Few-Shot Private Object Localization via Support Token Matching

Junwen Pan[1], Dawei Lu[1], Xin Wan[1], Rui Zhang[1], Kai Huang[1], Qi She[1]
[1]ByteDance, Beijing, China
panjunwen@bytedance.com

Abstract

Few-shot localization aims to locate visual objects in images using a minimal set of annotated examples, which is particularly useful for rare categories under long-tailed distributions in real-world applications. Traditionally, few-shot algorithms have utilized episode training to simulate this setting by learning a metric space with optimized network parameters. However, these methods have been demonstrated sub-optimal performance in practical scenarios, such as localizing objects in photographs taken by blind or low-vision individuals. Conversely, the field of image classification has shown that metric spaces supported by massively pre-trained models could significantly enhance few-shot performance. Based on this insight, we introduce a few-shot localization baseline leveraging the power of massive pre-trained visual features. Our solution demonstrates that large scale pretraining could cultivate a more robust metric space for few-shot localization, eliminating the complexity on trivial episode training or multi-phase fine-tuning. This approach secured the top-3 place in the few-shot private localization track of the 2024 VizWiz Grand Challenge.

1. Introduction

Object detection is a fundamental computer vision task with numerous practical applications. Region-level parsing plays a crucial role in assisting blind or low-vision users by allowing them to to magnify the regions of interests, remove unintended capture of private information, and edit their photographs [5]. However, traditional detection approaches heavily depend on extensive, well-annotated datasets, which may not always be accessible or practical for rare or novel categories. As set up by the competition of VizWiz few-shot private object localization, an intuitive approach is to construct models supported by a small number of each novel classes samples, which is known as few-shot detection.

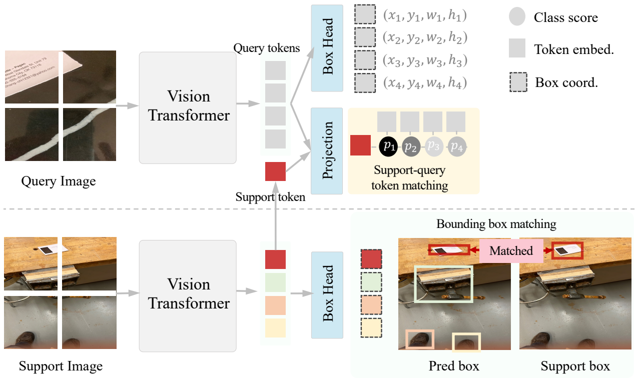

Figure 1. Overview of the Support Token Matching for few-shot localization. STM first matches the support boxes with class-agnostic boxes predicted by the foundation model. Then, the visual tokens belonging to the matched boxes are acquired and matched with the query image tokens for class-aware detection.

Previous few-shot detection approaches can be broadly categorized into fine-tuning-based [5], meta-learning-based [1] and foundation-model-based strategies [6]. Popular fine-tuning methods use a fairly simple two-stage approach, whereby the network is first trained on base categories and then partially fine-tuned on a mixture of base and novel categories. Despite their simplicity, these methods often struggle to maintain performance on base categories while prevent performance catastrophic forgetting on novel categories, posing a challenge for their application in real-world scenarios. Meta-learning based strategies utilize episode training to learn network parameters and classify query images with constructed prototypes from novel category examples, whose underlying assumption is metric learning, *i.e.*, learning more discriminative features and thus better distinguishing between different categories. However, while theoretically sound, these strategies has shown limitations in practical applications [5]. Recently, foundation models, *i.e.*, models with a large number of parameters pre-trained on noisy webly supervised or unsupervised datasets, have shown strong performance in few-shot learning [2], since such massive training yields a more general and robust metric space [4, 7]. Several explorations have shown that these foundational models also hold significant potential for few-shot detection [6]. However, harnessing this potential requires the integration of intricate region propagation techniques and post-processing steps which converting masks into bounding boxes. In VizWiz few-shot localization challenge, we propose to build a simple yet effective baseline based on the foundation model to get rid of the complexity from episodic training, or bounding box converting steps.

2. Method

Figure 1 illustrates the Support Token Matching (STM) process for few-shot localization, comprising two primary steps: 1) selecting support tokens based on box matching, and 2) detecting objects in query images through support-query token matching. It's important to note that this method presumes the foundation model effectively predicts the bounding boxes of arbitrary objects without relying on category-specific priors. Since DETR and vision-language pre-training are the de facto state-of-the-art in the fields of detection and few-shot learning respectively [7], we choose the OWL-ViT archetecture [3], a DETR-like model with a vision-language pre-trained Vision Transformer, as our foundation model.

Bounding Box Matching. The goal of bounding box matching is to apply the support box provided by the human to construct the corresponding prototype features as a classifier of the category. To this end, we first establish the spatial relationship between human annotations and the predicted boxes, which in turn enables us to find the token in the feature space that represents the support object. Specifically, we compute the Intersection over Union (IoU) for the given support box against all predicted boxes as the matching metric.

Support Token Selection. Following bounding box matching, an adaptive IoU threshold is set at 80% of the maximum IoU score to identify relevant predicted boxes. This threshold ensures the selection includes all boxes that substantially overlap with the highest matching box, maintaining focus on the most relevant spatial features. The final selection of the support tokens is determined by comparing the embeddings of the boxes that meet this IoU threshold. To identify the most representative token, we assess the dissimilarity of each token against the mean of candidates. The token exhibiting the least similarity to the mean—indicating distinctiveness—is selected as the optimal representative for the object.

Support-Query Token Matching. Once selecting the support tokens representing the object categories, the next stage is to match these tokens with the embeddings extracted from the query image. We derive the classification score by computing cosine similarity between each query token and the selected support embedding for each predicted bounding boxes on the query set.

Table 1. Results on *novel query* set of few-shot private object localization track.

Foundation Model	Ensemble	mAP_{50}	$\text{mAP}_{50:95}$
ViT-Base	✓	28.53	19.86
ViT-Large		22.22	14.90
ViT-Large	✓	**30.34**	20.08

3. Experiments

Results. The VizWiz few-shot private object localization dataset comprises 16 private object categories. The challenge server conducts evaluations in a 1-shot setting on the novel query set, which contains 1,072 query images. As shown in Table 1, the STM strategy—without complex training or post-processing—achieved decent performance. Increasing the model size led to a 1.83% improvement in mean Average Precision (mAP), validating the scaling law of foundation models in few-shot detection.

Discussion. We conducted a preliminary evaluation of the category leakage in pre-training datasets and found that 1/16 categories was explicitly trained on. Similar to challenges faced in the community of zero-&few-shot classification [2], stringent evaluations have become increasingly challenging.

4. Conclusion

This report proposes a preliminary few-shot detection baseline, STM, based on a foundational detection model. The generality and robustness of the metric space from the foundational model were verified in the few-shot detection task.

References

[1] Mona Köhler and et.al. Few-shot object detection: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[2] Zhiqiu Lin, Samuel Yu, and et.al. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *CVPR*, 2023. 1, 2

[3] Neil Houlsby Matthias Minderer, Alexey Gritsenko. Scaling open-vocabulary object detection. *NeurIPS*, 2023. 2

[4] Maxime Oquab and et.al Darcet. Dinov2: Learning robust visual features without supervision, 2023. 1

[5] Yu-Yun Tseng, Alexander Bell, and Danna Gurari. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *ECCV*, 2022. 1

[6] Abdeslam Boularias Xinyu Zhang, Yuting Wang. Detect everything with few examples. *arXiv:2309.12969*, 2023. 1

[7] Renrui Zhang, Rongyao Fang, and et.al. Tip-adapter: Training-free clip-adapter for better vision-language modeling. abs/2111.03930, 2021. 1, 2