Performance of Convolution Neural Network on the Recognition of Speech Emotion and Images

Bin Zhang
Computer and Information, Hefei University of technology
Hefei, 230009, China
zhangbinh@yeah.net

Changqin Quan
Graduate School of System Informatics
Kobe, 657-8501, Japan
quanchqin@gmail.com

Fuji Ren
Institute of Science and Technology
Tokushima, 770-8506, Japan
ren@tokushim-u.ac.jp

Received (28 June 2015)
Revised (10 Oct. 2015)

In this paper, the performance of Convolution Neural Network (CNN) in image recognition and emotion recognition in speech will be presented. Feature extraction and selection in pattern recognition is an important issue and have been frequently discussed. Moreover, two-dimensional signals such as image and voice signals are hard to be modelled well by traditional models like SVM. The ability of CNN to characterize two-dimensional signals is prominent. And CNN can adaptively extract feature to eliminate the dependence on human subjectivity or experience. It mimics the effect of local filtering in visual cortex cells to dig local correlation in natural dimensional space. In this work, for the problems of the image recognition and emotion recognition in speech, CNN and SVM which is used as baseline for comparison of the recognition effect. Different kernel functions in SVM have been experimented for image recognition with, the best accuracy is 94.17%. However, the accuracy of using CNN is 95.5% (7291 pictures for train and 2007 pictures for test) with less time consuming. In the emotion recognition of speech, the accuracy of CNN is 97.6% corresponds to 55.5% by baseline model (4000 utterances for training, 1500 for validation, 500 for test). The experimental results showed that CNN can effectively extract features and its modeling capability for two-dimensional signals is prominent.

Keywords: Convolutional Neural Network, Emotional speech recognition, Image recognition.

1. Introduction

Image and voice are the most direct and most natural channels that people acquire information. If achievements in these two fields are applied on robots that can greatly improve the intelligence of the machine. In practice, in image recognition and speech
Performance of Convolution Neural Network

recognition we will encounter the feature selection problem [1] [2]. Common image features are composed of color feature, texture feature, shape feature, spatial relations characteristics [3]. The commonly used features in speech recognition are MFCC (Mel Frequency Cepstrum Coefficient), prosodic features, sound quality characteristics and acoustic features [4] [5]. Sometimes in order to acquire a better final result, these characteristics also be integrated appropriately [6]-[8].

Selection and integration of these features require human experience and subjective judgment so the result on these feature sets are not good enough. Therefore, in order to better identify emotions or image a large number of features are added. In emotion recognition in speech, to identify the emotion in the voice, the linguistic information [9] or emotional point information [7] are also added, such as context [10]-[12], keywords, etc. Typically, this will cause improvements of the recognition rate. However, there is also the phenomenon that the information of voice and text interfere with each other to influence the judgment. Besides, emotion is an important aspect of intelligence. The problem that we want to make the system to distinguish emotion leads scholars have to resort to our own emotional cognitive system. Bionics, biology have been used to detect emotion in the voice [13]. They utilize the physical structure of our human ear to generate suitable features for human ears perceived characteristics, such as MFCC, Lyon cochlear model [14], or to improve the model to enhance recognition performance. These methods are often easier for people to trust and adopt for they can be better understood. Due to more fully take into account the physical structure of the human ear, mimicking the physiological activity, the adaptability and stability of the system built by this method are better than general system. Such a large feature set leads to the use of dimensionality reduction methods, such as PCA [15], Fisher and the like. But the final result is not satisfactory [4].

And for multi-class emotional speech signal, no classification tools is especially suitable for multi-class classification. Thus, optimized classification tools and perfected classification strategies [16] may improve the accuracy of emotion recognition. But experiments show that this strategy is not robust and there is limited room for improvement [4] [16].

In this study, traditional model SVM in the image recognition and speech emotion recognition is as a baseline system. And CNN is a specially designed multi-layer perceptron to identify two-dimension shapes. Therefore dimensional information retained in waveform points is effectively utilized by CNN. CNN model due to its characteristics of adaptive feature extraction, it is applied for image recognition and emotion recognition in voice signals. In the emotional speech recognition, based on the test of two classic characteristics of the speech signal, we propose that directly use waveform points to characterize the emotional speech signals. It neither loss information, but also take advantage of the natural correlation information between the waveform to identify emotion. In image recognition, SVM and CNN models are used for image recognition. And we compare the recognition result before and after PCA.

The rest of the paper is organized as follows. In Section 2, we present the traditional
model SVM and the newly test model CNN. Experiments and the results are provided in Section 3. Finally, we summarize our findings in Section 4.

2. Tradition model and CNN

A. SVM

The baseline model chosen in this experiment is SVM, and Kernel function type are linear kernel, Polynomial kernel, sigmoid kernel, RBF (Radial Basis Function). SVM is designed for classification of two types but its performance of the multi-classification is poor in actual problems.

When choose C-SVC model, the decision function is

\[ P_{label} = \text{sgn}\left( \sum_{i=1}^{n} \alpha_i \cdot K(x_i, x) + b \right) \]  

(1)

where \( n \) is the number of support vector, \( K(x_i, x) \) is the kernel function and \( b \) is a constant.

When RBF kernel function is selected, the kernel function \( K(x, x) = \exp(-\gamma \| x - x \|^2) \) is

(2)

So the decision function in our problem is

\[ P_{label} = \text{sgn}\left( \sum_{i=1}^{n} \alpha_i \cdot \exp(-\gamma \| x_i - x \|^2) + b \right) \]

(3)

Similarly, the linear kernel \( K(x, x) \) is

\[ K(x_i, x) = x_i \cdot x \]

(4)

Polynomial kernel \( K(x, x) \) is

\[ K(x_i, x) = ((x_i \cdot x) + 1)^d \]

(5)

When selecting kernel function to solve practical problems, commonly used methods are as follows: First using a priori knowledge of experts to preselect kernel; and then using the Cross-Validation method, namely carrying a nuclear function selection and trying different kernels respectively. The kernel with the smallest induction error would be the best kernel.

B. CNN

Convolutional Neural Network 17 is a multi-layer neural network, each layer is composed by a plurality of two-dimensional plane, and each plane is composed by a plurality of individual neurons. The structure of LeNet-5 is shown in Fig. 118. Sparse, convolutional layers and max-pooling are the key of LeNet models 19. From Fig.1 there are five layers in LeNet5: 2 convolution layers, 2 sub-sampling layers and 1 fully connected MLP layer. Alternating convolution and max-pooling layers are composed of the lower-layers. The upper-layers however are fully-connected traditional MLP (hidden layer + logistic regression).
Performance of Convolution Neural Network

B.1 Sparse Connectivity

The schematic diagram of sparse connectivity is shown in Fig.2. The input of N-th layer is a subset of (N-1)-th layer. By strengthening a local connection between neurons in adjacent layers CNNs exploit spatially-local correlation. Such a structure is similar to a local filter which is able to generate the strongest response to the input pattern. However, if the increasing layers in the figure above it will lead to a non-linear filtering which responds a bigger space.

C. Image Recognition and Emotion Recognition in Speech

In pattern recognition perspective image recognition and emotion recognition in speech can be abstracted into the following form:

Specific to the field of image recognition and emotion recognition in speech, the process can be summarized as follows:

![Pattern Recognition Systems](image_url)

![Four steps of image recognition](image_url)
As shown in Figure 3 and Figure 4, their methodological framework are similar. They both go through feature extraction and screening, and these steps are critical to the entire recognition. Another key issue is that the ability of model to reconstruct signals. CNN used in this paper is able to solve the two key issues at the same time. The details will be shown in Section 3.

![Fig.4.Traditional method of emotion recognition in speech.](image)

### 3. Experiment

**A. Data preparation**

All experiments on speech were conducted on USPS dataset (The US Postal handwritten digit dataset). It is a frequently used data set in machine learning, artificial intelligence, and data mining. This dataset (scaled to [-1:1] instead of [0:2]) also appears in the book “The elements of statistical learning” 20 21. The dataset consists of 7291 pictures for train and 2007 pictures for test.

All experiments on speech were conducted on the emotion speech database recorded by Chinese Academy of Social Sciences (CASS). There are four speakers including two females and two males. Five emotions were included, they are angry, fear, happy, sad, and surprise. Each emotion corresponds to 1200 different utterances, a total of 1200*5 utterances. The signals are sampled at 16 kHz and transcribed in mono. Each sampling point is represented with 16bit. The utterances are between one or two seconds to reserve prominent parts in emotions.

**B. Experiment environment**

CNN is set as follows in emotion recognition in speech:
Table 1. The Set of CNN in Emotion Recognition in Speech

<table>
<thead>
<tr>
<th>Layer</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input level</td>
<td>100*100</td>
</tr>
<tr>
<td>Pool size</td>
<td>200 feature map, convolution window size : 5<em>5, pool window size: 2</em>2</td>
</tr>
<tr>
<td>Full connection layer 1</td>
<td>Hidden neurons:500</td>
</tr>
<tr>
<td>Full connection layer 2</td>
<td>Hidden neurons:500</td>
</tr>
<tr>
<td>Output layer</td>
<td>6 classification output</td>
</tr>
</tbody>
</table>

Table 2. The Set of CNN in Image Recognition

<table>
<thead>
<tr>
<th>Layer</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input level</td>
<td>16*16</td>
</tr>
<tr>
<td>Pool size</td>
<td>200 feature map, convolution window size : 3<em>3, pool window size: 2</em>2</td>
</tr>
<tr>
<td>Full connection layer 1</td>
<td>Hidden neurons:50</td>
</tr>
<tr>
<td>Full connection layer 2</td>
<td>Hidden neurons:100</td>
</tr>
<tr>
<td>Output layer</td>
<td>10 classification output</td>
</tr>
</tbody>
</table>

Table 3. Result on Image Recognition with SVM and CNN

<table>
<thead>
<tr>
<th>Kernel type</th>
<th>SVM</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear kernel</td>
<td>92.53</td>
<td></td>
</tr>
<tr>
<td>Polynomial kernel</td>
<td>92.87</td>
<td></td>
</tr>
<tr>
<td>sigmoid kernel</td>
<td>87.79</td>
<td></td>
</tr>
<tr>
<td>RBF</td>
<td>94.17</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Result on Emotion recognition in Speech with SVM and CNN

<table>
<thead>
<tr>
<th>Feature type</th>
<th>CNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>36.6</td>
<td></td>
</tr>
<tr>
<td>Prosodic Feature</td>
<td>20.1</td>
<td></td>
</tr>
<tr>
<td>Wave points</td>
<td>97.6</td>
<td></td>
</tr>
</tbody>
</table>
Table 1 can be explained or understood in the following way. Layer 0 (convolution layer + pooling layer): the points in each wave is reshaped to (100,100). wave_shape = (100,100), poolsize = (2, 2), and filter shape = (5, 5), filtering reduces the wave size to (100-5+1, 100-5+1) = (96, 96), maxpooling reduces this further to (96/2, 96/2) = (48, 48). Layer 1 (convolution layer + pooling layer): the output of layer 0 is the input of layer 1. And wave_shape = (48, 48), filter_shape = (5, 5), poolsize = (2, 2). With the same method of calculation as layer 0, filtering reduces the wave size to (48-5+1, 48-5+1) = (44, 44), maxpooling reduces this further to (44/2, 44/2) = (22, 22).

Layer 2 (HiddenLayer): the HiddenLayer is fully-connected, it operates on 2D matrices of shape. It is set as following: input = 50 * 22 * 22, output = 500.

Layer 3 (LogisticRegression Layer): it classify the values of the fully-connected sigmoidal layer. And the input of layer 3 is the output of layer 2 so input=500, output=5. The output of layer 3 is the prediction labels. CNN used in image recognition is set in table 2. And the baseline model SVM are applied with different kernel types: linear kernel, Polynomial kernel, sigmoid kernel, RBF (Radial Basis Function). With different kernels, the results are different and they are be shown in part C.

C. Recognition results

As for image recognition, the results on baseline model SVM and newly applied model CNN are shown in table 3 and table 4.

From the above table3, the effect of the baseline system SVM with RBF kernel function is the best. However, its results is 94.17% and not better than CNN with accuracy 95.5%. In addition, CNN avoids manual selection process for feature sets so that the result is more general in practical.

As for emotion recognition in speech, the result is even more amazing. It can be seen from the table4 that the waveform points directly applied as input for CNN which rely on adaptive modeling the results are surprisingly good (accuracy: 97.6%). This is in line with the characteristics of bionics in neural network: construct connection according to the phenomenon and learn connection weights in their own without human decisions, in other word the neural network determines what they can see in their own instead of us. It also explains why MFCC and prosodic feature are ineffective in recognizing emotions. If we cannot acquire features which is able to characterize the emotions in speech signals by priori knowledge or experiments, and directly apply the existing features regardless of its characterization capability for emotions, it is hard to say these features can effectively identify emotions.
4. Conclusion

In this paper, the CNN model is applied in image recognition and emotion recognition in speech. The speech waveform points are directly applied for optimized Convolutional Neural Network (CNN) to improve the performance in recognizing emotion in speech and CNN is applied to identify images. For image recognition, recognition results on CNN are better than them on SVM baseline system with better robustness. As for emotion recognition in speech, by contrasting MFCC feature and prosodic feature, the result of waveform points is surprisingly good and reaches 97.6%. Finally, we conclude that CNN has projecting modeling capabilities in two-dimensional signals.

Acknowledgments

Acknowledgments should come before the References. Funding information may also be included here. Our sincere thanks to MITPress, Elsvier, WorldScientific, with the reference of which this introduction document has been written.

References

This research has been partially supported by National Natural Science Foundation of China under Grant No.61472117, No. 61203312 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References


**Bin Zhang**

She received the BE degree in 2013 from Hefei University of Technology. Now she studies for Master’s degree in Hefei University of Technology. Her research interests include Speech Recognition, Speech Synthesis, Natural Language Processing, and Affective Analysis.

**Changqin Quan**

She was born in Wuhan, China, in 1978. She received her Ph.D degree from The University of Tokushima, Tokushima, Japan, in 2011, her M.E. and B.E. degrees from Huazhong Normal University, Wuhan, China, in 2005 and 2000. She is currently a professor in the Hefei University of Technology. Her research interests include natural language processing, machine learning, and affective computing.

**Fuji Ren**

He received the Ph.D. degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences. From 2001 he joined the faculty of engineering, the University of Tokushima as a professor. His research interests include Natural Language Processing, Artificial Intelligence, Language Understanding and Communication. He is a member of the IEICE, CAAI, IEEJ, IPSJ, JSAI, AAMT and a senior member of IEEE.