TEXT RECOGNITION IN IMAGES BASED ON TRANSFORMER WITH HIERARCHICAL ATTENTION

Yiwei Zhu, Shilin Wang, Zheng Huang and Kai Chen

School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai, China

Introduction

Recognizing text in images has been a hot research topic in computer vision for decades due to its various application. However, the variations in text appearance in term of perspective distortion, text styles, etc., cause great trouble in text recognition. In this paper, we propose a new network with hierarchical attention mechanism, which could be trained end-to-end by using only image-level annotations and could recognize complex scene text efficiently and sufficiently.

Contribution

In view of the recognition accuracy of previous methods is not well optimized and their recognition speed is slow, a new deep neural network structure is proposed. The main contributions of this paper are as follows: i) the Transformer structure is first applied in text recognition, which discards the RNN units. Thus the context information can be represented sufficiently while greatly reducing the training time; ii) the two-dimensional attention mechanism is proposed to extract the context information from images; and iii) considering the characteristics of images containing texts, a hierarchical attention scheme is proposed to speed up the training and inference procedures.

The proposed method

Fig.1 shows the overall architecture of the proposed Hierarchical Attention Transformer Network (HATN). The ResNet50 without the final stack is selected as image feature extraction network for its high representative ability. And the transformer network is responsible for the final output text sequence. And Considering the inherent characteristics of the images containing texts, three kinds of relationships need to be considered, including character level context, word level context and sentence level context, the hierarchical attention mechanism is added to the encoder end, like the colored rectangles in Fig.1. And as showed in the Fig.2, instead of compressing the feature vectors into one dimension as in the traditional methods, we maintain the two-dimensional structure. So that, the spatial information of image can be well-preserved, which is crucial for the subsequent text recognition.

Experiments and Conclusion

As showed int the table 1 and table 2, extensive experiments on seven public datasets with regular and irregular text arrangements are constructed, which show that our model not only could recognize the scene text accurately, but also has very fast prediction speed.

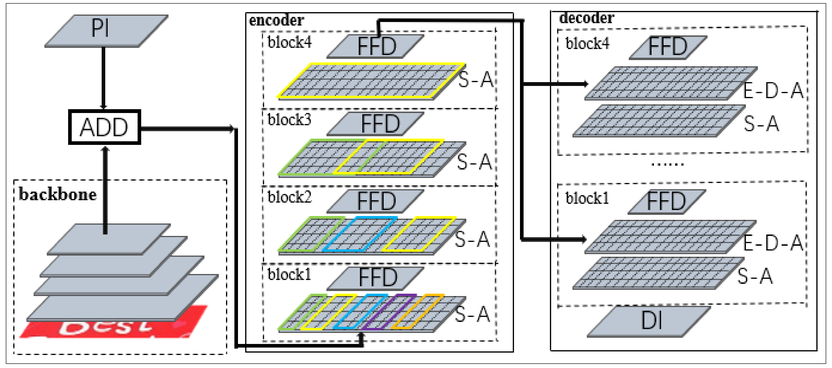
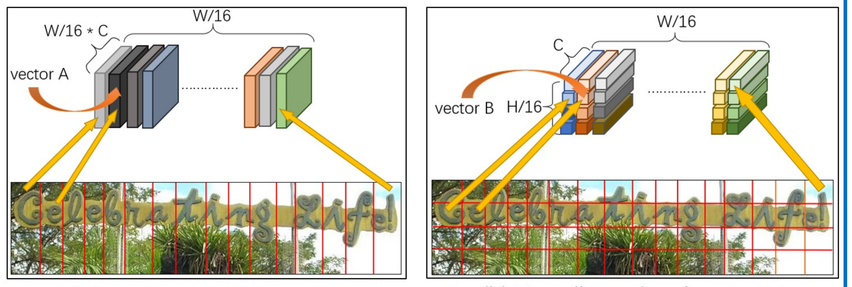


Fig.1 The overview of our network



(a) One dimensional structure (b) Two dimensional structure

Fig.2 Illustrations of 1D and 2D attention structure.

Method	IIIT 5K	SVT	IC03	IC13	SVT -P	CT80	IC 15
RCNN	78.2	80.8	89.4	86.7	66.8	54.9	N/A
RSAR	81.9	81.9	90.1	88.6	71.8	59.2	N/A
baseline of FAN	83.7	82.2	91.5	89.4	68.2	57.5	63.3
AON	87.0	**82.8**	91.5	N/A	73.0	**76.8**	68.2
w/o hierarchical	88.2	82.3	91.1	90.8	73.2	75.9	69.9
HATN	**88.6**	82.2	91.3	**91.1**	**73.5**	75.7	**70.1**
FAN	87.4	**85.9**	**94.2**	**93.3**	71.5	63.9	66.2

Table.1 Recognition accuracies on public datasets.

Method	Train time	Inference time
AON	90 hours	0.98s/90 images
FAN	120 hours	1s/ 90 images
baseline of FAN	85 hours	0.97s/90 images
w/o hierarchical	67 hours	0.85s/ 90 images
HATN	48 hours	0.68s/90 image

Table.2 Efficiency of different methods.

*HATN is the proposed network, and the w/o hierarchical is our baseline. FAN needs character-level annotation.