Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition (Supplementary Material)

Shancheng Fang Hongtao Xie* Yuxin Wang Zhendong Mao Yongdong Zhang
University of Science and Technology of China
{fangsc, htxie, zdmao, zhyd73}@ustc.edu.cn, wangyx58@mail.ustc.edu.cn

1. Analysis of SOTA method

Table 1. Ratios of failure cases (%) of SOTA method [4]. Strong/Weak Semantics is estimated by classifying text inside/outside Oxford Dictionaries. ed is the edit distance.

Strong Semantics			Weak Semantics		Illegible & Wrong Label
ed=1	ed=2	$ed \geq 3$	alphabet	digit	
23.2	11.0	11.6	26.9	3.6	23.7

We analyze the failure cases of current state-of-the-art (SOTA) method SRN [4], as the statistical data shown in Tab. 1. Among the failure cases, 45.8% text is with strong semantics, which should be successfully reasoned by linguistic rules. Specifically, 23.2% text is wrongly recognized by SRN where the edit distance between predicted text and ground truth is only 1. This denotes that there is still a big room for enhancing SOTA methods in language modeling, and therefore we propose the ABINet aiming to improve the ability of language modeling.

2. Gradient analysis

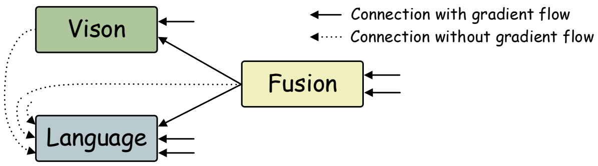
Vison
Fusion
Language
Connection with gradient flow
Connection without gradient flow

Figure 1. Analysis of gradient flow in ABINet.

As shown in Fig. 1, both vision and language models receive gradient from either direct supervision or backpropagation of fusion model, which ensures the independence of each model. Besides, the iteration can be regarded as a data augmentation method during training phase, which has no disturbance on the flow of gradient.

3. Analysis of Successful/Failure Cases

Figure 2 gives some examples which are successfully recognized by ABINet-LV$_{est}$ (bottom text) and wrongly recognized by ABINet-LV (top text). Failure cases of ABINet-LV may come from extremely long text, unusual font, irregular text, blurred images, etc. We directly resize all the images to 32 × 128, which is unfriendly to extremely long text due to the squeezed visual patch. However, we observe that learning from more data using a semi-supervised way can alleviate this problem obviously. Besides, even though we find that the language model (BCN) can assist the recognition of vision model dealing with unusual font, learning from more text images with various fonts is still a straightforward and effective way. We also find that some irregular text that is not recognized by ABINet-LV can be resolved by semi-supervised learning, which is the reason why ABINet-LV$_{est}$ achieves significant performance on CUTE dataset (Tab.6 in regular manuscript). In the last three rows of Figure 2, we visualize some hard examples caused by blurred images, which indicates that ABINet-LV$_{est}$ is able to recognize text under a hostile environment that even humans cannot read.

4. Experiment on Chinese Dataset

Table 2. Recognition accuracies on ICDAR2015 TRW.

Method	Character Accuracy(%)
CASIA-NLPR[5]	72.1
SCCM w/o LM[3]	76.5
SCCM[3]	81.2
2D-Attention [4]	72.2
CTC [4]	73.8
SRN [4]	85.5
ABINet	**87.1**

To validate the performance on non-Latin recognition, an additional experiment on ICDAR2015 Text Reading in the Wild Competition dataset (TRW15) [5] is conducted. Experimental setup follows the configuration of SRN [4].

*The corresponding author

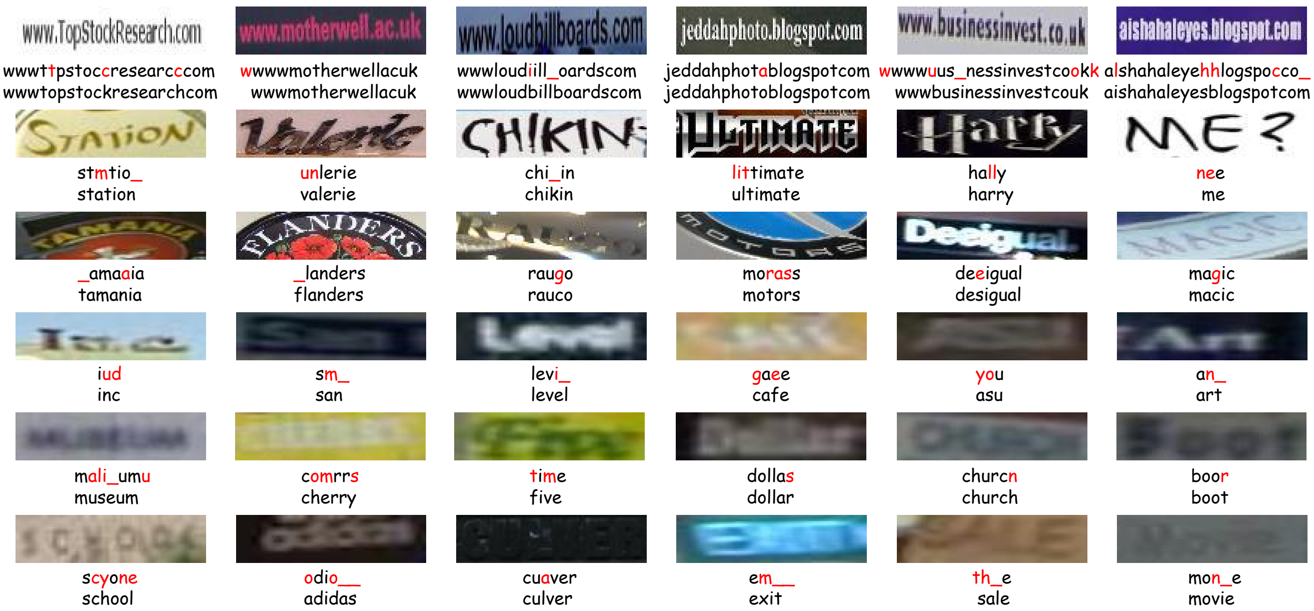

Figure 2. Recognition examples.

Specifically, ABINet is trained on a synthetic dataset with 4 million images, and the training sets of RCTW [1] and LSVT [2]. From the results in Tab. 2 we can see, ABINet obtains a 1.6% improvement on TRW15 dataset compared with SRN, showing that our ABINet is also robust to Chinese text recognition.

References

[1] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1429–1434. IEEE, 2017. 2

[2] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9086–9095, 2019. 2

[3] Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*, 2017. 1

[4] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12113–12122, 2020. 1

[5] Xinyu Zhou, Shuchang Zhou, Cong Yao, Zhimin Cao, and Qi Yin. Icdar 2015 text reading in the wild competition. *arXiv preprint arXiv:1506.03184*, 2015. 1