Supplementary Material – RobustScanner: Dynamically Enhancing Positional Clues for Robust Text Recognition

Anonymous ECCV submission

Paper ID 3160

In order to supplement our main submission, we provide additional materials in this document.

Ablation experiments of RandText benchmark. In this paper, we have synthesized the RandText benchmark to demonstrate the poor performance of the encoder-decoder with attention framework on contextless text sequence images. To eliminate the influence of font, background and color of generated RandText benchmark on the recognition results, we apply the same method to additionally synthesize a contextual benchmark of all text sequences of length 5 of 6 academical datasets(*i.e.* IIIT 5K-words [4], Street View Text [7], ICDAR 2013 [2], ICDAR 2015 [1], Street View Text Perspective [5], CUTE 80 [6]). Some examples can be found in Figure 1. The performance of RobustScanner and other representative encoder-decoder with attention based approaches are shown in Table 1. It can clearly be seen that the synthetic contextual text images can be accurately recognized by all these approaches which indicates that the poor performance of the encode-decode with attention framework on RandText is caused by the lack of contextual information.

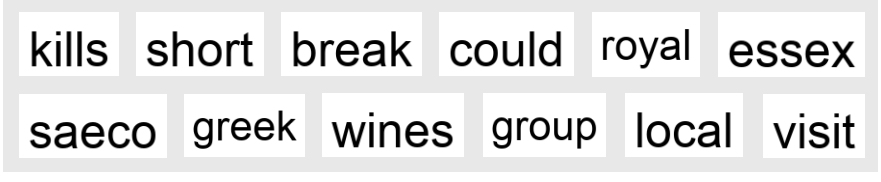


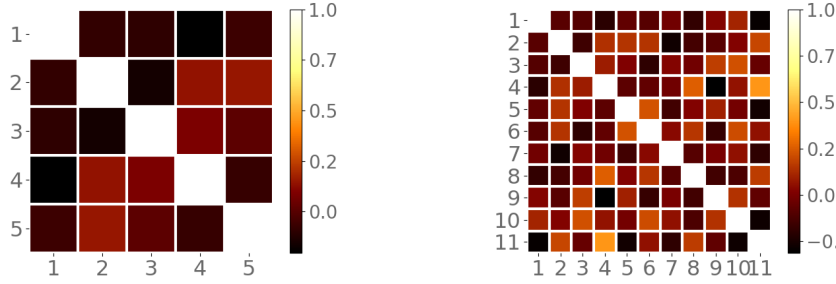
Fig. 1. Samples of the synthetic contextual text images.

Method	SAR [3]	DAN [9]	Wang *et al* [8]	RobustScanner
Accuracy	98.9	**1.0**	**1.0**	99.8

Table 1. Comparison on the synthetic benchmark.

The position encoding capability of the position enhancement branch. In order to demonstrate the effectiveness of our position enhancement branch in

terms of position encoding capability, we compute the averaged cosine similarity between the query feature vectors predicted by the position embedding layer as done in Section 3.2. The visualization of the averaged similarity matrices is shown in Figure 2. It has been shown that the query vectors of the same position are identical and can be distinguished from those of other positions. In light of the observation, we can conclude that the proposed position enhancement branch can effectively encode the positional information.



(a) $l = 5$ (b) $l = 11$

Fig. 2. Visualization of the averaged similarity matrix. The x-axis and the y-axis indicate the position index in sequences, while the color indicate the averaged similarity.

References

1. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Others: ICDAR 2015 Competition on Robust Reading. In: ICDAR. pp. 1156–1160. IEEE (2015)
2. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L.G.I., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. ICDAR pp. 1484–1493 (2013)
3. Li, H., Wang, P., Shen, C., Zhang, G.: Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. AAAI (2019)
4. Mishra, A., Alahari, K., Jawahar, C.V.: Scene text recognition using higher order language priors. In: BMVC-British Machine Vision Conference. BMVA (2012)
5. Quy Phan, T., Shivakumara, P., Tian, S., Lim Tan, C.: Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 569–576 (2013)
6. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Systems with Applications **41**(18), 8027–8048 (2014)
7. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision. pp. 1457–1464. IEEE (2011)
8. Wang, P., Yang, L., Li, H., Deng, Y., Shen, C., Zhang, Y.: A Simple and Robust Convolutional-Attention Network for Irregular Text Recognition. In: aXiv preprint (2019)
9. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: AAAI Conference on Artificial Intelligence (2020)