

Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation

Ruochen Fan¹[0000-0003-1991-0146], Qibin Hou²[0000-0002-8388-8708], Ming-Ming Cheng²[0000-0001-5550-8758], Gang Yu³[0000-0001-5570-2710], Ralph R. Martin⁴, and Shi-Min Hu¹[0000-0001-7507-6542]

¹ Tsinghua University, Beijing, China ffr16@mails., shimin@tsinghua.edu.cn

² Nankai University, Tianjin, China cmm@nankai.edu.cn, andrewhou@gmail.com

³ Megvii Inc., Beijing, China yugang@megvii.com

⁴ Cardiff University, Cardiff CF243AA, U.K. ralph@cs.cardiff.ac.uk

Abstract. Effectively bridging between image level keyword annotations and corresponding image pixels is one of the main challenges in weakly supervised semantic segmentation. In this paper, we use an instance-level salient object detector to automatically generate salient instances (candidate objects) for training images. Using similarity features extracted from each salient instance in the whole training set, we build a similarity graph, then use a graph partitioning algorithm to separate it into multiple subgraphs, each of which is associated with a single keyword (tag). Our graph-partitioning-based clustering algorithm allows us to consider the relationships between all salient instances in the training set as well as the information within them. We further show that with the help of attention information, our clustering algorithm is able to correct certain wrong assignments, leading to more accurate results. The proposed framework is general, and any state-of-the-art fully-supervised network structure can be incorporated to learn the segmentation network. When working with DeepLab for semantic segmentation, our method outperforms state-of-the-art weakly supervised alternatives by a large margin, achieving 65.6% mIoU on the PASCAL VOC 2012 dataset. We also combine our method with Mask R-CNN for instance segmentation, and demonstrated for the first time the ability of weakly supervised instance segmentation using only keyword annotations.

Keywords: Semantic segmentation, weak supervision, graph partitioning.

1 Introduction

Semantic segmentation, providing rich pixel level labeling of a scene, is one of the most important tasks in computer vision. The strong learning ability of convolutional neural networks (CNNs) has enabled significant progress in this field recently [5, 27, 29, 46, 47]. However, the performance of such CNN-based methods requires a large amount of training data annotated to pixel-level, e.g., PASCAL VOC [11] and MS COCO [28]; such data are very expensive to collect. As an approach to alleviate the demand for pixel-accurate annotations, weakly supervised semantic segmentation has drawn great attention recently. Such methods merely require supervisions of one or more of the

