Pushing the Performance Limit of Scene Text Recognizer without Human Annotation–Supplementary Material

Caiyuan Zheng[1,2*], Hui Li[3], Seon-Min Rhee[4], Seungju Han[4], Jae-Joon Han[4], Peng Wang[1,2†]
[1]School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, China,
[2]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, China,
[3]Samsung R&D Institute China Xi'an (SRCX),
[4]Samsung Advanced Institute of Technology (SAIT), South Korea
2020202704@mail.nwpu.edu.cn, {hui01.li, s.rhee, sj75.han, jae-joon.han}@samsung.com
peng.wang@nwpu.edu.cn

In this material, more ablation experimental results are presented using different hyper-parameters. Some more text recognition results are visualized as well. Experiments are still performed using SL_{sm} and RU_{sm}, with TRBA [1] adopted.

1. More Ablation Experiments

1.1. Effect of data augmentation strategies

Data augmentation plays an important role in semi-supervised learning [4–6]. In our framework, we employ StrongAug for online model and WeakAug for target model by default. Specially, StrongAug is borrowed from [3] and contains both color and geometry transformations, while WeakAug is much simple and only consists of color jitter. Their details are shown in Table 1. We conduct experiments to analyze the effect of different data augmentation strategies on model performance and summarize the results in Table 2. Compared with WeakAug, StrongAug is stronger and achieves better results in both supervised (89.9% *vs*. 87.5%) and semi-supervised (90.3% *vs*. 93.2%) settings. Those results show that stronger data augmentations found in supervised learning can be directly applied to semi-supervised learning and achieve better performance.

1.2. Effect of confidence threshold

In our framework, to alleviate the influence caused by noise samples in training process, we filter out noise samples based on their confidence scores calculated from target model. We set the confidence threshold $\beta_U = 0.5$ in our experiments. Here we conduct experiments to show the effect of different confidence thresholds.

As illustrated in Figure 1, a small threshold (0.1) will result in more samples participating in consistency training, but the final test accuracy is only 91.65%, as incorrect recognition makes trouble to model training. In contrast, a larger threshold (0.9) leads to fewer samples involved in consistency training, which cannot take full advantage of training images. Using a confidence threshold of 0.5 achieves the highest test accuracy (average score 93.23%), but 0.3 and 0.7 also work well, with average scores of about 93.2% and 93.2% respectively.

1.3. Effect of low entropy Softmax

We sharpen the output from target STR model $\mathbf{P}^{U_w}$ by using a low Softmax temperature τ as illustrated in Equation (3) in the main paper. We set $\tau = 0.4$ following UDA [5] without specific tuning for our model. A Softmax with $\tau = 1$ causes the average recognition score to fall to 92.1%.

1.4. Ablation on the size of unlabeled data

we train TRBA with 10%SL (SL_{sm}) and RU increasing from 20% to 100% (20% as interval). The resulting average recognition scores are 93.23%, 93.35%, 93.56%, 93.78%, 93.79% respectively. A slightly increasing trend is presented with the addition of unlabeled data.

1.5. Experiments with original TRBA setting

Our reproduced results by using TRBA in supervised training are higher than that reported in the original paper [1]. Besides the adopted StrongAug and improvements made by [2], we train TRBA with more iterations (250K *vs*. 200K), larger batch size (384 *vs*. 128), and larger learning rate ($1e-3$ *vs*. $5e-4$). Here we conduct experiments

*Part of the work was done when C.Zheng was an intern at SRCX.
†P. Wang is the corresponding author.

	Type	Range
Color transformation	Brightness	[0.8, 1.2]
	Contrast	[0.9, 1.1]
	Saturation	[0.9, 1.1]
	Hue	[-0.05, 0.05]

(a) WeakAug

	Type	Range
Color transformation	AutoContrast	-
	Brightness	[0.1, 1.9]
	Color	[0.1, 1.9]
	Contrast	[0.1, 1.9]
	Equalize	-
	Posterize	[4, 7.6]
	Solarize	[0, 230]
	SolarizeAdd	[-99, 99]
Geometry transformation	Rotate	[-14, -10] ∪ [10, 14]
	ShearX	[-0.18, -0.1] ∪ [0.1, 0.18]
	ShearY	[-0.18, -0.1] ∪ [0.1, 0.18]
	TranslateX	[-0.18, -0.1] ∪ [0.1, 0.18]
	TranslateY	[-0.18, -0.1] ∪ [0.1, 0.18]

(b) StrongAug.

Table 1. Details of WeakAug and StrongAug

Aug	Methods	IC13 IC15	SVT SVTP	IIIT CUTE	Avg
-	Sup	94.4 76.9	87.8 79.4	93.1 84.4	87.2
-	Ours	95.7 82.3	89.9 83.6	94.6 91.3	90.0
WeakAug	Sup	94.2 77.8	88.1 78.9	93.7 83.3	87.5
WeakAug	Ours	85.8 82.6	91.2 85.7	93.9 92.7	90.3
StrongAug	Sup	96.0 82.4	90.0 82.6	94.4 88.9	89.9
StrongAug	Ours	97.3 87.0	94.7 89.6	96.2 94.4	93.2

Table 2. Comparison with StrongAug with simple augmentation WeakAug on supervised learning (Sup) and semi-supervised training (Ours). "WeakAug" means replacing StrongAug with WeakAug in online model. "-" means training model without data augmentation.

following the experimental settings in [1] except some necessary changes needed by our consistency training framework. Specially, we adopt Adadelta optimizer with weight decay. StrongAug is kept for the online model of unsupervised branch, while no data augmentation is applied on the supervised branch and target model. Compared to the results reported in the original paper [1], our framework improves TRBA from 82.8% to 90.0% on average score. The specific results on each dataset are presented in Table 3. Note that we test on IC13_1015 and IC15_2077 here, following that used in [1] for fair comparison.

Method	IC13_1015 IC15_2077	SVT SVTP	IIIT CUTE	Avg
Original TRBA [1]	92.3 71.8	87.5 79.2	87.9 74.0	82.8
TRBA-cr (original setting)	94.9 84.0	93.4 89.0	91.5 92.3	90.0

Table 3. Experiments with original TRBA setting. Our framework steadily improves TRBA in this setting.

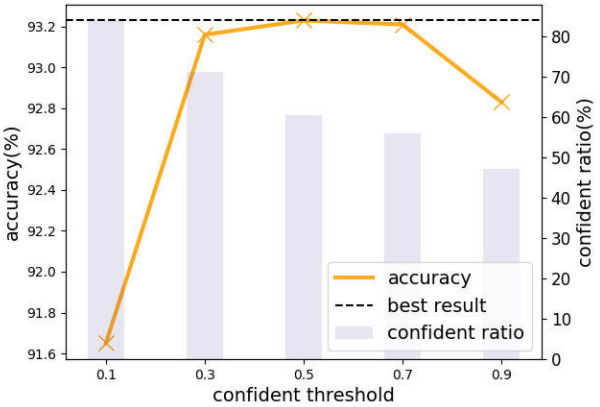


Figure 1. The impact of different confidence threshold on final test accuracy. Confident ratio is the ratio of unlabeled data that participates in consistency training (whose confidence score is higher than β_U). Setting β_U to 0.5 achieves the best result.

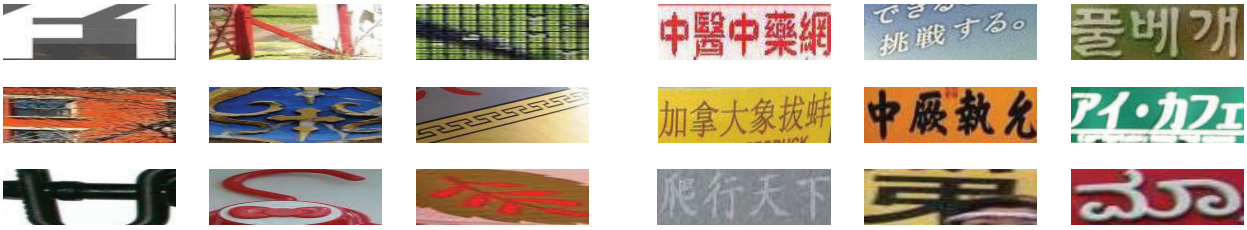
(a) Synthetic labeled data(SL) (b) Real unlabeled data(RU)

Figure 2. Examples of synthetic labeled data (SL) and real unlabeled data (RU). Domain bias exists between the two sets.

2. Examples of Training Images

Some examples from synthetic and real datasets are presented in Figure 2. Although synthetic data integrates with hundreds of fonts and complex backgrounds, there is still domain bias between two data. Figure 3 shows some more examples of our collected real word images. The unlabeled data is detected and cropped from images in various scenarios and is full of different styles and backgrounds. Most word images are normal, but there are also examples with non-character (Figure b) or non-latin characters (Figure c). The used filtering method based on confidence score helps to remove some noise samples but not all of them. Other advanced filtering approaches may be exploited to improve the quality of unlabeled data in the future.

(a) Normal examples.



(b) Non-character examples.

(c) Non-latin characters examples.

Figure 3. Examples of Real Unlabeled Data (RU).

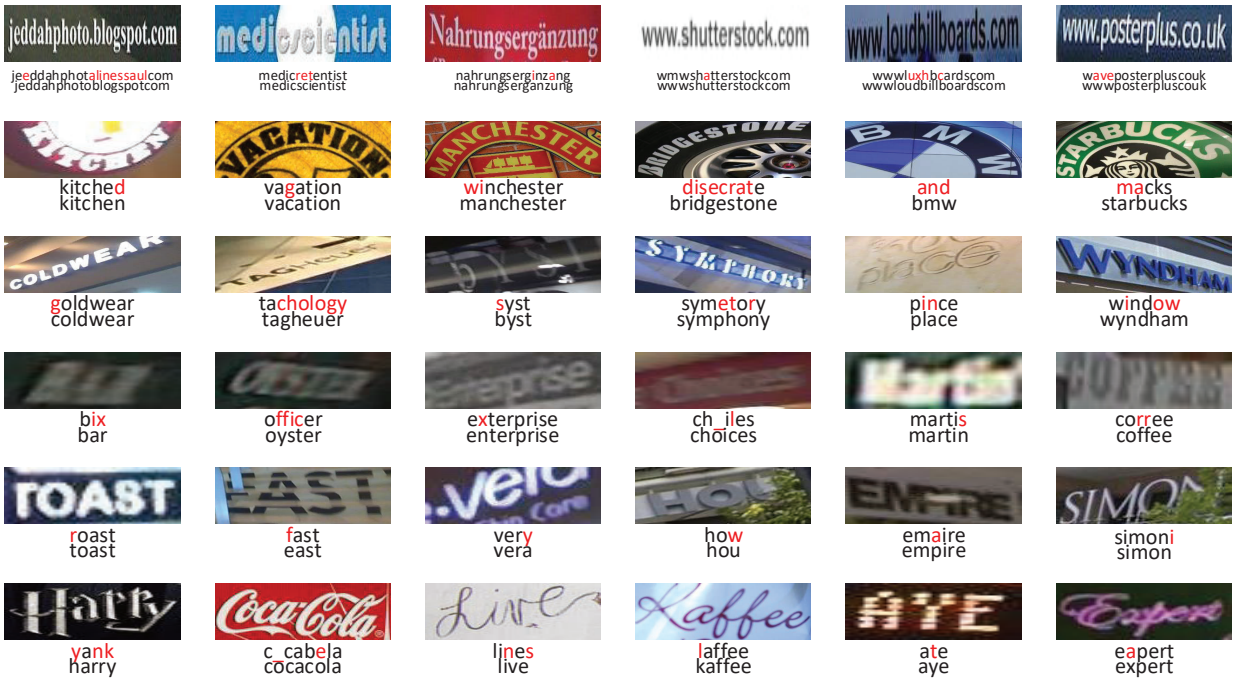


Figure 4. Examples that can be correctly recognized by TRBA after using our consistency regularization training method. The first line shows the recognition results by TRBA after fully supervised learning, which include mistakes (red characters). The second line are results after using our method.

3. Text Recognition Visualization

Figure 4 provides more examples which are wrongly recognized by TRBA in supervised setting but can be correctly recognized after using our consistency regularization training method. The failure of supervised method may be due to long character strings, curved or sloped shapes, blurred images, obscured or damaged images and various

font styles. With the proposed semi-supervised learning, those hard samples can be properly recognized.

References

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *ICCV*, pages 4714–4722, 2019. 1, 2

[2] Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *CVPR*, pages 3113–3122. Computer Vision Foundation / IEEE, 2021. 1

[3] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 3008–3017. Computer Vision Foundation / IEEE, 2020. 1

[4] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 1

[5] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 1

[6] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10684–10695, 2020. 1