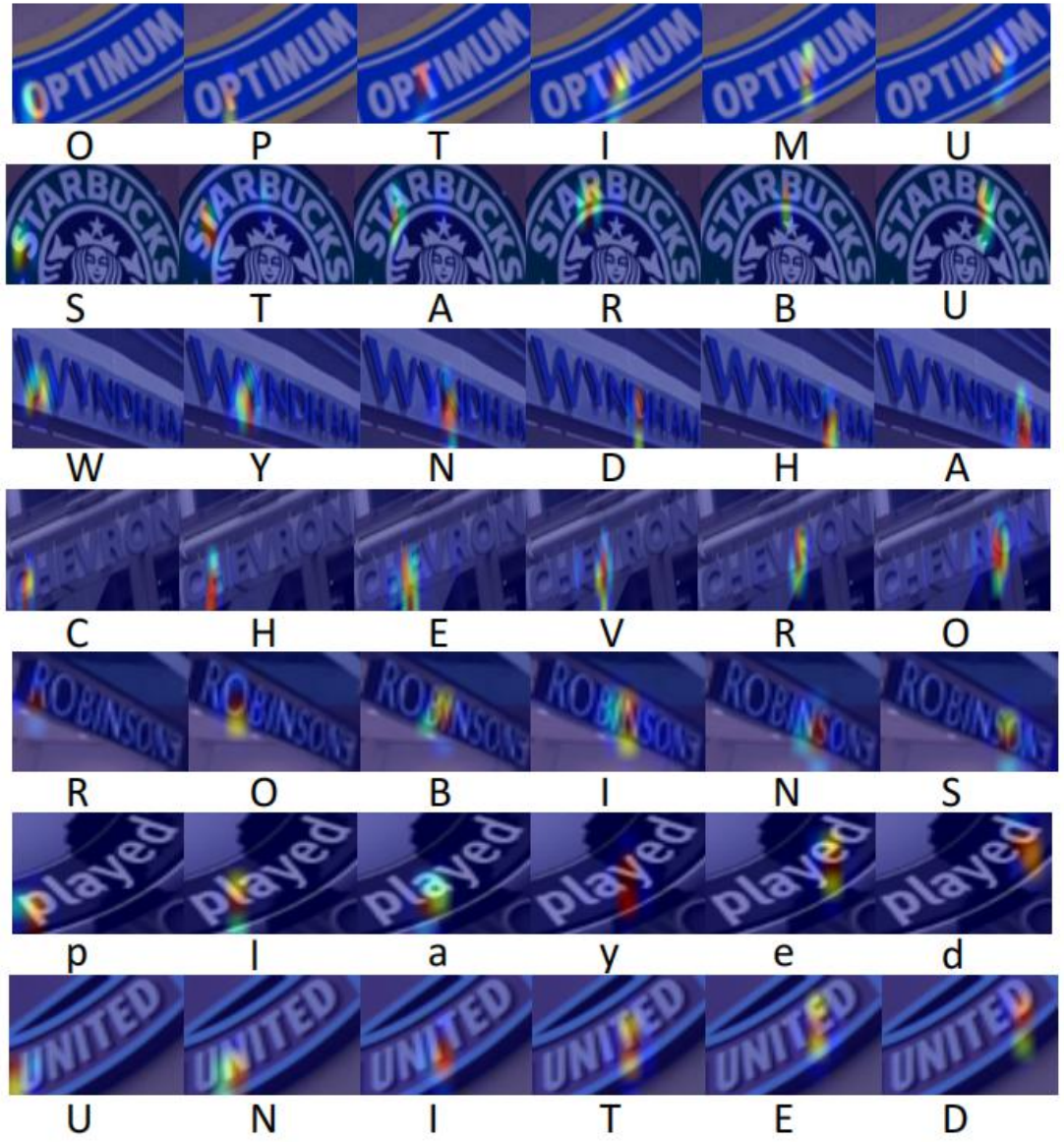
Weakly Supervised Attention Rectification for Scene Text Recognition

Chengyu Gu, Shilin Wang*, Yiwei Zhu, Zheng Huang, Kai Chen

Chengyu Gu

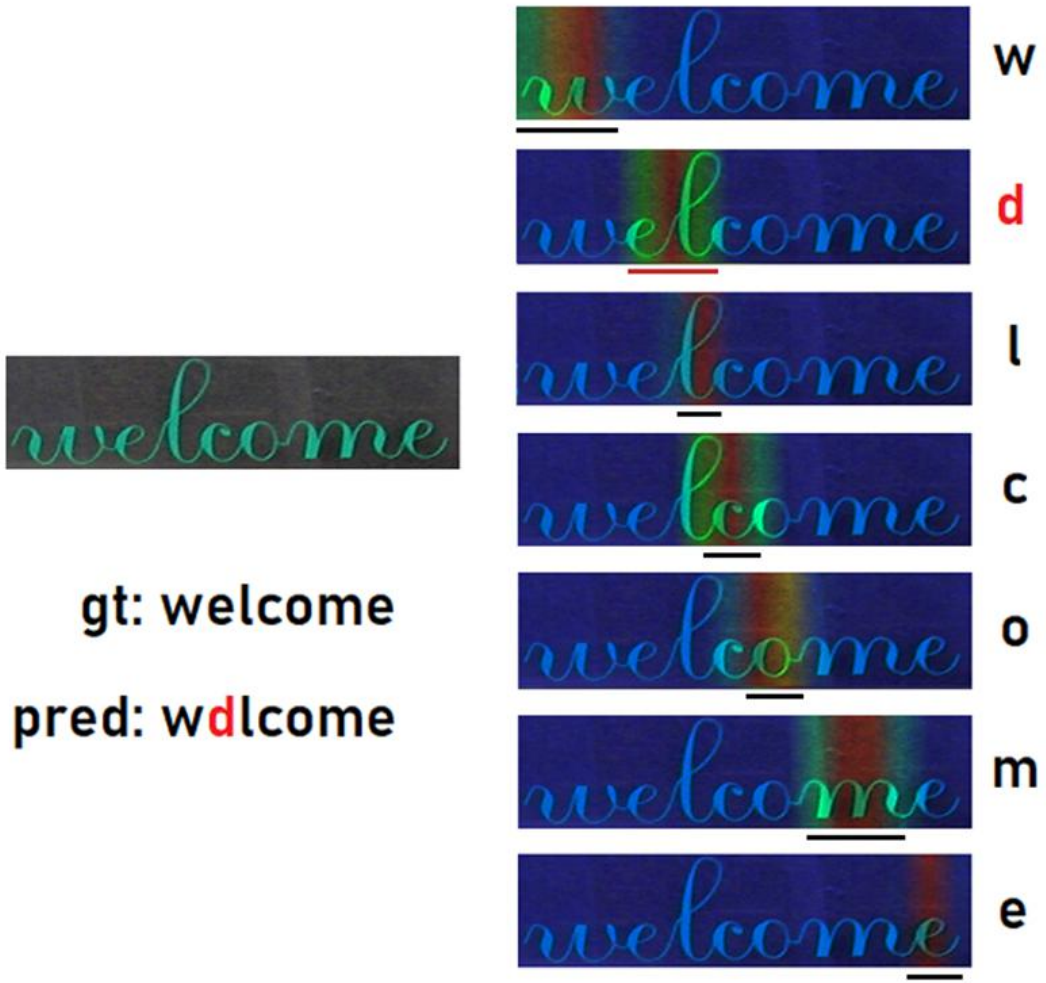
Attention Mechanism



Robustness
Interpretability

Attention Drift

The attention of the second character falls between the character ‘e’ and the character ‘l’, and the model mistakes these two characters for one character ‘d’.



Attention Drift

gt:2003

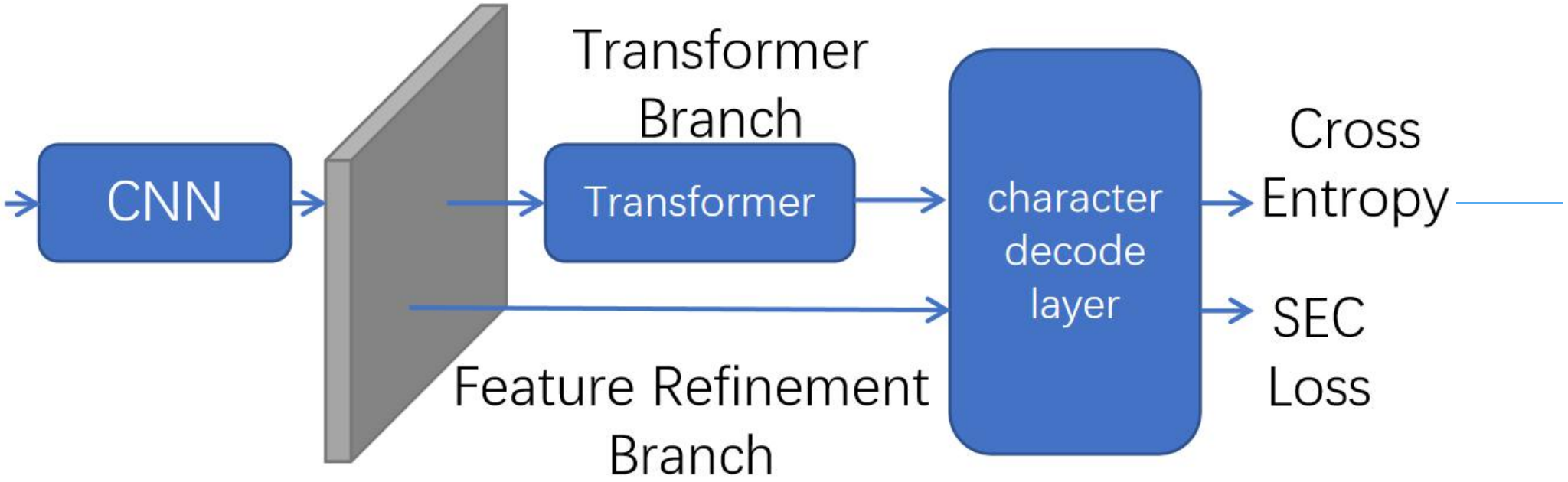
m	m	l	e	e	e	1	j	i	o	0	o	w	l	n	m	e	o	i	l	e	a	i	l	l
m	a	r	e	2	2	1	1	1	0	0	q	l	n	n	o	n	0	l	b	3	3	i	l	w
m	m	m	2	2	2	1	1	1	0	0	q	l	l	n	o	o	l	l	b	3	j	l	l	l
m	m	1	2	2	2	1	1	1	0	1	l	l	0	o	o	o	l	l	3	3	l	l	l	l
o	d	1	2	2	2	1	i	p	4	4	l	o	o	0	o	a	l	u	s	3	r	i	l	l
o	l	2	2	2	7	i	r	h	h	4	l	o	o	o	o	l	l	s	s	s	e	i	i	l
o	l	l	l	l	i	r	r	h	h	l	l	i	o	r	b	l	l	l	s	t	e	i	i	i
o	a	e	e	s	s	a	a	c	a	e	a	a	a	a	e	e	e	e	s	s	e	s	s	l

pred: 2103

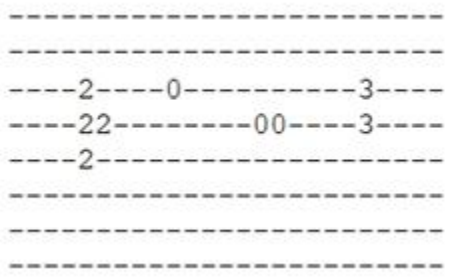
Noisy feature vectors in the background area can confuse the attention module

The noisy feature vectors are not supervised sufficiently for their low attention weights.

Method


CNN
Transformer Branch
Transformer
Feature Refinement Branch
character decode layer
Cross Entropy
SEC Loss

Spatial Existance Classification(SEC)



The existence probabilities for each char. (p_k)

0 :0.90
2: :0.95
3: :0.66
others :0.00

Cross-Entropy

Ground Truth:

0 :1
2 :1
3 :1
others :0

$$\mathrm{SEC}(\omega \mid \tau, S) = \frac{1}{|C|}\sum_{k=1}^{|C|}\mathrm{CE}(p_k, \delta(C_k \mid S))$$

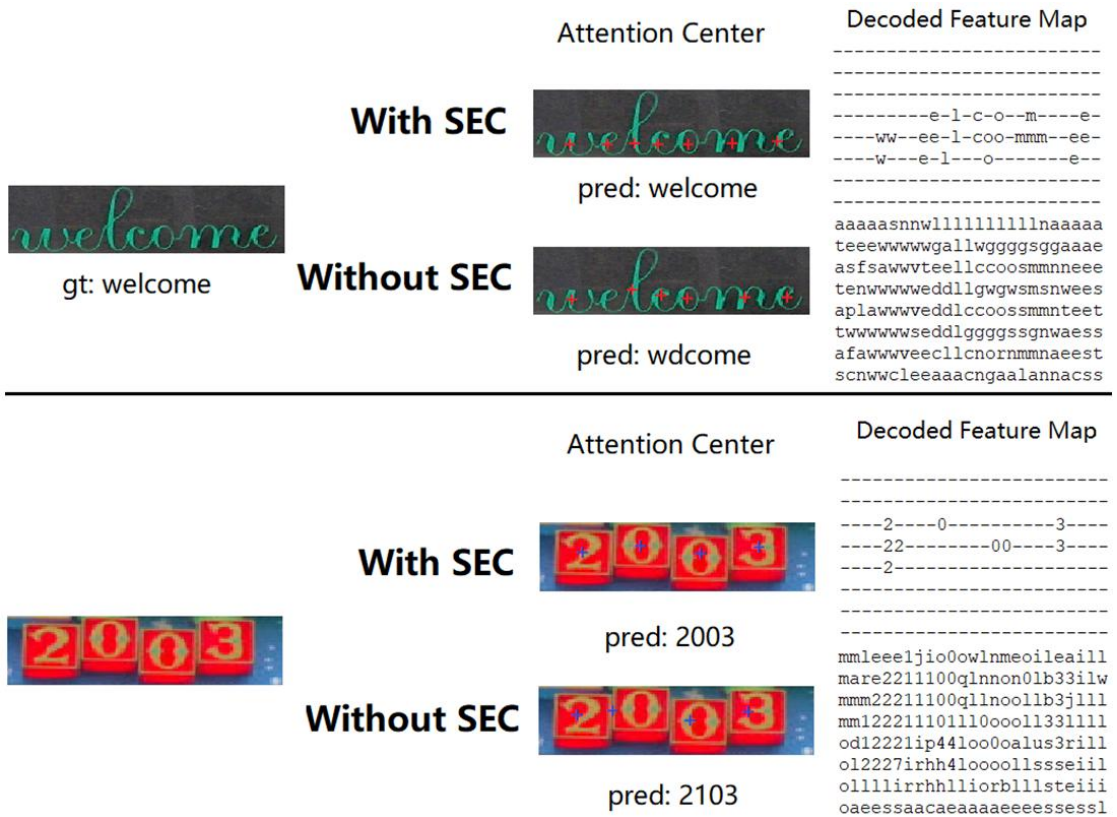
$$\delta(C_k \mid S) = \begin{cases} 1 \text{ if } C_k \in S \\ 0 \text{ if } C_k \notin S \end{cases}$$

where p_k represents the existence probability of the k-th character Ck in character set C, ω is the network parameters, and CE(•) is the cross-entropy loss function.

$$p_k = 1 - \prod_{x,y}(1 - p_k^{x,y})$$

where $p_k^{x,y}$ is the decoded feature map.

Experiments



method	IIIT5K	SVT	IC03	IC13	IC15	SVTP	CT80
RNN baseline	89.2	85.4	92.6	90.1	71.9	73.6	72.0
RNN SEC	89.9 (+0.7)	87.5 (+2.1)	92.6 (+0.0)	90.5 (+0.4)	72.6 (+0.7)	75.0 (+1.4)	78.5 (+6.5)
Transformer baseline	91.4	87.4	93.6	91.2	75.6	77.6	79.2
Transformer SEC	92. (+1.5)	89.6 (+2.2)	95.3 (+1.7)	93.6 (+2.4)	79.9 (+4.3)	82.2 (+4.6)	84.3 (+5.1)

Thank you

Weakly Supervised Attention Rectification for Scene Text Recognition

Chengyu Gu