Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry

Maximilian Nickel [1] **Douwe Kiela** [1]

Abstract

We are concerned with the discovery of hierarchical relationships from large-scale unstructured similarity scores. For this purpose, we study different models of hyperbolic space and find that learning embeddings in the Lorentz model is substantially more efficient than in the Poincaré-ball model. We show that the proposed approach allows us to learn high-quality embeddings of large taxonomies which can yield significant improvements over Poincaré embeddings, especially in low dimensions. Lastly, we apply our model to discover hierarchies in two real-world datasets: we show that an embedding in hyperbolic space can reveal important aspects of a company's organizational structure as well as historical relationships between language families.

1. Introduction

Hierarchical structures are ubiquitous in knowledge representation and reasoning. For example, starting with Linnaeus, taxonomies have long been used in biology to categorize and understand the relationships between species (Mayr, 1968). In social science, hierarchies are used to understand interactions in humans and animals or to analyze organizational structures such as companies and governments (Dodds et al., 2003). In comparative linguists, evolutionary trees are used to describe the origin of languages (Campbell, 2013), while ontologies are used to provide rich categorizations of entities in semantic networks (Antoniou & Van Harmelen, 2004). Hierarchies are also known to provide important information for learning and classification (Silla & Freitas, 2011). In cognitive development, the results of Inhelder & Piaget (1964) suggest that the classification structure in children's thinking is hierarchical in nature.

Hierarchies can therefore provide important insights into systems of concepts. However, explicit information about such hierarchical relationships is unavailable for many domains. In this paper, we therefore consider the problem of discovering concept hierarchies from unstructured observations, specifically in the following setting:

1. We focus on discovering pairwise hierarchical relations between concepts, where all superior and subordinate concepts are observed.
2. We aim to infer concept hierarchies only from pairwise similarity measurements, which are relatively easy and cheap to obtain in many domains.

Examples of hierarchy discovery that adhere to this setting include the creation of taxonomies from similarity judgments (e.g., genetic similarity of species or cognate similarity of languages) and the recovery of organizational hierarchies and dominance relations from social interactions.

To infer hierarchies from similarity judgments, we propose to model such relationships as a combination of two separate aspects: relatedness and generality. Concept A is a parent (a superior) to concept B if both concepts are related and A is more general than B. By separating these aspects, we can then discover concept hierarchies via hyperbolic embeddings. In particular, we build upon ideas of Poincaré embeddings (Nickel & Kiela, 2017) to learn continuous representations of hierarchies. Due to its geometric properties, hyperbolic space can be thought of as continuous analogue to discrete trees. By embeddings concepts in such a way that their similarity order is preserved, we can then identify (soft) hierarchical relationships from the embedding: relatedness is captured via the distance in the embedding space, while generality is captured via the norm of the embeddings.

To learn high-quality embeddings, we propose a new optimization approach based on the Lorentz model of hyperbolic space. The Lorentz model allows for an efficient, closed-form computation of the geodesics in hyperbolic spaces. This facilitates the development of an efficient optimizer that directly follows these geodesics, rather than doing a first-order approximation as in (Nickel & Kiela, 2017). It allows us also to avoid numerical instabilities that arise from the Poincaré distance. As we will show experimentally,

[1]Facebook AI Research, New York, NY, USA. Correspondence to: Maximilian Nickel <maxn@fb.com>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

this optimization method leads to a significantly improved embedding quality, especially in low dimensions. Simultaneously, we retain the attractive properties of hyperbolic embeddings, i.e., learning continuous representations of hierarchies via gradient-based optimization while scaling to large datasets.

The reminder of this paper is organized as follows. In Section 2, we discuss related work regarding hyperbolic and ordered embeddings. In Section 2, we introduce our model and algorithm to compute the embeddings. In Section 4 we evaluate the efficiency of our approach on large taxonomies. Furthermore, we evaluate the ability of our model to discover meaningful hierarchies on real-world datasets.

2. Related Work

Hyperbolic geometry has recently received attention in machine learning and network science due to its attractive properties for modeling data with latent hierarchies. Krioukov et al. (2010) showed that typical properties of complex networks (e.g., heterogeneous degree distributions and strong clustering) can be explained by assuming an underlying hyperbolic geometry and, moreover, developed a framework to model networks based on these properties. Furthermore, Kleinberg (2007) and Boguñá et al. (2010) proposed hyperbolic embeddings for greedy shortest-path routing in communication networks. Asta & Shalizi (2015) used hyperbolic embeddings of graph to compare the global structure of networks. Sun et al. (2015) proposed to learn representations of non-metric data in pseudo-Riemannian space-time, which is closely related to hyperbolic space.

Most similar to our work are the recently proposed Poincaré embeddings (Nickel & Kiela, 2017), which learn hierarchical representations of symbolic data by embedding them into an n-dimensional Poincaré ball. The main focus of that work was to model the link structure of symbolic data efficiently, i.e., to find low-dimensional embeddings via exploiting the hierarchical structure of hyperbolic space. Here, we build upon this idea and extend it in various ways. First, we propose a new model to compute hyperbolic embeddings in the Lorentz model of hyperbolic geometry. This allows us to develop an efficient Riemannian optimization method that scales well to large datasets and provides significantly better embeddings, especially in low dimensions. Second, we consider learning hierarchies from real-valued similarity scores, which generalize binary adjacency matrices as considered by Nickel & Kiela (2017). Third, we in addition to preserving similarity (e.g., local link structure), we also focus on recovering the correct hierarchical relationships between objects from the embedding.

Simultaneously to the present work, De Sa et al. (2018) analyzed the representation trade-offs for hyperbolic embeddings and proposed a new combinatorial embedding approach as well as a new approach to Multi-Dimensional Scaling (MDS) in hyperbolic space. Furthermore, Ganea et al. (2018) proposed a new model based on geodesically convex entailment cones for Poincaré embeddings to embed asymmetric relations.

Another related method is Order Embeddings (Vendrov et al., 2015), which was proposed to learn visual-semantic hierarchies over words, sentences, and images from ordered input pairs. In contrast, we are concerned with learning hierarchical embeddings from less supervision: namely, from *unordered* (symmetric) input pairs that provide no direct information about the partial ordering in the hierarchy.

Further work on embedding order-structures include Stochastic Triplet Embeddings (Van Der Maaten & Weinberger, 2012), Generalized Non-Metric MDS (Agarwal et al., 2007), and Crowd Kernels (Tamuz et al., 2011). In the context of word embeddings, Vilnis & McCallum (2015) proposed Gaussian Embeddings to learn improved representations. By mapping words to densities, this model is capable of capturing uncertainty, assymmetry, and (hierarchical) entailment relations.

To discover structural forms (e.g., trees, grids, chains) from data, Kemp & Tenenbaum (2008) proposed a model for making probabilistic inferences over a space of graph grammars. Recently, Lake et al. (2018) proposed an alternative approach to this work based on structural sparsity. Additionally, hierarchical clustering has a long history in machine learning and data mining (Duda et al., 1973). Bottom-up agglomerative clustering assigen each data point to its own cluster and then iteratively merges the two closest points according to a given distance measure (e.g., single link, average link, max link). As such, hierarchical clustering provides a hierarchical partition of the input space. In contrast, we are concerned with discovering hierarchical relationships directly between the input data points.

3. Methods

In the following, we describe our approach for learning continuous hierarchies from unstructured observations.

3.1. Hyperbolic Geometry & Poincaré Embeddings

Hyperbolic space is the unique, complete, simply connected Riemannian manifold with constant negative sectional curvature. There exist multiple equivalent[1] models for hyperbolic space and one can choose the model whichever is best suited for a given task. Nickel & Kiela (2017) based their approach for learning hyperbolic embeddings on the Poincaré ball

[1] Meaning that there exist transformations between the different models that preserve all geometric properties including isometry.

(a) Geodesics in the Poincaré disk.

(b) Lorentz model of hyperbolic geometry.

Figure 1: a) Geodesics in the Poincaré disk model of hyperbolic space. Due to the negative curvature of the space, geodesics between points are arc that are perpendicular to the boundary of the disk. For curved arcs, midpoints are closer to the origin of the disk (p1) than the associated points, e.g. (p3, p5). b) Points (p,q) lie on the surface of the upper sheet of a two-sheeted hyperboloid. Projection of points on the hyperboloid (p, q) onto the Poincaré disk.

model, due to its conformality and convenient parameterization. The Poincaré ball model is the Riemannian manifold $\mathcal{P}^n = (\mathcal{B}^n, g_p)$, where $\mathcal{B}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < 1\}$ is the *open* n-dimensional unit ball and where

$$g_p(\boldsymbol{x}) = \left(\frac{2}{1 - \|\boldsymbol{x}\|^2}\right)^2 g_e$$

The distance function on $\mathcal{P}$ is then defined as

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \operatorname{arcosh}\left(1 + 2\frac{\|\boldsymbol{x} - \boldsymbol{x}\|^2}{(1 - \|\boldsymbol{x}\|^2)(1 - \|\boldsymbol{y}\|^2)}\right) \quad (1)$$

It can be seen from Equation (1), that the distance within the Poincaré ball changes smoothly with respect to the norm of $\boldsymbol{x}$ and $\boldsymbol{y}$. This locality property of the distance is key for learning continuous embeddings of hierarchies. For instance, by placing the root node of a tree at the origin of $\mathcal{B}^n$, it would have relatively small distance to all other nodes, as its norm is zero. On the other hand, leaf nodes can be placed close to the boundary of the ball, as the distance between points grows quickly with a norm close to one.

3.2. Riemannian Optimization in the Lorentz Model

In the following, we propose a new method to compute hyperbolic embeddings based on the Lorentz model of hyperbolic geometry. The main advantage of the Lorentz parameterization is that it allows us to perform Riemannian optimization very efficiently. An additional advantage is that its distance function (see Equation (5)) avoids numerical instabilities that arise from the fraction in the Poincaré distance.

3.2.1. The Lorentz Model of Hyperbolic Space

In the following, let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n+1}$ and let

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_n y_n \quad (2)$$

denote the *Lorentzian scalar product*. The Lorentz model of n-dimensional hyperbolic space is then defined as the Riemannian manifold $\mathcal{L}^n = (\mathcal{H}^n, g_\ell)$, where

$$\mathcal{H}^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1, x_0 > 0\} \quad (3)$$

denotes the upper sheet of a two-sheeted n-dimensional hyperboloid and where

$$g_\ell(\boldsymbol{x}) = \begin{bmatrix} -1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}. \quad (4)$$

The associated distance function on $\mathcal{L}$ is then given as

$$d_\ell(\boldsymbol{x}, \boldsymbol{y}) = \operatorname{arcosh}(-\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}) \quad (5)$$

3.2.2. Riemannian Optimization

To derive the Riemannian SGD (RSGD) algorithm for the Lorentz model, we will first review the necessary concepts of Riemannian optimization. A Riemannian manifold $(\mathcal{M}, g)$ is a real, smooth manifold $\mathcal{M}$ equipped with a Riemannian metric g. Furthermore, for each $\boldsymbol{x} \in \mathcal{M}$, let $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ denote the associated *tangent space*. The metric g induces then a inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{x}} : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \times \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \to \mathbb{R}$. *Geodesics* $\gamma : [0, 1] \to \mathcal{M}$ are the generalizations of straight lines to Riemannian manifolds, i.e., constant speed curves

that are locally distance minimizing. The *exponential map* $\exp_{\boldsymbol{x}} : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \to \mathcal{M}$ maps a tangent vector $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ onto $\mathcal{M}$ such that $\exp_{\boldsymbol{x}}(\boldsymbol{v}) = \boldsymbol{y}$, $\gamma(0) = \boldsymbol{x}$, $\gamma(1) = \boldsymbol{y}$ and $\dot{\gamma}(0) = \frac{\partial}{\partial t}\gamma(0) = \boldsymbol{v}$. For a *complete manifold* $\mathcal{M}$, the exponential map is defined for all points $\boldsymbol{x} \in \mathcal{M}$.

Furthermore, let $f : \mathcal{M} \to \mathbb{R}$ be a smooth real-valued function over parameters $\theta \in \mathcal{M}$. In Riemannian optimization, we are then interested in solving optimization problems of the form

$$\min_{\theta \in \mathcal{M}} f(\theta) \tag{6}$$

Following Bonnabel (2013), we minimize Equation (6) using Riemannian SGD. In RSGD, updates to the parameters θ are computed via

$$\theta_{t+1} = \exp_{\theta_t}(-\eta \operatorname{grad} f(\theta_t)) \tag{7}$$

where $\operatorname{grad} f(\theta_t) \in \mathcal{T}_\theta\mathcal{M}$ denotes the *Riemannian gradient* and η denotes the learning rate.

For the Lorentz model, the tangent space is defined as follows: For a point $\boldsymbol{x} \in \mathcal{L}^n$, the tangent space $\mathcal{T}_{\boldsymbol{x}}\mathcal{L}^n$ consists of all vectors orthogonal to $\boldsymbol{x}$, where orthogonality is defined with respect to the Lorentzian scalar product. Hence,

$$\mathcal{T}_{\boldsymbol{x}}\mathcal{L}^n = \{\boldsymbol{v} : \langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathcal{L}} = 0\}$$

Furthermore, let $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{L}^n$. The exponential map $\exp_{\boldsymbol{x}} : \mathcal{T}_{\boldsymbol{x}}\mathcal{L}^n \to \mathcal{L}^n$ is then defined as

$$\exp_{\boldsymbol{x}}(\boldsymbol{v}) = \cosh(\|\boldsymbol{v}\|_{\mathcal{L}})\boldsymbol{x} + \sinh(\|\boldsymbol{v}\|_{\mathcal{L}})\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_{\mathcal{L}}} \tag{8}$$

To compute parameter updates as in Equation (6), we need additionally the Riemannian gradient of f at θ. For this purpose, we first compute the direction of steepest descent from the Euclidean gradient $\nabla f(\theta)$ via

$$\boldsymbol{h} = g_\ell^{-1} \nabla f(\theta) \tag{9}$$

Since g_ℓ is an involutory matrix (i.e., $g_\ell^{-1} = g_\ell$), the inverse in Equation (9) is trivial to compute. To derive the Riemannian gradient from $\boldsymbol{h} \in \mathbb{R}^{n+1}$, we then use the orthogonal projection $\operatorname{proj}_\theta : \mathbb{R}^{n+1} \to \mathcal{T}_\theta\mathcal{L}^n$ from the ambient Euclidean space onto the tangent space of the current parameter. This projection is computed as

$$\operatorname{proj}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{u} - \frac{\langle \boldsymbol{x}, \boldsymbol{u} \rangle_{\mathcal{L}}}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}}}\boldsymbol{x} = \boldsymbol{u} + \langle \boldsymbol{x}, \boldsymbol{u} \rangle_{\mathcal{L}}\boldsymbol{x}$$

since $\forall x \in \mathcal{H}^n : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1$. Using Equation (8) and section 3.2.2, we can then estimate the parameters θ using RSGD as in Algorithm 1.

Algorithm 1 Riemannian Stochastic Gradient Descent

Input Learning rate η, number of epochs T.
for $t = 1, \ldots, T$
$\quad \boldsymbol{h}_t \leftarrow g_{\theta_t}^{-1} \nabla f(\theta_t)$
$\quad \operatorname{grad} f(\theta_t) \leftarrow \operatorname{proj}_{\theta_t}(\boldsymbol{h}_t)$
$\quad \theta_{t+1} \leftarrow \exp_{\theta_t}(-\eta \operatorname{grad} f(\theta_t))$

3.2.3. Equivalence of models

The Lorentz and Poincaré disk model both have specific strengths: the Poincare disk provides a very intuitive method for visualizing and interpreting hyperbolic embeddings. The Lorentz model on the other hand is well-suited for Riemannian optimization. Due to the equivalence of both models, we can exploit their individual strengths simultaneously: points in the Lorentz model can be mapped into the Poincaré ball via the diffeomorphism $p : \mathcal{H}^n \to \mathcal{P}^n$, where

$$p(x_0, x_1, \ldots, x_n) = \frac{(x_1, \ldots, x_n)}{x_0 + 1} \tag{10}$$

Furthermore, points in $\mathcal{P}^n$ can be mapped into $\mathcal{H}^n$ via

$$p^{-1}(x_1, \ldots, x_n) = \frac{(1 + \|x\|^2, 2x_1, \ldots, 2x_n)}{1 - \|x\|^2}$$

We will therefore learn the embeddings via Algorithm 1 in the Lorentz model and visualize the embeddings by mapping them into the Poincaré disk using Equation (10). See also Figure 1b for an illustration of Lorentz model and its connections to the Poincaré disk.

3.3. Inferring Concept Hierarchies from Similarity

Nickel & Kiela (2017) considered the embedding of unweighted undirected graphs in hyperbolic space. Here, we extend this approach to a more general setting, where we intend to infer continuous hierarchies from *pairwise similarity measurements*.

Let $\mathcal{C} = \{c_i\}_{i=1}^m$ be a set of concepts and $X \in \mathbb{R}^{m \times m}$ be a dataset of pairwise similarity scores between these concepts. We also assume that the concepts can be organized according to an unobserved hierarchy $(\mathcal{C}, \preceq)$, where $c_i \preceq c_j$ defines a *partial order* over the elements of $\mathcal{C}$. Since partial order is a reflexive, anti-symmetric, and transitive binary relation, it is well suited to define hierarchical relations over $\mathcal{C}$. Furthermore, if $c_i \preceq c_j$ or $c_j \preceq c_i$, then the concepts c_i, c_j, then the concepts are *comparable* (e.g., located in the same subtree). Otherwise they are incomparable (e.g., located in different subtrees). For concepts $c_i \prec c_j$, we will refer to c_i as the superior and to c_j as the subordinate node.

Given this setting, our goal is then to recover the partial order $(\mathcal{C}, \preceq)$ from X. For this purpose, we separate the semantics of the partial order relation into two distinct aspects:

First, whether two concepts c_i, c_j are comparable (denoted by $c_i \sim c_j$) and, second, whether concept c_i is more *general* than c_j (denoted by $c_i \sqsubset c_j$). Combining both aspects provides us with the usual interpretation of partial order.

By explicitly distinguishing between the aspects of comparability and generality, we can then make the following structural assumptions on X to infer hierarchies from pairwise similarities: 1) Comparable (and related) concepts are more similar to each other than incomparable concepts (i.e., $X_{ij} \geq X_{ik}$ if $c_i \preceq c_j \wedge c_i \not\preceq c_k$); and 2) We assume that general concepts are similar to more concepts than less general ones. Both are mild assumptions given that the similarity scores X describe concepts that are organized in a latent hierarchy. For instance, assumption 1) is used in methods that use path-lengths in taxonomies to measure the *semantic similarity* of concepts (e.g., see (Resnik et al., 1999)).

It follows from assumption 1) the we want to preserve the similarity orderings in the embedding space in order to predict comparability. In particular, let $\boldsymbol{u}_i$ denote the embedding of c_i and let $\mathcal{N}(i,j) = \{k : X_{ik} < X_{ij}\} \cup \{j\}$ denote the set of concepts that are *less* similar to c_i then c_j (including c_j). Based only on pairwise similarities in X, it is difficult to make global decisions about the likelihood that $c_i \sim c_j$ is true. However, it follows from assumption 1) that we can make *local* decisions, i.e., we can infer that $c_i \sim c_j$ is more likely for all $c_k \in \mathcal{N}(i,j)$. For this purpose, let

$$\phi(i,j) = \underset{k \in \mathcal{N}(i,j)}{\arg\min}\, d(\boldsymbol{u}_i, \boldsymbol{u}_k)$$

be the nearest neighbor of c_i in the set $\mathcal{N}(i,j)$. We then learn embeddings $\Theta = \{\boldsymbol{u}\}_{i=1}^m$ by optimizing

$$\max_{\Theta} \sum_{i,j} \log \Pr(\phi(i,j) = j \mid \Theta) \tag{11}$$

where

$$\Pr(\phi(i,j) = j \mid \Theta) = \frac{e^{d(\boldsymbol{u}_i, \boldsymbol{u}_j)}}{\sum_{k \in \mathcal{N}(i,j)} e^{d(\boldsymbol{u}_i, vu_k)}}$$

For computational efficiency, we follow (Jean et al., 2015) and randomly subsample $\mathcal{N}(i,j)$ on large datasets.

Equation (11) is a ranking loss that aims to preserve the neighborhood structures in X. For each pair of concepts (i,j), this loss induces embeddings where (i,j) is closer in the embedding space than pairs (i,k) that are less similar. Since we compute the embedding in a metric space, we also retain transitive relations approximately. We can therefore identify the comparability of concepts $c_i \sim c_j$ by their distance $d(\boldsymbol{u}_i, \boldsymbol{u}_j)$ in the embedding.

Moreover, by optimizing Equation (11) in hyperbolic space, we are also able to infer the generality of concepts from their embeddings. According to assumption 2), we can can assume that general objects will be close to many different concepts. Since Equation (11) optimizes the local similarity ranking for all concepts, we can also assume that this ordering is preserved. We can see from Equation (1) that points with a small distance to many different points are located close to the center. We can therefore identify the generality of a concept c_i simply via the norm of its embedding $\|\boldsymbol{u}_i\|$.

We have now cast the problem of hierarchy discovery as a simple embedding problem whose objective is to preserve local similarity orderings

Table 1: Taxonomy Statistics. The number of edges refers to the full transitive closure of the respective taxonomy.

Taxonomy	Nodes	Edges	Depth
WORDNET Nouns	82,115	769,130	19
WORDNET Verbs	13,542	35,079	12
EUROVOC (en)	7,084	10,547	5
ACM	2,299	6,526	5
MESH	28,470	191,849	15

4. Evaluation

4.1. Embedding Taxonomies

In the following experiments, we evaluate the performance of the Lorentz model for embedding large taxonomies. For this purpose, we compore its embedding quality to Poincaré embeddings (Nickel & Kiela, 2017) on the following real-world taxonomies

WordNet WordNet (Miller & Fellbaum, 1998) is a large lexical database which, amongst other relations, provides hypernymy (is-a) relations. In our experiments, we embedded the noun and verb hierarchy of WordNet.

EuroVoc EuroVoc is a mulitlingual thesaurus maintained by the European Union. It contains keywords organized in 21 domains and 127 sub-domains. In our experiments, we used the English section of EuroVoc.

ACM The ACM computing classification system is a hierarchical ontology which is used by various ACM journals to organize subjects by area.

MeSH Medical Subject Headings (MeSH; (Rogers, 1963)) is a medical thesaurus which is provided by the U.S. National Library of Medicine. MeSH terminology is used to index and catalogue biomedical information in databases such as MEDLINE/PUBmed.

Statistics for all taxonomies are provided in Table 1.

In our evaluation, we closely follow setting of Nickel & Kiela (2017): First, we embed the undirected transitive

Table 2: Evaluation of Taxonomy Embeddings. MR = Mean Rank, MAP = Mean Average Precision ρ = Spearman rank-order correlation. $\Delta\%$ indicates the relative improvement of optimization in the Lorentz model.

		WORDNET Nouns			WORDNET Verbs			EUROVOC			ACM			MESH		
		2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
MR	Poincaré	90.7	5.09	4.12	10.71	1.39	1.35	2.83	1.25	1.23	4.14	1.8	1.71	61.11	14.05	12.8
	Lorentz	22.8	3.18	2.95	3.64	1.26	1.23	1.63	1.24	1.17	3.05	1.67	1.63	38.99	14.13	12.42
	$\Delta\%$	74.8	37.5	28.4	66.0	9.6	8.9	42.4	6.1	3.4	26.3	7.2	4.8	36.2	-0.5	2.9
MAP	Poincaré	11.8	82.8	86.5	36.5	91.0	91.2	64.3	94.0	94.4	69.3	94.1	94.8	19.5	76.3	79.4
	Lorentz	30.5	92.3	92.8	57.9	93.5	93.3	87.1	95.8	96.5	82.9	96.6	97.0	34.8	77.7	79.9
	$\Delta\%$	61.3	10.3	6.8	58.6	2.7	2.3	35.6	1.6	2.0	19.6	2.7	2.3	43.9	1.8	0.6
$\boldsymbol{\rho}$	Poincaré	13.8	57.2	58.5	11.0	54.1	55.1	37.5	57.5	61.4	59.8	63.5	62.9	42.2	69.9	74.9
	Lorentz	41.0	58.9	59.5	47.9	55.5	56.6	54.5	61.7	67.5	65.9	65.9	65.9	64.5	71.4	76.3

closure of these taxonomies, such that the hierarchical structure is not directly visible from the observed edges but has to be inferred. To measure the quality of the embedding, we compute for each observed edge (u, v) the corresponding distance $d(\boldsymbol{u}, \boldsymbol{v})$ in the embedding and rank it among the distances of all *unobserved* edges for u, i.e., among $\{d(\boldsymbol{u}, \boldsymbol{v}') : (u, v') \notin \mathcal{D}\}$. We then report the mean rank (MR) and mean average precision (MAP) of the ranking.

In addition, we also evaluate ho well the norm of the embeddings (i.e., our indicator for generality), correlates with the ground-truth ranks in the embedded taxonomy. Since different subtrees can have very different depths, we normalize the rank of each concept by the depth of its subtree and measure the Spearman rank-order correlation ρ of the normalized rank with the norm of the embedding. We compute the normalized rank in the following way: Let $\mathrm{sp}(c)$ denote the shortest path to the root node from c, and let $\mathrm{lp}(c)$ denote the longest path from c to any of its children.[2] The normalized rank of c is the given as

$$\mathrm{rank}(c) = \frac{\mathrm{sp}(c)}{\mathrm{sp}(c) + \mathrm{lp}(c)}.$$

To learn the embeddings in the Lorentz model, we employ the Riemannian optimization method as described in Section 3.2. For Poincaré embeddings, we use the official open-source implementation.[3] Both methods were cross-validated of identical sets of hyperparameters.

Table 2 shows the results of our evaluation. It can be seen that both methods are very efficient in embeddings these large taxonomies. However, the Lorentz model shows consistently higher-quality embeddings and especially so in low dimensions. The relative improvement of the two-dimensional Lorentz embeddings over the Poincaré embedding amounts to 78% on the WORDNET noun hierarchy and 42.4% on EUROVOC. Similar improvements can be observed on all taxonomies. Furthermore, on the most complex taxonomy (WORDNET nouns), the 10-dimensional Lorentz embeddings already out-performs the best reported numbers reported in (Nickel & Kiela, 2017) (which went up to 200 dimensions). This suggests that the full Riemannian optimization approach can be very helpful for obtaining good embeddings. This is especially the case in low dimensions where it is harder for the optimization procedure to escape from local minima.

4.2. Enron Email Corpus

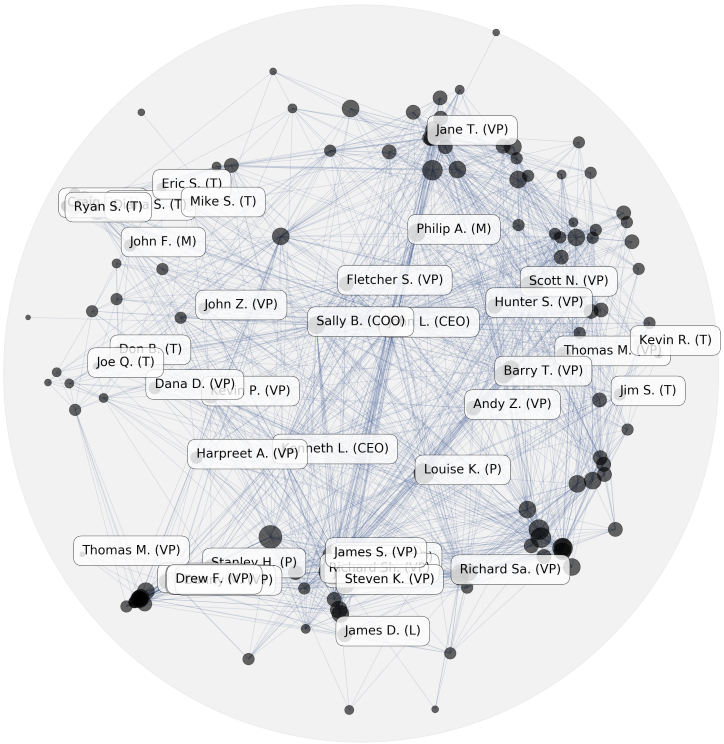
In addition to the taxonomies in Section 4.1, we are also interested in discovering hierarchies from real-world graphs that have not been generated from a clean DAG. For this purpose, we embed the communication graph of the Enron email corpus (Priebe et al., 2006) which consists of 125,409 emails that have been exchanged between 184 email addresses and 150 unique users.[4] From this data, we construct a graph where weighted edges represent the total number of emails that have been exchanged between two users. The dataset includes also the organizational roles for 130 users, which is largely based on the information collected by Shetty & Adibi (2005).

Figure 2a shows the two-dimensional embedding of this graph. It can be seen that the embedding captures important properties of the organizational structure. First, the nodes are approximately arranged according to the organizational hierarchy of the company. Executive roles such as CEOs, COOs, and (vice) presidents are embedded close to the origin, while other employees (e.g., traders and analysts) are

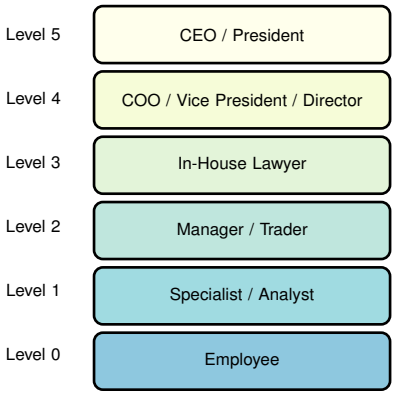
[2] Since all taxonomies in our experiments are DAGs, it is possible to compute the longest path in the graph

[3] Source code available at https://github.com/facebookresearch/poincare-embeddings

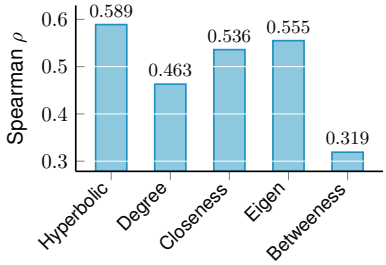
[4] This dataset has been created by Priebe et al. (2006) from the full Enron email corpus which has been released into public domain by the Federal Energy Regulatory Commission (FERC). The full corpus consists of 517,431 emails composed by 158 Enron employees.



(a) Embedding of the Enron communication graph



(b) Organizational hierarchy



(c) Rank-order correlation

Figure 2: Embedding of the Enron email corpus. Abbreviations in parentheses indicate organizational role: CEO = Chief Executive Officer, COO = Chief Operating Officer, P = President, VP = Vice President, D = Director, M = Manager, T = Trader. Blue lines indicate edges in the graph. Node size indicates node degree.

located closer to the boundary. Figure 2c shows the Spearman rank-order correlation of the norms of the embedding with the ground-truth organizational rank. It can be seen that the norm correlates well with the ground-truth ranking and is on-par or better than commonly-used centrality measures on graphs. We also observe that the embedding provides a meaningful clustering of users. For instance, the top of the disk shows a cluster of employees who worked as traders. Below that cluster (i.e., closer to the origin), are managers (e.g., John F.) and vice presidents (e.g., Fletcher S., Kevin B.) who have been associated with the trading arm of Enron. This illustrates that, in addition to the notion of rank in a hierarchy, the embedding provides also insight into the similarity of nodes within the hierarchy.

4.3. Historical Linguistics Data

The field of historical linguistics is concerned with the history of languages, their relations, and their origins. An important concept to determine the relations between languages are so-called *cognates*, i.e., words that are shared across different languages (but not borrowed) and which indicate common ancestry in the history of languages. To be classified as cognate, words must have similar meaning and systematic sound correspondences. Languages are assumed to be related if they share a large number of cognates.

The goal of our experiments was to discover the historical relationships between languages (which follow a hierarchical tree structure) by embedding cognate similarity data. For this purpose, we used the lexical cognate data provided by Bouckaert et al. (2012), which consists of 103 Indo-European languages and 6280 cognate sets in total. Since the number of cognate sets grew over time, not all languages are annotated with all possible sets. For this reason, we computed the cognate similarity between two languages in the following way. Let $c(\ell_1, \ell_2)$ denote the number of common cognates in languages ℓ_1, ℓ_2. Furthermore, let $a(\ell_1)$ denote the number of cognate annotations for ℓ_1. We then compute the cognate similarity of ℓ_1, ℓ_2 simply as

$$\mathrm{csim}(\ell_1, \ell_2) = \frac{c(\ell_1, \ell_2)}{\min(a(\ell_1), a(\ell_2))}.$$

Figure 3 shows a two-dimensional embedding of these cognate similarity scores. It can be seen that the embedding

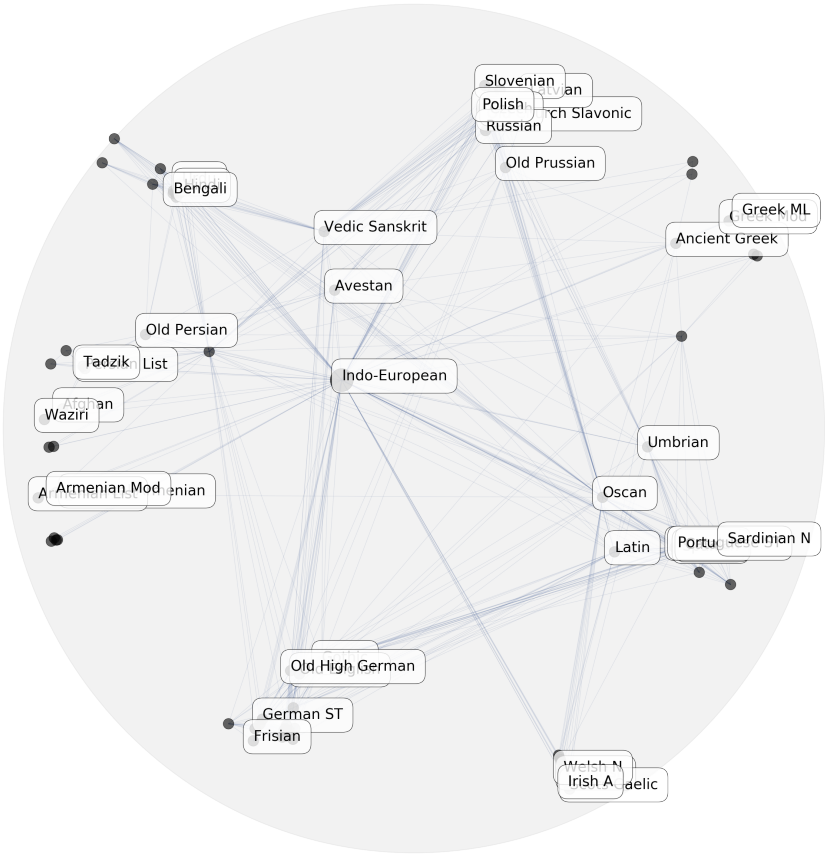


Figure 3: Embedding of the IELex lexical cognate data.

allows us to discover a meaningful hierarchy that corresponds well with the assumed origin of languages. First, the embedding clear clusters of high-level language families such as Celtic, Italic, Germanic, Balto-Slavic, Hellenic, Indic, and Iranian. Moreover, each of these cluster displays meaningful internal hierarchies such as (Gothic $\rightarrow$ Old High German $\rightarrow$ German), (Old Prussian $\rightarrow$ Old Church Slavonic $\rightarrow$ Serbo-Croation), (Latin $\rightarrow$ Italian), or (Ancient Greek $\rightarrow$ Greek). Closer to the center of the disc, we also find a number of ancient languages. For instance, Oscan and Umbrian are two extinct sister languages of Latin and located above the Italic cluster, Similarly, Avestan and Sanskrit are two ancient languages that separated early in the pre-historic era before 1800 BCE (Baldi, 1983). After separation, Avestan then developed in ancient Persia while Vedic-Sanskrit developed independently in ancient India. In the embedding, both languages are relatively close to the center and to each other. Furthermore, Avestan is close to the Iranian cluster while Vedic-Sanskrit is close to the Indian cluster.

5. Conclusion

We introduced a new method for learning continuous concept hierarchies from unstructured observations. We exploited the properties of hyperbolic geometry in such a way that we can discover hierarchies from pairwise similarity scores – under the assumption that concepts in the same subtree of the ground-truth hierarchy are more similar to each other than to concepts in different subtrees. To learn the embeddings, we developed an efficient Riemannian optimization approach based on the Lorentz model of hyperbolic space. Due to the more principled optimization approach, we were able to significantly improve the quality of the embeddings compared to the method proposed by Nickel & Kiela (2017) – especially in low dimensions. We further showed on two real-world datasets, that our method can discover meaningful hierarchies from nothing but pairwise similarity information.

Acknowledgments

The authors thank Joan Bruna, Martín Arjovsky, and Laurens van der Maaten for helpful discussions and suggestions.

References

Agarwal, Sameer, Wills, Josh, Cayton, Lawrence, Lanckriet, Gert, Kriegman, David, and Belongie, Serge. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*, pp. 11–18, 2007.

Antoniou, Grigoris and Van Harmelen, Frank. Web ontology language: Owl. In *Handbook on ontologies*, pp. 67–92. Springer, 2004.

Asta, Dena Marie and Shalizi, Cosma Rohilla. Geometric network comparisons. In Meila, Marina and Heskes, Tom (eds.), *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI*, pp. 102–110, 2015.

Baldi, Philip. *An introduction to the Indo-European languages*. SIU Press, 1983.

Boguñá, M, Papadopoulos, F, and Krioukov, D. Sustaining the internet with hyperbolic mapping. *Nature communications*, 1:62, 2010.

Bonnabel, Silvere. Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automat. Contr.*, 58(9): 2217–2229, 2013.

Bouckaert, Remco, Lemey, Philippe, Dunn, Michael, Greenhill, Simon J, Alekseyenko, Alexander V, Drummond, Alexei J, Gray, Russell D, Suchard, Marc A, and Atkinson, Quentin D. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097): 957–960, 2012.

Campbell, Lyle. *Historical linguistics*. Edinburgh University Press, 2013.

De Sa, Christopher, Gu, Albert, Ré, Christopher, and Sala, Frederic. Representation tradeoffs for hyperbolic embeddings. *arXiv preprint arXiv:1804.03329*, 2018.

Dodds, Peter Sheridan, Watts, Duncan J, and Sabel, Charles F. Information exchange and the robustness of organizational networks. *Proceedings of the National Academy of Sciences*, 100(21):12516–12521, 2003.

Duda, Richard O, Hart, Peter E, Stork, David G, et al. *Pattern classification*, volume 2. Wiley New York, 1973.

Ganea, Octavian-Eugen, Bécigneul, Gary, and Hofmann, Thomas. Hyperbolic entailment cones for learning hierarchical embeddings. *arXiv preprint arXiv:1804.01882*, 2018.

Inhelder, Bärbel and Piaget, Jean. *The growth of logic in the child*. Routledge & Paul, 1964.

Jean, Sébastien, Cho, Kyunghyun, Memisevic, Roland, and Bengio, Yoshua. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pp. 1–10, 2015.

Kemp, Charles and Tenenbaum, Joshua B. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

Kleinberg, Robert. Geographic routing using hyperbolic space. In *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pp. 1902–1909. IEEE, 2007.

Krioukov, Dmitri, Papadopoulos, Fragkiskos, Kitsak, Maksim, Vahdat, Amin, and Boguná, Marián. Hyperbolic geometry of complex networks. *Physical Review E*, 82 (3):036106, 2010.

Lake, Brenden M., Lawrence, Neil D., and Tenenbaum, Joshua B. The emergence of organizing structure in conceptual representation. *Cognitive Science*, 2018.

Mayr, Ernst. The role of systematics in biology. *Science*, 159(3815):595–599, 1968.

Miller, George and Fellbaum, Christiane. Wordnet: An electronic lexical database, 1998.

Nickel, Maximillian and Kiela, Douwe. Poincaré embeddings for learning hierarchical representations. pp. 6338–6347, 2017.

Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. Enron data set, 2006. URL http://cis.jhu.edu/~parky/Enron/enron.html.

Resnik, Philip et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130, 1999.

Rogers, FB. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116, 1963.

Shetty, J. and Adibi, J. Enron employee status report, 2005. URL http://www.isi.edu/~adibi/Enron/EnronEmployeeStatus.xls.

Silla, Carlos N and Freitas, Alex A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

Sun, Ke, Wang, Jun, Kalousis, Alexandros, and Marchand-Maillet, Stéphane. Space-time local embeddings. In *Advances in Neural Information Processing Systems 28*, pp. 100–108, 2015.

Tamuz, Omer, Liu, Ce, Belongie, Serge, Shamir, Ohad, and Kalai, Adam Tauman. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

Van Der Maaten, Laurens and Weinberger, Kilian. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pp. 1–6. IEEE, 2012.

Vendrov, Ivan, Kiros, Ryan, Fidler, Sanja, and Urtasun, Raquel. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Vilnis, Luke and McCallum, Andrew. Word representations via gaussian embedding. In *International Conference on Learning Representations (ICLR)*, 2015.