Language Matters: A Weakly Supervised Vision-Language Pre-training Approach for Scene Text Detection and Spotting

Chuhui Xue, Wenqing Zhang, Yu Hao,
Shijian Lu, Philip Torr, and Song Bai

Supplementary Material

1 Automatic Data Acquisition and Training from Web Images

Most existing scene text detection and spotting models are trained on fully-annotated data that are difficult to obtain from web images. Instead, the proposed weakly supervised pre-training approach can be simply applied to an automatic data acquisition and training pipeline by: (1) Extracting texts from web images by the existing OCR techniques; (2) Filtering out the less confident text instances (i.e. detected and recognized texts with low confident scores); (3) Pre-training a model on the collected web images and extracted text instances.

Table 1. Automatic data acquisition and training from web images: By pre-training on the automatically extracted images and texts from web, the proposed method can promote the existing scene text detectors significantly on TotalText and CTW1500 datasets.

Model	Total-Text			CTW1500		
	P	**R**	**F**	**P**	**R**	**F**
PSENet [9]	81.8	75.1	78.3	80.6	75.6	78.0
PSENet+SynthText	87.8	79.0	82.6	81.8	77.8	79.7
PSENet+Ours[SynthText]	90.7	80.8	85.5	86.3	79.6	82.8
PSENet+Ours[Web Images]	**92.2**	**82.4**	**87.0**	**87.5**	**79.9**	**83.5**

We conduct an experiment following this pipeline. We first extract texts from web images by using PSENet [9] for detection and Conformer [3] for recognition. Then, we filter out the less confident texts and non-text images, resulting in 40 million image-text pairs. Finally, we pre-train a model by using the proposed method and transfer the weights in the pre-trained model to fine-tune PSENet on Total-Text and CTW1500 datasets. As Table 1 shows, by automatically extracting data and pre-training, the proposed method significantly improves the performances of PSENet on Total-Text and CTW1500 datasets, demonstrating the effectiveness of the proposed method. This result also shows that the

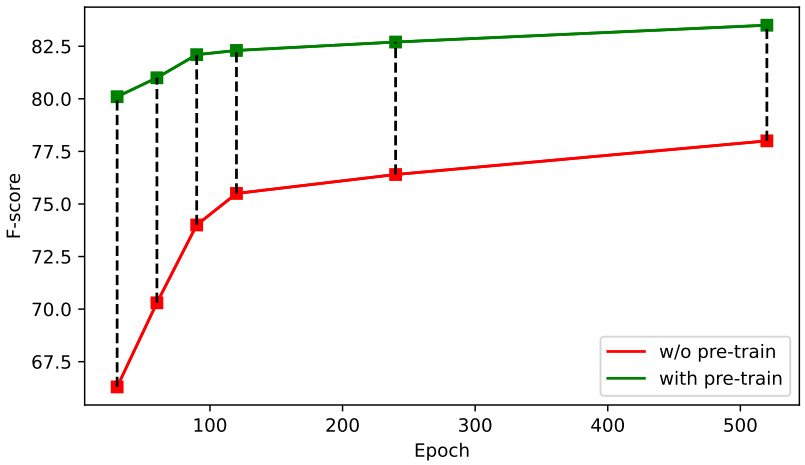

Fig. 1. By pre-training on web images, the model converges faster than the original model without pre-training.

scene text models can be effectively promoted by large-scale pre-training on web images.

Besides, the proposed pre-trained models effectively accelerate the convergence of the scene text model. As Fig. 1 shows, the scene text detector with pre-trained converges faster than the original model without pre-training.

2 Datsets

SynthText [4] contains more than 800,000 synthetic scene text images most of which are at word level with multi-oriented rectangular annotations. The texts are in English in SynthText dataset.

ICDAR2019-LSVT [8] consists of 450,000 images with mostly Chinese texts. 400,000 images are weakly annotated in which only the transcription of the text-of-interest in these images is provided. Besides, 50,000 images are fully annotated which are split into a training set with 30,000 images and a test set with 20,000 images.

CTW1500 [10] consists of 1,000 training images and 500 test images that contain 10,751 multi-oriented text instances of which 3,530 are arbitrarily curved. Most of the text instances are annotated at text-line level by using 14 vertices, where texts are largely in English and Chinese.

Total-Text [1] consists of 1,255 training images and 300 test images where texts are all in English. It contains a large number of multi-oriented curved text instances each of which is annotated at word level by using a polygon.

Fig. 2. Given sample images in the first row, the second row shows the corresponding attention maps in the image encoder. Rows 3-4 shows a context text and a contextless text as input, respectively, as well as the corresponding attention maps in the decoder and the predicted characters. The encoder and decoder effectively attend to the text and character regions, respectively.

ICDAR2015 [5] has 1000 training images and 500 test images which are collected by Google Glass and suffers from low resolution and motion blur. All text instances are annotated at word level using quadrilateral boxes.

3 Implementation Details

We fine-tune several scene text detectors and spotters for evaluation of the proposed method including: 1) PSENet [1] [9], 2) DB [2] [7], 3) FCENet [3] [12], 4)

[1] https://github.com/whai362/PSENet
[2] https://github.com/MhLiao/DB
[3] https://github.com/open-mmlab/mmocr

Table 2. Comparison with state-of-the-art scene text spotting techniques on **ICDAR2015**. '+oCLIP' refers to that the model are fine-tuned from the our pre-trained model on SynthText dataset. 'S', 'W', and 'G' refer to word spotting with strong, weak, generic lexicon for ICDAR2015. 'Full' refers to full lexicon for Total-Text.

Model	**ICDAR2015**		
	S	**W**	**G**
TextDragon [2]	86.2	81.6	68.0
Mask TextSpotter-V3 [6]	83.1	79.1	75.1
Mask TextSpotter-V3+oCLIP	**84.1**,	**79.5**	**75.1**

TextBPN [4] [11], and 5) Mask TextSpotter-v3 [5] [6]. The experiments are conducted by using the corresponding open-source codes. For DB, we replace the original network backbone (i.e. deformable ResNet-50) with ResNet-50 for better demonstration of the proposed method. For TextBPN, we follow the experimental settings reported in their paper to re-train the overall the model as the configuration files are not provided.

4 More Samples

The proposed method can attend to text regions with character awareness with language supervision only. Fig. 2 shows four more sample images as well as their attention maps. As Fig. 2 shows, the proposed model can effectively attend to the text regions and the missing character regions (corresponding to each input text instance). Especially in the last row of Fig. 2, the proposed decoder can attend to the regions of missing characters from contextless text instance, demonstrating the effectiveness of the proposed method on modelling the relations of visual and textual information.

5 More Experimental Results

Recent scene text spotters are usually evaluated on ICDAR2015 dataset under two evaluation metrics. We report our results on end-to-end spotting in the main manuscript, and additionally report the results under word spotting in Table 2.

References

1. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on. vol. 1, pp. 935–942. IEEE (2017)

[4] https://github.com/GXYM/TextBPN
[5] https://github.com/MhLiao/MaskTextSpotterV3

2. Feng, W., He, W., Yin, F., Zhang, X.Y., Liu, C.L.: Textdragon: An end-to-end framework for arbitrary shaped text spotting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9076–9085 (2019)
3. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
4. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2315–2324 (2016)
5. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 1156–1160. IEEE (2015)
6. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In: European Conference on Computer Vision. pp. 706–722. Springer (2020)
7. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proc. AAAI (2020)
8. Sun, Y., Ni, Z., Chng, C.K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., et al.: Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1557–1562. IEEE (2019)
9. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9336–9345 (2019)
10. Yuliang, L., Lianwen, J., Shuaitao, Z., Sheng, Z.: Detecting curve text in the wild: New dataset and new solution. arXiv preprint arXiv:1712.02170 (2017)
11. Zhang, S.X., Zhu, X., Hou, J.B., Liu, C., Yang, C., Wang, H., Yin, X.C.: Deep relational reasoning graph network for arbitrary shape text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9699–9708 (2020)
12. Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., Zhang, W.: Fourier contour embedding for arbitrary-shaped text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3123–3131 (2021)