Scene Text Recognition via a Modified CRNN Model

Abstract

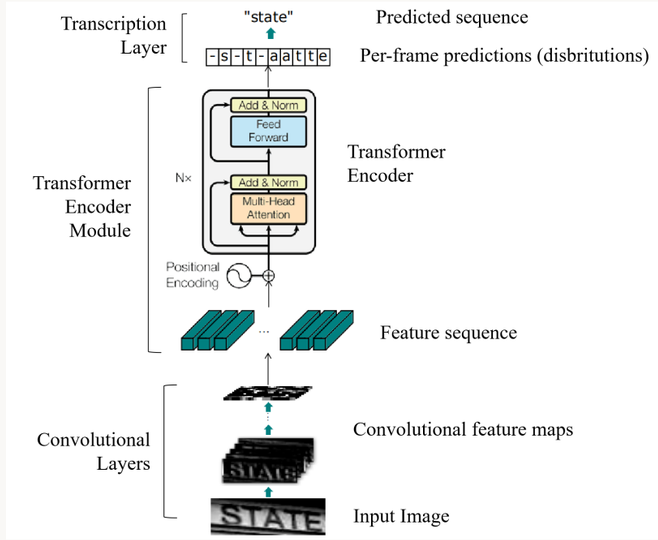
- Words in natural images often possess irregular shapes, which are caused by perspective distortion, curved character placement, etc.
- We replace the original bidirectional LSTM with a Transformer encoder to enhance the ability of modeling the long-term dependencies of texts.
- Our optimization of loss function and pre-training process further improve the recognition performance of the original CRNN model.

Motivation

- Huge challenges: complicated background, unbalanced illumination, arbitrary lengths, occluded text, and character segmentation.
- Methods: RNN-based, attention-based, language model, semantic information, multi-object network, end-to-end architecture.
- CRNN, an classic and popular approach for image-based sequence recognition, needs to speed up and be more practical.

Method

- Text recognition network



- Optimization of loss function
  - Connectionist Temporal Classification (CTC)
  - Weight normalization
  - Angle loss function
- Pre-training process
  - Data augmentation
  - "CNN+Transformer"+MLP
  - Soft-DTW distance and stop gradient operation

Conclusion

- The transformer encoder module obtains more context information and gains robustness, and the training process of the proposed method is highly parallel and efficient.
- With our optimization of loss function and the implementation pre-training, the text recognition network can further significantly increase recognition accuracy on scene text.
- The proposed scene text recognition system is competitive compared with the traditional "CRNN+CTC" methods.

Reference

[1] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, 1 Nov. 2017, doi: 10.1109/TPAMI.2016.26463.

[2] A. Graves, "Supervised sequence labelling", in Supervised sequence labelling with recurrent neural networks, Springer, 2012, p.

[3] X. Chen and K. He, "Exploring simple siamese representation learning", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–1.

Results

Method	IIIT5K	SVT	IC03	IC13
CRNN	78.2	80.8	89.4	86.7
CRNN+Transformer	82.90	81.72	90.66	88.96
CRNN+Transformer+Weight Normalization	**83.57**	82.69	91.23	89.06
Our method (CRNN+Transformer+Weight Normalization+Pre-training)	83.40	**83.15**	**91.69**	**90.34**

(a) Results of Our Proposed Network

Method	IIIT5K	SVT	IC03	IC13
EnEsCTC	82.0	80.6	92.0	**90.6**
ACE(1D, Cross Entropy)	82.3	82.6	**92.1**	89.7
Reinterpreting CTC	81.1	82.2	91.2	87.7
Our method	**83.40**	**83.15**	91.69	90.34

(b) Comparison with Mainstream Networks

Zhang Boyang, Cao Shiji

suikakon@163.com

Tsinghua-Berkerly Shenzhen Institute, Tsinghua University, China