Evaluating Human-Machine Representational Alignment in Hyperbolic Space

Otto Béla Márton[1], **Christina Sartzetaki**, **Pascal Mettes**, **Iris I. A. Groen**
Informatics Institute, University of Amsterdam, The Netherlands

Abstract

While humans naturally organize concepts hierarchically, this characteristic remains poorly represented in many deep neural networks (DNNs). This may lower generalisability and could cause DNNs to fail in unexpected ways. Virtually all DNNs make use of Euclidean geometry, but hyperbolic geometry is more naturally suitable for hierarchical structures. Using the THINGS dataset of human similarity judgments of object triplets, we examine the alignment between humans and Euclidean versus hyperbolic models, including both a hyperbolic version of a task-optimized DNN (CLIP) and a hyperbolic adaptation of sparse positive embeddings trained directly on the human behavioural data. Confirming the suitability of hyperbolic geometry, we find that the hyperbolic models predict human behavioural similarity judgments significantly better than their Euclidean counterparts.

Keywords: representational alignment; hyperbolic geometry; object similarity judgments; CLIP

Introduction

DNNs excel not only in computer vision, but are also increasingly being used as models for human visual processing (Kriegeskorte, 2015), with recent works showing particularly good alignment for a model trained with language supervision, Contrastive Language Image Pretraining (CLIP) with the human brain (Wang, Kay, Naselaris, Tarr, & Wehbe, 2023) and behavior (Bielawski, Devillers, Van De Cruys, & VanRullen, 2022). However, unlike humans, models (including CLIP) often fail to learn hierarchical conceptual structures and to generalize robustly (Muttenthaler et al., 2024). This may be because DNNs use Euclidean embeddings by default, where hierarchical relationships are inherently distorted, unlike in hyperbolic space (Nickel & Kiela, 2017). Here, we investigate whether representational alignment with humans is enhanced when using hyperbolic rather than Euclidean geometry.

Hyperbolic geometry is a non-Euclidean geometry in which space is negatively curved, making the space grow exponentially with distance from the origin, suitable for embedding hierarchical data. Another benefit of hyperbolic space is the concept of entailment cones, where the order in hierarchies (e.g. 'mammal' contains 'zebra') can be represented by spatial inclusion, with regions corresponding to broader concepts geometrically containing those of their descendants (Ganea, Bécigneul, & Hofmann, 2018). These findings have resulted in the development of hyperbolic versions of multimodal models (Mettes, Ghadimi Atigh, Keller-Ressel, Gu, & Yeung, 2024).

CLIP contains both an image encoder and a text encoder, which are trained using positive samples (matching image caption pairs) and negative samples. The aim of CLIP is to adjust the vision/text encoders such that the difference between the alignment of the positive and negative pairs is maximized (Radford et al., 2021). MERU is a hyperbolic extension of this architecture, with two major differences: 1) In the last layer, the embeddings are mapped from Euclidean space onto hyperbolic space and 2) it adds an entailment loss, enforcing the image embeddings to be in the entailment cone of the corresponding text embedding, since an image is more specific than the description (Desai, Nickel, Rajpurohit, Johnson, & Vedantam, 2023). Another hyperbolic alternative is HyCoCLIP, which also adds an intramodal entailment cone loss, in which more specific image/text segments (e.g. fresh flowers in a vase) must be in the entailment cone of less specific images/text-segments (e.g. fresh flowers) (Pal et al., 2024).

To determine the representational similarity between hyperbolic DNNs and humans, we use the THINGS dataset; a large-scale dataset of images of 1854 different nameable concepts (Hebart et al., 2023), containing 4.70 million human judgments of odd-one-out triplet similarity. Prior work shows these judgments are well captured by a Sparse Positive Object Similarity Embedding (SPoSE) model, which finds a sparse and interpretable set embedding for concept representations (Hebart, Zheng, Pereira, & Baker, 2020). Our contributions are twofold: first, we show that HyCoCLIP, but not MERU, better aligns with human odd-one-out similarity judgments. Second, we introduce the Hyperbolic Positive Object Embedding (HyPoE) architecture, a hyperbolic adaptation of SPoSE.

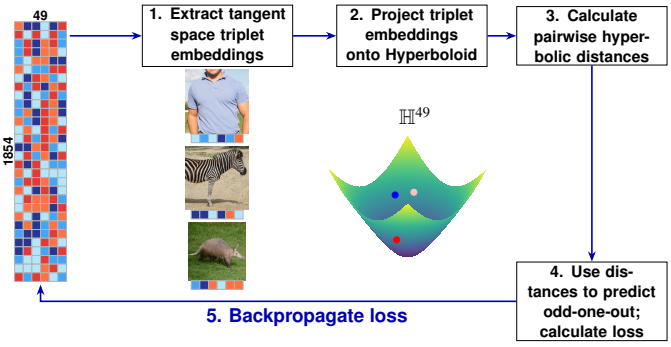

Figure 1: Overview of Hyperbolic Positive object Embedding (HyPoE) architecture[3]. Left: embedding matrix, middle: triplet images and representations; right: hyperboloid model of hyperbolic space ($\mathbb{H}$).

Methods

To test hyperbolic space's usefulness for encoding THINGS odd-one-out data directly, we introduce Hyperbolic Positive object Embeddings (HyPoE), a hyperbolic adaptation of

[1] mail@ottomarton.com

SPoSE. HyPoE learns hyperbolic object embeddings by starting with a random embedding matrix, extracting the triplet's embeddings, projecting them onto hyperbolic space, calculating the triplet's pairwise hyperbolic distances before ultimately predicting the odd-one-out. The loss function combines the prediction error, a parameterized L1-loss, and a positivity penalty:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{L1} + \gamma \mathcal{L}_{Pos} \tag{1}$$

With, $\mathcal{L}_{CE}$ as the cross-entropy loss based on the softmax prediction probabilities, $\mathcal{L}_{L1}$ penalizing the L1-norm of the embedding matrix to encourage sparsity, $\mathcal{L}_{Pos}$ penalizing negative embedding weights and λ and γ as tunable parameters. This loss is then backpropagated to update the embedding matrix (**Fig. 1**).

To determine the alignment between neural network representations and human conceptual organisation, we test whether the odd-one-out choice can be inferred from the network. For a triplet of images, let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ denote their respective embeddings in the model's last layer. As a distance function we used hyperbolic distance for hyperbolic embeddings and a negative dotproduct for Euclidean embeddings, as these distance metrics performed best in our tests.

$$d_L(\mathbf{v}_i, \mathbf{v}_j) = \cosh^{-1}\left(v_{i0}v_{j0} - \sum_{k=1}^{d} v_{ik}v_{jk}\right) \tag{2}$$

To determine the odd-one-out, we first compute the pairwise distances:

$$d_{12} = d(\mathbf{v}_1, \mathbf{v}_2), \quad d_{13} = d(\mathbf{v}_1, \mathbf{v}_3), \quad d_{23} = d(\mathbf{v}_2, \mathbf{v}_3). \tag{3}$$

We then sum the distances for each embedding:

$$s_1 = d_{12} + d_{13}, \quad s_2 = d_{12} + d_{23}, \quad s_3 = d_{13} + d_{23}. \tag{4}$$

The odd-one-out is then selected as the embedding with the biggest dissimilarity:

$$o = \arg\max_{i \in \{1,2,3\}} s_i. \tag{5}$$

The same distance metric has also be used to calculate the (dis)similarity between all pairs of n concepts, resulting in an n x n Representational Dissimilarity Matrix (RDM). We additionally extract embeddings from regular CLIP and MERU to create RDMs, and compare these to the human object similarities by computing the Spearman rank's correlation of the upper-right triangle of the RDMs.

Results

Our odd-one-out results (**Table 1**) show that HyCoCLIP outperforms CLIP and MERU, showing that its representations better align with human similarity judgments. Additionally, we find that hyperbolic SPoSE outperforms Euclidean SPoSE, further supporting the benefit of using hyperbolic embeddings for human similarity judgments.

HyCoCLIP has the highest correlation with SPoSE and HyPoE, followed by CLIP and finally by MERU (**Fig. 2**). The difference in correlation between SPoSE and HyPoE is bigger for the hyperbolic models (MERU, HyCoCLIP) than for CLIP. These results suggest that hyperbolic models capture different representational structures than Euclidean models.

The results for HyPoE are obtained from a run with a much lower value for lambda (λ = 0.00056 vs 0.008), resulting in denser representations than the Euclidean counterpart. However, upon manual inspection of the HyPoE dimensions the learned representations do still seem interpretable.

Model	**Accuracy**		
	Train	Validation	Test
CLIP ViT-B	51.70%	51.70%	54.37%
MERU ViT-B	51.50%	51.52%	54.60%
HyCoCLIP ViT-B	**52.25%**	**52.25%**	54.51%
SPoSE 49	63.98%	63.98%	64.74%
HyPoE 49	**64.78%**	64.04%	**65.77%**

Table 1: Odd-one-out accuracy for pretrained and trained embeddings on the train (4.1M triplets), validation (0.45M triplets) and a held-out test set (15640 triplets). Note that all sets are held out for the pretrained embeddings (CLIP, MERU, HyCoCLIP). Best results for each model type (trained vs pretrained) are bold if they are significantly better (paired t-test $p < 0.05$). For reference, the noise ceiling is 67.3%.

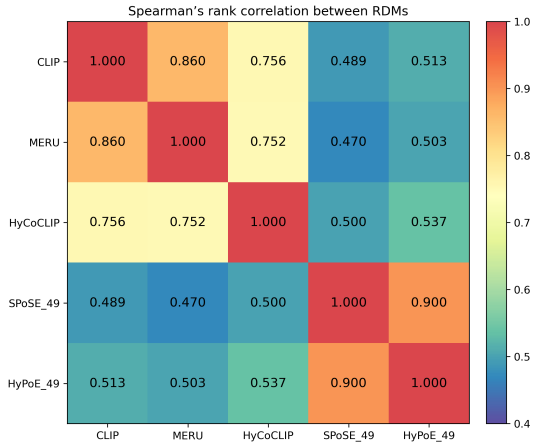

Figure 2: Representational similarity between embedding spaces. Each cell shows the Spearman rank correlation between RDMs derived from the models' embeddings.

Conclusion

Our results show higher performance for HyCoCLIP on the THINGS odd-one-out task, and higher representational similarity between HyCoCLIP and the human-behaviour-derived SPoSE and HyPoE embeddings. Together these results suggest that adding the inductive bias of hyperbolic geometry, which naturally accommodates hierarchies, may help align DNN and human concept representations.

References

Bielawski, R., Devillers, B., Van De Cruys, T., & VanRullen, R. (2022). When does clip generalize better than unimodal models? when judging human-centric concepts. In *7th workshop on representation learning (repl4nlp 2022)* (pp. 29–38).

Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., & Vedantam, S. R. (2023). Hyperbolic image-text representations. In *International conference on machine learning* (pp. 7694–7731).

Ganea, O., Bécigneul, G., & Hofmann, T. (2018). Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning* (pp. 1646–1655).

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., ... Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, *12*, e82580.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*(1), 417–446.

Mettes, P., Ghadimi Atigh, M., Keller-Ressel, M., Gu, J., & Yeung, S. (2024). Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, *132*(9), 3484–3508.

Muttenthaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M. C., ... Lampinen, A. K. (2024). Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*.

Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, *30*.

Pal, A., van Spengler, M., di Melendugno, G. M. D., Flaborea, A., Galasso, F., & Mettes, P. (2024). Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(12), 1415–1426.