AAAI-22

Context-based contrastive learning for scene text recognition

Xinyun Zhang, Binwu Zhu, Xufeng Yao, Qi Sun, Ruiyu Li, Bei Yu. *The Chinese University of Hong Kong & SmartMore*.

Introduction

Vocabulary Reliance

SIMON

FDDB

- ✦ Much more error-prone on out-of-vocabulary text, even though the image quality is good.

benchmarks / img number	IC13	SVT	IIIT	IC15	SVTP	CUTE80
in-vocabulary	1015	647	2588	1493	645	241
out-of-vocabulary	0	0	412	318	0	47

- ✦ Current benchmarks are infeasible to reveal models' out-of-vocabulary generalization ability.

Our Contributions:

- ✦ We propose a contrastive learning-based method, ConCLR, to improve attention-based recognizers' out-of-vocabulary generalization ability.
- ✦ We synthesize a benchmark, OutText, to fairly evaluate models' performance on unseen data.

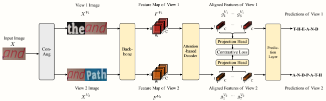
Benchmarks

CUTE IC13 IC15 IIIT SVT SVTP OutText

- ✦ Six common benchmarks plus OutText.
- ✦ All images in OutText are out-of-vocabulary and free of distortions.

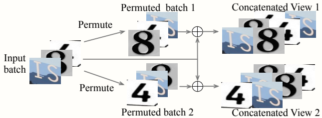
Context-based Contrastive Learning

Main Framework



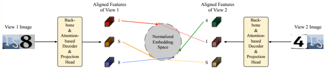
- ✦ Step 1: ConAug generates context-based augmented views.
- ✦ Step 2: Parallel attention-based decoder outputs the charater representations.
- ✦ Step 3: Contrastive loss is optimized on the projected character representations.

Context-based Data Augmentation



- ✦ Random permute and concatenate each batch of data.

Context-based Contrastive Loss



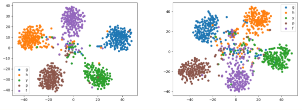
- ✦ Pull together the positive samples and push apart the negative samples.

Results

Analysis on seen and unseen data:

Method	IC13	SVT	IIIT	IC15	SVTP	CUTE	AVG	OutText
Baseline	94.7 NA	90.1 NA	96.5 85.2	85.9 63.9	82.9 NA	88.4 76.6	89.8	63.2
Baseline-ConCLR	**95.9** NA	**92.1** NA	96.6 **89.7**	**88.7** **64.8**	**85.7** NA	**90.0** **85.1**	**91.4**	**67.7**

Feature visualization



ConCLR Baseline

Ablation study: Concatenation

Concat	IC13	SVT	IIIT	IC15	SVTP	CUTE	AVG	OutText
SingleCat	95.4 NA	90.6 NA	96.8 88.4	87.6 64.2	83.3 NA	91.3 **87.2**	90.9	66.3
FixCat	95.2 NA	91.3 NA	**97.0** **89.9**	88.2 62.3	84.5 NA	**91.7** 85.1	91.2	67.2
RandCat	**95.9** NA	**92.1** NA	96.6 89.7	**88.7** **64.8**	**85.7** NA	90.0 85.1	**91.4**	**67.7**

Comparison with SOTA

Methods	Training Data	Annos	IIIT	IC13	SVT	IC15	SVTP	CUTE
ESIR Zhan and Lu 2019	MJ+ST	word	93.3	91.3	90.2	76.9	79.6	83.3
ASTER Shi, Bai, and Yao 2017	MJ+ST	word	93.4	91.8	89.5	76.1	78.5	79.5
RobustScanner Yue et al. 2020	MJ+ST	word	95.3	88.1	94.8	77.1	79.5	90.3
SAR Li et al. 2019	MJ+ST	word	91.5	91.0	84.5	69.2	76.4	83.3
DAN Wang et al. 2020	MJ+ST	word	94.3	93.9	89.2	74.5	80.0	84.4
SRN Yu et al. 2020	MJ+ST	word	94.8	95.5	91.5	82.7	85.1	87.8
SEED Qiao et al. 2020	MJ+ST	word	93.8	92.8	89.6	80.0	81.4	83.6
ABINet Fang et al. 2021	MJ+ST	word	96.2	97.4	93.5	86.0	89.3	89.2
ABINet-Vision†	MJ+ST	word	95.0	94.7	90.1	81.9	82.9	86.5
ABINet-Vision-ConCLR	MJ+ST	word	95.7	95.9	92.1	84.4	85.7	89.2
ABINet†	MJ+ST	word	95.7	**97.7**	93.9	84.9	88.5	87.5
ABINet-ConCLR	MJ+ST	word	**96.5**	**97.7**	**94.3**	**85.4**	**89.3**	**91.3**