International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01849-z

Semantic Contrastive Bootstrapping for Single-Positive Multi-label Recognition

Cheng Chen[1] · **Yifan Zhao**[2] · **Jia Li**[1,3]

Received: 3 October 2022 / Accepted: 12 July 2023

Abstract
Learning multi-label image recognition with incomplete annotation is gaining popularity due to its superior performance and significant labor savings when compared to training with fully labeled datasets. Existing literature mainly focuses on label completion and co-occurrence learning while facing difficulties with the most common single-positive label manner. To tackle this problem, we present a semantic contrastive bootstrapping (Scob) approach to gradually recover the cross-object relationships by introducing class activation as semantic guidance. With this learning guidance, we then propose a recurrent semantic masked transformer to extract iconic object-level representations and delve into the contrastive learning problems on multi-label classification tasks. We further propose a bootstrapping framework in an Expectation-Maximization fashion that iteratively optimizes the network parameters and refines semantic guidance to alleviate possible disturbance caused by wrong semantic guidance. Extensive experimental results demonstrate that the proposed joint learning framework surpasses the state-of-the-art models by a large margin on four public multi-label image recognition benchmarks. Codes can be found at https://github.com/iCVTEAM/Scob.

Keywords Multi-label image recognition · Single-positive label · Contrastive learning · Semantic guidance

1 Introduction

Recognizing multiple visual objects within one image is a natural and fundamental problem in computer vision, as it provides prerequisites for many downstream applications, including segmentation (Zhang et al., 2021b), scene understanding (Sener & Koltun, 2018), and attribute recognition (Jia et al., 2021). With the help of sufficient training annotations, existing research efforts (Rao et al., 2021; Chen et al., 2019b; Hu et al., 2020; Zhao et al., 2021; Chen et al., 2019a; Wang et al., 2016; Huynh & Elhamifar, 2020; Bucak et al., 2011; Carion et al., 2020) have undoubtedly made progress via supervised deep learning models. However, annotating all occurrences of candidate objects, especially small ones, is extremely tedious and labor-consuming, which also usually introduces incorrect noisy labels. Recent approaches towards this challenge prefer to use partial weak labels rather than full annotations, making data collecting considerably easier. In addition, Durand et al. (2019) have demonstrated that training sufficient weak labels shows more promising results than those trained with fully labeled but noisy datasets.

Motivated by this huge potential in multi-label learning, representative works tend to learn the co-occurrence correlations between instances (Wu et al., 2018; Chen et al., 2021, 2022). The other line of work attempts to refine the labeling matrix by pretraining on an accurate fully labeled dataset (Jiang et al., 2018; Chen et al., 2019a) or annotating additional negative training samples (Durand et al., 2019). Nevertheless, these works inevitably fail to handle extreme

Communicated by Jianfei Cai.

Cheng Chen and Yifan Zhao have contributed equally to this paper.

✉ Jia Li
jiali@buaa.edu.cn

Cheng Chen
chencheng1@buaa.edu.cn

Yifan Zhao
zhaoyf@pku.edu.cn

[1] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China

[2] School of Computer Science, Peking University, Beijing, China

[3] Peng Cheng Laboratory, Shenzhen, China

Published online: 13 August 2023