Hyperbolic Contrastive Learning for Visual Representations beyond Objects

Poster-TUE-PM-259

IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023

Songwei Ge[1*], Shlok Mishra[1*], Simon Kornblith[2], Chun-Liang Li[2], David Jacobs[1]

[1]University of Maryland, College Park [2]Google Research

Motivation

- We are using hyperbolic loss to learn this structure from real-world images!
- Representation learned by our model.

	Pre-train	Bbox	VOC	IN-100	IN-1k
MoCo-v2	COCO	-	64.79	64.84	51.17
HCL w/o $\mathcal{L}_{hyp}$	COCO	SS	73.13	73.84	54.21
HCL w/o $\mathcal{L}_{hyp}$	COCO	GT	75.55	76.22	54.52
HCL	COCO	SS	74.19	75.16	55.03
HCL	COCO	GT	**76.51**	**76.74**	**55.63**
MoCo-v2	OpenImages	-	69.95	72.80	54.12
HCL w/o $\mathcal{L}_{hyp}$	OpenImages	SS	71.82	75.33	56.58
HCL w/o $\mathcal{L}_{hyp}$	OpenImages	GT	73.79	77.36	57.57
HCL	OpenImages	SS	74.31	78.14	58.12
HCL	OpenImages	GT	**75.40**	**79.08**	**58.51**

- Linear evaluation results.

- Object images of the same class tend to gather near the center around similar directions, while the scene images are far away in these directions with larger norms.

Motivation

Motivation

Motivation

Motivation

- Objects from visually similar classes lie close to each other in the representation space.

* Wu, Zhirong, et al. "Unsupervised feature learning via non-parametric instance discrimination." *CVPR* 2018.

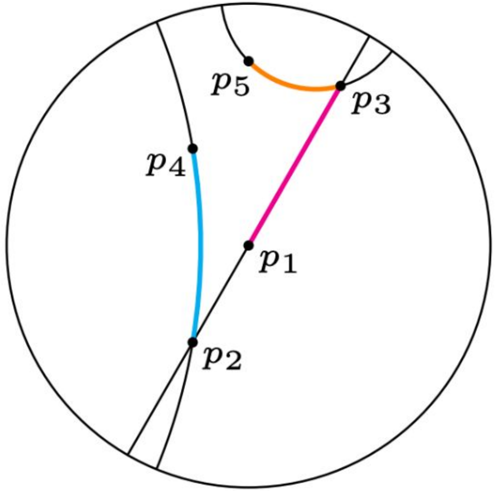
- Real world images have much more diversity and structure in them as compared to ImageNet images.

OpenImages Samples

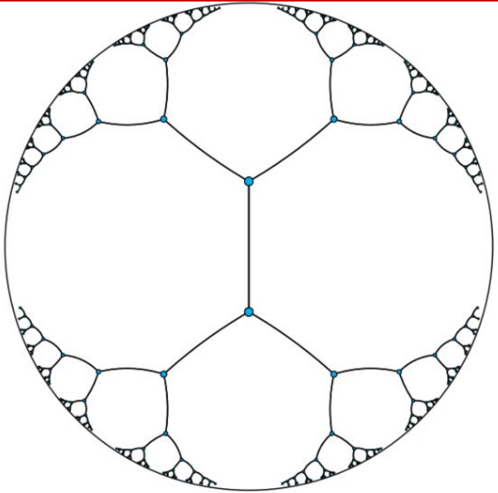
Background

- Hyperbolic distance:

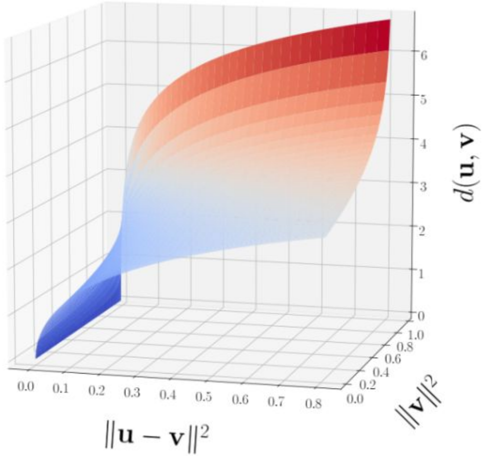
$$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)}\right).$$



(a) Geodesics of the Poincaré disk



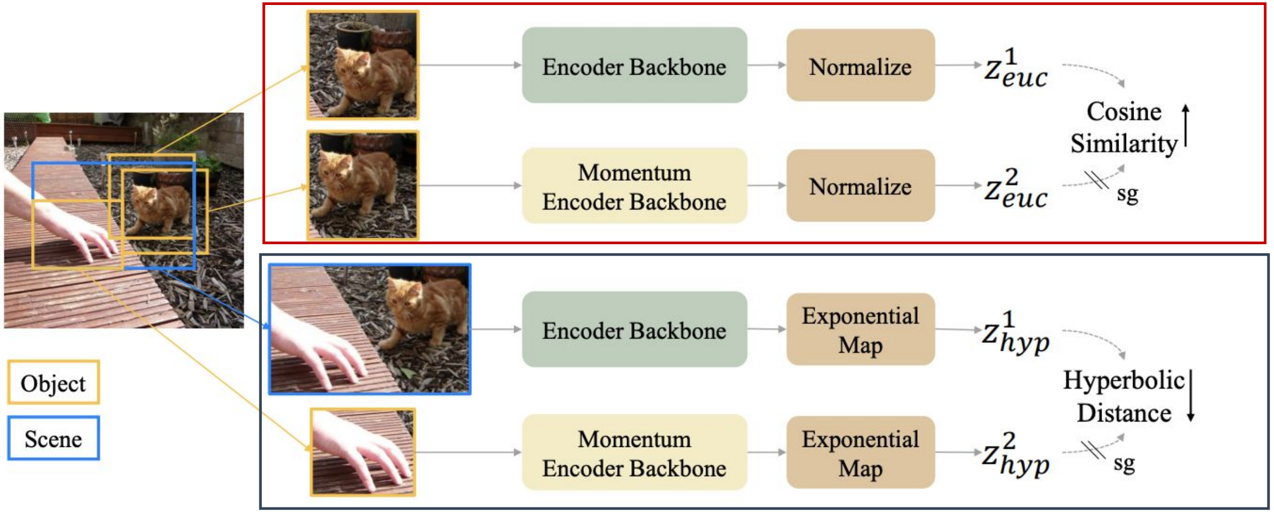
(b) Embedding of a tree in $\mathcal{B}^2$



(c) Growth of Poincaré distance

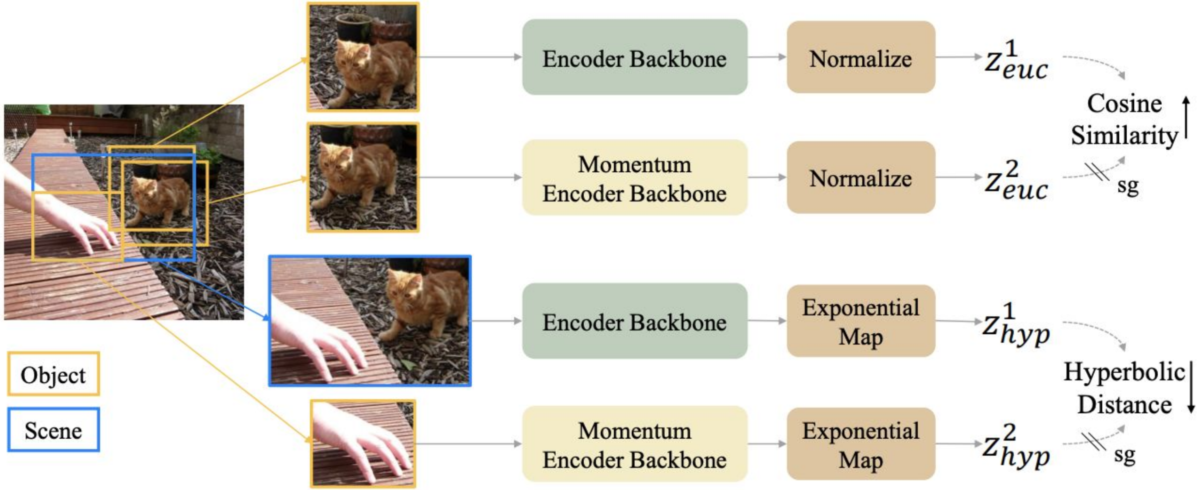
* Nickel, Maximillian, and Douwe Kiela. "Poincaré embeddings for learning hierarchical representations." *NeurIPS. 2017*

Proposed approach



- Two object regions are cropped to learn object representations with the euclidean loss.
- Scene region with a contained object region is used to learn scene representations using hyperbolic space.

Proposed approach



$$\mathcal{L}_{\text{hyp}} = -\log \frac{\exp\left(-\frac{d_{\mathbb{D}}(\mathbf{z}^1_{\text{hyp}}, \mathbf{z}^2_{\text{hyp}})}{\tau}\right)}{\exp\left(-\frac{d_{\mathbb{D}}(\mathbf{z}^1_{\text{hyp}}, \mathbf{z}^2_{\text{hyp}})}{\tau}\right) + \sum_n \exp\left(-\frac{d_{\mathbb{D}}(\mathbf{z}^1_{\text{hyp}}, \mathbf{z}^n_{\text{hyp}})}{\tau}\right)}$$

- Here z^1_{hyp} and z^2_{hyp} are the projected features on the Poincare ball.
- d_D is the riemannian distance on the Poincare ball.

Results using HCL.

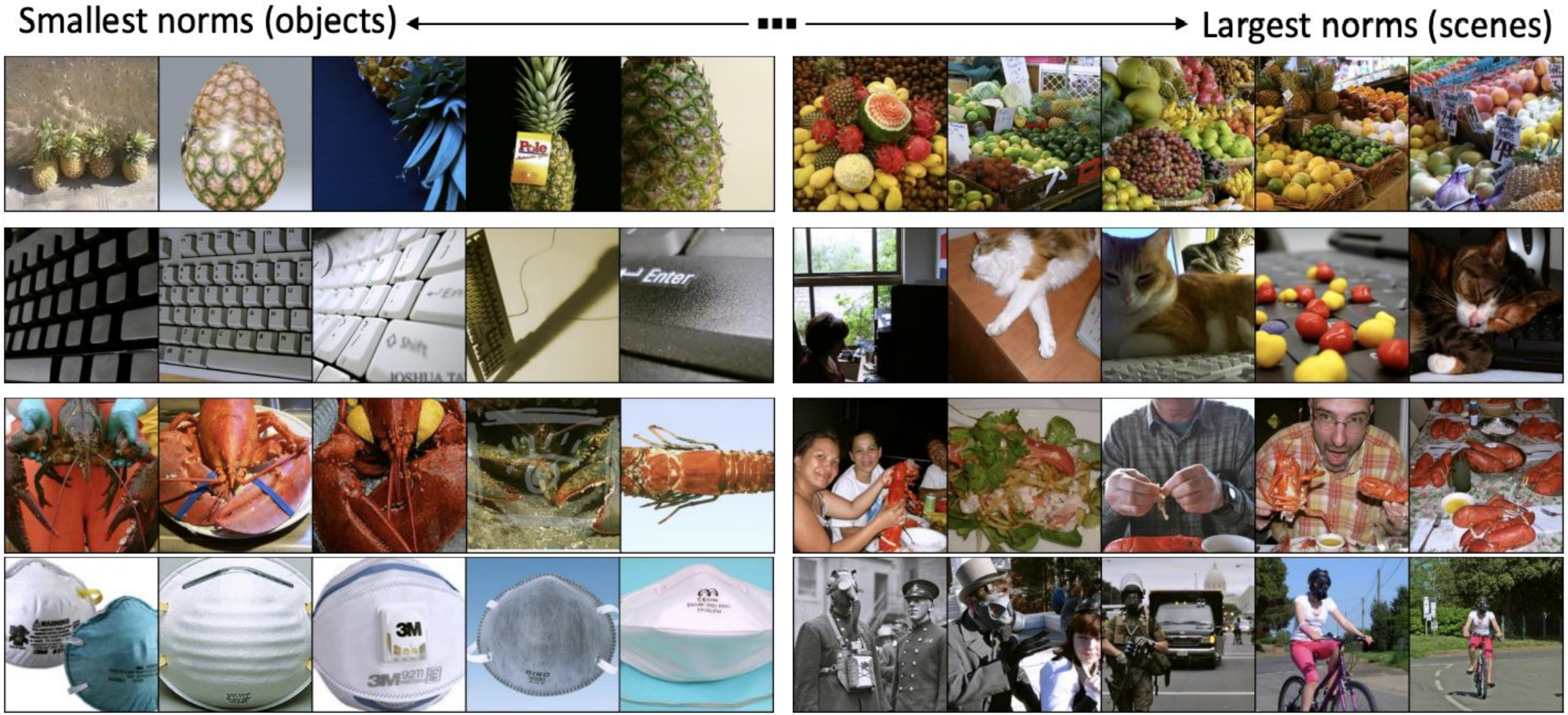
	Pre-train	Bbox	VOC	IN-100	IN-1k
MoCo-v2	COCO	-	64.79	64.84	51.17
HCL w/o $\mathcal{L}_{\text{hyp}}$	COCO	SS	73.13	73.84	54.21
HCL w/o $\mathcal{L}_{\text{hyp}}$	COCO	GT	75.55	76.22	54.52
HCL	COCO	SS	74.19	75.16	55.03
HCL	COCO	GT	**76.51**	**76.74**	**55.63**
MoCo-v2	OpenImages	-	69.95	72.80	54.12
HCL w/o $\mathcal{L}_{\text{hyp}}$	OpenImages	SS	71.82	75.33	56.58
HCL w/o $\mathcal{L}_{\text{hyp}}$	OpenImages	GT	73.79	77.36	57.57
HCL	OpenImages	SS	74.31	78.14	58.12
HCL	OpenImages	GT	**75.40**	**79.08**	**58.51**

- Linear evaluation results.

	AP^{b}	$\text{AP}^{\text{b}}_{50}$	$\text{AP}^{\text{b}}_{75}$	AP^{m}	$\text{AP}^{\text{m}}_{50}$	$\text{AP}^{\text{m}}_{75}$
MoCo-v2 pre-trained on COCO:						
Baseline	38.5	58.1	42.1	34.8	55.3	37.3
HCL w/o $\mathcal{L}_{\text{hyp}}$	39.7	60.1	43.4	36.0	57.3	38.8
HCL CC	**40.6**	**61.1**	**44.5**	**37.0**	**58.3**	**39.7**
Dense-CL pre-trained on COCO:						
Baseline	39.6	59.3	43.3	35.7	56.5	38.4
HCL w/o $\mathcal{L}_{\text{hyp}}$	41.3	61.5	44.7	37.5	59.5	40.4
HCL	**42.5**	**62.5**	**45.8**	**38.5**	**60.6**	**41.4**
ORL pre-trained on COCO:						
Baseline	40.3	60.2	44.4	36.3	57.3	38.9
HCL	**41.4**	**61.4**	**45.5**	**37.3**	**58.5**	**40.0**
Dense-CL pre-trained on OpenImages:						
Baseline	38.2	58.9	42.6	34.8	55.3	37.8
HCL w/o $\mathcal{L}_{\text{hyp}}$	41.1	61.5	44.4	37.2	58.3	39.7
HCL	**42.1**	**62.6**	**45.5**	**38.3**	**59.4**	**40.6**

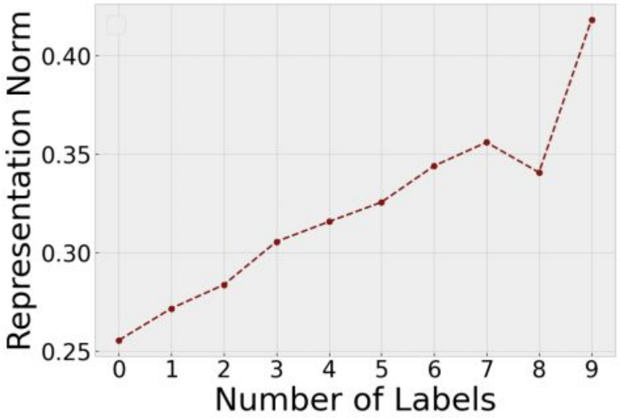
- Object Detection and Semantic Segmentation results on COCO.

Analysis



- The 5 images on the left have the smallest representation norms among all the images from the same class, and the 5 on the right have the largest norms.

Label Uncertainty quantification



- Average representation norms of images with different number of labels in ImageNet-ReaL