CS230

State-of-the-art Approaches for Handwriting-based Recognition on Gender and Handedness Using Deep Learning

Yanbang Wang
Department of Computer Science
Stanford University
ywangdr@stanford.edu

Jiangshan Li
Department of Electrical Engineering
Stanford University
jiangsli@stanford.edu

Tiancheng Cai
Department of Computer Science
Stanford University
caitch@stanford.edu

Abstract

Handwriting-based demographic identification, such as gender or handedness classification, has many applications in forensic biometrics and archaeology. Most of the existing methods heavily rely on handcrafted features combined with classic statistical methods. We report a deep-learning based method that achieved state-of-the-art performance on both English and Chinese handwriting dataset. Our model significantly improves over the existing best models by reducing their error rate by at least 30% in two different test settings. It also outperforms humans by a large margin. Meanwhile, we also conducted transfer learning to further boost performance on Chinese dataset. As yet another significant contribution, we managed to collect and publish TSEH (Tough Samples of English Handwritings dataset)[1] that simulate the complex and tough environment of application, which consist of 350 handwriting sample collected from 32 writers. Moreover, given the strong nature of application of the project, we also build a website to let readers to predict their own handwriting images.[2]

1 Introduction

Handwriting-based demographic identification is an important topic in document analysis and recognition. The "offline" version of the problem (which we focus on) is defined as the following: given an image of the handwriting of one person, and without any further knowledge, predict the person's demographic information, such as gender, handedness, age group, etc. Our work primarily focuses on prediction of gender and handedness. The task has many applications for different communities: police can use handwriting left at crime scenes to identify criminals' gender or handedness, and archaeologists also find such technology helpful in identifying figures of interest in history, etc. [1]

[1] Uploaded to github along with code

[2] We recorded a video to demonstrate this: https://drive.google.com/open?id=1esQpqV1EE_cyBKnG30v32GSLp13WuGpP

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Previous research largely fouced on integrating domain knowledge into the design of various statistical methods. Despite attempts of neural-network-based approaches in recent years, the possibility of building a relatively deep neural network still remains under-explored. Moreover, to our best knowledge no previous work report performance on the task on Chinese dataset. In light of this, we propose a deep-learning-based network architecture, with transfer learning (in some settings) to attack the problem. We also contribute to the community by publishing a new dataset collected by ourselves for measuring the performance in a much tougher environment.

2 Related work

2.1 Automatic Demographic Recognition

Automatic prediction of gender or handedness from handwriting involves image preprocessing, feature extraction and classification.[3] Most of the existing publications adopt traditional approach as for feature extraction and classification. Anil Thomas[4] trained the 80 most discriminative features by Gradient Boosting Decision Trees and averaged the probability outcomes to predict the gender. Liwicki[5] performed classification by Gaussian Mixture Method (GMM) and Support Vector Machine (SVM) with a detection rate of 67.57% on IAM-OnDB dataset. Gattal et al.[6] extracted the oriented Basic Image Features (oBIFs) before applying SVM, with a classification rate of 68% - 76% on a subset of QUWI dataset. Fourier descriptors[7] were also proposed for the tasks, which combine tangent and curvature information to classify gender from manuscript samples. Furthermore, N. Bi et al.[8] proposed kernel mutual information (KMI) approach to select and integrate some handwriting features and used SVM to identify the gender of writers, with an accuracy of 66.3% on a subset of QUWI dataset and 66.7% on Registration-Document-Form (RDF) database[9] in Chinese.

While the problem of demographics identification from handwriting has mostly been studied by traditional machine learning techniques, the application of deep network to this field is still novel. Ángel Morera et al.[10] claims to be the first to use a convolutional neural network for the task of demographic identification from handwriting in 2018. It achieves an accuracy of 80.72% and 90.70% on gender and handedness for the IAM dataset, and an accuracy of 68.90% and 70.91% on gender and handedness for the KHAFF dataset. Illouz E. et al.[11] proposes using CNN in a more end-to-end approach with less feature tuning and evaluate on their new HEBIU handwriting dataset in English and Hebrew and achieves an overall accuracy of 77% for both languages.

2.2 Human Performance

Illouz et al. [2] reported on 300 human volunteers who on average achieved 63.6% and 66.2% accuracy on English and Hebrew writings, respectively, when given a roughly balanced test set (male's and female's samples account for 50% each). Simlarly, Liwicki et al. reported in [3] that non-experts' performance on a balanced English test set is 63.88%.

3 Dataset and Features

There are some existing handwriting databases and benchmarks in English, Arabi, and Chinese.[12] Datasets that claims to come with writer's demographic information include IAM, QUWI[13], KHAFF, HCL2000[14], HIT-MW[15], CASIA-OLHWDBI[16]. However, we made vaious attempt throughout the quarter, and verified that only IAM and HIT-MW are freely available to public with reliable labels, using which we ran all the experiments.

IAM Dataset
The dataset contains more than 1,700 English writing examples from 221 writers, which amount to 13,049 extracted text lines. It also contains gender, handedness, country, language, education, etc. The extracted line images vary by sizes (height $\times$ width). The heights' medium number is 78, and the widths' is 989. Preprocessing (details elaborated in next section) results in 34,764 writing samples by males, and 16,326 writing samples by females. Grouped by handedness, there are 47,057 right-handed writing samples, and 4,033 left-handed ones.

HIT-MW Dataset
The dataset contains 853 Chinese handwriting samples from more than 780 writers, which amounts

to 8,664 text lines and 186,444 Chinese characters. It is labeled with gender. There are 2498 labeled images of text lines from 254 writers. Preprocessing results in 14,743 writing samples. 6,486 of them were written by males and 8,257 of them were written by females.

TSEH TSEH only has gender labels. Tiancheng, please include some stats of our dataset, especially the how many people are males or females

Preprocessing involves the following steps applied to both datasets: 1. Perform background whitening and Foreground darkening using OTSU Threshold in [6]. 2. Resize Images to have a unified height of 80, while keeping the aspect ratio unchanged. 3. Crop (long) line images to shorter writing clips with a uniform size of 80 * 320 using a sliding window of that size. 4. Perform data augmentation, including rotation, shifting, and morphological operation. 5. Centralize and normalize the cropped images.

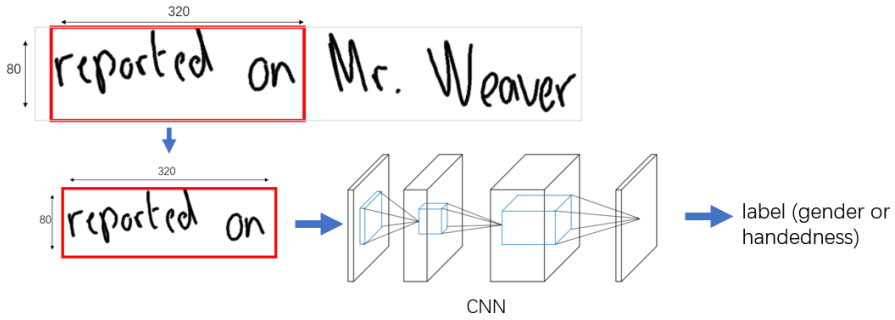

Figure 1: Data pipeline in basic task setting. The line image is sampled from IAM dataset.

4 Methods

4.1 Basic Network Architecture for IAM

One basic block of our network consists of 2 successive convolutional layer (with one ReLU nonlinearity), one maxpooling layer, and one 2D dropout layer. The whole network contains 4 basic blocks stacked together plus 3 fully connect layers, each with batch norms and dropouts. The final activation is Sigmoid. The detailed structure is attached to appendix in Figure 2.

4.2 Basic Network Architecture for HIT-MW

Each convolutional block of this network includes a convolutional layer, a batchnorm layer, an activation layer(ReLu), a dropout layer and a max pooling layer. This network has 5 basic convlutional blocks and 2 fully connected layers. Each fully connected layer is followed by a batchnorm layer, a ReLu and a dropout layer. The final activation is Sigmoid. The architecture is is attached to appendix in Figure 3.

4.3 Loss Functions

Binary classification loss is adopted in our training. Note that we use different weights for binary loss in different task settings (elaborated next subsection).

$$\mathcal{J} = -\frac{1}{m}\sum_{i=1}^{m}(\alpha y_i \log \hat{y}_i + \beta(1-y_i)\log(1-\hat{y}_i))$$

, where $\alpha = \beta = 1.0$ when training set and test set follows the same distribution, and $\alpha = \frac{r+1}{2r}$, $\beta = \frac{r(r+1)}{2r}$, r denotes the ratio of number of positive samples against negative samples. α and β are fixed as such with two reasons: The first reason is to cancel of the over(under) estimation of the majority (minority) class during training. The second reason is to keep the expectation of weight of

one sample to be consistently unit as $\alpha = \beta = 1.0$ does (so that we are able to roughly compare training accuracies of different settings by only looking comparing their training losses).

4.4 Motivation and Strategy for Transfer Learning

Our work only focused on transfering models from English dataset (IAM) to Chinese dataset (HIT-MW) for two reasons. The reason is that it is noticed our English dataset is about 6 times larger than the Chinese one, making it more reasonable to leverage the large volume of the English dataset to compensate the relative "shortage of data" in Chinese dataset, and not the other way around.

Due to the number of writing examples is small in HIT-MW, the network is hard to tune for this dataset. Deep network architecture tends to overfit the training data and small network architecture results in high bias. Therefore, we came up with transfer learning to improve the performance of prediction on Chinese datasets.

According to the model trained on IAM(large English dataset), we freeze the previous layers and only train the last few layers. The previous layers trained on large dataset can extract more precise features on handwriting. Fine-tuning the last few layers makes the network use these features to identify Chinese writers' gender. The number of layers to train is a hyperparameter. First train the last two layers, if the bias is higher than the smaller model trained on HIT-MW, then set another last two layers of pre-trained model to be trainable until the bias is low enough. Then use weight decay and add dropout layers to reduce the variance.

In the end, we fine-tuned all the fully-connected layers and the last convolutional layer, and got a more satisfying result compared with previous models.

5 Experiments, Results, and Analysis

We ran extensive experiments under many different settings. First, we trained (and validated) our model on IAM's train/val set and test it on the IAM's test set (under either balanced or unbalanced sample distribution), HIT-MW's test set, and TSEH. Note that we trained a model for gender and handedness prediction each. Second, the activities above were repeated except that training was run on HITWM's train/val set this time. Third, we carried out transfer learning by transfering the model trained on IAM to test on HIT-MW. The following subsections elaborate on each of them.

We define our accuracy = number of correctly classified samples / total number of samples, precision = $\frac{TP}{TP+FP}$, recall = $\frac{TN}{TN+FP}$ [3]

5.1 Training on IAM

The Experiments results are summarized in Table 1. In terms of hyper-parameters, We keep a consistent lr = 0.001, batch size = 256, and use Adam optimizer. As can be seen here, our model constantly outperforms the previous best models on both balanced and unbalanced dataset on gender recognition task. In terms of handedness recognition, we would like to point out that previous best models were all tested on highly skewed test sets, with right-handed samples accounting for over 90% of the test set. In our work, we instead only focus on tailoring our models to suit a balanced test environment, and would like to mention that we are the first to predict handedness on the balanced test set of IAM. Also notice that when models were originally trained on English, accuracy drops significantly to about 0.5 when the tested on Chinese dataset. Performance on TSEH (which has a mixture of Chinese and English), in comparison, is much better.

5.2 Training on HIT-MW

The experimental results are shown in Table 2. The training set accuracy is 0.782, the validation set accuracy is 0.755. Initialize weights by Xavier initialization initialize bias as 0. Use Adam optimization with weight decay = 1e-2 (has an effect of regularization to reduce variance). batch size = 128. learning rate = 1e-4. The number of filters in each layer is chosen referring to the parameters of VGG16. Every time set the next convolutional layer, the number of filters goes up by a factor of 2.

[3] We calculate precision for males and females independently, and then weighted average them. The same is for calculation of recall. TP: true positves, FP: false postives, TN: true negatives, FN: False negatives

Test Set	Task	Accuracy	Precision	Recall	F1 Score
IAM (Balanced)	Handedness	0.743	0.746	0.743	0.744
IAM (Skewed)	Handedness	0.923	0.712	0.792	0.776
IAM (Balanced)	Gender	0.714	0.717	0.714	0.715
IAM (Skewed)	Gender	0.859	0.845	0.828	0.836
HIT-MW	Gender	0.567	0.515	0.504	0.514
TSEH	Gender	0.639	0.588	0.435	0.500
Previous Best results:					
IAM (Balanced)	Gender	0.689	0.685	-	-
IAM (Skewed)	Gender	0.807	0.780	0.789	0.784
IAM (Skewed)	Handedness	0.907	0.690	0.780	0.724
Humans	Gender	0.639	-	-	-

Table 1: Performance results when training on IAM Dataset and test on different others.

Gender	Precision	Recall	F1 Score
male	0.712	0.698	0.705
female	0.854	0.891	0.872
No previous Best results:			

Table 2: Performance results when training on HIT-MW Dataset.

To reduce variance, set dropout rate = 0.3 (smaller rate makes the network tend to overfit, larger rate cause higher bias) add batchnorm layer after each convolutional layer and fully-connected layer. Batch norm has a slight regularization effect and make the network robust to covariate shift.

5.3 Transfer learning

We fine-tuned all the fully-connected layers and the last convolutional layer of the model pre-trained on IAM. We increased the test set accuracy to 0.844 and increased the validation set accuracy to 0.801. The validation accuracy of previous models is 0.782.

6 Conclusion & Future Work

In this work, we introduced how we achieved state-of-the-art performance of demographic recognition on both English and Chinese, and on both handedness and gender prediction tasks, using deep learning with a relatively large CNN. We also carefully designed an approach to transfer knowledge learned from large English dataset to boost prediction accuracy on Chinese dataset. Extensive experiments were conducted under many different settings, and the results verify that our work does outperform a number of previous best results. We also contribute to the public TSEH as a "tough" dataset for testing gender reconition performance. In terms of future work, we suggested more investigation into how to boost the accuracy on TSEH, since be able to really apply the model to realities, we need to figure out a better way to transfer from the lab environment set up by IAM and HITMW to the "dirty" realities.

7 Contributions

Yanbang Wang:

1. Design and implement the whole data preprocessing pipeline.
2. Build and tune all the models that were trained on IAM dataset.
3. Collect 30% of TSEH dataset.

4. Write the majority of final report.

Jiangshan Li:

1. implement data preprocessing for HIT-MW.
2. design the model for HIT-MW.
3. implement transfer learning from IAM to HIT-MW dataset.
4. Collect 30% of TSEH dataset.

Tiancheng Cai:

1. implement data preprocessing for HIT-MW.
2. design the model for HIT-MW.
3. implement transfer learning from IAM to HIT-MW dataset.
4. Collect 30% of TSEH dataset.

References

[1] Rehman, A., Naz, S. & Razzak, M.I. Multimed Tools Appl (2019) 78: 10889. https://doi.org/10.1007/s11042-018-6577-1

[2] Illouz E., (Omid) David E., Netanyahu N.S. (2018) Handwriting-Based Gender Classification Using End-to-End Deep Neural Networks. In: Kůrková V., Manolopoulos Y., Hammer B., Iliadis L., Maglogiannis I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018. Lecture Notes in Computer Science, vol 11141. Springer, Cham

[3] Liwicki, M., A. Schlapbach, H. Bunke. Automatic Gender Detection Using On-Line and Off-Line Information. – Pattern Analysis and Applications, Vol. 14, 2011, No 1, pp. 87-92.

[4] N. Bouadjenek, H. Nemmour, and Y. Chibani, "Histogram of Oriented Gradients for writer's gender, handedness and age prediction," in Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, INISTA 2015, IEEE, Madrid, Spain, August 2015.

[5] Ángel Morera, Ángel Sánchez, José Francisco Vélez, and Ana Belén Moreno, "Gender and Handedness Prediction from Offline Handwriting Using Convolutional Neural Networks," Complexity, vol. 2018, Article ID 3891624, 14 pages, 2018. https://doi.org/10.1155/2018/3891624.

[6] Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." IEEE transactions on systems, man, and cybernetics 9.1 (1979): 62-66.

Appendix

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 32, 79, 319]	544
Conv2d-2	[-1, 32, 78, 318]	16,416
ReLU-3	[-1, 32, 78, 318]	0
MaxPool2d-4	[-1, 32, 40, 160]	0
Dropout2d-5	[-1, 32, 40, 160]	0
Conv2d-6	[-1, 64, 39, 159]	32,832
Conv2d-7	[-1, 64, 38, 158]	65,600
ReLU-8	[-1, 64, 38, 158]	0
MaxPool2d-9	[-1, 64, 20, 80]	0
Dropout2d-10	[-1, 64, 20, 80]	0
Conv2d-11	[-1, 128, 19, 79]	131,200
Conv2d-12	[-1, 128, 18, 78]	262,272
ReLU-13	[-1, 128, 18, 78]	0
MaxPool2d-14	[-1, 128, 10, 40]	0
Dropout2d-15	[-1, 128, 10, 40]	0
Conv2d-16	[-1, 256, 9, 39]	524,544
Conv2d-17	[-1, 256, 8, 38]	1,048,832
ReLU-18	[-1, 256, 8, 38]	0
MaxPool2d-19	[-1, 256, 5, 20]	0
Dropout2d-20	[-1, 256, 5, 20]	0
Linear-21	[-1, 2048]	52,430,848
ReLU-22	[-1, 2048]	0
BatchNorm1d-23	[-1, 2048]	4,096
Dropout-24	[-1, 2048]	0
Linear-25	[-1, 500]	1,024,500
ReLU-26	[-1, 500]	0
BatchNorm1d-27	[-1, 500]	1,000
Dropout-28	[-1, 500]	0
Linear-29	[-1, 10]	5,010
ReLU-30	[-1, 10]	0
BatchNorm1d-31	[-1, 10]	20
Dropout-32	[-1, 10]	0
Linear-33	[-1, 1]	11
Sigmoid-34	[-1, 1]	0

Total params: 55,547,725

Figure 2: Network architecture in detail.

chinese.png

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 32, 80, 320]	320
BatchNorm2d-2	[-1, 32, 80, 320]	64
ReLU-3	[-1, 32, 80, 320]	0
Dropout2d-4	[-1, 32, 80, 320]	0
MaxPool2d-5	[-1, 32, 40, 160]	0
Conv2d-6	[-1, 64, 40, 160]	18,496
BatchNorm2d-7	[-1, 64, 40, 160]	128
ReLU-8	[-1, 64, 40, 160]	0
Dropout2d-9	[-1, 64, 40, 160]	0
MaxPool2d-10	[-1, 64, 20, 80]	0
Conv2d-11	[-1, 128, 20, 80]	73,856
BatchNorm2d-12	[-1, 128, 20, 80]	256
ReLU-13	[-1, 128, 20, 80]	0
Dropout2d-14	[-1, 128, 20, 80]	0
MaxPool2d-15	[-1, 128, 10, 40]	0
Conv2d-16	[-1, 256, 10, 40]	295,168
BatchNorm2d-17	[-1, 256, 10, 40]	512
ReLU-18	[-1, 256, 10, 40]	0
Dropout2d-19	[-1, 256, 10, 40]	0
MaxPool2d-20	[-1, 256, 5, 20]	0
Conv2d-21	[-1, 512, 5, 20]	1,180,160
BatchNorm2d-22	[-1, 512, 5, 20]	1,024
ReLU-23	[-1, 512, 5, 20]	0
Dropout2d-24	[-1, 512, 5, 20]	0
MaxPool2d-25	[-1, 512, 2, 10]	0
Linear-26	[-1, 2048]	20,973,568
BatchNorm1d-27	[-1, 2048]	4,096
ReLU-28	[-1, 2048]	0
Dropout-29	[-1, 2048]	0
Linear-30	[-1, 2048]	4,196,352
BatchNorm1d-31	[-1, 2048]	4,096
ReLU-32	[-1, 2048]	0
Dropout-33	[-1, 2048]	0
Linear-34	[-1, 1]	2,049
Sigmoid-35	[-1, 1]	0

Total params: 26,750,145
Trainable params: 26,750,145
Non-trainable params: 0

Figure 3: Network architecture in detail.