Hyperbolic Learning with Synthetic Captions for Open-World Detection

Fanjie Kong[1]*, Yanbei Chen[2]†, Jiarui Cai[2], Davide Modolo[2]
[1]Duke University [2]AWS AI Labs
fk43@duke.edu, {yanbec,cjiarui,dmodolo}@amazon.com

A. Implementation details

In Table A, we present the learning schedules when fine-tuning on different pre-training datasets. Note that when adding GoldG, or RefC in row 2, 3, our model is initialized from the checkpoints in row 1, 2 respectively to continue model training. In Table B, we present the learning schedules when fine-tuning on different downstream datasets. During both pre-training and fine-tuning, the learning rate is decayed by 0.1 at 1/3 and 2/3 of the training progress.

Model	Dataset(s)	Iterations	batch size	learning rate
HyperLearner	O365	900,000	64	0.0002
	O365,GoldG	+500,000	64	0.0002
	O365,GoldG,RefC	+500,000	64	0.0002

Table A. Learning schedules on different pre-training datasets.

Model	Dataset(s)	Iterations	batch size	learning rate
HyperLearner	COCO	100,000	64	0.0001
	ODinW	50,000	64	0.0001
	RefCOCO	100,000	32	0.0001

Table B. Learning schedules on different downstream datasets.

B. Computation cost of hyperbolic learning

To mitigate the high computational cost of hyperbolic learning, we make the following design choices. First, we choose to only lift the last-layer Euclidean features into hyperbolic space, instead of using a hyperbolic neural network. This approach circumvents computationally intensive operations, such as Frechet mean calculations, as noted in [5].

Second, we adopt the Lorentz model as [8], rather than the Poincaré ball model employed in [2]. The Poincaré ball model is operationally intensive due to its computationally heavy Gyrovector operations, which necessitate a series of intermediate calculations. In contemporary deep learning frameworks, these calculations substantially increase the overhead of computation graph, as described in [3]. In contrast, the Lorentz model offers a more computationally efficient alternative for hyperbolic geometry, which employs four-vector operations (Euclidean axes plus a time dimension), thus learning only one additional dimension. Notably, the time dimension can be readily determined based on the curvature of the hyperboloid, as detailed in [2].

Third, thanks to the two design choices made above, the only additional computational costs are hyperbolic functions, including "sinh" and "cosh". Fortunately, PyTorch efficiently optimizes these operations, particularly through techniques such as operation fusion. This technique combines multiple pointwise operations – including arithmetic and trigonometric functions – into a single kernel, thus further enhancing computational efficiency, as detailed in [2].

C. Discussion on evaluation benchmarks

Open-vocabulary detection evaluation benchmarks. It is worth noting that existing works in open-vocabulary detection (OVD) [1, 4, 6, 11–15] consider the evaluation setting where only *novel* (*unseen*) object classes are presented in the evaluation benchmarks. Specifically, the OVCOCO and OVLVIS benchmarks are built by splitting the class labels into two sets: seen and unseen. The model is trained on the seen classes and tested on the unseen classes.

Open-world detection evaluation benchmarks. In contrast, our open-world detection task consider the evaluation setting where any object classes (both seen and unseen) are presented in the evaluation benchmarks, which aligns with detection in the open world. Moreover, these object classes could be described by keywords (class labels) or free-form texts. Our evaluation setting also poses a unique challenge of understanding novel concepts in free-form texts.

D. Additional results

Additional quantitative results on ODinW. Table C shows the additional results of 13 individual datasets on ODinW benchmark of our proposed approach, when using different evaluation modes (zero-shot or fine-tune/full-shot) and different pre-training data. Results show the benefit of using more data for pre-training, and further fine-tuning on different target datasets to improve generalization.

Additional qualitative results. Figure A presents more qualitative examples using our model to localize objects described using free-form texts to specify certain object at-

*Work done during an internship with AWS AI Labs.

†Corresponding author.

Model	Evaluation Mode	Data	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	PascalVOC	pistols	pothole	Raccoon	Shellfish	Thermal	Vehicles	mAP
HyperLearner	zero-shot	O365	16.63	27.35	72.53	12.47	41.08	63.68	55.88	23.03	15.31	42.68	21.50	40.51	59.92	36.05
	fine-tune	O365	41.53	50.72	75.59	68.41	83.45	72.34	84.39	99.77	52.08	61.00	43.09	75.35	59.42	67.31
	zero-shot	O365, GoldG	27.25	35.34	74.38	18.12	58.50	68.95	64.12	27.03	17.06	56.01	29.36	60.72	60.24	44.74
	fine-tune	O365, GoldG	42.48	52.10	76.34	74.66	90.43	76.53	87.94	99.77	53.00	63.58	39.52	78.52	61.25	69.57

Table C. Detailed results of zero-shot and fine-tune transfer on 13 individual datasets on ODinW. Metric: mAP over 13 datasets.

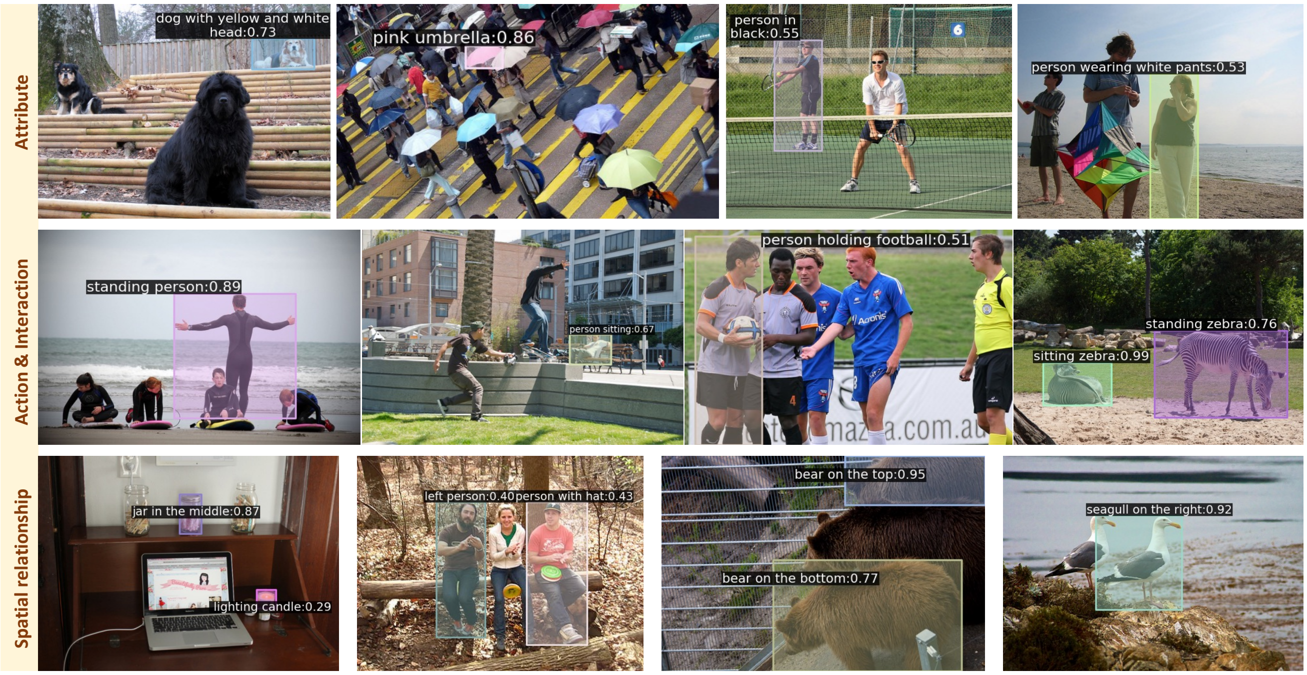

Figure A. More qualitative results on open-world detection using free-form texts, where the free-form texts specify objects with attribute, action/interaction, and spatial relationships.

tribute, action/interaction, and spatial relationships. Examples illustrate the strong generalizability of our model in understanding and recognizing novel concepts in open world.

E. Discussion on quality of synthetic captions

Generally, our VLM, BLIP2 6.7b, generates synthetic captions that accurately describe the objects in the image, as indicated by [6]. Furthermore, our experimental result reveals that the synthetic captions generated by VLM can significantly improve the performance of our model across various open-world object detection benchmarks. These observations qualitatively demonstrates the quality of synthetic captions for training open-world object detection models. In this section, we carry out a quantitative study for the quality of synthetic captions generated by VLMs, by determining the ratio of noisy or hallucinated objects in generated captions.

To quantify the noise level in our synthetic captions (i.e., % incorrect objects described), we adopt the Caption Hallucination Assessment with Image Relevance (CHAIR) metric from [7, 10] as our noise-level measurement for synthetic captions. It can be formulated as the percentage of incorrectly described objects,

$$\text{noise } \% = \frac{1}{m}\sum_{i=1}^{m}\frac{N_{\text{incorrect objects}}}{N_{\text{total objects}}} \times 100\% \tag{1}$$

,where $N_{\text{incorrect objects}}$ denotes the number of hallucinated objects in the caption, $N_{\text{total objects}}$ denotes the number of total objects in the caption, and m is the total number of synthetic captions collected from the dataset. To obtain $N_{\text{incorrect objects}}$ and $N_{\text{total objects}}$ in (Eq. (1)), we detect the synonyms of object classes within the synthetic captions using WordNet synsets [9]. If an object or its synonym is mentioned in the caption, but the detection ground truth does not include this object, it is classified as an incorrect object. Our result indicates a noise level of 16.3% on COCO 2017 test set, using BLIP2 6.7b. Despite the presence of noise and hallucinated contents in the synthetic captions, we have proposed a hyperbolic learning strategy to mitigate the issue of training with noisy synthetic captions. The hyperbolic learning loss learns visual and caption embeddings hierarchically, imposing an entailment relationship between captions and region embeddings, rather than directly aligning them. Moreover, we suggest that employing an enhanced pre-trained VLM for captioning could further mitigate the noise level.

F. Future work

Building upon our current work in open-world object detection, there are several promising avenues for future exploration. First, it is essential to develop a new benchmark that evaluates open-world object detection using both keywords and free-form text, overcoming the limitations of current benchmarks that merely focus on either keywords (e.g., COCO, LVIS, ODinW) or free-form texts (e.g., RefCOCO). Second, training image captioning models that generate object-centric synthetic captions presents an intriguing research direction to provide better open-world knowledge in object detection. Third, integrating various data modalities or information sources to augment the model's detection capabilities is also an interesting direction, such as providing a sketch of a target object to localize in the open world.

References

[1] Han-Cheol Cho, Won Young Jhoo, Wooyoung Kang, and Byungseok Roh. Open-vocabulary object detection using pseudo caption labels. *arXiv preprint arXiv:2303.13040*, 2023. 1

[2] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023. 1

[3] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, 2023. 1

[4] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 1

[5] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *CVPR*, 2022. 1

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2

[8] Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the fréchet mean. In *ICML*, 2020. 1

[9] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[10] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2

[11] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 1

[12] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, 2023.

[13] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022.

[14] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021.

[15] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 1