ReADS: A Rectified Attentional Double Supervised Network for Scene Text Recognition

Qi Song, Qianyi Jiang, Nan Li, Rui Zhang, and Xiaolin Wei
Meituan, Beijing, China

INTRODUCTION

- Both CTC and Attentional sequence recognition (Attn) have inherent drawbacks for scene text recognition.
- We propose a novel double supervised network which predicts text from both image inherent texture and semantic context by CTC and Attn.
- Our proposed method achieves the state-of-the-art performance on both regular and irregular scene text benchmarks.

METHOD

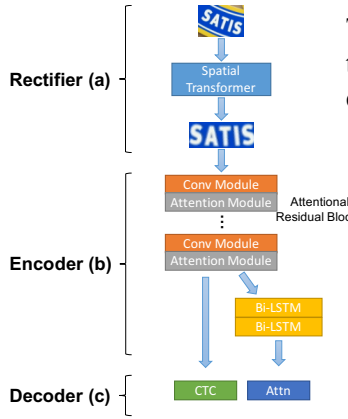

Fig1 The architecture of the ReADS

The ReADS is composed of three parts, the **rectifier**, the **encoder**, and the **decoder**.

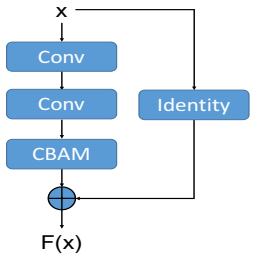

Fig2 The attentional residual block

- **Rectifier:** we adopt an STN with a predicted TPS. The STN consists of three parts which are the localization network, the grid generator and the sampler.
- **Encoder:** we introduce attention mechanisms(CBAM) into the encoder and elaborate two branches to extract features. The first directly passes the visual features to the decoder and the second processes the visual features by stacked Bi-LSTMs before feeding it into the decoder.

$$M_c(F) = sigmoid(Conv(F^c_{avg}) + Conv(F^c_{max})).$$

$$M_a(F) = sigmoid(f^{3\times3}([F^s_{avg}; F^s_{max}])).$$

- **Decoder:** We adopt two kinds of techniques in the decoding phase, namely CTC and Attn, to take both advantages of them. The CTC is responsible for recognition using inherent texture features, While Attn mainly focuses on semantic context modeling. The loss is calculated as follows,

$$L_{total} = L_{Attn} + \lambda L_{CTC},$$

- Inference: During inference, we choose the character sequence from Attn as the final result because the accuracy of the Attn branch is always better than that of CTC branch under any experimental conditions.

ABLATION STUDIES

We conduct two sets of experiments for ablation studies. The first is to analyze the impact of some modules in the network. The second is to verify the effectiveness of double supervised branches.

Modules		Regular Text				Irregular Text			
Rectifier	Attentions	IIIT5K	SVT	IC03	IC13	IC15-2077	IC15-1811	SVTP	CUTE
		89.4	87.6	94.8	93.1	70.4	75.0	76.7	80.2
	✓	90.0	90.0	95.3	93.4	74.3	79.2	80.3	77.4
✓		90.1	90.3	94.6	92.3	72.8	78.4	81.2	**83.7**
✓	✓	**91.0**	**91.2**	**96.1**	**94.5**	**75.1**	**80.4**	**83.3**	**83.7**

Table1 Results of using different modules

Branches		Regular Text				Irregular Text			
Attn	CTC	IIIT5K	SVT	IC03	IC13	IC15-2077	IC15-1811	SVTP	CUTE
	✓	88.6	87.3	92.4	90.3	72.1	76.5	77.1	78.8
✓		**91.0**	90.6	94.3	93.3	**75.7**	80.2	**84.2**	82.3
✓	✓	**91.0**	**91.2**	**96.1**	**94.5**	75.1	**80.4**	83.3	**83.7**

Table2 Results of using different supervised branches

RESULTS

Method	Regular Text				Irregular Text			
	IIIT5K	SVT	IC03	IC13	IC15-2077	IC15-1811	SVTP	CUTE
Jaderberg *et al.* 2014 [4]	-	80.7	93.1	90.8	-	-	-	-
Shi *et al.* 2016 [1]	78.2	80.8	89.4	86.7	-	-	-	-
Shi *et al.* 2016 [16]	81.9	81.9	90.1	88.6	-	-	71.8	59.2
Liu *et al.* 2016 [2]	83.3	83.6	89.9	89.1	-	-	73.5	-
Gao *et al.* 2017 [3]	81.8	82.7	89.2	88.0	-	-	-	-
Cheng *et al.* 2018 [21]	87.0	82.8	91.5	-	68.2	-	73.0	76.8
Liu *et al.* 2018 [30]	83.6	84.4	91.5	90.8	60.0	-	73.5	-
Shi *et al.* 2019 [17]	93.4	**93.6**	94.5	91.8	-	76.1	78.5	79.5
Liao *et al.* 2019 [31]	92.0	82.1	-	91.4	-	-	-	78.1
Zhan & Lu *et al.* 2019 [32]	93.3	90.2	-	91.3	-	76.9	79.6	83.3
Luo *et al.* 2019 [18]	91.2	88.3	95.0	92.4	68.8	-	76.1	77.4
Gao *et al.* 2019 [33]	89.9	87.2	93.3	92.9	74.5	-	76.4	70.8
Baek *et al.* 2019 [34]	87.9	87.5	94.4	92.3	71.8	77.6	79.2	74.0
Liu *et al.* 2019 [35]	85.2	85.5	92.9	90.3	65.7	71.8	74.4	-
Wan *et al.* 2020 [36]	**94.7**	90.6	-	93.9	-	75.2	79.2	81.3
Wang *et al.* 2020 [37]	90.5	82.2	-	-	-	-	-	83.3
Ours	91.0	91.2	**96.1**	**94.5**	**75.1**	**80.4**	**83.3**	**83.7**

Table3 Results of our model compared with other proposed models.

- Our method gets five first, one second and one competitive results on a total of seven benchmarks.
- Our double branch supervised network outperforms any single branch supervised versions on most benchmarks. Some qualitative cases are illustrated on the left.

CONCLUSION

- we present a novel approach named ReADS, a rectified attentional double supervised scene text recognizer.
- It is equipped with Attn and CTC decoders for semantic context and visual texture modeling.
- Future works: Merging predictions from CTC and Attn branches. Combining scene text recognition with NLP techniques.