Report of ICPR MTWI Challenge 3

Team: nelslip(iflytek&ustc)
Jianshu Zhang*,Yixing Zhu*, Jun Du*, Lirong Dai*, Mingjun Chen†, Jiajia Wu† and Jinshui Hu†
*National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, P. R. China
†iFLYTEK Research, Hefei, Anhui, P. R. China

I. Methodology

This challenge requires the recognition system to recognize text lines existed in one big image, therefore we first detect possible text lines in the image and then recognize them as single text lines.

A. Detection

As for the detection system, our method is based on FPN [1], and we augmented the network as PANet [2] does. Because text line may be arbitrary quadrilateral, we use SLPR [3] to fit the outline of the text line, this method isometrically regress the points on the edge of text by using the vertically and horizontally sliding lines which can solve ambiguity of four vertices label, then we use its outline to generate a oriented rectangle for RoIRotate, so we adopt Cascade R-CNN as [4] with two steps, in the first step we propose a horizontal rectangle, in the second step we propose a oriented rectangle with SLPR, finally, we regress four quadrilateral vertices.

B. Recognition

As for the recognition system, we adopt the attention based encoder-decoder model for this challenge. The encoder is a multi-layer convolutional network that learns to encode input web images and maps them to high-level visual features. The decoder is a recurrent neural network (RNN) with gated recurrent units (GRU) that converts these high-level features into output strings one character at a time. We employ DenseNet [5] architecture as the CNN encoder. The implementation of attention based encoder-decoder model by using DenseNet as encoder is illustrated in [6]. To adapt this model to the OCR problem and capture the document's temporal layout, we also incorporate a new source encoder layer in the form of a multi-row bidirectional GRU applied before the application of attention. Most importantly, we employ the Radical Analysis Network (RAN) [7] since Chinese characters dominate the whole dataset in this challenge. RAN recognize Chinese characters by firstly identifying radicals, analyzing two-dimensional spatial structures among them and then generating IDS (Ideographic Description Sequences, could be found in wikipedia) caption of Chinese characters. The manner of treating a Chinese character as a composition of radicals rather than a single character class largely reduces the size of vocabulary and enables RAN to possess the ability of recognizing unseen Chinese character classes. To implement RAN on text line recognition, we append an end-of-word (eow) flag after each Chinese character IDS caption. Besides, some text line images are hard to be recognized in RGB channels, so the raw input of our encoder not only contains RGB information of images but also contains HSV information. Finally, since this challenge also requires the system to recognize some rotated text lines, we include image rotation as data augmentation.

II. Experimental details

A. Detection

We use some tricks to improve the performance of our detection system, including image rotation, multi-scale training & testing, when model assembling, we use ResNeXt-101 (32*8d) [8] pre-trained on ImageNet as backbone network and we only use a single model. All experiments were implemented in Caffe2 by using the NVIDIA GTX 1080Ti GPU.

B. Recognition

As for the recognition system, we employ DenseNet-BC ($L = 171, k = 24$) [5] and two stacked bidirectional GRU layers (each layer has 512 units) as the encoder. The decoder is a unidirectional GRU layer (256 units) equipped with coverage based attention model [9]. We train 12 models with different parameters initialization, where 3 models utilize HSV information and 3 models utilize image rotation as data augmentation. During testing, we combine these 12 models by employing an ensemble way at each beam search step. The experiment are all implemented with Theano 0.10.0 and an NVIDIA Tesla P40 24G GPU.

III. Dateset

The official training dataset contains 10,000 images and we split them into 9,000 for training and the other 1,000 for validation. We did not use any extra data for training our detection system.

When training our recognition system, we crop text lines from raw images and the official training set contains 128,210 text lines and the validation set contains 15,288 text lines. Besides, we also collect about 250,000 text lines in the natural scenes and label them manually. The additional 250,000 text lines are all used for training, including 180,000 Chinese character text lines and 70,000 English character text lines.

IV. TEAM INTRODUCTION

The team leader is Jianshu Zhang, his email address: xysszjs@mail.ustc.edu.cn, telephone number: 15856910468, address: West campus of University of Science and Technology of China, No.8 apartment building, Room 640, No.443 HuangShan Road, Hefei, Anhui, 230027, China.

This team is a cooperation between National Engineering Laboratory for Speech and Language Information Processing in University of Science and Technology of China (USTC-NELSLIP) and iFLYTEK Research. We hope to have others' attention on our technology for addressing OCR and detection problem by participating in this competition.

REFERENCES

[1] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, vol. 1, no. 2, 2017, p. 4.

[2] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *arXiv preprint arXiv:1803.01534*, 2018.

[3] Y. Zhu and J. Du, "Sliding line point regression for shape robust scene text detection," *arXiv preprint arXiv:1801.09969*, 2018.

[4] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," *arXiv preprint arXiv:1712.00726*, 2017.

[5] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

[6] J. Zhang, J. Du, and L. Dai, "Multi-scale attention with dense encoder for handwritten mathematical expression recognition," *arXiv preprint arXiv:1801.03530*, 2018.

[7] J. Zhang, Y. Zhu, J. Du, and L. Dai, "RAN: Radical analysis networks for zero-shot learning of Chinese characters," *arXiv preprint arXiv:1711.01889*, 2017.

[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5987–5995.

[9] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, vol. 71, pp. 196–206, 2017.