Accept the Modality Gap: An Exploration in the Hyperbolic Space

Sameera Ramasinghe Violetta Shevchenko Gil Avraham Ajanthan Thalaiyasingam
Amazon, Australia

Abstract

Recent advancements in machine learning have spotlighted the potential of hyperbolic spaces as they effectively learn hierarchical feature representations. While there has been progress in leveraging hyperbolic spaces in single-modality contexts, its exploration in multimodal settings remains under explored. A recent work has sought to transpose Euclidean multimodal learning techniques to hyperbolic spaces, by adopting a geodesic distance based contrastive loss. However, we show both theoretically and empirically that such spatial proximity based contrastive loss significantly disrupts hierarchies in the latent space. To remedy this, we advocate that the cross-modal representations should accept the inherent modality gap between text and images, and introduce a novel approach to measure cross-modal similarity that does not enforce spatial proximity. Our approach shows remarkable capabilities in preserving unimodal hierarchies while aligning the two modalities. Our experiments on a series of downstream tasks demonstrate that a better latent structure emerges with our objective function while being superior in text-to-image and image-to-text retrieval tasks.

1. Introduction

Hierarchical structures are a fundamental component of the natural world. Multimodal foundational models that try to learn a holistic view of the world with a shared image and text representation, *e.g.*, CLIP [28], have primarily leveraged Euclidean and spherical geometries. However, the intrinsic geometric constraints of these spaces often fall short in capturing the complexity and granularity of hierarchical information. This limitation sparks a compelling case for hyperbolic spaces as they offer continuous approximation for tree-like hierarchical structures [6, 9, 12, 18, 34].

Although hierarchical embeddings in the hyperbolic space have been previously explored in unimodal settings [11, 12, 18], learning shared representations of different types of modalities – such as text and images – while preserving hierarchies remains under explored. Current multimodal models measure the similarity of cross-modal embeddings through spatial proximity in the underlying shared embedding space. Such spatial proximity based contrastive loss clusters matching concepts across modalities together, while pushing apart non-matching ones. Despite wide-spread usage, the alignment between image and text modalities is an ill-posed problem, and these learned representations are shown to have misalignment between modalities, defined as the *modality-gap* [23].

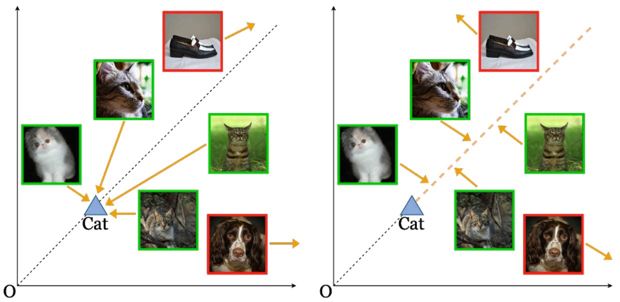

Figure 1. **Comparison between geodesic-based (left) and our angle-based contrastive losses (right) in the hyperbolic space.** Only the space component is shown for clarity. In our angle-based contrastive loss (details in Fig. 3), the images can be placed anywhere along the axis emanating from the text embedding (highlighted in yellow), which allows hierarchy among images.

We argue that modality gap is rooted in the intrinsic differences in the representational nature and information content of visual and linguistic data. Text, with its structured syntax and semantically rich lexicon, conveys abstract concepts and relationships explicitly. Images, in contrast, capture concrete instances of the world, expressing complex scenes and hierarchical relationships implicitly through visual cues. Thus, models that seek to minimize distance metrics between modalities may struggle to capture the nuanced associations between text and images. This often results in an oversimplified alignment that glosses over the rich, one-to-many correspondences from text to image.

A recent work [6] attempts to leverage this cross-modal hierarchy that text descriptions are more generic than images using an entailment loss, to learn a unified image-text representation in the hyperbolic space. Nevertheless, it still heavily relies on the spatial proximity based contrastive loss between image and text modalities, and suffers the same pitfalls as the Euclidean and spherical counterparts. In this

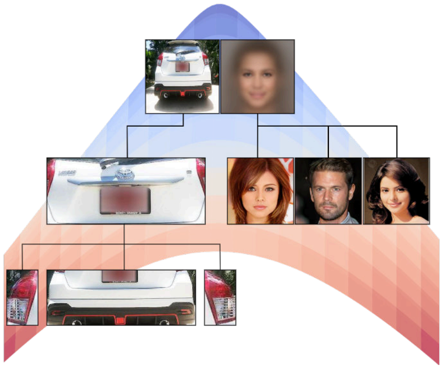

Figure 2. **Image hierarchy.** In visual domain, the hierarchy may arise from whole-to-fragment relations or by ambiguity of objects or identities in images.

work, we present theoretical and empirical evidence that the strategy of minimizing spatial proximity between modalities detrimentally impacts the hierarchical representation within both text and (specially) image embeddings.

Thus, we introduce a novel loss function that *accepts* the modality gap between image and text embeddings. We pivot from striving for spatial proximity in the latent space to leveraging a hyperbolic angle-based metric for assessing pairwise similarity. Our loss function not only allows embeddings to demonstrate better hierarchy, but also allows them to extend and better utilize the expanse of the hyperbolic space. Our idea is illustrated in Fig. 1 and a sample image hierarchy is shown in Fig. 2. Our contributions are as follows:

- We show that spatial proximity based contrastive loss for aligning image and text in the hyperbolic space is detrimental for preserving hierarchies. We show the existing work [6] that combines contrastive and entailment losses, face a fundamental mismatch in its objectives.
- We introduce a novel objective function that remedies the above issues by accepting the modality gap. Our objective preserves hierarchies in both language and visual concepts and utilize the expanse of the hyperbolic space better while aligning these two modalities.
- We show that the current extension of CLIP to hyperbolic space only performs well in near-Euclidean geometries with a low curvature, while ours is well suited for high curvature spaces.
- We perform extensive experiments on a series of downstream tasks to demonstrate that a better latent structure emerges with our objective function while being superior in text-to-image and image-to-text retrieval tasks.

2. Preliminaries

We briefly review essential concepts in hyperbolic geometry and the spatial proximity based contrastive loss below. We refer the reader to the textbook [29] and the CLIP paper [28] for more details on the respective topics.

2.1. Hyperbolic Spaces

Hyperbolic spaces are Riemannian manifolds with constant negative curvature and are fundamentally different to the Euclidean or spherical space which has zero or constant positive curvature, respectively. This enables unique properties such as the divergence of parallel lines and the exponential volume growth towards the boundary [1]. This volume growth property makes the hyperbolic space an ideal candidate for embedding hierarchical and graph structured data, and has found many machine learning applications.

2.1.1 Lorentz Model

The Lorentz model, also known as the Minkowski model, is a way to represent a hyperbolic space. It frames the d-dimensional hyperbolic space $\mathbb{H}^d$ with curvature c within an $(d+1)$-dimensional Euclidean space $\mathbb{R}^{d+1}$ as follows:

$$\mathbb{H}^d = \left\{ \mathbf{x} \in \mathbb{R}^{d+1} \mid \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{H}} = -1/c, x_0 > 0 \right\} \ , \tag{1}$$

where the Lorentzian inner product is defined as,

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}} = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i \ . \tag{2}$$

Here, the 0-th dimension of the vector is treated as the time component and the rest as the space component. From the definition of $\mathbb{H}^d$, the time component can be obtained from the space component as follows:

$$x_{\text{time}} = x_0 = \sqrt{1/c + \|\mathbf{x}_{\text{space}}\|^2} \ , \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm and $\mathbf{x}_{\text{space}} = \mathbf{x}_{1:d}$.

Geodesics. Geodesics in the hyperbolic space are the shortest paths between points, analogous to straight lines in Euclidean geometry. In the Lorentz model, geodesics are the intersections of planes through the origin with the hyperboloid. The geodesic distance between two points $\mathbf{x}, \mathbf{y}$ is,

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{y}) = \sqrt{1/c} \cosh^{-1}\left(-c \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}\right) \ . \tag{4}$$

Tangent Spaces. The tangent space at a point $\mathbf{x} \in \mathbb{H}^d$ in the hyperbolic space, is a Euclidean space that locally approximates the hyperbolic space around $\mathbf{x}$. Exponential and logarithmic maps are used to project a point from a tangent space to the hyperbolic space and vice versa. We defer the formulas for brevity and refer the reader to [26].

Centroid of Points. Obtaining the centroid of a set of points in the hyperbolic space is not as straightforward as the Euclidean setting. This is called the Einstein midpoint, and it is easier to obtain via converting to Klein coordinates [32]. Let $\mathbf{x} = (x_0, \mathbf{x}_{1:d}) \in \mathbb{H}^d$ be a point on the hyperboloid model, then it can be converted to Klein coordinates $\mathbf{k} \in \mathbb{K}^d$ and back via the following projections:

$$\Pi_{\mathbb{H}\to\mathbb{K}}(\mathbf{x}) = \frac{\mathbf{x}_{1:d}}{x_0} \; , \quad \Pi_{\mathbb{K}\to\mathbb{H}}(\mathbf{k}) = \frac{(1, \mathbf{k})}{\sqrt{c(1 - \|\mathbf{k}\|^2)}} \; . \quad (5)$$

Then the centroid takes the following form:

$$\text{Centroid}_{\mathbb{H}}\left(\{\mathbf{x}_j\}_{j=1}^N\right) = \Pi_{\mathbb{K}\to\mathbb{H}}\left(\sum_{j=1}^N \gamma_j \Pi_{\mathbb{H}\to\mathbb{K}}(\mathbf{x}_j) / \sum_{j=1}^N \gamma_j\right) \; , \quad (6)$$

where $\gamma_j = \frac{1}{\sqrt{1 - c\|\Pi_{\mathbb{H}\to\mathbb{K}}(\mathbf{x}_j)\|^2}}$ are the Lorentz factors.

2.2. Spatial Proximity based Contrastive Loss

The main idea of contrastive loss is to minimize the distance between matching datapoints (*i.e.*, positive pairs) while maximizing the distance between non-matching datapoints (*i.e.*, negative pairs). Formally, let $\mathcal{B}$ be a batch of N image-text pairs, the text-to-image contrastive loss for these samples can be written as:

$$L^{T\to I}(\mathcal{B}, \kappa) = -\frac{1}{N}\sum_{i=1}^N \log \frac{\exp(\kappa(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_{j=1}^N \exp(\kappa(\mathbf{x}_i, \mathbf{y}_j)/\tau)} \; , \quad (7)$$

where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$ denote the text and image embeddings corresponding to data sample i, $\kappa : \mathbb{R}^{d\times d} \to \mathbb{R}$ is the similarity function, and $\tau > 0$ is the temperature parameter.

In CLIP [28], a symmetric contrastive loss is used, *i.e.*, $L_{\text{contr}}(\mathcal{B}, \kappa) = L^{I\to T}(\mathcal{B}, \kappa) + L^{T\to I}(\mathcal{B}, \kappa)$. Here, the embeddings are normalized to be unit norm and κ is the cosine similarity. It is easy to see that CLIP embeddings are in the $(d-1)$-dimensional hypersphere. Since cosine similarity is inversely proportional to the geodesic distance in the hypersphere, we can conclude that CLIP measures the cross-modal similarity using spatial proximity.

Analogously, the above contrastive loss can be extended to the hyperbolic space by ensuring that the image and text embeddings are in $\mathbb{H}^d$ and using geodesic distance to measure proximity, *i.e.*, by setting $\kappa = -d_{\mathbb{H}}$. This loss is used in the recent work [6][1] together with an entailment loss to impose a hierarchy that text entails images. We discuss the interplay between these two losses in the next section.

3. Interplay between Losses

As previously discussed, contrastive loss in spherical or hyperbolic spaces try to *bridge* the modality gap between text and image embeddings with respect to the spatial proximity. Consequently, the objective is to pull matching concepts closer while non-matching ones should be pushed apart, regardless of the modality. Therefore, to encourage cross-modal hierarchy, Desai et al. [6] employ an entailment loss to enforce a partial hierarchy between image and text embeddings, which was originally proposed in [10]. The purpose is to enforce the prior that a text embedding is a more abstract concept compared to corresponding image embeddings. To this end, the entailment loss forces all the image embeddings matching with a text embedding inside a cone that emanates from the the text embedding.

Formally, the entailment loss between a text ($\mathbf{x}$) and image ($\mathbf{y}$) embedding is written as:

$$L_{\text{entail}}(\mathbf{x}, \mathbf{y}) = \max\left(0, \text{ext}(\mathbf{x}, \mathbf{y}) - \text{aper}(\mathbf{x})\right) \; . \quad (8)$$

Here, $\text{ext}(\cdot, \cdot)$ denotes the exterior angle between the text and the image embedding (α in Fig. 3), written as:

$$\text{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left(\frac{y_{\text{time}} + x_{\text{time}}\, c\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{H}}}{\|\mathbf{x}_{space}\|\sqrt{(c\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{H}})^2 - 1}}\right) \; , \quad (9)$$

and $\text{aper}(\cdot)$ is the aperture angle of the cone, which decreases as the norm of the space component increases, given by,

$$\text{aper}(\mathbf{x}) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\|\mathbf{x}_{space}\|}\right) \; . \quad (10)$$

Here K is a constant hyperparameter chosen to mitigate the discontinuity of cones near the origin.

Theoretically, contrastive loss pushes all image embeddings closer to the matching text embedding, reducing diversity of image embeddings, and hence discouraging hierarchies. Entailment loss does not prevent this collapse, therefore contrastive and entailment loss combination discourages hierarchies, especially in the image domain.

On the other hand, in practice, due to natural variations of text and images, and stochasticity, there exists some diversity in the latent space when trained using contrastive loss [28]. Nevertheless, we show that if there exists diversity across image embeddings, then it is highly likely that the entailment objective will be violated. We state this formally for a 2-dimensional case below.

Proposition 1. *Consider a set of points* $\{\mathbf{y}_i\}_{i=1}^n \in \mathbb{H}^2$ *and a point* $\mathbf{x} \in \mathbb{H}^2$*. Assume that the hyperbolic distance* $d_{\mathbb{H}}(\mathbf{x}, \mathbf{y}_i) = r$ *for all* i*. Then, for* $n > 1$*, to maximize the sum* $\sum_{i=1}^n \sum_{j=1}^n d_{\mathbb{H}}(\mathbf{y}_i, \mathbf{y}_j)$ *at least one* $\mathbf{y}_i$ *must reside outside the entailment cone originating from* $\mathbf{x}$.

(Proof in supplementary material). Intuitively, consider an example of a text embedding of a cat and a set of visually distinct images of cats. Because of the contrastive loss, the image embeddings are required to maintain a consistent distance to the text embedding, as all images depict cats. However, due to the visual diversity of images, it is imperative

[1] To our knowledge, this is the only instance of extending CLIP to the hyperbolic space.

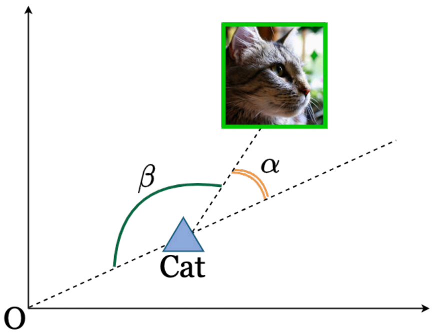

Figure 3. **A graphical illustration of our loss:** only the space component is shown for clarity. Our angle-based contrastive loss will maximize β and minimize α, encouraging the images and texts to be aligned, without forcing spatial proximity. The unimodal hierarchies are implicitly achieved by the model bias originating from the hyperbolic space.

to consider these image embeddings are maximally separated (if they are not separated, the model is forced to only learn abstract features that glosses over visual cues crucial for preserving hierarchy). This proposition indicates that in such scenarios, at least one image embedding will inevitably fall outside the entailment cone, highlighting a fundamental discrepancy between the objectives of contrastive loss and entailment losses. This discrepancy between the losses is observed empirically in the original paper [6] as well as in our experiments.

In the subsequent analysis, we demonstrate that the simultaneous minimization of both losses necessitates a rapid decrease in the area occupied by image embeddings (*i.e.*, collapse) within the hyperbolic space. This constraint significantly limits the ability to utilize the full extent of the hyperbolic space, thereby impeding the establishment of a hierarchical structure. We state this formally below.

Proposition 2. *Consider a set of points* $\{\mathbf{y}_i\}_{i=1}^n \in \mathbb{H}^2$ *and a point* $\mathbf{x} \in \mathbb{H}^2$*. Let* $d_{max} = \max_i(d_{\mathbb{H}}(\mathbf{x}, \mathbf{y}_i))$*. Further, let* $\{\mathbf{y}_i\}_{i=1}^n$ *be contained within the entailment cone originating from* $\mathbf{x}$*. Then, as* d_{max} *decreases, the total area of the spread of the image embeddings decreases exponentially.*

(Proof in the supplementary material). To intuitively understand the implications, let us revisit the earlier "cat" example. When both losses converge, all image embeddings are confined within the entailment cone, and the convergence of the contrastive loss reduces the maximum geodesic distance from these embeddings to the text embedding. Consequently, the spatial domain occupied by the image embeddings undergoes an exponential contraction. This scenario is detrimental to maintaining any form of image hierarchy based on spatial proximity, as it necessitates a tight clustering of the embeddings within an exponentially diminishing area.

4. Our Approach: Accept the Modality Gap

The discussion thus far highlighted the limitations inherent in applying geodesic-based contrastive losses, including the entailment cone loss, within hyperbolic spaces. A key insight emerges from these limitations is that they stem from striving to spatially bridge the gap between two fundamentally distinct modalities. To remedy this, we present a novel hypothesis: these challenges can be effectively addressed by adopting an objective function that *acknowledges the modality gap*, employing an alternative approach to measure cross-modal concept similarity.

Based on this hypothesis, we introduce a unique constraint where all matching concepts (whether unimodal or cross-modal) are aligned along a specific geodesic emanating from the origin of the hyperbolic space. Within this framework, image embeddings are positioned further from their textual counterparts, reflecting their greater specificity compared to text. Consequently, the deviation angle between two concepts along this geodesic becomes a new metric for evaluating their conceptual distance.

Figure 3 graphically illustrates our approach. To exemplify, consider aligning the image embedding for a "cat" with its corresponding textual descriptor. Our objective is to minimize the angle α while simultaneously maximizing the angle β, thereby achieving an optimal alignment between the two modalities. Similarly, for non-matching concepts, α will be maximized and β will be minimized. The angle α in the hyperbolic space is computed using Eq. (9), *i.e.*, $\alpha(\mathbf{x}, \mathbf{y}) = \text{ext}(\mathbf{x}, \mathbf{y})$ and $\beta(\mathbf{x}, \mathbf{y}) = \pi - \alpha(\mathbf{x}, \mathbf{y})$. Note that the sum of angles on a geodesic passing through the origin adds up to π. Then our angle-based contrastive loss can be written as:

$$L_{\text{angle}}(\mathcal{B}) = L^{T\rightarrow I}(\mathcal{B}, -\alpha) + L^{T\rightarrow I}(\mathcal{B}, \beta) \,, \tag{11}$$

where $L^{T\rightarrow I}$ denotes the text-to-image contrastive loss in Eq. (7). The similarity function κ is replaced in the above equation with angles α and β such that the contrastive loss minimizes the angle α while maximizing β for matching pairs in the batch $\mathcal{B}$, and vice versa.[2] Furthermore, our loss is asymmetric as we impose the entailment relationship that text is more generic than an image. Therefore, it can be regarded as a smoothed, contrastive version of the entailment loss (Eq. (8)), as we are aligning similar concepts along the axis of cones. Note that both the terms of L_{angle} are satisfied mutually, hence, there is no mismatch between their objectives. In our experiments, L_{entail} alone did not yield meaningful results, highlighting the importance of our smooth, contrastive version and the underlying idea behind it.

To further encourage better distribution of embeddings on the hyperbolic manifold, we apply a soft regularizer at

[2] Theoretically, minimizing α maximizes β. Nevertheless, explicitly maximizing β imposes a stronger bias, analogous to the symmetric contrastive loss.

the distribution level. Specifically, we impose that centroid of text embeddings should be closer to the origin than the centroid of the image embeddings. Formally, let $\mathbf{x}_e$, $\mathbf{y}_e$ be the Einstein midpoint (*i.e.*, centroid, see Eq. (6)) of a set of text and image embeddings, respectively. Then, our regularization takes the following form:

$$\mathcal{L}_{\text{centroid}} = \left\| \mathbf{x}_e - \tfrac{1}{\sqrt{c}} \cosh^{-1}(cq) \right\| + \left\| \mathbf{y}_e - \tfrac{1}{\sqrt{c}} \cosh^{-1}(cp) \right\| \ , \tag{12}$$

where $\| \cdot \|$ is the Euclidean norm and $p > q$ to ensure that the centroid of images are further from the origin than text. Then, our final loss is,

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{angle}} + \lambda \, \mathcal{L}_{\text{centroid}} \ , \tag{13}$$

where $\lambda > 0$ is the trade-off hyperparameter.

We follow the parametrization of [6], in that we encode only the space component of the Lorentz model in the tangent space using the neural network. The exponential map to project to the hyperboloid takes the following form:

$$\mathbf{x}_{\text{space}} = \frac{\sinh\left(\sqrt{c}\|\mathbf{v}\|\right)}{\sqrt{c}\|\mathbf{v}\|} \mathbf{v} \ , \tag{14}$$

where $\mathbf{v} \in \mathbb{R}^d$ and $\| \cdot \|$ is the Euclidean norm. The corresponding time component is obtained via Eq. (3), and $\mathbf{x} = (x_{\text{time}}, \mathbf{x}_{\text{space}}) \in \mathbb{H}^d$ is used to compute the loss.

5. Related Work

Embeddings in Hyperbolic Spaces. Hyperbolic geometry enables exponential volume expansion towards the boundary [1], enabling embedding hierarchical structures more efficiently. This feature has produced a surge of works that exhibit such hierarchical relationships including molecular structures [38], action recognition [9], 3D data [2, 34], text data [11, 31, 40] and images [13, 18]. Hyperbolic embeddings can be learned using standard deep learning layers [14, 17] and performing hyperbolic projection [25] or via hyperbolic neural networks [11]. To further incorporate a tree-like inductive bias, Ganea et al. [10] introduce an entailment loss which forces child nodes to be contained within the cone emanating from the parent node embedding. Further, two recent works [12, 39] utilize contrastive learning [3] but with minimising the geodesic distance objective in the hyperbolic space. Recently, hyperbolic embeddings have been adopted for the cross-modality setting [6] and a combination of contrastive and entailment loss is used. In this work, we discuss the interplay between these losses, their drawbacks, and introduce a new loss to mitigate them.

Joint Multimodal Learning. Joint multimodal learning comprises a vast literature [35]. Early pretraining works in large multimodal models emerged from unimodal pretraining [3, 7, 37]. The underlying principle for these approaches is constructing pretraining tasks that result in a structured embedding space that is useful for downstream tasks. Extending above to multimodal settings assumes an extra weak supervision signal with paired images and text. The task in multimodal learning then comprises pretraining tasks using paired information [20, 36]. Most notably, CLIP [28] and ALIGN [15] train a contrastive objective which aligns text and image embeddings. Rapid progress has been made in the recent years to extend these encoder only multi-modal models to include a text decoder [21] or integrate with a large language model [22, 24].

6. Experiments

In this section, we first highlight the superior performance of our model in zero-shot classification and retrieval tasks, outperforming both MERU [6] and CLIP [28]. Subsequently, we delve into a comparative analysis, illustrating the enhanced visual-text hierarchy achieved by our model. Then, we expose a critical examination of MERU's loss function, emphasizing its limitations in handling curved spaces. Finally, we demonstrate the more efficient utilization of embedding space by our model.

For a fair comparison, we ensured identical architectures across our model, MERU, and CLIP, with both ours and MERU incorporating the same additional projection layer into the hyperbolic space. As the image encoder, we employed ViT-S [4], and for the text encoder, we used the model from [33]. In line with MERU, we trained all the models from scratch utilizing the Redcaps dataset [5]. However, we noted that some image links were removed from the original dataset, which led us to use an $\sim$ 8M subset of the original dataset ($\sim$ 12M). For detailed information on hyperparameters, and training, please refer to the supplementary material.

6.1. Zero-shot Image Classification

The learned correspondences between images and text enable zero-shot image classification by embedding class labels as text prompts [6, 8, 28]. We first embed a set of prompts for each class label, projecting them into the hyperbolic space, and then computing the average of the α values (exterior angle as in Fig. 3) against the image embedding. The class corresponding to the minimum average α is subsequently selected as the predicted class. For MERU, we used the maximum average Lorentzian inner product to select the predicted class as proposed in the original work. A comparison of Top-1 accuracy across various datasets is presented in Table 1. Note that some datasets, which contain classes that are underrepresented by Redcaps cause poor performance across all the models. As also discussed in [6], this can be alleviated by using a large scale dataset for training. As indicated, our model achieves better

	ImageNet	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	MNIST	CLEVR	PCAM	SST2
CLIP	28.5	70.1	55.8	24.2	29.7	21.9	**1.5**	13.0	64.3	56.3	**50.7**	86.3	22.0	19.6	3.7	9.2	**14.2**	50.7	50.0
MERU	28.6	71.6	53.7	25.0	30.3	21.8	1.3	11.1	**66.5**	57.5	50.3	87.1	29.6	23.3	3.7	10.2	12.4	**55.2**	50.0
Ours	**29.7**	**71.7**	**61.8**	**27.1**	**32.3**	**22.5**	0.8	**14.0**	65.0	**57.6**	46.9	**88.0**	**34.7**	**24.8**	**4.1**	**10.5**	13.7	50.0	50.0

Table 1. **Zero-shot image classification.** Our model demonstrates overall better performance compared to both MERU and CLIP.

	text → image				*image → text*			
	COCO		Flickr		COCO		Flickr	
	R5	R10	R5	R10	R5	R10	R5	R10
CLIP	22.7	31.4	30.8	41.9	24.8	33.6	30.2	39.5
MERU	23.4	32.1	33.1	42.3	26.5	35.3	**31.4**	**42.3**
Ours	**23.7**	**32.7**	**33.2**	**42.7**	**32.6**	**42.5**	30.7	39.7

Table 2. **Zero-shot image and text retrieval.** We show overall better performance over both MERU and CLIP.

results in 13 out of 19 datasets, demonstrating that the representations produced by our approach are superior overall.

6.2. Zero-shot Retrieval

We assess the retrieval performance of our model against CLIP and MERU using two well-known benchmarks: COCO and Flickr30K, which contain 5000 and 1000 images, respectively, each paired with five captions. Our test splits are identical to Desai et al. [6]. During inference, we order a set of candidate image/text embeddings based on their exterior angle relative to a given text/image query embedding. Table 2 reports recall@5,10 of our model and the reproduced CLIP and MERU baselines on these benchmarks. Hyperbolic representations of our model perform best on 3 out of 4 tasks.

6.3. Vision and Language Hierarchies

The primary advantage of embedding features in the hyperbolic space is to capture the implicit hierarchies within each data modality. As discussed in Sec. 4, geodesic based contrastive losses can severely impede these hierarchies, in contrast to the proposed method. To demonstrate this, we conduct a series of experiments next.

Visual hierarchy. Visual hierarchies can emerge in two ways: **1)** either through the relationship between a composition of objects and the objects therein, or **2)** through deteriorated/abstract images that may be associated with multiple identities or classes [18] (see Fig. 2). Below, we demonstrate our method's superior ability to capture both types of hierarchies.

			Curvatures					
			0.1	0.2	0.5	1.0	2.0	3.0
Car Parts	depth-1	MERU	19.2	5.9	-	-	-	-
		Ours	**94.0**	**96.8**	96.8	92.7	93.3	96.2
	depth-2	MERU	0.6	0.0	-	-	-	-
		Ours	**31.7**	**37.9**	35.4	38.5	40.4	37.9
Open Images	depth-1	MERU	31.1	30.3	-	-	-	-
		Ours	**69.1**	**69.8**	69.8	70.4	71.4	70.6
	depth-2	MERU	10.5	11.0	-	-	-	-
		Ours	**33.2**	**35.0**	33.4	35.3	36.8	35.2

Table 3. **Image hierarchy accuracy (%).** Our method significantly outperforms MERU on both datasets.

		Curvatures					
		0.1	0.2	0.5	1.0	2.0	3.0
depth-1	MERU	88.1	84.6	-	-	-	
	Ours	**93.4**	**92.0**	91.5	93.6	90.3	92.6
depth-2	MERU	58.1	55.3	-	-	-	-
	Ours	**73.5**	**70.0**	69.0	69.0	69.5	70.5

Table 4. **Text hierarchy accuracy (%).** Our method further improves text hierarchies.

To evaluate the first type of hierarchies, we create visual hierarchy chains using two datasets: Car Parts Segmentation [27] and OpenImages [19]. In Car Parts, we extracted images of parts using segmentation masks in a hierarchical order, *e.g.,* "car → hood → headlights", to create hierarchical chains. Similarly, for Open Images, we created hierarchical chains by extracting objects using bounding boxes (see supplementary). For Car Parts, we created 161 depth-2 chains and 811 depth-1 chains. For Open Images, we created $188,261$ depth-2 chains and $166,143$ depth-1 chains. For evaluating the models, we arranged these objects/parts by their embedding proximity to the origin, and assessed the accuracy of this ordering against the known hierarchical structure (objects deeper in the hierarchy should be embedded further away from `ROOT`, *i.e.*, origin). Our approach

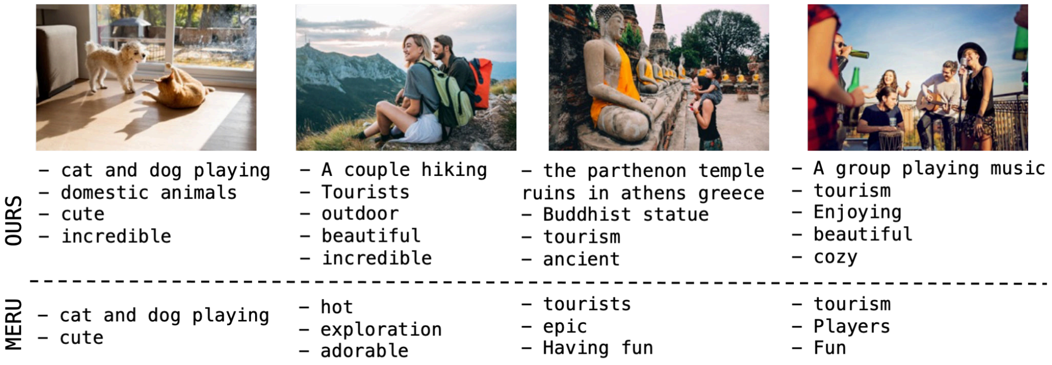

Figure 4. **An illustration of the superior text hierarchy of our model.** We retrieve multiple text descriptions while traversing from an image embedding to [ROOT]. Our model is able to retrieve richer hierarchical text descriptions compared to MERU.

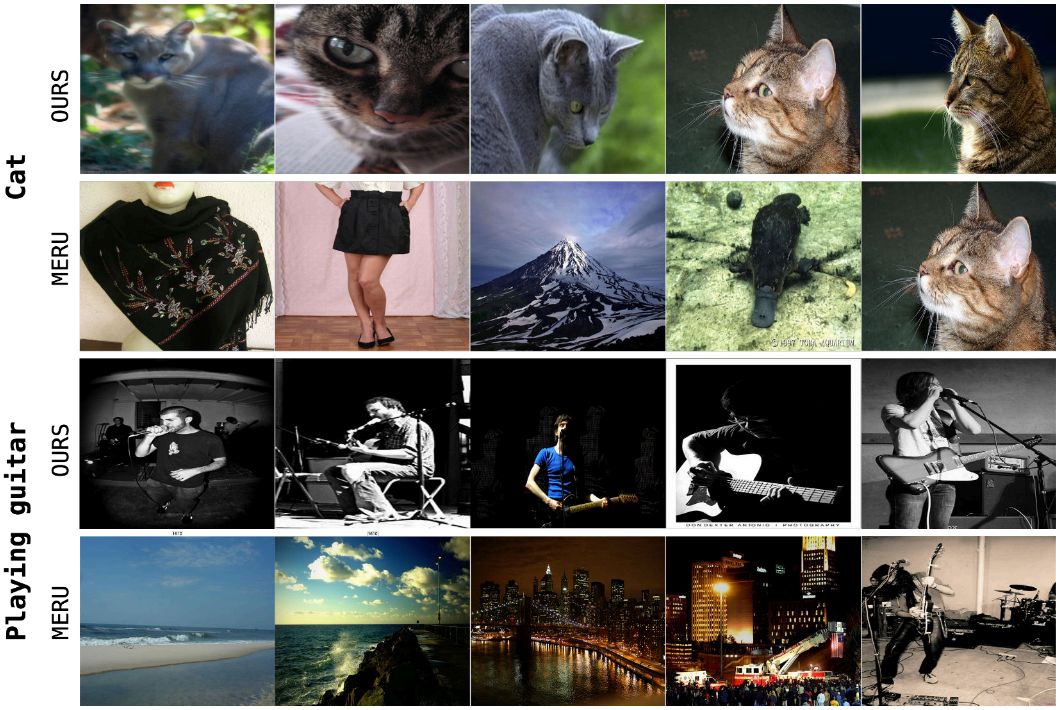

Figure 5. **Visual hierarchy as a measure of uncertainty in image retrieval**. As evident, when the distance to the [ROOT] increases (left → right), our model retrieves similar images with an increasing hierarchical order where the text prompt is better described. In contrast, retrieval results from MERU does not preserve the hierarchy or the visual similarity.

significantly surpasses MERU in preserving image hierarchies, as evidenced in Table 3.

To evaluate the second type of hierarchies, we first compute the 1000 most prominent eigenfaces from the Celeb-HQ dataset [16]. Then, we randomly sampled 1000 examples from Celeb-HQ and plotted histograms of the geodesic distance to the [ROOT] for these two distributions (see Fig. 6). The results illustrate that our model accurately positions the eigenfaces in close proximity to [ROOT], in contrast to MERU, which tends to place both distributions around a common mean value.

Image hierarchy as an uncertainty measure. Preserving hierarchies results positioning more abstract/ambiguous objects nearer to the [ROOT], and more specific objects closer to the boundary. Thus, the distance from the origin serves as a natural indicator of uncertainty. This attribute can be particularly advantageous for retrieval tasks, as we shall demonstrate next. For text embeddings, *e.g.*, *cat*, *playing guitar*, we find the 40 closest image embeddings per each text by utilizing the Lorentzian inner product for MERU and the angle α for ours. We then sort image embeddings according to their distances to the [ROOT] and pick top five

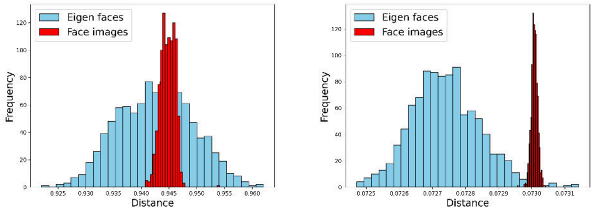

Figure 6. **The distribution of distances to the origin of eigen faces and face images.** In MERU (left), the distributions are centered around an approximately common mean, whereas ours (right) shows a clear distinction between the distributions.

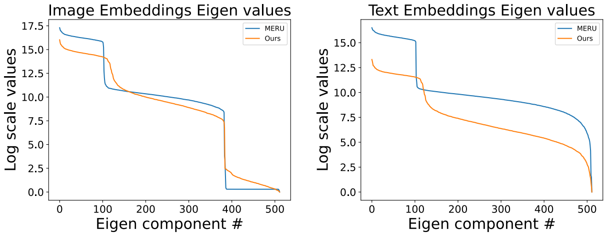

Figure 7. **Curves of the log scaled eigenvalues of image and text embeddings.** Note the order 2 magnitude difference for the largest eigenvalues for MERU. This phenomenon points to high density concentration along a small number of dimensions.

embeddings. We apply deduplication based on the distance to the `[ROOT]`. Two examples are in Fig. 5. As shown, our model consistently presents images that not only resemble each other more closely but also exhibit a progressive hierarchical relationship as one moves further from the `[ROOT]` (from left to right). Importantly, note that the text prompt is better described as the image gets further away from the `[ROOT]`, demonstrating reduced uncertainty. In comparison, MERU's retrieval outcomes fail to uphold either the hierarchical progression or visual coherency. See supplementary for more examples.

Text hierarchy. MERU showed that it can preserve satisfactory text hierarchy. We show that this effect is further improved by our objective function. Our first experiment follows Desai et al. [6]'s image traversal. We traverse from image embeddings, extracted from GettyImages[3], interpolating 50 equally spaced steps along the geodesic connecting their embedding vectors to `ROOT`. We use every interpolated step embedding as a query to retrieve the nearest neighbor from a set of text embeddings that are extracted from *pexels.com* and YFCC dataset [30]. We use an expanded set of texts compared to the version used in [6] to facilitate a broader range of images for both models. As shown in Fig. 4, our model has learned richer hierarchies. In the next experiment, we created a set of text hierarchy chains using the above collected text captions as "adjectives → nouns → captions" and measure the predicted hierarchy similar to the image chain experiments above. As Table 4 depicts, we achieve better results compared to MERU.

	Curvatures					
	0.1	0.2	0.5	1.0	2.0	3.0
MERU	28.4	20.9	-	-	-	-
Ours	**29.6**	**29.4**	28.9	29.7	28.5	29.1

Table 5. **Accuracy (%) on ImageNet zero-shot classification.** Note the significant drop of MERU when the curvature increases.

6.4. Effect of Curvature

As discussed, there exists a fundamental mismatch in employing a geodesic contrastive loss in hyperbolic space. Further solidifying this insight, we empirically observed that geodesic contrastive loss compels MERU towards a lower (trainable) curvature, eventually being clipped at 0.1. In contrast, our loss function increases the curvature, suggesting its suitability for curved spaces (see supplementary for loss curves). To substantiate this further, we trained models using fixed curvatures. As detailed in Table 5, MERU failed to converge for curvatures ≥ 0.5, whereas our model demonstrated almost constant accuracy. This behaviour is consistent across unimodal hierarchy experiments as well.

6.5. Better Space Utilization

We observed that the embeddings of MERU tend to concentrate within narrow cones, leading to under-utilization of the embedding space. This phenomenon has also been observed in CLIP [23]. We found that our loss function facilitates better spatial utilization, likely due to the dispersion of embeddings resulting from the angle-based contrastive loss. To illustrate this, we plot the eigenvalue distributions of the embeddings and noted that MERU exhibits orders of magnitude higher energy in the first few eigenvectors (Fig. 7), suggesting a greater concentration within a narrow space.

7. Conclusion

We showed that while hyperbolic spaces are useful at preserving hierarchies in single-modality settings, their potential in multimodal settings entails unique challenges. We introduced a novel loss function that accepts the modality gap and enables hierarchical structure in both image and text embeddings while better aligning these modalities. The presented insights promise to enhance the adoption of hierarchical representations in multimodal settings.

[3] https://www.gettyimages.com

References

[1] Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*. Springer Science & Business Media, 2013. 2, 5

[2] Jiaxin Chen, Jie Qin, Yuming Shen, Li Liu, Fan Zhu, and Ling Shao. Learning attentive and hierarchical representations for 3d shape recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 105–122. Springer, 2020. 5

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[4] X Chen, S Xie, and K He. An empirical study of training self-supervised vision transformers. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629. 5

[5] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 5

[6] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 1, 2, 3, 4, 5, 6, 8

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[8] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591, 2013. 5

[9] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. *arXiv preprint arXiv:2303.06242*, 2023. 1, 5

[10] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018. 3, 5

[11] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018. 1, 5, 11

[12] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6840–6849, 2023. 1, 5

[13] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2022. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 5

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7

[17] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 5

[18] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6418–6428, 2020. 1, 5, 6

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 6

[20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11336–11344, 2020. 5

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 5

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5

[23] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1, 8

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 5

[25] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 5

[26] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018. 2

[27] Kitsuchart Pasupa, Phongsathorn Kittiworapanya, Napasin Hongngern, and Kuntpong Woraratpanya. Evaluation of

deep learning algorithms for semantic segmentation of car parts. *Complex & Intelligent Systems*, pages 1–13, 2021. 6

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5

[29] John G Ratcliffe, S Axler, and KA Ribet. *Foundations of hyperbolic manifolds*. Springer, 1994. 2

[30] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 8

[31] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 5

[32] Abraham A Ungar. *Analytic hyperbolic geometry: Mathematical foundations and applications*. World Scientific, 2005. 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[34] Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5176–5185, 2023. 1, 5

[35] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, pages 1–36, 2023. 5

[36] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 5

[37] Gokul Yenduri, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Rutvij H Jhaveri, Weizheng Wang, Athanasios V Vasilakos, Thippa Reddy Gadekallu, et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*, 2023. 5

[38] Zhen Yu, Toan Nguyen, Yaniv Gal, Lie Ju, Shekhar S Chandra, Lei Zhang, Paul Bonnington, Victoria Mar, Zhiyong Wang, and Zongyuan Ge. Skin lesion recognition with class-hierarchy regularized hyperbolic embeddings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–603. Springer, 2022. 5

[39] Yun Yue, Fangzhou Lin, Kazunori D Yamada, and Ziming Zhang. Hyperbolic contrastive learning. *arXiv preprint arXiv:2302.01409*, 2023. 5

[40] Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. Hypertext: Endowing fasttext with hyperbolic geometry. *arXiv preprint arXiv:2010.16143*, 2020. 5