Hyperbolic Variational Auto-encoder for Remote Sensing Scene Classification

M. Hamzaoui, L. Chapel, M. T. Pham, S. Lefèvre

Université Bretagne-Sud, IRISA UMR 6074, 56000 Vannes, France

manal.hamzaoui@univ-ubs.fr

Résumé

Les espaces hyperboliques ont récemment attiré l'attention de la communauté de l'apprentissage automatique du fait qu'ils permettent de mieux représenter les données hiérarchiques que les espaces Euclidiens. En conséquence, de nombreux modèles populaires d'apprentissage automatique, tels que les Réseaux de Neurones de Graphes (GNNs) et les Auto-Encodeurs Variationnels (VAEs), ont été généralisés avec succès afin de représenter des données dans les espaces hyperboliques. Dans cet article, nous cherchons à savoir si les promesses faites par les différents travaux utilisant les espaces hyperboliques peuvent être atteintes dans le contexte des données de télédétection. À notre connaissance, il s'agit de la première évaluation des bénéfices des espaces hyperboliques dans la communauté de la télédétection. Nous nous focalisons particulièrement sur le problème de classification de scènes d'images de télédétection, dans lequel les exemples sont des images dont les étiquettes sémantiques sont généralement intrinsèquement structurées. Nous utilisons un Auto-Encodeur Variationnel pour projeter les données dans un espace latent hyperbolique et nous analysons l'organisation de l'espace induit en fonction de la structure des labels. Nous supervisons également l'apprentissage du VAE afin de guider la construction de l'espace latent en fonction de la hiérarchie des classes. Nous réalisons des expériences sur le jeu de données de télédétection PatternNet et effectuons une évaluation sur une tâche de classification, en prenant en compte la distance hiérarchique inter-classes. Les résultats expérimentaux indiquent que l'espace hyperbolique n'améliore pas la précision globale de la classification par rapport à un espace Euclidien, mais permet d'améliorer légèrement les performances lorsque l'on considère la distance entre l'étiquette prédite et la vraie étiquette dans la hiérarchie des labels.

Mots Clef

Espaces hyperboliques, Auto-Encodeur variationnel, télédétection, classification de scènes, labels hiérarchiques.

Abstract

Hyperbolic spaces have recently attracted attention in the machine learning community as they better handle hierarchical data than Euclidean spaces. Consequently, several popular machine learning models, such as Graph Neural Networks (GNNs) and Variational Auto-Encoders (VAEs), have been successfully generalized for data embedding in hyperbolic spaces. In this paper, we investigate whether the promises given by the various works that use hyperbolic spaces can be fulfilled in a remote sensing data context. To our knowledge, this is the first evaluation of the benefits of the hyperbolic spaces in the Remote Sensing community. We specifically focus on the remote sensing image scene classification problem, in which samples are images whose semantic labels usually have an intrinsic hierarchical structure. We use a Variational Auto-Encoder to project the data in a hyperbolic latent space and study the structure of the induced space w.r.t. the structure of the labels. We also supervise the VAE training in order to drive the latent space construction according to the class hierarchy. We carry out experiments on the remote sensing dataset PatternNet and perform an evaluation on a classification task, taking into account the inter-class hierarchical distance. Experimental results show that the Hyperbolic Space does not improve the global classification accuracy when compared with an Euclidean Space, but allows one to slightly improve the deviation among the misclassified examples when taking into account the distance between the predicted and the actual label in the label hierarchy.

Keywords

Hyperbolic spaces, Variational Auto-Encoder, remote sensing, scene classification, hierarchical labels.

1 Introduction

In most machine learning applications, the learning is performed on an Euclidean space, mostly because it has convenient mathematical properties, such as vectorial structures or closed forms for computing distances. Nevertheless, in many domains, real-world data do not possess an Euclidean structure [1] but can rather be represented with a hierarchical structure. In that case, they cannot be embedded in an Euclidean space with low distortion [2]. In the opposite, hyperbolic spaces [3] are manifolds that have been shown to represent efficiently hierarchical data in many applications, e.g. link prediction [3, 4], image embedding [12, 13], hierarchical clustering [11] or

word emdedding [10]. As such, there have been considerable recent works that use hyperbolic spaces to learn data representations, and various machine learning methods have been adapted to that setting. Among them, one can cite the hyperbolic-SVM [5] or hyperbolic neural network [6]. Other studies have provided a generalization of normal distributions on Hyperbolic spaces that can be used to build and learn a probabilistic model like Variational Auto-Encoder (VAE) [7, 8]. The proposed Hyperbolic VAEs have been used to embed images into a hyperbolic latent space then infer their underlying hierarchical structure. These methods have been validated on MNIST and Atari 2600 Breakout datasets, showing that the H-VAE is able to retrieve their hierarchical nature. However, those datasets are simple and do not reflect complex scenarii as in real-world images. Moreover, the MNIST dataset is not hierarchical whereas real-world images can show hierarchical structures, either within the image [9], or between the images, when a hierarchy of classes is available [12, 13, 14]. Considering these particularities, it is a challenging task to use these hyperbolic VAEs to embed real-world images and expect the latent space to be hierarchically organized according to the image hierarchy. An interesting study [15] was suggesting to rather guide the VAE learning in order to drive the construction of its latent space such that it reflects a given class hierarchy.

In this paper, we aim to investigate whether the promises given by H-VAE can be fulfilled in a remote sensing data context. We focus especially on the scene classification problem [18, 21], which is an active research topic in the field of computer vision applied to earth observation. Its aim is to classify scene images into a set of classes according to the image contents and has been applied in a wide variety of scenarii (land use/land cover mapping, disaster relief, etc.). In many contexts, the images' labels are not flat but are rather hierarchically organized, depending on the level of details that is required. To the best of our knowledge, this is the first evaluation of hyperbolic spaces in the remote sensing community. In contrary to several works that have showed the interest of working in hyperbolic spaces when the data are hierarchical, we could not show the interest of using such a space in the remote sensing scene classification application.

The rest of this paper is structured as follows. After providing some details about the hyperbolic spaces, Section 2 describes the H-VAE that will be used to provide an embedding of the images. Section 3 details our motivation and how we guide the VAE learning. Section 4 discusses the experimental results. Conclusion and future works are given in Section 5.

2 Hyperbolic Variational Auto-encoders

In this section, we briefly recall the required background on Variational Auto-Encoders in the Euclidean space. We then review the Lorentz model of hyperbolic geometry, in which we outline some of the mathematical preliminaries that are needed to define the Hyperbolic version of VAE.

2.1 Variational Auto-Encoder

Variational Auto-Encoder (VAE) [16] is a probabilistic generative model relevant to representation learning in which we aim to learn good representations, such as interpretable representations or representations that give a better generalization [7]. A VAE model is composed of two components: a (stochastic) encoder that embeds observations x into a low dimensional latent space $z \in Z$, and a decoder generating observations $\hat{x}$ from this latent space. Formally, the VAE consists of a probabilistic decoder defined as a likelihood function $p_\theta(x^{(i)}|z)$ and parameterized by θ which generates data $\hat{x}^{(i)}$ given the latent variable z as well as a posterior distribution $q_\phi(z|x^{(i)})$ that can be considered as a probabilistic encoder parameterized by ϕ.

The parameters ϕ and θ are learned simultaneously by maximizing the evidence lower bound (ELBO). ELBO is defined for each observation $x^{(i)}$ by:

$$\begin{aligned}\log p_\theta(x^{(i)}) \geq & \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] \\ & - D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)),\end{aligned} \tag{1}$$

where the first term after the inequality encourages the decoder to learn to reconstruct the observation, and the second is a regularization term that promotes output representations to follow a predefined distribution, $\mathbb{E}$ and D_{KL} being respectively the expectation and the Kullback–Leibler (KL) divergence. Usually, $p_\theta(z)$ is chosen as a standard Normal distribution with mean zero and variance one.

In practice, we approximate the reconstruction term using a Monte Carlo estimator:

$$\begin{aligned}&\mathbb{E}_{z\sim q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] \approx \\ &\qquad \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x^{(i)}|g_\phi(\epsilon^{(i,l)}, x^{(i)})),\end{aligned} \tag{2}$$

where L is the number of samples per data point $x^{(i)}$, $g_\phi(\epsilon^{(i,l)}, x^{(i)}) = \mu_\phi^{(i)} + \sigma_\phi^{(i)} \odot \epsilon^{(i,l)}$ is the reparameterization trick, $\odot$ indicates an element-wise product and $\epsilon^{(i,l)} \sim \mathcal{N}(0, I)$ is a random noise vector. $\mu_\phi^{(i)}$ and $\sigma_\phi^{(i)}$ are outputs of the encoder, representing respectively the mean and the standard deviation of the target distribution.

The regularization term D_{KL} encourages the approximate posterior $q_\phi(z|x^{(i)})$ to be close to the prior $p_\theta(z)$ and is defined as:

$$\begin{aligned}D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) &= \mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right] \\ &= -\frac{1}{2}\sum_{j=1}^{J}\left[1 + \log({\sigma_{j,\phi}^{(i)}}^2) - {\sigma_{j,\phi}^{(i)}}^2 - {\mu_{j,\phi}^{(i)}}^2\right],\end{aligned} \tag{3}$$

where J is the dimension of z, $\mu_{j,\phi}^{(i)}$ and $\sigma_{j,\phi}^{(i)}$ denote the j^{th} element of the encoder outputs.

2.2 Hyperbolic Geometry

Hyperbolic spaces have recently attracted a lot of attention in the machine learning community as they are more suitable to handle hierarchical data than Euclidean spaces. This is thanks to their geometric properties that make the space growing exponentially with distance from the origin unlike the Euclidean space which grows polynomially [3] (see Figure 1). Within the Riemannian geometry framework, hyperbolic spaces are manifolds and several models of n dimensional spaces exist. In this paper, we consider a particular model of Hyperbolic spaces that has recently become very successful, namely the Lorentz model (also known as the hyperboloid model). It is computationally attractive as it has a simple closed-form distance function as well as analytical forms for the exponential map, logarithmic map and parallel transport, preliminary mathematical notions that will be necessary to adjust the Euclidean VAE to hyperbolic spaces. We now briefly review the Lorentz model (Figure 2). We do not assume a background in Riemannian geometry, readers can refer to [8] for more details.

Lorentz Model. The Lorentz Model is a Riemannian manifold defined as $\mathcal{L}^n = (\mathcal{H}^n, g_l)$, where g_l is the Riemannian metric tensor and $\mathcal{H}^n$ denotes the upper sheet of a two sheeted n-dimentional hyperboloid:

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

where $\langle ., .\rangle_{\mathcal{L}}$ is the Lorentzian inner product, also known as the metric tensor, defined as: for $x, y \in \mathbb{R}^{n+1}$, $\langle x, y\rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i$.
We note that for any point $x = (x_0, x') \in \mathbb{R}^{n+1}$

$$x \in \mathcal{H}^n \Leftrightarrow x_0 = \sqrt{1 + ||x'||^2}.$$

The origin of the hyperbolic space is referred as a one-hot vector $\mu_0 = [1, 0, \ldots, 0] \in \mathcal{H}^n$. In addition, the shortest path between two points $x, y \in \mathcal{H}^n$ is given by the geodesic distance defined as:

$$d_l(x, y) = \text{arcosh}(-\langle x, y\rangle_{\mathcal{L}}).$$

Exponential and logarithmic map. Working in hyperbolic spaces is not easy: it requires generalizing basic operations, such as vector addition, matrix-vector multiplication and vector translation to these spaces, which is not trivial or sometimes even impossible [6]. A simple and straightforward way to accomplish this is to move the data from a hyperbolic space to a tangent space, a local Euclidean space in which the operations are constructed as in Euclidean space [6]. To switch respectively from and to the hyperbolic space, exponential (resp. logarithmic) map is employed. We formalize these notions as follows.

The tangent space $\mathcal{T}_\mu \mathcal{H}^n$ at point $\mu \in \mathcal{H}^n$ can be described as a subspace of $\mathbb{R}^{n+1}$. Formally, it is represented by a set of points $u \in \mathbb{R}^{n+1}$ satisfying the orthogonality relation with respect to the Lorentzian product:

$$\mathcal{T}_\mu \mathcal{H}^n = \{u \in \mathbb{R}^{n+1} \mid \langle u, \mu\rangle_{\mathcal{L}} = 0\}.$$

Note that $\mathcal{T}_{\mu_0} \mathcal{H}^n$, the tangent space at the origin, consists of points $u \in \mathbb{R}^{n+1}$ with $u_0 = 0$ and $||u||_{\mathcal{L}} = \langle x, y\rangle_{\mathcal{L}} = ||u||_2$.

The exponential map This is a function that projects a tangent space vector $u \in \mathcal{T}_\mu \mathcal{H}^n$ onto the hyperbolic space $\mathcal{H}^n$. It is defined locally and only projects a small neighbourhood of the tangent space origin μ onto its neighbourhood in the hyperbolic space. The exponential map of the Lorentz model is then given by:

$$\begin{aligned} \exp_\mu &: \mathcal{T}_\mu \mathcal{H}^n \to \mathcal{H}^n \\ exp_\mu(u) &= \cosh(||u||_{\mathcal{L}}).\mu + \sinh(||u||_{\mathcal{L}}).\frac{u}{||u||_{\mathcal{L}}}. \end{aligned} \tag{4}$$

The logarithmic map Also known as the inverse exponential map. It is defined as:

$$\begin{aligned} \log_\mu &: \mathcal{H}^n \to \mathcal{T}_\mu \mathcal{H}^n \\ \log_\mu(z) &= \exp_\mu^{-1}(z) = \frac{\text{arcosh}(\alpha)}{\sqrt{\alpha^2 - 1}}(z - \alpha\mu), \end{aligned} \tag{5}$$

where $z,\ \mu \in \mathcal{H}^n$ and $\alpha = -\langle \mu, z\rangle_{\mathcal{L}}$.

Parallel transport. For any couple of points $\mu, \nu \in \mathcal{H}^n$, parallel transport from ν to μ is a map that carries a vector $v \in \mathcal{T}_\nu \mathcal{H}^n$ along the geodesic to their corresponding vector $v' \in \mathcal{T}_\mu \mathcal{H}^n$ while preserving its metric tensor i.e. $\langle PT_{\nu\to\mu}(v), PT_{\nu\to\mu}(u)\rangle_{\mathcal{L}} = \langle v, u\rangle_{\mathcal{L}}$. For the Lorentz model, this map is given by

$$\text{PT}_{\nu\to\mu}(v) = v + \frac{\langle \mu - \alpha\nu, v\rangle_{\mathcal{L}}}{\alpha + 1}(\nu + \mu), \tag{6}$$

where α is defined as above. The inverse parallel transport $PT^{-1}_{\nu\to\mu}$ simply carries back the vector in $\mathcal{T}_\mu \mathcal{H}^n$ to $\mathcal{T}_\nu \mathcal{H}^n$ along the geodesic and is defined as:

$$v = \text{PT}^{-1}_{\nu\to\mu}(u) = \text{PT}_{\mu\to\nu}(u). \tag{7}$$

Equipped with all these mathematical preliminaries about the Lorentz model, we can now describe the Hyperbolic VAE.

2.3 Hyperbolic Variational Auto-Encoder

Hyperbolic Variational Auto-Encoder (H-VAE) is a variant of VAE (we choose the E-VAE notation for Euclidean VAE) in which the latent variables are defined on a Hyperbolic space. As such, it is necessary to adapt the normal distribution defined for an Euclidean space so that it operates in a Hyperbolic space. A wrapped normal distribution was proposed by [8] for the Lorentz model, which we denote $\mathcal{G}(\mu, \Sigma)$, where $\mu \in \mathcal{H}^n$ and Σ are defined as positive. Sampling from this distribution can be summarized in 3 steps:

(1) sample a vector from the Gaussian distribution $\tilde{v} \sim \mathcal{N}(0, \Sigma)$ and interpret it as an element of the tangent space at the origin μ_0, $v = [0, \tilde{v}] \in \mathcal{T}_{\mu_0} \mathcal{H}^n$;

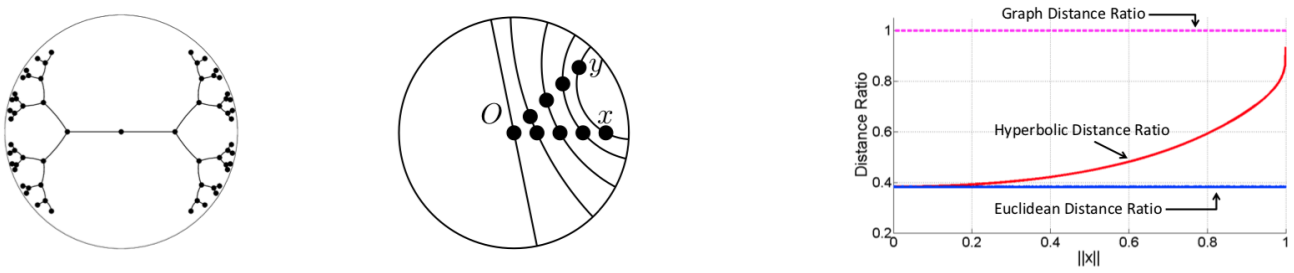

Figure 1 – Left: Embeddings of a binary tree in Poincaré disk (one of the Hyperbolic models), Right: Geodesic and distances. The distance $d_H(x, y)$ approaches $d_H(x, O) + d_H(O, y)$ as x and y move towards the outside of the disk (Figure from [2]).

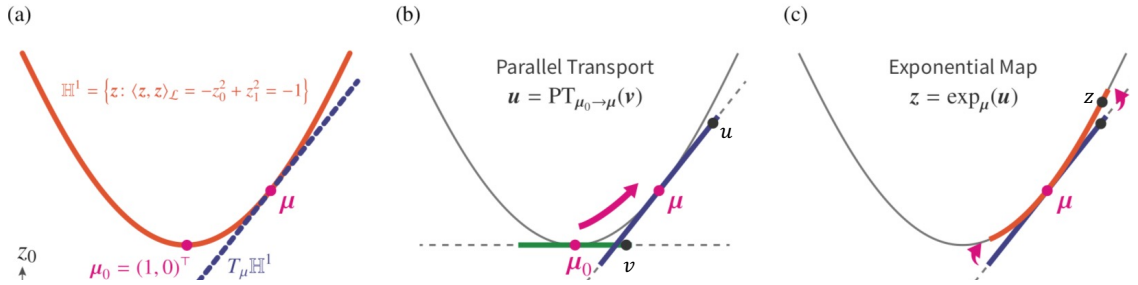

Figure 2 – (a) in red the one-dimensional Lorentz model $\mathcal{H}^1$ and its tangent space $\mathcal{T}_\mu \mathcal{H}^1$ (blue). (b) Parallel transport that carries $b \in \mathcal{T}_{\mu_0} \mathcal{H}^1$ (green) to $u \in \mathcal{T}_\mu \mathcal{H}^1$ (blue) while preserving $||.||_{\mathcal{L}}$. (c) Exponential map projects the $u \in \mathcal{T}_\mu \mathcal{H}^1$ (blue) to $z \in \mathcal{H}^1$(red) (Figure from [8]).

(2) parallel transport $v \in \mathcal{T}_{\mu_0} \mathcal{H}^n$ to the tangent space of the desired location μ, $u = PT_{\mu_0 \to \mu}(v)$;

(3) use $\exp_\mu$ to map the transported vector u from the tangent space $\mathcal{T}_\mu \mathcal{H}^n$ to the manifold $\mathcal{H}^n$, $z = \exp_\mu(u)$.

This sampling strategy is used in the H-VAE as a reparameterization trick. Therefore our hyperbolic latent variables $z^{(i)} \sim q_\phi(z|x^{(i)})$ are defined as:

$$z^{(i)} = g_\phi(v^{(i)}, \mu_\phi^{(i)}) = \exp_{\mu_\phi^{(i)}}(PT_{\mu_0 \to \mu_\phi^{(i)}}(v^{(i)})), \quad (8)$$

where $v^{(i)} = [0, \tilde{v}^{(i)}], \tilde{v}^{(i)} \sim \mathcal{N}(0, \Sigma_\phi^{(i)})$, $\Sigma_\phi^{(i)}$ and $\mu_\phi^{(i)}$ are outputs of the encoder. $\mu_\phi^{(i)}$ is assured to be in $\mathcal{H}^n$ by applying $\exp_{\mu_0}$ to the final layer of the encoder.
The Kullback-Leibler divergence must also be adapted to the hyperbolic space. According to eq. (3), we only need to redefine the logarithmic probability density function which is then given, in Lorentz space, as:

$$\log q_\phi(z|x^{(i)}) = \log p(v^{(i)}) - (n-1) \log \left(\frac{\sinh(||u^{(i)}||_{\mathcal{L}})}{||u^{(i)}||_{\mathcal{L}}} \right) \quad (9)$$

and

$$\log p_\theta(z) = \log p(v_0^{(i)}) - (n-1) \log \left(\frac{\sinh(||u_0^{(i)}||_{\mathcal{L}})}{||u_0^{(i)}||_{\mathcal{L}}} \right), \quad (10)$$

where $q_\phi(z|x^{(i)})$ and $p_\theta(z)$ are the wrapped normal distributions. $p(v)$ is the normal distribution in the tangent space at the origin μ_0.

$u^{(i)} = \exp^{-1}_{\mu_\phi^{(i)}}(z)$ is the projection of the hyperbolic embedding z into the tangent space $\mathcal{T}_{\mu_\phi^{(i)}} \mathcal{H}^n$.

$v^{(i)} = PT^{-1}_{\mu_0 \to \mu_\phi^{(i)}}(u^{(i)})$ is the transported vector from the tangent space $\mathcal{T}_{\mu_\phi^{(i)}} \mathcal{H}^n$ to the tangent space at the origin $\mathcal{T}_{\mu_0} \mathcal{H}^n$.

$u_0^{(i)} = \exp^{-1}_{\mu_0}(z) = PT^{-1}_{\mu_0 \to \mu_0}(u_0^{(i)}) = v_0^{(i)}$ is the projected vector of the hyperbolic embedding z in the tangent space at the origin $\mathcal{T}_{\mu_0} \mathcal{H}^n$.

3 When H-VAE meets remote sensing scene classification

3.1 Motivation

Remote Sensing scene classification aims at categorizing aerial or satellite images into a discrete set of meaningful classes based on the images' content. Those labels are often semantic and can been organized hierarchically. For instance, an image can be classified as *freeway*, *highway* or *transportation*, depending on the level of details we focus on. Remote sensing scene classification is a challenging task; scene images can exhibit confusing visual similarity between different classes, significant intra-class variation that may even be greater than the inter-class variance and similar semantic classes may show significant visual dissimilarity [20]. For instance, although a *freeway* and a *runway* have visual similarities, they are semantically unrelated. In contrast, the *runway* and the *airplane* classes

may be visually distinct but semantically similar as they may both be instances of the *Airport* class. As such, state-of-the-art methods do not build on raw pixels but rather start by defining efficient features to rely on for classification task [21]. In this paper, we rely on the VAE that has been successfully used to extract meaningful features in that context and perform a classification step on the resulting embedding. As the labels are intrisically hierarchical, we aim in this paper at studying what is the impact of considering a Hyperbolic Embedding rather than an Euclidean one. We study two settings: i) are the Hyperbolic VAE embeddings reflecting the hierarchical organization of the classes? (as it is done in [7]) ii) when guiding the construction of the embeddings with the taxonomy of classes, does the hyperbolic space provide a better organization of the space, and then better classification performances? We specify that in this study, we consider a simple VAE architecture to limit its impact on the conclusions.

3.2 Label-driven VAE learning

VAE only considers visual information when learning image embeddings. Here, we detail the incorporation of the hierarchical class structure into the VAE learning process so as to supervise and guide the construction of the latent space Z (see Figure 4). To do this, and following [15, 4], we use a class hierarchy-based pairwise similarity measurement between images, which aims to bring semantically similar images closer together and distancing them from those that are less similar.

We therefore drive the construction of our latent space Z by optimizing the Soft Local Ranking (SLR) loss defined as :

$$\mathcal{L}_{\text{SLR}}(x^{(i)}, \mathcal{T}; \phi) = \sum_{i,j} \log \Pr(x^{(i)}, x^{(j)}; \phi),$$

where

$$\Pr(x^{(i)}, x^{(j)}; \phi) = \frac{e^{-d_l(\mu_\phi^{(i)}, \mu_\phi^{(j)})}}{\sum_{j' \in \mathcal{N}(i,j)} e^{-d_l(\mu_\phi^{(i)}, \mu_\phi^{(j')})}}, \tag{11}$$

where γ is where $\mu_\phi^{(i)}$ is the hyperbolic mean of the input image $x^{(i)}$, $d_l(\mu_\phi^{(i)}, \mu_\phi^{(j)})$ is the Lorentzian distance between $\mu_\phi^{(i)}$ and $\mu_\phi^{(j)}$.

$\mathcal{N}(i,j)$ is the set of images semantically less similar to $x^{(i)}$ than $x^{(j)}$ including $x^{(j)}$, which is given by $\mathcal{N}(i,j) = \{j' : d_\mathcal{T}(l^{(i)}, l^{(j')}) > d_\mathcal{T}(l^{(i)}, l^{(j)})\} \cup \{j\}$ where $d_\mathcal{T}(l^{(i)}, l^{(j)})$ is the path-length between $l^{(i)}$ and $l^{(j)}$, labels of images $x^{(i)}$ and $x^{(j)}$ respectively, in the class hierarchy $\mathcal{T}$.

We then formulate our label-driven VAE for scene image embedding as:

$$\arg\max_{\phi,\theta} \left(\mathcal{L}_{\text{ELBO}}(x; \phi, \theta) + \gamma \, \mathcal{L}_{\text{SLR}}(x; \mathcal{T}, \phi) \right), \tag{12}$$

The first term is the VAE objective which embeds the scene images based on their visual similarity, while the second term is the SLR objective detailed above.

The label-driven VAE objective can thus be detailed as:

$$\begin{aligned} \arg\max_{\phi,\theta} (& \mathbb{E}_{z \sim q_\phi(z|x^{(i)})} \left[\log p_\theta(x^{(i)}|z) \right] \\ & - \beta \, D_{KL} \left(q_\phi(z|x^{(i)}) || p_\theta(z) \right) \\ & + \gamma \, \mathcal{L}_{\text{SLR}}(x, \mathcal{T}; \phi)), \end{aligned} \tag{13}$$

where $x^{(i)}$ are scene images, ϕ and θ are VAE parameters, β and γ are the scaling hyperparameters controlling the weight relative to the KL divergence and SLR during training.

4 Experimental study

In the following experiments, we evaluate the quality of the hyperbolic VAE image embedding for the scene classification task in both unguided and guided scenarii. We first introduce a remote sensing scene dataset whose labels are hierarchically organized. Then, we describe the network architecture and the parameter settings. Finally, in order to assess the interest of the hyperbolic embedding in our context, we compare the classification results based on a k-nearest neighbor applied on the embeddings provided by an E-VAE, an H-VAE and a label-driven H-VAE.

4.1 Experimental setup

Dataset and implementation. PatternNet [18] is a high-resolution remote sensing dataset for scene classification and retrieval. It contains $30\,400$ images of 256×256 pixels with a resolution ranging from 0.062 to 4.693 m per pixel. The dataset covers 38 classes of scenes, such as an airplane, baseball field, basketball court, beach, bridge, and cemetery, that are organized hierarchically [14] and a subtree of the 3-level label tree is depicted in Figure 3.

To investigate the relevance of the hyperbolic geometry for remote sensing scene classification, we randomly select for each class 100 images for the training set, 50 images for the validation set and 80 images for the test set. For all experiments, we report the average and standard deviations over three runs. We implemented the described approach based on [15, 8].

Architecture of the VAE and classification method. For both the E-VAE and the H-VAE, we choose the same following architecture. Both the encoder and the decoder are composed of 5 convolutional layers and a linear layer, each convolutional layer is followed by a batch normalization layer and a Leaky ReLU activation, except for the decoder last convolutional layer which is followed by a sigmoid activation. The input size is set to 64×64. The latent space dimension d of the embedding z is set to 8, 16, 32, 64 and 128 respectively. Note that this architecture is very simple compared to the one used in remote sensing [17, 21] but here we are not looking for high performance.

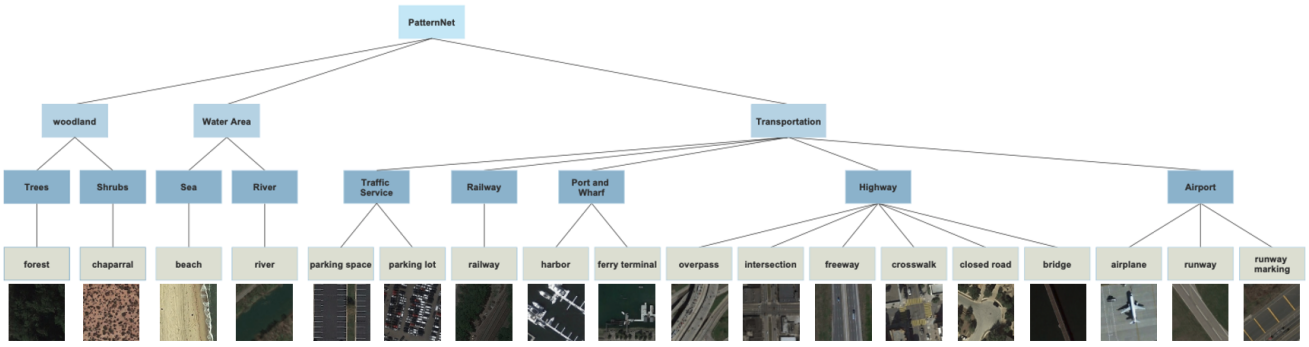

Figure 3 – Sub-branch of the category tree of the Remote Sensing PatternNet dataset. The leaves correspond to classes, the length of a single edge is equal to 0.5, and the distance between two given classes can take one of the following values: 0, 1, 2 and 3.

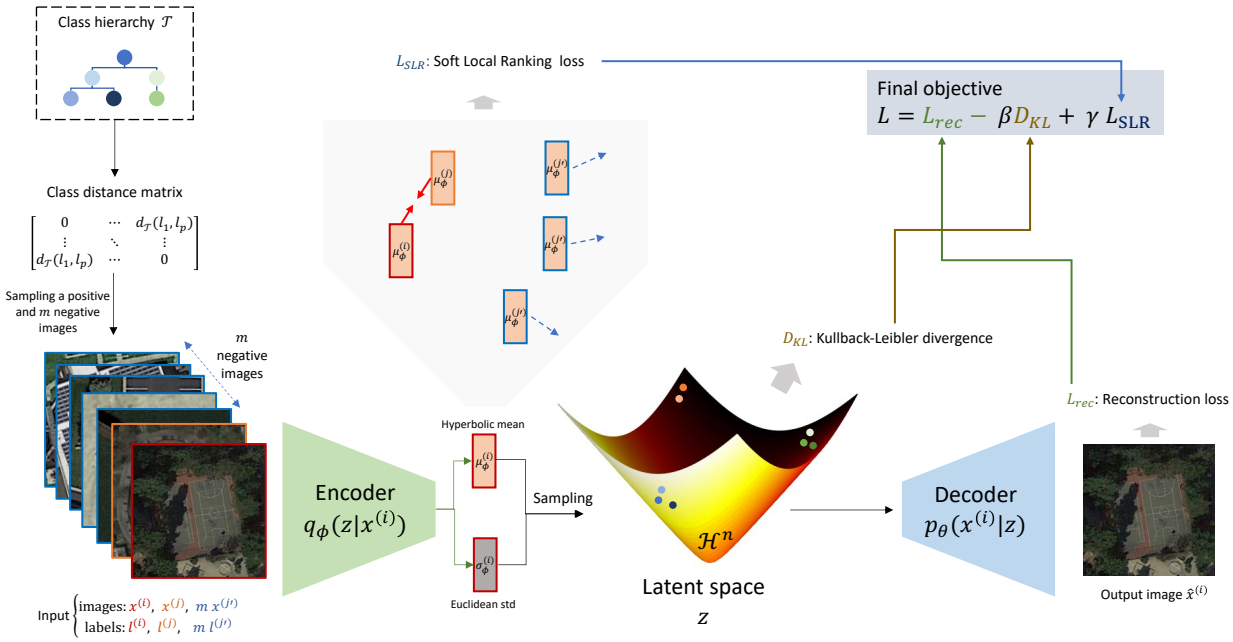

Figure 4 – Overview of the Hyperbolic VAE for remote sensing image embeddings.

We report the performances of the models for the following parameter values, as they allowed us to get the best results. The Adam optimizer [19] acts as our optimizer with a constant learning rate of $1e^{-3}$. The models are trained with mini-batches of size 64 for 350 epochs, an early stopping of 50 epochs and $m = 10$ negative samples to optimize the SLR term. The ELBO term is approximated by Monte Carlo (MC) estimation with $K = 1$. β and γ scaling hyperparameters weights of the KL divergence and the SLR loss were chosen experimentally and set to $5e^{-5}$ and $1e^{-3}$ respectively. For classification, we adopt, without any loss of generality, the k-Nearest Neighbors (k-NN) classifier with k set to 1.

Evaluation metrics. We consider two different evaluation metrics to assess the quality of the VAE embeddings. The first one is the classification accuracy computed at each level of the hierarchy (except at the root). The second one is a distance-based metric that measure the coherency of the classes w.r.t. the hierarchy, which is an adaptation of the mAD@k (mean Average Deviation at cutoff k) metric proposed for scene retrieval evaluation in [14]. This metric measures, for the misclassified samples, how far they are from the actual class considering a distance computed on the label tree (the smaller the better). It is defined as follows:

$$mAD = \begin{cases} 0, & \text{if } |N| = 0; \\ \frac{1}{|N|} \sum_{i=1}^{|N|} d_{\mathcal{T}}(l^{(i)}, pl^{(i)}) & \text{otherwise,} \end{cases} \quad (14)$$

where N is the set of misclassified labels, $|N|$ is the number of misclassified labels, $l^{(i)}$ and $pl^{(i)}$ are the true and the predicted labels, respectively. $d_{\mathcal{T}}(i, j)$ is the path-length between labels i and j in the class hierarchy $\mathcal{T}$; the path-length between two classes is defined as the total edge

weight (a single edge is equal to 0.5) of the path between the two associated leaves and can take one of the following values: 0, 1, 2 and 3.

4.2 Results

Table 1 reports the classification accuracy at different levels of the PatternNet class hierarchy.

When comparing E-VAE and H-VAE, we observe that E-VAE outperforms H-VAE in terms of classification accuracy across different levels and dimensions. In opposite to what was observed in several papers in the literature [7, 8, 15], H-VAE does not provide a latent space that better reflects the label's hierarchy than E-VAE. While the performances of E-VAE increase or remain stable with the latent space dimension, we surprinsingly observe that the H-VAE gets worst performances for dimension greater than 32. We conjecture that this is due to the numerical instability of the hyperbolic space which was observed in [22] but this requires a further investigation. Guiding the H-VAE learning with the SLR term improves the performance of H-VAE but still with scores lower than the (un-guided) E-VAE. Interestingly enough, it avoids the degradated performances when increasing the dimension of the latent space.

We now study how the misclassified images are organized w.r.t. the label's hierarchy. Table 2 gives the mAD performance. Again, we observe that E-VAE outperforms H-VAE. We also note that the performances drop when the dimension increases. Nevertheless, when guiding the latent space, one can note that it has similar or sometimes better performances that the E-VAE.

5 Conclusion

Hyperbolic embeddings have captured the attention of the machine learning community thanks to their ability to represent more efficiently hierarchically-structured data, showing better results than Euclidean embeddings in many domains. In this work, we investigate, for the first time, the ability of hyperbolic spaces in a remote sensing scene classification context, in which scene semantic labels have an intrinsic hierarchical structure. We first performed a feature extraction step, considering a Variational Auto-encoder, to embed the scene images, and drove the learning of the latent space such that it fits the label's hierarchy. Evaluating the hyperbolic latent space in a classification context, we did not highlight the advantages of these hyperbolic spaces over Euclidean ones.

In future works, we plan to investigate the use of more complex architectures in order to define more efficient features to describe the scenes. Indeed, in remote sensing, deep networks such as VGG, AlexNet, ResNet are generally used for image feature extraction [17]. We also plan to define dedicated loss functions that bring out the hierarchical aspect of the data more efficiently.

Acknowledgement

This work was supported by the ANR Multiscale project under the reference ANR-18-CE23-0022.

References

[1] M. M. Bronstein, J. Bruna, Y. LeCun, A.Szlam, & P. Vandergheynst, Geometric deep learning: going beyond euclidean data, *IEEE Signal Processing Magazine*, vol. 34, pp. 18-42, 2017.

[2] F.Sala, C.De Sa, A. Gu, & C. Ré, Representation tradeoffs for hyperbolic embeddings, *International Conference on Machine Learning (ICML)*, pp. 4460-4469, 2018.

[3] M. Nickel, & D. Kiela, Poincaré embeddings for learning hierarchical representations, *Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 6338-6347, 2017.

[4] M. Nickel, & D. Kiela, Learning continuous hierarchies in the lorentz model of hyperbolic geometry, PMLR. *International Conference on Machine Learning (ICML)*, vol. 80, pp. 3779-3788, 2018.

[5] H. Cho, B. DeMeo, J. Peng, & B. Berger, Large-Margin Classification in Hyperbolic Space, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1832-1840, 2019.

[6] O. E. Ganea, G. Bécigneul, & T.Hofmann, Hyperbolic neural networks, *Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 5350–5360, 2018.

[7] E. Mathieu, C. L. Lan, C. J. Maddison, R.Tomioka, & Y. W. Teh, Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders, *Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 12544-12555, 2019.

[8] Y. Nagano, S. Yamaguchi, Y.Fujita, & M. Koyama, A wrapped normal distribution on hyperbolic space for gradient-based learning, *International Conference on Machine Learning (ICML)*, pp. 4693-4702, 2019.

[9] Y. Cui, L. Chapel, & S. Lefèvre, A subpath kernel for learning hierarchical image representations, *International Workshop on Graph-Based Representations in Pattern Recognition*, pp. 34-43, 2015.

[10] A.Tifrea, G. Bécigneul, & O. E. Ganea, Poincaré GloVe: Hyperbolic Word Embeddings, *International Conference on Learning Representations (ICLR)*, 2019.

[11] N. Monath, M. Zaheer, D. Silva, A. McCallum, & A. Ahmed, Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space, *International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 714-722, 2019.

	Geometry	Hierarchy Level	Latent Space Dimension d				
			8	16	32	64	128
VAE	E-VAE	Level 4	13.23 ± 0.7	17.62 ± 0.6	20.03 ± 0.2	19.09 ± 0.5	17.64 ± 0.5
		Level 3	19.56 ± 1.2	23.25 ± 0.6	25.38 ± 0.4	24.10 ± 0.3	22.52 ± 0.5
		Level 2	31.40 ± 0.7	35.34 ± 0.5	37.59 ± 0.4	37.39 ± 0.4	37.18 ± 0.2
	H-VAE	Level 4	10.05 ± 0.4	10.99 ± 0.5	10.67 ± 2.8	7.97 ± 1.1	7.05 ± 1.2
		Level 3	16.73 ± 0.4	17.43 ± 0.1	13.66 ± 4.5	11.54 ± 3.0	9.17 ± 2.6
		Level 2	29.32 ± 0.4	30.38 ± 1.8	24.02 ± 10.3	19.40 ± 7.9	16.82 ± 7.4
VAE+SLR	H-VAE	Level 4	11.99 ± 1.1	14.23 ± 0.6	15.21 ± 0.5	11.73 ± 1.3	10.06 ± 0.5
		Level 3	18.04 ± 1.3	20.76 ± 0.6	20.34 ± 0.6	17.21 ± 1.8	17.30 ± 0.6
		Level 2	29.81 ± 1.7	33.68 ± 0.8	35.93 ± 1.0	35.82 ± 1.9	35.15 ± 0.7

Table 1 – 1-NN classification accuracy of different VAE models on the PatternNet dataset (the higher the better) at different levels of the class hierarchy; level 4 represents the leaves of the hierarchy and thus the PatternNet classes. Results are averaged over 3 runs.

	Geometry	Latent Space Dimension				
		8	16	32	64	128
VAE	E-VAE	2.72 ± 0.01	2.72 ± 0.01	2.71 ± 0.01	2.71 ± 0.01	2.70 ± 0.01
	H-VAE	**2.71** ± 0.01	2.71 ± 0.01	2.82 ± 0.11	2.84 ± 0.10	2.87 ± 0.08
VAE+SLR	H-VAE	2.72 ± 0.01	**2.70** ± 0.01	**2.70** ± 0.02	**2.67** ± 0.02	**2.64** ± 0.01

Table 2 – Values of mAD on PatternNet dataset (the smaller the better). Results are averaged over 3 runs.

[12] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I.Oseledets, & V. Lempitsky, Hyperbolic image embeddings, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6418-6428, 2020.

[13] A. Dhall, A. Makarova, O. Ganea, D. Pavllo, M. Greeff, & A. Krause, Hierarchical image classification using entailment cone embeddings, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 836-837, 2020.

[14] Y. Liu, Y. Liu, C. Chen, & L. Ding, Remote-sensing image retrieval with tree-triplet-classification networks, *Neurocomputing*, vol. 405, pp. 48-61, 2020.

[15] K. Yu, S. Visweswaran, & K. Batmanghelich, Semi-supervised hierarchical drug embedding in hyperbolic space, *Journal of Chemical Information and Modeling*, vol. 60, pp. 5647-5657, 2020.

[16] D. P. Kingma, & M. Welling, Auto-Encoding Variational Bayes, *International Conference on Learning Representations (ICLR)*, 2014.

[17] G. Cheng, J. Han, & X. Lu, Remote sensing image scene classification: Benchmark and state of the art, *Proceedings of the IEEE*, vol. 105, pp. 1865-1883. 2017.

[18] W. Zhou, S. Newsam, C. Li, & Z. Shao, PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval, *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 197-209, 2018.

[19] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations (ICLR)*, 2016.

[20] X. Yu, X. Wu, C. Luo, & P. Ren, Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework, *GIScience & Remote Sensing* vol. 54, pp. 741-758, 2017.

[21] G. Cheng, X. Xie, J. Han, L. Guo and G. S. Xia, Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities, *IEEE JSTARS*, 13, pp.3735-3756, 2020.

[22] T. Yu, & C. De Sa, Numerically accurate hyperbolic embeddings using tiling-based models, *Neural Information Processing Systems (NeurIPS)*, pp. 2021–2031, 2019.