RegionCLIP: Region-based Language-Image Pretraining

Yiwu Zhong[1]*, Jianwei Yang[2], Pengchuan Zhang[2], Chunyuan Li[2], Noel Codella[3],
Liunian Harold Li[4], Luowei Zhou[3], Xiyang Dai[3], Lu Yuan[3], Yin Li[1], Jianfeng Gao[2]
[1]University of Wisconsin-Madison, [2]Microsoft Research, [3]Microsoft Cloud + AI, [4]UCLA
{yzhong52, yin.li}@wisc.edu,{jianwei.yang, penzhan, chunyl, ncodella,
luozhou, xidai, luyuan, jfgao}@microsoft.com, {liunian.harold.li}@cs.ucla.edu

Appendices

In appendices, we provide additional information for our main paper, including implementation details of network architecture, visualization of zero-shot inference with a large number of target categories from LVIS dataset, and experiment results for ablation study.

A. Additional Implementation Details

During pretraining, the visual encoder of our model was initialized by a pretrained CLIP model. During transfer learning, the visual backbone of our detector was initialized by our pretrained visual encoder. Both our visual encoder and backbone adopt the implementation from CLIP [2] whose ResNet is slightly different from the standard ResNet [1]. According to the public code base of CLIP[1], there are three architecture changes: (1) three "stem" convolutions with an average pooling are used instead of one "stem" convolution with a max pooling. (2) Anti-aliasing strided convolutions were used where an average pooling was prepended to convolutions with stride larger than 1. (3) The final pooling layer is a self-attention layer instead of an average pooling.

B. Additional Visualization

Our pretrained models can predict the customized object concepts by simply replacing the language embeddings of target categories. Fig. 1 visualizes results of zero-shot inference with ground-truth boxes and 1203 categories from LVIS dataset, instead of the small set of 65 categories from COCO dataset. We show the *top-3* predictions for each region with their confidence scores.

As shown by the successful cases in Fig. 1, our pretrained model can correctly recognize the image regions while the CLIP model often fails to predict the correct labels (*e.g.*, "teddy bear" is predicted by our model with a high confidence score 99.5%). Interestingly, other than the most-confident category, our model can also predict reasonable categories with top-3 scores (*e.g.*, "bear" in 1st exam-

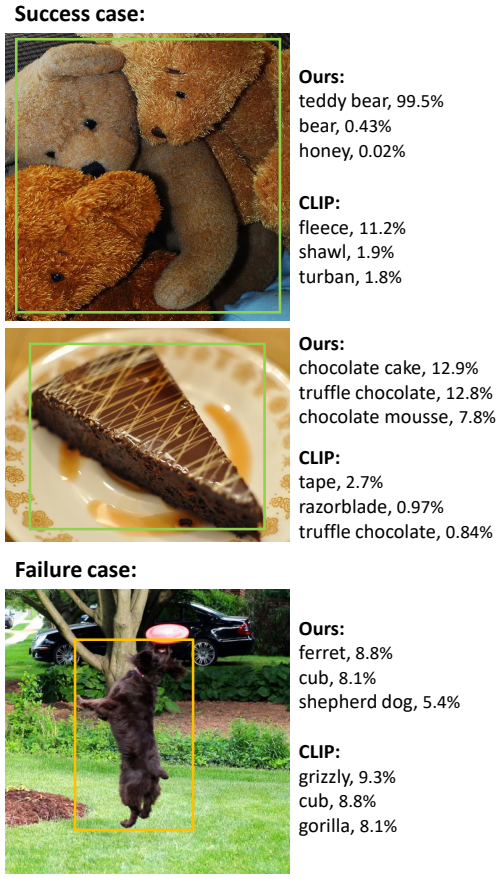

Figure 1. Visualization of zero-shot inference on COCO dataset with *ground-truth boxes*. Without finetuning, the pretrained models are asked to predict 1203 categories from LVIS dataset. We show the top-3 predicted categories from our pretrained model and pretrained CLIP model. (Image IDs: 776, 13597, 17029)

*Work done as an intern at Microsoft Research.
[1]https://github.com/openai/CLIP

Pretraining Dataset	Concept Pool Source	COCO Zero-shot Inference		COCO Generalized (17+48)		
		All (RPN)	All (GT)	Novel	Base	All
COCO Cap	COCO Cap	28.0	62.8	26.8	54.8	47.5
CC3M	COCO Cap	26.8	61.4	31.4	57.1	50.4
CC3M	CC3M	26.5	60.8	29.1	56.0	49.0

Table 1. Ablation study on the pretraining datasets and the source of concept pool.

Teacher Backbone	Student Backbone	COCO Zero-shot Inference		COCO Generalized (17+48)		
		All (RPN)	All (GT)	Novel	Base	All
RN50	RN50	26.8	61.4	31.4	57.1	50.4
RN50x4	RN50	29.3	65.6	30.8	57.3	50.4
RN50x4	RN50x4	29.6	65.5	39.3	61.6	55.7

Table 2. Ablation study on COCO with different teacher and student models in pretraining. All models are pretrained on CC3M.

Focal Scaling	COCO Generalized (17+48)		
	Novel	Base	All
	22.6	58.5	49.1
✓	31.4	57.1	50.4

Table 3. Ablation study on effects of focal scaling during transfer learning for object detection.

ple and "truffle chocolate" in 2nd example). Even in the failure case where both CLIP and our model fail to recognize the dog as most-confident category, our model can still recognize the image region as visually similar concepts (*e.g.*, "ferret" and "cub") or a fine-grained type of dog (*e.g.*, "shepherd dog"). On the contrary, CLIP predicts less visually similar concepts, such as "grizzly" and "gorilla".

C. Additional Ablation Study

We report additional ablation studies following the ablation setup in our main paper, and report results on both transfer learning and zero-shot inference.

Pretraining dataset and concept pool. Table 1 probes into the effects of pretraining dataset and concept pool. Using COCO Cap or concepts from COCO achieves better results for zero-shot inference (62.8 vs. 61.4 vs. 60.8 AP50 with GT boxes), as COCO Cap shares the same images as the detection task. However, the model pretrained on CC3M achieves significant boost on transfer learning (50.4 vs. 47.5 vs. All AP50), potentially due to its exposure to rich visual concepts in CC3M.

Teacher model and student model. Table 2 studies the effects of using different teacher and student models. Compared with the default setting at first row, using ResNet50x4 as the teacher model can largely improve the zero-shot inference performance (+4.2 AP50 with GT boxes). However, in the transfer learning setting, the performance using a stronger teacher remains roughly the same (both are 50.4 AP50 for All). When we further replace the student model with ResNet50x4, the transfer learning performance is significantly boosted (+5.3 AP50 for All), but the zero-shot inference performance remains (29.6 vs. 29.3 AP50 with RPN boxes). Based on these results, we conjecture that zero-shot inference performance relies on the teacher model that guides the region-text alignment, while transfer learning is more likely constrained by the capacity of student model.

Focal scaling. Table 3 studies the effects of focal scaling during transfer learning. With focal scaling, the finetuned detector achieves a better balance between novel categories and base categories on COCO dataset. We conjecture that the detector overfits to the small set of base categories in COCO (*e.g.*, 48 base categories), which hurts the generalization on novel categories. Focal scaling effectively alleviates the potential overfitting.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.