Mohit Jain
B.Tech. and M.S. in Computer Science
International Institute of Information Technology Hyderabad, India
mohit.jain@research.iiit.ac.in || +91-7842662536

SceneText Recognition

Recognizing text from natural scene images

OVERVIEW

Reading text in natural scenes is relatively a harder task compared to printed or handwritten text recognition. The problem has been drawing increasing research interest in recent years. This can partially be attributed to the rapid development of wearable and mobile devices such as smart phones, digital cameras, and the latest tech like google-glass and self-driving cars, where scene text is a key module to a wide range of practical and useful applications. While the recognition of text in printed documents has been studied extensively in the form of Optical Character Recognition (OCR) Systems, these methods generally don't generalize very well to a natural scene setting where factors like inconsistent lighting conditions, variable fonts, orientations, background noise and image distortions add to the problem complexity.

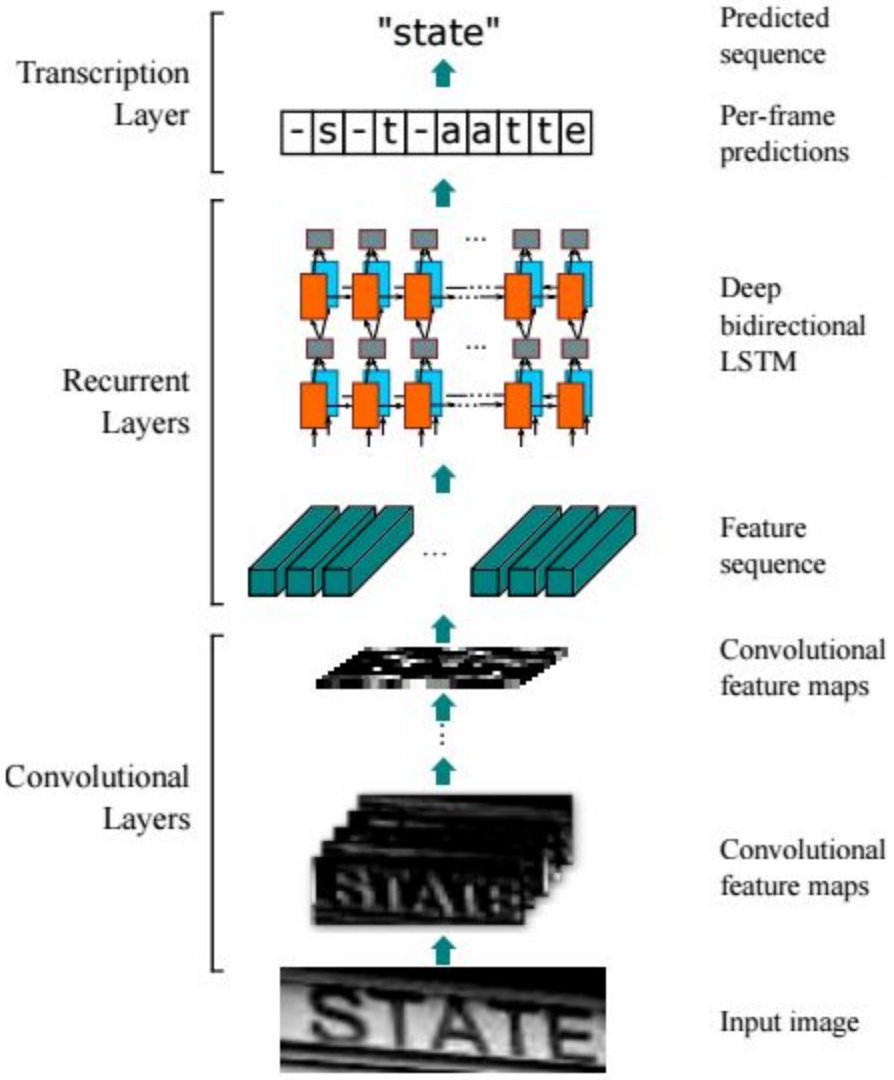
GOALS

1. End-to-end trainable, in contrast to most of the existing algorithms whose components are separately trained and tuned.
2. Handle sequences in arbitrary lengths, involving no character segmentation or horizontal scale normalization.
3. Not be confined to any predefined lexicon and make the solution unconstrained to language.
4. Generate an effective yet much smaller model, which is more practical for real-world application scenarios.

SOLUTION

The network architecture of proposed CNN-RNN hybrid solution, as shown in Figure below, consists of three components, including the convolutional layers, the recurrent layers, and a transcription layer, from bottom to top.

The convolutional layers automatically extract a feature sequence from each input image. On top of the convolutional network, a recurrent network is built for making prediction for each frame of the feature sequence, outputted by the convolutional layers. The transcription layer at the top of CRNN is adopted to translate the per-frame predictions by the recurrent layers into a label sequence. Though CRNN is composed of different kinds of network architectures (eg. CNN and RNN), it can be jointly trained with one loss function.



SPECIFICATIONS

The model is created using Torch7 deep learning framework. It takes the model 1 day/million-images to train while using a single 12GB Nvidia TitanX GPU with a 64GB RAM CPU machine. The images are loaded and stored to a LMDB dataset for faster and easier File I/O.

To train the network, synthetic images were created by masking texts of various fonts, sizes, colors, perspective distortions and noise with images of natural scenes. We used 6 million synthetic images to train the model and the performance was tested on standard scene text benchmarking datasets, IIIT-5k and SVT.

We obtain close to 80% word-level accuracy on both of these datasets which is the current state-of-the-art in this paradigm. We also extend this project to set the benchmark accuracies for 9 other Indian scripts namely, Hindi, Marathi, Gurmukhi, Telugu, Tamil, Kannada, Urdu, Arabic and Bengali.

This project is implemented using the works of Baoguang Shi et al. [CRNN] submitted in CVPR 2015.