Draft paper about few-shot object detection

Abstract

(abstract here)

1 Introduction

Web-scale training data allowed to achieve impressive performance on vision tasks, sometimes claiming to reach or surpass human performance [1, 2, 3]. However, in some scenarios data scarcity is an unavoidable limitation. Examples of such scenarios are labeling assistants, detecting rare animals using trail camera or rare vehicles on the road [4], detecting thousands of different goods on the shelves [5, 6], and embodied agents that learn new objects from their own experience. More generally speaking, in most vision applications increasing data-efficiency would be a benefit. The current performance of vision models on few-shot learning tasks still remains very far from human level [7, 8].

Consider the following scenario: an embodied agent learns new, previously unknown objects in order to later perform some actions with them, or learn to avoid them, or draw conclusions from their presence. This is what humans frequently do throughout life. New objects may be either concept-level («a headphones») or more fine-grained («my headphones of special model») and should be learnt from one or few image or video samples, since agent learns from its visual experience. A large part of few-shot learning studies addresses image-level tasks (classification or retrieval), however, in many cases this may be not enough. Embodied agents should not only classify, but also locate the object in complex scenes and separate object instances. On the other side, pixel-level instance segmentation may be redundant, ambiguous and also hard to annotate, so approximate localization seems the optimal middle ground for testing the ability to learn new objects from few samples. Further we focus on object detection with rectangular bounding boxes.

Existing benchmarks for few-shot object detection [7, 11, 13] contain unsystematic set of classes and usually do not contain fine-grained classes. However, the few-shot object detection task can take many different forms. For example, the object of interest may be a subclass of some already known object, a part of some known object or a completely new object; the object may have exact shape and texture or may be a wide concept that exhibit high intra-class variation. The training setup may be different: the training data may come in form of iconic or in-the-wild views, text description may be provided or not, etc. The properties of test distribution may also be different: on the test set, the ability to recognize several overlapping instances may be required or not; the ability to distinguish an object from very similar objects may be required or not. If we look at many existing object detection datasets [7, 9, 10, 11, 12, 13], we found that the described factors of variation in task specifics vary greatly. We believe that different cases are better to study separately.

In this work, we collect a few-shot object detection benchmark that consists of 100 diverse classes that follow a specific taxonomy. This allow not only to get average performance, but to study model performance in several different cases. Moreover, we provide several subsets for each class to study different scenarios, for example generalizing from iconic views in the train set to in-the-wild test images. We limit ourselves to cases when objects are large and unoccluded enough to be clearly recognizable by humans. Finally, we argue that the standard mAP metrics is not suitable for few-shot object detection, since selecting the optimal threshold in few-shot regime is non-trivial and should be estimated under a given threshold value to simulate real application.

Traditionally few-shot learning requires new, unseen classes, but self-supervised or image-text pretrained models literally have seen everything in varying degrees. So, it is not clear what is meant by «seen» and «unseen» classes, and we omit this requirement. Instead, for each of 100 objects we try to collect and report its approximate frequency in common pretraining datasets.

(paragraph about model testing and results)

We also extend our benchmark to study some important questions related to the use of synthetic data. Firstly, it's interesting to check out if testing in virtual environments may replace testing on natural images. We supplement our benchmark with synthetic data of the same taxonomy and check the model performance on them. Secondly, we compare the detection performance in 3 cases: simple scenes, natural complex scenes and synthetic complex scenes.

Few-shot learning is also interesting from the theoretical point of view, since it is related to the part-whole hierarchy understanding in vision models [18]. A new, previously unknown object is usually a new combination of known parts and properties (yellow bottle with blue rectangular logo), or a part of a known object (a black cap on human's head). This suggests that if a model sees the world as parts and their combinations, learning new object will be data-efficient and parameter-efficient if the object is a part of some known object. Recently, some studies propose pretrained models that are claimed to learn robust part representations [19]. Few-shot object detection seems a good task to validate this property.

To summarize, while most previous works in few-shot vision were focused on algorithms, our work is focused mostly on data. Our goal was not to compare many few-shot algorithms, but to compare several algorithms on few-shot tasks with different specifics.

In short, our contributions are as follows:

- We provide a taxonomy of few-shot object detection tasks with 11 factors of variation.
- We collect a few-shot object detection benchmark that consists of 100 classes divided into several groups, several subsets for each class, several testing scenarios and metrics.
- (item about model testing and results)

The rest of the paper is organized as follows. In Section 2 we review the related work, in Section 3 we discuss factors of variation in object detection task and develop the structure of our benchmark, in Section 4 we describe the benchmark and data collection process, in Section 5 we describe a set of models and the test results, then we conclude in Section 6.

2 Related work

(not finished)

Existing benchmarks for few-shot object detection can be divided into two categories. The first benchmarks category is subsets of large-scale datasets LVIS [11] and COCO [14]. This category is limited in several aspects. Firstly, few-shot classes may overlap with usual pretraining classes and hence fair testing may require custom pretraining. Secondly, they do not contain diverse domains and hence does not allow cross-domain testing. Thirdly, they contain only objects of the same hierarchy level: for example, it may contain the class «human», but will never contain more fine-grained classes «human head» or «human in the red uniform».

The second category of few-shot object detection benchmarks are specifically assembled benchmarks [7, 13]. These benchmarks are currently small and not systematic: they consist of random sets of objects and their distributions, thus not allowing for systematic testing of factors of variation. For example, there is still no benchmark for few-shot detection of fine-grained classes with specific fixed shape and texture – a task that humans solve, for example, to buy items in a store and use them.

In general, every object detection dataset may be used for few-shot learning and testing. They may come in form of large-scale datasets with many classes annotated with bounding boxes [11, 12, 14, 22, 24, 25], large-scale datasets with boxes and detailed descriptions [23, 26, 27] or collections of several medium sized detection datasets [9, 10].

The few-shot object detection task is closely related to several another tasks: few-shot classification, image retrieval, long-tailed object detection and open set object detection.

Few-shot learning also relates to the distributional shift robustness. Training set contains only a few backgrounds, positions and lightning conditions, generalizing to novel object variations, context and conditions may be required for high performance on a test set.

See detailed related work in Appendix A.

3 Task specifics and taxonomy

(not finished)

Interestingly, in few-shot setting the amount of negative examples (image regions that are not an object of interest) is usually not specified at all, however, it may largely affect performance. Learning without negative samples (or with not enough of them) is a specific scenario that may require custom engineering.

4 Benchmark collection

(not finished)

We collect our data from a web-scale image-text datasets CommonPool (aka DataComp xlarge) [15] and OBELICS [16]. To collect data, we used plain text search and tried to avoid model-assisted search (for example, KNN search by CLIP embeddings) to avoid model biases; for the same reason we did not use LAION-5B dataset [17] that is already CLIP-filtered. However, text-based search induces human biases.

Using text search is difficult to collect in-the-wild images of objects in complex scenes. For example, for the query «badger» we usually find images in which the badger is large, centered and is not surrounded by other objects. However, the model performance in simple and complex scenes may be different. So, it's interesting to test if synthetic complex scenes made by concatenating photos yield the same performance or not.

This may come in forms of bounding boxes, keypoints, oriented bounding boxes, bounding circles [20], amodal 3D bounding boxes [21] etc.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv, 1502.01852.

[2] Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. arXiv, 2106.07411.

[3] Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., & Brendel, W. (2022). The bittersweet lesson: data-rich models narrow the behavioural gap to human vision. J. Vis., 22(14), 3273. doi: 10.1167/jov.22.14.3273

[4] Li, N., Song, F., Zhang, Y., Liang, P., & Cheng, E. (2022). Traffic Context Aware Data Augmentation for Rare Object Detection in Autonomous Driving. arXiv, 2205.00376.

[5] Osokin, A., Sumin, D., & Lomakin, V. (2020). OS2D: One-Stage One-Shot Object Detection by Matching Anchor Features. arXiv, 2003.06800.

[6] Chen, F., Zhang, H., Li, Z., Dou, J., Mo, S., Chen, H., ...Savvides, M. (2022). Unitail: Detecting, Reading, and Matching in Retail Scene. arXiv, 2204.00298.

[7] Lee, K., Yang, H., Chakraborty, S., Cai, Z., Swaminathan, G., Ravichandran, A., & Dabeer, O. (2022). Rethinking Few-Shot Object Detection on a Multi-Domain Benchmark. arXiv, 2207.11169.

[8] Bar, A., Wang, X., Kantorov, V., Reed, C. J., Herzig, R., Chechik, G., ...Globerson, A. (2021). DETReg: Unsupervised Pretraining with Region Priors for Object Detection. arXiv, 2106.04550.

[9] Ciaglia, F., Zuppichini, F. S., Guerrie, P., McQuade, M., & Solawetz, J. (2022). Roboflow 100: A Rich, Multi-Domain Object Detection Benchmark. arXiv, 2211.13523.

[10] Bai, H., Mou, S., Likhomanenko, T., Cinbis, R. G., Tuzel, O., Huang, P., ...Cao, M. (2023). VISION Datasets: A Benchmark for Vision-based InduStrial InspectiON. arXiv, 2306.07890.

[11] Gupta, A., Dollár, P., & Girshick, R. (2019). LVIS: A Dataset for Large Vocabulary Instance Segmentation. arXiv, 1908.03195.

[12] Zhang, Y., Sun, Q., Zhou, Y., He, Z., Yin, Z., Wang, K., ...Liu, Z. (2022). Bamboo: Building Mega-Scale Vision Dataset Continually with Human-Machine Synergy. arXiv, 2203.07845.

[13] Xiong, W. (2022). CD-FSOD: A Benchmark for Cross-domain Few-shot Object Detection. arXiv, 2210.05311.

[14] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ...Dollár, P. (2014). Microsoft COCO: Common Objects in Context. arXiv, 1405.0312.

[15] Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., ...Schmidt, L. (2023). DataComp: In search of the next generation of multimodal datasets. arXiv, 2304.14108.

[16] Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., ...Sanh, V. (2023). OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. arXiv, 2306.16527.

[17] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ...Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv, 2210.08402.

[18] Hinton, G. (2021). How to represent part-whole hierarchies in a neural network. arXiv, 2102.12627. Retrieved from https://arxiv.org/abs/2102.12627v1

[19] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ...Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. arXiv, 2304.07193.

[20] Yang, H., Deng, R., Lu, Y., Zhu, Z., Chen, Y., Roland, J. T., ...Huo, Y. (2020). CircleNet: Anchor-free Detection with Circle Representation. arXiv, 2006.02474.

[21] Qin, Z., Wang, J., & Lu, Y. (2018). MonoGRNet: A Geometric Reasoning Network for Monocular 3D Object Localization. arXiv, 1811.10247.

[22] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., ...Sun, J. (2019). Objects365: A Large-Scale, High-Quality Dataset for Object Detection. 2019 IEEE/CVF International Conference on

Computer Vision (ICCV). IEEE. doi: 10.1109/ICCV.2019.00852

[23] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ...Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Int. J. Comput. Vision, 123(1), 32–73. doi: 10.1007/s11263-016-0981-7

[24] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE. doi: 10.1109/CVPR.2009.5206848

[25] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ...Ferrari, V. (2020). The Open Images Dataset V4. Int. J. Comput. Vision, 128(7), 1956–1981. doi: 10.1007/s11263-020-01316-z

[26] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., & Murphy, K. (2015). Generation and Comprehension of Unambiguous Object Descriptions. arXiv, 1511.02283.

[27] Wang, W., Shi, M., Li, Q., Wang, W., Huang, Z., Xing, L., ...Qiao, Y. (2023). The All-Seeing Project: Towards Panoptic Visual Recognition and Understanding of the Open World. arXiv, 2308.01907.