CIGIT & XJTLU submission to ICDAR 2019-ArT Task 2.1

Zhaohong Guo[1] Hui Xu[1] Qiufeng Wang[2] Xiangdong Zhou[1] Yu Shi[1]
[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[2]Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University
xuhui@cigit.ac.cn

Abstract

This report introduces our submission to ICDAR2019-ArT Task 2.1. Our approach utilize a Spatial Transformer Network (STN) with Thin-Plate Spline (TPS) transformation to rectify the original text, and then employ an Attention-based Sequence Recognition Network to recognize the rectified text.

1. Introduction

Our solution follows the strategy of Spatial Transformer Network (STN) and Attention-based Sequence Recognition Network [2].

2. Dataset

2.1. Training datasets

We use the public datasets Synth90k and SynthText to train our models. We crop the words with annotated bounding boxes in SynthText and obtain 7 million word images, and Synth90k contains 9 million word images.

2.2. Fine-tuning datasets

IIIT5K, SVT, SVT-P, CUTE80, COCO-Text, ICDAR2013, ICDAR2015, ICDAR2019-MLT and 10k Latin text line from ICDAR2019-ArT are used to fine-tune the model. ICDAR2019-MLT contains 60k Latin word images. Note that, the image which has the aspect ratio of 3:1 or larger is removed.

2.3. Validation datasets

Excluding 10k Latin text line used as fine-tuning dataset. The validation datasets consist of 25k Latin text line from ICDAR2019-ArT.

3. Approach

3.1. Spatial Transformer Network (STN)

The STN follows the framework proposed in [1]. In our settings, the localization network consists of 4 convolutional layers, 3 pooling layers, a global average pooling layer, and 2 fully-connected layers.

3.2. Attention-based Sequence Recognition Network (ASRN)

The major structure of the ASRN is a CNN-BLSTM framework. We adopt a one-dimensional attention mechanism at the top of CRNN. A 45-layer residual network is adopted as the convolutional feature extractor. Following the residual network are two layers of Bidirectional LSTM (BiLSTM). To generate the character sequence, we employ a bidirectional decoder which consists of two decoders(attentional LSTMs) with opposite directions.

3.3. Implementation details

Training The model is trained from scratch. We adopt ADADELTA as the optimizer. The model is trained by batches of 256 examples for 400,000 iterations over Training datasets and 7,000 iterations over Fine-tuning datasets. In training stage, the learning rate is set to 1.0 initially and decayed to 0.1 at step 320,000. In fine-tuning stage, we randomly rotate the image with an angle in the range of $[-15°, 15°]$ and apply Piecewise Affine Transformation with annotated polygons to image in ICDAR2019-ArT dataset, the learning rate is set to 0.01.
Testing For image which has the aspect ratio of 2:1 or larger, we rotate it ±90 degrees. The highest-scored recognition result will be chosen as the final output.
All the input images are resized to 32×100. The model that generates the submitted results is fine-tuned by batches of 256 examples for 4,500 iterations over all the fine-tuning datasets and validation datasets, and the learning rate is set to 0.5.

References

[1] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 1

[2] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai. Robust scene text recognition with automatic rectification. In *CVPR*, pages 4168–4176, 2016. 1