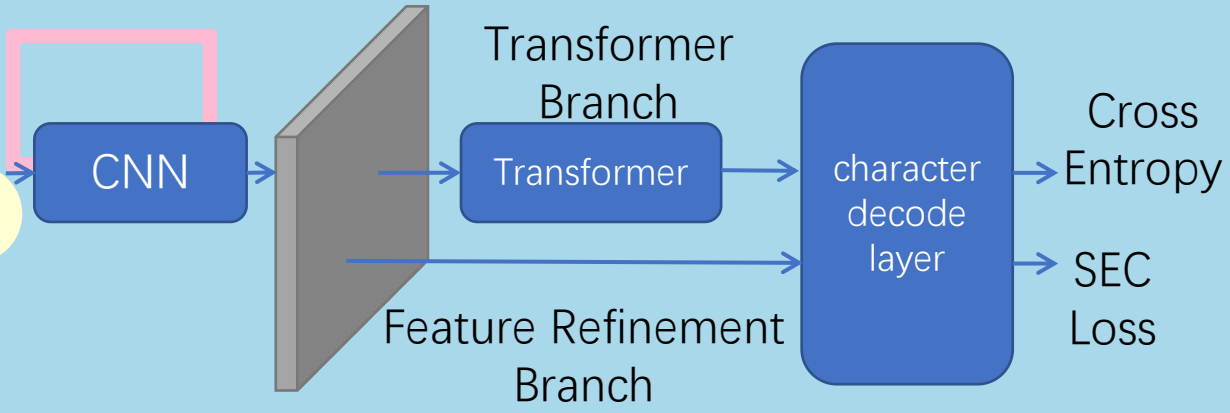
Weakly Supervised Attention Rectification for Scene Text Recognition

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

Motivation

1. The feature vectors in the background region can not be well supervised due to their very low attention weights when training.
2. The noise points in the feature map can confuse the attention module and lead to attention drift.
3. The quality of the feature map determine the correctness of prediction results when attention dirft happen.

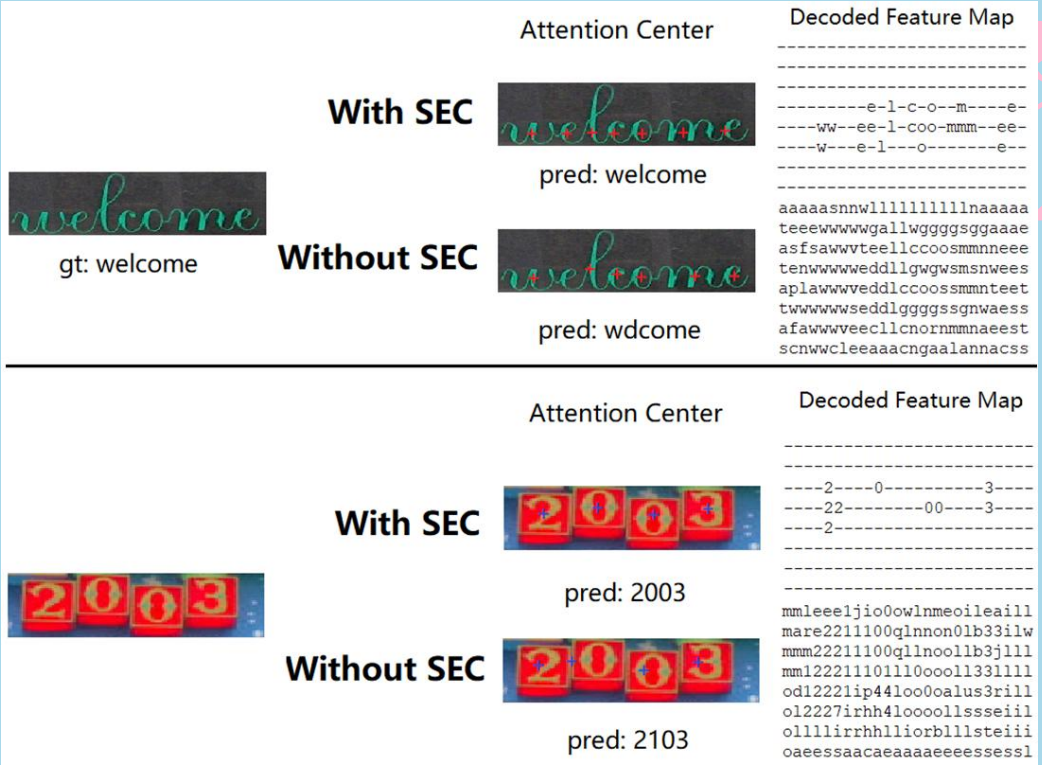
Method



1. Denoise feature map with SEC Loss on 2D feature map.
2. Both the Transformer branch and the feature refinement branch share the same character decode layer.
3. With the supervision of Spatial Existence Classification (SEC) loss, the attention module (RNN or Transformer) can better align the character area.

SEC Loss

1. For each character in the character set, calculate its probability of appearing in the decoded feature map.
2. Encode the annotation to binary code for each character in the character set according to its appearence in the annotation.
3. Calculate the cross-entropy between value from 1. and 2.



Results

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)

1. By supervising the feature map with SEC Loss, the model based on attention mechanism(i.e. RNN attention and Transformer attention) have constant improvment.
2. The improvements on irregular datasets are larger than that on regular datasets because irregular samples have more disturbing factors.
3. Our method achieve comparable performance with other SOTA methods.