BABEŞ-BOLYAI UNIVERSITY CLUJ-NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
SPECIALIZATION Applied Computational Intelligence

DISSERTATION THESIS

Semi-supervised Active Learning for Object Detection

Supervisor
Prof. Dr. Diosan Laura

Author
Novacean Nicoleta-Ligia

2022

ABSTRACT

Abstract: un rezumat în limba engleză cu prezentarea, pe scurt, a conținutului pe capitole, punând accent pe contribuțiile proprii și originalitate

Contents

Chapter 1

Introduction

Introducere: obiectivele lucrării și descrierea succintă a capitolelor, prezentarea temei, prezentarea contribuției proprii, respectiv a rezultatelor originale și menționarea (dacă este cazul) a sesiunii de comunicări unde a fost prezentată sau a revistei unde a fost publicată.

Chapter 2

Bibliographic Study

The impressive performance of machine learning models in the field of computer vision is partially due to the availability of large annotated datasets. However, for certain tasks, the annotation process is highly time-consuming. In the case of object detection, annotations consist of a pair of bounding box location and label for every object within in image and may take up to 10 seconds per object [RLFF15].

To reduce the labeling cost, two of the studied methods are Active Learning (AL) and Semi-Supervised Learning (SSL). With Active Learning, the core idea is that machine learning models can achieve better performance with fewer training samples if allowed to choose which samples to learn. For Semi-Supervised Learning, models are trained in a context of limited supervision, with both labeled and unlabeled data. Both techniques were initially applied for image classification tasks, but recently substantial progress was made in applying them for object detection. Moreover, research was also done on combining these two strategies for image classification, in the attempt to further reduce the labelling cost [GZY+19].

2.1 Semi-Supervised Learning for Object Detection

In recent years, semi-supervised learning has received a lot of attention as it allows models to take advantage of the large amounts of unlabeled data readily available. Generally, these methods fall in one of the following two categories:

Self-training: A model is trained on labeled data and, afterwards, used to generate predictions for unlabeled data. For unlabeled samples where the prediction confidence exceeds a given threshold, the prediction is considered the ground truth. The produced labels are known as *pseudo-labels*. These methods are highly dependent upon the selected threshold. Adding too many samples can reduce the stability of the training process, while adding too few samples may not bring a considerable

performance improvement.

Consistency regularization: Consistency-based methods encourage invariant predictions for different perturbations of unlabeled input samples. These methods also have a regularization effect, as they help smooth the distribution.

These two methodologies were successfully and widely applied for image classification. Given the laborious and time-consuming process required by object detection annotations, a surge of attention was also dedicated to the application of Semi-Supervised Learning in Object Detection (SSOD). Initial SSOD strategies tried to use techniques from image classification. However, it turned out the methods cannot be directly applied on object detection, mainly because of the following reasons:

- Lack of consideration of the localization precision, hence the bounding box regression component of the object detection task.
- Class imbalance, a long-established challenge in object detection, that refers to both foreground-background imbalance and foreground-foreground imbalance. For example, in the 1% SSOD COCO-standard dataset, background instance account for about 90% of the training instances.

Consequently, Semi-Supervised Learning for Object Detection techniques tried to address object detection specific challenges. Despite the increase in SSOD related research and methods, there is still a significant performance gap between the SSOD methods and the fully supervised alternative.

The current state-of-the-art methods in SSOD rely on the self-training paradigm, hence on pseudo-labeling, often mixed with consistency regularization via the weak-strong data augmentation scheme. The self-training scheme is multi-stage, generally consisting of the supervised training an initial detector, followed by pseudo-label generation and re-training using the obtained pseudo-labeled annotated data. The performance of these approaches is highly dependent upon the quality of the pseudo-labels. If pseudo-labels are noisy, the confidence of the network on incorrect predictions increases as training evolves. This effect is called confirmation bias, or noise accumulation.

One of the first papers on SSOD that employed the self-training paradigm mixed with consistency regularization is STAC [SZL+20]. STAC trains an initial object detector that is afterwards used to generate predictions for unlabeled data. After NSM, predictions are filtered based on the classification confidence using a high threshold. Then, strong data augmentations are applied to unlabeled data and the model is further trained on labeled and augmented unlabeled data, with the associated pseudo-labels. One of the shortcomings of STAC is that the initial object detector network, used to generate pseudo-labels, is not updated during training. To address this is-

sue, Instant Teaching [ZYW+21] updates the labels during training. Furthermore, an reduce the effect of the confirmation bias, the authors propose a co-rectification scheme, named Instant-Teaching*, that uses two models with the same architecture and different weights to rectify false predictions.

Both STAC [SZL+20] and Instant Teaching [ZYW+21] suffer from the class imbalance issue. With pseudo-labeling methods, the deep neural network is biased towards dominant and, consequently, confident classes. The inherent class imbalance in object detection is, then, emphasized by these methods. To address this issue, Unbiased Teacher [LMH+21] uses the Teacher-Student dual learning framework where the Student and Teacher networks are trained together, in a mutually beneficial process. The Teacher network is responsible for pseudo-labels generation, used afterwards to train the Student network. At the same time, the Student updates the Teacher using Exponential Moving Average techniques. Under this formulation, the Teacher can be seen as a temporal ensemble of Student networks, which reduces the confirmation bias. In the Teacher's classification head, the focal loss is used to alleviate the class imbalance.

Besides the classification component, object detection networks have a localization or bounding box regression component. However, most SSOD approaches do not take into account any information about how precise or confident the localization is. Soft Teacher [XZH+21] introduces box jittering as a bounding box reliability measure. Pseudo-foreground candidate boxes are jittered and the variance of the confidence is used as reliability measure. Furthermore, Soft Teacher also introduces a simple mechanism to address over confident classes: the classification loss of

2.2 Active Learning for Object Detection

Active Learning (AL) represents a set of strategies that reduce the amount of required annotated data by selecting the most informative data for the training process. The selected samples are than annotated by human annotators. While AL has seen great progress in image classification, it is still under-developed for object detection problems.

Initial approaches to Active Learning for Object Detection, referred to as Active Object Detection (AOD), tried to apply techniques from image classification based active learning. In this direction, losses were simply sorted or image-level uncertainty was estimated based on pixel-level information. However, for instance level based active learning, there exist very few solutions. Such an example is MI-AOD [YWF+21], which treats the images in the active learning pool as instance bags and feature anchors in images as instances. Then, it estimates the image uncertainty by re-weighting instances, as per the multiple instances learning approach.

Model name	Methodology	Network	Data augmentations	Characteristics
STAC [SZL+20]	• Self-training with pseudo-labels • Consistency regularization via data augmentation	Faster R-CNN with FPN and ResNet-50	Strong: Color transformations, Geometric transformations, Cutout	• **Unsupervised loss**: OD supervised loss for pseudo-labels
Instant Teaching [ZYW+21]	• Self-training with pseudo-labels • Consistency regularization via strong-weak data augmentation scheme	Faster R-CNN with FPN and ResNet-50, continously trained	Weak: not mentioned Strong: Color+Cutout, Mixup, Mosaic	• **Unsupervised loss**: OD supervised loss for pseudo-labels
Unbiased Teacher [LMH+21]	• Self-training with pseudo-labels • Consistency regularization via strong-weak data augmentation scheme	Teacher: Faster R-CNN with FPN and ResNet-50, EMA of Student; Student: Faster R-CNN with FPN and ResNet-50	Weak: horizontal fliping; Strong: Color jittering, Gaussian blur, Cutout	• **Unsupervised loss**: OD supervised loss for pseudo-labels • Focal loss
RPL [LWSD21]	• Self-training with pseudo-labels • Consistency regularization via strong-weak data augmentation scheme	Teacher: Faster R-CNN with FPN and ResNet-50, EMA of Student; Student: Faster R-CNN with FPN and ResNet-50	Weak: not mentioned; Strong: Color, Blur, Cutout	• **Unsupervised loss**: OD supervised loss for pseudo-labels • Certainty-aware pseudo labels • Dynamic thresholding & Weighted losses per class
Soft Teacher [XZH+21]	• Self-training with pseudo-labels • Consistency regularization via strong-weak data augmentation scheme	Teacher: Faster R-CNN with FPN and ResNet-50, EMA of Student; Student: Faster R-CNN with FPN and ResNet-50	Weak: not mentioned; Strong: Color jittering, Cutout, FixMatch	• **Unsupervised loss**: OD supervised loss for pseudo-labels • Box jittering to improve BB reliability

Table 2.1: Semi-Supervised Learning for Object Detection: Overview of existing algorithms

Recently, a paper addressed the AOD problem in a consistency-based scenario [YZYC21]. One of the biggest challenges in object detection is finding an acquisition metric that reflects both the classification and bounding box regression components of the detection task. The authors of [YZYC21] propose an approach based on consistency across multiple data augmentations. More precisely, they compute the consistency between augmented and non-augmented images in terms of intersection over union and weighted class distributions. For the data augmentation part, horizontal flipping, cutout, down-sizing and rotation are employed.

2.3 Semi-Supervised Active Learning

Active Learning (AL) can suffer from the ”cold start” problem, especially in contexts with a very small amount of data. The ”cold start” problem refers to having a wrongly biased initial model trained for computing the acquisition metrics, which afterwards has an impact on the entire active learning pipeline. One of the strategies that can be employed to alleviate the ”cold start” problem is semi-supervised active learning. Using unlabeled samples in the initial training process can regularize the network and stabilize predictions, whilst improving the performance in the selection part of AL. Consequently, one can conclude that active learning and semi-supervised learning can be used in a joint setup. Furthermore, semi-supervised training can be applied not only in the initial burn-in phase, but throughout the entire training process.

2.3.1 Consistency-based Semi-Supervised Active Learning for Image Classification

In [GZY+19], the authors propose a framework to unify active learning and semi-supervised learning for image classification. The resulting framework, named Active Semi-Supervised Learning for Image Classification (ASSL), proposes a consistency-based selection metric for AL, that is coherent with the unsupervised training objective used by SSL.

In the semi-supervised part, consistency regularization is used. Thus, every input image is distorted using a random augmentation. Then, the KL-divergence between the augmented and non-augmented predictions is computed, representing the unsupervised loss. The model in trained in this setup until convergence is obtained.

For the active learning pipeline, there exists a pool of unlabeled data from which samples are selection for annotation. for every unlabeled samples, a set of random augmentations are applied and the variance of the probability distributions is com-

puted. This forms the selection metric for AL: the higher the variance, the more uncertain the model is about the correct label of an image. The label itself does not matter in this context, but only the smoothness in predictions in a neighbourhood defined by augmentations.

Experiments performed in this setup indicate that a very important component for the performance of this framework is the coherence between the semi-supervised objection and active acquisition function. Ablation studies indicate that when consistency-based unsupervised losses are joint with feature space distribution or entropy-based acquisition metrics, the performance decreases.

Chapter 3

Analysis and Theoretical Study

Chapter 4

Design and Implementation

Chapter 5

Experimental Results

Chapter 6

Conclusions

Bibliography

[GZY+19] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Arik, Larry Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost, 10 2019.

[LMH+21] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection, 02 2021.

[LWSD21] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry Davis. Rethinking pseudo labels for semi-supervised object detection, 05 2021.

[RLFF15] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: Human-machine collaboration for object annotation. pages 2121–2131, 06 2015.

[SZL+20] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection, 05 2020.

[XZH+21] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher, 06 2021.

[YWF+21] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection, 04 2021.

[YZYC21] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection, 03 2021.

[ZYW+21] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework, 03 2021.