

Gate-based Bidirectional Interactive Decoding Network for Scene Text Recognition

Yunze Gao, Yingying Chen, Jinqiao Wang and Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
No. 95, Zhongguancun East Road, Beijing 100190, China
University of Chinese Academy of Sciences, Beijing 100049, China
{yunze.gao,yingying.chen,jqwang,luhq}@nlpr.ia.ac.cn

ABSTRACT

Scene text recognition has attracted rapidly increasing attention from the research community. Recent dominant approaches typically follow an attention-based encoder-decoder framework that uses a unidirectional decoder to perform decoding in a left-to-right manner, but ignoring equally important right-to-left grammar information. In this paper, we propose a novel Gate-based Bidirectional Interactive Decoding Network (GBIDN) for scene text recognition. Firstly, the backward decoder performs decoding from right to left and generates the reverse language context. After that, the forward decoder simultaneously utilizes the visual context from image encoder and the reverse language context from backward decoder through two attention modules. In this way, the bidirectional decoders perform effective interaction to fully fuse the bidirectional grammar information and further improve the decoding quality. Besides, in order to relieve the adverse effect of noises, we devise a gated context mechanism to adaptively make use of the visual context and reverse language context. Extensive experiments on various challenging benchmarks demonstrate the effectiveness of our method.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**; *Computer vision problems*; Artificial intelligence.

KEYWORDS

Text recognition; Bidirectional interactive decoder; Gated mechanism

ACM Reference Format:

Yunze Gao, Yingying Chen, Jinqiao Wang and Hanqing Lu. 2019. Gate-based Bidirectional Interactive Decoding Network for Scene Text Recognition. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*, November 3–7, 2019, Beijing, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3357384.3358135>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '19, November 3–7, 2019, Beijing, China

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6976-3/19/11...\$15.00

<https://doi.org/10.1145/3357384.3358135>





Input Image	Ground Truth	Forward Prediction	Backward Prediction
	marlboro	marleoro	mrailboro
	subs	tabs	sube
	leve	lete	have
	visible	visiale	visible

Figure 1: Recognition examples of attention-based encoder-decoder models with different decoding directions. Blue and red characters are correctly and mistakenly recognized characters, respectively.

3–7, 2019, Beijing, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3357384.3358135>

1 INTRODUCTION

Scene text recognition has been an active research topic in computer vision due to the practical value in real-world applications. With the advent of deep neural networks, there has been substantial progress on this topic within the last few years. Current approaches [1, 2, 6, 9, 10] typically follow an attention-based encoder-decoder framework, which consists of a Convolutional Neural Network (CNN)/Long-Short Term Memory (LSTM) based image encoder and an attentional Recurrent Neural Network (RNN) based word decoder.

In general, most approaches [1, 2, 6, 9] adopt a unidirectional decoder that generates the word sequence in the left-to-right direction. At each step, the decoder predicts the next character based on the previously generated characters. Although RNN has the ability to capture the language context during decoding, the reverse language context still cannot be exploited. Besides, once errors occur in previous predictions, the quality of subsequent predictions would be undermined due to the negative impact of the noises. Intuitively, the reverse language context could provide complementary information, which is crucial for word predictions. As observed in Fig. 1, some characters are recognized incorrectly, but can be accurately predicted by another decoder in the opposite direction. Consequently, it is important to investigate how to effectively incorporate reverse language context into the decoder to improve recognition performance.

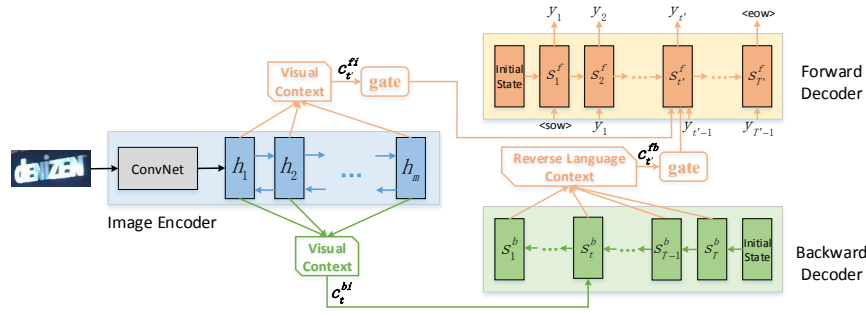


Figure 2: Overview of our Gate-based Bidirectional Interactive Decoding Network. “sow” and “eow” represent the “start-of-word” and “end-of-word” tokens, respectively.

In this paper, we propose a novel Gate-based Bidirectional Interactive Decoding Network for scene text recognition. Specifically, the forward decoder can make full use of the reverse language context through an effective interaction with the backward decoder. First, the backward decoder performs decoding as the standard decoder but in the right-to-left manner, where the generated hidden states contain the reverse language context. Then, the forward decoder produces the character sequence in the left-to-right direction, where two attention modules are simultaneously utilized to capture the visual context and reverse language context, respectively. In this way, the forward decoder is able to fully exploit the complementary backward grammar information. Thus, the rich bidirectional language contexts work together for more accurate predictions. Besides, in order to avoid the effect of noises, we design a flexible gated context mechanism to decide how much information the forward decoder gets from the visual context and the reverse language context.

The main contributions are summarized as follows:

- (1) We propose a novel Gate-based Bidirectional Interactive Decoding Network, which fully leverages the bidirectional language contexts through the interaction between decoders with opposite directions.
- (2) We design a gated context mechanism to adaptively exploit the visual context and reverse language context, relieving the negative impact of noises.
- (3) Extensive experiments on challenging datasets demonstrate the superiority of our approach over existing methods.

2 METHODS

The overview of our Gate-based Bidirectional Interactive Decoding Network is shown in Fig. 2. First, we combine CNN and BLSTM to encode the input image into a sequence of feature vectors, denoted as $h = (h_1, h_2, \dots, h_m)$. Then the backward decoder performs decoding in the right-to-left direction. Based on the visual features and backward hidden states, the forward decoder generates the character sequence through two attention modules and gated context mechanism.

2.1 Bidirectional Interactive Decoder

Given the sequential features generated by image encoder, the backward decoder first performs decoding with an attention module in a right-to-left manner. Specifically, at step t , the decoder dynamically weights the image features as follows:

$$e_{t,j}^{bi} = w_{bi}^T \tanh(W_{bi} s_{t+1}^b + U_{bi} h_j + b_{bi}) \quad (1)$$

$$\alpha_{t,j}^{bi} = \frac{\exp(e_{t,j}^{bi})}{\sum_{j=1}^{|h|} \exp(e_{t,j}^{bi})} \quad (2)$$

$$c_t^{bi} = \sum_{j=1}^{|h|} \alpha_{t,j}^{bi} h_j \quad (3)$$

where c_t^{bi} is the weighted visual context vector, s_{t+1}^b is the previous hidden state of the backward decoder, W_{bi} , U_{bi} , w_{bi} and b_{bi} are the learnable parameters. Then RNN updates the hidden state as follows:

$$s_t^b = RNN(\overleftarrow{y}_{t+1}, s_{t+1}^b, c_t^{bi}) \quad (4)$$

where \overleftarrow{y}_{t+1} is the previous predicted character. After that, we can obtain the probability distribution over label space and the conditional probability of the reverse decoding sequence:

$$p(\overleftarrow{y}_t | h, \overleftarrow{y}_{t+1}) = \text{softmax}(V_b^T s_t^b) \quad (5)$$

$$p(\overleftarrow{y} | h) = \sum_{t=1}^T \log p(\overleftarrow{y}_t | h, \overleftarrow{y}_{t+1}) \quad (6)$$

The hidden state sequence produced by the backward decoder serves as the input of the subsequent forward decoder, which contains the reverse language context. There are two reasons why we choose the hidden state sequence instead of the character sequence. The first is that sampling character is not differentiable, and the second is that hidden states contain more information and are able to alleviate the negative impacts of prediction errors.

Then the forward decoder conducts decoding under the guidance of the visual context and reverse language context. Specifically, the visual context is captured through weighting the image features with attention mechanism. At t' -th step,

$$e_{t',j}^{fi} = w_{fi}^T \tanh(W_{fi} s_{t'-1}^f + U_{fi} h_j + b_{fi}) \quad (7)$$

Table 1: Ablation studies by changing the structure of decoder and the hyper-parameter of objective function.

Model Configuration			λ	SVT	IIIT5k	IC03	IC13	IC15	SVTP
Backward Decoder	Interactive	Gated Context							
				85.5	89.7	92.1	88.7	69.4	71.7
✓			0.5	85.8	91.1	92.6	90.1	70.1	73.4
✓	✓		0.5	86.6	91.5	93.4	90.1	72.4	74.6
✓	✓	✓	0.5	87.0	92.1	94.0	92.0	73.6	76.1
✓	✓	✓	0.9	84.7	89.9	93.0	89.0	70.4	72.0
✓	✓	✓	0.7	87.2	91.6	93.5	91.4	72.3	75.0
✓	✓	✓	0.3	87.6	92.6	93.7	90.8	74.2	77.2
✓	✓	✓	0.1	87.5	92.7	93.1	91.4	73.0	75.1

$$\alpha_{t',j}^{fi} = \frac{\exp(e_{t',j}^{fi})}{\sum_{j=1}^{|h|} \exp(e_{t',j}^{fi})} \quad (8)$$

$$c_{t'}^{fi} = \sum_{j=1}^{|h|} \alpha_{t',j}^{fi} h_j \quad (9)$$

where $c_{t'}^{fi}$, $s_{t'-1}^f$ are the visual context vector and the previous hidden state of the forward decoder separately, W_{fi} , U_{fi} , w_{fi} and b_{fi} are the learnable parameters. Likewise, the reverse language context is utilized by another attention module:

$$e_{t',j}^{fb} = w_{fb}^T \tanh(W_{fb} s_{t'-1}^f + U_{fb} s_j^b + b_{fb}) \quad (10)$$

$$\alpha_{t',j}^{fb} = \frac{\exp(e_{t',j}^{fb})}{\sum_{j=1}^{|s^b|} \exp(e_{t',j}^{fb})} \quad (11)$$

$$c_{t'}^{fb} = \sum_{j=1}^{|s^b|} \alpha_{t',j}^{fb} s_j^b \quad (12)$$

in which $c_{t'}^{fb}$ is the reverse language context vector for forward decoder, W_{fb} , U_{fb} , w_{fb} and b_{fb} are the learnable parameters.

Then the hidden state of the forward decoder is updated:

$$s_{t'}^f = RNN(y_{t'-1}, s_{t'-1}^f, c_{t'}^{fi}, c_{t'}^{fb}) \quad (13)$$

where $y_{t'-1}$ is the previous predicted character. Besides, the conditional probability of the forward decoding sequence is:

$$p(y_t|h, y_{t-1}) = \text{softmax}(V_f^T s_{t'}^f) \quad (14)$$

$$p(y|h) = \sum_{t'=1}^{T'} \log p(y_{t'}|h, y_{t'-1}) \quad (15)$$

In this way, the bidirectional language contexts are effectively fused to improve the decoding quality.

2.2 Gated Context Mechanism

As shown in Eq.13, the visual context and the reverse language context are simultaneously used as the input of forward decoder. We observe that the visual information may be confusing in the case of occlusions, blurring and heavy touching. On the other hand, there may exist prediction errors in the reverse language context. To relieve the effects of mistakes, we need to filter the noises. Motivated by the above analyses, we elaborate a flexible gated context mechanism to decide how much information the forward decoder wants to get from the visual context and the reverse language context.

Concretely, the visual information and the reverse language information are separately controlled by two adaptive gates. Specifically, the gates are generated as follows:

$$G_I = \sigma(W_{gi} s_{t'-1}^f + U_{gi} c_{t'}^{fi} + V_{gi} c_{t'}^{fb} + b_{gi}) \quad (16)$$

$$G_B = \sigma(W_{gb} s_{t'-1}^f + U_{gb} c_{t'}^{fi} + V_{gb} c_{t'}^{fb} + b_{gb}) \quad (17)$$

where G_I and G_B are the gate vectors for visual and reverse language context separately, W_{gi} , U_{gi} , V_{gi} , b_{gi} , W_{gb} , U_{gb} , V_{gb} , b_{gb} are the learnable parameters, σ is the sigmoid activation.

Then we model the new adaptive context as a mixture of the visual context and the reverse language context:

$$c_{t'} = G_I \odot \tanh(c_{t'}^{fi}) + G_B \odot \tanh(c_{t'}^{fb}) \quad (18)$$

where \odot denotes element-wise product. Then we use the gated context to replace the original inputs and rewrite Eq.13:

$$s_{t'}^f = RNN(y_{t'-1}, s_{t'-1}^f, c_{t'}) \quad (19)$$

Thus, the visual and reverse language context can be adaptively modulated, and the impact of noise is effectively relieved.

2.3 Model Training and Testing

Denote the training set as \mathcal{X} , which consists of pairs of image \mathcal{I} and label \mathcal{Y} . The objective function is formulated as:

$$\mathcal{L} = - \sum_{(\mathcal{I}, \mathcal{Y}) \in \mathcal{X}} (\lambda p(\mathcal{Y}|h) + (1 - \lambda) p(\tilde{\mathcal{Y}}|h)) \quad (20)$$

where $\tilde{\mathcal{Y}}$ is obtained by inverting \mathcal{Y} , and λ is a tunable parameter in the range $[0, 1]$.

During testing, we choose the generated word of forward decoder. For lexicon-based recognition, we select the nearest lexicon word to replace predicted word under edit distance.

3 EXPERIMENTS

3.1 Datasets and Experimental Settings

Several public datasets are used for evaluation, including Street View Text, IIIT5K, ICDAR 2003, ICDAR 2013, ICDAR 2015 and Street View Text Perspective. The synthetic datasets are used for training, including the 8-million Synth90k [4] and the 7-million SynthText [3].

There are six convolutional blocks in the image encoder. The detailed configurations are $[3, 32] \times 1$, $[1, 32; 3, 32] \times 3$, $[1, 64; 3, 64] \times 4$, $[1, 128; 3, 128] \times 6$, $[1, 256; 3, 256] \times 6$, $[1, 512; 3, 512] \times 3$. The residual connection is used

Table 2: Scene text recognition accuracies on standard benchmarks. “50”, “1000” and “Full” represent the size of lexicon used for lexicon-based recognition, and “None” represents lexicon-free recognition. “*” represents the methods trained with both word-level and character-level annotations.

Methods	SVT		IIIT5k			IC03			IC13	IC15	SVTP
	50	None	50	1k	None	50	Full	None	None	None	None
Jaderberg <i>et al.</i> [5]	95.4	80.7	97.1	92.7	-	98.7	98.6	93.1	90.8	-	-
Shi <i>et al.</i> [8]	97.5	82.7	97.8	95.0	81.2	98.7	98.0	91.9	89.6	-	66.8
Liu <i>et al.</i> [7]	-	87.6	-	-	83.6	-	-	93.3	93.7	-	73.5
Lee and Osindero [6]	96.3	80.7	96.8	94.4	78.4	97.9	97.0	88.7	90.0	-	-
Shi <i>et al.</i> [9]	95.5	81.9	96.2	93.8	81.9	98.3	96.2	90.1	88.6	-	71.8
Cheng <i>et al.</i> [2]	96.0	82.8	99.6	98.1	87.0	98.5	97.1	91.5	-	68.2	73.0
*Cheng <i>et al.</i> [1]	97.1	85.9	99.3	97.5	87.4	99.2	97.3	94.2	93.3	66.2	71.5
*Yang <i>et al.</i> [10]	95.2	-	97.8	96.1	-	97.7	-	-	-	-	75.8
GBIDN (Ours)	97.2	87.6	99.5	98.6	92.6	99.0	97.9	93.7	90.8	74.2	77.2
GBIDN+stn (Ours)	97.7	89.5	99.6	98.7	93.6	99.4	98.3	94.2	91.9	77.1	80.6
GBIDN+90k Dict (Ours)	-	90.1	-	-	-	-	-	95.8	93.6	-	82.3

except the first one. Downsampling is performed by 2×2 stride convolutions in the second and third blocks. The stride is changed to 2×1 in the last three blocks. Following the convolutional layers is two-layer BLSTM with 256 hidden units per LSTM. Both the backward and forward decoder adopt one-layer LSTM with 256 hidden units. The input images are resized to 32×100 . The label space contains 94 classes, including 10 digits, 52 case sensitive letters, and 32 punctuations. The model takes 24.6ms recognizing an image.

3.2 Ablation Studies

The ablation studies are shown in Table 1. As observed, the introduction of backward decoder significantly improves the performance. Comparing with the independent bidirectional decoders with two parallel opposite decoding branches that do not interact, the bidirectional interactive decoder performs better. Moreover, the gated context mechanism further improves the recognition performance. Furthermore, we should conduct relatively more constraints on the backward decoder.

3.3 Comparison with Existing Methods

We compare our method with the state-of-the-arts as in Table 2. Our method performs better than most existing methods. In particular, we achieve significant improvements over the methods with unidirectional decoder [1, 2, 6, 9]. We also use an STN module following [9] to fairly compare with other methods specially for irregular text. We substantially outperforms [2, 9, 10] on all benchmarks. As [5, 7] benefited from a 90K dictionary, we also report results using the same dictionary to post-process the predictions on SVT, IC03 and IC13. Particularly, with the same dictionary, we outperform [7] by 2.5 percents on SVT and 2.5 percents on IC03.

4 CONCLUSION

In this paper, we propose a Gated Bidirectional Interactive Decoding Network for scene text recognition. We introduce a bidirectional interactive decoder into the traditional attention-based encoder-decoder framework. Besides, we design a gated

context mechanism to adaptively modulate the visual and reverse language context. Extensive experiments validate the effectiveness of our method.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China 61772527, 61806200.

REFERENCES

- [1] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE International Conference on Computer Vision*. 5076–5084.
- [2] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. 2018. AON: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5571–5579.
- [3] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2315–2324.
- [4] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227* (2014).
- [5] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision* 116, 1 (2016), 1–20.
- [6] Chen-Yu Lee and Simon Osindero. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2231–2239.
- [7] Wei Liu, Chaofeng Chen, and Kwan-Yee K Wong. 2018. Char-net: A character-aware neural network for distorted scene text recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [8] Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence* 39, 11 (2017), 2298–2304.
- [9] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4168–4176.
- [10] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. 2017. Learning to Read Irregular Text with Attention Mechanisms. In *IJCAI*. 3280–3286.