Chinese Character Image Clustering and Classification Based on Object Embedding Model (Student Abstract)

Mengting Wang, Xun Liang, Yang Xue
School of Information, Renmin University of China, Beijing 100872, China
{mtw, xliang, xueyang}@ruc.edu.cn

Abstract

We proposed Image2Vec, an unsupervised algorithm that learns feature representations for variable-size pieces of Chinese character images.

Introduction

In recent ten years, the methods represented by CNN supervised learning have made great progress in writer identification, and the precision has reached more than 99%. Nevertheless, the precision of the unsupervised learning methods is still at a low level of about 60%. This paper studies the problem of font recognition based on unsupervised learning method. The font recognition on Chinese characters is a subset of the broader domain of writer identification. Whereas supervised learning methods for Chinese character writer identification have made great achievements, unsupervised learning for this task is still a very difficult problem because the ability of feature extraction in unsupervised learning is weaker. Unsupervised learning methods for the writer identification include TF-IDF and subtractive clustering method (Singh and Sundaram 2015), three-layer LDA (Shivram et al. 2013), and paramagnetic clustering and spatial temporal neural network for Farsi writer identification (Sangdehi and Faez 2009). It is found that the success rate of writer identification based on unsupervised learning is very low in the case of a few training samples.

We proposed an unsupervised learning model to map images into feature vectors, called Image2Vec. We are inspired by the "distribution hypothesis" (words with similar contexts have similar semantics) (Harris 1954). In certain type of images, pixels with similar context also have similar meanings, so each pixel can be regarded as a word. In Image2Vec, we use the object embedding model to construct the vectors in the high-dimensional space according to the image characteristics. Lastly, we use manifold learning and logical regression algorithm to test the model. The classification results outperformed the existing unsupervised learning writer identification methods.

Model

In this section, we introduce our model Image2Vec. Given a series of images $\{M_i\}$ (i = 1, 2, ..., N) , Image2Vec maps the items into the real space $\mathbb{R}^d$. As showed in Figure 1, Image2Vec has two main steps: the preparation step and object embedding step.

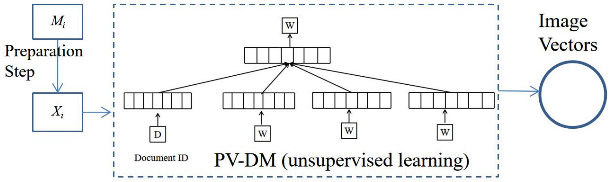

Figure 1: Image2Vec model.

Preparation Step. The input of the preparation step is a set of pictures. The convert method of the existing PIL platform are applied to convert the image into a matrix. Each pixel is represented by 8 bits, 0 for black, 255 for white. The conversion formula is as follows:

$$L = R * 299/1000 + G * 587/1000+ B * 114/1000 \quad (1)$$

The output of this step is the document collection $\{X_i\}$ (i = 1, 2, ..., N). Each document corresponds to an image matrix.

Object Embedding Step. Here we adopt the PV-DM model proposed in the document embedding model. The algorithm is based on the idea of Word2Vec model. Word2Vec is based on a large-scale deep learning model, as showed in Figure 2. Document embedding model is for one-dimensional sequential data. Although images are two-dimensional data, our image form is single, for the main body is text and the text in each image is sequential (a.k.a. not rotated), so the information of x-axis is enough to represent the characteristics of the image and the information of *y*-axis is not extracted here. The input of this step is the document collection $\{X_i\}$ (i =1, 2, ..., N). The outputs are N d-dimensional vectors, with each dimension representing a hidden feature of the image.

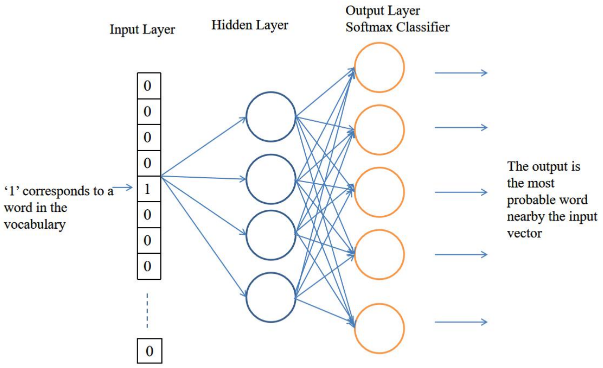
Input Layer
Hidden Layer
Output Layer
Softmax Classifier
0
0
0
0
1
0
0
0
0
'1' corresponds to a word in the vocabulary
The output is the most probable word nearby the input vector

Figure 2: Network model for natural language processing.

Dataset #		Precision	Recall	F1	# of test samples
1		83	71	70	16
2	Song	**100**	**95**	**98**	65
	Kai	90	88	89	41
	Hei	89	95	92	60

Table 1: Results of logistic regression classification (%).

Experiments

To investigate the effectiveness of our model, we tested our model by unsupervised manifold learning algorithm and supervised classification algorithm. Using the geometric properties of object embedding vectors, we applied T-SNE to reduce the vector dimension for visual observation. The machine learning algorithm we selected for object embedding vector classification is logical regression.

We conducted the experiment on two data sets. Dataset 1 is the pictures of the works of four famous regular script calligraphers Zongyuan Liu, Xun Ouyang, Zhenqing Yan and Mengfu Zhao, as showed in Figure 3. Dataset 2 is the pictures of characters in Kai, imitation Song Dynasty and Hei style. The result of dimension-reduction clustering on dataset 2 is shown in Figure 4, which shows the similarity differences of pictures caused by different stroke thickness, stroke shape differences and other factors. The results of classification are shown in Table 1. The recognition result of imitation Song Dynasty style in the dataset 2 is the best, which indicates that the more characteristic the font, the higher the recognition rate. With the similar number of samples, the mean writer recognition rate of subtractive clustering method with 10 prototypes is only 61.3% (Singh and Sundaram 2015). When the sample size is more than 9000, the accuracy of SPC method is only 79% (Sangdehi and Faez 2009), which is lower than the results of our model.

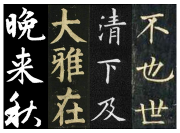

Figure 3: Four Calligraphers' works. From left to right are the works of Zhao, Yan, Ouyang and Liu.

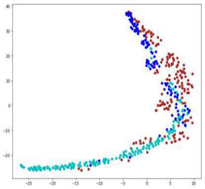

Figure 4: Manifold learning dimension-reduction clustering. Blue, brown and green dots represent the imitation Song, Kai and Hei style font respectively.

Conclusions

In this paper, we proposed Image2Vec, an unsupervised model that learns vector representations for variable-size pieces of Chinese character images. Unsupervised learning for Chinese character images is an arduous task, nonetheless, the experimental results of clustering and classification show that our method is effective.

Without the labeled training sample set, the unsupervised method in this paper can automatically distinguish the styles of four famous calligraphers in Chinese history. This theoretically reveals the potential ability of unsupervised deep learning method in fine-grained image recognition.

Our model has the potential to identify the works of less famous or unknown calligraphers and achieve calligrapher identification that can surpass the expert level, accomplishing the fine-grained recognition in unsupervised learning situation for the first time, which is a great step towards the automatic writer identification without labels.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62072463, 71531012), and State Key Laboratory of Digital Publishing Technology of Founder Group. Xun Liang is the corresponding author of this paper.

References

Harris, Z. 1954. Distributional Structure. Word, 10, 146-162. doi.org/10.1080/00437956.1954.11659520.

Sangdehi, S.A.T., Faez, K. 2009. Writer Identification Using Super Paramagnetic Clustering and Spatio Temporal Neural Network. Lecture Notes in Computer Science, Vol. 5856: 669-676.doi.org/10.1007/978-3-642-10268-4_79.

Shivram, A.; Ramaiah, C.; Govindaraju, V. 2013. A hierarchical Bayesian Approach to Online Writer Identification. IET Biometrics, Vol. 2(4): 191-198.

Singh, G., Sundaram, S. 2015. A Subtractive Clustering Scheme for Text-Independent Online Writer Identification. In Proc. of the 2015 Int. Conf. Document Analysis and Recognition, 311-315.