A Semi-Supervised Learning Approach for Traditional Chinese Scene Text Detection

Chia-Fu Yeh and Mei-Chen Yeh

Department of Computer Science and Information Engineering
National Taiwan Normal University

Traditional Chinese Scene Text Detection

Samples of traditional Chinese scene texts. Source: the traditional Chinese scene text dataset provided by MOE AI competition and labeled data acquisition project.

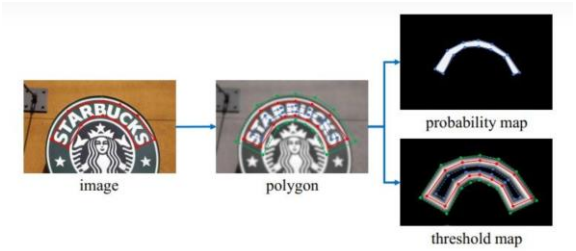
Traditional Chinese Scene Text Detection

Samples of traditional Chinese scene texts. Source: the traditional Chinese scene text dataset provided by MOE AI competition and labeled data acquisition project.

EAST (*CVPR*'17)

AE TextSpotter (*ECCV*'20)

DBNet (*AAAI*'20)



- Current methods focus primarily on English texts.
- 5,000 to 6,000 characters in a typical Chinese dictionary
- The cost of annotating training samples is high.

A semi-supervised traditional Chinese scene text detector

Synthesize new samples from annotated samples

Burn-In Stage

Teacher-Student Mutual Learning Stage

(a) Train Detector

Labeled Data

Detector

(b) Generate pseudo labels and teacher student mutual learning

Labeled + Synthesized Data

Strong Data Augmentations

Student

Unlabeled Data

Teacher

Pseudo Labels

EMA Update

Liu et al., Unbiased teacher for semi-supervised object detection. *International Conference on Learning Representations*, 2021.

Data Augmentations

- Color Jitter
- Scaling
- Horizontal Flip
- Mosaic

Original

Horizontal Flip

Color Jitter

Scaling

Mosaic

Data Synthesis

Experiments

Dataset (AI CUP 2021)

	#labeled samples	#unlabeled samples
Training	2,800	6,000
Validation	1,200	
Test	500	
Total	4,500	6,000

Evaluation Metrics

$$\textbf{precision} = \frac{\#\text{correct predictions}}{\#\text{predictions}} \qquad \textbf{recall} = \frac{\#\text{correct predictions}}{\#\text{ground truth}} \qquad \textbf{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Experimental results

Base detection model: YOLOv5m (small)

Method	Precision	Recall	F1-score
Supervised	89.47	69.77	78.40
Multi-stage	89.28	72.80	80.20 (+1.80)
End-to-end (ours)	88.82	76.07	81.95 (+3.55)

Base detection model: YOLOv5m6 (large)

Method	Precision	Recall	F1-score
Supervised	88.76	78.47	83.30
Multi-stage	89.06	80.26	84.43 (+1.13)
End-to-end (ours)	87.66	83.06	85.30 (+2.00)

Ablation experiments

Effects of Mosaic

Mosaic	Labeled Data	Precision	Recall	F1-score
		88.67	69.72	78.06
✔		86.44	73.02	79.16
	✔	88.48	72.16	79.49
✔	✔	**88.82**	**76.07**	**81.95**

Effects of Synthesized Data

Amount of synth.	Precision	Recall	F1-score
Burn-In Stage			
0x	89.47	69.77	78.40
1x	88.20	72.46	79.56
2x	89.54	72.27	79.98
3x	89.68	72.78	80.35
Teacher-Student Mutual Learning Stage			
0x	88.82	76.07	81.95
1x	88.85	76.35	82.13
2x	89.53	75.85	82.13

Summary

- We present a semi-supervised traditional Chinese scene text detector.
- A teacher student mutual learning framework is developed in which pseudo labels computed from unlabeled data can be refreshed and reused.
- Data reconstitution including data synthesis and mosaic further improves the detection performance.