Supplemental Material for Towards End-to-End Unified Scene Text Detection and Layout Analysis

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, Michalis Raptis
Google Research
{longshangbang,qinb,dpantele,bissacco,yasuhisaf,mraptis}@google.com

Figure 1. Failure case where some paragraphs are not fully recovered.

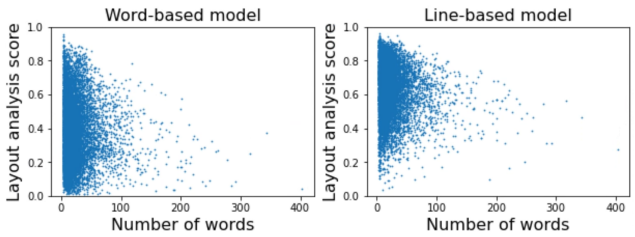

Figure 2. Each dot represents one image in the validation set. The x-axis is the number of words in each image. The y-axis is the layout analysis score of the proposed models.

A. Model demonstration

Fig. 3 shows more results of the unified detector.

A.1. Failure cases

Fig. 1 shows a typical failure case where the paragraph of the title of news paper is not fully recovered. Note that, the proposed PQ-like metric is able to penalize the missing line ('*Commission told to*') by the IoU score between ground-truth and prediction.

Method	Text detection branch	Layout analysis branch	Latency (ms)
GCN-PP [3]	Text detection branch of unified detector	GCN	644
Mask-RCNN-Cluster		Mask-RCNN [1]	689
MaX-DeepLab-Cluster		MaX-DeepLab [2]	883
Unified Detector		Layout branch of unified detector	**521**

Table 1. Latency of different methods

B. Latency

Tab. 1 shows the latency of different methods, measured on a V100 accelerator. The unified model achieves the best latency, while other methods are slower due to multi-step pipeline design and multiple separate modules.

C. Word based model v.s. line based methods

In Fig. 2, we show the layout scores against the number of words in images. Both models perform worse when images contain larger amounts of words. The line-based model is more robust to higher word numbers.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[2] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 1

[3] Renshen Wang, Yasuhisa Fujii, and Ashok C Popat. Post-ocr paragraph recognition by graph convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 493–502, 2022. 1

Figure 3. Demonstration of the model. In each pair of images, line detection and paragraph clustering are shown in the left and right respectively.