Handwritten Chinese Sentence Recognition and Classification Based on Deep Convolutional Neural Networks

Chenyu Wang, Ganquan Wen, Jiali Ge, Runzhou Han, Yuchen Wang
wangcy@bu.edu, wengq@bu.edu, ivydany@bu.edu, rzhan@bu.edu, wangyc95@bu.edu

Figure 1. Handwritten Chinese characters

1. Task

In this project, we are going to study the performance of deep neural network on recognizing the image of sentences written in Chinese characters and give semantic classification on them. The work can be divided into a vision recognition part and a sentence classification part and both of them can be achieved with a deep convolutional neural network (CNN). If we have enough time, we will try to explore the performance of a recurrent neural network (RNN) and long short-term memory (LSTM) on longer handwritten sentences.

2. Related Work

Deep learning simulates human brains to establish neural networks. Networks simulate the mechanism of human brains to explain the data. Handwritten character recognition is one of the popular research fields of deep learning. Amounts of groups have contributed to the topic.

Handwritten Digits Recognition base on Improved LeNet5 mainly combine CNN and SVM together to realize the handwritten digits recognition [1]. The paper replaced the last two layers of the LeNet5 structure with SVM classifier. The main point here is: Apply CNN to extract feature vectors and use SVM to classify them. In addition, the paper used stochastic diagonal Levenberg-Marquardt to accelerate the training process of CNN. The fault rate of the new method is 0.85% which is lower than that of CNN and SVM. The potential flaw of the method is LetNet5 has more feature maps, a large fully connected layer and has a distributed representation to encode the categories at the output layers, rather than the more traditional “1” of “N” code.

A very high accuracy handwritten character recognition system for Farsi/Arabic digits combined a number of Convolutional Neural Networks with gradient descent training algorithm [2]. It is concentrated on two main contributions. The first one is automating extraction of input pattern’s features by using a CNN for Farsi digits and the second one is fusing the results of boosted classifiers to compensate the recognizers' errors. It used IFH-CDB database and applied two preprocess method to deal with the training result of the approved network. The potential flaw of the method is that it is quite similar to that of the MNIST network model. So it inherits the potential flaw of the MNIST network model.

Convolutional Neural Network Committees For Handwritten Character Classification applied group of convolution neural network which means 7 GPU train different datasets simultaneously [3]. Then classify due to the output of 7 networks. The false rate of the new method is clearly lower than the result of a single neural network. The potential flaw of the method is that the calculation needs amounts of time and it has a highly GPU demand.

3. Approach

Step I: Image segmentation

To extract the characters in an image of a sentence, it is necessary to locate the boundary of the sentence first, and then locate the boundary of each separate characters. This step can be done with Python.

The input image is expected to be clean (for example, a sentence written on white paper). After binarization, a picture can be treated as a matrix consists of 0 and 1. We count the number of ones of each row and column and sum them up respectively. The rows/columns where the summation is zero (or choose a threshold greater than zero as there may exist overlaps between two characters) can be regard as boundaries.

Step II: Character recognition

A deep CNN will be trained on GPUs and will involve Tensorflow library. A configuration from Yuhao Zhang's work can be an ideal reference [4]. In Zhang's paper, he ran most of the experiments on a small number of class of dataset because of the limited GPU power. The performance of different depth is also discussed and 11 layers is discovered to have the best performance. In our work, we may use Zhang's implementation and try to tune it by ourselves.

Step III: Sentence classification

Based on Yoon Kim's paper, with a simple CNN, we can obtain excellent sentence classification results with only a few parametric adjustments and static vectors [5]. Figure 2 is the architecture of Kim's model.

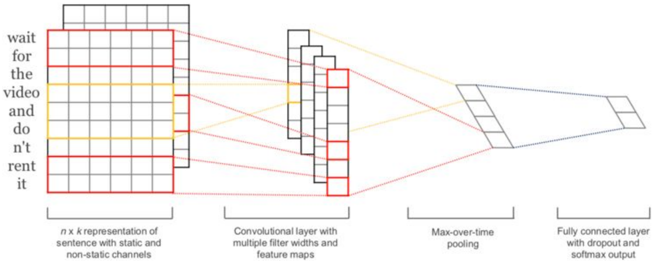
wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 2: Kim's model architecture with two channels for an example sentence.

In this part, as characters in Chinese sentence are not separated, a pretreatment is needed to split the sentences into the characters.

Basically, we plan to do a classification to judge whether the sentence conveys positive emotion or negative emotion. If we have abundant time, we will try to increase the number of classes and divide the sentences into more detailed classes.

The software involved: Python 3, Tensorflow 0.12, Numpy.

4. Dataset and Metric

The dataset we are using for character recognition is CASIA-HWDB1.1 which is developed by the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA). CASIA-HWDB1.1 contains 3755 Chinese characters which mean that the number of labels is 3755. There are 897,758 samples in the training set and 223,991 samples in the test set, both of them are stored in gnt files. However, it is not feasible to use the gnt file to train our deep learning model, so, converting the gnt file to png or jpg format would be the very first step for our data preprocessing. Furthermore, adjusting the size of input images would be the next step for data preprocessing since the size of input images should be the same. Moreover, applying some image processing methods like contrast enhancement and smoothing to the input images could possibly improve the performance of the model.

The dataset we are using for sentence classification is a refined set of news headlines from Sougou. This dataset is a compilation of news from news portals such as Today's headlines, Sina, Baidu News, etc. We plan to pre-classified the set into a positive emotion class and a negative emotion class using LDA algorithm.

The metric for our project would be CCR(Correct Classification Rate), we would like to see how many samples in the test set will be classified correctly.

5. Proposed Timeline and Roles

Task	**Deadline**	**Lead**
Preprocess data for training	10/20/18	Chenyu Wang
Image segmentation(extract the characters in an image of a sentence)	10/31/18	Jiali Ge
Apply CNN to character recognition	11/20/18	Runzhou Han
Using CNN to do sentence classification	12/01/18	Yuchen Wang
Compute metrics and analyze errors	12/07/18	Ganquan Wen
Adjust parameters to improve accuracy; Prepare report and presentation	12/09/18	All

References

[1] Yu N, Jiao P, Zheng Y. *Handwritten digits recognition based on improved LeNet5*. Control and Decision Conference on IEEE, 2011: 1135-1139.

[2] Ahranjany S, Razzazi F, Mohammad H. *Ghassemian: A very high accuracy handwritten character recognition system for Farsi/Arabic digits using Convolutional Neural Networks*. BICTA, 2010.

[3] Ciresan D, Meier U, Gambardella L, Schmidhuber J. *Convolutional Neural Network Committees for Handwritten Character Classification*. ICDAR, 2011

[4] Zhang Y. *Deep Convolutional Network for Handwritten Chinese Character Recognition*.(2015).

[5] Kim Y. *Convolutional Neural Networks for Sentence Classification*. EMNLP, 2014.