DENSE CHAINED ATTENTION NETWORK FOR SCENE TEXT RECOGNITION

Yunze Gao, Yingying Chen, Jinqiao Wang, Ming Tang and Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190
University of Chinese Academy of Sciences, Beijing, China, 100049
{yunze.gao, yingying.chen, jqwang, tangm, luhq}@nlpr.ia.ac.cn

ABSTRACT

Reading text in the wild is a challenging task in computer vision. Scene text suffers from various background noise, including shadow, irrelevant symbols and background texture. In order to reduce the disturbance of background noise, we propose a dense chained attention network with stacked attention modules for scene text recognition. Each attention module learns the attention map that is adapted to corresponding features to enhance the foreground text and suppress the background noise. Besides, the attention branch is designed with the convolution-deconvolution structure which rapidly captures global information to guide the discriminative feature selection. We stack multiple attention modules to gradually refine the attention maps and capture both the low-level appearance feature and the high-level semantic information. Extensive experiments on the standard benchmarks, the Street View Text, IIIT5K, and ICDAR datasets validate the superiority of the proposed method. The dense chained attention network achieves state-of-the-art or highly competitive recognition performance.

***Index Terms*—** text recognition, attention, convolution-deconvolution

1. INTRODUCTION

Reading Scene text is to recognize the text in natural images, and has been receiving considerable attention and playing an important role in a variety of computer vision tasks. Reading text in the wild can extract rich semantic information that is highly relevant to scene or object and therefore has been widely applied in street sign reading in the driverless vehicle, automatic license plate recognition, assistant technologies for the blind, robot navigation, scene understanding and image retrieval. However, suffering from various appearance, distortion, low resolution, blurring and disturbance of background noise, text recognition in unconstrained environment is still a challenging problem.

Traditional methods [1, 2] recognized scene text by first detecting individual character and then recognizing each cropped character with convolutional neural network. However, a large amount of inter-character and intra-character confusion will reduce the performance of the entire recognition network greatly. Therefore, these approaches heavily depend on an accurate character detector. Recently, some approaches adopted an end-to-end framework for scene text recognition, without detecting characters. Jaderberg *et al.* [3] formulated the scene text recognition as an image classification problem. Each class corresponds to a word in a pre-defined lexicon composed of around 90k words. However, it is difficult for this method to generalize to other situations with huge number of classes, due to the oversized pre-defined dictionary and the requirement for large amount of training samples.

Recent studies [4, 5, 6, 7] regarded scene text recognition as a sequence recognition problem. Shi *et al.* [4] proposed Convolutional Recurrent Neural Network (CRNN) that combined convolutional network and recurrent network to model the spatial dependencies. In [5], a recurrent network with attention mechanism was constructed to decode feature sequence and predict labels recurrently. Shi *et al.* [6] adopted a convolutional-recurrent structure in the encoder to learn the sequential dynamics.

However, for text in natural images, there often exists some disturbances, including shadow, irrelevant symbols and background texture. The scene text with various appearance is often confused by these factors. Existing approaches are incapable of extracting discriminative feature which is robust to various background noise. Considering attention mechanism could selectively focus on the salient regions of the objects and enhance the representation of relevant parts. We design a dense chained attention network to enhance the representation of foreground text and suppress background noise. The attention module learns the soft weights for features, which plays an important guiding role in the process of feature learning. Besides, stacking multiple attention modules gradually refines attention maps. And different attention modules generate corresponding attention weights adaptively. Furthermore, we use a convolution-deconvolution structure in the attention branch to rapidly capture global information within a

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

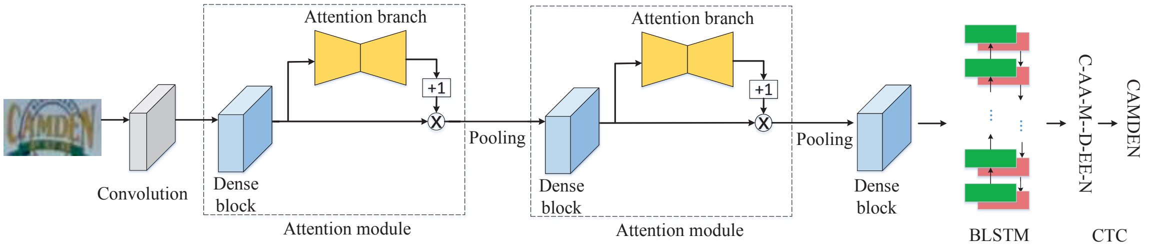

Fig. 1. Overview of the proposed dense chained attention network.

larger receptive field and generate better attention mask.

We evaluate our approach on the challenging benchmarks, Street View Text, IIIT-5K and ICDAR datasets. It is observed that our method not only achieves state-of-the-art or highly competitive performance, but also effectively suppresses the response of background noise while enhancing the representation of foreground text.

2. THE PROPOSED APPROACH

2.1. Network Achitecture

The overview of our dense chained attention network is illustrated in Figure 1. The stacked attention modules extract robust feature representation for the entire image. Then we convert the feature maps into a sequential representation. Specifically, the three-dimensional feature map is cut into 2D maps along its width and then each map is flattened into a vector. In this way, each element in the feature sequence corresponds to a local region of the word image and can be viewed as the feature representation of the region. Then bidirectional long short-term memory (BLSTM) learns the context information and models the sequential dependencies within the feature sequence. Next, the generated sequence is transformed into a sequence of probability distributions. Finally, Connectionist Temporal Classification (CTC) converts the probability distributions into the label.

2.2. Dense Chained Attention Module

To extract the discriminative feature representation, we design an attention module to suppress the response of background noise while enhancing the representation of foreground text. For the attention module, we utilize the dense connectivity [8] in the basic convolutional block to improve the flow of information and gradient propagation. There exist direct connections between all layers in the dense block. Therefore, each layer can get the information from all preceding layers and transmit its message to all subsequent layers. Taking the dense block as input, the attention branch generates the soft attention weights adapted to corresponding features.

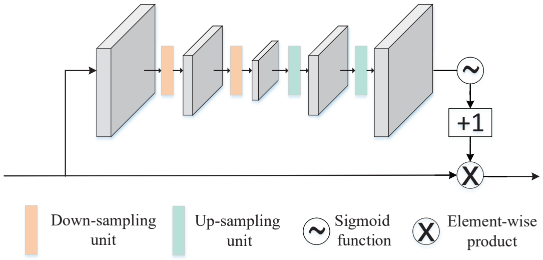

Fig. 2. The structure of attention branch.

As shown in Figure 2, the attention branch is designed with the convolution-deconvolution structure so that the global information can be rapidly captured to guide the discriminative feature selection. Specifically, each down-sampling unit contains a max pooling layer and a convolutional layer. Correspondingly, each up-sampling unit contains a bilinear interpolation layer and a convolutional layer. The feature maps are first down-sampled to the lowest spatial resolution. A series of stacked down-sampling units increase the receptive field rapidly and collect the global information. Then a symmetrical architecture with up-sampling units is applied to recover the resolution to the original size. During up-sampling, the high-level semantic information is expanded to guide corresponding features in each position. Then the values are normalized by a sigmoid function as the attention weights.

Considering the attention weights range from zero to one, the element-wise product between feature maps and attention maps may cause severe degradation of useful information. Especially, repeated element-wise product will cause significantly information loss. Therefore, we add 1 on the values of attention maps to enhance the original features. In this way, the good properties of original features could be maintained. Next, the modified attention maps as soft weights are fused on corresponding feature maps with element-wise product. Thus, the attention maps not only reduces the response of background noise, but also keeps the disciriminative information of original features.

Furthermore, we stack multiple attention modules to gradually refine the attention maps. If the learned attention map cannot correctly focus on the foreground text, the subsequent attention modules can modify the attention-aware features. Besides, different attention modules generate the attention maps adapted to the corresponding features. The low-level attention module mainly concentrates on the appearance including edge, color and texture, while the high-level attention module contains more semantic information. With the attention mechanism, the feature encoder benefits from the noise suppression to obtain a more discriminative representation.

2.3. Sequence Decoding

In the process of sequence decoding, the output sequence $\mathbf{y} = (y_1, y_2, \cdots, y_w)$ by BLSTM is transformed into the label sequence, using the CTC proposed by [9]. Defining L as the set of 36 classes including all English alphanumeric characters, we get the final label space $L^{'} = L \cup \{blank\}$, in which the extra $blank$ denotes the class of observing no character. Given the probability distribution, the conditional probability of the sequence π is

$$p(\pi|\mathbf{y}) = \prod_{t=1}^{w} y_{\pi_t}^{t} \tag{1}$$

where $y_{\pi_t}^{t}$ denotes the probability of emtting label π_t at step t. Then a many-to-one mapping $\mathcal{B}$ merges the repeated continuous labels to a single one and then remove the $blank$ labels. Furthermore, the probability of the final output sequence is formulated as the sum of the conditional probabilities of all π corresponding to it:

$$p(l|\mathbf{y}) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi|\mathbf{y}) \tag{2}$$

Given the training set $\mathcal{D} = \{I_i, l_i\}$, where I_i and l_i represent the word image and the corresponding ground truth label, respectively. The objective function is formulated as the sum of the negative log likelihood of the probabilities for target labels:

$$\mathcal{O} = - \sum_{(I_i, l_i) \in \mathcal{D}} \log p(l_i|\mathbf{y}_i) \tag{3}$$

During inference, for lexicon-free recognition, we emit the label with the highest probability at each step and then use $\mathcal{B}$ to generate the final label. For lexicon-based recognition, we adopt an approximate method by comparing the edit distance between the predicted sequence in the lexicon-free setting and words in the lexicon, then choosing the word with the smallest edit distance as the output label.

3. EXPERIMENT

Several public datasets are used for the evaluation, including Street View Text, IIIT5K, ICDAR 2003 and ICDAR 2013. For training data, our model is trained purely on the synthetic dataset released by [10], without any extra data for finetuning. Following the evaluation protocol in [1], we perform recognition on word images that contain only alphanumeric characters and at least three characters.

3.1. Implementation Details

The dense chained attention network contains two attention modules. Specifically, there are six convolutional layers with filter size 3 and stride 1 in each dense block. And the growth rate, that is the number of output feature maps for each layer, is 18 in all experiments. Besides, 3,2 down-sampling units are used in the two attention branches, respectively. All the convolution are performed with zero padding, ReLU activation and batch normalization. The BLSTM contains 2 layers with 256 units per LSTM. In the process of training and testing, the word images are resized to 32 × 100 with gray scale. We adopt the msra [20] as the weight initialization method. The proposed network is implemented with Tensorflow.

3.2. Attention Module

To validate the effectiveness of the attention modules, we compare the performance of networks with and without attention mechanism. Keep the other parts unchanged, we only remove the attention branches for comparison. Besides, we conduct experiments on networks with different depths to evaluate the generalization. As shown in Table 2, the networks with attention mechanism consistently outperform the networks without attention, which proves the effectiveness of the method. Additionally, we observe that the improvements brought by attention mechanism on ICDAR datasets are not as significant as that on SVT and IIIT5k. ICDAR13 dataset has plenty of images suffering from blurring and non-uniform illumination. This increases the difficulty in focusing on the foreground text. Some incorrect samples are presented in Figure 3. Furthermore, we visualize the attention maps of some examples in Figure 3. As shown in Figure 3, in most cases, the attention maps focus on the foreground text to be recognized and effectively reduce the response of background noise.

3.3. Comparisons with State-of-the-art Methods

We evaluate our method on the above four public datasets and compare it with state-of-the-art algorithms in Table 1. Since Cheng *et al.*[21] uses large additional training datasets with character level annotations, we do not do comparison with the results of [21].

For lexicon-free recognition, our network achieves the state-of-the-art or highly competitive performance. Specifi-

Table 1. Scene text recognition accuracies on the benchmark datasets. “50”, “1000” and “Full” represent the size of lexicon used for constrained recognition, and “None” represents unconstrained recognition. “∗”[3] is not lexicon-free strictly, due to the output sequence is constrained to a 90k dictionary.

Methods	SVT		IIIT5k			IC03			IC13
	50	None	50	1k	None	50	Full	None	None
Wang *et al.*[1]	57.0	-	-	-	-	76.0	62.0	-	-
Mishra *et al.*[11]	73.2	-	64.1	57.5	-	81.8	67.8	-	-
Bissacco *et al.*[2]	90.4	78.0	-	-	-	-	-	-	87.6
Yao *et al.*[12]	75.9	-	80.2	69.3	-	88.5	80.3	-	-
Rodriguez-Serrano *et al.*[13]	70.0	-	76.1	57.4	-	-	-	-	-
Jaderberg *et al.*[14]	86.1	-	-	-	-	96.2	91.5	-	-
Gordo [15]	91.8	-	93.3	86.6	-	-	-	-	-
∗Jaderberg *et al.*[3]	95.4	80.7	97.1	92.7	-	98.7	**98.6**	**93.1**	**90.8**
Jaderberg *et al.*[16]	93.2	71.7	95.5	89.6	-	97.8	97.0	89.6	81.8
Shi *et al.*[4]	97.5	82.7	97.8	95.0	81.2	98.7	98.0	91.9	89.6
Shi *et al.*[6]	95.5	81.9	96.2	93.8	81.9	98.3	96.2	90.1	88.6
Lee *et al.*[5]	96.3	80.7	96.8	94.4	78.4	97.9	97.0	88.7	90.0
Liu *et al.*[17]	95.5	83.6	97.7	94.5	83.3	96.9	95.3	89.9	89.1
He *et al.*[7]	92.0	-	94.0	91.6	-	97.0	94.4	-	-
Yang *et al.*[18]	95.2	-	97.8	96.1	-	97.7	-	-	-
Wang and Hu[19]	96.3	81.5	98.0	95.6	80.8	**98.8**	97.8	91.2	-
Ours	**97.7**	**83.9**	**99.1**	**97.2**	**83.6**	98.6	96.6	91.4	89.5

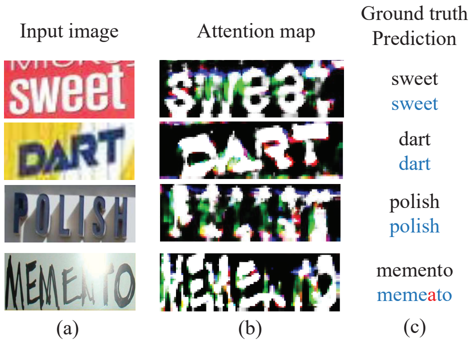

Fig. 3. Visualization of the attention maps and the recognition results. (a) Input images. (b) Attention maps. (c) The ground truth and the recognized results. The blue and red characters represent the correctly and incorrectly recognized characters, respectively.

cally, we obtain the best result on SVT and IIIT5k. It is worth noting that the model in [3] only can deal with the words in its 90k dictionary, which is not lexicon-free recognition strictly. Different from [3], our approach is able to recognize random word strings and is not restricted by a fixed dictionary.

For lexicon-based recognition, our method consistently outperforms other approaches on several benchmarks. The significant improvement validates the effectiveness of our method. Moreover, it is observed that IIIT5k contains plenty of images suffering from background noise, which proves the superiority of our method in suppressing noise. Besides, we behind [3] on ICDAR03 dataset. However, [3] benefits from the pre-defined large dictionary as mentioned before. Therefore, our results are still competitive compared with the state-of-the-arts.

Table 2. Lexicon-free scene text recognition accuracies on standard benchmarks. ”Depth17” and ”Depth23” represent the network with 17 layers and 23 layers, respectively.

Method	SVT	IIIT5k	IC03	IC13
Depth17	81.9	80.4	90.7	89.5
Depth17+attention	83.0	83.0	90.8	88.7
Depth23	82.4	82.9	89.2	89.9
Depth23+attention	83.9	83.6	91.4	89.5

4. CONCLUSION

In this paper, we propose a dense chained attention network for scene text recognition, with stacked attention modules to learn robust representation. The attention mechanism significantly suppresses background noise and improves the performance. Besides, the proposed network can be trained end-to-end with the word level annotations. The extensive experimental results on the benchmarks demonstrate the superiority of our approach compared with the state-of-the-art methods.

5. REFERENCES

[1] Kai Wang, Boris Babenko, and Serge Belongie, "End-to-end scene text recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1457–1464.

[2] Alessandro Bissacco, Mark Cummins, Yuval Netzer, and Hartmut Neven, "Photoocr: Reading text in uncontrolled conditions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.

[3] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[4] Baoguang Shi, 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*, 2016.

[5] Chen-Yu Lee and Simon Osindero, "Recursive recurrent nets with attention modeling for ocr in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2231–2239.

[6] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai, "Robust scene text recognition with automatic rectification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168–4176.

[7] Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang, "Reading scene text in deep convolutional sequences.," in *AAAI*, 2016, pp. 3501–3508.

[8] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, vol. 1, p. 3.

[9] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[10] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.

[11] Anand Mishra, Karteek Alahari, and CV Jawahar, "Scene text recognition using higher order language priors," in *BMVC 2012-23rd British Machine Vision Conference*. BMVA, 2012.

[12] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4042–4049.

[13] Jose A Rodriguez-Serrano, Albert Gordo, and Florent Perronnin, "Label embedding: A frugal baseline for text recognition," *International Journal of Computer Vision*, vol. 113, no. 3, pp. 193–207, 2015.

[14] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman, "Deep features for text spotting," in *European conference on computer vision*. Springer, 2014, pp. 512–528.

[15] Albert Gordo, "Supervised mid-level features for word image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2956–2964.

[16] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep structured output learning for unconstrained text recognition," *arXiv preprint arXiv:1412.5903*, 2014.

[17] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han, "Star-net: A spatial attention residue network for scene text recognition.," in *BMVC*, 2016, vol. 2, p. 7.

[18] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles, "Learning to read irregular text with attention mechanisms," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3280–3286.

[19] Jianfeng Wang and Xiaolin Hu, "Gated recurrent convolution neural network for ocr," in *Advances in Neural Information Processing Systems*, 2017, pp. 334–343.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[21] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5086–5094.