A simple technical report about the Foundational Few-Shot Object Detection Challenge

Qibo Chen, Jianyue Ge, Weizhong Jin, Li Yu
China Mobile(Zhejiang) Research & Innovation Institute
c18857824339@gmail.com

Abstract

A technical report on our using method on the Foundational Few Shot Object Detection Challenge, divided into 5 parts: model, pre training data, fewshot fine-tuning method, sample completion strategy

1. Model

Our model Instruction DINO (ISD) is an unpublished algorithm, so we will only provide a general introduction to the model. Our model is based on the DETR[1] detector architecture and achieves open set object detection by introducing text modality. Our model refers to the early fusion of image and text information in the encoding section of Grounded DINO[2]. Our practical experience shows that this not only reduces the difficulty of image text modal alignment during the training process, but also significantly improves prompt tuning effect after pre-training. Our visual backbone uses Swin L, and the text encoder uses EVA02 CLIP L[3].

2. Pre-training data

ISD model pre-training data includes: O365v2[4], COCO2017, LVIS[5], GoldG[6], VG[7], OpenImages-V6[8], V3Det[9], PhraseCut[10], RefCOCO[11], RefCO-CO+[11], RefCOCOg[11], gRef-COCO[12],

We divide the training into two stages. The first stage only trains the detection data, and the second stage uses all the data. ISD uses sentence level text representation vectors, so we have done some processing on the ground-ing format data. We transformed the grounding training format into a description of a single object using QWen Max[13] on the Flicker dataset in GoldG. The GQA[14] in goldG is regenerated through scene maps.

3. Fewshot fine-tuning method

We visualized the official provided 10 shot training JSON and found that there will only be one category of annotations in a single image, even if other categories of objects are widely present in the image. This special situation makes the traditional closed set fine-tuning method unable to work properly, as it requires training all categories. Therefore, we used a flexible training format to fine tune the model. We have tried two negative sample strategies: the first is to randomly sample an indefinite number of negative sample texts from the remaining category list during the training process, and the second is to use VLM (CLIP[15], TAP[16], LLava[17]) to generate reliable negative sample texts for different images through predefined word list classification and generation methods. Through practice, Strategy 2 can improve the final indicator by 1.8 mAP compared to Strategy 1. Of course, these negative sample strategies are mainly due to our model using early fusion, which makes it more sensitive to negative samples during model training compared to post fusion algorithms such as OWL-ViT[18].

In addition, during the fine-tuning process, we also tried three settings. Setting 1 only fine-tuned the visual part, which includes visual encoder, neck, encode, decode. Setting 2 only fine-tuned the text encoder. Setting 3 prompt tuning. In the training process, we found an interesting phenomenon that only fine-tuning the visual part perform-ed poorly in leaderboard test but had better visual perfor-mance in the training set. This indicates that fine-tuning the visual part of the pre-trained model is more prone to overfitting data, while fine-tuning the text encoder and prompt tuning will have better generalization in the test set. Our prompt tuning uses a text encoder for initialization and multiple vectors to represent a single category. Compared to fine-tuning the text encoder, it introduces more text context information, which makes prompt tuning perform best in the test set.

Therefore, our fine-tuning method ultimately adopts prompt fine-tuning and negative sample sampling strategy two. This fine-tuning method helped our model improve 5~6mAP on the baseline of zero shot testing. We are unable to provide specific values as the ranking has already been closed.

4. Sparse annotation completion

In the visualization of 10 shot training data, we found that the visualization effect of only fine-tuning the visual part was significantly better than that of prompt tuning and only fine-tuning the text encoder. We suspect that it is

caused by sparse annotations in the training data, so we first train the visual part and use it to reason on the training data, set a threshold of 0.7, and combine the original 10 shot annotation information for completion. After obtaining the complete annotation file, we trained it through prompt tuning in the third section and aggregate the prompt vectors of the last 5 epochs.

By completing the annotations, we achieved a further improvement of 6~7mAP and ultimately achieved 31.574 mAP.

References

[1] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[C] International Conference on Learning Representations. 2020.

[2] Liu S, Zeng Z, Ren T, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection[J]. arXiv preprint arXiv:2303.05499, 2023.

[3] Fang Y, Sun Q, Wang X, et al. Eva-02: A visual representation for neon genesis[J]. arXiv preprint arXiv:2303.11331, 2023.

[4] Shao S, Li Z, Zhang T, et al. Objects365: A large-scale, high-quality dataset for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8430-8439.

[5] Gupta A, Dollar P, Girshick R. Lvis: A dataset for large vocabulary instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5356-5364.

[6] Kamath A, Singh M, LeCun Y, et al. Mdetr-modulated detection for end-to-end multi-modal understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1780-1790.

[7] Krishna R, Zhu Y, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 2017, 123: 32-73.

[8] Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[J]. International journal of computer vision, 2020, 128(7): 1956-1981.

[9] Wang J, Zhang P, Chu T, et al. V3det: Vast vocabulary visual detection dataset[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 19844-19854.

[10] Wu C, Lin Z, Cohen S, et al. Phrasecut: Language-based image segmentation in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10216-10225.

[11] Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 69-85.

[12] Liu C, Ding H, Jiang X. Gres: Generalized referring expression segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 23592-23601.

[13] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[14] Hudson D A, Manning C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6700-6709.

[15] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[16] Pan T, Tang L, Wang X, et al. Tokenize Anything via Prompting[J]. arXiv preprint arXiv:2312.09128, 2023.

[17] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.

[18] Minderer M, Gritsenko A, Stone A, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022[J]. arXiv preprint arXiv:2205.06230, 2022, 2.