25th International Conference on Pattern Recognition

A Multi-head Self-relation Network for Scene Text Recognition

Junwei Zhou, Hongchao Gao, Jiao Dai, Dongqin Liu, Jizhong Han
Institute of Information Engineering
Chinese Academy of Sciences, Beijing, China

Introduction

- The text embedded in scene images can be seen everywhere in our lives. However, recognizing text from natural scene images is still a challenge because of its diverse shapes and distorted patterns.
- The scene text recognition aims to retrieve all text strings from scene text images. It is a high-level and complicated task that translate between two different forms of information: Computer vision and Natural Language Processing.
- Extracting a meaningful text representation has become a challenge due to the complex environment in the irregular scene text recognition task such as uneven illumination, positional changes and so on.

Objectives

- A multi-head self-relation layer is proposed in this paper, which can be used to capture the relation between visual feature map cells. And the multi-head self-relation layer can learn 2-D attention in spatial.
- The MSRN is trained under the end-to-end training mode without any additional pixel-level or character-level supervision information, it means that the multi-head self-relation layer is weakly supervised by the cross entropy loss on the final predictions.
- The proposed multi-head self-relation layer in this paper will not change the size of the feature map. if the multi-head self-relation layer is removed, the model can still work without any changes, so that it can be embedded in other models..

Method

The overall framework consists of two main components: MSRN(encoder) and multi-head attention(decoder).

- Firstly, the input includes a text image and its ground-truth sequence. Then the deep visual representations of the text image can be extracted by using the proposed Multi-head Self-relation Network in this paper.
- Secondly, the output of the last three multi-head self-relation layers is taken as the input of the decoder, resulting in three text representations.
- Finally, the text representations are concatenated and used to predict the result sequence.

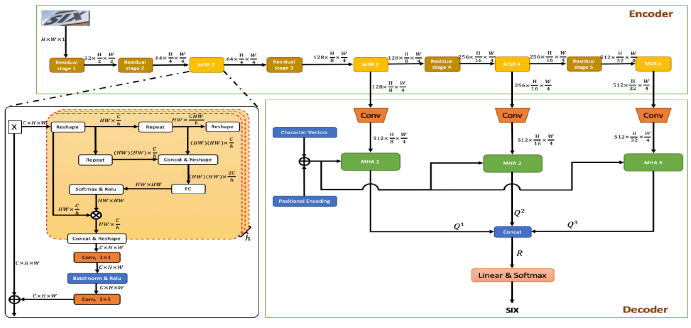

Fig 1: Overall Framework

Algorithm: Multi-head Self-relation Layer

Input: Visual feature vectors generated by a convolutional layer

1: Transform the visual feature vectors $X = \{x_1, x_2, \cdots\cdots, x_i, \cdots\cdots, x_n\}$ into higher-level features $E = \{e_1, e_2, \cdots\cdots, e_i, \cdots\cdots, e_n\}$ by a learnable weight W.
2: Get the relation coefficient between each vector.
3: **for** head = 1: head numbers **do**:
4: **for** i = 1: n **do** :
5: $z_{ij} = gConcat(e_i, e_j)$
6: $\alpha_{ij} = \frac{exp(LeakyReLU(z_{ij}))}{\sum_{m=1}^{n} exp(LeakyReLU(z_{im}))}$
7: $h_i = \sum_{j=1}^{n} \alpha_{ij} e_j$
8: **end for**
9: **end for**
10: Then all the heads' features are concatenated resulting in $t_i = Concat(h_i^1, \cdots\cdots, h_i^{nh})$. #nh means head numbers
11 : **Output**: $y_i = x_i + W1ReLU(BN(W2t_i))$

Results

TABLE I
UNDER THE DIFFERENT NUMBERS OF SELF-RELATION HEADS OUR NETWORK RECOGNITION ACCURACY, THE TRAINING DATASET IS SYSTH90K. THE MSRN CONTAINS MSR2, MSR3, MSR4, AND MSR5. THE DECODER INCLUDES MHA1, MHA2, MHA3.

Head number	IIIT5K	SVT	IC03	IC13	CUTE80	IC15	SVTP
1	79.4	78.5	88.5	89.2	56.9	63.8	67.8
4	79.2	81.6	92.2	90.8	**64.2**	67.0	72.8
8	**82.1**	**82.4**	**92.8**	**91.6**	63.9	**68.6**	**73.3**
16	79.4	81.3	88.4	88.3	56.6	64.9	71.1

TABLE II
COMPARISON OF DIFFERENT MSRN SETTINGS. √ IN THE COLUMN NAMED MSRi MEANS THE NETWORK CONTAINS THE i-TH MSR LAYER, THE TRAINING DATASET IS SYSTH90K, THE HEAD NUMBER IS 8. THE DECODER INCLUDES MHA1, MHA2, MHA3.

MSR2	MSR3	MSR4	MSR5	IIIT5K	SVT	IC03	IC13	CUTE80	IC15	SVTP
			√	80.3	80.4	89.5	89.3	59.0	65.9	72.5
		√	√	**82.3**	**82.4**	90.9	89.9	63.5	66.4	72.0
√		√	√	81.5	80.1	**93.4**	91.5	**63.9**	67.0	73.0
	√	√	√	79.1	79.2	90.1	89.1	59.2	65.1	69.5
√	√	√	√	82.1	**82.4**	92.8	**91.6**	**63.9**	**68.6**	**73.3**

TABLE III
COMPARE THE PERFORMANCE UNDER DIFFERENT DECODER SETTINGS. √ IN THE COLUMN NAMED MHAi MEANS THE DECODER CONTAINS THE i-TH MHAi BLOCK. THE TRAINING DATASET IS SYSTH90K, THE MSRN CONTAINS MSR2, MSR3, MSR4 AND MSR5. THE HEAD NUMBER IS 8.

MHA1	MHA2	MHA3	IIIT5K	SVT	IC03	IC13	CUTE80	IC15	SVTP
		√	80.5	79.5	89.9	88.8	59.4	63.7	71.2
	√	√	80.1	77.9	88.8	88.0	59.0	62.9	65.7
√		√	81.8	81.5	90.1	90.5	64.2	67.1	72.3
√	√	√	**82.1**	**82.4**	**92.8**	**91.6**	**63.9**	**68.6**	**73.3**

TABLE IV
COMPARE WITH OTHER METHODS, ALL SCORES ARE IN LEXION-FREE MODE. 90K MEANS SYNTH90K DATASET ST MEANS SYNTHTEXT DATASET, 'STADD' MEANS SYNTHADD DATASET. ST+ MEANS INCLUDING CHARACTER-LEVEL ANNOTATIONS OR USING ADDITIONAL DATASETS.

Method	Convnet,Data	Regular datasets				Irregular datasets		
		IIIT5K	SVT	IC03	IC13	CUTE80	IC15	SVTP
Shi et al.	VGG,90K	78.2	80.8	89.4	86.7	-	-	-
*Shi et al.	VGG,90K	81.9	81.9	90.1	88.6	-	59.2	71.8
Lee et al.	VGG,90K	78.4	80.7	88.7	90.0	-	-	-
Jaderberg et al.	VGG,90K	-	80.7	93.1	90.8	-	-	-
Wang et al.	-,90k	80.8	81.5	91.2	-	-	-	-
Cheng et al.	ResNet, 90k+ST+	87.4	85.9	94.2	93.3	-	70.6	-
Cheng et al.	-, 90k+ST+	87.0	82.8	91.5	-	76.8	68.2	73.0
Shi et al.	ResNet,90k+ST+	**93.4**	**93.6**	94.5	91.8	79.5	76.1	78.5
Luo et al.	-,90k+ST	91.2	88.3	95.0	92.4	77.4	68.8	76.1
Zhan et al.	ResNet,90k+ST	93.3	90.2	-	91.3	**83.3**	76.9	79.6
Li et al.	ResNet,90k+ST+	91.5	84.5	-	91.0	**83.3**	69.2	76.4
MSRN(ours)	ResNet,90k+ST+	91.9	89.7	**95.1**	**95.2**	79.5	**78.1**	**81.8**

- Table I shows that when the head number is 8, the recognition network gets the best performance.
- Table II shows that when we use MSR2, MSR3, MSR4, and MSR5, the performance is best.
- As shown in Table III, when the decoder includes MHA1, MHA2, and MHA3, our model gets the best recognition accuracy.
- From Table IV we can also learn that the performance of our method improves more on irregular datasets than that on regular datasets

Conclusion

- In this paper, we propose a novel multi-head self-relation network, which can extract the relationship between each feature map cell.
- In our recognition network, a feature map is treated as a graph and each cell is treated as a node of the graph, then a correlation matrix is learnt to guide the nodes states updating.
- The performance of our recognition network shows that the MSRN is effective.

Email: daijiao@iie.ac.cn