UHCone: Universal Hyperbolic Cone For Implicit Hierarchical Learning

Menglin Yang [1] **Jiahong Liu** [2] **Irwin King** [2] **Rex Ying** [1]

Abstract

Hierarchical structures play a vital role in numerous fields, from linguistics, biology, and network science to computer vision, as they represent asymmetric dependencies that are crucial for acquiring high-quality representations and inductive bias. The hyperbolic entailment cone is an effective geometric approach for preserving these relationships by optimizing child nodes to reside within their parent's hyperbolic entailment cone. However, this method necessitates prior information on superior-subordinate hierarchical relationships, which significantly restricts its generality in most real-world data where such prior is implicit and unknown. To address this limitation, we propose the universal hyperbolic cone (UHCone), an effective algorithm designed to capture implicit hierarchical structures in data, making it suitable for a wide range of real-world scenarios. Our approach utilizes the hyperbolic embedding to infer hierarchical relationships first and then reinforce them with cone constraints. This method eliminates the need for prior information on superior-subordinate hierarchies, enabling broader application scenarios. We evaluated the UHCone algorithm on various applications and consistently observed an improvement over baseline methods and the largest improvement up to 4.71%, demonstrating its effectiveness and versatility in capturing implicit hierarchical relationships.

1. Introduction

Hierarchical relationships, including forms of general-specific, class-subclass, group-member, and whole-part, are common types of asymmetric structure found in various applications (Nickel & Kiela, 2017; 2018; Sha et al., 2016; dos Santos & Gonçalves, 2019; Berman, 2019; Nurek & Michalski, 2020; Khrulkov et al., 2020; Li et al., 2019a; Ma et al., 2023) such as textual entailment, biological taxonomies, organizational structures grouping, and image understanding. Accurately modeling these relationships is essential for obtaining high-quality representations and further improving downstream tasks.

In modeling the hierarchical relationships, previous studies have demonstrated the effectiveness of geometric techniques (Xiong et al., 2023). Their success is primarily attributed to the inherent biases in these geometric methods. For example, region-based geometric methods (Abboud et al., 2020; Boratko et al., 2021; Zhang et al., 2022; Vilnis et al., 2018; Ganea et al., 2018b; Bai et al., 2021; Dhall et al., 2020; Özçep et al., 2020; Suzuki et al., 2019), like cones, boxes and discs, model hierarchical linkage via representing the object as a geometric element in embedding space and impose containment constraints, resulting in inherent biases that reflect human intuitions about hierarchies. However, they assume the prior knowledge of the direction or entailment of the hierarchy is known or given.

Implicit hierarchical structures, characterized by structural asymmetries like popularity, relevance, evolution, influence, adaptation, or distribution, frequently occur in real-world data such as social networks, flight networks, recommendation systems, images, etc. Figure 1 demonstrates the network structures in different domains. The flight network, depicted in the left subfigure, exhibits a hub-and-spoke topology where the central airport connects multiple destinations, indicating its pivotal role. Similarly, in the WordNet of food, depicted in the middle subfigure, the abstract word "food" is associated with multiple specific words, exhibiting a clear entailment relationship. Moreover, the right subfigure shows that some noisy images contain more patterns than clear ones, as observed in various generative models such as the diffusion model (Croitoru et al., 2023), where high-dimensional data is refined from Gaussian noise to clear images. In image understanding, noisy images are more likely to convey higher-level semantics since they can contain patterns from multiple specific images (Khrulkov et al., 2020). Capturing implicit hierarchical structures is crucial for achieving high-quality representations and performance in downstream tasks.

[1]Department of Computer Science, University of Yale [2]Department of Computer Science and Engineering, The Chinese University of Hong Kong. Correspondence to: Menglin Yang <menglin.yang@yale.edu>.

Accepted as an extended abstract for the Geometry-grounded Representation Learning and Generative Modeling Workshop at the 41st International Conference on Machine Learning, ICML 2024, Vienna, Austria.

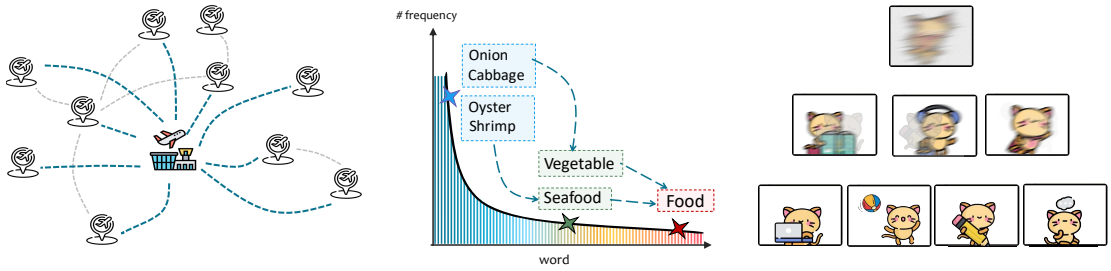

Figure 1: Illustration data structures in different domains. The flight network (left) displays a hub-and-spoke topology with a central airport connecting multiple destinations, indicating its crucial role. The WordNet of food (middle) shows how the abstract concept of "food" is linked to several specific words, exhibiting a clear entailment relationship. The right subfigure demonstrates that noisy images can contain more patterns than clear ones, as observed in various generative models, including the diffusion model, where high-dimensional data is refined from Gaussian noise to clear images.

Challenges However, in the absence of explicitly defined hierarchical relationships, previous methods are inadequate and inapplicable. The direct utilization of these methods in such scenarios often results in suboptimal or even worse performance. Consequently, there is a pressing need for a novel approach that can effectively model hierarchical relationships within the context of general data.

Proposed Work In this work, we present the universal hyperbolic cone (UHCone), an effective algorithm crafted to capture implicit hierarchical structures. Our approach stems from the observation that hyperbolic embedding (Nickel & Kiela, 2017; 2018) intrinsically induce a norm on each node, which is defined as the distance between the node and the origin point. We exploit the fact that nodes higher up in the hierarchy possess smaller hyperbolic norms than their offspring. Based on this property, we obtain underlying hierarchical relationships between pairs of nodes and impose cone constraints on these relationships, which in turn promotes more accurate hierarchical bias.

Contributions In summary, the main contributions of the study are two-fold: *First*, we introduce a novel algorithm, UHCone, which can effectively model implicit hierarchical structures in more general data without explicit hierarchical annotations. *Second*, we evaluated the UHCone on benchmarks, WordNet, and images. The performance consistently improved over baselines, up to 4.71%. Besides, the experimental findings show that directly adding the hyperbolic cone method to non-directional datasets led to a significant drop in performance.

2. Limitation of HCone

In this section, we point out that HCone is unable to model any undirected relationship, as demonstrated in Proposition 2.1[1], which demonstrates that two entities in a pair mutually point to each other in an undirected graph or in non-graph datasets like two images, where there is no clear entailment information that indicates which entity entails the other, as illustrated in Figure 1.

Proposition 2.1. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected acyclic graph, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges such that $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Let $\mathcal{E}_{\text{train}}$ be a symmetric and non-empty subset of $\mathcal{E}$. Specifically, since $\mathcal{E}_{train}$ is symmetric, for any pair $(u, v) \in \mathcal{E}_{train}$, it follows that $(v, u) \in \mathcal{E}_{\text{train}}$. Assume that $\mathcal{L}_{\text{HCone}}$ is a loss function defined in Equation (10). Then, $\mathcal{L}_{\text{HCone}}$ cannot yield non-zero values defined in Equation (10). In other words, HCone is incapable of modeling the symmetric relationship $\{(u, v), (v, u)\}$ in $\mathcal{E}_{train}$.*

3. UHCone: Universal Hyperbolic Entailment Cones

The hyperbolic entailment cone is a useful technique for encoding hierarchical relationships, but it has a fundamental limitation in that it requires prior knowledge of the entailment relationship. Although hierarchical structures are present in many real-world datasets, the relationships between entities are always unknown. To address this limitation, we propose UHCone, a novel method that enhances the generalizability of the hyperbolic entailment cone to a wider range of scenarios. Our approach is based on the idea of leveraging the hierarchical inductive bias produced by hyperbolic embedding to infer implicit hierarchical relationships at first. Then this inferred information guides the hyperbolic cone containment, which in turn boosts the hierarchical inductive bias.

Supposing we have data $\mathcal{X}$ and data split $\mathcal{X}_{\text{train}}$, $\mathcal{X}_{\text{val}}$, $\mathcal{X}_{\text{test}}$. We begin by training a model f_θ for k epochs in hyperbolic space to obtain the embedding $\boldsymbol{x}$. For the following training, we first compute the hyperbolic norm ℓ_x of the data point x

[1]Due to page limit, the HCone method (Equation (10)) is introduced in Appendix B.

for all $x \in \mathcal{X}_{\text{train}}$:

$$\ell_x = d_{\mathbb{H}}(\boldsymbol{x}, \boldsymbol{o}). \tag{1}$$

Then, to stabilize the training, we define the following normalization factor:

$$\tilde{\ell}_x = \frac{\ell_x - \ell_{\min}}{\ell_{\min} - \ell_{\max}}, \tag{2}$$

where $\ell_{\min}$, $\ell_{\max}$ denote the minimal and maximal values of the hyperbolic norm, respectively. We use the normalized norm $\tilde{\ell}_x$ to infer the hierarchical relationship between pairs of data points. The relationship score s_{xy} between points x and y is calculated as:

$$s_{xy} = \text{sign}\left(\frac{\left|\tilde{\ell}_x - \tilde{\ell}_y\right|}{\tilde{\ell}_y - \tilde{\ell}_x + \epsilon} \cdot \max\left(\left|\tilde{\ell}_x - \tilde{\ell}_y\right| - \alpha, 0\right)\right), \tag{3}$$

where $|\cdot|$ represents abs function, ϵ is a tiny positive constant, e.g., 10^{-6}, in our implementation. α is a hyperparameter indicates the difference margin and $s_{xy} \in \{-1, 0, 1\}$. A score of 1 indicates that x is a higher-level node, -1 denotes that y is the higher-level node, and 0 implies that the nodes are close to each other and considered to be at the same hierarchical level. Based on the inferred hierarchical relationships, we reorder each pair of nodes (x, y) as follows:

$$r(x, y) = \begin{cases} (x, y), & s_{xy} > 0 \\ \text{None}, & s_{xy} = 0. \\ (y, x), & s_{xy} < 0 \end{cases} \tag{4}$$

Incorporating this information into our algorithm, we refine the loss function in Equation (10) as

$$\mathcal{L}_{\text{UHCone}} = \sum_{(x,y)\in P} E(r(x, y)) - \sum_{(x',y')\in N} E(r(x', y')). \tag{5}$$

This optimization serves as an auxiliary task and optimizes downstream tasks, aiding the model in acquiring accurate local hierarchical relationships. The optimization objectives enable us to establish a more precise entailment relationship, creating a positive feedback loop that enhances hierarchical inference in the subsequent training. As the proposed approach focuses on the embedding level and does not involve models or downstream tasks, it can be applied in diverse application scenarios. Moreover, the computational requirements are minimal, with the primary complexity resulting from the norm calculation, which has a time complexity of $O(N)$ (assuming we have N data points). With parallel computation, this can be reduced to $O(1)$ if our GPU resources scale linearly with the number of nodes.

Corollary 3.1. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected acyclic graph as outlined in Proposition 2.1, and consider $\mathcal{L}_{\text{UHCone}}$ as the designated loss function with its associated energy function $E(u, v)$. Given any vertex pair $\{u, v\}$ in $\mathcal{E}$ exhibiting an implicit asymmetric relationship, where either u entails v or v entails u, the proposed UHCone approach ensures that if $(u, v) \in \mathcal{E}_{train}$, the energy function will satisfy the criterion: $E(\boldsymbol{u}, \boldsymbol{v}) \neq E(\boldsymbol{v}, \boldsymbol{u})$. This criterion highlights an inherent asymmetry in representing hierarchical relationships between any vertex pair $\{(u, v), (v, u)\}$. Consequently, UHCone possesses the capability to learn and represent implicit entailment relationships.*

4. Experiments

We evaluate the effectiveness of our proposed method in domains with no explicit hierarchies, including undirected WordNet and images.

Experiments on undirected WordNet. We first make a comparison on the undirected WordNet - Mammal and Noun datasets. Note that the undirected WordNet - Mammal is different from the dataset used in (Ganea et al., 2018b) where there are no given entailment relationships for model to learn. For detailed data processing and training processing, please refer to the Appendix E.

For comparison, we select two types of base models, the first one is the the Euclidean and hyperbolic shallow models used in (Chami et al., 2019). Besides, we select a graph base model (He et al., 2020) since the dataset can be viewed as a graph. For the implicit hierarchical learning method, we compare with the HCone method (Ganea et al., 2018b), where we apply this cone function directly without additional operations. The other is the gating HCone (GHCone), where we set a trainable MLP on paired nodes to infer their hierarchies.

The experimental results are presented in Table 1, which shows the comparison of AUC scores on the Mammal and Noun datasets in WordNet for different dimensions (2, 5, and 10). From Table 1, we can observe that the proposed UHCone `consistently` outperforms the Euclidean and hyperbolic baselines across all dataset settings. It has also been discovered that the introduction of HCone and GHCone resulted in a significant reduction in performance. This is because HCone is a directional method, and without any direction information, it can lead to an inaccurate entailment relationship. Additionally, the trainable gating method is not efficient in learning hierarchical information. To facilitate a better understanding of the effectiveness of the proposed method, we presented the entire test performance on the WordNet Mammal dataset in Figure 2. It is clear that adding different cones after 100 epochs results in an immediate difference. This demonstrates the effectiveness of our model in learning implicit hierarchical relationships in the absence of explicit entailment information.

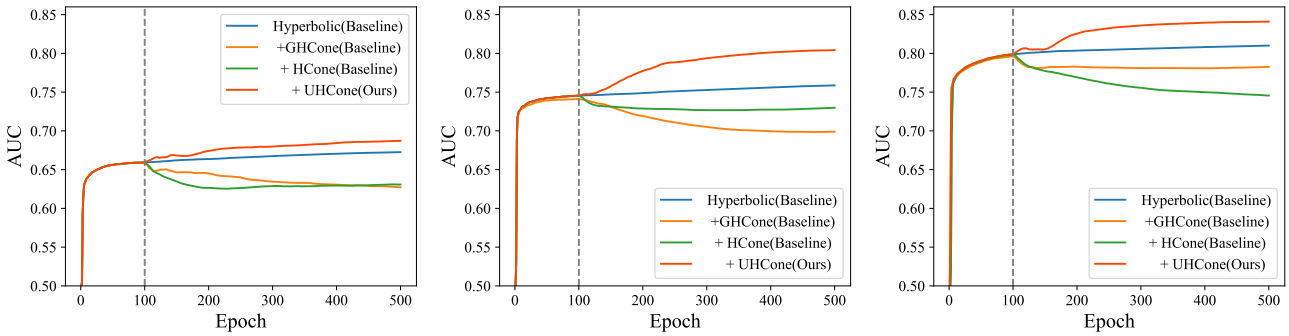

Figure 2: AUC score changes on the undirected WordNet Mammal test dataset along with epochs. The first, second, and third subfigure denotes performance on Mammal with dimensions 2, 5 and 10, respectively. The vertical dash line at the 100th epoch denotes we started adding UHCone, HCone, and Gating methods where we use the hyperbolic shallow model as the base model.

Table 1: Comparison of AUC scores on the Mammal and Noun datasets in WordNet. The highest-performing model is highlighted in bold.

Shallow-based Model						
	Mammal			Noun		
Dimension	2	5	10	2	5	10
Euclidean	56.45	63.76	65.81	54.08	54.82	59.17
Hyperbolic	60.67	72.45	72.71	62.16	68.19	71.98
HCone	58.21	70.07	70.08	58.80	62.64	66.62
GHCone	60.04	70.97	70.02	60.01	64.54	68.11
UHCone	**61.93**	**74.49**	**75.84**	**63.08**	**69.02**	**72.52**
Improvement	+2.08%	+2.82%	+4.30%	+1.48%	+1.22%	+0.75%
Graph-based Model						
Euclidean	62.31	69.29	75.61	62.66	66.81	71.31
Hyperbolic	67.33	75.61	81.54	66.29	71.25	78.12
DHCone	62.62	72.75	78.42	64.91	69.62	78.12
GHCone	63.63	74.13	79.65	64.87	69.82	76.87
UHCone	**68.15**	**79.17**	**84.51**	**67.53**	**73.09**	**80.31**
Improvement	+1.22%	+4.71%	+3.64%	+1.87%	+2.58%	+2.80%

Table 2: Accuracy of different models on few-shot image classification tasks with 1-shot 5-way task, 5-shot 5-way task on Caltech-UCSD Birds (CUB) dataset. The results are reported with 95% confidence intervals. For each task, the best-performing method is highlighted, and the performances of baselines are taken from (Khrulkov et al., 2020). The results in grey are reproduced by us.

Baselines	Embedding Net	1-Shot 5-Way	5-Shot 5-Way
MatchingNet (Vinyals et al., 2016)	4 Conv	61.16 ± 0.89	72.86 ± 0.70
MAML (Finn et al., 2017)	4 Conv	55.92 ± 0.95	72.09 ± 0.76
ProtoNet (Snell et al., 2017)	4 Conv	51.31 ± 0.91	70.77 ± 0.69
MACO (Hilliard et al., 2018)	4 Conv	60.76	74.96
RelationNet (Sung et al., 2018)	4 Conv	62.45 ± 0.98	76.11 ± 0.69
Baseline++ (Chen et al., 2019)	4 Conv	60.53 ± 0.83	79.34 ± 0.61
DN4-DA (Li et al., 2019b)	4 Conv	53.15 ± 0.84	81.90 ± 0.60
Hyperbolic ProtoNet (Khrulkov et al., 2020)	4 Conv	64.02 ± 0.24	82.53 ± 0.14
Hyperbolic ProtoNet*	4 Conv	65.01 ± 0.24	81.94 ± 0.15
UHCone (Hyperbolic ProtoNet)	4 Conv	**65.60** ± 0.24	**82.61** ± 0.14

Experiments on Image dataset. In the image domain, we follow the methods and experimental settings presented in (Khrulkov et al., 2020) and evaluate the proposed models in the few-shot image classification task on the Caltech-UCSD Birds (CUB) dataset (Wah et al., 2011), which involves classifying new data with limited labeled samples. The few-shot learning task is formulated as N-way K-shot, with N representing the number of classes to classify and K the number of available samples per class. The 1-shot 5-way and 5-shot 5-way tasks were considered in our experiments, keeping the same with settings in (Khrulkov et al., 2020). The detailed data processing and training process are given in Appendix E.

We report the average performance and the 95% confidence interval in Table 2, and for baseline results, we take them from (Khrulkov et al., 2020). The results show that the proposed method outperforms the existing baselines on both 1-shot and 5-shot 5-way tasks. In particular, the proposed method achieved an average accuracy of 65.60 ± 0.24 for the 1-shot 5-way task and 82.61 ± 0.14 for the 5-shot 5-way task. These results demonstrate the effectiveness of integrating UHCone into image embedding for few-shot learning.

5. Conclusion

In this work, we proposed UHCone, a novel approach to capture implicit hierarchical structures in data by leveraging the hyperbolic norm of nodes. Our method infers hierarchical relationships from the hyperbolic norm and imposes cone constraints based on the inferred relationships, which in turn promotes more accurate hierarchical bias through a positive feedback loop. UHCone is simple yet effective and can be employed as a plug-in to facilitate learning of hierarchical structures independent of the hyperbolic model or task.

References

Abboud, R., Ceylan, I., Lukasiewicz, T., and Salvatori, T. Boxe: A box embedding model for knowledge base completion. *NeurIPS*, 33:9649–9661, 2020.

Athiwaratkun, B. and Wilson, A. G. Hierarchical density order embeddings. *arXiv preprint arXiv:1804.09843*, 2018.

Bai, Y., Ying, Z., Ren, H., and Leskovec, J. Modeling heterogeneous hierarchies with relation-specific hyperbolic cones. *NeurIPS*, 34:12316–12327, 2021.

Berman, J. J. *Taxonomic guide to infectious diseases: understanding the biologic classes of pathogenic organisms*. Academic Press, 2019.

Boratko, M., Zhang, D., Monath, N., Vilnis, L., Clarkson, K. L., and McCallum, A. Capacity and bias of learned geometric embeddings for directed graphs. *NeurIPS*, 34: 16423–16436, 2021.

Chami, I., Ying, Z., Ré, C., and Leskovec, J. Hyperbolic graph convolutional neural networks. In *NeurIPS*, pp. 4868–4879, 2019.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

Chen, Y., Yang, M., Zhang, Y., Zhao, M., Meng, Z., Hao, J., and King, I. Modeling scale-free graphs for knowledge-aware recommendation. *WSDM*, 2022.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *TPAMI*, 2023.

Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., and Vedantam, S. R. Hyperbolic image-text representations. In *ICML*, pp. 7694–7731. PMLR, 2023.

Dhall, A., Makarova, A., Ganea, O., Pavllo, D., Greeff, M., and Krause, A. Hierarchical image classification using entailment cone embeddings. In *CVPR workshops*, pp. 836–837, 2020.

dos Santos, A. A. and Gonçalves, W. N. Improving pantanal fish species recognition through taxonomic ranks in convolutional neural networks. *Ecological Informatics*, 53:100977, 2019.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pp. 1126–1135. PMLR, 2017.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. In *NeurIPS*, pp. 5345–5355, 2018a.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, pp. 1646–1655. PMLR, 2018b.

Gulcehre, C., Denil, M., Malinowski, M., Razavi, A., Pascanu, R., Hermann, K. M., Battaglia, P., Bapst, V., Raposo, D., Santoro, A., et al. Hyperbolic attention networks. In *ICLR*, 2019.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. LightGCN: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pp. 639–648, 2020.

Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C. D., and Hodas, N. O. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018.

Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., and Lempitsky, V. Hyperbolic image embeddings. In *CVPR*, pp. 6418–6428, 2020.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Li, J., Gao, Y., Bing, L., King, I., and Lyu, M. R. Improving question generation with to the point context. In *EMNLP-IJCNLP*, pp. 3216–3226, 2019a.

Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., and Luo, J. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*, pp. 7260–7268, 2019b.

Liu, J., Yang, M., Zhou, M., Feng, S., and Fournier-Viger, P. Enhancing hyperbolic graph embeddings via contrastive learning. In *NeurIPS 2nd SSL workshop*, 2022.

Liu, Q., Nickel, M., and Kiela, D. Hyperbolic graph neural networks. In *NeurIPS*, pp. 8230–8241, 2019.

Ma, Y., Song, Z., Hu, X., Li, J., Zhang, Y., and King, I. Graph component contrastive learning for concept relatedness estimation. In *AAAI*, pp. 13362–13370. AAAI Press, 2023.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, pp. 6338–6347, 2017.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pp. 3779–3788, 2018.

Nurek, M. and Michalski, R. Combining machine learning and social network analysis to reveal the organizational structures. *Applied Sciences*, 10(5):1699, 2020.

Özçep, Ö. L., Leemhuis, M., and Wolter, D. Cone semantics for logics with negation. In *IJCAI*, pp. 1820–1826, 2020.

Sala, F., De Sa, C., Gu, A., and Re, C. Representation tradeoffs for hyperbolic embeddings. In *ICML*, pp. 4460–4469, 2018.

Sarkar, R. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

Sha, L., Chang, B., Sui, Z., and Li, S. Reading and thinking: Re-read LSTM unit for textual entailment recognition. In *COLING*, pp. 2870–2879, 2016.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *NeurIPS*, 30, 2017.

Sun, J., Cheng, Z., Zuberi, S., Pérez, F., and Volkovs, M. HGCF: Hyperbolic graph convolution networks for collaborative filtering. In *WWW*, pp. 593–601, 2021.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *CVPR*, pp. 1199–1208, 2018.

Suzuki, A., Nitanda, A., Wang, J., Xu, L., Yamanishi, K., and Cavazza, M. Generalization error bound for hyperbolic ordinal embedding. In *ICML*, pp. 10011–10021. PMLR, 2021a.

Suzuki, A., Nitanda, A., Xu, L., Yamanishi, K., Cavazza, M., et al. Generalization bounds for graph embedding using negative sampling: Linear vs hyperbolic. *NeurIPS*, 34:1243–1255, 2021b.

Suzuki, R., Takahama, R., and Onoda, S. Hyperbolic disk embeddings for directed acyclic graphs. In *ICML*, pp. 6066–6075. PMLR, 2019.

Tseng, A., Yu, T., Liu, T. J., and De Sa, C. Coneheads: Hierarchy aware attention. *arXiv preprint arXiv:2306.00392*, 2023.

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Vilnis, L. and McCallum, A. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014.

Vilnis, L., Li, X., Murty, S., and McCallum, A. Probabilistic embedding of knowledge graphs with box lattice measures. *arXiv preprint arXiv:1805.06627*, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *NeurIPS*, 29, 2016.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Xiong, B., Nayyeri, M., Jin, M., He, Y., Cochez, M., Pan, S., and Staab, S. Geometric relational embeddings: A survey. *arXiv preprint arXiv:2304.11949*, 2023.

Yang, M., Zhou, M., Kalander, M., Huang, Z., and King, I. Discrete-time temporal network embedding via implicit hierarchical learning in hyperbolic space. In *KDD*, pp. 1975–1985, 2021.

Yang, M., Li, Z., Zhou, M., Liu, J., and King, I. Hicf: Hyperbolic informative collaborative filtering. In *KDD*, pp. 2212–2221, 2022a.

Yang, M., Zhou, M., Liu, J., Lian, D., and King, I. HRCF: Enhancing collaborative filtering via hyperbolic geometric regularization. In *WWW*, 2022b.

Yang, M., Zhou, M., Xiong, H., and King, I. Hyperbolic temporal network embedding. *TKDE*, 2022c.

Yang, M., Zhou, M., Pan, L., and King, I. κHGCN: Tree-likeness modeling via continuous and discrete curvature learning. In *KDD*, pp. 2965–2977, 2023.

Zhang, D., Boratko, M., Musco, C., and McCallum, A. Modeling transitivity and cyclicity in directed graphs via binary code box embeddings. *NeurIPS*, 35:10587–10599, 2022.

Zhang, R., Khan, A. A., and Grossman, R. L. Evaluation of hyperbolic attention in histopathology images. In *BIBE*, pp. 773–776. IEEE, 2020.

A. Related Work

Geometric Embedding Recent advances in the field of geometric embeddings (Xiong et al., 2023) have been largely focused on the representation of intricate data structures, especially those that involve asymmetric relationships. Geometric embeddings utilize geometric objects with intricate structures to represent data elements, including boxes (Abboud et al., 2020; Boratko et al., 2021; Zhang et al., 2022), entailment cones (Ganea et al., 2018b; Bai et al., 2021; Dhall et al., 2020; Özçep et al., 2020; Tseng et al., 2023), discs (Suzuki et al., 2019), densities (Athiwaratkun & Wilson, 2018; Vilnis & McCallum, 2014), and elements of hyperbolic geometry (Nickel & Kiela, 2017; 2018; Ganea et al., 2018a), etc. By leveraging the rich geometric structure of embedding objects, these methods can provide more expressive and effective representations of data elements in various contexts, including but not limited to graph representation learning, recommendation systems, and natural language processing.

Hyperbolic Embedding Hyperbolic embedding is one of the geometric embeddings, which embeds objects in a continuous, low-dimensional hyperbolic space, showing impressive performance (Nickel & Kiela, 2017; 2018; Chami et al., 2019; Liu et al., 2019; Yang et al., 2023), less distortion (Sarkar, 2011; Sala et al., 2018) and smaller generalization error (Suzuki et al., 2021a;b) in hierarchical and scale-free structured data. Hyperbolic space can be regarded as a continuous tree structure (Krioukov et al., 2010; Sarkar, 2011), implicitly capturing the hierarchical relationships. In recent years, the field of hyperbolic learning has seen a growing interest in various areas, particularly in the areas of lexical entailment (Nickel & Kiela, 2017; Gulcehre et al., 2019; Sala et al., 2018), image embedding (Khrulkov et al., 2020; Zhang et al., 2020; Desai et al., 2023), graph embedding (Gulcehre et al., 2019; Chami et al., 2019; Liu et al., 2019; Yang et al., 2021; 2022c; Liu et al., 2022) and recommender systems (Sun et al., 2021; Yang et al., 2022b;a; Chen et al., 2022).

Cone Embedding Cone embedding builds upon the idea of order embedding (Vendrov et al., 2015), which represents a partially ordered set. Ganea et al. (Ganea et al., 2018b) extend this idea to hyperbolic space, where nodes are modeled as cones. By leveraging the increased expressive power of hyperbolic space for tree-like graphs and the asymmetry and inductive bias of region-based representations, this approach offers superior performance over traditional methods. To capture multiple heterogeneous hierarchies, Bai et al. (Bai et al., 2021) introduced ConeE, which employs cone containment constraints in different subspaces of the hyperbolic embedding space. Tseng et al. (Tseng et al., 2023) proposed a new attention in the Transformer which is defined by hyperbolic entailment cones. Cone embedding and its extensions represent a promising avenue for hierarchical learning (Desai et al., 2023). However, current methods are designed specifically for directed graphs or explicitly hierarchical relationships and lack generality. In this work, we address this limitation, enabling its applicability to a wider range of real-world scenarios.

B. Preliminaries

Hyperbolic Embedding Hyperbolic geometry is a non-Euclidean geometry characterized by a constant negative curvature, whereby the curvature describes the degree of deviation of a geometric manifold from Euclidean space. The hyperbolic space comprises multiple different models, and these models are mutually isometric. In our work, we utilize the Poincaré ball model to develop our approach. However, it is noteworthy that our methodology is not bound by any specific model and can be readily applied to other models. An n-dimensional Poincaré ball model centered at origin with negative curvature $\kappa(\kappa < 0)$ is defined as $\mathbb{H}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < -1/\kappa\}$, where $\|\cdot\|$ is L_2 norm. For any point pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{H}^n, \boldsymbol{x} \neq \boldsymbol{y}$, the distance on this manifold is defined as:

$$d_{\mathbb{H}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{2}{\sqrt{|\kappa|}} \tanh^{-1}\left(\sqrt{|\kappa|}\,\|-\boldsymbol{x} \oplus_\kappa \boldsymbol{y}\|\right), \tag{6}$$

where $\oplus_\kappa$ is Möbiuous addition, and it is defined as:

$$\boldsymbol{x} \oplus_\kappa \boldsymbol{y} = \frac{\left(1 - 2\kappa\langle \boldsymbol{x}, \boldsymbol{y}\rangle_2 - \kappa\|\boldsymbol{y}\|_2^2\right)\boldsymbol{x} + \left(1 + \kappa\|\boldsymbol{x}\|_2^2\right)\boldsymbol{y}}{1 - 2\kappa\langle \boldsymbol{x}, \boldsymbol{y}\rangle_2 + \kappa^2\|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2}. \tag{7}$$

In particular, for each point $\boldsymbol{x} \in \mathbb{H}^n$, its distance to origin $\boldsymbol{o}$, $d_{\mathbb{H}}(\boldsymbol{x}, \boldsymbol{o}) = 2\tanh^{-1}(\|\boldsymbol{x}\|)$ is the induced hyperbolic norm. For each point $\boldsymbol{x} \in \mathbb{H}^n$, the tangent space $\mathcal{T}_{\boldsymbol{x}}\mathbb{H}$ provides a local linear approximation of $\mathbb{H}^n$ at $\boldsymbol{x}$. Besides, the exponential map $\exp_{\boldsymbol{x}} : \mathcal{T}_{\boldsymbol{x}}\mathbb{H}^n \to \mathbb{H}^n$ help project the embedding from tangent space to hyperbolic space and logarithmic map $\log_{\boldsymbol{x}} : \mathbb{H}^n \to \mathcal{T}_{\boldsymbol{x}}\mathbb{H}$ do the inverse. The tangent space enables direct vector operations, like addition and scalar multiplication. Conversely, the logarithmic map $\log_{\boldsymbol{x}} : \mathbb{H}^n \to \mathcal{T}_{\boldsymbol{x}}\mathbb{H}$ maps vectors in $\mathbb{H}^n$ to vectors in $\mathcal{T}_{\boldsymbol{x}}\mathbb{H}^n$.

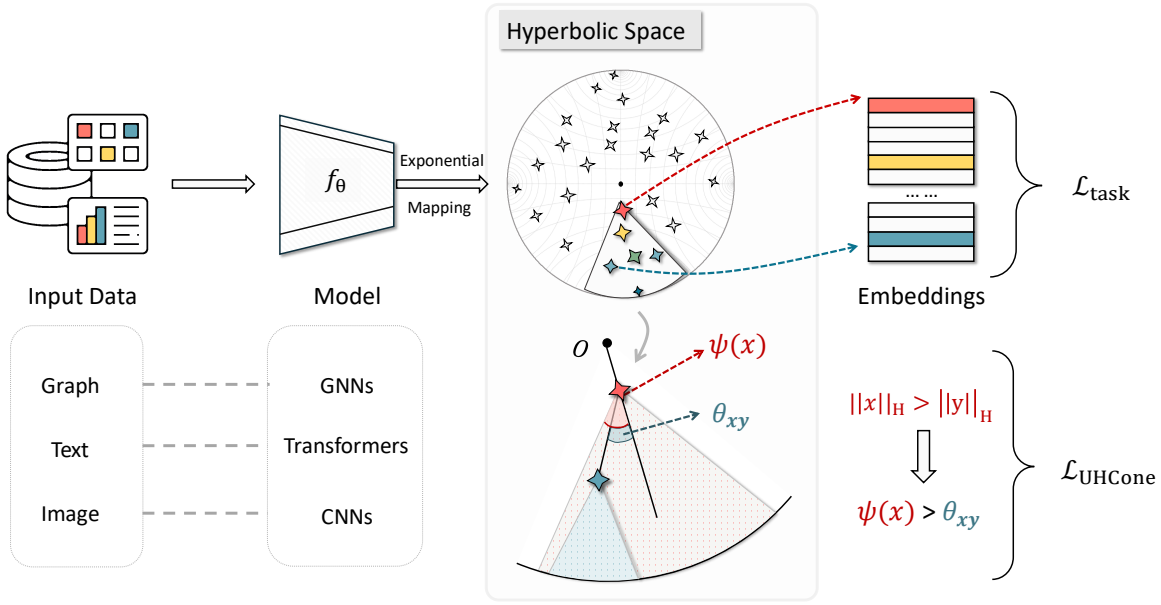

Figure 3: Overview of the UHCone framework. Input data in various formats (i.e., graphs, text, or images) is processed by suitable neural network models, such as GNNs, transformers, or CNNs, to obtain an effective object representation. The UHCone approach extracts the hierarchical information from the embedding norm and applies the cone energy function.

Hyperbolic space exhibits a fundamental characteristic that distinguishes it from Euclidean space: the volume of space expands exponentially rather than polynomially. This remarkable property endows hyperbolic space with a natural geometric prior for representing hierarchical structures and power-law distributed data (Krioukov et al., 2010). Specifically, this prior is leveraged by embedding high-level nodes with small norms, resulting in their proximity to the origin, while low-level nodes are optimized with large norms and positioned relatively further away from the origin.

Hyperbolic Entailment Cone (HCone) By leveraging the hyperbolic space, we can represent nodes as cones (Ganea et al., 2018b), which enables us to capture partial ordering induced by hierarchical relations. This approach combines the advantages of hyperbolic space in representing tree-like structures with the asymmetry and inductive bias of region-based cone representations.

The cone at the apex $\boldsymbol{x}$ is denoted by $\mathfrak{S}_{\boldsymbol{x}}^{\psi(\boldsymbol{x})}$ where $\psi(\boldsymbol{x})$ specific the half-aperture of cone at $\boldsymbol{x}$. The $\psi(\boldsymbol{x})$ remain unchanged irrespective of the angular coordinate of its apex $\boldsymbol{x}$, and are solely determined by the norm of $\boldsymbol{x}$, i.e., $\psi(\boldsymbol{x}) = \psi(\boldsymbol{y})\,(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}^n \backslash \{\mathbf{0}\}$, s.t. $\|\boldsymbol{x}\| = \|\boldsymbol{y}\|)$. The objective is to model partial order using the containment relationship between cones. Specifically, the entailment cones adhere to transitivity, which can be expressed as: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}^d \backslash \{\mathbf{0}\} : \boldsymbol{y} \in \mathfrak{S}_{\boldsymbol{x}}^{\psi(\boldsymbol{x})} \Rightarrow \mathfrak{S}_{\boldsymbol{y}}^{\psi(\boldsymbol{y})} \subseteq \mathfrak{S}_{\boldsymbol{x}}^{\psi(\boldsymbol{x})}$. Given a dataset $\mathcal{X}$ and an hierarchical entailment pairs $(x, y) \in \mathcal{X}$, supposing x entails y or y is a subclass of x, to encourage y being in the cone of x in the embedding space, we define the angle $\theta_{\boldsymbol{xy}}$ between the half-lines $\overrightarrow{\boldsymbol{xy}}$ and $\overrightarrow{\boldsymbol{ox}}$, that is:

$$\theta_{\boldsymbol{xy}} = \cos^{-1}\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle \left(1 + \|\boldsymbol{x}\|^2\right) - \|\boldsymbol{x}\|^2 \left(1 + \|\boldsymbol{y}\|^2\right)}{\|\boldsymbol{x}\|\|\boldsymbol{x} - \boldsymbol{y}\|\sqrt{1 + \|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 - 2\langle \boldsymbol{x}\boldsymbol{y}\rangle}}\right), \tag{8}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the hyperbolic embeddings of point $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. Then, to satisfy the transitivity of nested angular cones and symmetric conditions, we have the following expression of hyperbolic entailment cone at apex $\boldsymbol{x} \in \mathbb{H}^n$:

$$\mathfrak{S}_{\boldsymbol{x}}^{\psi(\boldsymbol{x})} = \{\boldsymbol{y} \in \mathbb{H}^n \mid \theta_{\boldsymbol{xy}} \leq \psi(\boldsymbol{x})\}. \tag{9}$$

To achieve the above goal, the model can be trained with max-margin loss function (Ganea et al., 2018b; Bai et al., 2021):

$$\mathcal{L}_{\text{HCone}} = \sum_{(\boldsymbol{x},\boldsymbol{y}) \in P} E(\boldsymbol{x}, \boldsymbol{y}) + \sum_{(\boldsymbol{x}',\boldsymbol{y}') \in N} \max\left(0, \gamma - E\left(\boldsymbol{x}', \boldsymbol{y}'\right)\right), \tag{10}$$

where P and N denote sets of positive and negative edge samples, respectively. The function $E(\boldsymbol{x}, \boldsymbol{y})$ is given by: $E(\boldsymbol{x}, \boldsymbol{y}) = \max(0, \theta_{\boldsymbol{xy}} - \psi(\boldsymbol{x}))$, which measures the penalty of a wrongly classified pair $(\boldsymbol{x}, \boldsymbol{y})$.

C. Proof of Proposition 2.1

Proof. We aim to show that the HCone model, using the $\mathcal{L}_{\text{HCone}}$ loss function, cannot simultaneously optimize a symmetric relationship. To demonstrate this, we consider the optimization of positive samples, meaning we need to simultaneously minimize $E(\boldsymbol{u}, \boldsymbol{v})$ and $E(\boldsymbol{v}, \boldsymbol{u})$.

Given the definitions of $E(\boldsymbol{u}, \boldsymbol{v})$ and $E(\boldsymbol{v}, \boldsymbol{u})$, this optimization problem is equivalent to minimizing both $\theta_{\boldsymbol{uv}} - \psi(\boldsymbol{u})$ and $\theta_{\boldsymbol{vu}} - \psi(\boldsymbol{v})$. In order to achieve this minimization, we must satisfy the following conditions:

- $\theta_{\boldsymbol{uv}} - \psi(\boldsymbol{u}) \leq 0$, which implies $\|\boldsymbol{u}\| \leq \|\boldsymbol{v}\|$
- $\theta_{\boldsymbol{vu}} - \psi(\boldsymbol{v}) \leq 0$, which implies $\|\boldsymbol{v}\| \leq \|\boldsymbol{u}\|$.

Combining conditions (1) and (2), we obtain $\|\boldsymbol{u}\| = \|\boldsymbol{v}\|$. This means that, under the minimization constraint, in the optimal state, vertices u and v coincide at the same point. However, this result contradicts the assumption that u and v are distinct vertices with a symmetric relationship in $\mathcal{E}_{train}$. Thus, we conclude that the HCone model is unable to optimize a symmetric relationship simultaneously. □

D. Proof of Corollary

Proof. Given the implicit hierarchical relationship between the vertex pair u, v, where either u entails v or v entails u, there arises a norm disparity such that either $\|\boldsymbol{u}\| > \|\boldsymbol{v}\|$ or $\|\boldsymbol{v}\| > \|\boldsymbol{u}\|$. Referring to Equation (4), this disparity introduces an asymmetric bias. Consequently, when applied to Equation (5), this bias ensures distinct energy scores, leading to the conclusion that $E(\boldsymbol{u}, \boldsymbol{v}) \neq E(\boldsymbol{v}, \boldsymbol{u})$. □

E. Experimental Settings

In this work, we use different datasets for evaluations. In the following, we give more details about dataset description and data split.

In WordNet Embedding, we use the subset mammal and noun for evaluation. WordNet is a lexical database for the English language that organizes words into sets of synonyms called sunsets. The mammal subset in WordNet is a collection of synsets that encompasses words related to mammals, which are a class of warm-blooded vertebrates typically characterized by the presence of hair or fur, a four-chambered heart, and the production of milk to nourish their young. The noun hierarchy in WordNet is organized around the concept of hypernymy, which refers to a type-of or is-a relationship between two synsets. Based on previous research (Ganea et al., 2018b), we eliminate the root of the tree as it doesn't provide significant information and only has trivial edges to predict. The remaining set is divided into validation (5%), test (5%), and train (90%). To remove existing entailment relationships in the dataset, we reverse all edges and add them to the training set, ensuring it is symmetric. To enhance the validation and test parts, we include additional negative pairs. For every true edge (u, v), we randomly select five negative corrupted pairs: five pairs (u′, v) and five pairs (u, v′) that are not connected in the complete transitive closure. These negative pairs are then added to the respective negative set. We also use k times negative pairs and $k \in \{1, 5, 10\}$. For link prediction in WordNet, we adopt a Fermi-Dirac decoder, aligning with the methodology presented in (Ganea et al., 2018b).

To obtain the embedding for each word in WordNet, we also used a graph-based encoder with residual connection, which is similar to the encoder (Sun et al., 2021). We searched for the cone weight in the scope of $\{0.01, 0.1, 0.2, 0.5\}$ and set the level margin and score margin to 0.1 and 0.5, respectively. We also found that their values didn't have a significant impact on the results. For baselines and our proposed methods, we used the same training and evaluation protocol. In particular, HCone, GHCone used the same training dataset as UHCone, i.e., applying the bidirectional edge for training. In the first phase of hyperbolic embedding, the training objective is

$$L_{\text{lp}} = \frac{1}{|E|} \sum_{(i,j) \in E} -\log p(\mathbf{z}_i, \mathbf{z}_j) + \frac{1}{|E|} \sum_{(i,j) \notin E'} \log p(\mathbf{z}_i, \mathbf{z}'_j), \tag{11}$$

where E is the edge set and $p(\cdot)$ is the Fermi-Dirac function, indicating the probability of two hyperbolic nodes $\mathbf{u}, \mathbf{v}$ have a link or not, which is defined as:

$$p(\mathbf{u}, \mathbf{v}) = \left[\exp\left(d_{\mathcal{H}}(\mathbf{u}, \mathbf{v})^2 - r\right)/t + 1\right]^{-1}, \tag{12}$$

The loss function is to maximize the probability of two nodes if they are linked in the training set while minimizing the probability of two nodes if they are not linked in the training set.

In the domain of image embedding, our evaluation is applied on the Caltech-UCSD Birds (CUB) dataset. This dataset, specifically curated for fine-grained classification tasks, comprises 11,788 images distributed across 200 distinct bird species. In alignment with the methodology adopted in prior studies (Khrulkov et al., 2020), we partition the dataset such that half of the classes (100 out of 200) are designated for training. The remaining classes are evenly divided between validation and testing, each receiving 50 classes. Considering the relative simplicity of the dataset, we employ a 4-Convolutional (4-Conv) backbone, as previously utilized in the work of Khrulkov et al. (Khrulkov et al., 2020). This choice of architecture is deemed sufficient for the task at hand without introducing unnecessary complexity.

Unlike graphs, there are no links between images. To apply UHCone in image Embedding, within each training batch, we randomly select $2k$ samples from the same class as k positive pairs and $2k$ samples from different classes as k negative pairs, resulting in a total of $2k$ pairs per batch. This integration enables us to insert UHCone into various image embedding tasks. We show that the proposed method can also improve the performance of hyperbolic ProtoNet (Khrulkov et al., 2020) for few-shot learning. Different from the standard ProtoNet (Snell et al., 2017), which computes the prototype of each class in Euclidean space, hyperbolic ProtoNet (Khrulkov et al., 2020) computes the class prototype in hyperbolic space using hyperbolic midpoint. We followed previous work (Khrulkov et al., 2020) for the experimental settings on image embedding and searched for the cone weight in the scope of $\{0.01, 0.1, 0.2, 0.5\}$. We set the level margin and score margin to 0.1 and 0.5, respectively.