Uber-Text: A Large-Scale Dataset for Optical Character Recognition from Street-Level Imagery

Ying Zhang[1,2] Lionel Gueguen[1] Ilya Zharkov[1] Peter Zhang[1] Keith Seifert[1]
Ben Kadlec[1]
[1]Uber Technologies.
[2]MILA, Université de Montréal
zhangy@umontreal.ca, {lgueguen, zharkov, peizha, kseifert, bkadlec}@uber.com

Abstract

Optical Character Recognition (OCR) approaches have been widely advanced in recent years thanks to the resurgence of deep learning. The state-of-the-art models are mainly trained on the datasets consisting of the constrained scenes. Detecting and recognizing text from the real-world images remains a technical challenge. In this paper, we introduce a large-scale OCR dataset Uber-Text, which contains (1) streetside images with their text region polygons and the corresponding transcriptions, (2) 9 *categories indicating the business name text, street name text and street number text, etc, (3) a set containing over* 110*k images, (4)* 4.84 *text instances per image on average. We show the challenge of the task and the dataset via evaluating the prevalent methods, which proves the significance of the dataset and motivates the future work in this field of study.*

1. Introduction

Text detection and recognition from images is crucial for the digitalization and understanding of our world. The ability to detect and interpret text in imagery is a broad ranging problem. Numerous studies on OCR have been conducted in the past decades[6, 2], but often in the context of laboratory conditions with canonically located text and low resolution images.

Natural and street view images are characterized by high degree of variability, which could easily make the cutting-edge methods fail[5]. The lack of training data acquired in varied and unconstrained conditions is primarily accounted for the limited results. To support and facilitate this specific field of study, we introduce a large-scale OCR dataset, Uber-Text[1], which consists of 117969 street-level images with 571534 annotated text regions acquired in 6 US cities. Samples from the dataset are depicted in Figure 1.

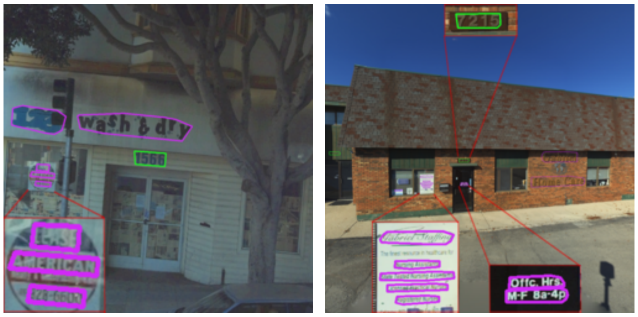

Figure 1: Two examples from Uber-Text dataset with the corresponding truth highlighted. These examples show the variabilities of the dataset in terms of text type, location, font and category and image background.

In this paper, we assess the accurancy performance of the state-of-the-art detectors, namely Faster R-CNN [4] and YOLO [3], in addition, we adopt an end-to-end text sequence recognition method, which could be considered as a generalization of the previous work [1].

2. Uber-Text Dataset

Uber-Text dataset has been obtained by capturing 117969 images by the Bing Maps Streetside program deployed in 6 US cities over the course of 2 years. Then, these images have been labeled by a team of image analysts. The captured text would be categorized into street number, street name, business name, license plate, phone number, secondary unit designator, street number range, traffic sign and others (none). Non-ASCII, non-English and unrecognizable characters are labeled as asterisk.

Each image is associated to a text file, following the representations adopted by the ICDAR 2015 Robust Reading competition. The differences from the ICDAR 15 include: (1) we provide accurate polygon coordinates, (2) a category

[1] available at `https://s3-us-west-2.amazonaws.com/uber-common-public/ubertext/index.html`

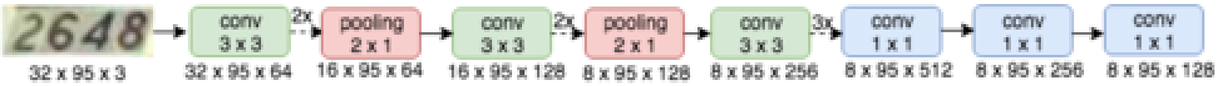

Figure 2: CNN structure of the adopted model. 2×/ 3× above the dashed arrow represents times of repetition of the convolution

has been appended classifying the text instances, (3) no "Do Not Care" regions and the words less than 3 characters are considered. The dataset is subdivided in two subsets, accounting for the corresponding image sizes which can be either approximate 1K×1K pixels or 4K×4K pixels. The dataset incorporates 571534 labeled text instances spread in 117969 images, which averages to 4.84 instances per image.

3. Experiments and Results

We evaluate Faster R-CNN and YOLO on Uber-Text for detection and provide the recognition baseline via the adopted model.

3.1. Text Bounding Box Detection

We use fstrcnn-1k and fstrcnn-4k denote the Faster R-CNN models trained on 1K×1K and 4K×4K subset. Comparison experiment on a smaller anchor scales set (64^2, 128^2, 256^2) is conducted. YOLO is evaluated on 1K×1K subset solely since YOLO has the deficiency of localizing small objects as the author noted in [3].

mean Average Precision (mAP) with 50% IoU is adopted as the evaluation metric. As shown in the table 1, Faster R-CNN achieves 51.1% mAP on 1Kx1K subset while YOLO obtain 49.5% mAP. With the smaller anchor scales set, the performance of faster R-CNN gets sightly better.

Model	mAP (%)
fstrcnn-1K	51.1
fstrcnn-1K (small scale)	53.7
fstrcnn-4K	9.0
fstrcnn-4K (small scale)	10.4
yolo-1K	49.5

Table 1: Detection results of Faster R-CNN and YOLO on Uber-Text test subsets.

3.2. Text Recognition

Our model is composed of a deep CNN, bi-directional LSTM layers and Connectionist Temporal Classification (CTC). The CNN module is shown in Figure 2. 3 bi-directional LSTM layers with 250 units in each direction follow the fully-connected layers and CTC objective is applied on top. The result are shown in table 2.

Model	Accuracy (%)
end-to-end	56.4
human performance	$\geqslant 90$

Table 2: Recognition result of the proposed model on Uber-Text compared to the human evaluation.

4. Discussion

We introduce a dataset which contains over 110k images collected from natural scenes. This is the first large-scale dataset from street-level imagery which could serve as a new benchmark for the OCR tasks. The performance of the state-of-the-art methods indicate the limitation of the current approaches in the real-world condition and also evidence the challenge and significance of our dataset. We anticipate Uber-text would facilitate the research in the field of study.

References

[1] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang. Reading scene text in deep convolutional sequences. *arXiv preprint arXiv:1506.04395*, 2015.

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[5] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[6] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.