Hyperbolic Contrastive Learning for Visual Representations beyond Objects

Songwei Ge, Shlok Mishra, et al.

University of Maryland, Google
2023 IEEE/CVPR

October 18, 2024

Presented by Angel Huang

Introduction

Purpose

- The paper aims to improve contrastive learning for multi-object scenes.

Intuition

- In addition to making single-object representations close when objects are similar, this paper seeks to encourage multi-object scenes that share similar objects to also be close in the representation space.
- **Problem**: A given number of objects can be composed into exponentially many possible scenes, making it challenging to learn scene representations in the Euclidean space.
- **Solution**: The proposed method maps the scene representations to a hyperbolic space that follow a hierarchical structure, which is better at handling combinatorial explosion, and proposes a hyperbolic loss.

Potential Applications (Computer vision)

- image classification
- object detection
- semantic segmentation
- vision tasks that involve the interaction between scenes and objects in a zero-shot fashion

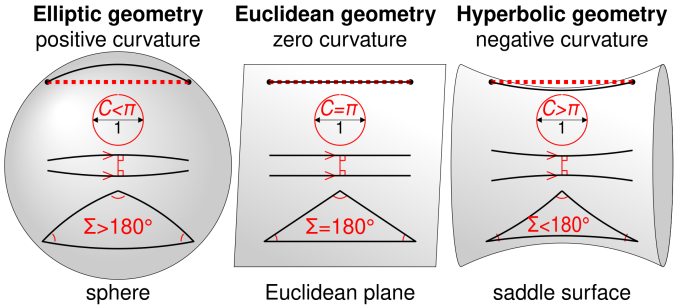
Background

Contrastive Learning Framework

- **Objective**: Learn representations by pulling positive pairs (similar samples) close together and pushing negative pairs (dissimilar samples) apart in the embedding space.
- **Self-supervised**: No labels needed—positive pairs can be created using augmentations
- Object representations are learned using a **Euclidean contrastive loss**.
- **Combinatorial explosion problem**: In tasks involving scenes (multiple objects), the number of possible combinations increases exponentially, which requires extremely large dimensions in Euclidean space to avoid distortion.
- **Hyperbolic Geometry**: Provides an efficient way to embed tree-like or hierarchical structures.

Background

Hyperbolic Space is a complete, connected Riemannian manifold with constant negative sectional curvature.



Poincaré ball is one of the commonly used hyperbolic models. It can be viewed as a natural counterpart of the hypersphere as it allows all directions, unlike the other models that have constraints on the directions.

Method

Riemannian Distance in Poincaré ball

$$d_{\mathbb{D}}(p, q) = 2r \tanh^{-1}\left(\frac{\| -p \oplus q\|}{r}\right)$$

Distance between two points p and q in hyperbolic space (the Poincaré ball). It replaces the usual Euclidean distance with hyperbolic distance, where:

- r is the radius of the Poincaré ball.
- $\| -p \oplus q\|$ is the hyperbolic distance between the two points in the Poincaré ball model. ($\oplus$ is the Möbius addition, a special vector addition).
- $\tanh^{-1}$ (inverse hyperbolic tangent) captures the curved geometry of hyperbolic space.

Methods

Object-centric scene hierarchy

- visually similar **objects** (classes of objects) are the **root** nodes
- **scene** images (instances of the object) are the **descendants**

Figure 1. Illustration of the representation space learned by our models. 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.

Methods

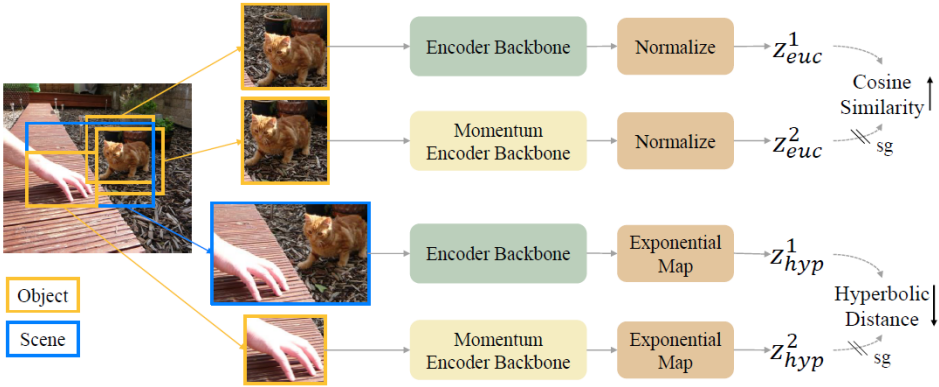


Figure 2. Our **H**yperbolic **C**ontrastive **L**earning (HCL) framework has two branches: given a scene image, two object regions are cropped to learn the object representations with a loss defined in the Euclidean space focusing on the representation directions. A scene region as well as a contained object region are used to learn the scene representations with a loss defined in the hyperbolic space that affects the representation norms.

Methods

Euclidean Loss (to learn object representations)

$$L_{\text{euc}} = -\log\left(\frac{\exp\left(z^1_{\text{euc}} \cdot z^2_{\text{euc}}/\tau\right)}{\exp\left(z^1_{\text{euc}} \cdot z^2_{\text{euc}}/\tau\right) + \sum_n \exp\left(z^1_{\text{euc}} \cdot z^n_{\text{euc}}/\tau\right)}\right)$$

This is the standard contrastive loss function used in contrastive learning frameworks like MoCo, where positive pairs are pulled closer together, and negative pairs are pushed apart in the Euclidean space.

- z_1^{euc} and z_2^{euc} are the embeddings (representations) of two views of the same object in Euclidean space. These are the "positive pairs."
- z_n^{euc} refers to the embeddings of negative samples, which are representations of different objects.
- dot product between the two vectors measures how similar they are in Euclidean space.
- τ is a temperature parameter that controls the scaling of the similarity scores. Lower τ sharpens the distribution, focusing more on confident samples.

Methods

Hyperbolic Contrastive Loss (to learn scene representations)

$$L_{\text{hyp}} = -\log\left(\frac{\exp\left(-\frac{d_{\mathbb{D}}(z_1,z_2)}{\tau}\right)}{\exp\left(-\frac{d_{\mathbb{D}}(z_1,z_2)}{\tau}\right) + \sum_n \exp\left(-\frac{d_{\mathbb{D}}(z_1,z_n)}{\tau}\right)}\right)$$

This contrastive loss encourages positive pairs (e.g., scene-object pairs) to stay close and negative pairs (non-related objects or scenes) to stay apart:

- $d_{\mathbb{D}}(z_1, z_2)$ is the hyperbolic distance between two embeddings z_1 and z_2.
- τ is a temperature parameter that controls the scaling of the distances.
- The summation term includes negative samples that the model needs to push away from positive pairs.

Method

Overall Loss Function

$$L = L_{\text{euc}} + \lambda L_{\text{hyp}}$$

The total loss combines the Euclidean contrastive loss for object representations and the hyperbolic contrastive loss for scene representations.

- L_{euc}: The standard contrastive loss applied to objects in Euclidean space.
- L_{hyp}: The hyperbolic contrastive loss applied to scenes, preserving their relationship with objects.
- λ: A scaling parameter that balances the two loss components.

Method: Additional Detail

Time-consumption

- Calculating hyperbolic loss itself takes nearly the same time as a normal contrastive loss.
- The only overhead in training is one additional forward pass to get scene representations.
- MoCo takes 0.616 sec/iter while HCL takes 0.757 sec/iter under 4 P6000 GPUs.

Experimental Results: AP for Bounding Box and Masks

HCL consistently improves both object detection and semantic segmentation tasks across multiple contrastive learning baselines by pre-training on multi-object datasets COCO and OpenImages. (a general-purpose add-on)

- **Object detection** (colums 1-3) predict bounding box and class label (eg. person, car)
- **Semantic segmentation** (columns 4-6) predict pixel-level segmentation masks
- Metric: Average of AP across multiple Intersection over Union (IoU) thresholds.
- eg. AP50: a detection or segmentation is considered correct if it overlaps with the ground truth by at least 50%.

	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^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**

HCL Model Properties: Label Uncertainty Quantification

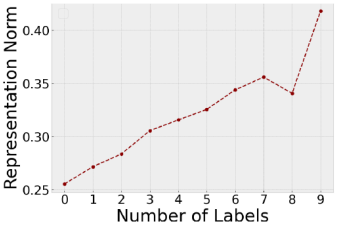


Figure 3. Average representation norms of images with different number of labels in ImageNet-ReaL.

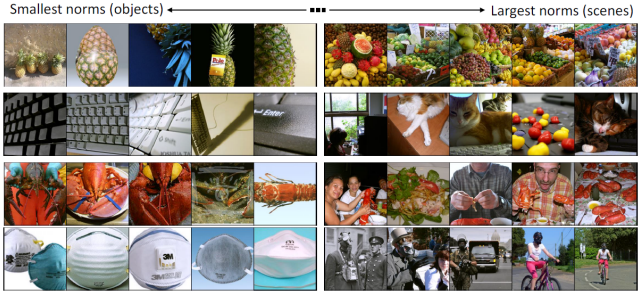


Figure 4. Images from ImageNet training set. The 5 images on the left have the smallest representation norms among all the images from the same class, and the 5 on the right have the largest norms.

- Representation Norm = Length of the embedding vector in hyperbolic space
- Indicates how general or specific a concept is
- Strong correlation between the representation norm and the number of labels per image

HCL Model Properties: Out-of-Context Detection

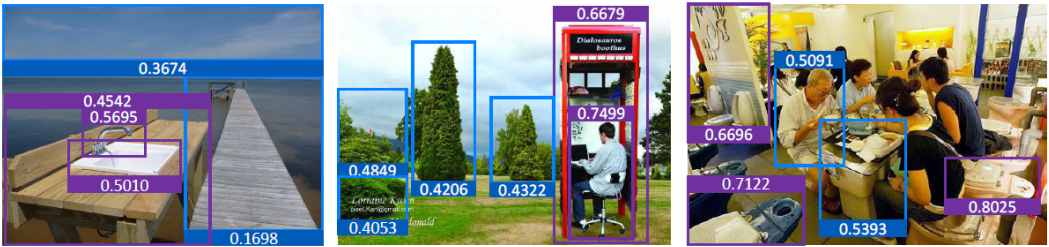


Figure 5. Out-of-context images from the SUN09 dataset. The bounding box of each object and its hyperbolic distance to the scene are shown. Regular objects are in blue and out-of-context objects are in purple. Note that the out-of-context objects tend to have large distances.

- Out-of-context objects generally have a large distance, i.e. smaller similarity, to the overall scene image.

Conclusions

- Proposed a hyperbolic contrastive loss that regularizes scene representations so that they follow an object-centric hierarchy.
- HCL learned representations transfer better than representations learned using vanilla contrastive loss on various downstream tasks, including object detection, semantic segmentation, and linear classification.
- The magnitude of representation norms effectively reflects the scene-objective relationship.

Conclusions

Strength

- Utilizes the hyperbolic space to formulate the scene-object relationship and modeling hierarchy by representation norm is smart and flexible.
- Can be used as a general-purpose add-on to other contrastive learning methods.
- Seems easy to implement by adding a contrastive loss.

Weakness

- AP improvement of 1% (from ORL) is not considered huge in the field.
- Lack of generalization to non-visual data (text or structured data)
- Computational overhead in hyperbolic space when dealing with non-visual data?

"Our goal is not to develop another state-of-the-art self-supervised learning method but a step towards learning representations for images depicting not just objects."

Recommendations

- Results compelling? – Yes, small but consistent improvement over current contrastive learning methods.
- Recommend reading? – Yes, clearly written.
- Recommend implementing? – Yes, a general add-on to contrastive learning methods.
- Implement in my research? – Need to translate to clinical setting. Eg. Contrastive learning across two modalities (structured and unstructured data). EHR data as general root, doctor's notes as detailed and contextual instances; data from the same encounter of a patient are positive samples and data from different patients are negative?)
- Github: https://github.com/shlokk/HCL/tree/main