ReELFA: A Scene Text Recognizer with Encoded Location and Focused Attention

Qingqing Wang[*,+], Wenjing Jia[+], Xiangjian He[+], Yue Lu[*], Michael Blumenstein[+], Ye Huang[+], Shujing Lyu[*]
[*]*Department of Computer Science and Technology*
East China Normal University, Shanghai, China
[+]*Faculty of Engineering and Information Technology*
University of Technology Sydney, Sydney, Australia
qingqing.wang-1@student.uts.edu.au

***Abstract*—LSTM and attention mechanism have been widely used for scene text recognition. However, existing LSTM-based recognizers usually convert 2D feature maps into 1D space by flattening or pooling operations, resulting in the neglect of spatial information of text images. Additionally, the attention drift problem, where models fail to align targets at proper feature regions, has a serious impact on the recognition performance of existing models. To tackle the above problems, in this paper, we propose a scene text Recognizer with Encoded Location and Focused Attention, *i.e.*, ReELFA. Our ReELFA utilizes one-hot encoded coordinates to indicate the spatial relationship of pixels and character center masks to help focus attention on the right feature areas. Experiments conducted on benchmark datasets IIIT5K, SVT, CUTE and IC15 demonstrate that the proposed method achieves comparable performance on the regular, low-resolution and noisy text images, and outperforms state-of-the-art approaches on the more challenging curved text images.**

***Index Terms*—attention LSTM; encoded location; center masks; attention drift.**

I. Introduction

Text is an important way to convey information and knowledge. Scene text recognition has been studied extensively since 1990s due to its great application potentials, such as image retrieval, automatic navigation, assistance for the blind and car plate recognition etc. Challenges of this task mainly arise from poor image qualities (including low resolution, blur, skew, uneven illumination, etc.) and unconstrained text appearances (in terms of sizes, fonts, colors, directions, backgrounds, etc.). In the past decades, though many efforts have been made, scene text recognition is still unsolved and attracts numerous attentions from research community.

Currently, leading solutions to scene text recognition are all based on deep learning techniques, among which Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Connectionist Temporal Classification (CTC) and attention-based decoders are most widely used [1]–[8]. Specifically, CNN and LSTM are usually employed to extract deep features from input images and perform encoding or prediction, while CTC and attention-based decoders are the most popular sequential transcription models.

However, LSTM is an idea borrowed from speech recognition and machine translation, where the inputs are 1D vectors, rather than 2D feature maps. Therefore, to adapt LSTM to scene text recognition, 2D feature maps have to be down-sampled or flattened into 1D space, as shown in Fig. 1, which severely damages the valuable spatial correlation information of text images. As claimed in [9], LSTM-based recognizers can only achieve good performance on horizontal or nearly horizontal texts. As for curved or skewed text, the performance is far from being satisfactory. On the other hand, for the sequential transcription models, CTC takes the prediction results of individual frames as input and searches for the optimal sequential outputs while considering all the possible paths via a forward-backward algorithm. By contrast, the attention-based decoder directly produces sequential outputs from input features via auto-regressive connections, and as claimed in [8], has achieved better performance than CTC. Therefore, the state-of-the-art approaches usually take attention-Gated Recurrent Units (attention-GRU) or attention-LSTM as their decoders. However, as pointed out in [2], the attention-based decoder suffers from the 'attention drift' problem, *i.e.*, models cannot align targets at proper feature regions due to poor image quality, complex backgrounds or crowed character appearance.

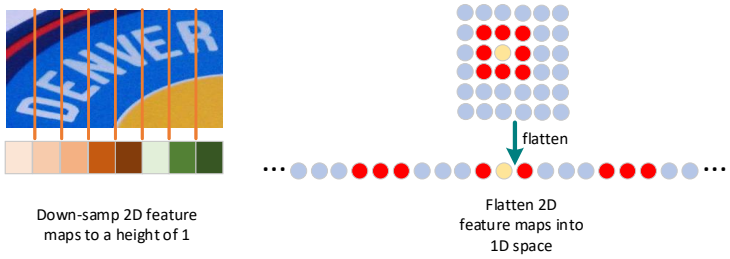

Fig. 1. Convert 2D feature maps into 1D space to adapt LSTM to scene text recognition

Showing the contributions of the work in this paper, we design an efficient scene text recognizer with encoded location and focused attention, *i.e.*, ReELFA, to tackle the aforementioned problems. Inspired by Wojna et al. [8], when flattening 2D feature maps into 1D space, we attach the one-hot encoded coordinates for individual pixels to indicate their spatial relationships. Besides, at the CNN-based feature extraction stage, we embed a second decoder branch to learn the character center masks, aiming to assist the subsequent

module to focus their attentions at proper feature areas.

The rest of the paper is organized as follows. Section II briefly introduces the existing scene text recognizers. Section III describes the proposed approach in details. The designed experiments and conclusion are drawn in Sections IV and V, respectively.

II. Related works

Traditional methods usually address the scene text recognition issue from a bottom-up perspective, namely detecting and recognizing single characters with sliding window-based or connected component-based techniques, followed by integrating individual characters into words with consideration of language models [10], [11]. In these methods, handcrafted features and classifiers like support vector machine (SVM) are widely used. Ye et al. [12] performed a comprehensive review for the methods following this fashion.

Around 2015, deep learning was introduced to the field of computer vision and surpassed traditional methods significantly in a variety of tasks. Since then, recognizers based on deep learning techniques have become the dominant solutions to scene text recognition. At first, CNN was simply used as character classifier or word classifier [13], [14] in a group of Deep Convolutional Neural Network (DCNN) recognizers because of its extraordinary representation ability and convenient end-to-end trainable structure. Afterwards, inspired by sequence-to-sequence prediction tasks such as machine translation and speech recognition, the Recurrent Neural Network (RNN)-based networks emerged and became popular.

For instance, CRNN [4] utilized CNN for deep feature exaction and bi-LSTM for frame-level prediction. Then, a CTC-based sequential transcription module was assembled to produce sequential outputs. R^2AM [15] also followed the same idea, and exploited a combination of recursive CNN and recurrent CNN to capture longer data dependencies when extracting deep features with CNN. Though CRNN and R^2AM outperformed DCNN recognizers by a large margin, they were not capable of handling irregular texts (curved or skewed texts) well. To tackle this problem, RARE [5] employed a Spatial Transformer Network (STN) to rectify input images into more 'readable' ones before feeding them into the followed recognizer. The same as CRNN, CNN and bi-LSTM were also used in RARE, but instead of frame-level prediction, the bi-LSTM in RARE was employed to further encode frames into feature vectors because it leveraged attention-GRU as its sequential transcription model, which directly generated sequential predictions from features. Inspired by RARE, STN-OCR [16] also took advantage of STN, but for a different purpose, *i.e.*, sampling potential text regions from input images.

Recently, Cheng et al. [2] raised the problem of 'attention drift' and pointed out that it was the bottleneck of existing attention-based recognizers. They designed a Focusing Attention Network (FAN) to address this issue and improved the recognition performance significantly. Additionally, Cheng et al. [3] also proposed another recognizer named AON for irregular text recognition. AON extracted horizontal, vertical and character placement features from input images, and employed a filter gate to filter out irrelevant features before feeding them into the following attention based decoder. SqueezedText proposed in [17] was a real-time scene text recognizer that utilized binary convolutional encoder-decoder network (B-CEDNet) to alleviate the computational burden. Firstly, C (number of character classes) salience maps was generated by the B-CEDNet. Then, thresholding and binary morphologic filtering were performed to obtain character sequences, which were fed into a Bi-RNN network later to produce sequential outputs. As we can see, so many recognizers have relied on LSTM. As aforementioned, these models neglect the spatial correlation information of 2D text images, so they cannot handle the irregular text well. To address this issue, Liao et al. [9] proposed a network named Character Attention Fully Convolutional Network (CA-FCN), which produced salience maps from 2D feature maps. Finally, the sequential outputs were inferred from these salience maps with some empirical rules. CA-FCN also calculated attention maps to highlight the foreground and weaken the background to further boost the recognition performance.

In this work, by considering the drawback of current LSTM-based and attention-based recognizers, we propose ReELFA, which sequential transcription module sequentially decodes the outputs from deep features, one-hot encoded location and character center masks. More details of the proposed network are presented next.

III. Methodology

A. Overview

As illustrated in Fig. 2, our proposed network consists of two parts, *i.e.*, an encoder-decoder feature extraction module and an attention-LSTM-based sequence transcription module. Inspired by [9], the feature extraction module takes VGG-16 as its backbone network and is assembled with two decoder branches: one is used as normal to extract deep features from input images, and the other is specially designed to learn character center masks, which are supposed to help subsequent module to focus attentions at proper areas. Afterwards, together with one-hot encoded coordinates, the extracted deep features and character center masks are fed into the attention-LSTM-based decoder to produce final sequential predictions.

B. Structure of the Proposed ReELFA

Since 1D feature vectors are usually with poor representation ability for non-horizontal texts, in this work, we generate 2D feature maps with a fully convolutional encoder-decoder network. Motivated by [9], we take the VGG-16 without pooling layers at the stage-4 and stage-5 as our backbone network, and embed two deformable convolution layers at the decoder stage, given their flexible receptive fields [18]. Assuming the size of input images is $H \times W \times 3$, our final generated feature maps, indicated by F in Fig. 2, are of size

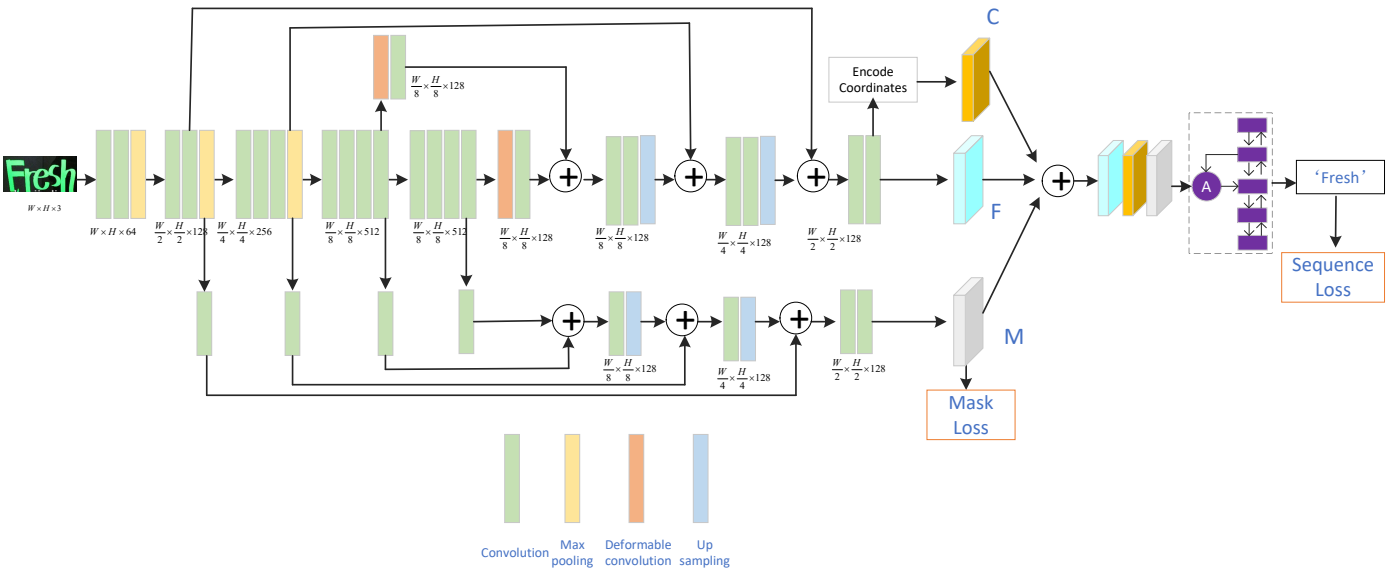

Fig. 2. The structure of our proposed ReELFA network

$\frac{H}{2} \times \frac{W}{2} \times C$, where H, W and C denote the height, width and number of channels of the feature maps, respectively.

Character Center Masks: In [9] and [19], in order to highlight the relevant pixels and suppress the irrelevant pixels, attention maps (character center masks) are generated at the encoder stage and fused with feature maps with the way shown in Eq. 1, where A denotes the attention map and $\otimes$ means pixel-wise multiplication. In fact, this idea can also be leveraged to tackle the 'attention drift' problem because the centers of characters are exactly where the attentions should be placed. Towards this end, we produce character center masks to help focus attentions at proper areas. However, in contrast to [9] and [19], we generate only one group of masks, indicated by M in Fig. 2, via a second decoder branch, rather than multiple groups of masks at the low-level encoder stage. In addition, in our experiments, we find that directly concatenating feature maps F and center masks M is able to achieve better performance and faster convergence speed than pixel-wise multiplication. Therefore, instead of using Eq. 1, Eq. 2 is applied to combine F and M in our work, where $\oplus$ is the concatenation operation.

$$F_o = F \otimes (1 + A) \tag{1}$$

$$F_o = F \oplus M \tag{2}$$

One-hot Encoded Location: LSTM takes sequential feature vectors as input, thus 2D feature maps have to be flattened into 1D space before proceeding to the LSTM-based modules, so the spatial relationships of pixels are disturbed. To retain such important information, inspired by Wojna et al. [8], we propose to utilize one-hot encoded coordinates to make the LSTM 'location aware'. As shown in Fig. 3, where f_k, c_k and m_k represent the extracted features, one-hot encoded coordinates and character center masks, respectively, the encoded coordinates c_k of pixel P_1 is closer to that of adjacent pixels P_2 and P_3 when comparing with P_4, who has longer distance to P_1 than other pixels.

Attention-LSTM-based Sequence Transcription: At the end of our proposed network, an attention-LSTM-based sequence transcription model is exploited to generate target sequential outputs $(y_1, y_2, ..., y_N)$ from the input feature vectors $[f_1, f_2, ..., f_K]$, the center masks $[m_1, m_2, ..., m_K]$ and the encoded coordinates $[c_1, c_2, ..., c_K]$, where $f_k \in R^L$, $m_k \in R^H$ and $c_k \in R^T$, and K is the length of sequential feature vectors. The procedure can be formulated in Eq. 3 as:

$$\begin{aligned} u_t &= \sum_{k=1}^{K} \alpha_{t,k}(f_k + c_k + m_k) \\ x_t &= W_y \overline{y}_{t-1} + W_{u1} u_{t-1} \\ (o_t, s_t) &= LSTM(x_t, s_{t-1}) \\ \widetilde{o}_t &= softmax(W_o o_t + W_{u2} u_t) \\ \widetilde{y}_t &= \arg\max_y \widetilde{o}_t(y), \end{aligned} \tag{3}$$

where u_t and $\widetilde{y}_t$ are the weighted features and prediction results at time t, and x_t, o_t and s_t denote the inputs, outputs and states of the LSTM at time t. $\overline{y}_{t-1}$ is the ground truth y_{t-1} at the training stage equals to the prediction result $\widetilde{y}_{t-1}$ at the inference stage. The attentions of the k^{th} feature vector at time t are denoted by $\alpha_{t,k}$, and can be derived from Eq. 4 by:

$$\begin{aligned} a_{t,k} &= V_a^T tanh(W_s s_t + W_f f_k + W_c c_k + W_m m_k) \\ \alpha_t &= softmax_k(a_{t,k}). \end{aligned} \tag{4}$$

C. Training

From Fig. 2 we can see that the loss function L of our proposed ReELFA consists of two parts, *i.e.*, the sequence

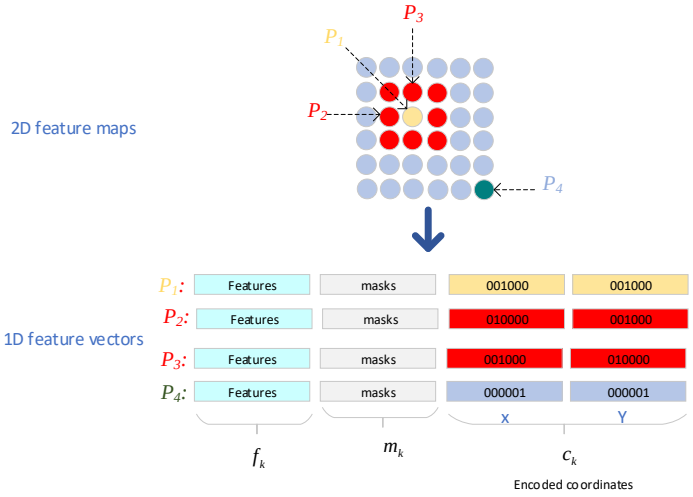

Fig. 3. Illustration of one-hot encoded location.

transcription loss L_s and the mask loss L_m, as expressed in Eq. 5, where $\widetilde{y}$ and $\widetilde{m}$ are the predicted sequential label and mask, while $\widehat{y}$ and m are the ground truth label and mask. To regularize the proposed model and make it more adaptable, when calculating L_s, the original sequential label y^{OneHot} is smoothed to $\widehat{y}$ with the method proposed in [20] and described in Eq. 6, where the label smoothing weight ϵ is set to 0.1 in our experiments. As for the mask loss L_m, we calculate it with Eq. 7 when setting the foreground and background pixels to be 1 and 0, respectively.

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

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

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

Additionally, the ground truth of character center masks are required to optimize the proposed network. Assuming that $b = (x, y, w, h)$ represents the bounding box of an character, where (x, y), w and h denote the center coordinate, width and height of the bounding box, we shrink b to $\overline{b} = (x, y, r \times w, r \times h)$ with a ratio of r (set to 0.25 in our experiments) to obtain the ground truth of individual masks.

IV. Experiments

In this section, we evaluate the performance of our proposed ReELFA on both regular and irregular text. Word-level accuracy is used as the measurement of performance, just as other compared approaches did. Related datasets, implementation details and comparison results with the state-of-the-art approaches are described in details below.

A. Datasets

We train the proposed ReELFA on the synthetic dataset SynthText [21] without fine-tuning on individual real-world datasets, and evaluate the corresponding performance on four widely used benchmarks named IIIT5K, SVT, IC15 and CUTE.

- **SynthText** is a synthetic dataset created by Gupta et al. [21] for scene text detection. It contains 800,000 training images, from which about 7 million word images can be cropped for the recognition task. Bounding boxes and transcriptions are provided for text presented in scene images of this dataset.
- **IIIT5K** is provided by Mishra et al. [22]. This dataset consists of 3,000 test images obtained from the web, and for individual images, one 50-word lexicon and one 1000-word lexicon are provided.
- **SVT** is collected from the Google Street View by Wang et al. [23]. Totally, 647 low-resolution and noisy images are included.
- **IC15** is short for ICDAR 2015 dataset [24]. 2,077 cropped scene text images, including 200 irregular ones (arbitrary-oriented, curved or perspective), are included in this dataset.
- **CUTE** is proposed in [25] and only contains 288 images, but most of them are severely curved. Therefore, it is more challenging when compared with other datasets.

B. Implementation Details

In this work, all the input images are resized to 64×256 while preserving the aspect ratio. Adam optimizer is adopted to optimize the proposed network with an initial learning rate of 1e-4 and a batch size of 32. The learning rate is decreased with a decay factor of 0.5 per epoch, and the training will be terminated once the learning rate is below 1e-6. Additionally, the number of LSTM units is set to 256 and the LSTM values are clipped to 10. The maximum length of output sequence is set to 20, including one Start token and one EOS token. Totally, we have 39 character classes in our experiments, *i.e*, 26 alphabets, 10 digitals, 1 Start token, 1 EOS token and 1 special token for other symbols.

C. Evaluation of the Proposed Recognizer

We compare our proposed ReELFA with the state-of-the-art approaches on the aforementioned regular text dataset IIIT5K, low-resolution and noisy text dataset SVT and IC15, as well as the curved text dataset CUTE. Comparison results are listed in Table I.

Since the existing state-of-the-art networks are trained with different data, readers should keep in mind that the SynthText [21] dataset learns its text color palette from word images cropped from IIIT5K. By contrast, another widely used 4 million synthetic text image dataset provided by Jaderberg et al. [26] is generated with a 50k-word lexicon derived from the ICDAR and SVT datasets, and blended with cropped word images randomly sampled from these two datasets. Therefore, the recognition performance on IIIT5K will benefit from the usage of SynthText dataset, while Jaderberg's [26] 4 million training samples will contribute more to the recognition performance on ICDAR and SVT datasets.

TABLE I

RESULTS OBTAINED WITH DIFFERENT METHODS ON VARIOUS DATASETS. 'IIIT5K_NONE' INDICATES THAT NO LEXICON IS USED, WHILE 'IIIT5K_50' AND 'IIIT5K_1K' MEANS LEXICON WITH 50 AND 1K WORDS ARE USED RESPECTIVELY. 'OURS_NOEL' AND 'OURS_NOFA' REPRESENT OUR MODEL WITHOUT THE ENCODED LOCATION AND FOCUSED ATTENTION RESPECTIVELY. '*' MEANS THE WORD IMAGES CONTAINING NON-ALPHANUMERIC CHARACTERS ARE REMOVED FROM THE TEST DATASET.

Methods	IIIT5K_None	IIIT5K_50	IIIT5K_1k	SVT	CUTE	IC15
FAN [2]	87.4	99.3	97.5	**85.9**	63.9	66.2
AON [3]	87.0	*99.6*	*98.1*	82.8	76.8	*68.2*
CRNN [4]	78.2	97.6	94.4	80.8	-	-
(Gao et al.)* [1]	83.6	99.1	97.2	*83.9*	-	-
(Gao et al.)* [19]	81.8	99.1	97.9	82.7	-	-
RARE [5]	81.9	96.2	93.8	81.9	59.2	-
STN-OCR* [16]	86.0	-	-	79.8	-	-
SqueezedText(binary) [17]	86.6	96.9	94.3	-	-	-
SqueezedText(full-precision) [17]	87.0	97.0	94.1	-	-	-
R^2AM [15]	78.4	96.8	94.4	80.7	-	-
CA-FCN [9]	**92.0**	**99.8**	**98.9**	82.1	78.1	-
Ours_noEL	87.8	99.3	*98.1*	78.2	75.7	66.6
Ours_noFA	89.8	99.2	97.9	79.8	*81.6*	66.9
ReELFA (proposed)	*90.9*	99.2	*98.1*	82.7	**82.3**	**68.5**

Comparison with Attention-LSTM-based Models: Among the methods listed in Table I, RARE [5], AON [3] and FAN [2] use the combination of bi-LSTM and attention mechanism in the sequence transcription module. FAN and AON are more recent works than RARE, while RARE and AON are specially designed for irregular text recognition. Additionally, a focusing network is designed in FAN to tackle the problem of 'attention drift'. Moreover, our ReELFA and RARE [5] are trained with only SynthText dataset, while FAN and AON are trained with both SynthText dataset and Jaderberg's dataset [26].

From Table I we can see that RARE [5] is significantly surpassed by FAN [2], AON [3] and our proposed ReELFA on all datasets, and FAN [2] achieves the best performance on the SVT dataset. However, on the regular and curved text datasets IIIT5K and CUTE, our proposed ReELFA achieves the best performance, even without assistance from Jaderberg's dataset [26]. Especially, on the CUTE dataset, we get an accuracy of 82.3%, which is 17.4% and 5.5% higher than FAN and AON, respectively. Therefore, our proposed ReELFA is more robust to curved text recognition. As for the IC15 dataset, AON [3] has obtained the best performance of 68.2%, which is slightly better than our 67.6%. But it is notable that FAN and AON has taken the advantage of the prior knowledge of ICDAR and SVT datasets by leveraging Jaderberg's 4 million samples [26].

Comparison with Other Models: Methods without using attention-LSTM also achieve promising performance in the field of scene text recognition, as shown in Table I. In these methods, R^2AM [15], CRNN [4], CA-FCN [9] and both Gao's methods [1], [19] are trained with SynthText dataset, while SqueezedText [17] and STN-OCR [16] are trained with text images generated by new rending engines.

Apparently, CA-FCN [9] and our proposed ReELFA are on the first and second places on the IIIT5K dataset with accuracies of 92.0% and 90.9%, respectively, which outperform other methods significantly. For the low-resolution and noisy dataset SVT, even though Gao et al. [1] reported a higher accuracy of 83.9%, we cannot say their model is more robust than CA-FCN [9] and ours because their model is evaluated on an incomplete dataset, where word images containing non-alphanumeric characters or with less than three characters are removed. Finally, on the challenging curved text dataset CUTE, our ReELFA achieves the best performance of 82.3%, which is 4.2% higher than CA-FCN [9].

The Importance of EL and FA: To highlight the importance of our proposed encoded location and focused attention modules, we also conduct ablation experiments on two baseline models. The first one named 'Ours_noEL' in Table I is the version without encoded location module and the second one named 'Ours_noFA' is the version without focused attention module. The rest of these two baseline models' configurations are just the same as our ReELFA.

From Table I, we can see that when the one-hot encoded location module is removed, the accuracies on IIIT5K, SVT, CUTE and IC15 datasets have decreased by 3.1% (from 90.9% to 87.8%), 4.5% (from 82.7% to 78.2%), 6.6% (from 82.3% to 75.7%) and 1.9% (from 68.5% to 66.6%), respectively. The significant performance degradation evidences the importance of spatial correlation information to scene text recognition, especially to curved text recognition, and the effectiveness of our proposed strategy.

Moreover, when the attention focusing module is removed, the performance on IIIT5K, SVT, CUTE and IC15 datasets has dropped by 1.1% (from 90.9% to 89.8%), 2.9% (from 82.7% to 79.8%), 0.7% (from 82.3% to 81.6%) and 1.6% (from 68.5% to 66.9), respectively. Although the performance gap between 'Ours_noFA' and our proposed ReELFA is not as large as that between 'Ours_noEL' and ReELFA, the recognition accuracies on both regular and irregular texts are improved to certain degrees when the focused attention module is deployed. Therefore, the current attention-based models do suffer from the 'attention drift' problem, which can be alleviated by focusing attentions on the centers of characters.

V. Conclusion

LSTM and attention mechanism have been widely used in scene text recognition. However, existing LSTM-based models have often neglected the spatial correlation information of 2D text images, and the attention-based models suffer from the 'attention drift' problem. In this paper, we have proposed a focused attention module and an encoded location module to tackle these problems. Our proposed model, named as ReELFA, has been evaluated on both regular and irregular datasets, *i.e.*, IIIT5K, SVT, IC15 and CUTE. The experimental results have demonstrated that the proposed recognizer is able to achieve comparable performance on the regular, low-resolution and noisy text datasets, and outperforms the state-of-the-art approaches significantly on the more challenging curved text dataset.

Acknowledgment

This work was supported by China Scholarship Council (No. 201706140138) and Shanghai Natural Science Foundation (No. 19ZR1415900).

References

[1] Y. Gao, Y. Chen, J. Wang, M. Tang, and H. Lu, "Dense chained attention network for scene text recognition," in *International Conference on Image Processing*, 2018, pp. 679–683.

[2] Z. Cheng, F. Bai, Y. Xu, and G. Zheng, "Focusing attention: towards accurate text recognition in natural images," in *IEEE International Conference on Computer Vision*, 2017, pp. 5086–5094.

[3] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "Aon: towards arbitrarily-oriented text recognition," in *International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5571–5579.

[4] 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 Transaction on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[5] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168–4176.

[6] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 5238–5246.

[7] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: fast oriented text spotting with a unified network," *CoRR*, vol. arXiv preprint arXiv: 1801.01671v2, 2018.

[8] Z. Wojna, A. Gorban, D. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, "Attention-based extraction of structured information from street view imagery," in *International Conference on Document Analysis and Recognition*, 2017, pp. 844–850.

[9] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, and X. Bai, "Scene text recognition from two-dimensional perspective," in *AAAI*, 2019.

[10] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," in *International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2961–2968.

[11] T. Novikova, O. Barinova, P. Kohli, and V. Lempitsky, "Large-lexicon attribute consistent text recognition in natural images," in *ECCV*, 2012, pp. 752–765.

[12] Q. Ye and D. Doermann, "Text detection and recognition in imagery: a survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1480–1500, 2015.

[13] M. Jaderberg, A.Vedaldi, and A. Zisserman, "Deep features for text spotting," in *ECCV*, 2014.

[14] M. Haderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, pp. 1–20, 2016.

[15] C. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for ocr in the wild," in *International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2231–2239.

[16] C. Bartz, H. Yang, and C. Meinel, "Stn-ocr: a single neural network for text detection and recognition," *CoRR*, vol. arXiv preprint arXiv: 1707.08831v1, 2017.

[17] Z. Liu, Y. Li, F. Ren, W. Goh, and H. Yu, "Squeezedtext: a real-time scene text recognition by binary convolutional encoder-decoder network," in *AAAI*, 2018, pp. 7194–7201.

[18] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[19] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Reading scene text with attention convolutional sequence modeling," *CoRR*, vol. arXiv preprint arXiv: 1709.04303v1, 2017.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[21] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localization in natural images," in *International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324.

[22] A. Mishra, K. Alahari, and C. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2687–2694.

[23] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *IEEE International Conference on Computer Vision*, 2011, pp. 1457–1464.

[24] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura *et al.*, "Icdar 2015 competition on robust reading," in *International Conference on Document Analysis and Recognition*, 2015, pp. 1156–1160.

[25] A. Risnumawan, P. Shivakumara, C. Chan, and C. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, pp. 8027–8048, 2014.

[26] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *arXiv preprint arXiv:1412.1842*, 2014.