Text-Attentional Convolutional Neural Network for Scene Text Detection.

Recent deep learning models have demonstrated strong capabilities for classifying text and non-text components in natural images. They extract a high-level feature globally computed from a whole image component (patch), where the cluttered background information may dominate true text features in the deep representation. This leads to less discriminative power and poorer robustness. In this paper, we present a new system for scene text detection by proposing a novel text-attentional convolutional neural network (Text-CNN) that particularly focuses on extracting text-related regions and features from the image components. We develop a new learning mechanism to train the Text-CNN with multi-level and rich supervised information, including text region mask, character label, and binary text/non-text information. The rich supervision information enables the Text-CNN with a strong capability for discriminating ambiguous texts, and also increases its robustness against complicated background components. The training process is formulated as a multi-task learning problem, where low-level supervised information greatly facilitates the main task of text/non-text classification. In addition, a powerful low-level detector called contrast-enhancement maximally stable extremal regions (MSERs) is developed, which extends the widely used MSERs by enhancing intensity contrast between text patterns and background. This allows it to detect highly challenging text patterns, resulting in a higher recall. Our approach achieved promising results on the ICDAR 2013 data set, with an F-measure of 0.82, substantially improving the state-of-the-art results.