similar to relationship between ears and brain. And it discriminates not image that has instantaneous data but music that has temporal data. We hope AI to expand the range of labor force.

2. Proposed method

Deep learning consists of input layers, convolutional layers, pooling layers, fully connected layers and output layers. Convolutional layers make many layers from input layers and augment feature quantity. Pooling layers have filter and pick up maximal value. Fully connected layers connect every neuron in one layer to every neuron in another layer. We focus on pooling layers. Usual pooling layers pick up and transmit maximum value. We change the value to the difference of maximum value and minimum value as Fig. 1.

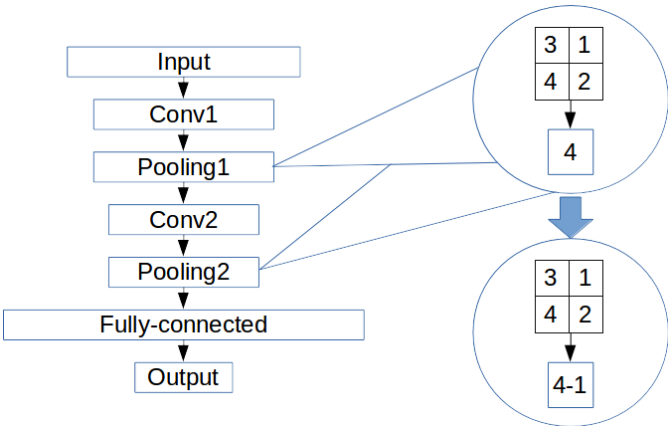

Figure 1: Changing system in pooling layers.

The ear uses hair sells to convert fluctuation to electrical signals. The firing of hair sell is based on displacement of hair. We consider that it is the most difference between eyes and ears. So we change pooling layers as proposed method.

3. Simulation results

In simulation, network discriminates music scales. We prepare data set for learning and test from youtube (URL: https://www.youtube.com/watch?v=IVdp0uBdrMM). We convert MP3 files from youtube to WAV files with 44,100 Hz sampling frequency. The number of data is 1,360,291. And we divide data set into each 5000 data.

The amplitude of waveform means just sound volume. It is not necessary to discriminate music scales. So we normalize the maximum value of data to be 1 in each data. Finally, we change all data to positive. Because the activation function in network outputs 0 if input is negative. Figure. 2 shows the activation function in network.

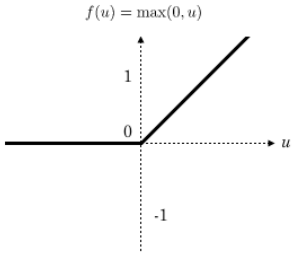

Figure 2: Rectified Linear Unit.

The number of input data is 5000 data in a single learning, and we input data in line. We define the method that is picking up maximum value in pooling layer as conventional method.

Table 1: The discriminant accuracy

	conventional	proposed
accuracy[%]	82.1	83.4

From Table 1, the proposed method is only a little bit better than the conventional method.

4. Conclusions

We change pooling system in pooling layer in deep learning for temporal data. Conventional pooling layers pick up and transmit maximum value. We change the transmitting value to the difference of maximum value and minimum value as proposed method. Because the ear converts fluctuation to electrical signal based on displacement of hair.

In simulation result, the proposed method obtains a little bit better accuracy than the conventional method. We consider that the difference of maximum value and minimum value has more information than maximum value for temporal data.

In future work, we make networking component more and more like the relationship between ear and brain for discriminating temporal data.