A Simple and Strong Baseline for Irregular Text Recognition

源码地址：lua语言

Paper：2019, 机构

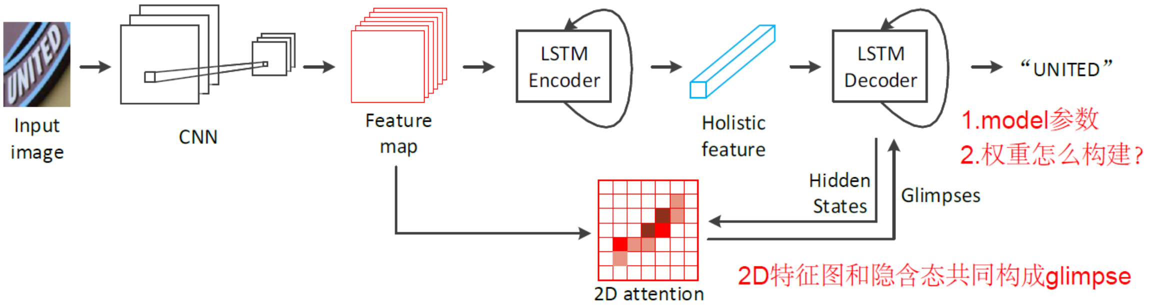
摘要

本文提出一种简单有效的baseline模型，通过字符级的监督标注，识别自然场景中不规则的文本。模型由31层的ResNet，LSTM编解码框架和一个2D-Attention模块构成。在不规则和规则的benchmarks上都达到STA水平。

追溯相关工作

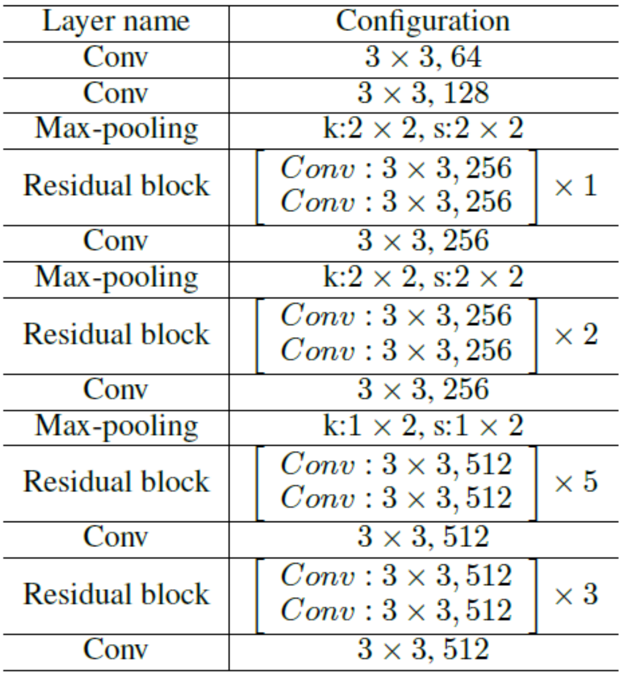
pass

模型细节设计



CNN层提取特征：

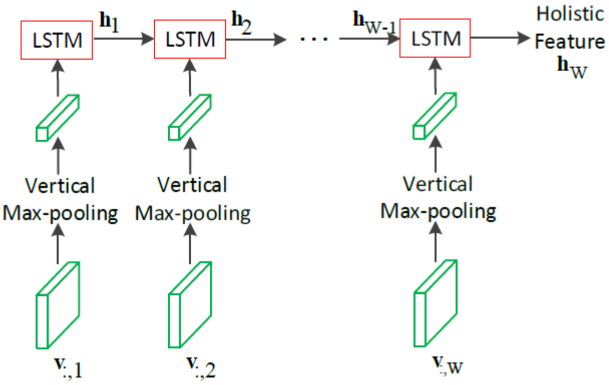
- 采用31层ResNet，所有kerner大小是3x3。做残差块时，输入输出维度不同，shutcut中加1x1卷积；
- 采用2x2和2x1池化。在水平方向保留更多信息，有利于对狭窄字符的识别；
- 输入整张图片，保持比例压缩到固定高度，宽度可变；
- 输出2D特征图，用于1）提取图片整体特征；2）作为Attention模块的上下文环境；
- **CNN细节结构**：卷积层stride和padding都是1，池化层没有padding；



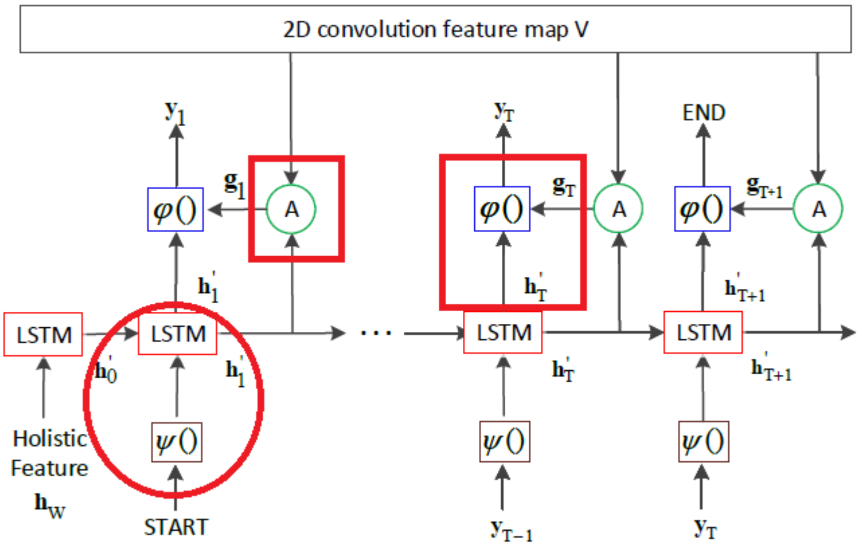
Layer name	Configuration
Conv	$3 \times 3, 64$
Conv	$3 \times 3, 128$
Max-pooling	k:2×2, s:2×2
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 256 \\ Conv : 3 \times 3, 256 \end{bmatrix} \times 1$
Conv	$3 \times 3, 256$
Max-pooling	k:2×2, s:2×2
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 256 \\ Conv : 3 \times 3, 256 \end{bmatrix} \times 2$
Conv	$3 \times 3, 256$
Max-pooling	k:1×2, s:1×2
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 512 \\ Conv : 3 \times 3, 512 \end{bmatrix} \times 5$
Conv	$3 \times 3, 512$
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 512 \\ Conv : 3 \times 3, 512 \end{bmatrix} \times 3$
Conv	$3 \times 3, 512$

LSTM编解码：

- 编码器：两层LSTM含512个隐含单元，每个时间步接受一列（最大池化后）特征数据，遍历W步后，第2层LSTM输出的隐含态hw（长度固定）作为整体的特征表示，用于初始化解码器



- 解码器：另一个两层LSTM（512单元），接受hw作为初始隐含态，每个时间步，接受上一步的输出（预测值和隐含态）作为输入，输入以one-hot向量+线性层的方式构建。训练时用gt序列代替预测序列作为输入。每一步的输出，是将当前LSTM的隐含态hi和Attention模块的输出gi拼接起来，再加一个线性层（把特征表示成类别空间，中文OCR类别空间通常很大）做softmax。



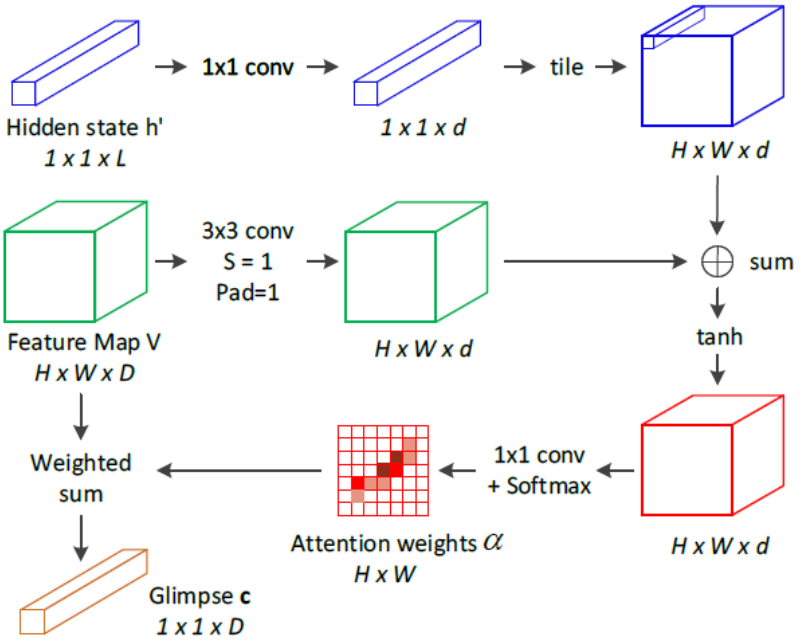
2D-Attention模块：

代替文本校正，自动适应形状，方向和排布不规则的文本。

- 考虑像素位置的相关性，构建Attention权重时加入8-领域的位置信息；

$$\begin{cases} \mathbf{e}_{ij} = \tanh(\mathbf{W}_v \mathbf{v}_{ij} + \boxed{\sum_{p,q\in\mathcal{N}_{ij}} \tilde{\mathbf{W}}_{p-i,q-j}\cdot \mathbf{v}_{pq}} + \mathbf{W}_h \mathbf{h}'_t), \\ \alpha_{ij} = \mathrm{softmax}(\mathbf{w}_e^T \cdot \mathbf{e}_{ij}), \\ \mathbf{g}_t = \sum_{i,j}\alpha_{ij}\mathbf{v}_{ij}, \quad i=1,\dots,H, \quad j=1,\dots,W. \end{cases}$$

- 通过3x3的卷积实现领域操作，结合膨胀后的隐向量（[H，W，d]维）得到eij。再用1x1卷积+softmax得到2D的权重图；



其他细节

实验数据集：Syn90k，SVTP，CUTE80， COCO-Text

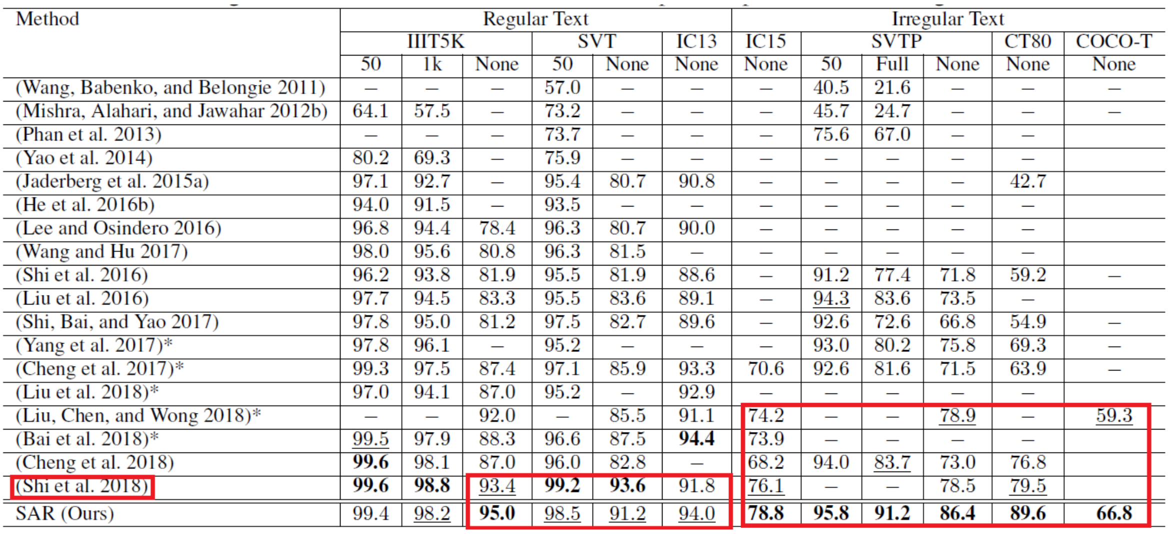
训练参数：

- 交叉熵Loss，ADAM优化器
- batch=32，lr_init=0.001, 每1w步衰减0.9， 直到0.00001

实验结果

在公共benchmarks下的精度

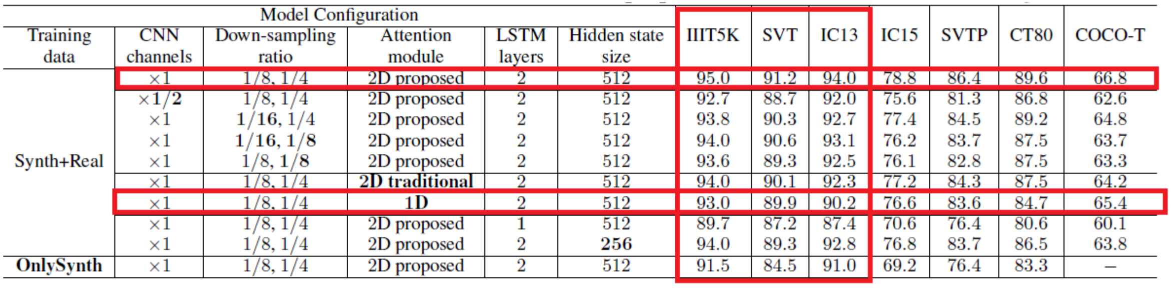
- 规则文本+有字典；shi et al.2018-Aster表现最好（本文也不差）
- 不规则文本；本文表现很好



Method	Regular Text						Irregular Text					
	IIIT5K			SVT		IC13	IC15	SVTP			CT80	COCO-T
	50	1k	None	50	None	None	None	50	Full	None	None	None
(Wang, Babenko, and Belongie 2011)	–	–	–	57.0	–	–	–	40.5	21.6	–	–	–
(Mishra, Alahari, and Jawahar 2012b)	64.1	57.5	–	73.2	–	–	–	45.7	24.7	–	–	–
(Phan et al. 2013)	–	–	–	73.7	–	–	–	75.6	67.0	–	–	
(Yao et al. 2014)	80.2	69.3	–	75.9	–	–	–	–	–	–	–	
(Jaderberg et al. 2015a)	97.1	92.7	–	95.4	80.7	90.8	–	–	–	–	42.7	
(He et al. 2016b)	94.0	91.5	–	93.5	–	–	–	–	–	–	–	
(Lee and Osindero 2016)	96.8	94.4	78.4	96.3	80.7	90.0	–	–	–	–	–	
(Wang and Hu 2017)	98.0	95.6	80.8	96.3	81.5	–	–	–	–	–	–	
(Shi et al. 2016)	96.2	93.8	81.9	95.5	81.9	88.6	–	91.2	77.4	71.8	59.2	–
(Liu et al. 2016)	97.7	94.5	83.3	95.5	83.6	89.1	–	94.3	83.6	73.5	–	
(Shi, Bai, and Yao 2017)	97.8	95.0	81.2	97.5	82.7	89.6	–	92.6	72.6	66.8	54.9	–
(Yang et al. 2017)*	97.8	96.1	–	95.2	–	–	–	93.0	80.2	75.8	69.3	–
(Cheng et al. 2017)*	99.3	97.5	87.4	97.1	85.9	93.3	70.6	92.6	81.6	71.5	63.9	–
(Liu et al. 2018)*	97.0	94.1	87.0	95.2	–	92.9	–	–	–	–	–	
(Liu, Chen, and Wong 2018)*	–	–	92.0	–	85.5	91.1	74.2	–	–	78.9	–	59.3
(Bai et al. 2018)*	99.5	97.9	88.3	96.6	87.5	**94.4**	73.9	–	–	–	–	
(Cheng et al. 2018)	**99.6**	98.1	87.0	96.0	82.8	–	68.2	94.0	83.7	73.0	76.8	
(Shi et al. 2018)	**99.6**	**98.8**	93.4	**99.2**	**93.6**	91.8	76.1	–	–	78.5	79.5	
SAR (Ours)	99.4	98.2	**95.0**	98.5	91.2	94.0	**78.8**	**95.8**	**91.2**	**86.4**	**89.6**	**66.8**

消融实验：

- 加入真实图片提高近10个点；
- 2D-Attention比1D提高2~3个点；
- 减少CNN和LSTM隐含层数量会降低性能;(CNN更明显)
- 改变下采样率（改变特征图大小）会导致性能小幅下降



Model Configuration												
Training data	CNN channels	Down-sampling ratio	Attention module	LSTM layers	Hidden state size	IIIT5K	SVT	IC13	IC15	SVTP	CT80	COCO-T
Synth+Real	×1	1/8, 1/4	2D proposed	2	512	95.0	91.2	94.0	78.8	86.4	89.6	66.8
	×1/2	1/8, 1/4	2D proposed	2	512	92.7	88.7	92.0	75.6	81.3	86.8	62.6
	×1	**1/16, 1/4**	2D proposed	2	512	93.8	90.3	92.7	77.4	84.5	89.2	64.8
	×1	**1/16, 1/8**	2D proposed	2	512	94.0	90.6	93.1	76.2	83.7	87.5	63.7
	×1	1/8, **1/8**	2D proposed	2	512	93.6	89.3	92.5	76.1	82.8	87.5	63.3
	×1	1/8, 1/4	**2D traditional**	2	512	94.0	90.1	92.3	77.2	84.3	87.5	64.2
	×1	1/8, 1/4	**1D**	2	512	93.0	89.9	90.2	76.6	83.6	84.7	65.4
	×1	1/8, 1/4	2D proposed	**1**	512	89.7	87.2	87.4	70.6	76.4	80.6	60.1
	×1	1/8, 1/4	2D proposed	2	**256**	94.0	89.3	92.8	76.8	83.7	86.5	63.8
OnlySynth	×1	1/8, 1/4	2D proposed	2	512	91.5	84.5	91.0	69.2	76.4	83.3	–