Scene Text Recognition with Automatic Rectification

Hongzhi Liu

Jun 12, 2018

Abstract

In natural scenes, text appears on various kinds of objects, e.g. road signs, billboards and product packaging. It carries rich and high-level semantic information that is important for image understanding. Recognizing text in images facilitates many real-world applications, such as geolocation, driverless car and image-based machine translation. For these reasons, scene text recognition has attracted great interest from the community. Today, I read a thesis written by Baoguang Shi, who is in the School of Electronic Information and Communications from Huazhong University of Science and Technology. His team propose a Robust text recognizer with Automatic REctification, namely RARE, a recognition model that is robust to irregular text. State-of-the-art or highly-competitive performance achieved on several benchmarks well demonstrates the effectiveness of the proposed model.

1. Overview of Recognition Model

In recent years, a rich body of literature concerning scene text recognition has been published. Among the traditional methods, many adopt bottom-up approaches, where individual characters are firstly detected using sliding window, connected components or Hough voting [12]. Following that, the detected characters are integrated into words by means of dynamic programming. Other work adopts top-down approaches, where text is directly recognized from entire input images, rather than detecting and recognizing individual characters. Some recent work models the problem as a sequence recognition problem, where text is represented by character sequence. Su and Lu [9] extract sequential image representation, which is a sequence of HOG [2] descriptors, and predict the corresponding character sequence with a recurrent neural network (RNN). Shi's method also adopts the sequence prediction scheme but they further take the problem of irregular text into account.

Usually, a text recognizer works best when its input images contain tightly-bounded regular text. This motivates us to apply a spatial transformation prior to recognition, in order to rectify input images into ones that are more "readable" by recognizers. In this paper, the team propose a recognition method that is robust to irregular text [8]. Specifically, they construct a deep neural network that combines a Spatial Transformer Network (STN) [3] and a Sequence Recognition Network (SRN). An overview of the model is given in Fig. 1.

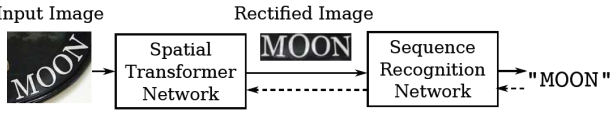

Figure 1. Schematic overview of RARE, which consists a spatial transformer network (STN) and a sequence recognition network (SRN). The STN transforms an input image to a rectified image, while the SRN recognizes text. The two networks are jointly trained by the back-propagation algorithm [4]. The dashed lines represent the flows of the back-propagated gradients.

2. Model of RARE

RARE is a speciallydesigned deep neural network, which consists of a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN). Overall, the model Shi formulates takes an input image I and outputs a sequence $l = (l_1, \ldots, l_T)$, where l_t is the t-th character, T is the variable string length.

2.1. Spatial Transformer Network

The STN transforms an input image I to a rectified image I' with a predicted TPS transformation. It follows the framework proposed in [3]. As illustrated in Fig. 2, it first predicts a set of fiducial points via its localization network. Then, inside the grid generator, it calculates the TPS transformation parameters from the fiducial points, and generates a sampling grid on I. The sampler takes both the grid and the input image, it produces a rectified image I' by sampling on the grid points.

A distinctive property of STN is that its sampler is differentiable.Therefore, once they have a differentiable localization network and a differentiable grid generator, the STN

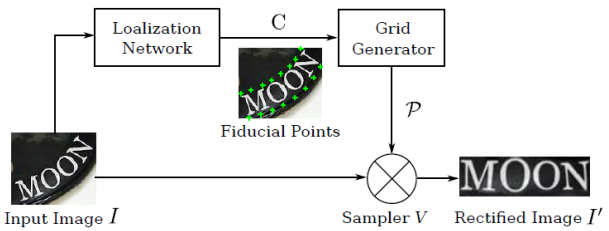

Figure 2. Schematic overview of RARE, which consists a spatial transformer network (STN) and a sequence recognition network (SRN). The STN transforms an input image to a rectified image, while the SRN recognizes text. The two networks are jointly trained by the back-propagation algorithm [4]. The dashed lines represent the flows of the back-propagated gradients.

can back-propagate error differentials and gets trained.

2.2. Sequence Recognition Network

Since target words are inherently sequences of characters, the team model the recognition problem as a sequence recognition problem, and address it with a sequence recognition network. The input to the SRN is a rectified image I', which ideally contains a word that is written horizontally from left to right. They extract a sequential representation from I', and recognize a word from it.

In their model, the SRN is an attention-based model [1], which directly recognizes a sequence from an input image. The SRN consists of an encoder and a decoder. The encoder extracts a sequential representation from the input image I'. The decoder recurrently generates a sequence conditioned on the sequential representation, by decoding the relevant contents it attends to at each step.

An easy approach for extracting a sequential representation for I' is to take local image patches from left to right, and describe each of them with a CNN. However, this approach does not share the computation among overlapping patches, thus inefficient. Besides, the spatial dependencies between the patches are not exploited and leveraged. Instead, following [7], they build a network that combines convolutional layers and recurrent networks. The network extracts a sequence of feature vectors, given an input image of arbitrary size.

As illustrated in Fig. 3, at the bottom of the encoder is several convolutional layers. They produce feature maps that are robust and high-level descriptions of an input image. The output sequence is $h = (h_1, \ldots, h_L)$, where $L = W_{conv}$.

3. Evaluation of the Model

Shi evaluates their model on a number of standard scene text recognition benchmarks, paying special attention to

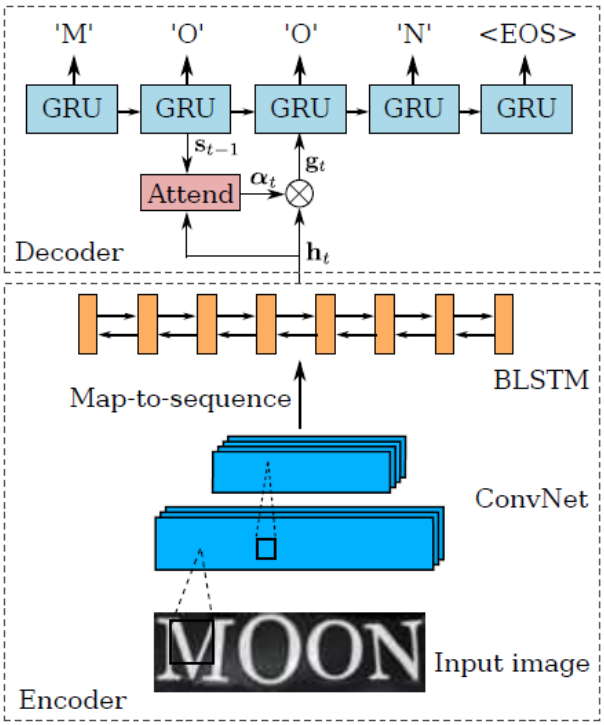

Figure 3. Structure of the SRN, which consists of an encoder and a decoder. The encoder uses several convolution layers (ConvNet) and a two-layer BLSTM network to extract a sequential representation (h) for the input image. The decoder generates a character sequence (including the EOS token) conditioned on h.

recognition performance on irregular text. First they evaluate the model on some general recognition benchmarks, which mainly consist of regular text, but irregular text also exists. Next, they perform evaluations on benchmarks that are specially designed for irregular text recognition. For all benchmarks, performance is measured by word accuracy.

To validate the effectiveness of the rectification scheme, they evaluate RARE on the task of perspective text recognition. SVT-Perspective [6] is specifically designed for evaluating performance of perspective text recognition algorithms. Text samples in SVT-Perspective are picked from side view angles in Google Street View, thus most of them are heavily deformed by perspective distortion. SVT-Perspective consists of 639 cropped images for testing. Each image is associated with a 50-word lexicon, which is inherited from the SVT [10] dataset. The team use the same model trained on the synthetic dataset without fine-tuning. For comparison, they test the CRNN model [7] on SVT-Perspective.

Tab. 1. summarizes the results. In the second and third columns, the team compare the accuracies of recognition with the 50-word lexicon and the full lexicon. Their method outperforms, which is a perspective text recognition

Table 1. Recognition accuracies on SVT-Perspective [6]. Recognition accuracies on SVT-Perspective."50" and "Full" represent recognition with 50-word lexicons and the full lexicon respectively. "None" represents recognition without a lexicon.

Method	50	Full	None
Wang et al. [10]	40.5	26.1	-
Mishra et al. [5]	45.7	24.7	-
Wang et al. [11]	40.2	32.4	-
Phan et al. [6]	75.6	67.0	-
Shi et al. [7]	92.6	72.6	66.8
RARE	91.2	77.4	71.8

method, by a large margin on both lexicons. However, this gap is partially due to that we use a much larger training set than [6]. In the comparisons with [7], which uses the same training set as RARE, they still observe significant improvements in both the Full lexicon and the lexicon-free settings.

Furthermore, RARE outperforms by a even larger margin on SVTPerspective. The reason is that the SVT-perspective dataset mainly consists of perspective text, which is inappropriate for direct recognition. Their rectification scheme can significantly alleviate this problem.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 2

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[3] M. Jaderberg, K. Simonyan, and A. Zisserman. Spatial transformer networks. In *NIPS*, 2015. 1

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 1, 2

[5] A. Mishra, K. Alahari, and C. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012. 3

[6] T. Quy Phan, P. Shivakumara, S. Tian, and C. Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, 2013. 2, 3

[7] 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. *IEEE TPAMI*, 2017. 2, 3

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

[9] B. Su and S. Lu. Accurate scene text recognition based on recurrent neural network. In *ACCV*, 2014. 1

[10] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *ICCV*, 2011. 2, 3

[11] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *ICPR*, 2012. 3

[12] C. Yao, X. Bai, B. Shi, and W. Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *CVPR*, 2014. 1