Sparse Hyperbolic Convolutional Networks with Enhanced Object Localization via GradCAM Analysis

Anonymous ICCV submission

Paper ID *****

Abstract

Hyperbolic spaces model hierarchical structures within data. Studies have demonstrated that spatial representations in the hippocampus are structured within hyperbolic spaces to optimize efficiency[17]. We explore the use of hyperbolic convolutional networks with sparsity constraints (L1 and Top-k) and analyze the significance of features in the images for classification tasks using GradCAM. We show that applying sparsity constraints to hyperbolic convolutional networks yields performance comparable to established benchmarks and results in greater interpretability. This work develops sparse hyperbolic representations, enhancing interpretability in AI systems.

1. Introduction

Deep convolutional neural networks have revolutionized computer vision by learning hierarchical feature representations that capture complex visual patterns. However, traditional CNNs operate exclusively in Euclidean space, fundamentally limiting their ability to model the inherent hierarchical structures present in visual data [6, 12]. Real-world images exhibit rich hierarchical organizations—from fine-grained textures to object parts, from parts to complete objects, and from objects to complex scenes—that would benefit from geometric spaces designed to naturally accommodate such tree-like structures. Recent neuroscience experiments reveal that spatial representations in CA1 hippocampal neurons of rats organize within hyperbolic spaces, enabling efficient coding that dynamically expands over time[17].

Hyperbolic geometry, characterized by constant negative curvature, offers a compelling alternative to Euclidean representations. Unlike flat Euclidean space, hyperbolic space exhibits exponential volume growth, making it particularly well-suited for embedding hierarchical data with minimal distortion. Recent advances in hyperbolic neural networks have demonstrated significant improvements in tasks involving hierarchical data, such as knowledge graphs and social networks [4]. However, the application of hyperbolic geometry to standard computer vision tasks remains largely underexplored, with most existing work focusing on specialized domains or requiring architectural constraints that limit practical applicability.

A critical challenge in modern deep learning is interpretability. Experimental evidence in neurosciecne suggests that the energy budget tries to drive the brain towards energy efficient neural codes and wiring patterns resulting in sparse codes[1]. Sparsity mechanisms offer a promising solution by selectively retaining only the most informative features while eliminating redundant parameters [9]. Two primary approaches have emerged: L1 regularization, which naturally induces sparsity through geometric properties of the L1 norm, and Top-K selection, which provides direct control over sparsity levels by retaining only the most significant activations [8]. While these techniques have been extensively studied in Euclidean neural networks, their application to hyperbolic architectures remains unexplored.

Understanding and interpreting the decision-making processes of deep neural networks has become increasingly important as these models are deployed in critical applications. Gradient-weighted Class Activation Mapping (GradCAM) has emerged as a powerful tool for providing visual explanations by highlighting regions in input images that contribute most significantly to model predictions [14]. However, existing interpretability methods are designed exclusively for Euclidean networks and do not account for the unique geometric properties and constraints of hyperbolic space. This limitation prevents us from understanding how hyperbolic networks make decisions and whether their purported advantages in hierarchical modeling translate to improved attention mechanisms in computer vision tasks.

Our Contributions. In this work, we address these limitations by introducing the first comprehensive framework for sparse hyperbolic convolutional neural networks with enhanced interpretability. Our key contributions are:

1. **Sparse Hyperbolic CNNs:** We present novel implementations of L1 regularization and Top-K sparsity

mechanisms specifically designed for hyperbolic convolutional neural networks operating in the Lorentz model which act on the activations making the activations sparser. Our approach maintains the geometric constraints of hyperbolic space while achieving sparsification.

2. **Hyperbolic GradCAM:** We extend gradient-weighted class activation mapping to work with hyperbolic neural networks by decomposing gradients and activations into temporal and spatial components that respect the underlying Lorentzian geometry. This enables visual interpretation of sparse hyperbolic network decisions for the first time.
3. **Comprehensive Comparative Analysis:** We provide the first systematic comparison between sparse Euclidean ResNet architectures and their hyperbolic counterparts using both traditional performance metrics and visual explanation analysis. Our experiments on CIFAR-10 demonstrate that sparse hyperbolic networks consistently achieve superior object localization compared to their Euclidean equivalents.

Our experimental results on CIFAR-10 demonstrate that hyperbolic CNNs with both L1 and Top-K sparsity constraints outperform their Euclidean counterparts in terms of object localization quality, as evidenced by GradCAM visualizations that show more precise and semantically meaningful attention patterns. The sparse hyperbolic networks maintain competitive classification accuracy while requiring significantly fewer computational resources, making them particularly attractive for resource-constrained applications.

Broader Impact. This work opens new avenues for research at the intersection of non-Euclidean geometry, sparse neural networks, and interpretable AI. By demonstrating that hyperbolic geometry can enhance both performance and interpretability in computer vision tasks, we provide a foundation for developing more efficient and explainable deep learning systems. The improved object localization capabilities revealed through our GradCAM analysis suggest that hyperbolic networks may be particularly valuable for applications requiring precise spatial understanding, such as medical imaging, autonomous navigation, and scene understanding.

The rest of the paper is organized as follows: Section 2 provides essential background on hyperbolic geometry and the theoretical foundations underlying our approach. Section 3 reviews related work in hyperbolic neural networks, sparsity mechanisms, and visual explanation methods. Section 4 details our methodology for implementing sparse hyperbolic CNNs and extending GradCAM to hyperbolic space. Section 5 presents comprehensive experimental results comparing sparse hyperbolic and Euclidean networks on CIFAR-10, and Section 6 concludes with discussions of implications and future directions.

2. Background

This section outlines the key theoretical foundations underlying our work: hyperbolic geometry and its relevance for deep learning, hyperbolic convolutional neural networks (HCNNs), sparsity mechanisms in neural representations, and gradient-based visual explanation methods. Together, these components motivate and enable the design of interpretable and efficient hyperbolic models for visual recognition tasks.

2.1. Hyperbolic Geometry for Deep Learning

Hyperbolic geometry is a non-Euclidean space of constant negative curvature, offering a natural inductive bias for representing hierarchical and tree-like structures often found in linguistic and visual data [12?]. A distinguishing property of hyperbolic space is its exponential volume growth with radius, which contrasts with the polynomial growth of Euclidean space, enabling compact embeddings of hierarchical data.

Lorentz Model. We adopt the Lorentz (or hyperboloid) model for its numerical stability in optimization and compatibility with Riemannian geometry toolkits [6, 10]. The d-dimensional hyperbolic space $\mathbb{H}^d$ is realized as:

$$\mathbb{H}^d = \left\{ x \in \mathbb{R}^{d+1} : \langle x, x \rangle_L = -1,\ x_0 > 0 \right\} \tag{1}$$

where the Lorentzian inner product is defined as:

$$\langle x, y \rangle_L = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i \tag{2}$$

Key operations include the exponential map $\exp_x^L : T_x\mathbb{H}^d \to \mathbb{H}^d$ and logarithmic map $\log_x^L : \mathbb{H}^d \to T_x\mathbb{H}^d$, which bridge the manifold and its tangent space:

$$\exp_x^L(v) = \cosh(\|v\|_L)x + \sinh(\|v\|_L)\frac{v}{\|v\|_L} \tag{3}$$

$$\log_x^L(y) = d_L(x, y) \cdot \frac{y + \langle x, y \rangle_L x}{\|y + \langle x, y \rangle_L x\|_L} \tag{4}$$

where $d_L(x, y) = \operatorname{arccosh}(-\langle x, y \rangle_L)$ is the Lorentzian geodesic distance.

2.2. Hyperbolic Convolutional Neural Networks

While standard convolutional neural networks (CNNs) operate in Euclidean space, their representational capacity is limited when modeling inherently hierarchical visual structures. Hyperbolic CNNs extend standard convolutions to curved spaces by operating in tangent spaces via Riemannian mappings [4, 15].

A typical hyperbolic convolution consists of three stages:

$$\tilde{f}(y_i) = \log_x^L(f(y_i)) \quad \text{(Project features to tangent space)} \tag{5}$$

$$\tilde{g}(x) = \sum_i k_i \tilde{f}(y_i) \quad \text{(Euclidean-like convolution)} \tag{6}$$

$$g(x) = \exp_x^L(\tilde{g}(x)) \quad \text{(Map back to manifold)} \tag{7}$$

For computational efficiency, they adopt a linearized kernel formulation by expressing 2D convolution as:

$$\text{LConv2d}(x) = \text{LFC}(\text{Unfold}(x)) \tag{8}$$

where Unfold extracts spatial patches and LFC denotes Lorentz fully connected operations. Temporal components are handled via a rescaling procedure:

$$x_{\text{time}}^{\text{rescaled}} = \sqrt{\sum x_{\text{time}}^2 - (k_{\text{len}} - 1) \cdot \kappa} \tag{9}$$

To maintain numerical stability and preserve the manifold geometry, batch normalization is performed in the tangent space. Given input x, we compute the Fréchet mean μ and perform:

$$x_T = \log_\mu^L(x) \tag{10}$$

$$\hat{x}_T = \gamma \frac{x_T - \mu_T}{\sqrt{\sigma_T^2 + \epsilon}} + \beta \tag{11}$$

$$\hat{x} = \exp_\mu^L(\hat{x}_T) \tag{12}$$

Here, μ_T and σ_T^2 are the mean and variance in the tangent space, and γ, β are learnable affine parameters.

Finally, classification is performed using hyperbolic hyperplanes defined in Lorentz space. For each class c with parameters (a_c, z_c), the class logit is computed as:

$$w_{t,c} = \sinh(\sqrt{\kappa^{-1}} a_c) \|z_c\| \tag{13}$$

$$w_{s,c} = \cosh(\sqrt{\kappa^{-1}} a_c) z_c \tag{14}$$

$$\text{logit}_c = -\langle w_c, x \rangle_L \tag{15}$$

2.3. Gradient-weighted Class Activation Mapping (GradCAM)

GradCAM [14] is a widely used technique for visual model explanation. It highlights input regions that most influence a model's prediction for a specific class c, based on gradient information. Given a feature map A^k and the gradient of the output score y^c with respect to A^k, the class-specific importance weight is computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A_{ij}^k} \tag{16}$$

The GradCAM localization map is then given by:

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{17}$$

In our work, we generalize GradCAM to hyperbolic settings by accounting for curvature and the temporal-spatial decomposition inherent in Lorentzian embeddings. This allows us to evaluate the interpretability of sparse hyperbolic networks through visual explanations that respect the geometry of the representation space.

3. Related Work

Our work lies at the intersection of hyperbolic geometry in vision, sparse neural networks, and interpretability techniques. We briefly review the most relevant contributions across these domains.

3.1. Hyperbolic Geometry in Computer Vision

Hyperbolic geometry has shown promise in computer vision due to its exponential volume growth and capacity to model hierarchies [11, 12]. Chami et al. [4] demonstrated hyperbolic graph neural networks preserve hierarchical information better than Euclidean counterparts.

Building on these insights, Schwethelm et al. [2] proposed HCNN, a fully Lorentzian convolutional network capable of hyperbolic batch normalization and classification. Earlier efforts, such as Das et al. [16], used expansive convolutions in the Poincaré disk for theoretical generalization guarantees. However, these works focus on dense architectures and do not explore sparsity or interpretability.

3.2. Sparsity in Neural Networks

Sparsity improves both efficiency and interpretability. L1 regularization promotes sparsity by penalizing the L_1 norm of activations [7, 9], while also enhancing disentanglement [13]. In contrast, Top-K selection methods such as Top-KAST [8] enforce fixed-ratio sparsity during training and inference without gradient masking.

Although these techniques are well-studied in Euclidean settings, their adaptation to non-Euclidean spaces—especially in the Lorentz model—remains largely unexplored. Our work bridges this gap by introducing both L1 and Top-K sparsity in hyperbolic CNNs.

3.3. Visual Explanation Techniques

GradCAM [14] and its variants [5] are widely used to visualize CNN decision processes by highlighting class-relevant regions. These methods, however, are restricted to Euclidean activations.

While a few hybrid approaches have explored combining GradCAM with techniques like LRP [3], no existing work extends GradCAM to hyperbolic networks. We propose

Hyperbolic GradCAM to fill this gap, enabling manifold-aware interpretation of sparse Lorentz-based models.

4. Methods

Building on Prior Work. Leveraging the Lorentz model's stability and the effectiveness of fully hyperbolic convolutional architectures [2, 4, 6, 10], we adopt this foundation to construct our hyperbolic networks. Our contributions extend this line of work by introducing sparsity-driven mechanisms for disentanglement and interpretability in hyperbolic space, along with a novel adaptation of GradCAM tailored to the Lorentzian geometry.

4.1. Sparsity-Induced Interpretable Representations in Hyperbolic Networks

To promote interpretability in hyperbolic space, we introduce sparsity into our model via two mechanisms: L1 regularization and Top-K activation masking. Sparse representations have been shown to improve interpretability and generalization [7, 9, 13], and we adapt these principles to the Lorentzian manifold.

L1 Regularization in Tangent Space. Given hyperbolic activations $h \in \mathbb{H}^d$, we encourage sparsity by applying an L1 penalty to their tangent-space projections:

$$\mathcal{L}_{\text{sparse}} = \mathcal{L}_{\text{task}} + \lambda \| \log_{\mathbf{0}}^{L}(h) \|_1 \tag{18}$$

At inference, we apply soft thresholding to enforce sparsity explicitly:

$$h_{\text{sparse}} = \exp_{\mathbf{0}}^{L} \left(\text{SoftThreshold}(\log_{\mathbf{0}}^{L}(h), \tau) \right) \tag{19}$$

Top-K Activation Masking. To impose structured sparsity, we also experiment with forwarding only the top-k tangent activations where we select $\rho\%$ of activations from the total number of activations.

$$k = \lfloor \rho \cdot n \rfloor, \quad \text{TopK}_{\rho}(x)_i = \begin{cases} x_i & \text{if } |x_i| \text{ in top-}k \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

$$h_{\text{topk}} = \exp_{\mathbf{0}}^{L} \left(\text{TopK}_{\rho}(\log_{\mathbf{0}}^{L}(h)) \right) \tag{21}$$

Gradients are propagated through the discrete Top-K operation via straight-through estimation:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial \text{TopK}(x)} \cdot \mathbb{I}_{\text{selected}} \tag{22}$$

By sparsifying hyperbolic representations, we aim to get interpretable features, reduce redundancy, and better understand how different geometric components contribute to model predictions.

4.2. Hyperbolic GradCAM for Visual Explanation

To evaluate the interpretability benefits of sparsity in hyperbolic neural networks, particularly for vision tasks, we extend the well-established GradCAM technique [14] to the Lorentzian setting. Our proposed *Hyperbolic GradCAM* respects the manifold structure and disentangles spatial and temporal contributions to enable geometry-aware visualizations.

Temporal-Spatial Decomposition. Given hyperbolic activations $A \in \mathbb{R}^{H \times W \times C}$ and gradients $G \in \mathbb{R}^{H \times W \times C}$ in Lorentz space (with $C \geq 2$) where H, W, C refer to the height, width and number of channels of the outputs of the filters, we decompose each into temporal and spatial components:

$$A_{\text{time}} = A[:, :, 0], \quad A_{\text{space}} = A[:, :, 1:] \tag{23}$$

$$G_{\text{time}} = G[:, :, 0], \quad G_{\text{space}} = G[:, :, 1:] \tag{24}$$

Curvature-Aware Importance Scoring. We compute class-discriminative importance by combining curvature-scaled temporal correlation and spatial alignment:

$$I_{\text{time}} = |G_{\text{time}} \cdot A_{\text{time}}| \cdot (1 + 0.1\kappa) \tag{25}$$

$$I_{\text{space}} = \|G_{\text{space}}\|_2 \cdot \|A_{\text{space}}\|_2 \tag{26}$$

$$\text{HypGradCAM} = \alpha I_{\text{time}} + \beta I_{\text{space}} \tag{27}$$

The weights (α, β) are adjusted by layer depth to reflect the increasing semantic abstraction of deeper layers:

$$(\alpha, \beta) = \begin{cases} (0.05, 1.0) & \text{shallow layers} \\ (0.1, 1.0) & \text{intermediate layers} \\ (0.15, 0.9) & \text{deep layers} \end{cases} \tag{28}$$

Sparsity-Aware Emphasis. To maintain visual clarity when sparse activation constraints are imposed, we enhance the spatial importance map:

$$\xi_{\text{spatial}}^{\text{sparse}} = \xi_{\text{spatial}} \cdot (1 + 0.2(1 - \rho)) \tag{29}$$

This modulation compensates for reduced activation spread and ensures that salient features remain visible under strong sparsity levels.

By integrating Hyperbolic GradCAM with our sparsity mechanisms, we are able to visualize how disentangled features emerge in the hyperbolic representation space and assess their contribution to model decisions.

5. Results

In this section, we comprehensively evaluate the impact of sparse activation mechanisms on hyperbolic neural networks. Our analysis proceeds along two main dimensions: (i) quantitative performance, where we measure top-1 classification accuracy across different architectural variants, and (ii) interpretability, where we assess model behavior using adapted visual explanation techniques such as Hyperbolic GradCAM.

Due to computational limitations, our experiments primarily utilize the ResNet-18 backbone and are assessed on the CIFAR-10 and CIFAR-100 benchmark datasets. We explore Euclidean, fully hyperbolic (Lorentzian), and hybrid architectures, incorporating sparsity via L1 regularization or Top-K activation masking. These evaluations are designed to elucidate not only the performance trade-offs associated with sparsity in hyperbolic networks, but also its impact on the interpretability and structure of the learned representations.

5.1. Quantitative Performance Evaluation on CIFAR-10 and CIFAR-100

We evaluate the performance of Euclidean, Lorentzian (fully hyperbolic), and hybrid architectures with and without sparsity mechanisms on CIFAR-10 and CIFAR-100 datasets. Table 1 reports Top-1 accuracy (%) for each variant. Sparsity is introduced using L1 regularization or Top-K masking, and the hybrid model follows the configuration described in [15] where blocks with high hyperbolicity (e.g., 1 and 3) are replaced with Lorentz blocks while others remain Euclidean.

Despite the imposition of strong sparsity constraints—through L1 regularization or Top-K masking—our models maintain competitive or even superior accuracy compared to their dense counterparts. For instance, the Euclidean model with Top-K sparsity at $\rho = 0.01$ achieves a Top-1 accuracy of 95.79% on CIFAR-10, surpassing the dense baseline. Similarly, both Lorentzian and hybrid architectures exhibit strong robustness to sparsification, particularly on CIFAR-100. These results demonstrate that hyperbolic geometry facilitates compact, expressive representations, with sparsity introducing negligible performance degradation while providing greater interpretability as shown in subsection 5.2.

From a neuroscientific perspective, sparse representations are considered a hallmark of efficient information encoding in the brain. In particular, early visual cortex (V1) has been shown to operate with overcomplete, sparse codes to maximize information content while minimizing metabolic cost [13]. Sparse activations also contribute to disentangling latent factors, reducing interference between features, and enhancing generalization [7, 9]. The resilience of our sparse models thus aligns with the biological principle that efficient perception arises not from exhaustive activation, but from selective, high-precision responses.

These results motivate deeper investigation into the interpretability and semantic structure of sparse hyperbolic representations. In the following section, we employ Hyperbolic GradCAM to visualize how sparsity shapes the geometry of class-relevant features and enhances our ability to interpret model predictions.

5.2. Hyperbolic GradCAM analysis

To assess the qualitative interpretability benefits of hyperbolic models, we visualize the GradCAM heatmaps generated from Euclidean and fully hyperbolic CNNs. Figure 1 shows side-by-side comparisons on the same input image. We observe that while the Euclidean GradCAM tends to produce broader, often diffused attention regions that may highlight irrelevant background areas, the Hyperbolic GradCAM yields sharper, spatially localized, and semantically focused activations, concentrating more effectively on the discriminative regions (e.g., the contours and head of the frog).

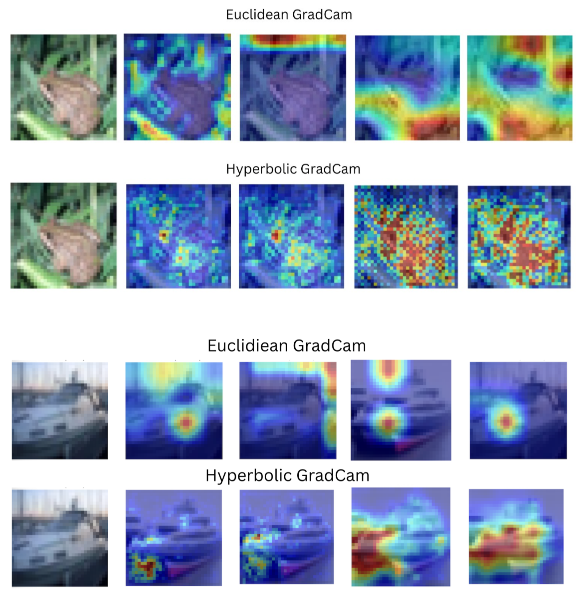

Figure 1. Comparison of GradCAM visualizations between standard Euclidean CNNs (top row) and fully hyperbolic CNNs (bottom row). The hyperbolic variant focuses more sharply on the object of interest, yielding more interpretable and compact saliency maps. Additional examples are shown below.

We hypothesize that this difference stems from the hyperbolic model's intrinsic capacity to encode hierarchical relations. Instead of merely identifying low-level discriminative patterns, the hyperbolic geometry allows the network to learn global structural cues those that define what makes an object a "frog" in a taxonomic or conceptual sense, beyond superficial texture or contrast differences. This aligns with the theory that hyperbolic spaces are better suited to represent hierarchical or tree-like data structures [6, 12]. Such behavior hints at a shift from learning purely class-discriminative saliency to capturing conceptual part-whole semantics, which may offer more cognitively aligned interpretations of the model's decision process.

ICCV
#*****

Domain	Variant	CIFAR-10	CIFAR-100
Euclidean	Baseline (ResNet-18)	95.14	**77.93**
	+ L1 Sparse (all layers)	95.46	77.85
	+ Top-K Sparse ($\rho = 0.1$)	95.19	77.35
	+ Top-K Sparse ($\rho = 0.01$)	**95.79**	77.91
Lorentz	Baseline (Hyp-ResNet19)	95.20	8.00
	+ L1 Sparse	94.97	77.41
	+ Top-K Sparse ($\rho = 0.1$)	95.17	**78.14**
	+ Top-K Sparse ($\rho = 0.01$)	95.17	77.94
Hybrid	Baseline (Hybrid ResNet)	95.24	**78.24**
	+ L1 Sparse	**95.36**	77.93
	+ Top-K Sparse ($\rho = 0.1$)	95.32	77.75
	+ Top-K Sparse ($\rho = 0.01$)	95.26	77.98

Table 1. Top-1 accuracy (%) on CIFAR-10 and CIFAR-100 across Euclidean, Lorentzian, and Hybrid variants with different sparsity mechanisms. Top-K sparsity at $\rho = 0.01$ achieves the best performance in Euclidean settings, while hybrid and Lorentzian models show strong results on CIFAR-100.

5.3. Analysis of activation sparsity in Hyperbolic CNN

The visualizations in Figure 1 demonstrate that hyperbolic neural networks inherently exhibit more localized and semantically aligned attention compared to Euclidean CNN. Building on this geometric advantage, we now investigate whether explicitly enforcing activation sparsity can further sharpen these representations. Our goal is to examine whether sparse activations encourage the network to focus on the most critical, high-salience features, thereby enhancing interpretability without compromising performance.

This line of inquiry is grounded in the hypothesis that activation sparsity can act as a form of structural inductive bias, promoting disentanglement in the latent space and improving the selectivity of GradCAM attributions. In doing so, we aim to bridge architectural expressiveness (via hyperbolic geometry) with functional parsimony (via sparsity), both of which are known to contribute to interpretable representations in biological systems [7, 13].

Figure 2 demonstrates the qualitative effects of applying sparsity to hyperbolic CNNs via L1 and Top-k activation constraints. Across all configurations, we observe a consistent sharpening of GradCAM heatmaps as sparsity increases. Specifically:

- **L1 Sparse Hyperbolic GradCAM** shows moderately focused attention with denoised activations that remain semantically relevant and follow object contours.
- **Top-k Sparse** variants highlight salient object regions more aggressively, producing concentrated and interpretable maps.
- **Harder Top-k** (with lower ρ) further localizes attention to core features, although occasionally at the cost of contextual cues.

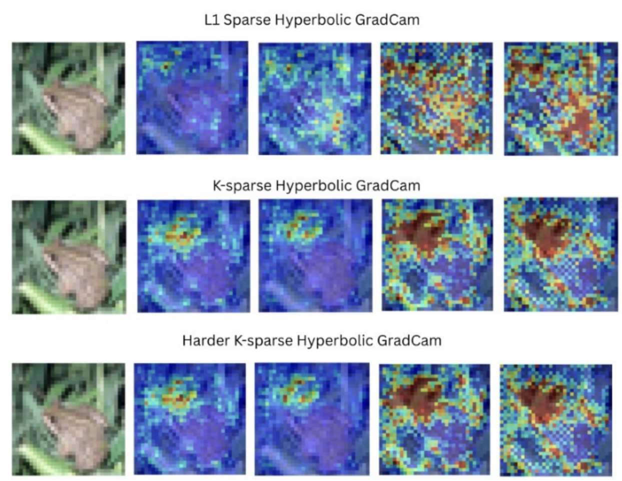

Figure 2. GradCAM visualizations for hyperbolic CNNs with different activation sparsity mechanisms. From top to bottom: L1 sparse, Top-k sparse ($\rho = 0.1$), and harder Top-k sparse ($\rho = 0.01$). Each row shows the original image followed by activation maps from successive layers.

These results align with our hypothesis that sparsity aids in feature selection by suppressing irrelevant activations and enhancing signal-to-noise ratio in geometric representations. In particular, hyperbolic networks benefit from this effect by leveraging their natural hierarchy-preserving structure to amplify semantically meaningful activations.

Sparse activation in hyperbolic neural networks improves the interpretability of internal representations without degrading performance. The resulting GradCAM maps are not only visually sharper but also better aligned with object boundaries and key discriminative regions. This supports the use of sparsity as a cognitively inspired prior and reinforces its potential to yield more explainable and structured feature learning in non-Euclidean spaces.

5.4. Quantitative metrics for GradCam analysis

To better understand the interpretability benefits of sparse activation mechanisms, we evaluate GradCAM-based visual explanations using five key metrics: **Robustness**, **Faithfulness**, **Localization**, **Complexity**, and **Interpretability**. *Robustness* measures the stability of the saliency maps under perturbations, where higher values imply more consistent explanations. *Faithfulness* quantifies how well the saliency map aligns with the model's true decision-making process (e.g., via input occlusion). *Localization* evaluates the sharpness and spatial concentration of salient regions, indicating how focused the explanations are. *Complexity*, in contrast, is minimized; more negative values denote simpler and less noisy saliency maps. Finally, *Interpretability* is an aggregate score indicating how comprehensible the explanations are to humans, combining fidelity and sparsity-based heuristics.

From Table 2, it is evident that sparse variants, especially the **L1 Sparse** model, outperform the standard hyperbolic network across most metrics. It achieves the highest **Robustness**, **Faithfulness**, and **Interpretability**, while also having the lowest (i.e., best) **Complexity**. Interestingly, both **Top-0.1%** and **Top-0.01%** sparsity levels exhibit superior **Localization** scores compared to the baseline, suggesting sharper and more spatially focused attention maps.

These results provide compelling evidence that sparse hyperbolic networks not only preserve but often enhance interpretability across multiple axes. This underscores a strong case for further investigating sparse activation mechanisms—not merely as regularization tools, but as principled methods for improving model transparency and alignment with cognitively relevant priors.

6. Conclusion and discussion

We introduce Hyperbolic GradCAM, a novel interpretability framework that extends gradient-based visual explanations to hyperbolic convolutional networks. By incorporating Lorentzian geometric structure and disentangling spatiotemporal components, this approach enables—for the first time—geometrically principled visualizations of hyperbolic models.

Complementing this, we explore sparse hyperbolic CNNs using L1 regularization and Top-K activation masking. These models achieve classification performance comparable to both Euclidean and fully hyperbolic baselines. Crucially, our Hyperbolic GradCAM analysis reveals that sparse hyperbolic networks yield enhanced interpretability, producing sharper and more semantically meaningful attention maps.

Our findings highlight that hyperbolic representations—especially when combined with sparse activations—can lead to more expressive and interpretable models, bridging the gap between powerful geometric modeling and human-aligned understanding. As a promising direction for future work, we aim to investigate how sparsity may promote disentanglement in hyperbolic feature spaces and whether this contributes to the emergence of more structured and semantically aligned representations. Such insights could open up new possibilities for principled feature-level explanations in non-Euclidean deep learning systems.

Table 2. GradCAM evaluation metrics for layer 15 across standard and sparse hyperbolic networks trained on CIFAR-100. Bold values indicate best performance per metric (excluding complexity, where lower is better).

Model	Robustness ↑	Faithfulness ↑	Localization ↑	Complexity ↓	Interpretability ↑
Standard	0.556 ± 0.173	0.148 ± 0.096	0.062 ± 0.046	-16.952 ± 5.278	0.682 ± 0.025
L1 Sparse	**0.702 ± 0.134**	**0.233 ± 0.068**	0.063 ± 0.032	**-17.807 ± 2.113**	**0.699 ± 0.023**
Top-0.1% Sparse	0.699 ± 0.154	0.140 ± 0.102	**0.066 ± 0.037**	-16.262 ± 5.344	0.664 ± 0.047
Top-0.01% Sparse	0.694 ± 0.158	0.140 ± 0.102	**0.066 ± 0.037**	-16.262 ± 5.344	0.664 ± 0.047

References

[1] David Attwell and Simon B. Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21:1133 – 1145, 2001. 1

[2] Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. *arXiv preprint arXiv:2303.15919*, 2023. 3, 4

[3] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016. 3

[4] Ines Chami, Aditya Wolf, Frederic Sala, Sujith Ravi, and Christopher Ré. Hyperbolic graph convolutional neural networks. *NeurIPS*, 2019. 1, 2, 3, 4

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 3

[6] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *NeurIPS*, 2018. 1, 2, 4, 5

[7] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011. 3, 4, 5, 6

[8] Siddhant Jayakumar, Razvan Pascanu, Jack Rae, Simon Osindero, and Erich Elsen. Top-kast: Top-k always sparse training. *Advances in Neural Information Processing Systems*, 33:20744–20754, 2020. 1, 3

[9] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018. 1, 3, 4, 5

[10] Emilien Mathieu, Maximilian Nickel, and Douwe Kiela. Continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, 2019. 2, 4

[11] Pascal Mettes, Elise Van der Pol, and Cees Snoek. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019. 3

[12] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 2017. 1, 2, 3, 5

[13] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 3, 4, 5, 6

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 3, 4

[15] Shinnosuke Shimizu, Kento Hikita, Yusuke Yamada, et al. Hyperbolic convolutional neural networks with hybrid curvature. *CVPR*, 2022. 2, 5

[16] Max Van Spengler, Erwin Berkhout, and Pascal Mettes. Poincaré resnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5419–5428, 2023. 3

[17] H. Zhang, P.D. Rich, A.K. Lee, and T.O. Sharpee. Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience. *Nature Neuroscience*, 26 (1):131–139, 2023. 1