Sample-aware Data Augmentor for Scene Text Recognition

Guanghao Meng, Tao Dai, Shudeng Wu, Bin Chen, Jian Lu, Yong Jiang, Shu-Tao Xia
mgh19@mails.tsinghua.edu.cn, daitao.edu@gmail.com, xiast@sz.tsinghua.edu.cn

Problem

DNN-based recognizers require a huge amount of labeled data for training, but data collection and annotation is usually cost-expensive and time consuming in practice. Existing models attempt to alleviate such problems by using data augmentation for training. However, we found that existing data augmentation strategies for scene text images suffer from the problems of under- and over-diversity, due to the complexity of text contents and shapes.

Method

We propose a sample-aware data augmentor to balance the *diversity* and *affinity* of samples. Our data augmentor consists of three parts: gated module (GM), affine transformation module (ATM), and elastic transformation module (ETM). Specifically, GM can choose the transformation type adaptively based on input samples. ATM aims to keep the affinity of samples by performing the linear transformation on samples, while ETM aims to improve the local diversity of samples by the non-linear transformation on samples. In addition, we design a loss function for the data augmentor based on the learning progress of the scene text recognizer, and the data augmentor and the recognizer can be optimized jointly.

Contribution

We propose a sample-aware data augmentation framework for scene text images. To the best of our knowledge, this is the first work that integrates the affine and the elastic transformation methods in a unified framework; Our data augmentor mainly consists of three parts: gated module, affine transformation module, and elastic transformation module. Moreover, we design a loss function for the data augmentor based on the learning progress of the scene text recognizer; Extensive experiments on various benchmarks show that our data augmentation framework significantly improves the performance of the scene text recognizer.

References

[1] Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation policies from data." arXiv preprint arXiv:1805.09501 (2018).

[2] Shi, Baoguang, et al. "Aster: An attentional scene text recognizer with flexible rectification." IEEE transactions on pattern analysis and machine intelligence 41.9 (2018): 2035-2048.

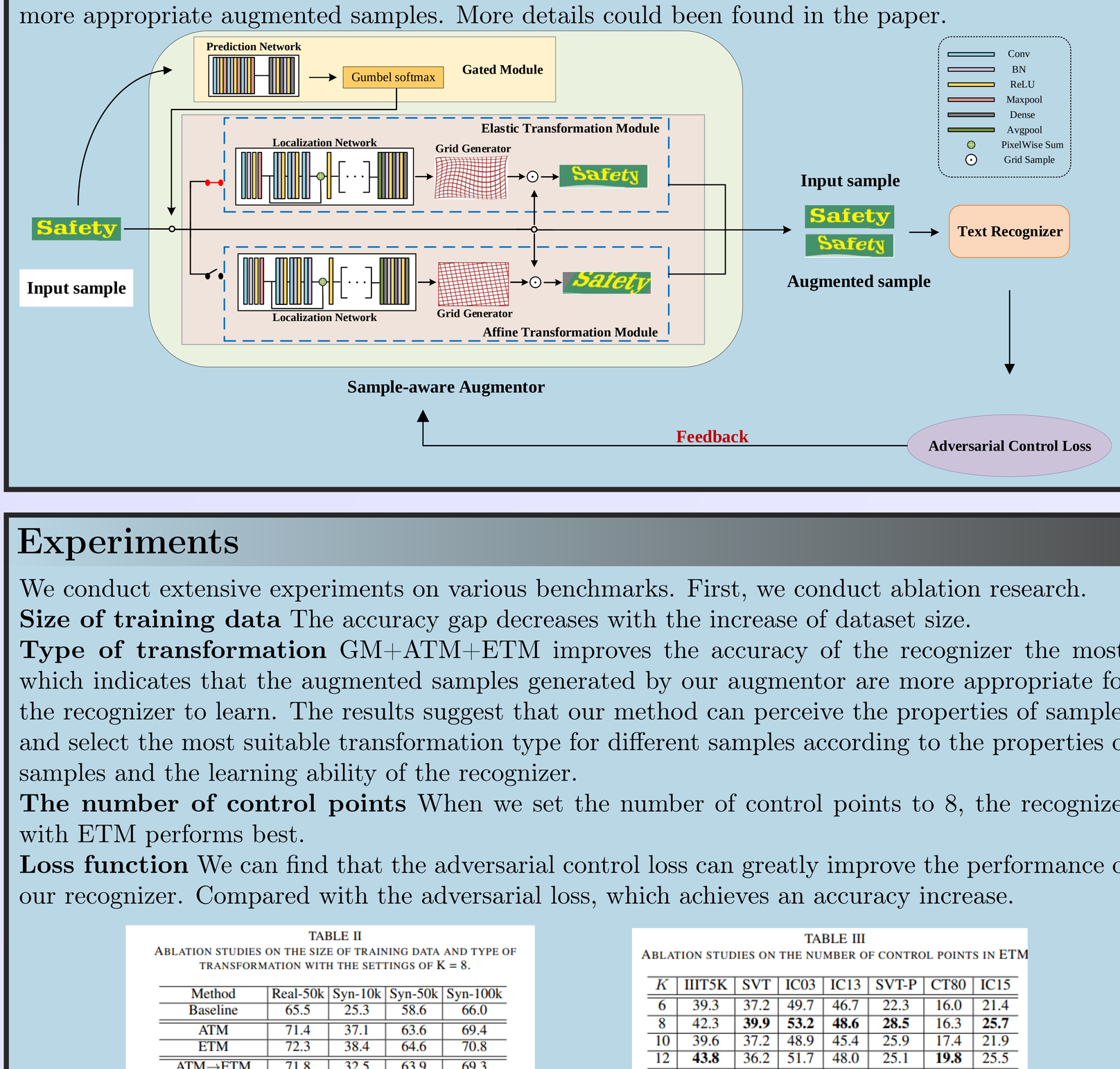
[3] Luo, Canjie, et al. "Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant 61771273, Guangdong Basic and Applied Basic Research Foundation on 2019A1515110344, the China Postdoctoral Science Foundation under Grant 2019M660645, the R&D Program of Shenzhen under Grant JCYJ20180508152204044, and the project "PCL Future Greater-Bay Area Network Facilities for Large-scale Experiments and Applications (LZC0019)".

Proposed Architecture

A reasonable data augmentor for scene text recognition should take both diversity and affinity of augmented samples into account. To this end, we propose a sample-aware data augmentor. The overall architecture of our data augmentation framework is shown in Fig. 1, which mainly consists of three parts: gated module (GM), affine transformation module (ATM), and elastic transformation module (ETM). We send input samples to the gated module to predict the transformation type. According to the prediction of the gated module, the corresponding transformation will be performed. We utilize the spatial transformer network (STN) to perform the differentiable image transformation. Finally, the augmented sample and the input sample are sent to the recognizer for recognition. In addition, we design a loss function for the data augmentor based on the learning progress of the scene text recognizer, and thus the data augmentor and the recognizer can be optimized jointly. The loss obtained by the recognizer is fed back to the augmentor to guide the data augmentor to generate more appropriate augmented samples. More details could been found in the paper.



Experiments

We conduct extensive experiments on various benchmarks. First, we conduct ablation research.
Size of training data The accuracy gap decreases with the increase of dataset size.
Type of transformation GM+ATM+ETM improves the accuracy of the recognizer the most, which indicates that the augmented samples generated by our augmentor are more appropriate for the recognizer to learn. The results suggest that our method can perceive the properties of samples and select the most suitable transformation type for different samples according to the properties of samples and the learning ability of the recognizer.
The number of control points When we set the number of control points to 8, the recognizer with ETM performs best.
Loss function We can find that the adversarial control loss can greatly improve the performance of our recognizer. Compared with the adversarial loss, which achieves an accuracy increase.

TABLE II
ABLATION STUDIES ON THE SIZE OF TRAINING DATA AND TYPE OF TRANSFORMATION WITH THE SETTINGS OF K = 8.

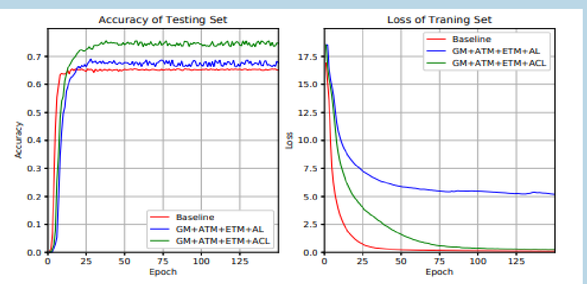
Method	Real-50k	Syn-10k	Syn-50k	Syn-100k
Baseline	65.5	25.3	58.6	66.0
ATM	71.4	37.1	63.6	69.4
ETM	72.3	38.4	64.6	70.8
ATM→ETM	71.8	32.5	63.9	69.3
ETM→ATM	71.3	33.7	64.3	70.3
ATM+ETM	73.4	39.5	65.1	71.0
GM+ATM+ETM	**74.6**	**41.6**	**66.0**	**71.8**

TABLE III
ABLATION STUDIES ON THE NUMBER OF CONTROL POINTS IN ETM

K	IIIT5K	SVT	IC03	IC13	SVT-P	CT80	IC15
6	39.3	37.2	49.7	46.7	22.3	16.0	21.4
8	42.3	**39.9**	**53.2**	**48.6**	**28.5**	16.3	**25.7**
10	39.6	37.2	48.9	45.4	25.9	17.4	21.9
12	**43.8**	36.2	51.7	48.0	25.1	**19.8**	25.5
14	38.5	37.6	52.4	44.8	25.6	13.9	24.3

TABLE IV
ABLATION STUDIES ON THE LOSS FUNCTION. THE TRAINING DATASET IS REAL-50K.

Loss function	ATM	ETM	GM+ATM+ETM
ℓ_A	68.2	67.4	69.1
ℓ_{AC}	**71.4**	**72.3**	**74.6**



Finally, we combine state-of-the-art recognizers with our method to show the effectiveness of our augmentor. When the ASTER is equipped with our method, it has an increase of 1.3%, 1.4%, and 4.1% on IC15, SVT-P, and CT80 respectively. Our method has an increase of 1.3% on the IC15 dataset, higher than the increase of 0.3% in Luo et al.[3]. That means, compared with the improvement with Luo et al.[3], the recognizer equipped with our method can increase about 18 correctly recognized images. Although the improvement with our method on SVTP and CT80 is slightly lower than that of Luo et al.[3], considering the size of the datasets, there are only the distinguish of one image on the two datasets. More augmented samples are shown in Fig.5.

TABLE I
WORD ACCURACY ON IRREGULAR TEXT. "REIM" IS OUR RE-IMPLEMENTED ASTER. "*" DENOTES THE RESULT IS FROM RE-IMPLEMENTATION OF OTHERS.

Method	Irregular Text		
	IC15	SVT-P	CT80
ASTER* [3]	75.8	77.7	79.9
+ Luo et al.[3]	76.1	79.2	84.4
ASTER(*ReIm*)	77.3	80.3	80.6
+ours	**78.6**	**81.7**	**84.7**

Fig. 5. Visualization of augmented samples (right) on scene text images.