Sample-aware Data Augmentor for Scene Text Recognition

Guanghao Meng, Tao Dai, Shudeng Wu, Bin Chen, Jian Lu, Yong Jiang, Shu-Tao Xia

Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China

mgh19@mails.tsinghua.edu.cn, daitao.edu@gmail.com, xiast@sz.tsinghua.edu.cn

CATALOG

- Motivation
- Contribution
- Architecture
- Experiments

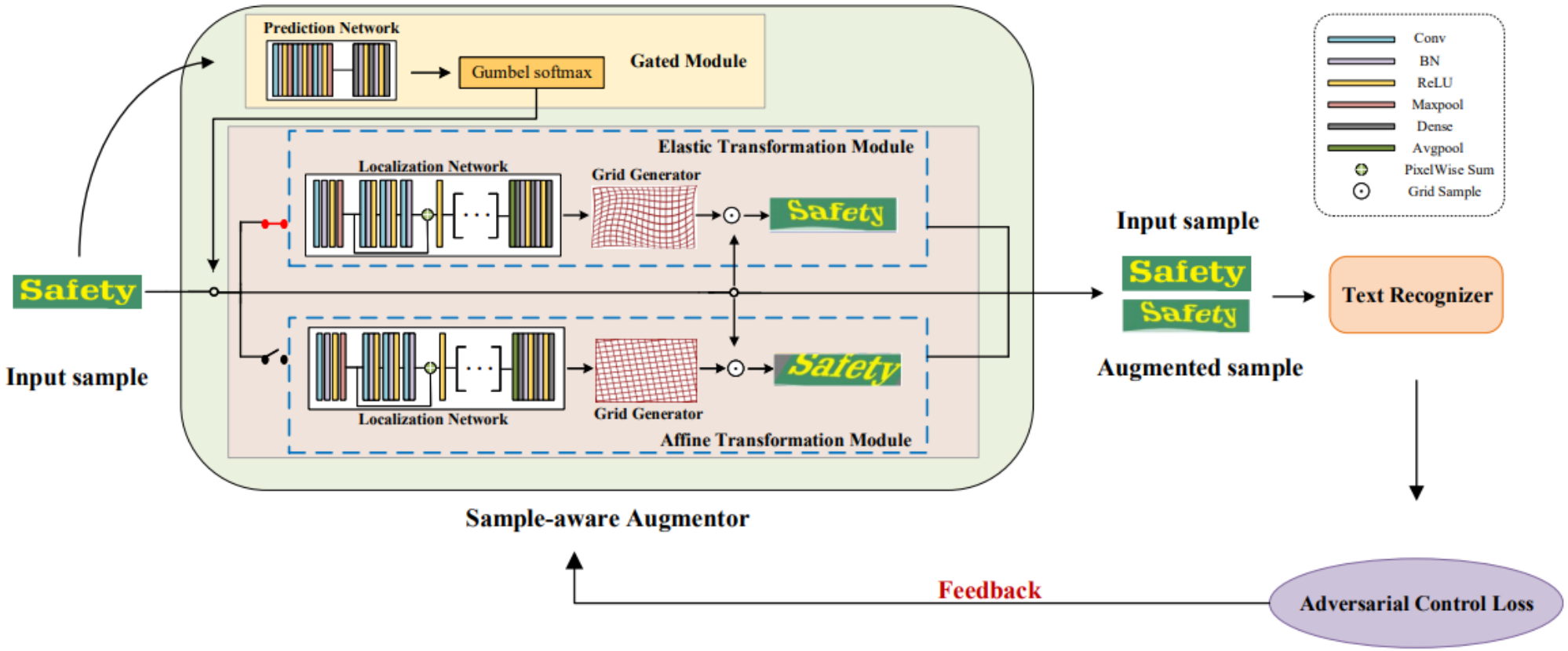
Motivation

- 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.
- 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.
- To address this issue, we propose a sample-aware augmentor to balance the diversity and affinity of samples.

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. Thus our data augmentor adaptively generates the augmented samples based on the properties of training samples and the recognition capability.
- Extensive experiments on various benchmarks show that our data augmentation framework significantly improves the performance of the state-of-the-art scene text recognizer.

Architecture



The overall architecture of our data augmentation framework.

Architecture

- Gated Model

$$X = \arg\max_{k \in \{1,2\}} (\alpha_k + G_k) \qquad \hat{X}_k = soft\max\left((\alpha_k + G_k)/\tau\right)$$

- Affine Transformation Module

$$A_\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix}$$

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

$$I_i' = \sum_h^H \sum_w^W I_{hw} \max(0, 1 - |x_i^s - w|) \max(0, 1 - |y_i^s - h|)$$

Architecture

- Elastic Transformation Module

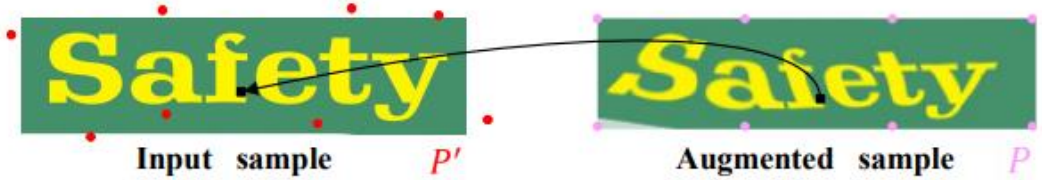


Fig. 3. Text image transformation with TPS. The black arrow represents the transformation Γ_θ, connecting the points on the augmented sample and points on the input sample. P' is the control points on the input sample. P is the control points on the augmented sample.

$$\Gamma_\theta = \left(\Delta_{P'}^{-1}\begin{bmatrix} P^T \\ 0^{3\times 2} \end{bmatrix}\right)^T$$

$$\Delta_{P'} = \begin{pmatrix} 1^{K\times 1} & P'^T & \varepsilon \\ 0 & 0 & 1^{1\times K} \\ 0 & 0 & P' \end{pmatrix}$$

$$s_i = \Gamma_\theta \cdot \hat{s}_i'$$

- Adversarial control Loss(ACL)

$$\ell_{AC} = |1.0 - \exp\left[L\left(P'\right) - \alpha L\left(P\right)\right]|$$

$$\alpha = \max(1, \exp(\sum_{k=1}^{K} \hat{y}_k \cdot y_k^G))$$

Experiments

- **Ablation Study**

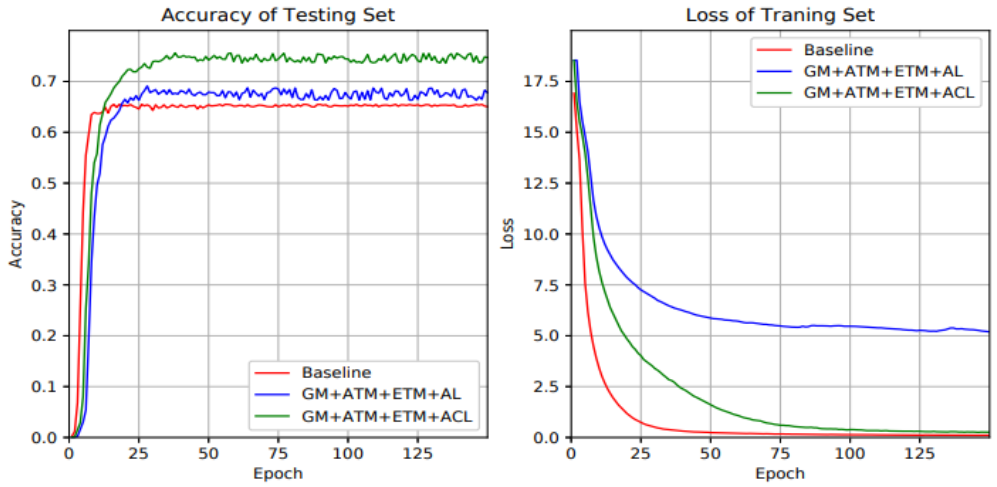


TABLE II
ABLATION STUDIES ON THE SIZE OF TRAINING DATA AND TYPE OF TRANSFORMATION WITH THE SETTINGS OF K = 8.

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**

Experiments

- **Integration with State-of-the-art Methods**

TABLE V
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
Shi, Bai, and Yao [32]	-	66.8	54.9
Shi et al. [33]	-	71.8	59.2
Liu et al. [34]	-	73.5	-
Yang et al. [35]	-	75.8	69.3
Cheng et al. [36]	70.6	71.5	63.9
Liu, Chen, and Wong [37]	60.0	73.5	-
Cheng et al [38]	68.2	73.0	76.8
Bai et al. [39]	73.9	-	-
Liu et al. [40]	-	73.9	62.5
Luo, Jin, and Sun [41]	68.8	76.1	77.4
Liao et al. [2]	-	-	79.9
ASTER [16]	76.1	78.5	79.5
ASTER* [10]	75.8	77.7	79.9
+ Luo et al. [10]	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.

THANKS