

Paper Reading

卢宁¹

2020 年 11 月 8 日

- 1 Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning
- 2 Graph Convolution for Multimodal Information Extraction from Visually Rich Documents

Table of Contents

- 1 Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning
- 2 Graph Convolution for Multimodal Information Extraction from Visually Rich Documents

动机

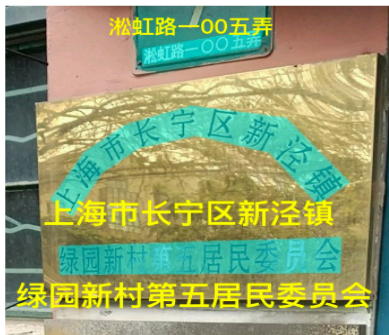
- ① 中文字符比拉丁字符要多很多
- ② 中文字符的分布特别不平衡
- ③ 每个字符的训练样本有限

- ① 提出了更大的中文新数据集 **C-SVT**，比已有最大的中文数据集大 14 倍
- ② 提出了一种利用完全和弱标注信息的，基于部分监督的端到端可训练中文识别网络。

- ① 430000 中文数据集，比现有的数据大 14 倍
- ② 全标注：标注了精确文本区域位置和内容。弱标注：标注了 ROI 的区域 mask 和文字
- ③ 完全监督学习：只用全标注数据。弱监督学习：只用弱标注数据。部分监督：两者混着用。

数据集

全标注



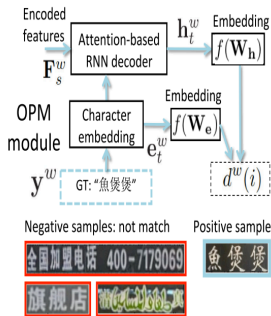
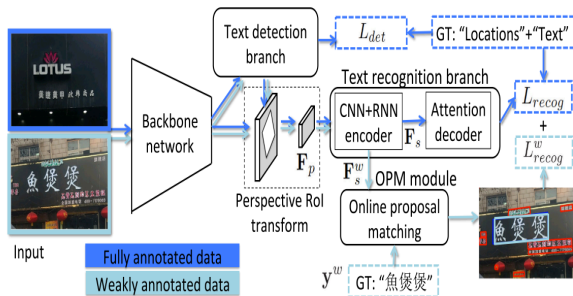
train:val:test 4:1:1, 29966 images, 243537 text lines, 1509256 chars

弱标注

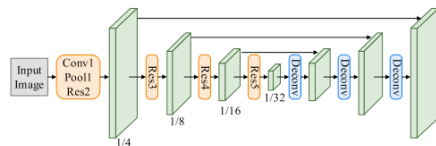


400000 images, 5 M chars

总体框架



Text Detection Branch



FOTS¹ like: ResNet 50 (shared backbone) + FPN

$$L_{det} = L_1 + L_{reg}$$

¹Xuebo Liu et al., 2018, FOTS: Fast Oriented Text Spotting with a Unified Network

Perspective ROI Transform

固定高度，等宽高比进行透视变换，与 ROIRotate 不同的是参数是可学习的。

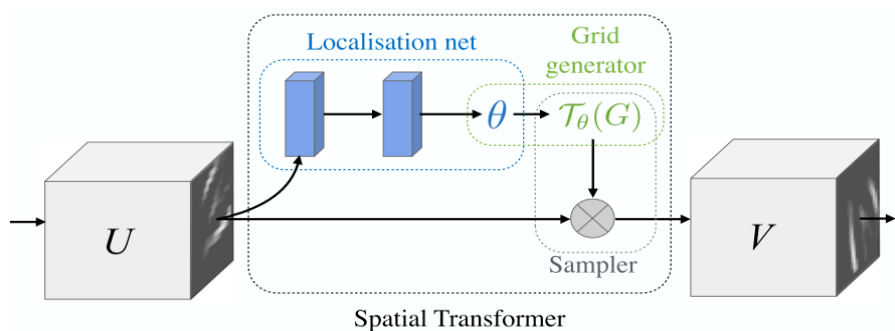
Spatial Transformer Network¹

$$\begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} = \mathcal{T}_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & 1 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (1)$$

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (2)$$

¹Jaderberg, Max et al., 2015, Spatial Transformer Networks

Perspective ROI Transform



Perspective ROI Transform

变换矩阵 M 是直接通过将三个基础变换旋转，缩放和平移组合成一个仿射变换而成的。（为什么不学习？）

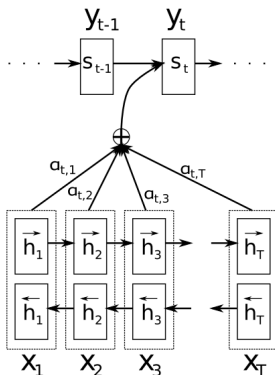
ROI Rotate¹

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \\ &= s \begin{bmatrix} \cos \theta & -\sin \theta & t_x \cos \theta - t_y \sin \theta \\ \sin \theta & \cos \theta & t_x \sin \theta + t_y \cos \theta \\ 0 & 0 & \frac{1}{s} \end{bmatrix} \end{aligned} \quad (3)$$

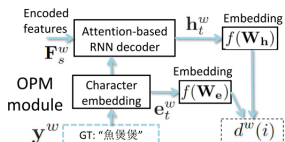
¹Xuebo Liu et al., 2018, FOTS: Fast Oriented Text Spotting with a Unified Network

Text Recognition Branch

Soft-attention Based Bidirectional GRUs. Please refer to [code](#)



Online Proposal Matching



$$d^w(i) = \frac{1}{T^w} \sum_{t=1}^{T^w} \|f(h_t^w, W_h) - f(e_t^w, W_e)\| \quad (4)$$

$$L_{opm} = \frac{1}{N} \sum_{i=1}^N [s^w(i)]^2 \quad (5)$$

where $s^w(i) = d^w(i)$ if the text proposal $P^w(i)$ is a positive sample that matches the keyword y^w , otherwise $s^w(i) = \max(0, 1 - d^w(i))$.

Negative samples: not match



Positive sample



Fully and Weakly Supervised Joint Training

Fully Supervised Training

$$L_{full} = L_{det} + \beta L_{recog} \quad (6)$$

Partially Supervised Training

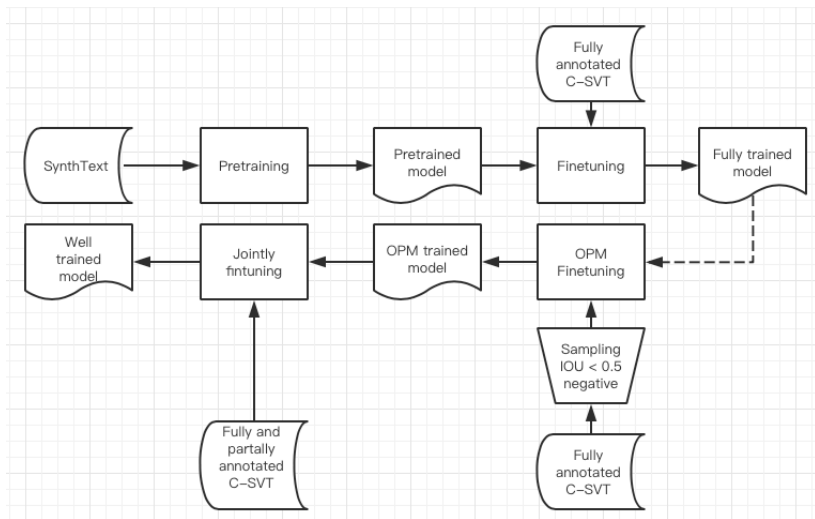
$$L_{total} = L_{det} + \beta (L_{recog} + L_{recog}^w) \quad (7)$$

$$L_{recog}^w = \frac{1}{\sum_{i=1}^N m(i)} \sum_{i=1}^N m(i) l_{recog}^w(i) \quad (8)$$

$$m(i) = 1 \text{ if } d_m(i) \leq \tau \text{ otherwise } m(i) = 0$$

$$l_{recog}^w(i) = -\frac{1}{T^w} \sum_{t=1}^{T^w} \log p(\mathbf{y}_t^w | \mathbf{y}_{t-1}^w, \mathbf{h}_{t-1}^w, \mathbf{c}_t^w) \quad (9)$$

Training Pipeline



实现细节

- ① random scale $[0.5, 1.0, 2.0, 3.0]$ \rightarrow random crop \rightarrow resize longer side to 512 \rightarrow padding to 512×512 with 0
- ② roi transform to 8×64 , less padding and larger bilinear resizing
- ③ weakly annotated image resize to 512×512 with padding
- ④ 8 GPUs NVIDIA Tesla P40, 16 batch size with 32 proposals when fully supervised, 8 fully and 8 weakly with 32 proposals when partially supervised
- ⑤ Adam with learning rate 10^{-4} $\lambda = 0.01$ $\beta = 0.02$

实验结果




Method		Training data	Valid								Test							
			Detection			End-to-end					Detection			End-to-end				
			R %	P %	F %	R %	P %	F %	AED	R %	P %	F %	R %	P %	F %	AED		
EAST ^[46] +Attention ^[35]		Train	71.74	77.58	74.54	23.89	25.83	24.82	22.29	73.37	79.31	76.22	25.02	27.05	25.99	21.26		
EAST ^[46] +CRNN ^[34]		Train	71.74	77.58	74.54	25.78	27.88	26.79	20.30	73.37	79.31	76.22	26.96	29.14	28.0	19.25		
End2End	 	Train	72.70	78.21	75.35	26.83	28.86	27.81	20.01	74.60	80.42	77.40	27.55	29.69	28.58	19.68		
		Train + 4.4K Extra Full	72.98	78.46	75.62	28.03	30.13	29.04	19.62	74.95	80.84	77.79	28.77	31.03	29.85	19.06		
		Train + 10K Extra Full	73.23	76.69	74.92	29.91	31.32	30.60	18.87	75.13	78.82	76.93	30.57	32.07	31.30	18.46		
End2End-PSL	 $\frac{1}{12}$	Train + 25K Weak	72.93	79.37	76.01	29.44	32.04	30.68	19.47	74.72	81.39	77.91	30.18	32.87	31.46	18.82		
		Train + 50K Weak	73.09	79.36	76.10	29.96	32.53	31.19	19.20	74.80	81.32	77.93	30.56	33.22	31.83	18.72		
		Train + 100K Weak	73.17	78.50	75.74	30.55	32.78	31.63	18.97	75.04	80.41	77.63	31.19	33.43	32.27	18.28		
		Train + 200K Weak	73.26	78.64	75.85	31.31	33.61	32.41	18.54	75.14	80.68	77.81	32.01	34.38	33.15	18.12		
		Train + 400K Weak	73.31	79.73	76.38	31.80	34.58	33.13	18.14	75.21	81.71	78.32	32.53	35.34	33.88	17.59		

Table of Contents

- 1 Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning
- 2 Graph Convolution for Multimodal Information Extraction from Visually Rich Documents

- ① 同样的文字在不同位置代表不同的语义，同样的位置不同模板代表不同语义
- ② 鲁棒性不足（模板，拍照环境）
- ③ 无法编码空间依赖特性

- ① 第一个使用 GCN 来编码文字框结构和文本特征来做 VRD 的 IE 任务
- ② 实验表明，结合结构和文字特征的方法会比 baseline 方法更好。

数据集

VATI（增值税发票）

16 个字段（购买方，出售方，日期，总额），统一模板，有一些干扰

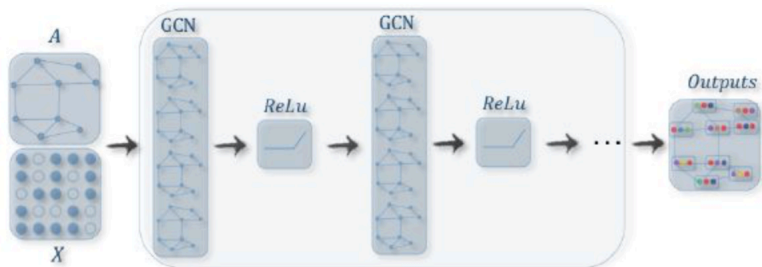
IPR（国际购物小票）

4 个字段（发票号，商户名，购买方，总额），146 个模板

Train:Val:Test 0.75:0.15:0.15

GCN 简单介绍

图卷积网络将卷积运算从传统数据（例如图像）推广到图数据。其核心思想是学习一个函数映射 $f(\cdot)$ ，通过该映射图中的节点 v_i 可以聚合它自己的特征 x_i 与它的邻居特征 $x_j \in N(v_i)$ 来生成节点 v_j 的新表示。图卷积网络是许多复杂图神经网络模型的基础，包括基于自动编码器的模型、生成模型和时空网络等。



条件随机场的定义

如果随机变量 Y 构成一个由无向图 $G = (V, E)$ 表示的马尔可夫随机场, 对任意节点 $v \in V$ 都成立, 即

$$P(Y_v | X, Y_w, w \neq v) = P(Y_v | X, Y_w, w \sim v)$$

对任意节点 v 都成立, 则称 $P(Y|X)$ 是条件随机场。式中 $w \neq v$ 表示 w 是除 v 以外的所有节点, $w \sim v$ 表示 w 是与 v 相连接的所有节点。

线性链条件随机场定义

设两组随机变量, $\mathbf{X} = (X_1, \dots, X_n)$ 和 $\mathbf{Y} = (Y_1, \dots, Y_n)$, 那么线性链条件随机场的定义为:

$$P(Y_i | \mathbf{X}, Y_i, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) = P(Y_i | \mathbf{X}, Y_{i-1}, Y_{i+1})$$

特征函数

线性链条件随机场的参数化形式

$$P(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)$$

$$Z(x) = \sum_y \left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)$$

其中

t_k 是定义在 **边**上的特征函数，称为转移特征

s_l 是定义在 **结点**上的特征函数，称为状态特征

注意到这种表达就是不同特征的加权求和形式， t_k, s_l 都依赖于位置，是局部特征函数。

CRF 简单介绍

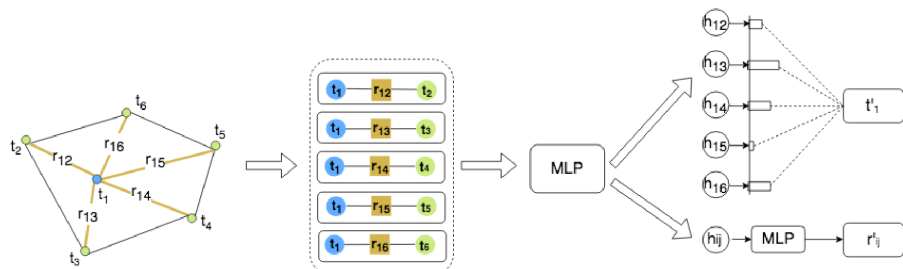
- ① 前向后向算法, 计算 $P(Y = y_i|x)$ 和 $P(Y_{i-1} = y_{i-1}, Y_i = y_i|x)$
 - ② 解码算法, 找到一个概率最大的最优序列, 维特比算法
- 具体细节, 请参考 [here](#)

Mathematically, a document D is a tuple (T, E) , where $T = t_1, t_2, \dots, t_n$, $t_i \in T$ is a set of n text boxes/nodes, $R = \{r_{i1}, r_{i2}, \dots, r_{ij}\}$, $r_{ij} \in R$ is a set of edges, and $E = T \times R \times T$ is a set of **directed edges (dense?)** of the form (t_i, r_{ij}, t_j) where $t_i, t_j \in T$ and $r_{ij} \in R$. In our experiments, every node is connected to each other.

初始特征抽取

- ① 节点信息：通过 BiLSTM 编码文本信息，字符级别输入
- ② 边信息： $r_{ij} = [x_{ij}, y_{ij}, w_i/h_i, h_j/h_i, w_j/h_i]$ 归一化

GCN



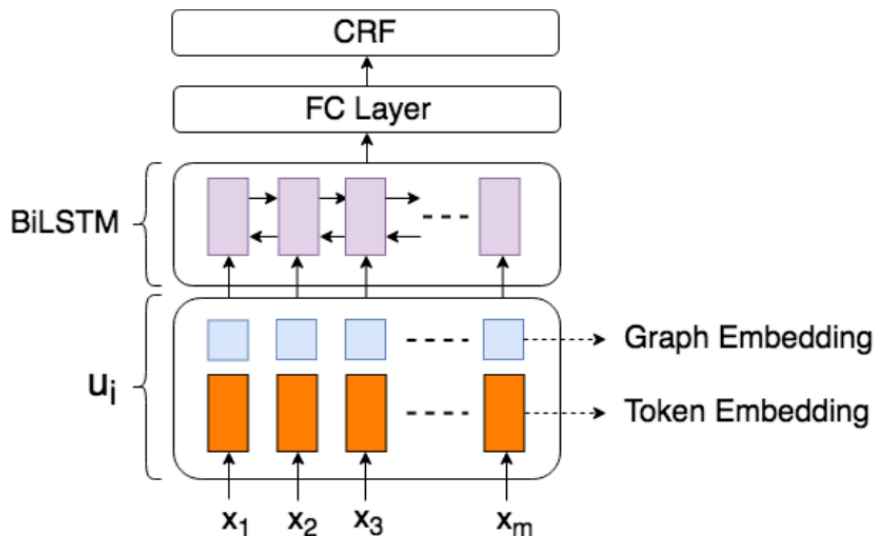
$$\mathbf{h}_{ij} = g(\mathbf{t}_i, \mathbf{r}_{ij}, \mathbf{t}_j) = \text{MLP}([\mathbf{t}_i \parallel \mathbf{r}_{ij} \parallel \mathbf{t}_j]) \quad (10)$$

$$\mathbf{t}'_i = \sigma \left(\sum_{j \in \{1, \dots, n\}} \alpha_{ij} \mathbf{h}_{ij} \right) \quad (11)$$

$$\alpha_{ij} = \frac{\exp(\text{LeakyRelu}(\mathbf{w}_a^T \mathbf{h}_{ij}))}{\sum_{j \in \{1, \dots, n\}} \exp(\text{LeakyRelu}(\mathbf{w}_a^T \mathbf{h}_{ij}))} \quad (12)$$

$$\mathbf{r}'_{ij} = \text{MLP}(\mathbf{h}_{ij}) \quad (13)$$

BiLSTM-CRF with graph embedding



首先标注关键字段的值和位置，然后使用 OCR 系统跑一遍数据，将两个结果的框进行匹配，如果匹配超过某个阈值就认为匹配上了，然后再通过字符串匹配方法对 OCR 系统出来的文字内容进行 IOB 标注。

实验结果

Model	VATI	IPR
Baseline I	0.745	0.747
Baseline II	0.854	0.820
BiLSTM-CRF + GCN	0.873	0.836

Entities	Baseline I	Baseline II	Our model
Invoice #	0.952	0.961	0.975
Date	0.962	0.963	0.963
Price	0.527	0.910	0.943
Tax	0.584	0.902	0.924
Buyer	0.402	0.797	0.833
Seller	0.681	0.731	0.782

实验结果

Configurations/Datasets	VATI	IPR
Full model	0.873	0.836
w/o vis. features	0.808	0.775
w/o text features	0.871	0.817
w/o attention	0.872	0.821

Model	VATI	IPR
BiLSTM-CRF + GCN	0.873	0.836
+ Multi-task	0.881	0.849

