Robust Lexicon-Free Confidence Prediction for Text Recognition

Qi Song, Qianyi Jiang, Rui Zhang and Xiaolin Wei

Meituan,
Beijing, China

Paper ID: 971

Introduction

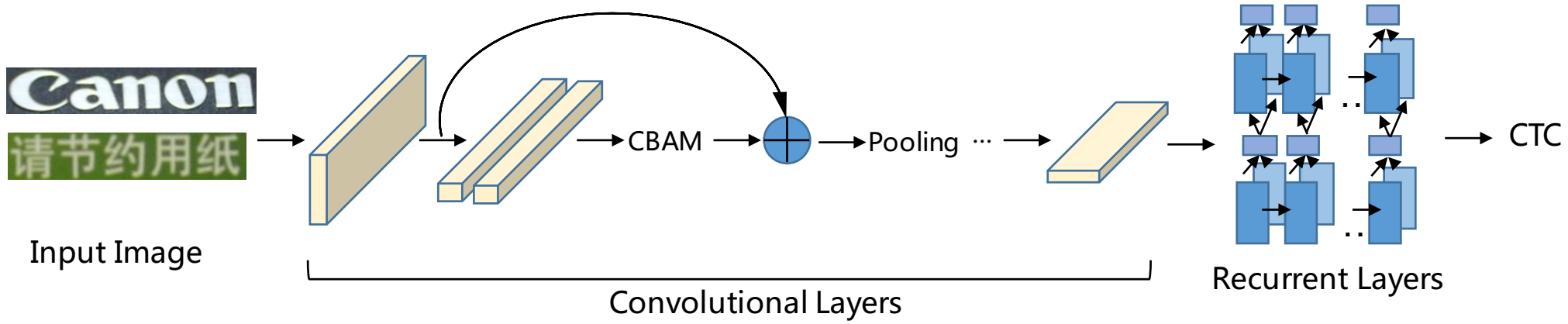
Optical Character Recognition (OCR) is booming in recent years. As we all know, the text recognition results are vulnerable to slight perturbation in input images, thus a method for measuring how reliable the results are is crucial.

In this paper, We propose a coarse-to-fine method for lexicon-free OCR confidence prediction which can be embedded with any text sequence recognition networks.

Method

A. Text Recognizer

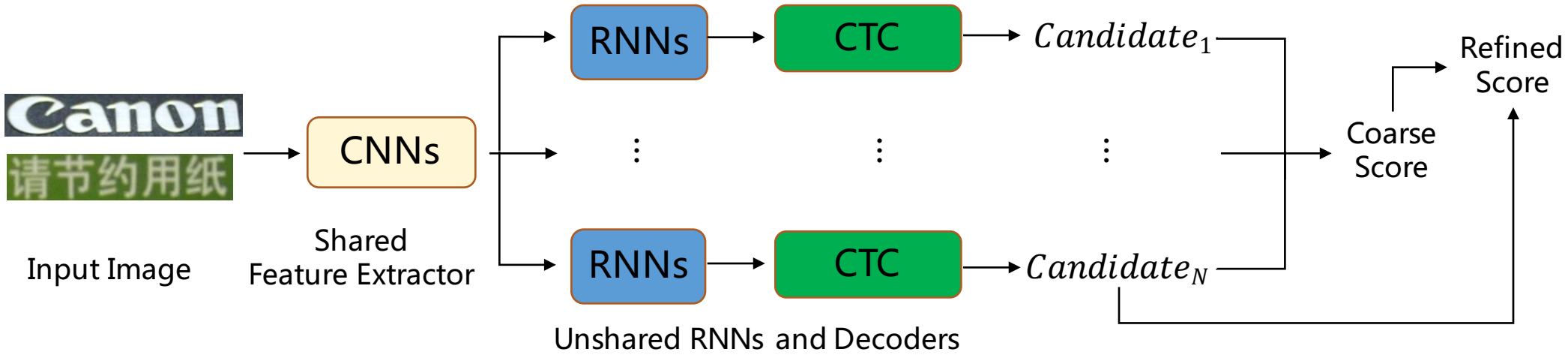
the recognizer is an encoder-decoder structure network. The encoder consists of two components, a CNN based visual feature extraction module and an RNN based semantic modeling module. In the decoding stage, a CTC layer is utilized for transcription.



Method

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 C(c) can be presented as a conditional probability, C(c) = p(c|x).

- Coarse Scoring: $C_c(\mathbf{c}) = \frac{K}{N}.$



The architecture of the SIMO for inference.

Method

- Refined Scoring: After acquiring K valid candidates, the conditional probabilities of Top-1 probable character sequence p(s|x) can be calculated from valid candidates. the final refined score is evaluated as follows：

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

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

Evaluation Results

TABLE II
AUC OBTAINED BY DIFFERENT METHODS. THE FIRST TWO ROWS ARE TEXT RECOGNITION ACCURACIES OF THE SIMO WITH DIFFERENT DECODING APPROACHES.

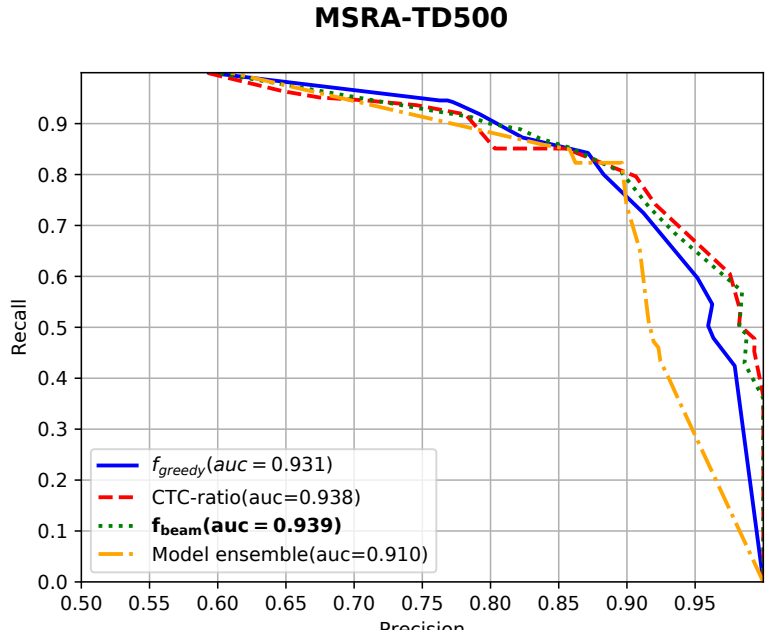
Accuracy/AUC	IC13	SVT	ALIF	MSRA-TD500
Beam searching	75.38±0.59%	56.23±0.64%	85.5±1.0 %	59.76±0.49%
Greedy searching	75.00±0.55%	55.61±0.49%	85.4±1.0 %	59.75±0.86%
CTC-ratio	**0.979**	0.889	0.971	0.938
Ours, f_{greedy}	0.971	0.877	0.969	0.931
Ours, f_{greedy} (model ensemble)	-	-	-	0.910
Ours, f_{beam}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

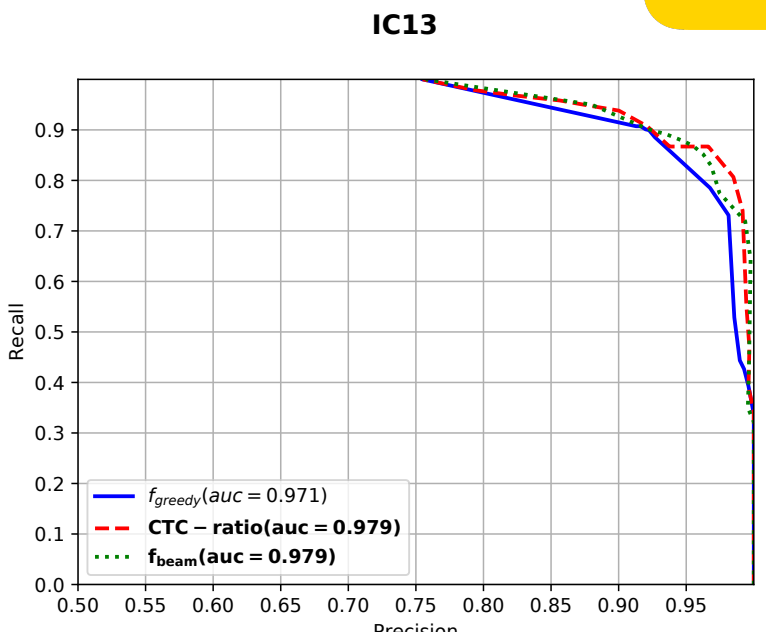
Evaluation Results

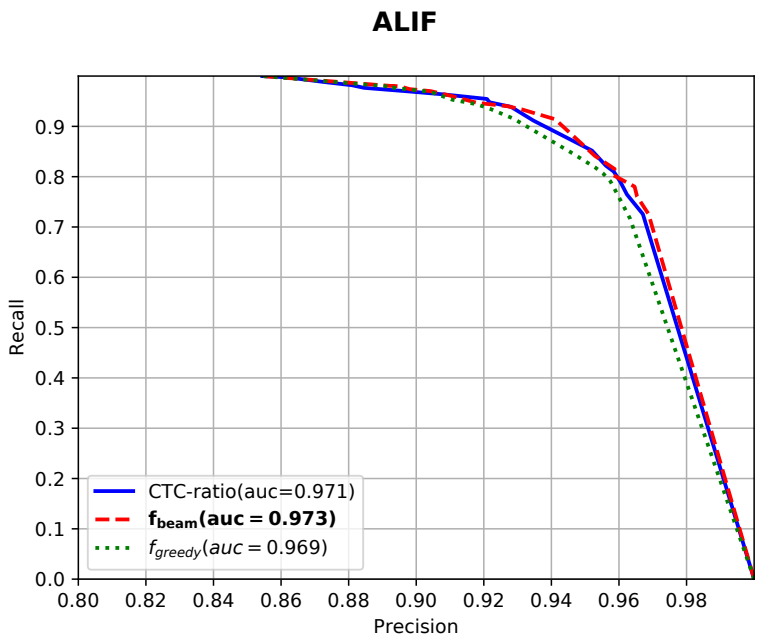
TABLE III
THE AVERAGE, TOP50, TOP90, TOP99 AND TOP999 TIME PERFORMANCE ON DIFFERENT DATASETS. THE TOP* INDICATES THE RUNNING TIME AT THE * PERCENT IN A ASCENDING SORTING.

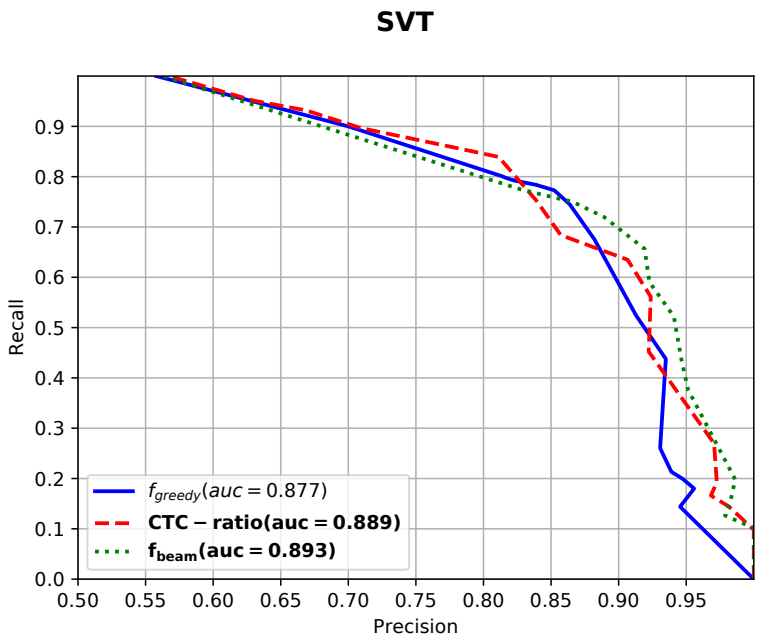
Dataset	Time (s)	Method		
		CTC-ratio	f_{beam}	f_{greedy}
IC13	average	**0.041**	0.156	0.046
	top50	**0.033**	0.132	**0.033**
	top90	0.053	0.129	**0.049**
	top99	0.080	0.361	**0.076**
	top999	0.293	0.775	**0.229**
SVT	average	**0.039**	0.143	0.049
	top50	**0.031**	0.121	0.032
	top90	0.049	0.208	**0.047**
	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**

Evaluation Results









Conclusion

- We present a coarse-to-fine framework that consists of two stages: For the first stage, a solution named SIMO is proposed to calculate a coarse score. For the second stage, a transform function is invented to refine the coarse score.
- 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.