ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096434

A3S: ADVERSARIAL LEARNING OF SEMANTIC REPRESENTATIONS FOR SCENE-TEXT SPOTTING

Masato Fujitake

Fast Accounting Co., Ltd.

fujitake@fastaccounting.co.jp

ABSTRACT

Scene-text spotting is a task that predicts a text area on natural scene images and recognizes its text characters simultaneously. It has attracted much attention in recent years due to its wide applications. Existing research has mainly focused on improving text region detection, not text recognition. Thus, while detection accuracy is improved, the end-to-end accuracy is insufficient. Texts in natural scene images tend to not be a random string of characters but a meaningful string of characters, a word. Therefore, we propose adversarial learning of semantic representations for scene text spotting (A3S) to improve end-to-end accuracy, including text recognition. A3S simultaneously predicts semantic features in the detected text area instead of only performing text recognition based on existing visual features. Experimental results on publicly available datasets show that the proposed method achieves better accuracy than other methods.

Index Terms— Scene-text spotting, Document analysis, Deep learning

1. INTRODUCTION

Scene-text spotting is a task for detecting and recognizing texts in natural scene images. It has recently attracted attention for its usefulness in real-world applications such as document understanding and automated driving [1–4]. The task consists of two main processes: first, to predict the coordinate positions of text areas in natural scene images, and second, to recognize texts in the detected regions. Accurate detection of text regions in natural scene images is difficult due to various variations, such as curved text strings and complex layouts. For this reason, many works have been proposed in recent years that have greatly improved the accuracy of text region detection [3, 5, 6]. However, end-to-end accuracy has not been sufficiently improved due to errors in text recognition. To recognize text in scene images, most of the methods proposed in recent years rely on only visual information in the detected text area. However, such an approach is vulnerable to change in various fonts, styles, colors, and shapes, which leads to misspelled recognized texts.

In natural scene images, many texts have semantic information like meaningful words and sentences, and random character strings are rare. Therefore, we considered using semantic information from language models can help to reduce text recognition errors. In this paper, we propose a novel method, **A**dversarial learning of **S**emantic representations for **S**cene-text **S**potting, A3S. In text recognition, A3S simultaneously predicts the text area's semantic features instead of directly predicting text from the visual features obtained from images. The predicted semantic features of the text are trained to match the ones of a pre-trained language model. Similar studies that utilize visual and semantic features have densely matched in the same space [7]. Still, since these features are strictly different, we propose to match them flexibly through adversarial learning. This approach improves end-to-end accuracy by predicting features that consider semantic information without relying excessively on visual information.

Our method improves the end-to-end accuracy in scene-text spotting. Furthermore, it achieves state-of-the-art accuracy on several public datasets. In summary, the contributions of this paper are summarized as follows:

- We propose A3S, an adversarial learning method for scene text spotting. To the best of our knowledge, this is the first work that jointly leverages semantic representations for scene-text spotting.
- With a simple but effective approach, we improve 6.9% accuracy on the CTW1500 dataset. We also archive state-of-the-art accuracy on several benchmarks.

2. RELATED WORKS

Scene-Text Spotting. Scene-text spotting requires text detection [8] and recognition [9] simultaneously. Two-stage approaches [10] are proposed, which individually develop detection and recognition modules and join them during inference. However, multi-steps may require detailed tuning, leading to sub-optimal performance and time consumption. On the other hand, recent literature focuses on end-to-end methods [1, 5, 11, 12], which train both modules simultaneously.

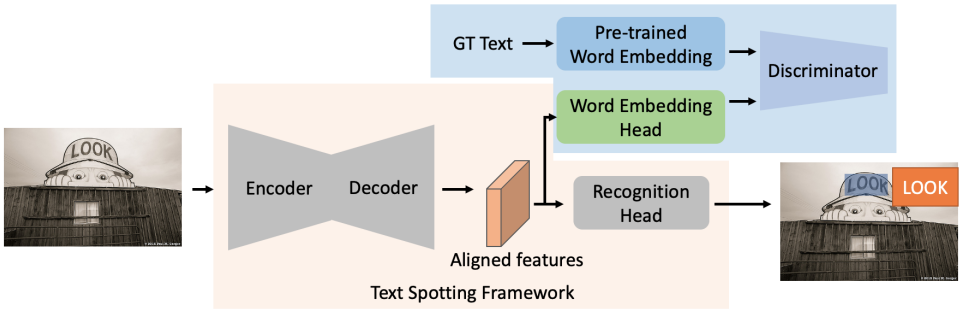

Fig. 1: The architecture of the proposed method, adversarial learning of semantic representations for text spotting (A3S). It consists of the text spotting framework [5], the word embedding head, the pre-trained word embedding, and the discriminator. The word embedding head predicts the semantic features from the aligned detection candidate features obtained inside the text spotting framework. The discriminator optimizes to close the predicted and semantic features from the pre-trained word embedding.

Some methods [6,13] proposed special RoI operations to sample features in the text area. TESTER [1] proposes query-based decoders to remove complex RoI operations. Mask TextSpotter series detect and recognize text instances of arbitrary shapes by segmenting the text regions [3]. ABCNet series [5] introduces parametric Bezier curve representations for curved texts. SwinTextSpotter [14] proposes a recognition conversion mechanism to explicitly guide text localization. Although they improve detection accuracy by focusing on detecting arbitrary-shaped text areas, the end-to-end accuracy is insufficient. In contrast, our work focuses on improving text recognition on scene-text spotting. Although GLASS [15] proposed an attention mechanism recently to fuse visual global and local information for text recognition accuracy, it only relies on vulnerable visual features. Our method exploits the semantic representations of the text to improve recognition.

Visual-Semantic embedding. Semantic information provides different information from the visual one obtained from images. Therefore, some studies proposed using both representations simultaneously. Methods have been proposed for embedding them in the same Euclidean space [7] or estimating one feature from the other and fusing them together [16]. However, they utilize different representations simultaneously and directly, which may interfere with each other. In contrast, we enable flexible learning by introducing adversarial learning.

3. METHOD

The overall architecture of A3S is presented in Fig 1, which consists of four components: (1) a baseline text spotting framework [5]; (2) word embedding head; (3) pre-trained word embedding; and (4) a discriminator for adversarial learning. In the following, we briefly explain the text spotting framework and introduce the details of the proposed modules and optimization.

Table 1: Structure of Word Embedding Head. For all convolutional layers, the padding size is restricted to 1. n, c, h, and w represent the batch size, the channel size, and the height and width of the outputted features, respectively. d means the output dimension size of the pre-trained word embedding.

Layers	Parameters (kernel size, stride)	Output Size (n, c, h, w)
conv. layers w/ ReLU $\times$ 2	(3,1)	(n, 256, h, w)
average pool for h	–	(n, 256, 1, w)
fc layers w/ ReLU $\times$ 2	–	(n, d)

3.1. Text Spotting Framework

We employ ABCNet v2 [5] as the baseline text spotting framework, as shown in the orange-colored area in Fig 1. It utilizes a single-shot, anchor-free convolutional neural network as the detection framework and the lightweight attention mechanism for recognition. Based on the predicted detection results, the features are passed to the text recognition head through BezierAlign [5] to precisely align the visual features of each arbitrarily shaped text area. The framework adapts end-to-end optimization for detection and recognition.

3.2. Word Embedding Head

We propose a word embedding head to enable feature prediction that considers semantic information in a text-spotting framework. It is a simple neural network shown in Table 1 that estimates semantic features from visual features obtained by detection. It is optimized through adversarial learning to close the output of the features by the pre-trained word embedding. Word Embedding Head and Recognition Head share the detected visual features to enable text recognition with semantic representation. We set the output dimension of the head identical to the pre-trained word embedding.

3.3. Pre-trained Word Embedding

To generate the ground truth semantic information of text in images, we utilize a pre-trained language model, BERT [17]. It is a transformer-based language model [18], which leverages an attention mechanism that learns contextual relations between words in a text. The ground truth text to be detected is encoded through BERT, which outputs a semantic vector to the subsequent discriminator. We use the last hidden states of BERT as the semantic vector. If the text has multiple words, the output vector is mean pooled. The pre-trained word embedding's weights are fixed.

3.4. Adversarial learning

By adversarial learning [19], we optimize the word embedding head with the predicted semantic features and the ones from pre-trained language models. We train the head by training a discriminator that identifies whether features are the predicted ones or the ones from pre-trained word embedding. A two-class classification using binary cross-entropy loss function is performed with the predicted features as 0 and the ones of the pre-trained word embedding as 1. A two-class classifier with three fully-connected layers was used as the discriminator.

3.5. Optimization

The training method follows the text spotting framework [5], except for adversarial learning. We prepare semantic features as ground truth from the ground truth text in images to train the word embedding head and the discriminator. We generate features with all the semantic vector elements equal to zero for false-positive detections. It is then optimized according to the adversarial learning procedure [19].

The whole loss function consists of three parts, which are defined as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{\mathrm{det}} + \beta\mathcal{L}_{\mathrm{rec}} + \gamma\mathcal{L}_{\mathrm{adv}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{det}}$ and $\mathcal{L}_{\mathrm{rec}}$ are the detection and recognition loss function in text spotting framework [5]. The loss function $\mathcal{L}_{\mathrm{adv}}$ is the adversarial learning loss [19]. α, β and γ are the balance weights for $\mathcal{L}_{\mathrm{det}}$, $\mathcal{L}_{\mathrm{rec}}$ and $\mathcal{L}_{\mathrm{adv}}$, respectively.

4. EXPERIMENTS

4.1. Dataset and Evaluation

We evaluate the end-to-end text spotting accuracy of the proposed method on several standard benchmarks. We follow the standard evaluation protocols, which are based on F-measure evaluation [21–23]. The benchmark datasets used for the experiments in this paper are briefly introduced below.

CTW1500 [21]: is a curved scene benchmark, with 1,000 images for training and 500 images for testing.

Table 2: End-to-End text spotting performance comparison on CTW1500, ICDAR 2015, and Total-Text datasets. Following the standard evaluation protocol, we report the end-to-end results over two lexicons: "None" and "Full" on CTW1500 and Total-Text. "None" means that no lexicons are provided, and the "Full" lexicon provides all words in the test set. In ICDAR1500, "S", "W", and "G" mean recognition with the strong, weak, and generic lexicon, respectively.

Methods	CTW1500		ICDAR2015			Total-Text	
	None	Full	S	W	G	None	Full
Text Dragon [6]	38.7	72.4	82.5	78.3	65.1	48.8	74.8
Mask TextSpotter v3 [3]	–	–	83.3	78.1	74.2	71.2	78.4
PGNet [20]	–	–	83.3	78.3	63.5	63.1	–
ABCNet v2 [5]	57.5	77.2	82.7	78.5	73.0	70.4	78.1
TESTR [1]	56.0	81.5	**85.2**	79.4	73.6	73.3	83.9
SwinTextSpotter [14]	51.8	77.0	83.9	77.3	70.5	74.3	84.1
GLASS [15]	–	–	84.7	80.1	76.3	76.6	83.0
Ours	**64.4**	**82.3**	84.8	**83.7**	**79.6**	**79.4**	**85.1**

ICDAR 2015 [22]: is collected as scene text images containing low resolution and containing small text instances. It contains 1,000 training and 500 testing images.

Total-Text [23]: is another curved text benchmark, which consists of 1,255 training images and 300 testing images with multiple orientations.

4.2. Implementation Details

We follow the implementation procedure of ABCNet v2 [5]. The backbone of the network is based on ResNet-50-FPN [24]. For the dataset, The model is pre-trained on a mixture of 150k synthesized data [3], 7k MLT data [25], and the corresponding training data of each dataset. The pre-trained model is then fine-tuned on the training set of the target dataset. We use the BERT base model (uncased) pre-trained on Wikipedia provided by hugging face as a pre-trained word embedding model. We optimize our model using SGD with an image batch size of 8 on 4 A100 GPUs. We train the model until 260k iterations with the initial learning rate of 10^{-2}, which reduces to 10^{-3} and 10^{-4} at the 160k-th and 220k-th iteration, respectively. Note that We set the hyper parameters to be $\alpha = 1$, $\beta = 1$, $\gamma = 0.6$ in experiments.

4.3. Main Results

In this section, we compare the performances of the proposed model with state-of-the-art methods on public datasets. As shown in Table 2, we have summarized various text spotting methods. We see that A3S surpasses the recent competitive methods [1, 5, 14, 15] on most of the metrics.

In CTW1500, our method achieves 64.4 and 82.3 for "None" and "Full", respectively. In particular, the accuracy on "None", accuracy without lexicons, is significantly improved by 6.9 points compared to the baseline ABCNet v2. This indicates that A3S is more effective when lexicons are unavailable.

(a) Baseline (ABCNet v2 [5])

(b) Our proposed method (A3S)

Fig. 2: Visualized end-to-end text spotting results of the proposed method and baseline on CTW1500. The detected area is shown in blue, and the predicted text and its confidence score are shown in the upper left corner of the detected area. We see that the proposed method reduces both detection and text recognition errors.

In ICDAR2015 and Total-Text, the proposed method reaches competitive accuracy. Similar to CTW1500, we confirm that A3S performs well when unavailable lexicons. The proposed method performs better than GLASS [15], which enhances features before text recognition based on the global information in an input image. It is suggested that not only visual information but also semantic one is effective in scene-text spotting.

We visualize the scene-text spotting result of ours and baseline (ABCNet v2 [5]) in Fig. 2. We see an improvement in detection and recognition compared to the baseline.

4.4. Ablation Study

To confirm the effectiveness of the proposed method in detail, we conducted ablation studies on CTW1500.

Impact of Adversarial Learning. We investigated the impact of adversarial learning on semantic representations. Table 3 shows the accuracy results for the baseline, the complete proposed method, and the proposed model with different loss functions to match visual and semantic representations. Models (a) and (b) show the models where the loss function is replaced by L1-Norm and L2-Norm, respectively. Methods with semantic representations show accuracy improvement, but adversarial learning has a significant effect. Since L1- and L2-Norm directly join different representations of images and language, the influence from the pre-trained language model is likely too significant. Conversely, adversarial learning matches both representations indirectly by making them similar and thus is expected to be flexible and optimal.

Table 3: Analysis of adversarial learning for semantic representations on CTW1500.

Model	Method	None	Full
ABCNet v2 (baseline) [5]	–	57.5	77.2
model (a)	L1-Norm	58.3	79.4
model (b)	L2-Norm	59.5	79.9
complete model	Adversarial learning	**64.4**	**82.3**

Table 4: Impact of pre-trained word embedding on CTW1500. For GloVe, we utilized "glove-wiki-gigaword-300" from gensim.

Model	Word embedding	None	Full
model (c)	GloVe [26]	61.1	80.2
complete model	BERT [17]	**64.4**	**82.3**

Effect of Pre-trained Word Embedding. We investigated the impact of the pre-trained word embedding method in A3S. The proposed method uses the pre-trained BERT [17]. We examine the effect of using another word embedding model, GloVe [26], which learns embedding based on global word-word co-occurrence statistics from a corpus. Table 4 shows the difference in accuracy depending on the language model. The model with GloVe is denoted as the model (c). In both cases, we can see that using the language model improves accuracy compared to the baseline in Table 3. We see that the context-aware embedding method, BERT, has better accuracy, confirming the effectiveness of word embedding that utilizes context in scene-text spotting.

5. CONCLUSION

In this paper, we propose a novel scene-text spotting method named A3S. This method utilizes not only the visual information in text recognition but also the semantic one from a language model in training. We proposed to leverage adversarial learning to connect visual and semantic representation flexibly. Experiments show that our method achieves consistent gain and competitive results on popular benchmarks.

6. REFERENCES

[1] X. Zhang, Y. Su, S. Tripathi, and Z. Tu, "Text spotting transformers," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9519–9528.

[2] L. Qiao, Y. Chen, Z. Cheng, Y. Xu, Y. Niu, S. Pu, and F. Wu, "Mango: A mask attention guided one-stage scene text spotter," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 2467–2476.

[3] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask textspotter v3: Segmentation proposal network for robust scene text spotting," in *Proceedings of European Conference on Computer Vision*. Springer, 2020, pp. 706–722.

[4] M. Fujitake and H. Ge, "Temporally-aware convolutional block attention module for video text detection," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 2021, pp. 220–225.

[5] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu, and H. Chen, "Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] W. Feng, W. He, F. Yin, X.-Y. Zhang, and C.-L. Liu, "Textdragon: An end-to-end framework for arbitrary shaped text spotting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9076–9085.

[7] J. A. Rodriguez-Serrano, F. Perronnin, and F. Meylan, "Label embedding for text recognition.," in *Proceedings of British Machine Vision Conference*, 2013, pp. 1–12.

[8] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2642–2651.

[9] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[10] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2017.

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

[12] Y. Zhang, W. Yang, Z. Xu, Y. Li, Z. Chen, and L. Huang, "Pointer networks for arbitrary-shaped text spotting," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 2375–2379.

[13] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: Fast oriented text spotting with a unified network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5676–5685.

[14] M. Huang, Y. Liu, Z. Peng, C. Liu, D. Lin, S. Zhu, N. Yuan, K. Ding, and L. Jin, "Swintextspotter: Scene text spotting via better synergy between text detection and text recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4593–4603.

[15] R. Ronen, S. Tsiper, O. Anschel, I. Lavi, A. Markovitz, and R. Manmatha, "Glass: Global to local attention for scene-text spotting," *arXiv preprint arXiv:2208.03364*, 2022.

[16] C. Cheng, B. Li, Q. Zheng, Y. Wang, and W. Liu, "Decoupling visual-semantic feature learning for robust scene text recognition," *arXiv preprint arXiv:2111.12351*, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of International Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[20] P. Wang, C. Zhang, F. Qi, S. Liu, X. Zhang, P. Lyu, J. Han, J. Liu, E. Ding, and G. Shi, "Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 2782–2790.

[21] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognition*, vol. 90, pp. 337–345, 2019.

[22] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al., "Icdar 2015 competition on robust reading," in *Proceedings of IEEE International Conference on Document Analysis and Recognition*, 2015, pp. 1156–1160.

[23] C.-K. Ch'ng, C. S. Chan, and C.-L. Liu, "Total-text: toward orientation robustness in scene text detection," *International Journal on Document Analysis and Recognition*, vol. 23, no. 1, pp. 31–52, 2020.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944.

[25] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J.-C. Burie, C.-l. Liu, et al., "Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019," in *Proceedings of IEEE International Conference on Document Analysis and Recognition*, 2019, pp. 1582–1587.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.