Scene Text Recognition

Xiangcheng Du

Scene Text Recognition

文本识别论文整理					
推荐阅读顺序	标题简称	标题	发表位置	推荐指数	简介
1	CRNN	An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition	TPAMI-2017	★★★★★	字符识别的开山之作，首次将CNN与RNN结合在一起，后续工作的模型结构基本沿用CRNN的框架
2	RARE	Robust scene text recognition with automatic rectification	CVPR-2016	★★★★★	文本矫正的开山之作，使用STN作为不规则文本的矫正器
3	MORAN	MORAN: A multi-object rectified attention network for scene text recognition	PR-2019	★★★	一个用于文本校正的框架
4	AON	AON: Towards arbitrarily-oriented text recognition	CVPR-2018	★★★★	简单且易于复现，解决了90度旋转与颠倒文本的识别问题
5	FAN	Focusing attention: Towards accurate text recognition in natural images	ICCV-2017	★★★★	在文本预测时，对注意力进行监督
6	SAR	Show, attend and read: A simple and strong baseline for irregular text recognition	AAAI-2019	★★★★★	将二维注意力图运用在OCR中，可以解决不规则文本问题
7		What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis	ICCV-2019	★★★★★	揭露了OCR领域论文的阴暗面
8		Text Recognition in the Wild: A Survey	Arxiv-2020	★★★★	文本识别的综述
9	CharNet	Char-Net: A character-aware neural network for distorted scene text recognition	CVPR-2018	★★	单字符级别的文本校正
10	DAN	Decoupled attention network for text recognition	CVPR-2020	★★★	使用CNN代替RNN，提供了一个新思路，CNN也能提取语义信息
11	GRCNN	Gated recurrent convolution neural network for OCR	NIPS-2017	★	提出了一个新backbone
12	Mask textspotter	Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary	TPAMI-2019	★★★	结合检测框架的端到端文本识别
13	NRTR	NRTR: A no-recurrence sequence-to-sequence model for scene text recognition	ICDAR-2019	★★★	将NLP的Transformer框架用于OCR中，实现预测并行化
14	SEED	SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition	CVPR-2020	★★★★	将语义信息结合进文本识别
15	SRN	Towards accurate scene text recognition with semantic reasoning networks	CVPR-2020	★★★	灵活的将语义信息结合进文本识别中，结构比较复杂
16	SCATTER	SCATTER: selective context attentional scene text recognizer	CVPR-2020	★★	使用多个LSTM加强语义信息
17		Separating content from style using adversarial learning for recognizing text in the wild	Arxiv-2020	★★★	鲜有的场景文本去背景文章
18		Synthetically Supervised Feature Learning for Scene Text Recognition	ECCV-2018	★★	提供了一个场景文本去背景的框架
19	LAL	LAL: Linguistically Aware Learning for Scene Text Recognition	ACM-2020	★★	结合语义信息
20		On Vocabulary Reliance in Scene Text Recognition	CVPR-2020	★★	研究文本识别中，out of vocab的问
21		Scene Text Image Super-Resolution in the Wild	ECCV-2020	★★★★	将识别与文本超分相结合的开山之作
22	PlugNet	PlugNet: Degradation Aware Scene Text Recognition Supervised by a Pluggable Super-Resolution	ECCV-2020	★★★	一个简单的文本超分框架
23		Exploring Font-independent Features for Scene Text Recognition	MM-2020	★★★	解决了场景文本中多字体的问题
24	Strokelets	Strokelets: A Learned Multi-Scale Representation for Scene Text Recognition	CVPR-2014	★	较早的文章，从笔画角度处出发
25		Sequence-to-sequence domain adaptation network for robust text image recognition	CVPR-2019	★★	解决了OCR中的域迁移问题
26		What Machines See Is Not What They Get: Fooling Scene Text Recognition Models with Adversarial Text Images	CVPR-2020	★★	研究OCR中的对抗样本
27		Scene text recognition from two-dimensional perspective	AAAI-2019	★	一个简单的二维注意力方案
28	EPAN	EPAN: Effective parts attention network for scene text recognition	NC-2020	★	多次注意力
29	REELFA	ReELFA: A scene text recognizer with encoded location and focused attention	ICDARW-2019	★	将位置编码结合进框架
30		Scene text recognition with sliding convolutional character models	ICCV-2017	★	使用CNN代替RNN

Outline

- **Background**
- Scene Text Detection
- Scene Text Recognition
- Applications
- Current Issues
- Future Trends

Background

- Definitions

Summary Booklet

Scene text detection

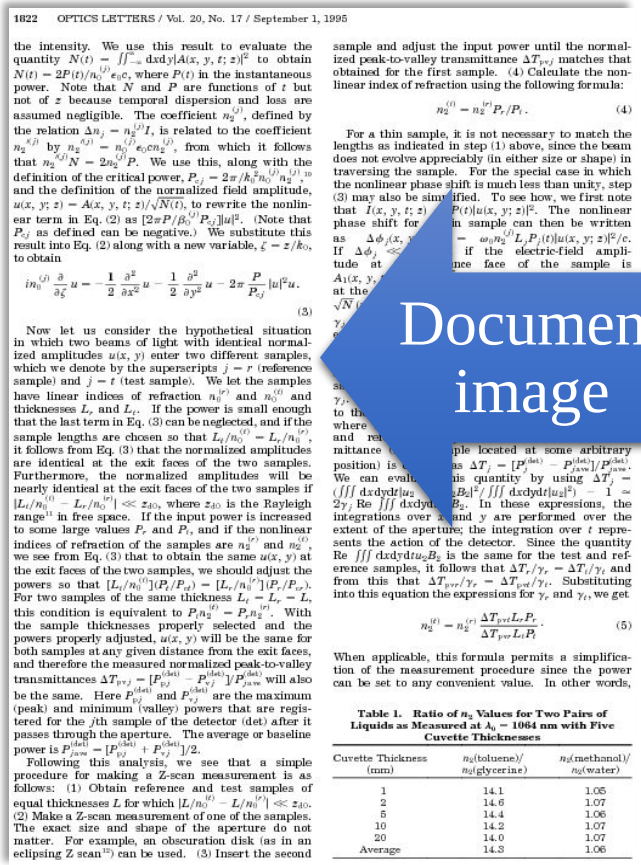
Predicting the presence of text and localizing each instance (if any), usually at word or line level, in natural scenes

End-to-end recognition

Scene text recognition

Converting text regions into computer readable and editable symbols

Background



1822 OPTICS LETTERS / Vol. 20, No. 17 / September 1, 1995

the intensity. We use this result to evaluate the quantity $N(t) = \iint_{-\infty} dxdy|A(x, y, t; z)|^2$ to obtain $N(t) = 2P(t)/n_0^{(j)}\epsilon_0 c$, where $P(t)$ in the instantaneous power. Note that N and P are functions of t but not of z because temporal dispersion and loss are assumed negligible. The coefficient $n_2^{(j)}$, defined by the relation $\Delta n_j = n_2^{(j)} I$, is related to the coefficient $n_2^{s(j)}$ by $n_2^{s(j)} = n_0^{(j)}\epsilon_0 c n_2^{(j)}$, from which it follows that $n_2^{s(j)} N = 2n_2^{(j)} P$. We use this, along with the definition of the critical power, $P_{cj} = 2\pi/k_0^2 n_0^{(j)} n_2^{(j)}$, and the definition of the normalized field amplitude, $u(x, y; z) = A(x, y, t; z)/\sqrt{N(t)}$, to rewrite the nonlinear term in Eq. (2) as $[2\pi P/\beta_0^{(j)} P_{cj}]|u|^2$. (Note that P_{cj} as defined can be negative.) We substitute this result into Eq. (2) along with a new variable, $\xi = z/k_0$, to obtain

$$in_0^{(j)}\frac{\partial}{\partial\xi}u = -\frac{1}{2}\frac{\partial^2}{\partial x^2}u - \frac{1}{2}\frac{\partial^2}{\partial y^2}u - 2\pi\frac{P}{P_{cj}}|u|^2 u. \quad (3)$$

Now let us consider the hypothetical situation in which two beams of light with identical normalized amplitudes $u(x, y)$ enter two different samples, which we denote by the superscripts $j = r$ (reference sample) and $j = t$ (test sample). We let the samples have linear indices of refraction $n_0^{(r)}$ and $n_0^{(t)}$ and thicknesses L_r and L_t. If the power is small enough that the last term in Eq. (3) can be neglected, and if the sample lengths are chosen so that $L_t/n_0^{(t)} = L_r/n_0^{(r)}$, it follows from Eq. (3) that the normalized amplitudes are identical at the exit faces of the two samples. Furthermore, the normalized amplitudes will be nearly identical at the exit faces of the two samples if $|L_t/n_0^{(t)} - L_r/n_0^{(r)}| \ll z_{R0}$, where z_{R0} is the Rayleigh range[11] in free space. If the input power is increased to some large values P_r and P_t, and if the nonlinear indices of refraction of the samples are $n_2^{(r)}$ and $n_2^{(t)}$, we see from Eq. (3) that to obtain the same $u(x, y)$ at the exit faces of the two samples, we should adjust the powers so that $[L_t/n_0^{(t)}](P_t/P_{ct}) = [L_r/n_0^{(r)}](P_r/P_{cr})$. For two samples of the same thickness $L_t = L_r = L$, this condition is equivalent to $P_t n_2^{(t)} = P_r n_2^{(r)}$. With the sample thicknesses properly selected and the powers properly adjusted, $u(x, y)$ will be the same for both samples at any given distance from the exit faces, and therefore the measured normalized peak-to-valley transmittances $\Delta T_{pvj} = [P_{pj}^{(det)} - P_{vj}^{(det)}]/P_{jave}^{(det)}$ will also be the same. Here $P_{pj}^{(det)}$ and $P_{vj}^{(det)}$ are the maximum (peak) and minimum (valley) powers that are registered for the jth sample of the detector (det) after it passes through the aperture. The average or baseline power is $P_{jave}^{(det)} = [P_{pj}^{(det)} + P_{vj}^{(det)}]/2$.

Following this analysis, we see that a simple procedure for making a Z-scan measurement is as follows: (1) Obtain reference and test samples of equal thicknesses L for which $|L/n_0^{(t)} - L/n_0^{(r)}| \ll z_{R0}$. (2) Make a Z-scan measurement of one of the samples. The exact size and shape of the aperture do not matter. For example, an obscuration disk (as in an eclipsing Z scan[12]) can be used. (3) Insert the second sample and adjust the input power until the normalized peak-to-valley transmittance ΔT_{pvr} matches that obtained for the first sample. (4) Calculate the nonlinear index of refraction using the following formula:

$$n_2^{(t)} = n_2^{(r)} P_r/P_t. \quad (4)$$

For a thin sample, it is not necessary to match the lengths as indicated in step (1) above, since the beam does not evolve appreciably (in either size or shape) in traversing the sample. For the special case in which the nonlinear phase shift is much less than unity, step (3) may also be simplified. To see how, we first note that $I(x, y, t; z) = P(t)|u(x, y; z)|^2$. The nonlinear phase shift for [illegible] in sample can then be written as $\Delta\phi_j(x, [illegible] = \omega_0 n_2^{(j)} L_j P_j(t)|u(x, y; z)|^2/c$. If $\Delta\phi_j \ll$ [illegible] if the electric-field amplitude at [illegible]nce face of the sample is $A_1(x, y,$ [illegible] at the [illegible] $\sqrt{N}$ [illegible] γ_j [illegible] to th[illegible] where [illegible] and re[illegible] mittance [illegible]ple located at some arbitrary position) is [illegible] as $\Delta T_j = [P_j^{(det)} - P_{jaw}^{(det)}]/P_{jave}^{(det)}$. We can evalu[illegible]is quantity by using $\Delta T_j = (\iiint dxdydt|u_2|^2 |B_2|^2 / \iiint dxdydt|u_2|^2) - 1 \approx 2\gamma_j \,\mathrm{Re}\iiint dxdyd[illegible]B_2$. In these expressions, the integrations over x and y are performed over the extent of the aperture; the integration over t represents the action of the detector. Since the quantity $\mathrm{Re}\iiint dxdydtu_2 B_2$ is the same for the test and reference samples, it follows that $\Delta T_r/\gamma_r = \Delta T_t/\gamma_t$ and from this that $\Delta T_{pvr}/\gamma_r = \Delta T_{pvt}/\gamma_t$. Substituting into this equation the expressions for γ_r and γ_t, we get

$$n_2^{(t)} = n_2^{(r)}\frac{\Delta T_{pvt} L_r P_r}{\Delta T_{pvr} L_t P_t}. \quad (5)$$

When applicable, this formula permits a simplification of the measurement procedure since the power can be set to any convenient value. In other words,

Table 1. Ratio of n_2 Values for Two Pairs of Liquids as Measured at $\lambda_0 = 1064$ nm with Five Cuvette Thicknesses

Cuvette Thickness (mm)	n_2(toluene)/ n_2(glycerine)	n_2(methanol)/ n_2(water)
1	14.1	1.05
2	14.6	1.07
5	14.4	1.06
10	14.2	1.07
20	14.0	1.07
Average	14.3	1.06

Document image

VS

Scene text image

- Scattered and sparse
- Multi-oriented
- Multi-lingual

Background

Different types of irregular scene text

Perspective text

Curved text

Vertical text

Outline

- Background
- **Scene Text Detection**
- Scene Text Recognition
- Applications
- Current Issues
- Future Trends

Scene Text Detection

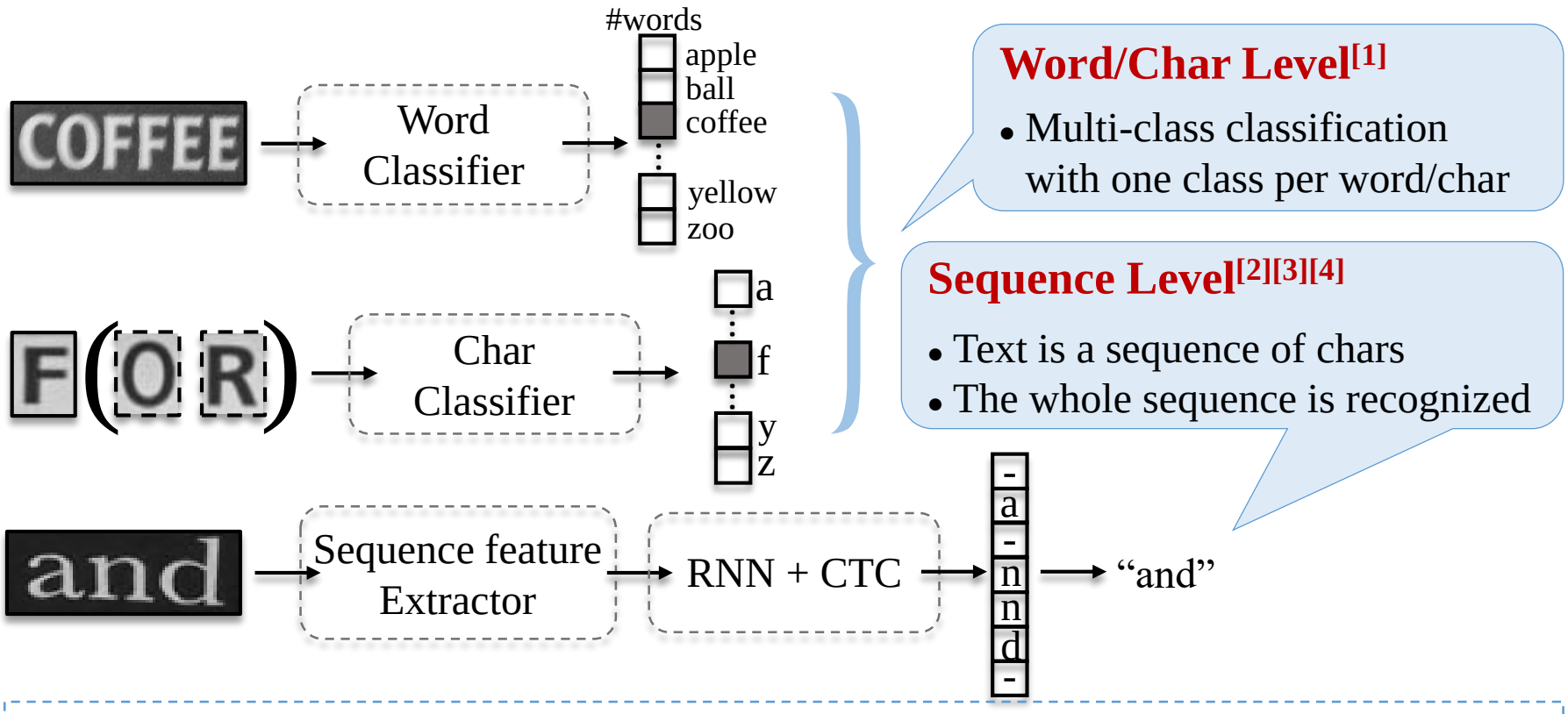
- M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.
- B. Shi et al. Detecting Oriented Text in Natural Images by Linking Segments. IEEE CVPR, 2017.
- Zhou et al., EAST: An Efficient and Accurate Scene Text Detector CVPR 2017
- Yixing Zhu, et al. TextMountain: Accurate Scene Text Detection via Instance Segmentation. Arxiv, 2018
- Shangbang Long, et al. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. ECCV, 2018
- Enze Xie, et al. Scene Text Detection with Supervised Pyramid Context Network. In AAAI 2019.
- Wang X, et al. Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation. CVPR, 2019.
- Baek Y, et al. Character Region Awareness for Text Detection. CVPR, 2019
- Yuxin Wang, et al. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. CVPR, 2020

Outline

- Background
- Scene Text Detection
- **Scene Text Recognition**
- Applications
- Current Issues
- Future Trends

Scene Text Recognition

Scene text recognition methods



[1] M. Jaderberg et al. Reading text in the wild with convolutional neural networks. IJCV, 2016.
[2] B. Su et al. Accurate scene text recognition based on recurrent neural network. ACCV, 2014.
[3] He et al. Reading Scene Text in Deep Convolutional Sequences. AAAI, 2016.
[4] Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

Scene Text Recognition

- **CRNN**[1] model for Regular Text Recognition
- **RARE**[2] model for Irregular Text Recognition
- **ASTER**[3] model for Irregular Text Recognition
- **AON**[4] model for Irregular Text Recognition
- **SAR**[5] model for Irregular Text Recognition

[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

[2] RARE: Shi B et al. Robust scene text recognition with automatic rectification. CVPR, 2016.

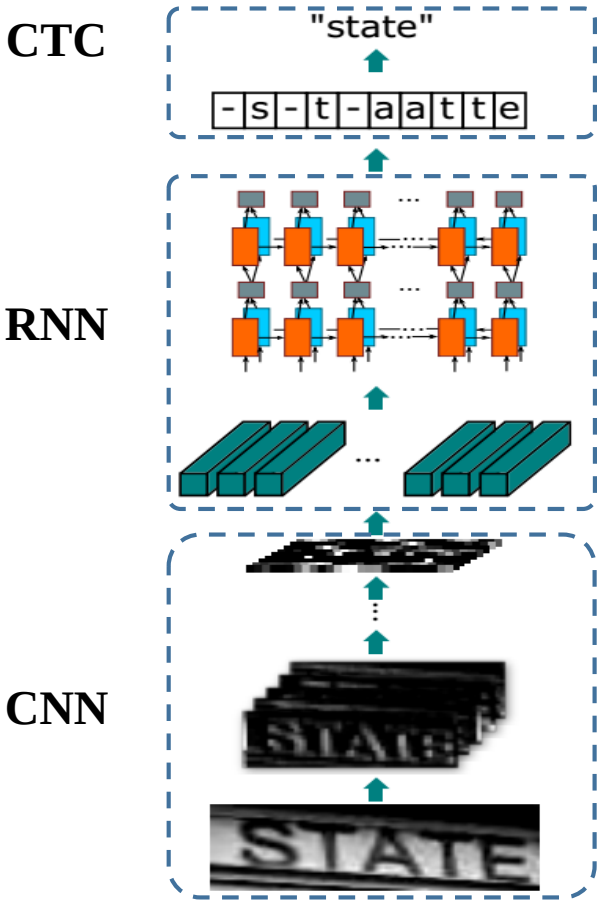
[3] Baoguang Shi, et al. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. TPAMI, 2018

[4] Cheng Z, Xu Y, Bai F, et al. Aon: Towards arbitrarily-oriented text recognition. CVPR, 2018

[5] Li H, Wang P, et al. Show, attend and read: A simple and strong baseline for irregular text recognition. AAAI. 2019

CRNN for Regular Text Recognition

The Network Architecture

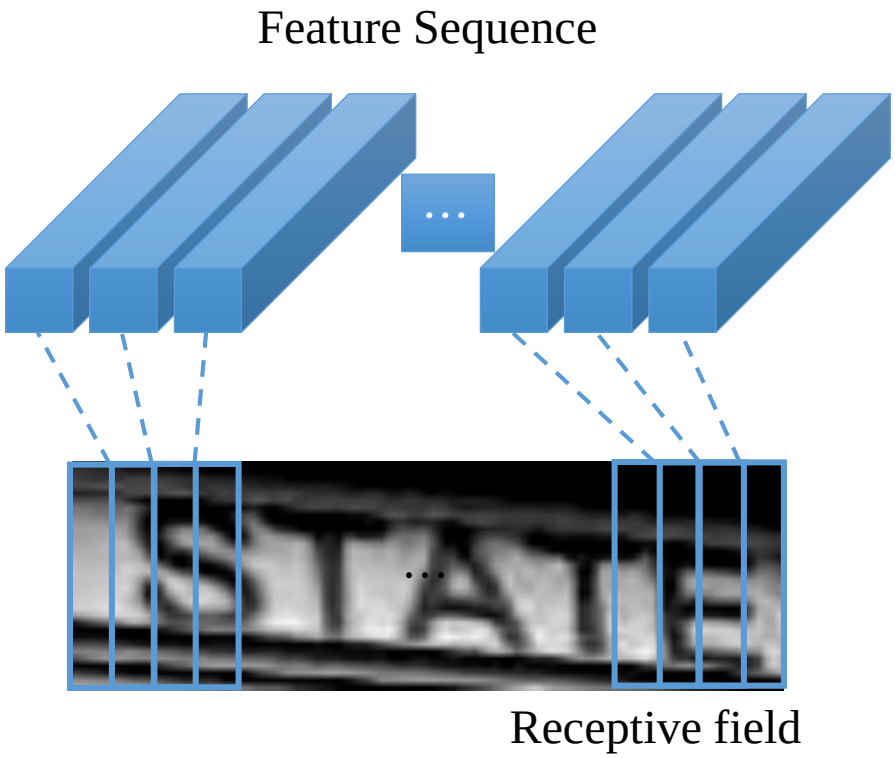


Network Structure

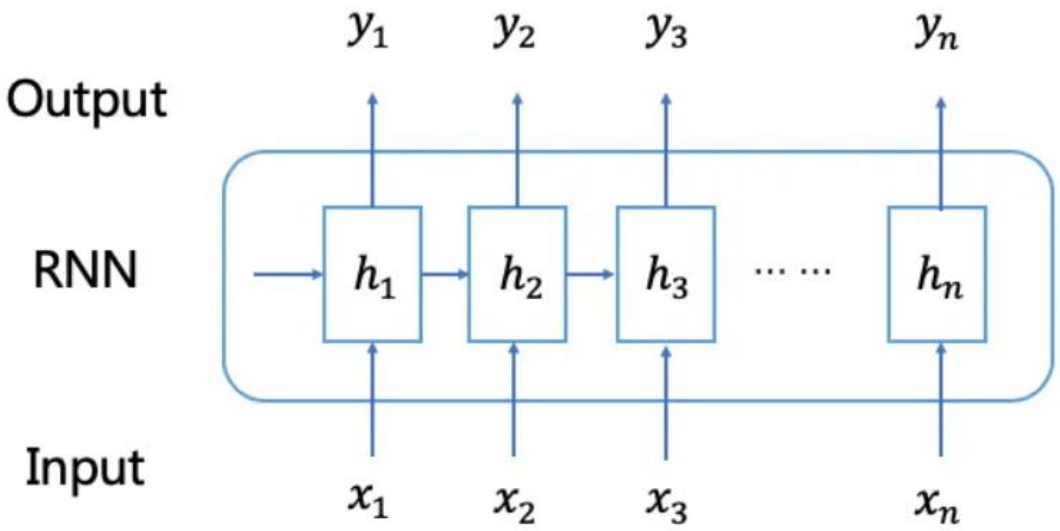
- Convolutional layers extract feature maps
- Convert feature maps into feature sequence
- Sequence labeling with LSTM
- Translate labels to text

CRNN for Regular Text Recognition

Sequence Modeling



CRNN for Regular Text Recognition

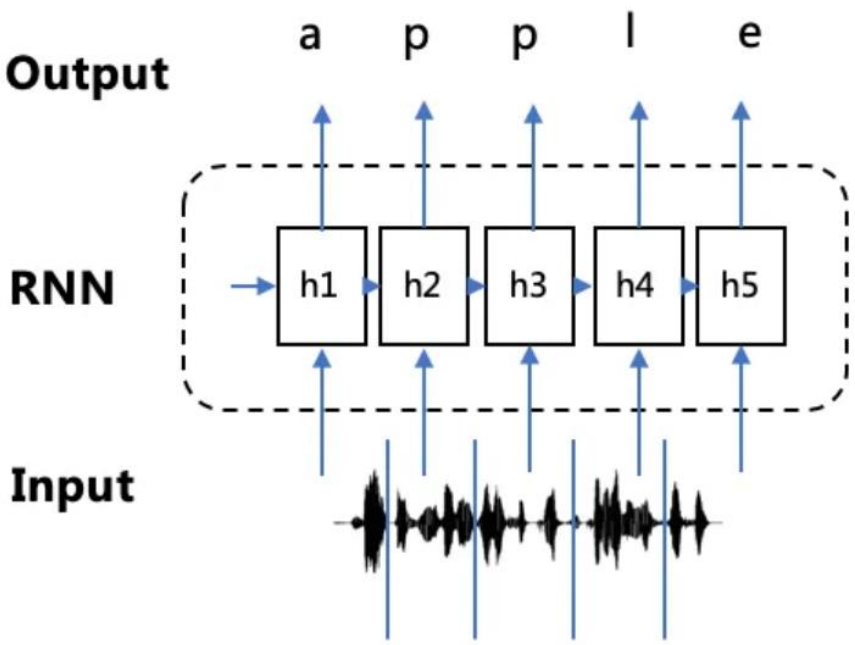


序列学习任务的依赖条件：输入序列和输出序列之间的映射关系已经事先标注

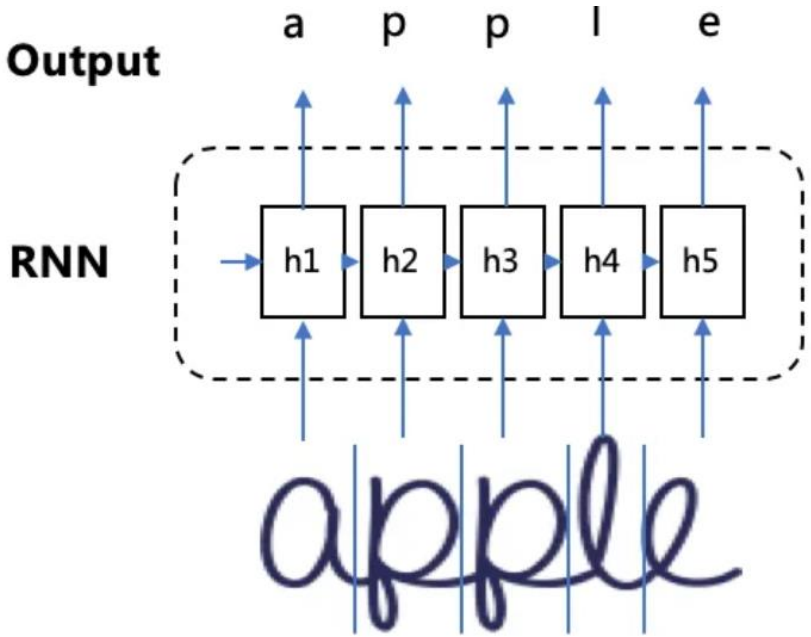
输入序列：The cat sat on the mat

输出序列： DT NN VBD IN DT NN

CRNN for Regular Text Recognition



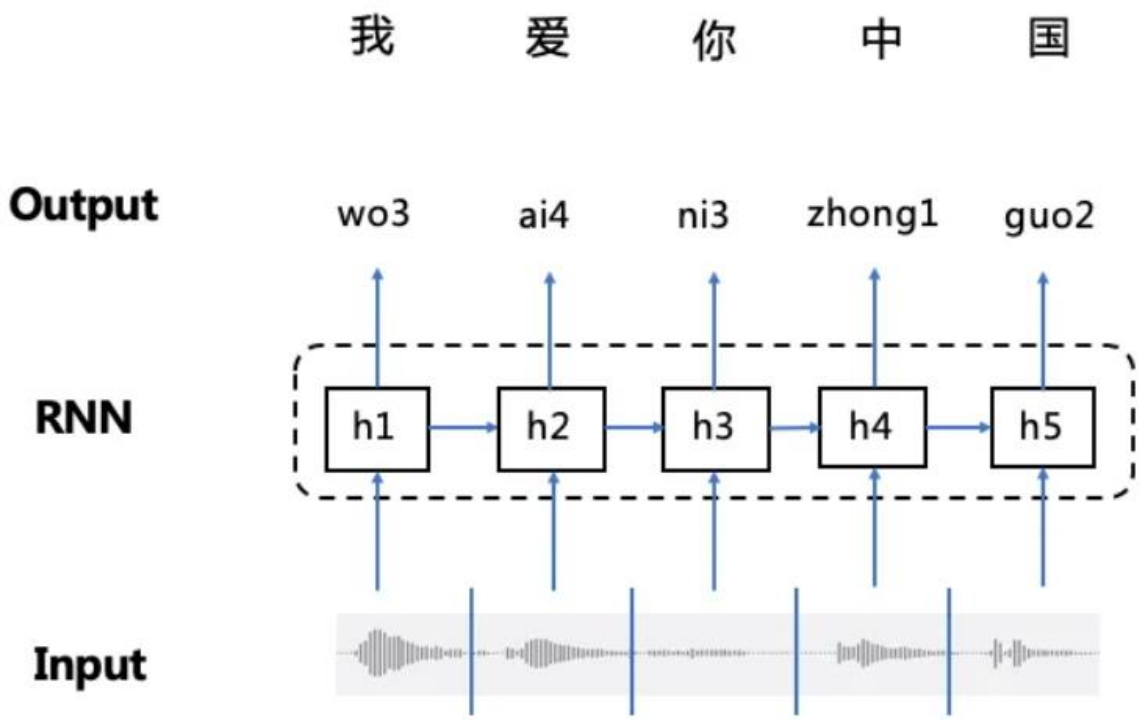
语音识别



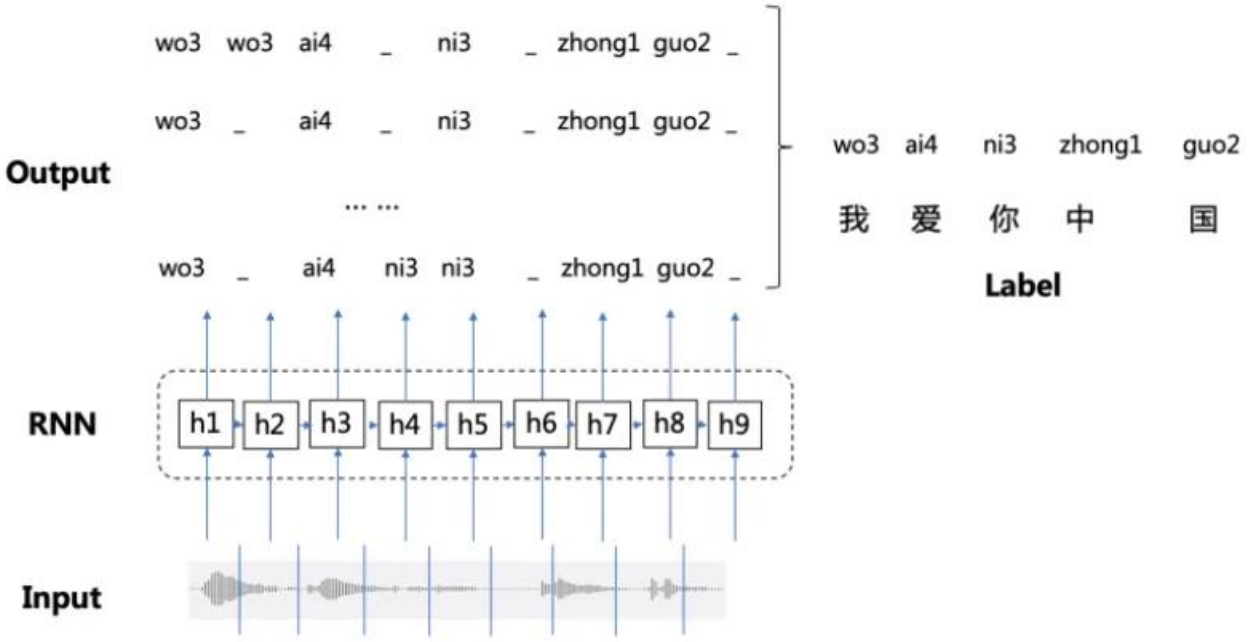
文本识别

音频数据和图像数据从现实世界中采集得到，这些数据就很难进行“分割”，很难获取到包含输入序列和输出序列映射关系的大规模训练样本。
因此，在这种条件下无法直接进行端到端的训练和预测

CRNN for Regular Text Recognition



CRNN for Regular Text Recognition



从输出序列中去除掉重复的元素以及间隔符，才可得到最终的音节序列，比如，“wo3 wo3 ai4 _ ni3 _ zhong1 guo2 _” 归一处理后得到“wo3 ai4 ni3 zhong1 guo2”。因此，输出序列和最终的label之间存在**多对一的映射关系**

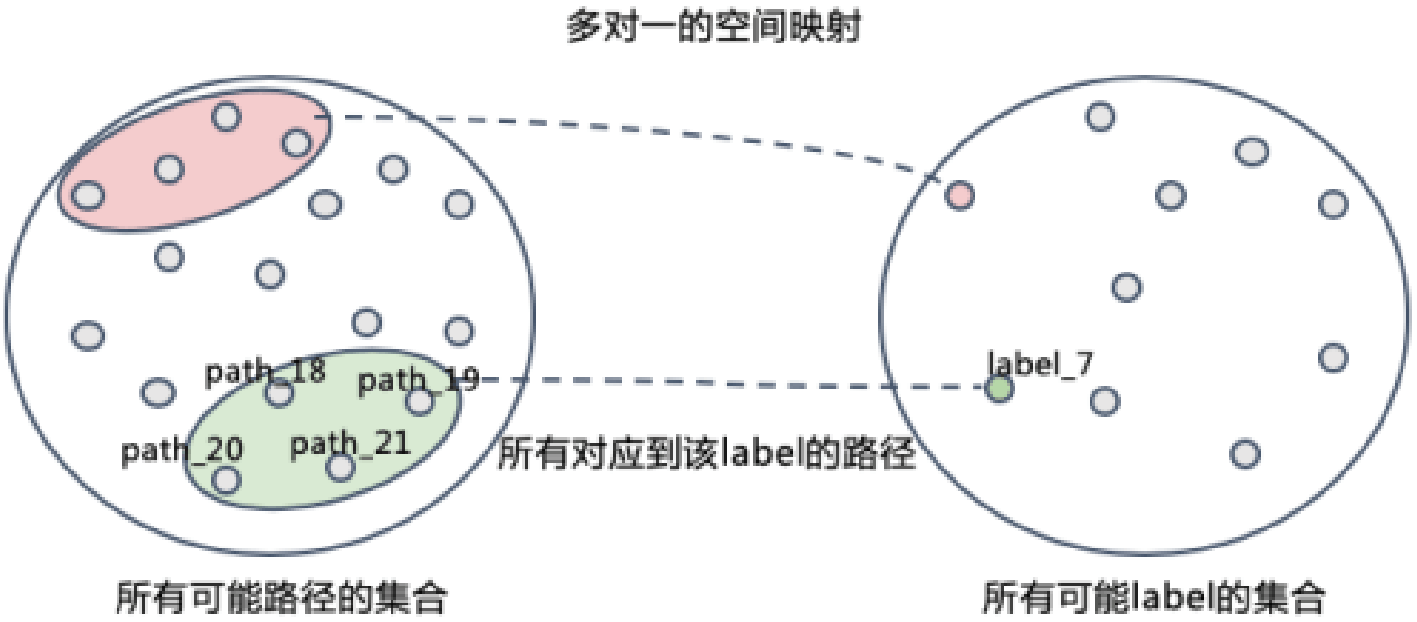
CRNN for Regular Text Recognition

Connectionist Temporal Classification (CTC)[1],直接对序列数据进行学习，而无需事先标注好训练数据中输入序列和输入序列的映射关系

RNN模型本质是对$\boldsymbol{p(z \mid x)}$建模，其中x表示输入序列，o表示输出序列，z表示最终的label，**o和z存在多对一的映射关系**，即：$\boldsymbol{p(z \mid x)}$=sum of all P(o|x)，其中o是所有映射到z的输出序列。因此，只需要穷举出所有的o，累加一起即可得到$\boldsymbol{p(z \mid x)}$，从而使得模型对最终的label进行建模

[1] Graves A, Fernández S, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. ICML, 2006

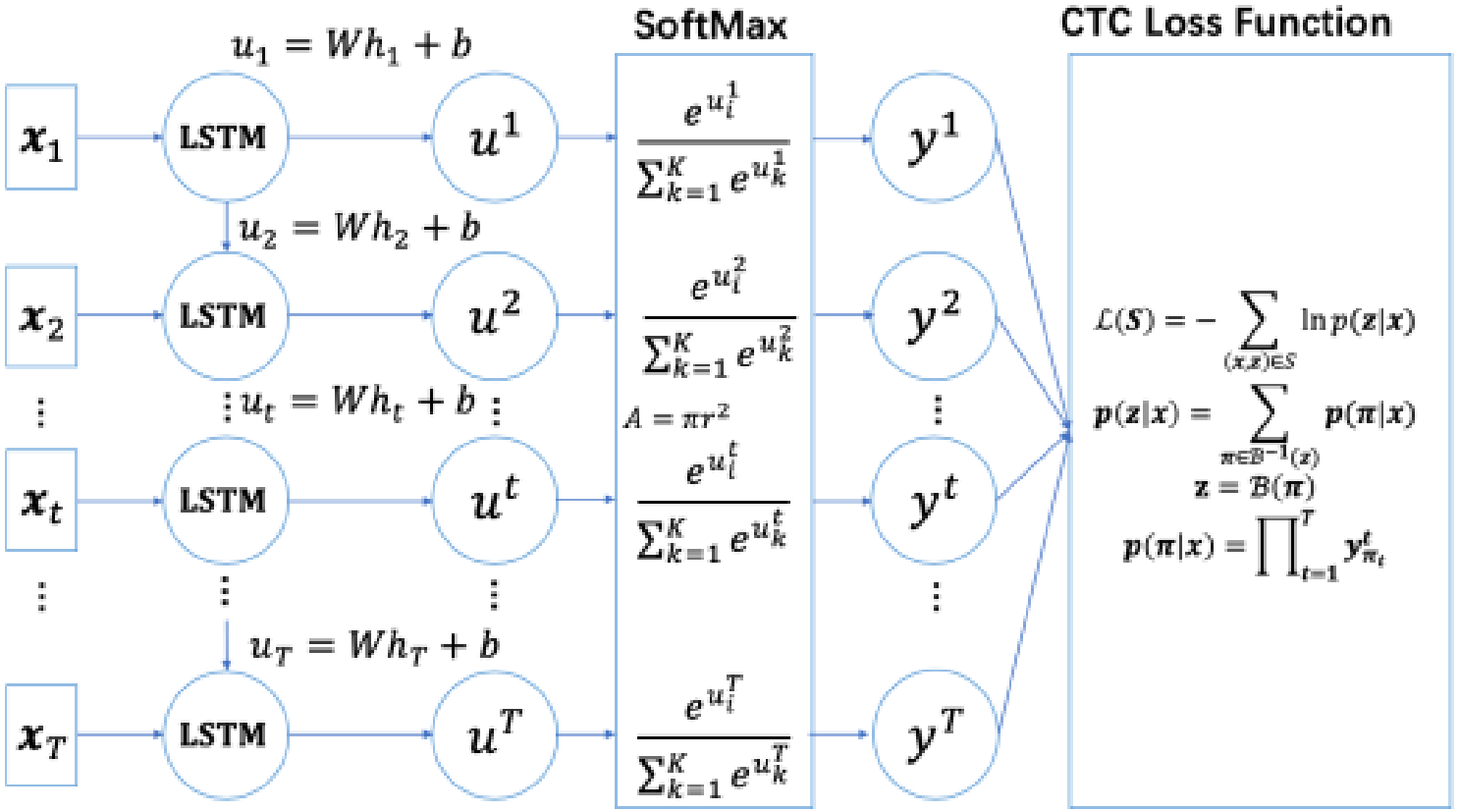
CRNN for Regular Text Recognition



某个label的概率等于对应其全部路径的概率总和：

$$p(label\_7) = p(path\_18) + p(path\_19) + p(path\_20) + p(path\_21)$$

CRNN for Regular Text Recognition



CRNN for Regular Text Recognition

Comparisons

Advantages

- End-to-end trainable
- Free of char-level annotations
- Unconstrained to specific lexicon
- 40~50 times less paramters than mainstream models
- Better or comparable performance with state-of-the-arts

Results(lexicon-free)

Method	IIIT5K	SVT	IC03	IC13
Bissacco et al. (ICCV13)	-	78.0	-	87.6
Jaderberg et al. (IJCV15)*	-	80.7	93.1	90.8
Jaderberg et al. (ICLR15)	-	71.7	89.6	81.8
Proposed	81.2	**82.7**	**91.9**	**89.6**

*is not lexicon-free, as its outputs are constrained to a 90k dictionary

RARE for Irregular Text Recognition

Motivation

Perspective and curved texts are hard to recognize!

(a) Perspective texts

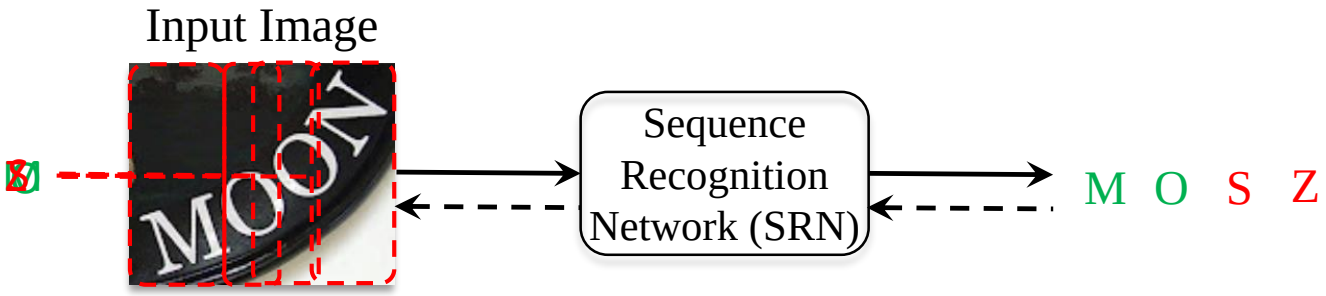
SVT-Perspective

(b) Curved texts

CUTE80

RARE for Irregular Text Recognition

Attention-based Sequence Recognition



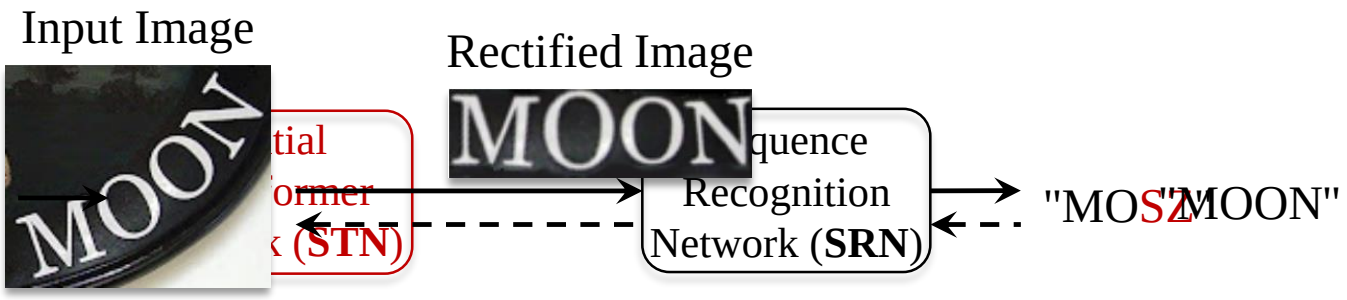
- **SRN**: an attention-based encoder-decoder framework
  - Encoder: ConvNet + Bi-LSTM
  - Decoder: Attention-based character generator

Results

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	93.6	91.8	68.2	62.5

RARE for Irregular Text Recognition

STN (Spatial Transform Network)[1] for Text Rectification

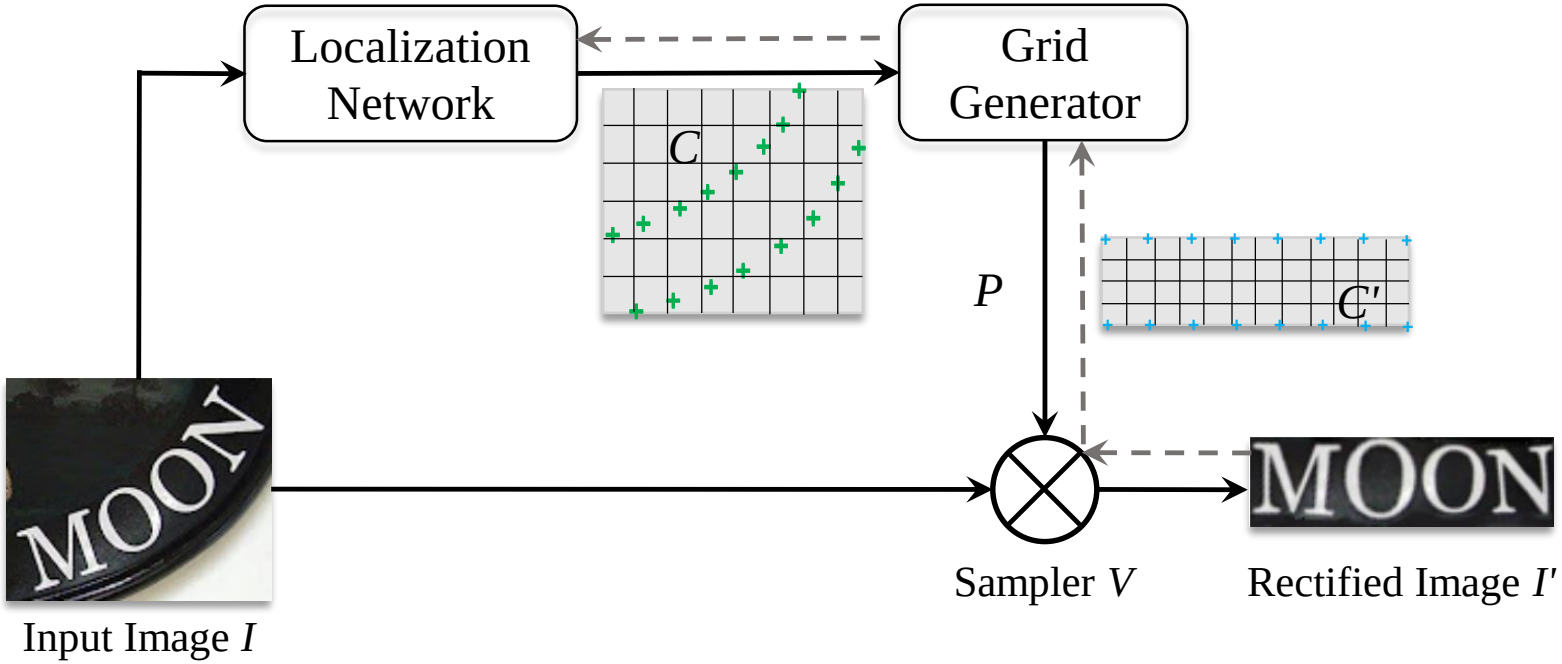


- An end-to-end trainable network
  - **STN**: rectifies images with spatial transformation
  - **SRN**: an attention-based encoder-decoder framework

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

RARE for Irregular Text Recognition

Spatial Transformer Network (STN)



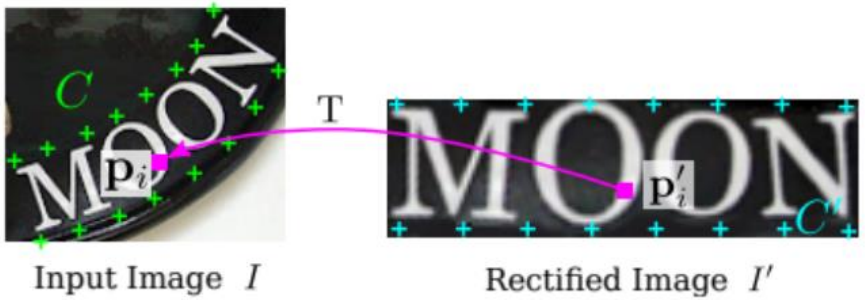
- **Localization Network**: A CNN that predicts the fiducial points.

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

RARE for Irregular Text Recognition

Localization Network

Type	Filters	Size	Output
Convolution	64	3*3	100*32
Convolution	128	3*3	100*32
Convolution	256	3*3	100*32
Convolution	512	3*3	100*32
Max-pooling	512	2*2	50*16
fc	-	-	1000
fc	-	-	1000
output	-	-	40

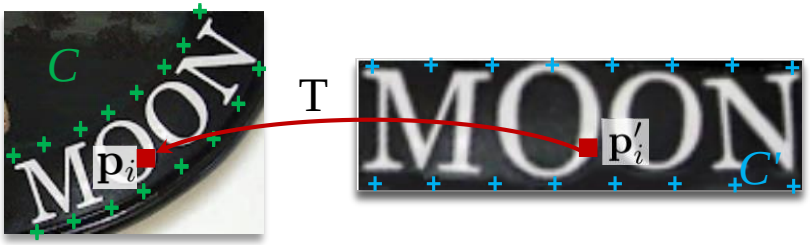


定义另一组 k 个基准点，
均匀分布在修正的照片的上下两侧

预测 k 个基准点的卷积网络需要 $2k$ 个输出

RARE for Irregular Text Recognition

Spatial Transformer Network (STN)



- Grid Generator: Computes a Thin-Plate-Spline (TPS) transform, **T**, from the fiducial points *C*.
- Sampler: TPS-Transform input image *I* into rectified *I'*.

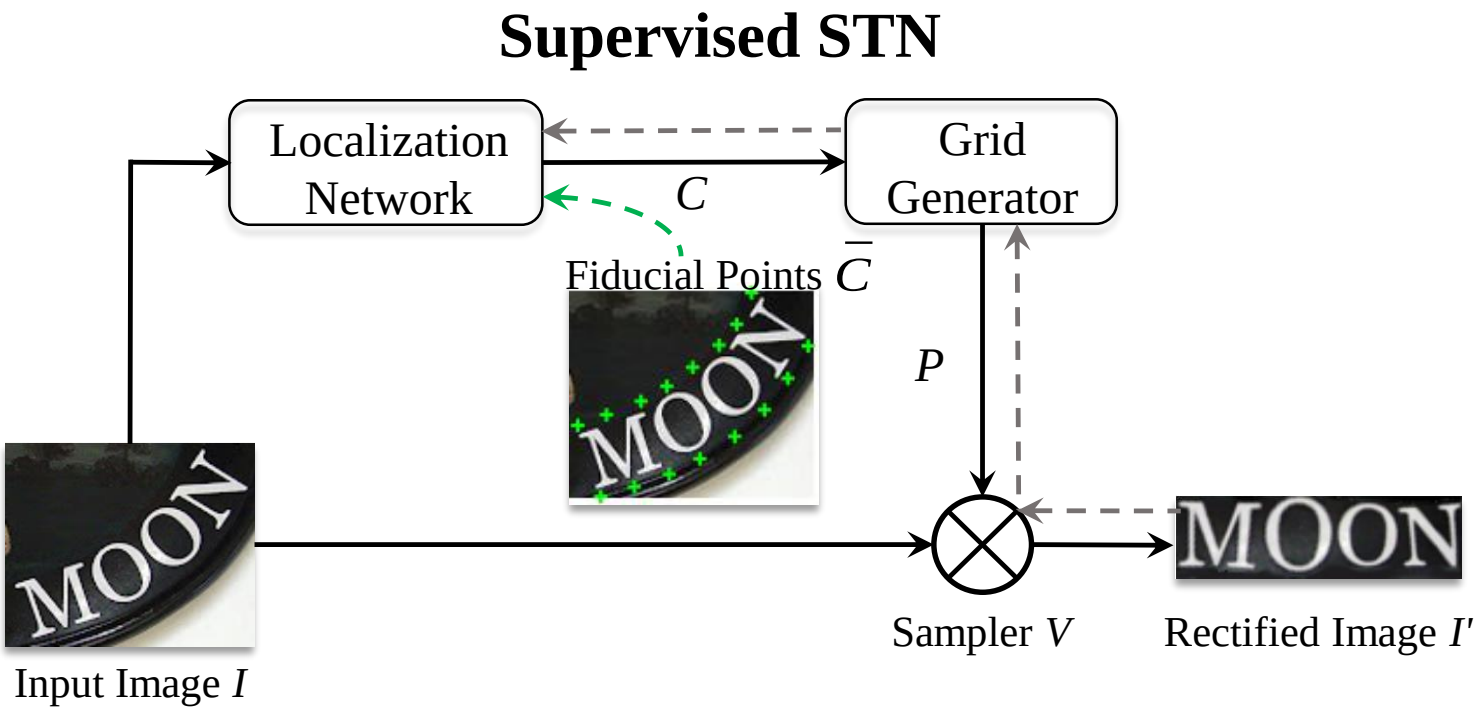
Results

Standard datasets | Deformable text datasets

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	**93.6**	91.8	68.2	62.5
STN+SRN	**88.2**	**86.7**	93.4	**92.7**	**76.8**	**76.7**

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

RARE for Irregular Text Recognition



- Synthetic dataset with fiducial points $\bar{C}$ to supervise the predicted C.

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	93.6	91.8	68.2	62.5
STN+SRN	88.2	86.7	93.4	92.7	76.8	76.7
STN(Supervised)+SRN	**88.8**	**87.9**	**94.1**	**94.0**	**77.7**	**78.8**

RARE for Irregular Text Recognition

Rectification Visualization

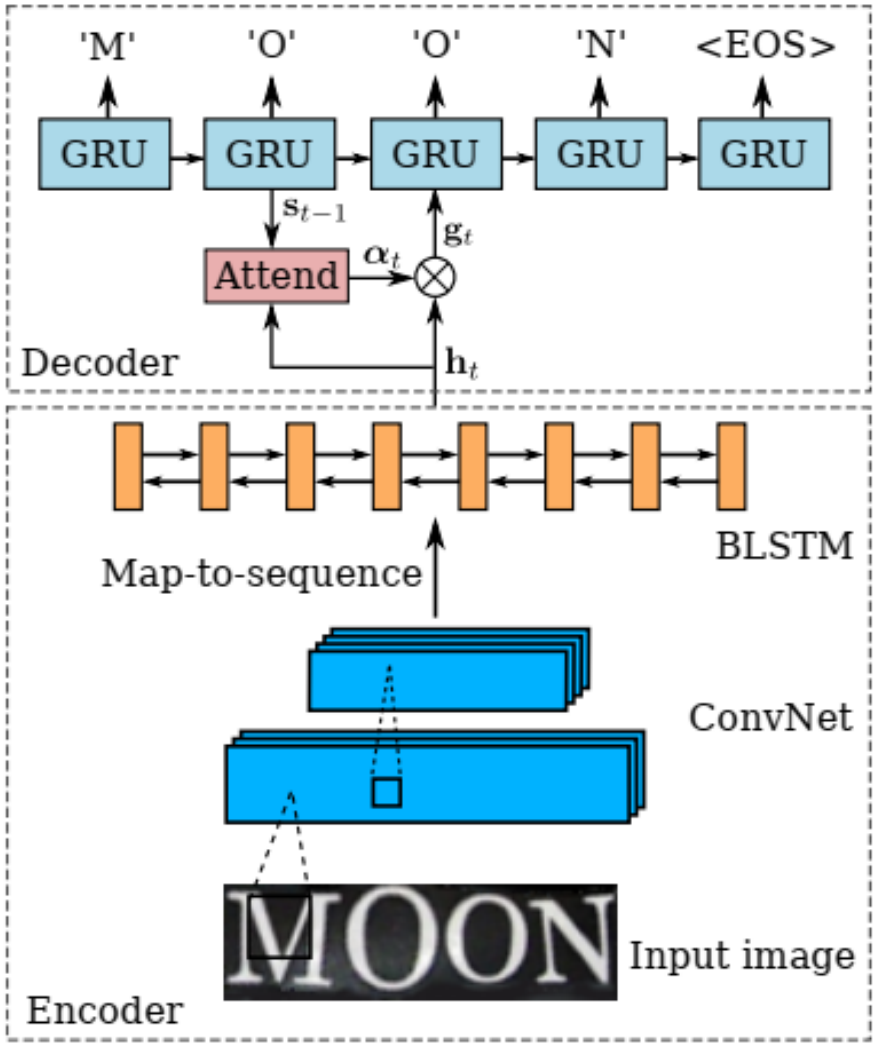
SVT-Perspective

Input	Rectified	Prediction Groundtruth
		restaurant restaurant
		quiznos quiznos
		sheraton sheraton
		mobil mobil
		jewelry jewelry
		public public

CUTE80

Input	Rectified	Prediction Groundtruth
		mercato marcato
		football football
		naval naval
		grove grove
		loka loka

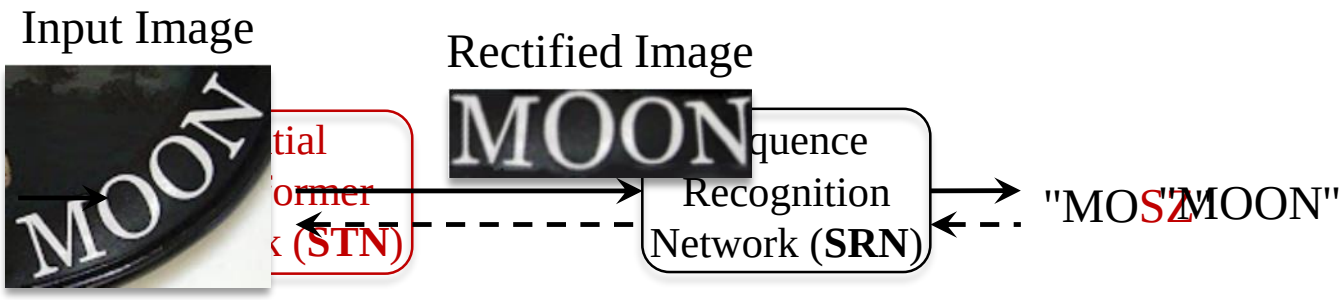
RARE for Irregular Text Recognition



The decoder recurrently generates a sequence of characters, conditioned on the sequence produced by the encoder.

ASTER for Irregular Text Recognition

STN (Spatial Transform Network)[1] for Text Rectification



- An end-to-end trainable network
  - **STN**: rectifies images with spatial transformation
  - **SRN**: an attention-based encoder-decoder framework

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

Symmetry-constrained Rectification Network for Scene Text Recognition. [Yang et al., ICCV2019]

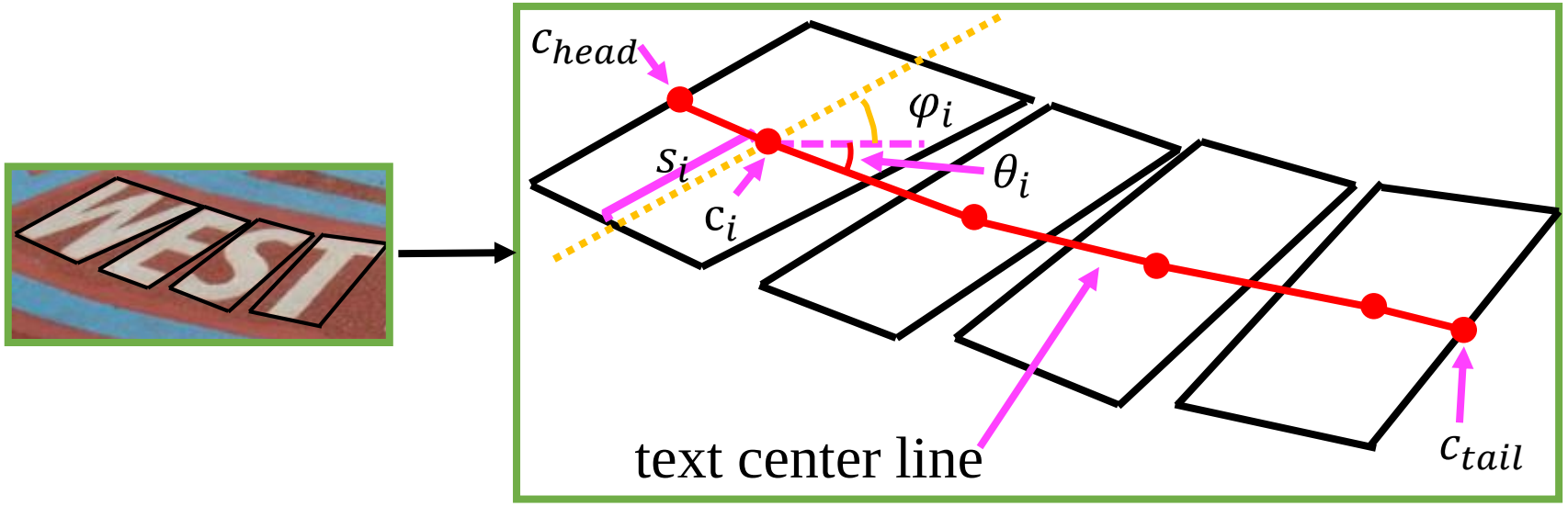
(a) Pipeline of ASTER

(b) Pipeline of ScRN

- Symmetry-constrained Rectification Network (**ScRN**) can yield more precise control points than **ASTER**, which adopts STN to predict control points in a weakly supervised way.

Symmetry-constrained Rectification Network for Scene Text Recognition. [Yang et al., ICCV2019]

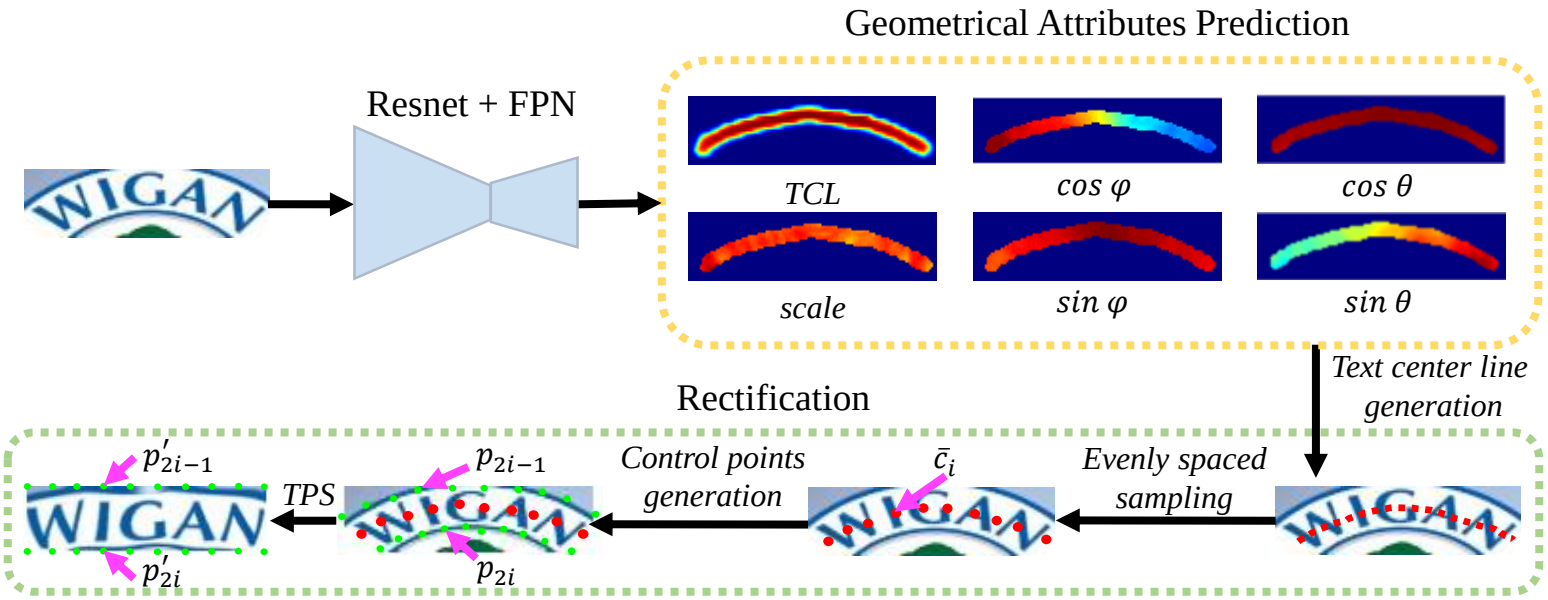
Geometrical Attributes



- Text center line $\boldsymbol{c}$: constructed via linking adjacent center points c_i.
- Scale $\boldsymbol{s}$: half the height of the character.
- Character orientation $\boldsymbol{\varphi}$: direction from the midpoint of the top edge to the midpoint of the bottom edge.
- Word orientation $\boldsymbol{\theta}$: tangential direction of $c_i \rightarrow c_{i+1}$.

Symmetry-constrained Rectification Network for Scene Text Recognition. [Yang et al., ICCV2019]

Rectification Process

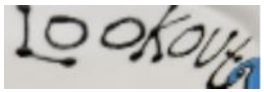


- Geometrical Attributes are used for rectification.
- Symmetry constraint is imposed.

Symmetry-constrained Rectification Network for Scene Text Recognition. [Yang et al., ICCV2019]

Rectification Visualization

Symmetry-constrained Rectification Network for Scene Text Recognition. [Yang et al., ICCV2019]

By ScRN	By ASTER	By ScRN By ASTER
		manchester for_____
		athletic athletic
		ballys ballf_
		salmon salmot
		bmw the
		bookstore soos____
		100kout _ook___

Regular datasets

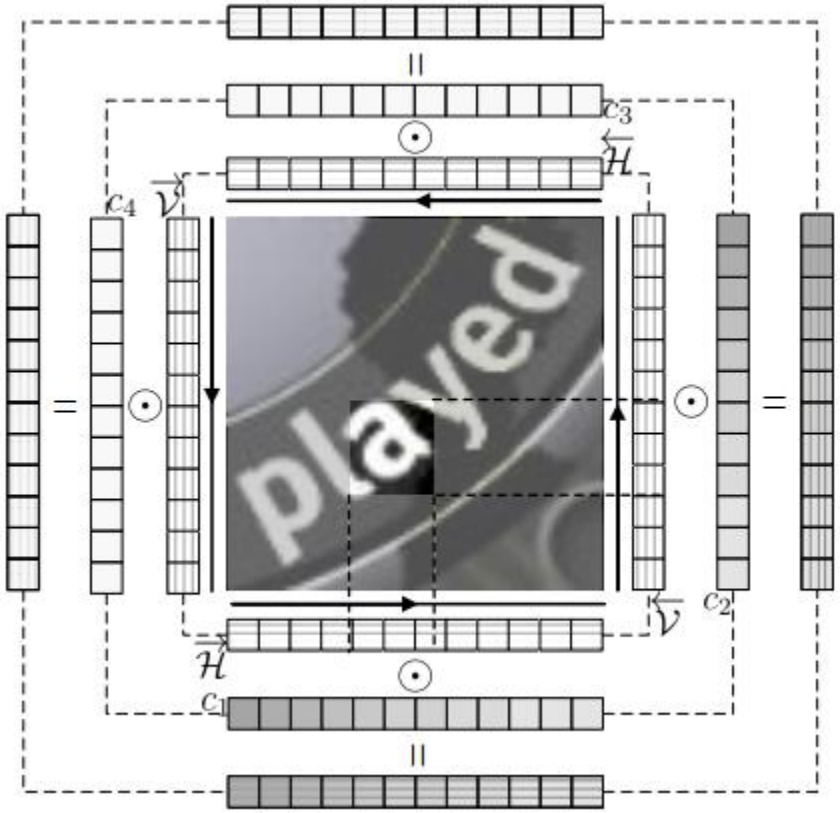
Method	IIIT5K	SVT	IC13
Liao et al.[1] (AAAI2019)	91.9	86.4	91.5
ASTER	93.4	**89.5**	94.5
ScRN	**94.4**	88.9	**95.0**

Irregular datasets

Method	IC15	SVTP	CUTE80
Liao et al.[1] (AAAI2019)	-	-	79.9
ASTER	76.1	78.5	79.5
ScRN	**78.7**	**80.8**	**87.5**

[1] Minghui Liao et al. Scene Text Recognition from Two-Dimensional Perspective. AAAI, 2019

AON for Irregular Text Recognition



$$\mathcal{H} = \begin{cases} \overrightarrow{\mathcal{H}} : (h_1, ..., h_L)^T, & left \rightarrow right \\ \overleftarrow{\mathcal{H}} : (h_L, ..., h_1)^T, & right \rightarrow left \end{cases}$$

$$\mathcal{V} = \begin{cases} \overrightarrow{\mathcal{V}} : (v_1, ..., v_L)^T, & top \rightarrow bottom \\ \overleftarrow{\mathcal{V}} : (v_L, ..., v_1)^T. & bottom \rightarrow top \end{cases}$$

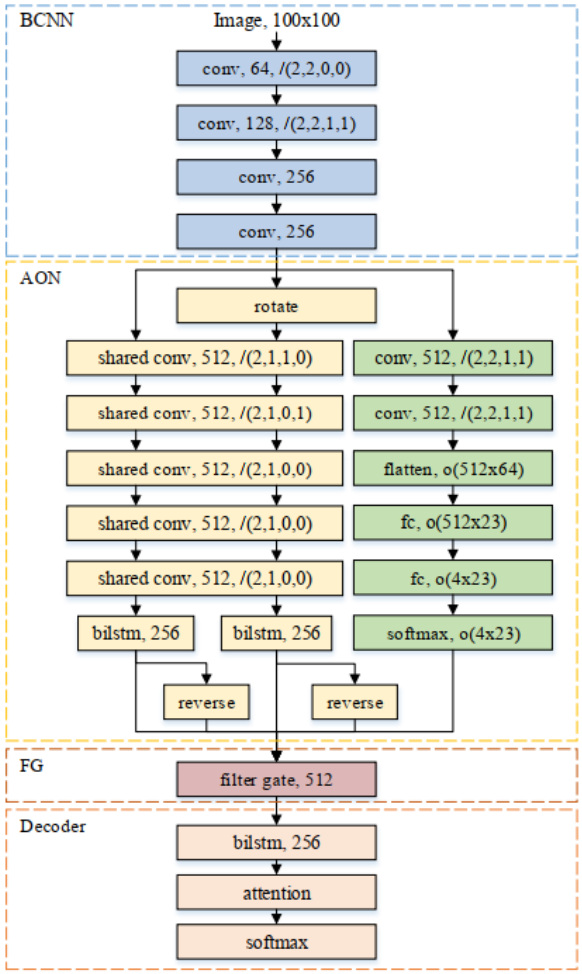
$$\mathcal{C} = (c_1, ..., c_L)^T.$$

$$\hat{h}'_i = [\overrightarrow{\mathcal{H}}_i \ \overleftarrow{\mathcal{H}}_i \ \overrightarrow{\mathcal{V}}_i \ \overleftarrow{\mathcal{V}}_i] c_i.$$

Then an activation operation is performed as follows:

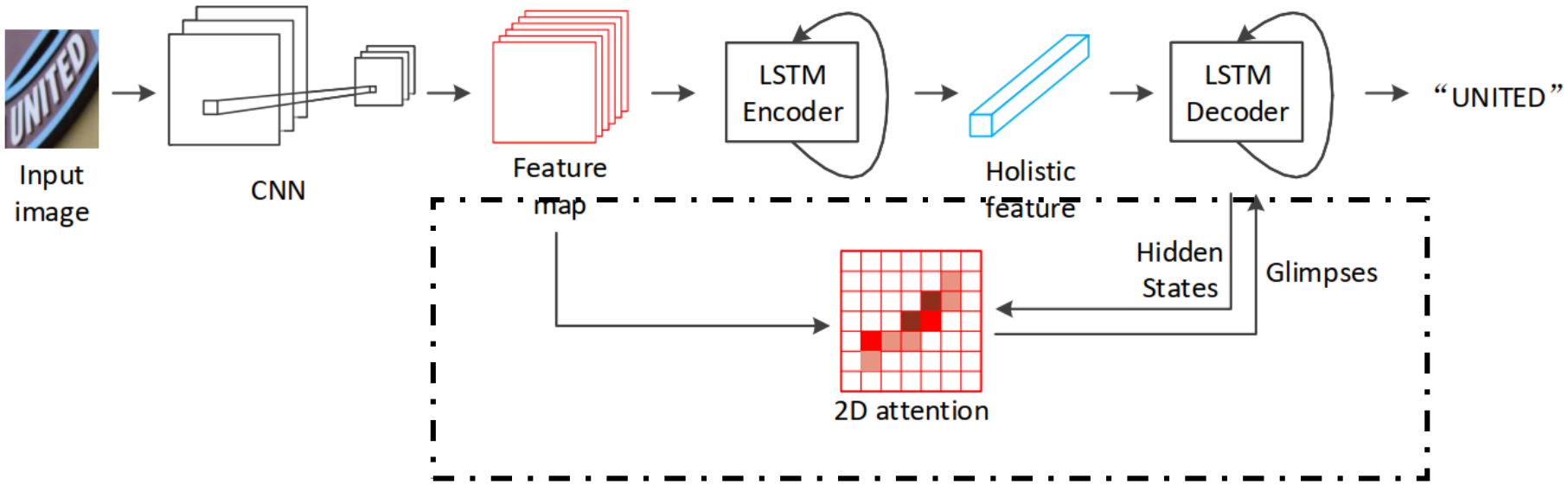
$$\hat{h}_i = \tanh(\hat{h}'_i).$$

AON for Irregular Text Recognition



- (BCNN) for low-level visual representation
- (AON) for capturing the horizontal, vertical and character placement features
- (FG) for combing four feature sequences with the character placement clues
- (Decoder) for predicting character sequence.

SAR for Irregular Text Recognition



- It is composed of a 31-layer ResNet, an LSTMbased encoder-decoder framework and a 2-dimensional attention module

Code is available at https://tinyurl.com/ShowAttendRead

SAR for Irregular Text Recognition

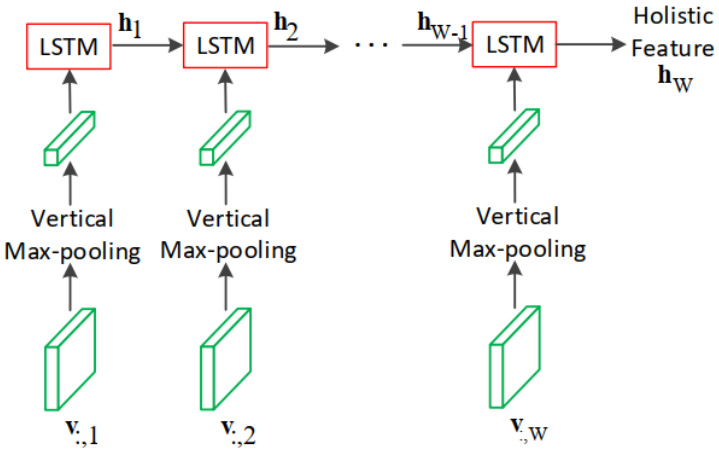


Figure 3: The structure of the LSTM encoder used in this work. $\mathbf{v}_{:,i}$ represents the ith column of the 2D feature map $\mathbf{V}$. At each time step, a column feature is firstly max pooled along the vertical direction, and then fed into LSTM.

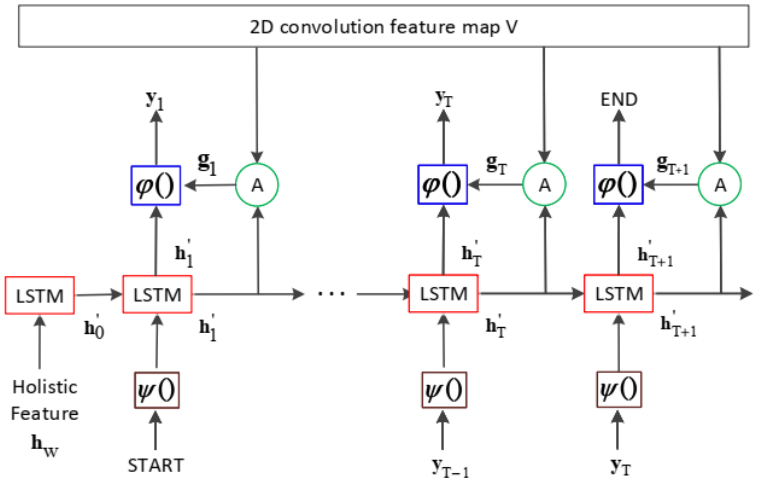
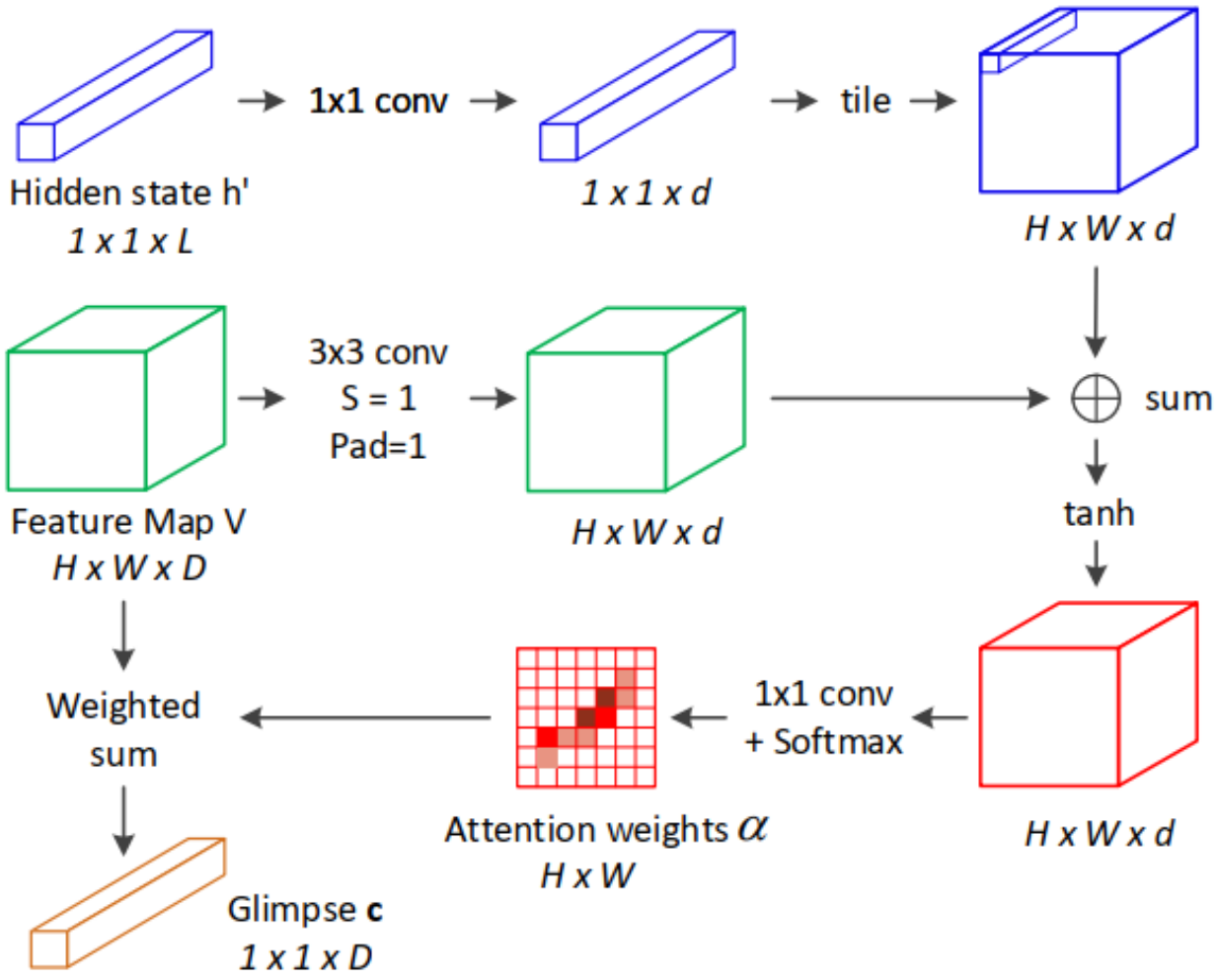
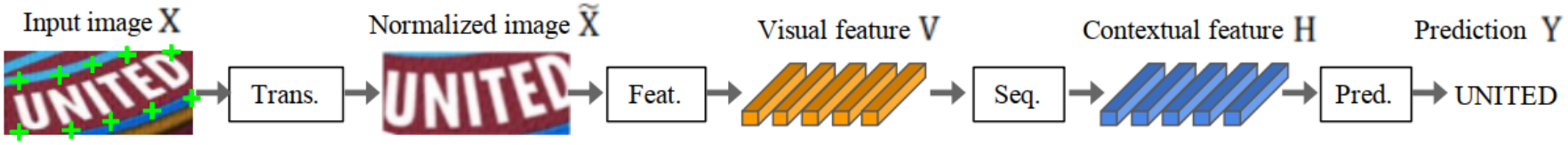


Figure 4: The structure of the LSTM decoder used in this work. The holistic feature $\mathbf{h}_W$, a "START" token and the previous outputs are input into LSTM subsequently, terminated by an "END" token. At each time step t, the output y_t is computed by $\varphi()$ with the current hidden state and the attention output as inputs.

SAR for Irregular Text Recognition



Baek J, Kim G, Lee J, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis. ICCV, 2019



- examine the inconsistencies of training and evaluation datasets, and the performance gap results from inconsistencies
- introduce a unified four-stage STR framework that most existing STR models fit into.
- analyze the module-wise contributions to performance in terms of accuracy, speed, and memory demand, under one consistent set of training and evaluation datasets.

Baek J, Kim G, Lee J, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis. ICCV, 2019

#	Trans.	Feat.	Seq.	Pred.	IIIT	SVT	IC03		IC13		IC15		SP	CT	Acc.	Time	params
					3000	647	860	867	857	1015	1811	2077	645	288	Total	ms	$\times 10^6$
1	None	VGG	None	CTC	76.2	73.8	86.7	86.0	84.8	81.9	56.6	52.4	56.6	49.9	69.5	1.3	5.6
2				Attn	80.1	78.4	91.0	90.5	88.5	86.3	63.0	58.3	66.0	56.1	74.6	19.0	6.6
3[1]			BiLSTM	CTC	82.9	81.6	93.1	92.6	91.1	89.2	69.4	64.2	70.0	65.5	78.4	4.4	8.3
4				Attn	84.3	83.8	93.7	93.1	91.9	90.0	70.8	65.4	71.9	66.8	79.7	21.2	9.2
5		RCNN	None	CTC	80.9	78.5	90.5	89.8	88.4	85.9	65.1	60.5	65.8	60.3	75.4	7.7	1.9
6[2]				Attn	83.4	82.4	92.2	92.0	90.2	88.1	68.9	63.6	72.1	64.9	78.5	24.1	2.9
7[3]			BiLSTM	CTC	84.2	83.7	93.5	93.0	90.9	88.8	71.4	65.8	73.6	68.1	79.8	10.7	4.6
8				Attn	85.7	84.8	93.9	93.4	91.6	89.6	72.7	67.1	75.0	69.2	81.0	27.4	5.5
9[4]		ResNet	None	CTC	84.3	84.7	93.4	92.9	90.9	89.0	71.2	66.0	73.8	69.2	80.0	4.7	44.3
10				Attn	86.1	85.7	94.0	93.6	91.9	90.1	73.5	68.0	74.5	72.2	81.5	22.2	45.3
11			BiLSTM	CTC	86.2	86.0	94.4	94.1	92.6	90.8	73.6	68.0	76.0	72.2	81.9	7.8	47.0
12				Attn	86.6	86.2	94.1	93.7	92.8	91.0	75.6	69.9	76.4	72.6	82.5	25.0	47.9
13	TPS	VGG	None	CTC	80.0	78.0	90.1	89.7	88.7	87.5	65.1	60.6	65.5	57.0	75.1	4.8	7.3
14				Attn	82.9	82.3	92.0	91.7	90.5	89.2	69.4	64.2	73.0	62.2	78.5	21.0	8.3
15			BiLSTM	CTC	84.6	83.8	93.3	92.9	91.2	89.4	72.4	66.8	74.0	66.8	80.2	7.6	10.0
16[5]				Attn	86.2	85.8	93.9	93.7	92.6	91.1	74.5	68.9	76.2	70.4	82.0	23.6	10.8
17		RCNN	None	CTC	82.8	81.7	92.0	91.6	89.5	88.4	69.8	64.6	71.3	61.2	78.3	10.9	3.6
18				Attn	85.1	84.0	93.1	93.1	91.5	90.2	72.4	66.8	75.6	64.9	80.6	26.4	4.6
19			BiLSTM	CTC	85.1	84.3	93.5	93.1	91.4	89.6	73.4	67.7	74.4	69.1	80.8	14.1	6.3
20				Attn	86.3	85.7	94.0	94.0	92.8	91.1	75.0	69.2	77.7	70.1	82.3	30.1	7.2
21		ResNet	None	CTC	85.0	85.7	94.0	93.6	92.5	90.8	74.6	68.8	75.2	71.0	81.5	8.3	46.0
22				Attn	87.1	87.1	94.3	93.9	93.2	91.8	76.5	70.6	78.9	73.2	83.3	25.6	47.0
23[6]			BiLSTM	CTC	87.0	86.9	94.4	94.0	92.8	91.5	76.1	70.3	77.5	71.7	82.9	10.9	48.7
24[7]				Attn	**87.9**	**87.5**	**94.9**	**94.4**	**93.6**	**92.3**	**77.6**	**71.8**	**79.2**	**74.0**	**84.0**	27.6	49.6

[1] CRNN. [2] R2AM. [3] GRCNN. [4] Rosetta. [5] RARE. [6] STAR-Net. [7] our best model.

Outline

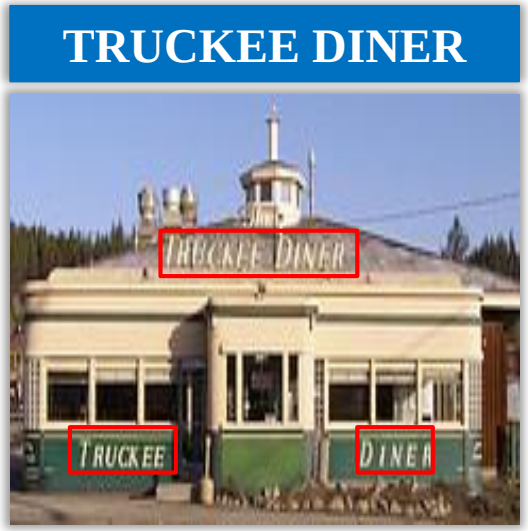
- Background
- Scene Text Detection
- Scene Text Recognition
- **Applications**
- Current Issues
- Future Trends

Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification

Fine-Grained Image Classification with Textual Cue

Motivations



(a)

(b)

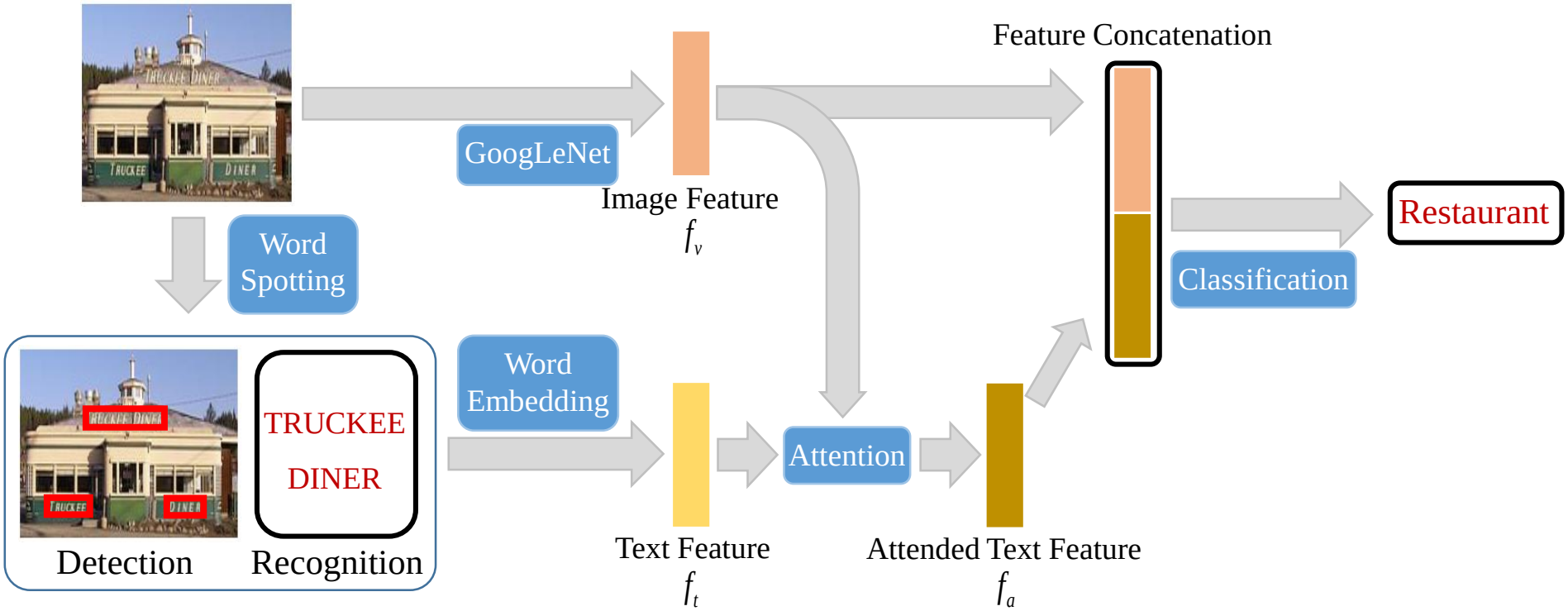
(c)

- Visual cues would group (a)-(b) whereas scene would group (b)-(c).
- Texts in images can improve the performance of fine-grained image classification.

[1] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv:1704.04613, 2017.

Fine-Grained Image Classification with Textual Cue

Pipeline



[1] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv:1704.04613, 2017.

Fine-Grained Image Classification with Textual Cue

Attention Model to Select Relevant Words

Repair shop

Hotel

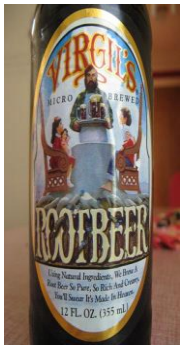
- Some irrelevant words to this Category

Fine-Grained Image Classification with Textual Cue

Con-Text dataset[1]

- 28 categories of **Scenes**
- 24,255 images in total

Drink Bottledataset[2]



- Selected from ImageNet
- 20 categories of **Drink Bottles**
- 18,488 images in total

[1] S. Karaoglu. et al. Con-text: text detection using background connectivity for fine-grained object classification. ACM2013
[2] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv2017.

Fine-Grained Image Classification with Textual Cue

Results: mAP(%) improvement on different datasets

Method	Dataset	
	Con-Text	Drink Bottle
GoogLeNet[1]	61.3	63.1
GoogLeNet + Textual Cue	**79.6 (+18.3)**	**72.8 (+9.7)**

[1] C. Szegedy, et al. Going deeper with convolutions. CVPR2015

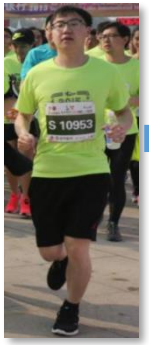
Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification

Number-based Person Re-Identification

- **Problem: hard to track and retrieve an athlete in a marathon game**

query



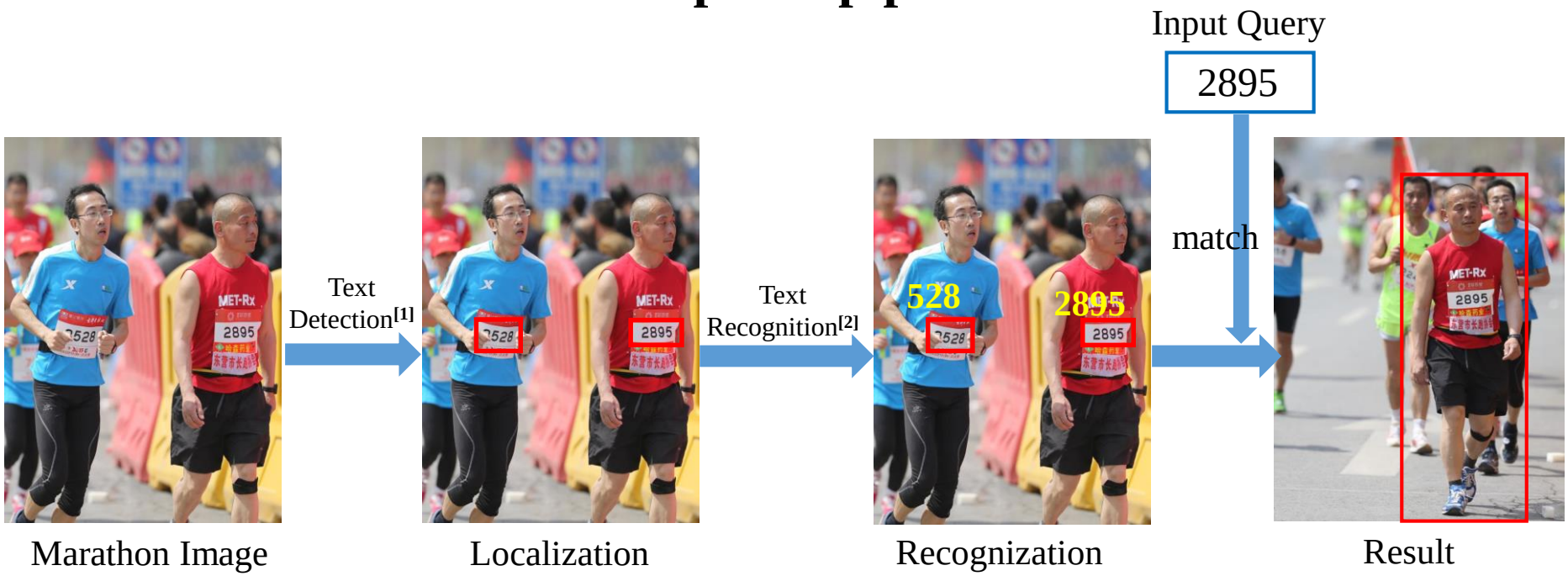
retrieval

Marathon image database

- **Motivation: every athlete has a unique racing bib number**

Number-based Person Re-Identification

Proposed pipeline



[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.
[2] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

Number-based Person Re-Identification

Marathon Dataset

8706 training images, 1000 testing images

Experimental Results

Identification accuracy rate(*Id_acc*): **85%**

$$Id\_acc = \frac{Num(correctly\ recognized\ persons)}{Num(total\ persons)}$$

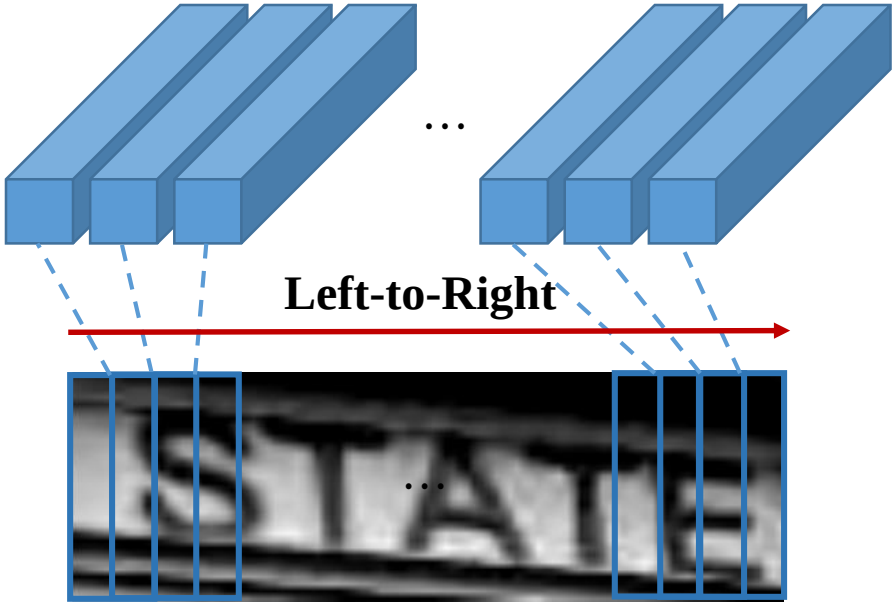
Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification

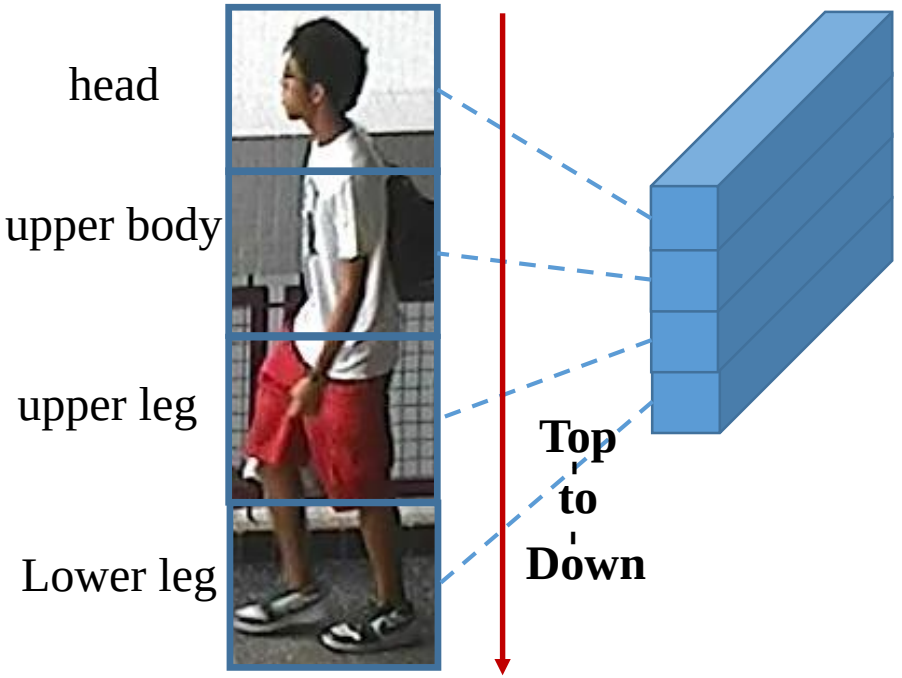
From Text Recognition to Person Re-Identification

Sequence Modeling

Text Recognition (**CRNN**)



Person Re-Idntification

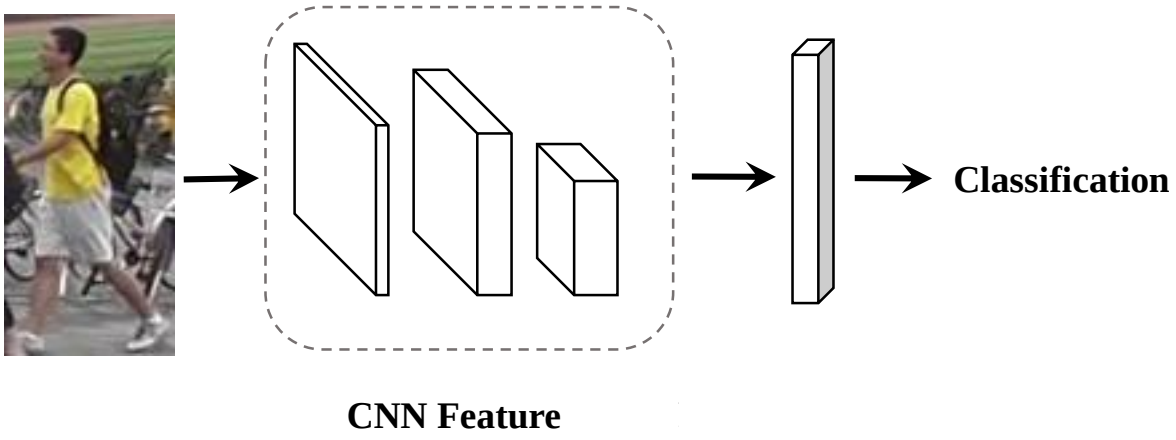


[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

From Text Recognition to Person Re-Identification

Model Architecture

CNN + LSTM



Results on Market1501[1]

Method	mAP(%)	R1(%)
CNN	59.8	81.4
CNN + LSTM	**65.5**	**85.8**

R1: given a query, precision of the top-1 similar image from gallery discriminated by model.

[1] Zheng et al. Scalable Person Re-identification: A Benchmark. ICCV 2015

From Text Recognition to Person Re-Identification

Retrival Results

CNN

… …

query

CNN+LSTM

… …

query

Outline

- Background
- Scene Text Detection
- Scene Text Recognition
- Applications
- **Current Issues**
- Future Trends

Unfair experimental settings

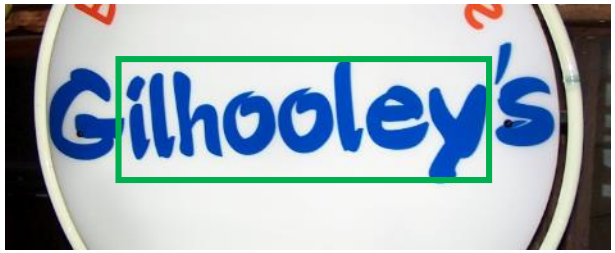
In some recent works, different experimental settings are used. It is hard to judge which one is the key factor for the final improvement.

Text Detection task, on ICDAR15

	Method A	Method B	Method C	TextField
F-score	87.1	87.5	86.8	82.4
Training Set	ICDAR17-MLT; ICDAR15;	SynthText; COCO-Text; ICDAR17-MLT ; 30K(Private data); ICDAR15;	SynthText; ICDAR15	SynthText; ICDAR15
Backbone	Resnet152	Resnet50	Resnet50	VGG16
Resolution of test image	2240*1260	1600*900	2240*1260	1280*720

Method A, B, C are published at top conference or top journal in 2019.

Imperfect evaluation protocol



Detection ✓

Recognition ✗

Detection ✓

Recognition ✗

Detection ✓

Recognition ✗

Detection ✓

Recognition ✗

IOU>0.5 is right for detection but making it difficult for recognition in some cases.

Limitations of current datasets

ICDAR13

- 462 images
- English
- Focused horizontal scene text
- Indoor(mall) and outdoor(street)

ICDAR15

- 1500 images
- English
- Incidental oriented text
- Indoor(mall) and outdoor(urban environments)

Limitations of current datasets

ICDAR17-MLT

- 2700 images per language
- 7 languages
- Oriented text
- Indoor(mall) and outdoor(urban environments)

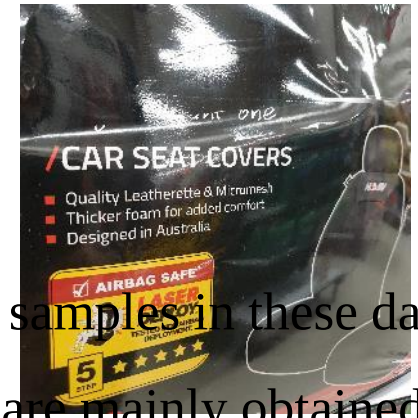
TotalText

- 1555 images
- English
- Horizontal, oriented and Curved text
- Indoor(shop) and outdoor(urban environments)

Limitations of current datasets

Larger benchmark with diverse and complex scenes is needed.

Book cover



Commodity character

Indicator

- The number of samples in these datasets is relatively small.
- These datasets are mainly obtained from a single scene.

Vehicle text

Merchant signs

Clothing characters

Outline

- Background
- Scene Text Detection
- Scene Text Recognition
- Applications
- Current Issues
- **Future Trends**

Future Trends

- Multilingual End-to-end text recognition
- Semi-supervised or weakly supervised text detection and recognition
- Text image synthesis (GAN)
- Unified framework for OCR and NLP
- Integrating Scene text and Image/Videos for many applications.
- A very large benchmark dataset like ImageNet including plentiful scenarios should be considered

Thank you !