Enhancing Scene Text Recognition with Visual Context Information

Ahmed Sabir
Universitat Politècnica de Catalunya
Barcelona, Spain
asabir@cs.upc.edu

Abstract

This thesis addresses the problem of improving text spotting systems, which aim to detect and recognize text in unrestricted images (e.g., a street sign, an advertisement, a bus destination, etc.). The goal is to improve the performance of off-the-shelf vision systems by exploiting the semantic information derived from the image itself. The rationale is that knowing the content of the image or the visual context can help to decide which words are the correct candidate words. For example, the fact that an image shows a coffee shop makes it more likely that a word on a signboard reads as *Dunkin* and not *unkind*.

We address this problem by drawing on successful developments in natural language processing and machine learning, in particular, learning to re-rank and neural networks, to present post-process frameworks that improve state-of-the-art text spotting systems without the need for costly data-driven re-training or tuning procedures.

Discovering the degree of semantic relatedness of candidate words and their image context is a task related to assessing the semantic similarity between words or text fragments. However, semantic relatedness is more general than similarity (e.g., *car*, *road*, and *traffic light* are related but not similar) and requires certain adaptations. To meet the requirements of these broader perspectives of semantic similarity, we develop two approaches to learn the semantic relatedness of the spotted word and its environmental context: word-to-word (object) or word-to-sentence (caption). In the word-to-word approach, word embedding based re-rankers are developed. The re-ranker takes the words from the text spotting baseline and re-ranks them based on the visual context from the object classifier. For the second, an end-to-end neural approach is designed to drive image description (caption) at the sentence-level as well as the word-level (objects) and re-rank them based not only on the visual context but also on the co-occurrence between them.

As an additional contribution, to meet the requirements of data-driven approaches such as neural networks, we propose a visual context dataset for this task, in which the publicly available COCO-text dataset[1] has been extended with information about the scene (including the objects and places appearing in the image) to enable researchers to include the semantic relations between texts and scene in their Text Spotting systems, and to offer a common evaluation baseline for such approaches.

[1]Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S. (2016). Coco-text: Dataset and benchmark for text detection and recognition in natural images. https://arxiv.org/pdf/1601.07140.pdf

Awarded by: Universitat Politècnica de Catalunya, Barcelona, Spain **on** 10 September 2020.
Supervised by: Lluís Padró and Francesc Moreno-Noguer.
Available at: https://upcommons.upc.edu/handle/2117/334952.

Selected Publications

Ahmed Sabir, Francesc Moreno-Noguer, and Lluís Padró. Enhancing text spotting with a language model and visual context information. In *Artificial Intelligence Research and Development*, pages 271–280. IOS Press, 2018a.

Ahmed Sabir, Francesc Moreno-Noguer, and Lluís Padró. Visual re-ranking with natural language understanding for text spotting. In *Asian Conference on Computer Vision*, pages 68–82. Springer, 2018b.

Ahmed Sabir, Francesc Moreno, and Lluís Padró. Semantic relatedness based re-ranker for text spotting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3451–3457, 2019.

Ahmed Sabir, Francesc Moreno-Noguer, and Lluís Padró. Textual visual semantic dataset for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 542–543, 2020.