SCIENCE CHINA
Information Sciences

February 2020, Vol. 63 120103:1–120103:14
https://doi.org/10.1007/s11432-019-2713-1

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

FACLSTM: ConvLSTM with focused attention for scene text recognition

Qingqing WANG[1,2], Ye HUANG[2], Wenjing JIA[2], Xiangjian HE[2],
Michael BLUMENSTEIN[2], Shujing LYU[1] & Yue LU[1,3*]

[1] *Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai* 200241*, China;*
[2] *Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney* 2007*, Australia;*
[3] *Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai* 200092*, China*

Received 30 July 2019/Revised 8 October 2019/Accepted 12 November 2019/Published online 15 January 2020

Abstract Scene text recognition has recently been widely treated as a sequence-to-sequence prediction problem, where traditional fully-connected-LSTM (FC-LSTM) has played a critical role. Owing to the limitation of FC-LSTM, existing methods have to convert 2-D feature maps into 1-D sequential feature vectors, resulting in severe damages of the valuable spatial and structural information of text images. In this paper, we argue that scene text recognition is essentially a spatiotemporal prediction problem for its 2-D image inputs, and propose a convolution LSTM (ConvLSTM)-based scene text recognizer, namely, FACLSTM, i.e., focused attention ConvLSTM, where the spatial correlation of pixels is fully leveraged when performing sequential prediction with LSTM. Particularly, the attention mechanism is properly incorporated into an efficient ConvLSTM structure via the convolutional operations and additional character center masks are generated to help focus attention on right feature areas. The experimental results on benchmark datasets IIIT5K, SVT and CUTE demonstrate that our proposed FACLSTM performs competitively on the regular, low-resolution and noisy text images, and outperforms the state-of-the-art approaches on the curved text images with large margins.

Keywords scene text recognition, convolutional LSTM, focused attention, spatial correlation, sequential prediction

Citation Wang Q Q, Huang Y, Jia W J, et al. FACLSTM: ConvLSTM with focused attention for scene text recognition. Sci China Inf Sci, 2020, 63(2): 120103, https://doi.org/10.1007/s11432-019-2713-1

1 Introduction

Scene text recognition has received considerable attention from the community of computer vision since text is an essential way to convey information and knowledge. Owing to the challenges posed by poor image qualities (e.g., low resolution, blur, and uneven illumination) and various text appearances (e.g., size, fonts, colors, directions, perspective view as well as complex background), as shown in Figure 1, though many efforts have been made in past decades, scene text recognition is still an unsolved task.

Inspired by speech recognition and machine translation, most of recent state-of-the-art approaches regard scene text recognition as a sequence-to-sequence prediction problem and widely adopt techniques like LSTM [1] and attention mechanism [2, 3] in their sequential transcription module. However, the LSTM used in these recognizers is the fully-connected-LSTM (FC-LSTM) that only takes stream signals like sentences or audio as inputs and connects them in a fully connected way, while scene text recognition

[*] Corresponding author (email: ylu@cs.ecnu.edu.cn)

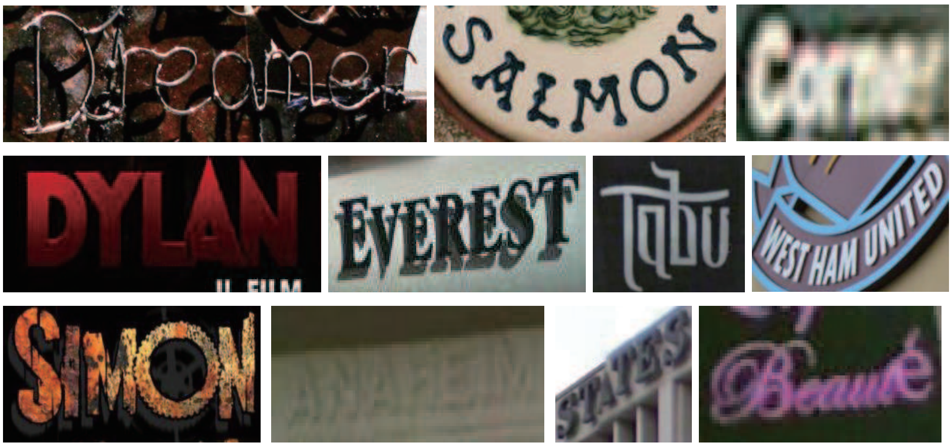

Figure 1 (Color online) Challenging samples of scene text recognition.

generates sequential outputs from 2-D images. To adapt FC-LSTM to scene text recognition, the most straightforward way is pooling 2-D feature maps to a height of one or flattening them into 1-D sequential feature vectors [4–8], as shown in Figure 2(a). Unfortunately, such operations could severely disrupt the valuable spatial correlation relationships among pixels, which is essential to computer vision tasks, especially to scene text recognition, where the structures of strokes are the key factors to discriminate characters. To retain such important spatial and structural information, researchers have also explored other alternative solutions. For example, STN-OCR [9] directly performed sequential prediction on 2-D feature maps with a fixed number of softmax classifiers; CA-FCN [10] generated character-level confidence maps with a fully convolutional network, as shown in Figure 2(b). However, compared with LSTM, these solutions often introduce additional parameters or post processing steps.

In this paper, we propose to address the issue of scene text recognition from the perspective of spatiotemporal prediction, where the spatial correlation information is taken into account when performing sequential prediction with LSTM. The convolution LSTM (ConvLSTM) proposed by Shi et al. [11] for precipitation nowcasting provides some insights on how to achieve this. In ConvLSTM, all of the fully connected operations are replaced by convolutional ones, so input feature maps are allowed to keep their 2-D shape when being fed into the ConvLSTM. Given this advantage, for the first time, we introduce ConvLSTM to scene text recognition and apply it in the sequential transcription module of our proposed recognizer.

However, in existing models, both FC-LSTM and ConvLSTM are used only for frame-level prediction and are incapable of producing sequential outputs from one single input image unless the connectionist temporal classification (CTC) [4, 7, 12] or attention mechanism [5, 6, 8, 13] is incorporated. To perform sequential prediction and, meanwhile, provide the model spatial awareness, we further improve ConvLSTM by embedding the attention mechanism into the structure. Notably, different from the existing attention-LSTM-based recognizers, where the attention mechanism and FC-LSTM are combined in a fully connected way, we properly integrate the attention mechanism into ConvLSTM with the convolutional operations. Moreover, as ConvLSTM extends 2-D operations into 3-D, the costs of computation and memory increase significantly. To achieve high efficiency, inspired by Liu et al. [14], we propose to assemble a bottleneck gate at the beginning of the proposed attention-equipped ConvLSTM, so that the internal feature map channels can be reduced.

Last but not the least, because existing attention-based recognizers often suffer from the 'attention drift' problem [5], i.e., they fail to align target outputs to proper feature areas, we propose to learn additional character center masks with a second decoder branch in the encoder-decoder feature extraction stage to assist the proposed network to focus attention on right feature areas. The experimental results conducted on benchmark datasets demonstrate that our proposed recognizer is able to achieve comparable performance with the state-of-the-art approaches on regular, low-resolution and noisy text and outperforms other methods significantly on the more challenging curved text images.

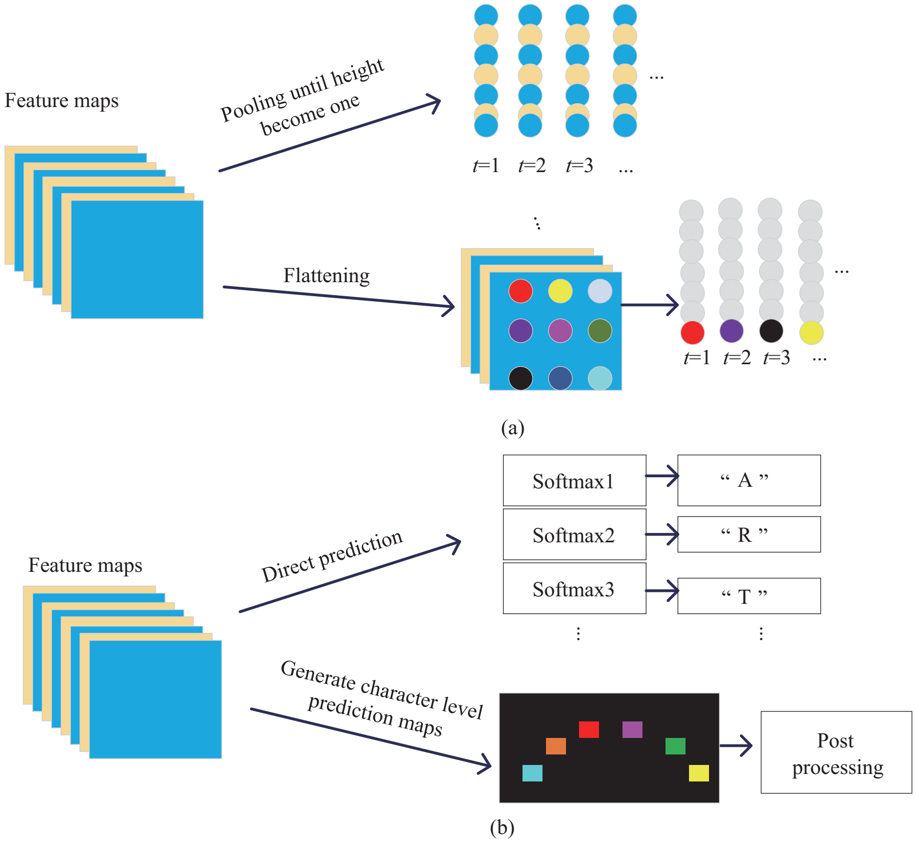

Figure 2 (Color online) Current solutions for scene text recognition. (a) Solutions with LSTM; (b) solutions without LSTM. When using LSTM, 2-D feature maps are usually converted to 1-D space by pooling or flattening operations. When the LSTM is not used, additional parameters or post precessing steps are involved.

The contributions made in this study are summarized as follows. (1) We propose to handle the scene text recognition problem from a spatiotemporal prediction perspective and for the first time introduce ConvLSTM to this application. (2) We design a ConvLSTM-based sequential transcription module, where the attention mechanism is harmoniously embedded into ConvLSTM with convolutional operations, and the bottleneck gate is assembled at the beginning of ConvLSTM to retain its efficiency. (3) We propose to learn additional character center masks to help the proposed network to focus attention on the center of characters.

In the rest of this paper, we first review the most related work in Section 2. Then, the details of our proposed approach and designed experiments are presented in Sections 3 and 4, respectively. Finally, the conclusion is given in Section 5.

2 Related work

The existing scene text recognizers can be grouped into two categories, i.e., the ones utilizing traditional techniques and the ones based on deep learning techniques. Methods belonging to the first category were mainly proposed before 2015, and follow a bottom-up routine, i.e., detecting and recognizing individual characters first, followed by word formation. Ye et al. [15] provided a comprehensive survey for these methods. By contrast, the deep learning-based recognizers depend on end-to-end trainable deep networks, where feature extraction and sequential translation are integrated into one unified framework. According to literature, the deep learning-based recognizers are now the dominant solutions to scene text recognition, and surpass traditional ones by large margins. Therefore, in this section, we only review recognizers applying deep learning techniques, along with ConvLSTM and related variants.

Methods based on LSTM. LSTM is widely used in the existing state-of-the-art recognizers for three purposes, i.e., producing frame-level predictions required by the subsequent sequential transcription module [4, 7], encoding sequential features with considering historical information [8, 16], and directly generating sequential predictions when cooperating with the attention mechanism [5, 6, 13, 16, 17]. For example, CRNN proposed by Shi et al. [7] was composed of three parts, i.e., the convolution module used to extract features from input images, a bi-LSTM layer built to make predictions for individual frames, and a CTC-based sequential transcription component utilized to infer sequential outputs from frame-level predictions. As clarified in [18], irregularly shaped art text presents frequently in our daily life, especially perspective text and curved text, which have posed enormous challenges for scene text recognition. To tackle this problem, Shi et al. [8] employed a bi-LSTM layer in their RARE to extract sequential feature vectors from input feature maps, followed by feeding these vectors into an attention-gated recurrent unit (GRU) module to generate label sequences. A highlight of RARE was its usage of spatial transformer network (STN) [19], which was responsible for rectifying images containing irregular texts and was widely adopted by subsequent recognizers like STN-OCR [9] and ASTER [16]. Afterwards, RARE was extended to ASTER [16] by modifying the architecture of rectification network. Note that, LSTM was used for both feature encoding and sequential transcription in ASTER. Lee et al. [17] combined a recursive CNN with a recurrent CNN in their R^2AM to capture long-term dependencies when extracting features from raw images, and then fed these features to an attention-RNN network for sequential transcription. Gao et al. [4, 12] designed two models to compare the performance of CNN and LSTM in terms of sequential feature encoding. According to their experiments, features extracted by LSTM were more powerful than those extracted by CNN. Cheng et al. [5, 6] combined LSTM with an attention mechanism in the sequential transcription module of their FAN and AON recognizers, but they criticized that the existing attention-based models often failed to align attention to right feature areas when performing prediction. Therefore, a focusing network was assembled in their FAN [5] to tackle this problem. AON [6] was specially designed for irregular text recognition. In this work, features were extracted from four directions, and then combined and filtered with a filter gate. Wojna et al. [13] utilized an attention-equipped LSTM to localize and recognize text from street view images. Their model was given location awareness by incorporating one-hot encoded spatial coordinates into the LSTM. Bai et al. [20] pointed out that exiting attention-based recognizers failed to align ground truth strings with attention's probability outputs, and this confused and misled the training process of the networks. To tackle this problem, they proposed edit probability (EP), which took the possible occurrences of missing and superfluous characters into consideration when estimating the probability of generating a string from the network's outputs. Su et al. [21, 22] converted text images into sequential signals via extracting their HOG features, and designed an ensembling technique to combine the outputs of two LSTM branches, so that better recognition performance could be achieved. Li et al. [23] pointed out that traditional attention mechanism was not able to produce accurate attention predictions, thus the recognition performance on irregular text images was largely compromised. To address this issue, they designed a 2-D attention module, where one LSTM was used to encode feature maps column by column to produce holistic features, and another was employed as usual to generate final sequential outputs. Note that, the LSTM used in all the methods mentioned above refers to traditional FC-LSTM, so the 2-D feature maps have to be mapped into 1-D space in order to adapt to the LSTM layers, and the attention mechanism has to be incorporated in a fully connected way. This severely damages the spatial and structural information of input images, which is essential to computer vision tasks such as scene text recognition.

Methods without LSTM. At the beginning of the deep learning era, a group of deep CNN (DCNN) recognizers [2, 24] were well developed and made breakthrough over traditional recognizers. For instance, Tian et al. [25] proposed two feature descriptors, i.e., co-occurrence HOG (Co-HOG) and convolutional Co-HOG, and combined them with CNN to perform scene text recognition on multiple languages. In these models, CNNs together with softmax classifiers were widely used for character or word classification. However, with the development of LSTM-based recognizers, the DCNN ones were quickly and significantly surpassed. Recently, some researchers argued that LSTM-based models were hard to train [12] and not able to achieve good performance on non-horizontal text [10], so explorations on models without LSTM

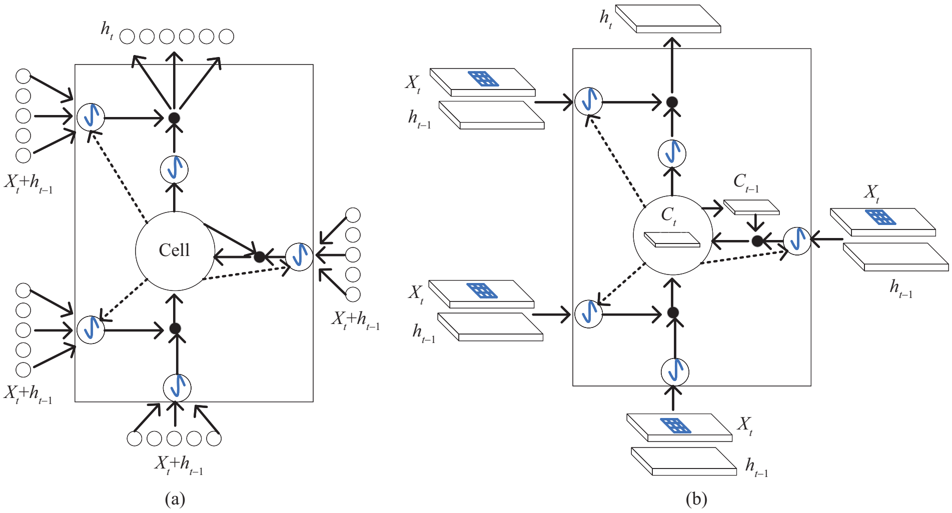

Figure 3 (Color online) Illustration of the FC-LSTM (a) and the ConvLSTM (b). The FC-LSTM is performed in 1-D space, while the ConvLSTM is performed in 2-D space.

started again. For instance, STN-OCR [9] utilized fully connected layers and a fixed number of softmax classifiers for sequential prediction; SqueezedText [26] employed a binary convolutional encoder-decoder network to generate salience maps for individual characters and then exploited a GRU-based bi-RNN for further correction; Liao et al. [10] proposed to address the scene text recognition issue from a 2-D perspective with a CA-FCN model, so that the spatial information could be taken into account when performing prediction. As proved in [27], the performance of object recognition has been largely fueled by the detection of salience regions. Therefore, in CA-FCN, a character attention module was utilized to produce pixel-level confidence maps for target characters, and then these maps were fed into a word formation module to generate word-level outputs.

Different from those LSTM-based approaches, recognizers without LSTM can better leverage the spatial information, but they also unavoidably introduce additional parameters or post processing steps in order to produce sequential outputs, such as the multiple classifiers used by STN-OCR [9] and the word formation module designed in [10].

ConvLSTM and related variants. As explained in [11], the main drawback of traditional FC-LSTM was its usage of full connections in the input-to-state and state-to-state transitions, which resulted in the neglect of spatial information. To retain such important information, ConvLSTM, proposed by Shi et al. [11], replaced all of the full connections of traditional FC-LSTM with convolutional operations, and extended the 2-D features and states into 3-D, as shown in Figure 3. Their experimental results demonstrated the superiority of ConvLSTM over traditional FC-LSTM. Thereafter, some variants of ConvLSTM have been developed for action recognition [28], object detection in videos [14], and gesture recognition [29, 30]. For example, Zhu et al. [30] combined ConvLSTM with the 3-D convolution in a multimodal model, and achieved promising gesture recognition performance. Li et al. [28] designed a motion-based attention mechanism and combined it with ConvLSTM in their VideoLSTM, which is proposed for action recognition in videos.

In our work, aiming to better consider the spatial and structural information of input images when performing sequential prediction with LSTM, for the first time, we propose an attention-equipped ConvLSTM structure in the sequential transcription module, and further design a focused attention module to help learn more accurate alignment between predicted characters and corresponding feature areas.

3 Methodology

As illustrated in Figure 4, our proposed FACLSTM, i.e., focused attention ConvLSTM, consists of two components, i.e., the CNN-based feature extraction module and the ConvLSTM-based sequential tran-

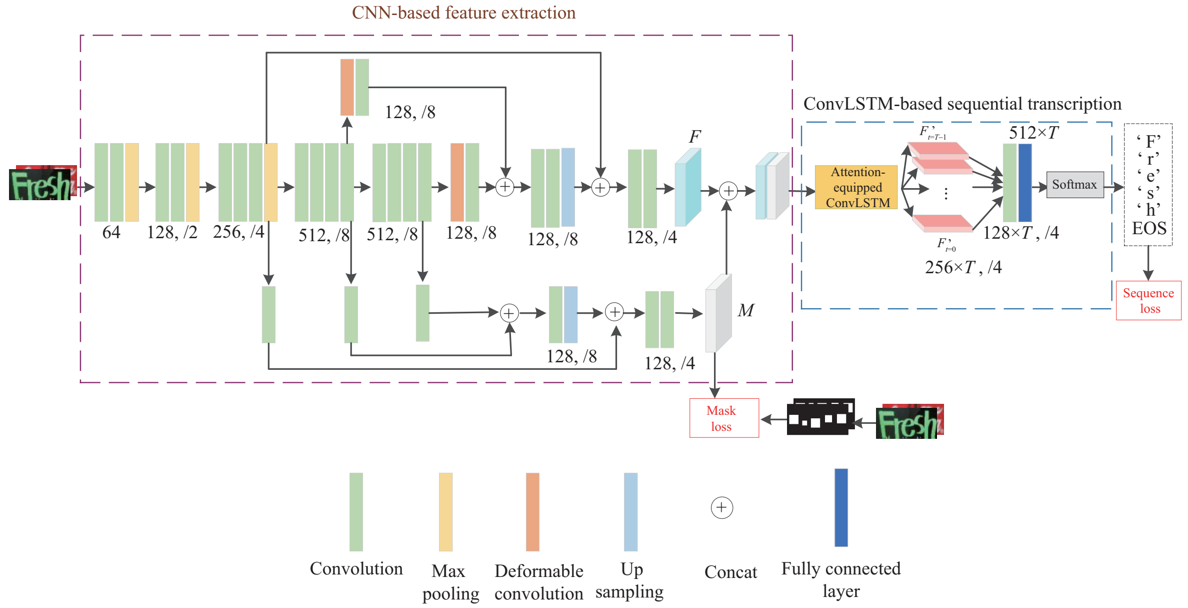

Figure 4 (Color online) Overview of proposed FACLSTM. F and M denote the extracted feature maps and character center masks. T groups of feature maps are produced by the proposed attention-eqipped ConvLSTM, where T is the maximal string length, and the followed softmax classifier is responsible for producing T groups of feature maps from extracted feature maps. Note that, the softmax classifier and previous fully connected layer are shared by the T groups of feature maps.

scription module. The feature extraction module is an encoder-decoder structure that takes VGG-16 as the backbone, while the sequential transcription module is a combination of ConvLSTM and attention mechanism. More details are presented as follows.

3.1 CNN-based feature extraction

Backbone. Similar to Liao's work [10], we take VGG-16 as the encoder of our feature extraction module, and remove the fully connected layers and pooling layers from the last two encoding stages. We also assemble two deformable convolutional layers [31] at stage-4 and stage-5 of the decoder given their flexible receptive fields. However, compared with Liao's network [10], the resolution of final feature maps is restored to a smaller size of $\frac{W}{4} \times \frac{H}{4} \times C$ in our FACLSTM, instead of the $\frac{W}{2} \times \frac{H}{2} \times C$ used in [10], considering the memory and computation cost. Here, W, H and C denote the width, height and channels of feature maps, respectively. In addition, we remove their character attention module set in the encoder stage, and meanwhile, design a focused attention module in the higher-level decoder stage so that more abstract and powerful character center masks can be extracted.

Focused attention module. As pointed out in [5], current attention-based models suffer from the 'attention drift' problem, i.e., they fail to obtain accurate alignment between target characters and related feature areas, especially in complicated and low-quality images. To tackle this problem, in the feature extraction module of the proposed FACLSTM, we assemble two decoder branches, of which one is used as normal for feature extraction and another is designed to learn additional character center masks as centers of text regions are always the key to scene text detection [32] and recognition [10]. These masks are expected to guide the subsequent attention module regarding where to focus. Obviously, for each timestep, the attention should be focused on the center of certain character. Moreover, these masks can also help to enhance foreground pixels and suppress background pixels.

In other recognizers [4,10,12], the feature maps F and maps A generated for other purposes are always combined with the element-wise multiplication $\otimes$ in the way of $F_{\text{out}} = F \otimes (1 + A)$. However, in our experiments we find that directly concatenating feature maps F and character center masks M achieves better performance, which means the subsequent attention-based module prefers to learn patterns from

F and M directly, rather than from their fused results. Therefore, direct concatenation $F_{\text{out}} = F \oplus M$ is used in our FACLSTM.

3.2 Sequential transcription module

As shown in Figure 4, our sequential transcription module starts with an attention-equipped ConvLSTM, by which T groups of feature maps with the size of $\frac{W}{4} \times \frac{H}{4} \times C$ are generated. Here, T is the predefined maximal string length. Afterwards, a 1×1 convolutional layer is applied to reduce the feature map channels, followed by a fully connected layer and a softmax classifier that are employed to sequentially predict T characters. Details of proposed sequential transcription module are presented below.

ConvLSTM. The structure of the traditional FC-LSTM [1] is illustrated in Figure 3(a), and related key formulations can be expressed as

$$\begin{aligned}
i_t &= f(w_{xi}x_t + w_{hi}h_{t-1} + w_{ci} \circ c_{t-1}),\\
f_t &= f(w_{xf}x_t + w_{hf}h_{t-1} + w_{cf} \circ c_{t-1}),\\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(w_{xc}x_t + w_{hc}h_{t-1}),\\
o_t &= f(w_{xo}x_t + w_{ho}h_{t-1} + w_{co} \circ c_t),\\
h_t &= o_t \circ \tanh(c_t),
\end{aligned} \tag{1}$$

where $\circ$ is the Hadamard product (i.e., element-wise multiplication), f denotes the activation function of input gate i_t, output gate o_t and forget gate f_t, and x_t, c_t and h_t represent input features, cell states and cell outputs, respectively.

As we can see, FC-LSTM takes 1-D sequential feature vectors as input, and calculates both the input-to-state and state-to-state transactions in a fully connected manner. Therefore, when applying it to computer vision tasks, the 2-D feature maps have to be mapped into 1-D space, during which the spatial correlation relationships among pixels are badly damaged.

To take advantage of such valuable spatial and structural information in computer vision tasks, Shi et al. [11] proposed ConvLSTM by incorporating convolutional structures into LSTM. As shown in Figure 3(b), all input features, gates, cell states and cell outputs are 3-D in ConvLSTM, and all of the input-to-state and state-to-state transactions are performed with the convolutional operations, instead of the fully connected ones. Thus, the key formulations of ConvLSTM can be written as

$$\begin{aligned}
i_t &= f(w_{xi} * x_t + w_{hi} * h_{t-1} + w_{ci} \circ c_{t-1}),\\
f_t &= f(w_{xf} * x_t + w_{hf} * h_{t-1} + w_{cf} \circ c_{t-1}),\\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(w_{xc} * x_t + w_{hc} * h_{t-1}),\\
o_t &= f(w_{xo} * x_t + w_{ho} * h_{t-1} + w_{co} \circ c_t),\\
h_t &= o_t \circ \tanh(c_t),
\end{aligned} \tag{2}$$

where $*$ denotes the convolutional operation.

Proposed attention-equipped ConvLSTM. The attention mechanism has achieved excellent performance in sequential prediction tasks, such as machine translation [2], speech recognition [3], as well as scene text recognition [5,6,13,16,17]. Especially, in the field of scene text recognition, it has been widely combined with FC-LSTM or GRU to produce more accurate predictions. On the other hand, LSTM is used only for frame-level prediction in the existing studies and is seldom utilized for producing sequential outputs from one single input image unless when combined with the CTC or attention mechanism.

Therefore, in this study, to adapt ConvLSTM to scene text recognition and, meanwhile, provide the proposed network location awareness, we incorporate the attention mechanism into ConvLSTM by weighting the input feature maps with attention scores derived from the cell states and cell outputs obtained at the previous timestep, as illustrated in Figure 5. In addition, to retain the efficiency of the proposed network, an additional bottleneck gate is assembled before the original input gate, forget gate and output gate to reduce the internal feature map channels.

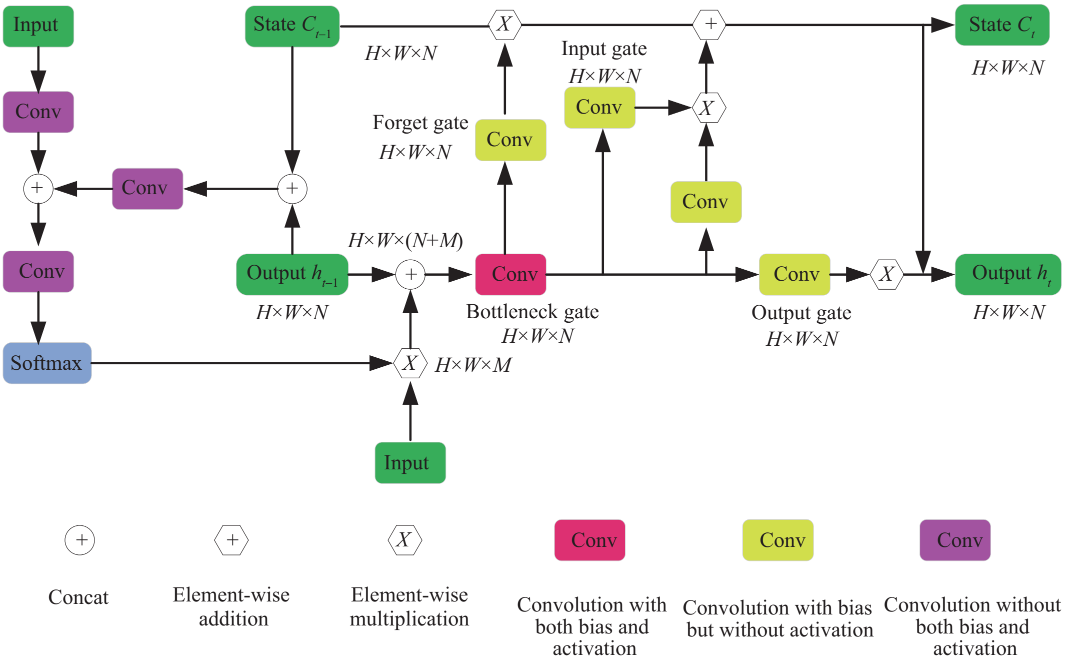

Figure 5 (Color online) Illustration of our proposed attention-equipped ConvLSTM, where the inputs are weighted by attention scores derived from previous cell states and cell outputs.

Eqs. (3) and (4) provide more details on how the cell outputs and the attention scores are calculated. Here, $[\cdot,\cdot]$ is the channel-wise concatenation, $R(\cdot)$ and $S(\cdot)$ denote the ReLU activation function and the Sigmoid function, respectively, and $\widehat{x}_t$ represents the weighted inputs computed by (4). Keep it in mind that all of the gates $\{b,i,o,f\}_t$, inputs $\widehat{x}_t$, cell states $c_{\{t,t-1\}}$ and cell outputs $h_{\{t,t-1\}}$ in (3) and (4) are in 3-D. Moreover, $w_{\{b,i,f,o,b2,h,x\}}$ and $\text{bias}_{\{b,i,f,o,f2,b2,y\}}$ are the involved network weights and biases, and x_t is the concatenation of feature maps F and character center masks M produced by aforementioned encoder-decoder feature extraction module.

$$\begin{aligned}
b_t &= R(w_b * ([\widehat{x}_t, h_{t-1}]) + \text{bias}_b),\\
i_t &= w_i * b_t + \text{bias}_i,\\
f_t &= w_f * b_t + \text{bias}_f,\\
o_t &= w_o * b_t + \text{bias}_o,\\
c_t &= S(f_t + \text{bias}_{f2}) \circ c_{t-1} + S(i_t) \circ R(w_{b2} * b_t + \text{bias}_{b2}),\\
h_t &= R(c_t) \circ S(o_t).
\end{aligned} \tag{3}$$

$$\begin{aligned}
h_{yt} &= [w_h * [c_{t-1}, h_{t-1}], (w_x * x)] + \text{bias}_y,\\
z_t &= w_z * \tanh(h_{yt}),\\
\text{attn}_t &= \text{softmax}(z_t),\\
\widehat{x}_t &= \text{attn}_t \circ x.
\end{aligned} \tag{4}$$

Once the cell outputs $H = \{h_1, h_2, \ldots, h_T\}$, $h_i \in \mathbb{R}^{M\times N\times C}$ are obtained from the proposed attention-equipped ConvLSTM, a 1×1 convolutional layer is applied to map them to $\widetilde{H} = \{\widetilde{h}_1, \widetilde{h}_2, \ldots, \widetilde{h}_T\}$, $\widetilde{h}_i \in \mathbb{R}^{M\times N\times \widetilde{C}}$ and $\widetilde{C} < C$, which is also used to improve model's efficiency, just like the bottleneck gate does. Afterwards, a fully connected layer and a softmax classifier are designed to generate the final sequential outputs $S = \{c_1, c_2, \ldots, c_T\}$ from $\widetilde{H}$, where c_i is from the predefined charset. Compared with STN-OCR [9], where multiple fully connected layers and multiple softmax classifiers are assembled for sequential transcription, in our FACLSTM, only one single fully connected layer and one softmax classifier are employed and shared by T groups of feature maps.

3.3 Training

Loss function. The objective function L of our proposed FACLSTM consists of two parts, i.e., the sequential prediction loss L_s and the mask loss L_m, as formulated in

$$L = L_s(\widehat{y}, \widetilde{y}) + \lambda L_m(m, \widetilde{m}), \tag{5}$$

where m, $\widetilde{m}$, $\widehat{y}$ and $\widetilde{y}$ are the ground truth masks, predicted masks, smoothed ground truth strings and predicted sequential outputs, respectively. λ is the coefficient used to balance the importance of the sequential prediction loss and the mask loss, and is set to 1 in our experiments. Additionally, the label smoothing method proposed by Szegedy et al. [33] is able to help regularize the proposed model. Therefore, given the one-hot encoded ground truth y^{OneHot}, we convert it to the smoothed version $\widehat{y}$ with

$$\widehat{y} = (1.0 - \epsilon) * y^{\text{OneHot}} + \epsilon * \left(\frac{1}{N_{\text{class}}}\right). \tag{6}$$

Moreover, for the ground truth masks m, we set the value of their foreground pixels (center of characters) and background pixels to 1 and 0, respectively. Thus, the mask loss L_m is calculated in the way of

$$L_m = 0.01 * \left\{1 - 2 * \left[\frac{\sum(m \otimes \widetilde{m})}{\sum m + \sum \widetilde{m}}\right]\right\}. \tag{7}$$

Generation of ground truth. Obviously, to optimize the proposed network, the ground truth of character center masks is required. Assuming $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ is the bounding box of individual characters, we use the same method as that in [10] to calculate the ground truth of the corresponding mask $g = (x^g_{\min}, y^g_{\min}, x^g_{\max}, y^g_{\max})$, as shown in

$$\begin{aligned}
&w = x_{\max} - x_{\min},\\
&h = y_{\max} - y_{\min},\\
&x^g_{\min} = (x_{\min} + x_{\max} - w * r)/2,\\
&x^g_{\max} = (x_{\min} + x_{\max} + w * r)/2,\\
&y^g_{\min} = (y_{\min} + y_{\max} - h * r)/2,\\
&y^g_{\max} = (y_{\min} + y_{\max} + h * r)/2.
\end{aligned} \tag{8}$$

Note that, the shrink ratio r is set to 0.25 in our experiments, instead of 0.5 used in [10].

4 Experiments

4.1 Datasets

We train the proposed FACLSTM network with 7 million synthetic images from SynthText dataset [34][1] without fine-tuning on individual real-word datasets, and evaluate the corresponding performance on three widely used benchmarks, including the regular text dataset IIIT5K, low-resolution and noisy text dataset SVT, and curved text dataset CUTE.

- SynthText is proposed by Gupta et al. [34] for scene text detection. The original dataset is composed of 800000 scene text images, each with multiple word instances. Texts in this dataset are rendered in different styles, and annotated with character-level bounding boxes. Overall, about 7 million text images are cropped for scene text recognition.
- IIIT5K is built by Mishra et al. [35]. This dataset consists of 3000 text images obtained from the web. Most of these images are regular, and for individual images, two lexicons are provided, including one 50-word lexicon and one 1000-word lexicon.

[1] http://www.robots.ox.ac.uk/~vgg/data/scenetext/.

Table 1 Result comparison across different methods and datasets a)b)

Method	LSTM	Samples	IIIT5K_None	IIIT5K_50	IIIT5K_1k	SVT	CUTE
FAN [5]	FC-LSTM	12M*	87.4	99.3	97.5	–	63.9
AON [6]	FC-LSTM	12M*	87.0	**99.6**	98.1	–	76.8
CRNN [7]	FC-LSTM	8M*	78.2	97.6	94.4	–	–
Gao et al. [4]	FC-LSTM	8M*	83.6	99.1	97.2	–	–
RARE [8]	FC-LSTM	8M*	81.9	96.2	93.8	–	59.2
$\mathrm{R^2AM}$ [17]	FC-LSTM	7M*	78.4	96.8	94.4	–	–
SqueezedText_binary [26]	FC-LSTM	1M	86.6	96.9	94.3	–	–
SqueezedText [26]	FC-LSTM	1M	87.0	97.0	94.1	–	–
CA-FCN [10]	No	7M	**92.0**	**99.8**	**98.9**	**82.1**	**78.1**
Gao et al. [12]	No	8M*	81.8	99.1	97.9	–	–
STN-OCR [9]	No	–	86.0	–	–	79.8	–
FLSTM_base1	FC-LSTM	7M	73.7	99.0	97.4	58.7	67.4
FAFLSTM_base2	FC-LSTM	7M	87.8	99.3	98.1	78.2	75.7
FACLSTM (proposed)	ConvLSTM	7M	**90.5**	99.5	**98.6**	**82.2**	**83.3**

a) Word-level recognition rate is used here. The two best recognizers on individual datasets are indicated in bold. IIIT5K_None, IIIT5K_50 and IIIT5K_1k denote no lexicon, 50-word lexicon and 1000-word lexicon, respectively.

b) Samples: the number of samples used for training individual models, where * means datasets derived from SVT are used.

- SVT is a very challenging dataset collected by Wang et al. [36] from the Google street view. Totally, 647 text images with low-resolution and noise are included.
- CUTE is released by Risnumawan et al. [37]. There are only 288 word images in this dataset, but most of them are seriously curved. Therefore, compared with other datasets, CUTE is more challenging.

4.2 Implementation details

In our experiments, all of the input images are scaled to a size of 64×256 with aspect ratio preserved. The maximal string length is set to 20, including one START token and one EOF token. This means up to 18 real characters are allowed within individual words. Our charset is composed of 39 characters, i.e., 26 alphabet letters, 10 digits, 1 START token, 1 EOS token and 1 special token for any other symbols. The Adam optimizer with an initial learning rate of 1E−4 is employed in our study to optimize the proposed network. Totally, the proposed FACLSTM is trained for five epochs, with learning rates of 1E−4, 1E−4, 5E−5, 1E−5, and 1E−6, respectively. Moreover, the kernel size and channels (N in Figure 5) of the convolutional operations in (3) and (4) are set to 3×3 and 256, respectively. Finally, the proposed network is implemented under the Tensorflow framework.

4.3 Experimental results

We evaluate the performance of our proposed FACLSTM on the aforementioned three benchmark datasets, and compare it with those of the state-of-the-art approaches. Table 1 presents the details of the comparison results. Note that, in this table, CA-FCN [10] and SqueezedText [26] are the two latest recognizers recently published in AAAI2019 and AAAI2018.

Comparison with methods based on the traditional FC-LSTM. As previously introduced, traditional FC-LSTM is widely used in existing recognizers. Among the methods listed in Table 1, RARE [8], AON [6], and FAN [5] combined FC-LSTM with the attention mechanism in the fully connected way when performing sequential transcription, while CRNN [7], $\mathrm{R^2AM}$ [17], Gao's model [4], and SqueezedText [26] utilized FC-LSTM for frame-level prediction, sequential feature encoding or other purposes. As shown in Table 1, our proposed FACLSTM outperforms these FC-LSTM-based methods by large margins on both regular text dataset IIIT5K (90.5% vs. 87.4%) and curved text dataset CUTE (83.33% and 76.8%) when no lexicon is used. It also achieves competitive performance on IIIT5K when 1000-word lexicon and 50-word lexicon are used. Apparently, handling the text recognition task from the spatiotemporal perspective with our ConvLSTM-based FACLSTM is more effective than casting it to a sequence-to-sequence prediction problem via FC-LSTM, no matter for regular or irregular text images. Note that our FACLSTM is optimized with less training samples than most of the listed FC-LSTM-based recognizers, except for R^{2}AM [17] and SqueezedText [26], and though AON [6] is specially designed for irregular text recognition, its recognition performance on CUTE is still 6.5% lower than that of our FACLSTM.

Readers should keep in mind that apart from the 4 million training images from SynthText, the recognizers named AON [6] and FAN [5] also employed additional 8 million images provided by Jaderberg et al. [38] for their training. Jaderberg's synthetic images are generated with a 50k-word lexicon that covers all the test words of ICDAR and SVT datasets, and blended with word images randomly-sampled from these two datasets. Thus, the recognition performance on SVT would benefit largely from the usage of Jaderberg's images because of this strong correlation. This is also proved by Liao's study [10], where a 4.3% accuracy improvement on SVT was achieved by their CA-FCN when additional 4 million images generated with Jaderberg's strategy were used. In this study, to demonstrate the generalizability and robustness of proposed FACLSTM, we only employ the SynthText dataset to train our network. Therefore, to give a fair comparison, we only compare FACLSTM with recognizers not utilizing SVT-derived training images, such as CA-FCN [10] and STN-OCR [9].

Comparison with Non-LSTM based methods. Considering the limitations of the traditional FC-LSTM on neglecting spatial and structural information and slow training convergence, CA-FCN [10], Gao's model [12], and STN-OCR [9] have also explored other non-LSTM solutions. Especially, CA-FCN [10] also addressed the recognition issue from the 2-D perspective by utilizing an FCN structure, and moreover, it used the same VGG-16 backbone and 7-million training images as our FACLSTM.

From Table 1, we can see that the accuracy of our proposed FACLSTM is 1.5% lower than that of the best recognizer, i.e., CA-FCN [10], on the regular text dataset IIIT5K. However, on the more challenging curved text dataset CUTE, we achieve an accuracy of 83.3%, which is 5.2% higher than that of CA-FCN [10]. As for the low-resolution and noisy dataset SVT, our FACLSTM performs slightly better than CA-FCN [10] with an accuracy of 82.2% (vs. 82.1% of CA-FCN [10]). Note that, CA-FCN [10] is not an end-to-end trainable system because in order to infer the final sequential outputs from the pixel-level predictions generated by their network, an empirical rule-based word formation module is required. By contrast, our FACLSTM is able to directly produce the final sequential outputs via the proposed ConvLSTM-based sequential transcription module. Admittedly, replacing FC-LSTM with Conv-LSTM will increase the memory cost. Therefore, to retain the efficiency, we up-sample feature maps to a small resolution of 1/4 in the decoder branches, instead of 1/2 used in CA-FCN. Undoubtedly, this small resolution will compromise the recognition accuracy to some extent, especially for small-size and low-resolution images from the IIIT5K and SVT datasets.

Effectiveness of the proposed focused attention module and ConvLSTM-based sequential transcription module. Furthermore, to highlight the effectiveness of our proposed focused attention module and ConvLSTM-based sequential transcription module, we compare the performance of our proposed FACLSTM with that of the following two baseline models.

- FLSTM_base1, which shares the same feature extraction module with our proposed FACLSTM, but removes the focused attention module. Besides, the sequential transcription module used in this model is the traditional attention-based FC-LSTM network, just as the one used in AON [6], FAN [5] and both Gao's models [4, 12].
- FAFLSTM_base2, which is built upon FLSTM_base1, but with the proposed focused attention module applied.

Apparently, from the comparison of FLSTM_base1 and FAFLSTM_base2, we can see that the recognition accuracies on IIIT5K, SVT, and CUTE datasets are elevated by 14.1%, 19.5%, and 8.4%, respectively when the proposed focused attention module is assembled. As illustrated in Figure 6, the focused attention module is able to accurately predict the character center masks because it is performed in the high-level decoder branch. The significant performance improvement demonstrates that these masks are effective to help the sequential transcription module to focus attention on the right character areas and suppress irrelevant background pixels. In addition, the image resolution of CUTE in much higher than

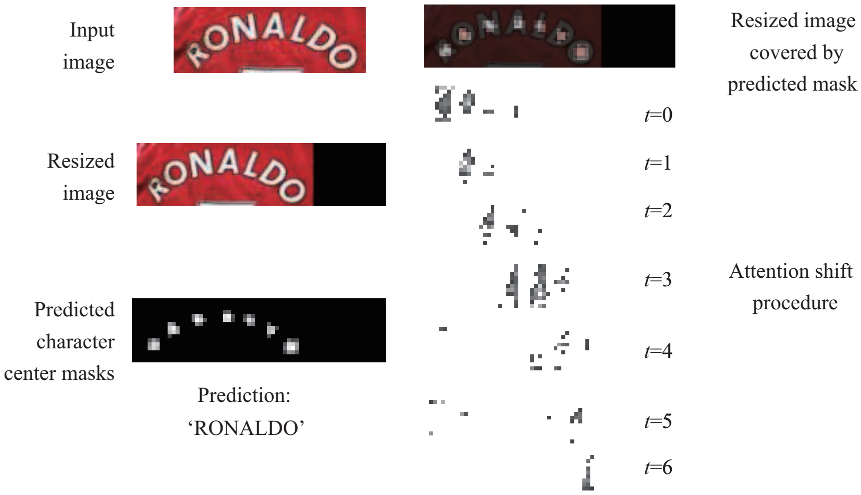
Input image
Resized image
Predicted character center masks
Prediction: 'RONALDO'
Resized image covered by predicted mask
Attention shift procedure
t=0
t=1
t=2
t=3
t=4
t=5
t=6

Figure 6 (Color online) Visualization results of predicted mask and attention shift procedure.

that of SVT and IIIT5K, and SVT is much noisier than the other two datasets. As claimed in [5,10], the attention-based recognizers perform poorly on low-quality images because of the 'attention drift' problem, and the scene text images suffer from noisy background badly, so the accuracy improvement is more on SVT and less on CUTE when the proposed focused attention module is utilized.

Moreover, from the comparison of FAFLSTM_base2 and FACLSTM, we can see that when the traditional attention-based FC-LSTM module is replaced by our proposed attention-ConvLSTM-based sequential transcription module, further 2.7%, 3.6%, and 7.6% improvements are achieved on IIIT5K, SVT, and CUTE, respectively. This means our FACLSTM is able to boost the recognition performance significantly by utilizing the proposed attention-ConvLSTM module to take benefits from the valuable spatial and structural information of text images. As clarified in [10], FC-LSTM only achieves good performance on horizontal or nearly horizontal text, and its performance on curved text is seriously limited because of the neglect of pixels' spatial correlation relationships. The huge performance improvement achieved by FACLSTM on CUTE evidences that our attention-ConvLSTM module is a good solution to this problem.

Therefore, we can say that both of the proposed focused attention module and attention-ConvLSTM module are effective. Note that the focused attention module can be removed from the network when datasets without character-level bounding box annotations are used for the training.

In summary, on the regular text dataset, our proposed FACLSTM outperforms all of listed FC-LSTM-based and non-LSTM-based recognizers, except CA-FCN, but on the more challenging curved text dataset, our FACLSTM surpasses all of the listed methods significantly with an accuracy of 83.3%, including CA-FCN (78.1%). Moreover, the comparisons with other two baseline models demonstrate the effectiveness of our proposed focused attention module and ConvLSTM-based sequential transcription module. Finally, we also give the visualization results of the predicted masks and the attention shift procedure, as shown in Figure 6. The comparison results of attention predicted by FACLSTM and FLSTM_base1 are shown in Figure 7. Note that our FACLSTM directly produces 2-D attention maps via the convolutional operations, while FLSTM_base1 generates 1-D attention vectors with the fully connected layers, just as other existing FC-LSTM-based recognizers did. These 1-D attention vectors are reshaped to 2-D maps in Figure 7 for an intuitional visualization. As we can see, the attention areas of FACLSTM is larger and more accurate, and the 'attention drift' problem is alleviated to some extent in our proposed FACLSTM.

5 Conclusion

Scene text recognition has been treated as a sequence-to-sequence prediction problem for quite a long time, and traditional FC-LSTM is widely used in current state-of-the-art recognizers. In this work,

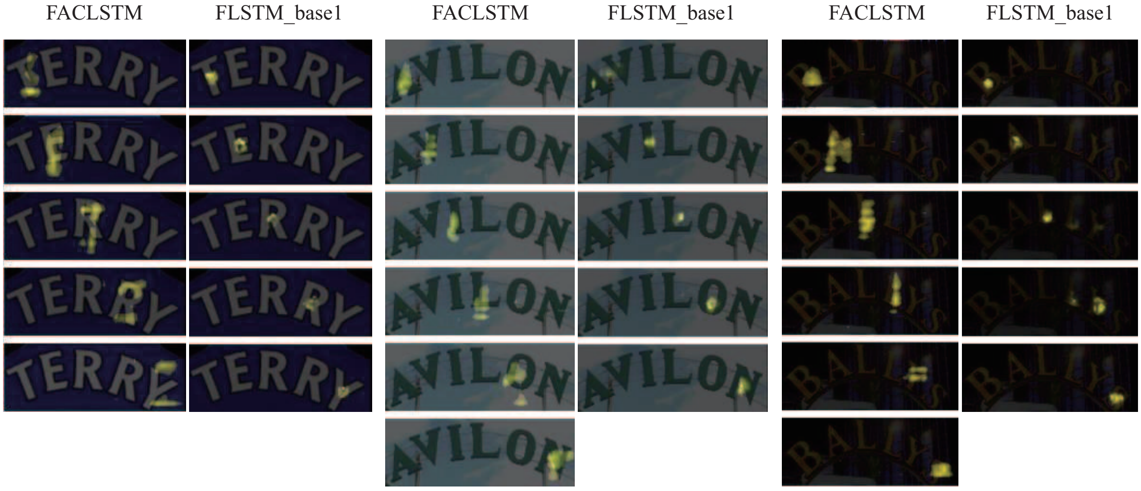

Figure 7 (Color online) Visualization results of attention predicted by FACLSTM and FLSTM_base1. Values of the attention maps are normalized and truncated for a better visualization. Note that FACLSTM directly produces 2-D attention maps, while FLSTM_base1 generates 1-D attention vectors, which are then reshaped to 2-D space.

we have demonstrated that scene text recognition is actually a spatiotemporal prediction problem and we have proposed to tackle this problem from the spatiotemporal perspective. Toward this end, we have presented an effective scene text recognizer named FACLSTM, where ConvLSTM was applied and improved by integrating the attention mechanism in the sequential transcription module, and a focused attention module has been designed at the encoder-decoder feature extraction stage. Experimental results have revealed that, our proposed FACLSTM is able to handle both regular and irregular (low-resolution, noisy and curved) text well. Especially for the curved text, our proposed FACLSTM has outperformed other advanced approaches by large margins. Thus, we can conclude that ConvLSTM is more effective in scene text recognition than the widely used FC-LSTM because the valuable spatial and structural information can be better leveraged when performing sequential prediction with ConvLSTM.

Acknowledgements This work was supported by China Scholarship Council (Grant No. 201706140138), Shanghai Natural Science Foundation (Grant No. 19ZR1415900), and Shanghai Knowledge Service Platform Project (Grant No. ZF1213).

References

1 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

2 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations, 2015

3 Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition. 2015. ArXiv: 1506.07503

4 Gao Y Z, Chen Y Y, Wang J Q, et al. Dense chained attention network for scene text recognition. In: Proceedings of International Conference on Image Processing, 2018

5 Cheng Z Z, Bai F, Xu Y L, et al. Focusing attention: towards accurate text recognition in natural images. In: Proceedings of IEEE International Conference on Computer Vision, 2017

6 Cheng Z Z, Xu Y L, Bai F, et al. AON: towards arbitrarily-oriented text recognition. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2018

7 Shi B G, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2298–2304

8 Shi B G, Wang X G, Lyu P Y, et al. Robust scene text recognition with automatic rectification. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2016

9 Bartz C, Yang H J, Meinel C. STN-OCR: a single neural network for text detection and recognition. 2017. ArXiv: 1707.08831v1

10 Liao M H, Zhang J, Wan Z Y, et al. Scene text recognition from two-dimensional perspective. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019

11 Shi X J, Chen Z R, Wang H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of Neural Information Processing Systems, 2015

12 Gao Y Z, Chen Y Y, Wang J Q, et al. Reading scene text with attention convolutional sequence modeling. 2017. ArXiv: 1709.04303v1

13 Wojna Z, Gorban A, Lee D, et al. Attention-based extraction of structured information from street view imagery. In: Proceedings of International Conference on Document Analysis and Recognition, 2017

14 Liu M, Zhu M L. Mobile video object detection with temporally-aware feature maps. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2018

15 Ye Q X, Doermann D. Text detection and recognition in imagery: a survey. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 1480–1500

16 Shi B G, Yang M K, Wang X G, et al. ASTER: an attentional scene text recognizer with flexible rectification. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 2035–2048

17 Lee C Y, Osindero S. Recursive recurrent nets with attention modeling for OCR in the wild. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2016

18 Bai X, Liao M K, Shi B G, et al. Deep learning for scene text detection and recognition (in Chinese). Sci Sin Inform, 2018, 48: 531–544

19 Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks. 2015. ArXiv: 1506.02025

20 Bai F, Cheng Z Z, Niu Y, et al. Edit probability for scene text recognition. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2018

21 Su B L, Lu S J. Accurate recognition of words in scenes without character segmentation using recurrent neural network. Pattern Recogn, 2017, 63: 397–405

22 Su B L, Lu S J. Accurate scene text recognition based on recurrent neural network. In: Proceedings of Asian Conference on Computer Vision, 2014

23 Li H, Wang P, Shen C H, et al. Show, attend and read: a simple and strong baseline for irregular text recognition. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019

24 Jaderberg M, Vedaldi A, Zisserman A. Deep features for text spotting. In: Proceedings of European Conference on Computer Vision, 2014

25 Tian S X, Bhattacharya U, Lu S J, et al. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. Pattern Recogn, 2016, 51: 125–134

26 Liu Z C, Li Y X, Ren F B, et al. SqueezedText: a real-time scene text recognition by binary convolutional encoder-decoder network. In: Proceedings of AAAI Conference on Artificial Intelligence, 2018

27 Huang T J, Tian Y H, Li J, et al. Salient region detection and segmentation for general object recognition and image understanding. Sci China Inf Sci, 2011, 54: 2461–2470

28 Li Z Y, Gavrilyuk K, Gavves E, et al. VideoLSTM convolves, attends and flows for action recognition. Comput Vision Image Underst, 2018, 166: 41–50

29 Zhang L, Zhu G M, Mei L, et al. Attention in convolutional LSTM for gesture recognition. In: Proceedings of Neural Information Processing Systems, 2018

30 Zhu G M, Zhang L, Shen P Y, et al. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 2017, 5: 4517–4524

31 Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks. In: Proceedings of International Conference on Computer Vision, 2017

32 Chen J, Lian Z H, Wang Y Z, et al. Irregular scene text detection via attention guided border labeling. Sci China Inf Sci, 2019, 62: 220103

33 Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2016

34 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localization in natural images. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2016

35 Mishra A, Alahari K, Jawahar C V. Top-down and bottom-up cues for scene text recognition. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, 2012

36 Wang K, Babenko B, Belongie S. End-to-end scene text recognition. In: Proceedings of International Conference on Computer Vision, 2011

37 Risnumawan A, Shivakumara P, Chan C S, et al. A robust arbitrary text detection system for natural scene images. Expert Syst Appl, 2014, 41: 8027–8048

38 Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition. 2014. ArXiv: 1412.1842