Robust Lexicon-Free Confidence Prediction for Text Recognition

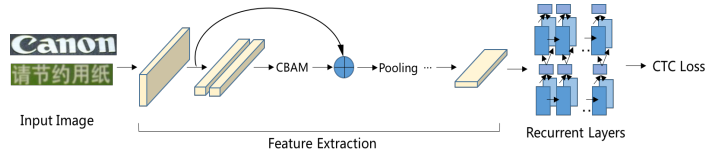
Qi Song, Qianyi Jiang, Rui Zhang and Xiaolin Wei
Meituan, Beijing, China

Introduction

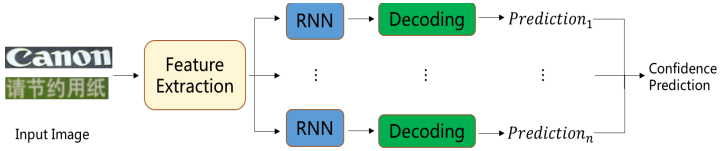
- OCR is an important computer vision task that reading text from images ,the text recognition results are vulnerable to slight perturbation in input images,
- The confidence prediction, as a validity estimation for the text recognition result, has not been fully researched.
- We propose a coarse-to-fine method for lexicon-free OCR confidence prediction which can be embedded with any text sequence recognition networks.

Methods

A. Text Recognizer: the recognizer is an encoder-decoder structure network as shown in (a)



(a) The architecture of the baseline recognizer



(b) The architecture of the SIMO for inference

B. Coarse-to-fine Confidence Prediction:

In the lexicon-free scenario, let x be the input image and c = [c1,c2,...,cD] of length D be the predicted string. The confidence score is C(c) = p(c|x)

- Coarse Scoring:

A Single-Input Multi-Output network (SIMO) is presented, the coarse confidence score is:

$$C_c(c) = \frac{K}{N}$$

- Refined Scoring:

After acquiring K valid candidates, the conditional probabilities of Top-1 probable character can be calculated, the final refined score is evaluated as follow

$$\bar{\mathbf{p}}(s|x) = \frac{1}{K}\sum_{i=1}^{K}\mathbf{p}(s_i|x),$$

$$C_r(\mathbf{c}) = \begin{cases} \min_{\bar{p}(s_i|x)\in\bar{\mathbf{p}}(s|x)} \bar{p}(s_i|x), & \text{if } C_c(\mathbf{c}) \geq \epsilon \\ 0 & otherwise \end{cases}$$

Algorithm 1 Procedure of training the SIMO for coarse scoring

Input:
The text images and annotations in trainset, $\mathbf{I}, \mathbf{A}$;
The parameters of SGDR, $\eta_{min}, \eta_{max}, T_0, \alpha$;
The number of models, N;
Output:
The SIMO, Θ_{simo};

1: Training a text recognition network with $\mathbf{I}, \mathbf{A}$, using SGDR, until converging to Θ_c;
2: Freezing CNNs in Θ_c and setting α to 1;
3: Continuing to train the models with N equal cosine periods;
4: Saving N models when the learning rate reaches η_{min} in every period;
5: Merging N models into one model Θ_{simo} with one input and N outputs;
6: **return** Θ_{simo};

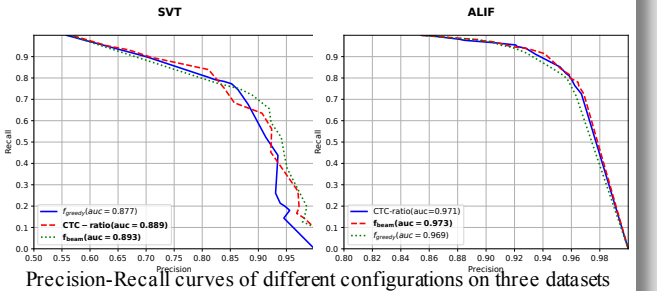
Results

Accuracy/AUC	IC13	SVT	ALIF	MSRA-TD500
Beam searching	75.38±0.59%	56.23±0.64%	85.5±1.0 %	59.76±0.49%
Greedy searching	75.00±0.55%	55.61±0.49%	85.4±1.0 %	59.75±0.86%
CTC-ratio	**0.979**	0.889	0.971	0.938
Ours, f_{greedy}	0.971	0.877	0.969	0.931
Ours, f_{greedy} (model ensemble)	-	-	-	0.910
Ours, f_{beam}	**0.979**	**0.893**	**0.973**	**0.939**

AUC OBTAINED BY DIFFERENT METHODS.

Dataset	Time (s)	Method		
		CTC-ratio	f_{beam}	f_{greedy}
IC13	average	**0.041**	0.156	0.046
	top50	**0.033**	0.132	**0.033**
	top90	0.053	0.129	**0.049**
	top99	0.080	0.361	**0.076**
	top999	0.293	0.775	**0.229**
SVT	average	**0.039**	0.143	0.049
	top50	**0.031**	0.121	0.032
	top90	0.049	0.208	**0.047**
	top99	**0.066**	0.296	**0.066**
	top999	**0.078**	0.341	0.079
ALIF	average	**0.040**	0.079	0.073
	top50	**0.019**	0.060	0.053
	top90	**0.113**	0.156	0.143
	top99	**0.218**	0.272	0.263
	top999	**1.650**	1.725	2.058
MSRA-TD500	average	7.163	57.051	**0.085**
	top50	5.927	47.431	**0.046**
	top90	13.113	104.086	**0.083**
	top99	25.859	211.020	**0.425**
	top999	33.337	299.460	**0.555**

THE AVERAGE, TOP50, TOP90, TOP99 AND TOP999 TIME PERFORMANCE ON DIFFERENT DATASETS



Precision-Recall curves of different configurations on three datasets

Conclusion

In this paper, we study the lexicon-free OCR confidence prediction. We present a coarse-to-fine framework that consists of two stages. Comprehensive experiments show the proposed framework is high competitive in both effectiveness and efficiency. Our framework can be applied in both Latin and non-Latin languages with different decoding approaches.