[Review] An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition (2016)

Taekyung Ki
tkki@jenti.ai

February 25, 2021

Outline

Introduction

Contibution

- An end-to-end trainable model, named **CRNN**.
- Handing sequences of **arbitrary lengths**, involving no character segmentation or horizontal scale normalization.
- Achieving remarkable performances in **lexicon-free** and **lexicon-based** scene text recognition tasks.

Introduction

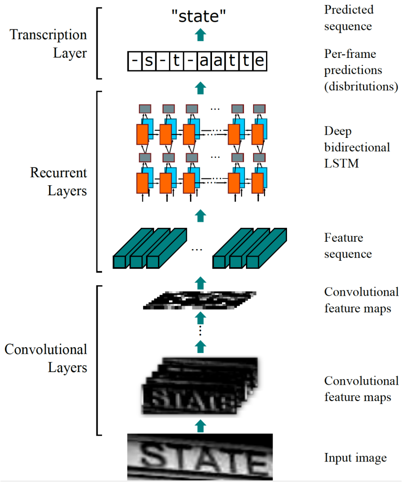


Figure: Overview of CRNN. There are mainly three steps: Convolutional layers, Recurrent layer, Transcription layer

Type	Configurations
Transcription	-
Bidirectional-LSTM	#hidden units:256
Bidirectional-LSTM	#hidden units:256
Map-to-Sequence	-
Convolution	#maps:512, k:2 $\times$ 2, s:1, p:0
MaxPooling	Window:1 $\times$ 2, s:2
BatchNormalization	-
Convolution	#maps:512, k:3 $\times$ 3, s:1, p:1
BatchNormalization	-
Convolution	#maps:512, k:3 $\times$ 3, s:1, p:1
MaxPooling	Window:1 $\times$ 2, s:2
Convolution	#maps:256, k:3 $\times$ 3, s:1, p:1
Convolution	#maps:256, k:3 $\times$ 3, s:1, p:1
MaxPooling	Window:2 $\times$ 2, s:2
Convolution	#maps:128, k:3 $\times$ 3, s:1, p:1
MaxPooling	Window:2 $\times$ 2, s:2
Convolution	#maps:64, k:3 $\times$ 3, s:1, p:1
Input	$W \times 32$ gray-scale image

Figure: Network configuration summary.

Sequential feature representations

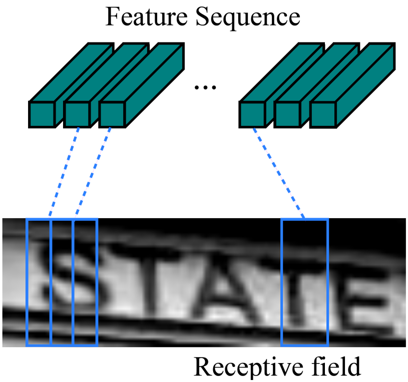


Figure: The receptive field. Each vector in the extracted feature sequence is associated with a receptive field on the input image, and can be considered as the feature vector of that field.

LSTM (Long-Short Term Memory)

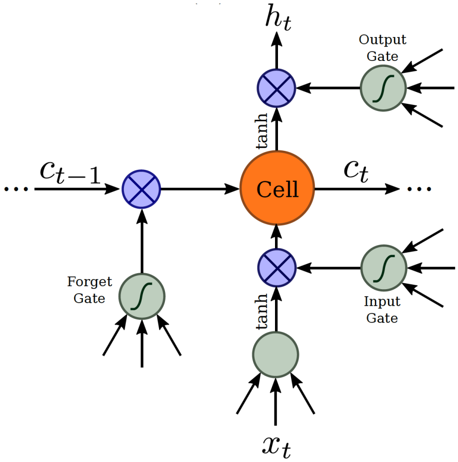


Figure: An illustration of LSTM (Long-Short Term Memory). It consists of a memory cell and three multiplicative gates, namely the input, output and forget gates.

BiLSTM

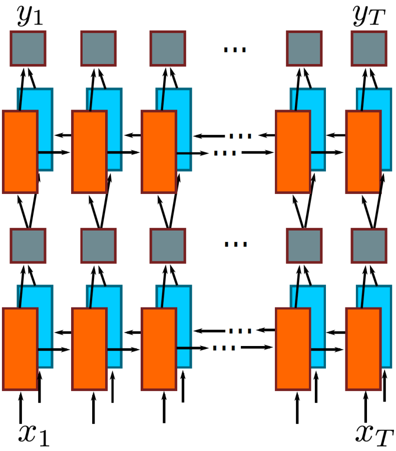


Figure: The structure of deep bidirectional LSTM in this paper. Stacking multiple bidirectional LSTM results is a deep BiLSTM.

Why RNN, especially... BiLSTM ?

1. A RNN has strong capability of capturing **contextual information** within a sequence.
2. A RNN can be back-propagates error differentials to its input.
3. A RNN is able to operate on sequence of arbitrary lengths, traversing from starts to ends.
4. But it suffers the problem of gradient vanishing problem.

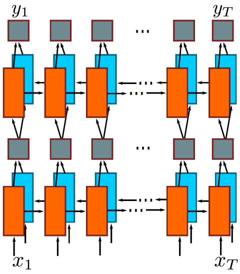
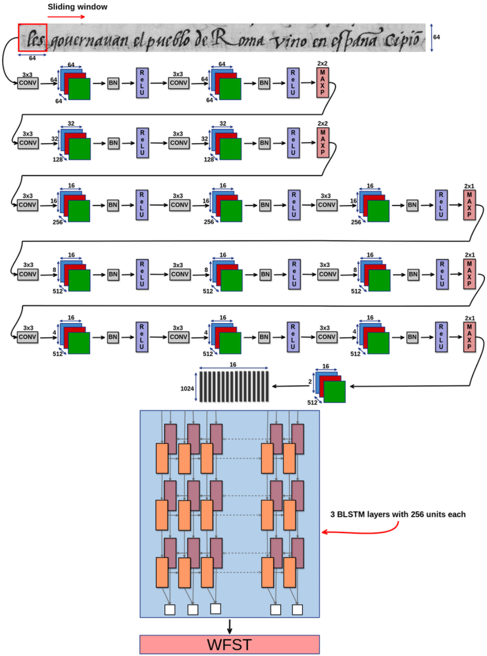


Figure: BiLSTM


Sliding window
3 BLSTM layers with 256 units each
WFST

Conditional Probability

Suppose $y = (y_1, y_2, \cdots, y_T)$ is an input of sequence length T. A sequence-to-sequence mapping function B maps π onto l by firstly removing the repeated labels, then removing the 'blanks'. Then the conditional probability is defined by

$$p(l|y) = \sum_{\pi:B(\pi)=l} p(\pi|y),$$

where the probability of π is defined as $p(\pi|y) = \prod_{t=1}^{T} y_{\pi_t}^t$, $y_{\pi_t}^t$ is the probability of having label π_t at time t.

- For example, B maps "–hh-e-l-ll-oo–" onto "hello" ("-" represents "blank").

Network Training

Denote the training dataset by $\chi = \{x_i, l_i\}$, where x_i is the training image and l_i is the ground truth label sequence. The objective is to minimize the negative log-likelihood of conditional probability of ground truth:

$$O := -\sum_{x_i, l_i} \log p(l_i | y_i),$$

where y_i is the sequence produced by the recurrent and convolutional layers from x_i.

Experimental Results

	E2E Train	Conv Ftrs	CharGT-Free	Unconstrained	Model Size
Wang *et al.* [34]	✗	✗	✗	✔	-
Mishra *et al.* [28]	✗	✗	✗	✗	-
Wang *et al.* [35]	✗	✔	✗	✔	-
Goel *et al.* [13]	✗	✗	✔	✗	-
Bissacco *et al.* [8]	✗	✗	✗	✔	-
Alsharif and Pineau [6]	✗	✔	✗	✔	-
Almazán *et al.* [5]	✗	✗	✔	✗	-
Yao *et al.* [36]	✗	✗	✗	✔	-
Rodrguez-Serrano *et al.* [30]	✗	✗	✔	✗	-
Jaderberg *et al.* [23]	✗	✔	✗	✔	-
Su and Lu [33]	✗	✗	✔	✔	-
Gordo [14]	✗	✗	✗	✗	-
Jaderberg *et al.* [22]	✔	✔	✔	✗	490M
Jaderberg *et al.* [21]	✔	✔	✔	✔	304M
CRNN	✔	✔	✔	✔	**8.3M**

Figure: Comparison among various methods.

Experimental Results

	IIIT5k			SVT		IC03				IC13
	50	1k	None	50	None	50	Full	50k	None	None
ABBYY [34]	24.3	-	-	35.0	-	56.0	55.0	-	-	-
Wang *et al.* [34]	-	-	-	57.0	-	76.0	62.0	-	-	-
Mishra *et al.* [28]	64.1	57.5	-	73.2	-	81.8	67.8	-	-	-
Wang *et al.* [35]	-	-	-	70.0	-	90.0	84.0	-	-	-
Goel *et al.* [13]	-	-	-	77.3	-	89.7	-	-	-	-
Bissacco *et al.* [8]	-	-	-	90.4	78.0	-	-	-	-	87.6
Alsharif and Pineau [6]	-	-	-	74.3	-	93.1	88.6	85.1	-	-
Almazán *et al.* [5]	91.2	82.1	-	89.2	-	-	-	-	-	-
Yao *et al.* [36]	80.2	69.3	-	75.9	-	88.5	80.3	-	-	-
Rodrguez-Serrano *et al.* [30]	76.1	57.4	-	70.0	-	-	-	-	-	-
Jaderberg *et al.* [23]	-	-	-	86.1	-	96.2	91.5	-	-	-
Su and Lu [33]	-	-	-	83.0	-	92.0	82.0	-	-	-
Gordo [14]	93.3	86.6	-	91.8	-	-	-	-	-	-
Jaderberg *et al.* [22]	97.1	92.7	-	95.4	80.7*	**98.7**	**98.6**	93.3	**93.1***	**90.8***
Jaderberg *et al.* [21]	95.5	89.6	-	93.2	71.7	97.8	97.0	93.4	89.6	81.8
CRNN	**97.6**	**94.4**	**78.2**	**96.4**	**80.8**	**98.7**	97.6	**95.5**	89.4	86.7

Figure: Recognition accuracy on four datasets. In the second row, 50, 1k, 50k, Full denote the lexicon used, and None denotes recognition without a lexicon.

Conclusions

1. CRNN that integrates the advantages of both CNN and RNN.
2. CRNN is able to take input images of varying dimensions.

References

[1] Baoguang Shi and Xiang Bai and Cong Yao, *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 11, 2298–2304, 2016

[2] Alex Graves and Santiago Fernández and Faustino Gomez and Jürgen Schmidhuber, *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks*, Proceedings of the 23rd International Conference on Machine Learning, 369–376, 2006