Hyperbolic Variational Auto-Encoder for Remote Sensing Embeddings

Manal Hamzaoui, Laetitia Chapel, Minh-Tan Pham, Sébastien Lefèvre

HYPERBOLIC VARIATIONAL AUTO-ENCODER FOR REMOTE SENSING SCENE EMBEDDINGS

Manal Hamzaoui, Laetitia Chapel, Minh-Tan Pham, Sébastien Lefèvre

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

ABSTRACT

The computer vision community is increasingly interested in exploring hyperbolic space for image representation, as hyperbolic approaches have demonstrated outstanding results in efficiently representing data with an underlying hierarchy. This interest arises from the intrinsic hierarchical nature among images. However, despite the hierarchical nature of remote sensing (RS) images, the investigation of hyperbolic spaces within the RS community has been relatively limited. The objective of this study is therefore to examine the relevance of hyperbolic embeddings of RS data, focusing on scene embedding. Using a Variational Auto-Encoder, we project the data into a hyperbolic latent space while ensuring numerical stability with a feature clipping technique. Experiments conducted on the NWPU-RESISC45 image dataset demonstrate the superiority of hyperbolic embeddings over the Euclidean counterparts in a classification task. Our study highlights the potential of operating in hyperbolic space as a promising approach for embedding RS data.

Index Terms— Remote sensing, scene embedding, underlying hierarchy, hyperbolic space, variational auto-encoder

1. INTRODUCTION

Traditionally, machine learning (ML) researches have focused mainly on approaches operating in Euclidean space, with less interest in other spaces, regardless of the nature of the data being manipulated and their specificities. This preference can be attributed to the convenient mathematical properties offered by Euclidean space, such as vectorial structures and closed forms for distance computations. However, it is worth noting that in many domains, real-world data does not have an *flat* structure but instead tends to exhibit a hierarchical structure. Recently, this trend has changed following the work of Nickel & Kiela [1], from which the authors propose to embed data with a latent hierarchy, particularly graphs, in hyperbolic space rather than Euclidean space. This publication marks a turning point in the ML community as it highlights the potential benefits of utilizing hyperbolic space for embedding hierarchical data structures or data with an underlying hierarchy. As such, several recent works have used this hyperbolic space to learn data representations in several applications such as word embedding, text classification, or image embedding [2]. Furthermore, various ML methods have been adapted to this space. Among them, we can mention hyperbolic Support Vector Machine (SVM) [3] or hyperbolic neural networks [4]. Other studies have provided a generalization of normal distributions on hyperbolic space that can be used to build and learn a probabilistic model like hyperbolic Variational Auto-Encoder (H-VAE) [5, 6].

Despite the popularity of hyperbolic space in the ML community, it has received rather limited attention in the remote sensing (RS) community, in spite of the hierarchical nature of RS data. The objective of this study is therefore to investigate the potential of hyperbolic embeddings in this context, in particular for scene images, while verifying whether the promises of the various works using hyperbolic space can be fulfilled. We rely on H-VAE which has been successfully used to embed data in hyperbolic space [5, 6] so that meaningful features can be extracted. H-VAEs are among the earliest studies dealing with images in a hyperbolic space. They were validated on the MNIST and Atari 2600 Breakout datasets by performing a classification step on the resulting embeddings, which showed that the H-VAE is able to better embed the data. Furthermore, despite the absence of an explicit hierarchy within these datasets, in particular MNIST dataset, a hierarchical structure was induced. This suggests that even better results can be anticipated for images that possess a genuine hierarchical arrangement, such as RS scene images.

2. H-VAE FOR REMOTE SENSING SCENE EMBEDDINGS

2.1. Motivations

RS scene images are naturally hierarchical. This is supported by a concept called the Gromov $\delta-$hyperbolicity, referred to as $\delta-$hyperbolicity for convenience, which enables us to measure the strength of the hierarchical information in a dataset. In practice, to quantify this information, we usually compute the scale-invariant metric δ_{rel} (the relative $\delta-$hyperbolicity)

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

which takes values in $[0, 1]$, the closer to zero the stronger the hierarchical information [7]. Furthermore, a low δ_{rel} value indicates that the data embedding space has an underlying hyperbolic geometry and that hyperbolic space would be suitable as an embedding space [2].

As we assume that RS data exhibit an underlying hierarchy, we adopt the procedure described in [7] and evaluate δ_{rel} for image scene embeddings of various RS scene datasets extracted by some reference CNNs pretrained on ImageNet.

Dataset	VGG16	ResNet18	GoogleNet	DenseNet
UCMerced	0.23	0.26	0.25	0.25
WHU-RS19	0.22	0.27	0.25	0.24
NWPU-RESISC45	0.23	0.28	0.24	0.25
AID	0.23	0.27	0.23	0.26
PatternNet	0.20	0.27	0.25	0.25

Table 1. The relative δ_{rel} values calculated for different RS image scene datasets. Results are averaged across 10 subsamples of size 1500. The standard deviation for all experiments does not exceed 0.03

Table 1 highlights the obtained δ_{rel} values for the five RS datasets including UCMerced, WHU-RS19, NWPU-RESISC45, AID and PatternNet [8]. We observe that the δ_{rel} values derived from these scene image datasets are closer to 0 than to 1 which results in a rather high degree of hyperbolicity, thus suggesting that hyperbolic space would be suited as an embedding space.

In this paper, we rely on H-VAE that has been successfully used to extract meaningful features from data with an underlying hierarchy and assess the quality of the embedding by performing a classification step.

2.2. Overall framework

Inspired from previous studies [5, 6], we adopt a hybrid architecture of the H-VAE in which the encoder and decoder networks are Euclidean networks and only the latent space of the VAE is hyperbolic. It is therefore necessary to extend the normal distribution to hyperbolic space. The wrapped normal distribution [6] is a generalization of this distribution to hyperbolic space. Furthermore, we add the Euclidean feature clipping technique [9] to avoid possible numerical instabilities of the model. We provide further details on feature clipping in the section below. The overall framework of the approach is illustrated in Figure 1.

2.3. Feature clipping

The majority of studies on hyperbolic space adopt hybrid "Euclidean-hyperbolic" architectures due to the generalization complexity of fundamental operations required for the extension of deep networks to this hyperbolic space. However, numerical issues leading to gradient vanishing often occur due to the passage between Euclidean and hyperbolic layers in these hybrid architectures. It was suggested in [9] that the Euclidean features should be clipped before moving to the hyperbolic layers, which allows to push the hyperbolic embeddings further away from the Poincaré ball boundary. The clipping technique allows hybrid architectures to cope with numerical problems, avoiding the vanishing gradient problem. Moreover, the behavior of hyperbolic networks becomes steadier, thus improving their performance. The feature clipping is defined as follows:

$$x_r^E = \min\left\{1, \frac{r}{||x^E||}\right\}.x^E, \tag{1}$$

where x_r^E is the clipped embedding of x^E which lies in the Euclidean space and r is the clipping value.

The VAE architecture has been extended to two hyperbolic models: the Poincaré Ball model [5] and the Lorentz model [6]. In our study, we consider the later one as it allows a better generalization of the normal distribution in the hyperbolic space [6]. The clipping here constrains the Euclidean embeddings which are in the origin tangent space to remain close to the origin in order to ensure the numerical stability of the Lorentz projection.

3. EXPERIMENTAL STUDY

This study aims to investigate whether hyperbolic space fulfills its promise in the context of RS and outperforms the Euclidean space. In this perspective, for both the E-VAE (Euclidean VAE) and the H-VAE, we adopt a very simple VAE architecture with regard to those used recently in the RS community [10]. We evaluate the quality of the resulting embeddings and the ability to discriminate between classes using a simple $1-$NN classifier.

3.1. Experimental setup

3.1.1. Dataset

The two models are learned on a subset of the NWPU-RESISC45 [11] RS scene dataset. All 45 classes are considered; for each of them, we randomly select 100 images for the training set, 50 images for the validation set and 80 images for the test set.

3.1.2. Implementation details

For both the E-VAE and the H-VAE, we adopt 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's last layer which is followed by a tanh activation. The input size of the encoder network is set to 64×64. The latent space dimension d of the embedding z is set to 8, 16, 32, 64 and 128, respectively.

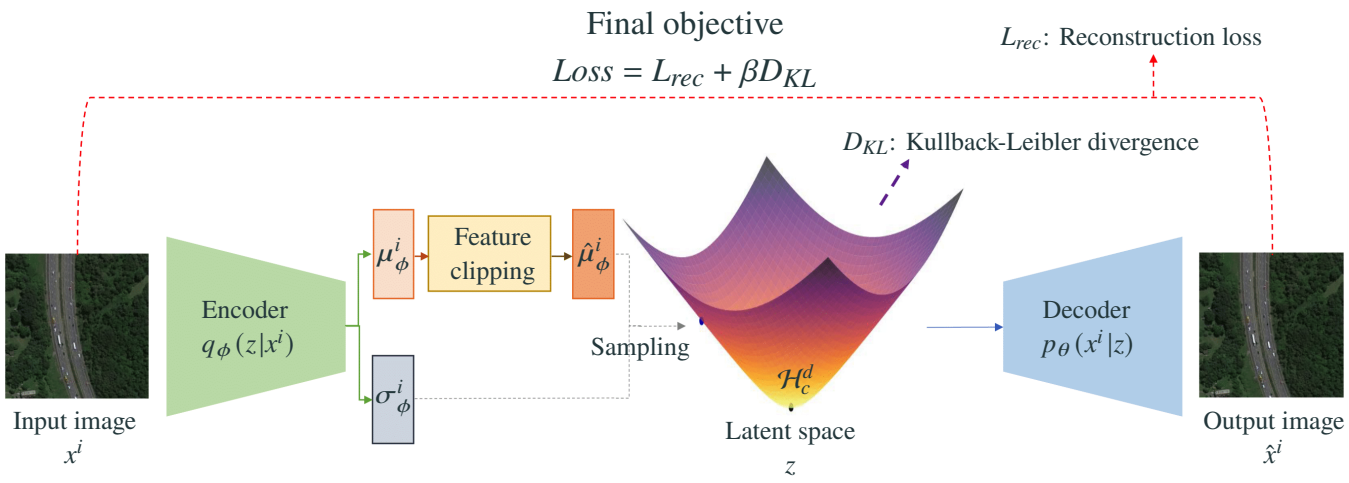

Fig. 1. Overview of the hyperbolic VAE for RS scene embeddings.

We use the Adam optimizer with a constant learning rate of $1e^{-3}$. The models are trained with mini-batches of size 64 for 1500 epochs with an early stopping at 50 epochs. The clipping hyper-parameter r is cross-validated.

3.2. Experimental Results

We compare our H-VAE with the E-VAE counterpart, as well as with the reference H-VAE model (without clipping). To do that, we evaluate the quality of the resulting embeddings of different VAEs and the ability to discriminate between classes using a simple $1-$NN classifier. Experiments are conducted on a subset of the NWPU-RESISC45 dataset and reported in Table 2, results are averaged over 3 runs. The reported scores correspond to models trained with hyper-parameters providing the best performance across different dimensions (clipping value $r = 1$).

Prior studies [5, 6] have demonstrated the superiority of H-VAE *w.r.t.* its Euclidean counterpart in various context (images and graphs). However, this observation does not hold in our RS context. We further investigate this behavior and we show that it is due to the numerical problems arising from hyperbolic projection operations that result in out-of-space embeddings. Our H-VAE, which uses the feature clipping technique, is therefore considerably steadier, allowing better representations to be learned and, consequently, outperforming the E-VAE. We also note that the low classification accuracy is obtained for all models due to the choice of the VAE architecture. RS data are complex and require very deep networks with a large amount of data to reach high performances. This was not used in this study since the focus was on comparing hyperbolic and Euclidean spaces rather than achieving the best results.

Impact of the clipping value

Figure 2 shows the $1-$NN classification accuracy values on the test set in function of the clipping value r.

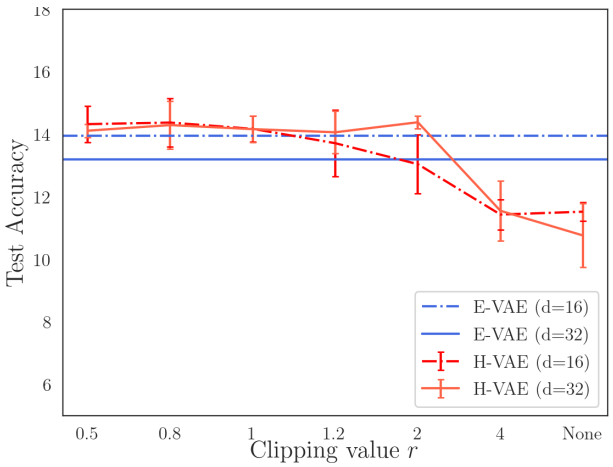

Fig. 2. 1-NN classification accuracy of different VAE models on a subset of the NWPU-RESISC45 RS scene dataset w.r.t. the clipping value r.

We observe that the H-VAE generally performs better with small values of the clipping hyper-parameter ($r < 1.2$). Larger clipping values often result in Euclidean tangent features far from the space origin. Nonetheless, in this scenario, performing such an operation necessitates a remarkably high floating point precision (i.e., a considerable number of bits) to adequately represent the resulting embeddings in the Lorentz model. This, however, is not feasible in PyTorch, as double precision is the highest floating-point number available, occupying 64 bits. The possibility of out-of-space embeddings therefore increases, leading to numerical instability of

Space	Clip r	Metric	Latent Space Dimension d				
			8	16	32	64	128
E-VAE	/	Overall acc	12.00 ± 0.15	13.96 ± 0.56	13.21 ± 0.21	12.08 ± 0.10	12.39 ± 0.32
		L3-acc	18.38 ± 0.42	20.64 ± 0.41	19.33 ± 0.13	17.93 ± 0.57	17.88 ± 0.24
		L2-acc	28.34 ± 0.57	31.49 ± 0.38	30.97 ± 0.16	29.96 ± 0.47	29.95 ± 0.49
H-VAE	None	Overall acc	11.38 ± 0.58	11.53 ± 0.30	10.77 ± 1.02	10.21 ± 0.96	11.33 ± 0.53
		L3-acc	17.83 ± 0.58	17.45 ± 0.45	15.90 ± 0.66	14.82 ± 0.90	16.43 ± 1.02
		L2-acc	28.58 ± 0.68	28.46 ± 0.33	27.85 ± 0.83	29.96 ± 1.71	$29,77 \pm 0,90$
	1	Overall acc	$\mathbf{12.36 \pm 0.53}$	$\mathbf{14.18 \pm 0.42}$	$\mathbf{14.17 \pm 0.42}$	$\mathbf{14.18 \pm 0.44}$	$\mathbf{12.87 \pm 0.54}$
		L3-acc	$\mathbf{18.80 \pm 0.61}$	$\mathbf{20.89 \pm 0.59}$	$\mathbf{20.50 \pm 0.22}$	$\mathbf{20.00 \pm 0.71}$	$\mathbf{18.39 \pm 0.82}$
		L2-acc	$\mathbf{28.46 \pm 0.40}$	$\mathbf{31.54 \pm 0.67}$	$\mathbf{31.91 \pm 0.43}$	$\mathbf{31.66 \pm 0.78}$	$\mathbf{30.11 \pm 0.73}$

Table 2. 1-NN classification results computed on the test set of the NWPU-RESISC45 dataset at different levels of the class hierarchy: overall acc represents the classification accuracy at the leaves (level 4) and thus the NWPU-RESISC45 classes; L3-acc and L2-acc give the accuracy at level 3 and level 2, respectively (the higher the better); Results are averaged over 3 runs.

the network, which is reflected by the significant decrease of classification scores across dimensions.

4. CONCLUSION

In this study, we explore the potential of hyperbolic space for embedding RS scene images, which typically exhibit an underlying hierarchical structure. We showed that our hyperbolic VAE better encodes the scene images, yielding improved latent space organization and superior performance compared to the Euclidean VAE. However, achieving this improvement requires careful tuning of the clipping value when learning the hyperbolic networks. As future work, we intend to investigate the impact of hyperbolic curvature as well as the Poincaré Ball model. Additionally, we plan to consider the use of a more dedicated loss function for hyperbolic space in order to fully take advantage of it ability to handle structured data such as RS images.

5. REFERENCES

[1] M. Nickel and D. Kiela, "Poincaré Embeddings for Learning Hierarchical Representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 6338–6347.

[2] W. Peng, T. Varanka, A. Mostafa, H. Shi, and G. Zhao, "Hyperbolic Deep Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 10023–10044, 2022.

[3] Hyunghoon Cho, Benjamin Demeo, Jian Peng, and Bonnie Berger, "Large-Margin Classification in Hyperbolic Space," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019, pp. 1832–1840.

[4] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada, "Hyperbolic Neural Networks++," in *International Conference on Learning Representations*, 2021.

[5] E. Mathieu, C. Le Lan, C.J. Maddison, R. Tomioka, and Y.W. Teh, "Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders," in *Advances in Neural Information Processing Systems*, 2019, pp. 12544–12555.

[6] Y. Nagano, S. Yamaguchi, Y. Fujita, and M. Koyama, "A wrapped normal distribution on hyperbolic space for gradient-based learning," in *International Conference on Machine Learning*, 2019, pp. 4693–4702.

[7] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I.V. Oseledets, and V.S. Lempitsky, "Hyperbolic Image Embeddings," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6417–6427.

[8] Suparna Dutta and Monidipa Das, "Remote sensing scene classification under scarcity of labelled samples - A survey of the state-of-the-arts," *Computers and Geosciences*, vol. 171, pp. 105295, 2023.

[9] Y. Guo, X. Wang, Y. Chen, and S.X. Yu, "Clipped Hyperbolic Classifiers Are Super-Hyperbolic Classifiers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1–10.

[10] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[11] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proceedings of the IEEE*, vol. 105, pp. 1865–1883, 2017.