Accept the Modality Gap: An Exploration in the Hyperbolic Space

Supplementary Material

A. Proof for propositions

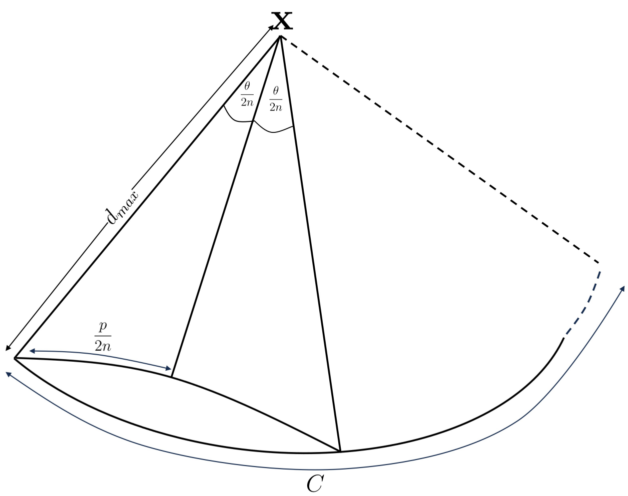

Figure 8

Proof for proposition 1: Consider the hyperbolic triangle formed by $\mathbf{x}, \mathbf{y}_i, \mathbf{y}_j$. Let the angle $\angle \hat{\mathbf{y}}_i \mathbf{x} \mathbf{y}_j / 2 = \theta_{i,j}$ and $d_{\mathbb{H}}(\mathbf{y}_i, \mathbf{y}_j)/2 = d_{i,j}$. By hyperbolic trigonometric relationships we get,

$$\sin(\theta_{i,j}) = \frac{\sinh(d_{i,j})}{\sinh(r)}$$

$$d_{i,j}(\theta_{i,j}, r) = \operatorname{arcsinh}(\sin(\theta_{i,j})\sinh(r))$$

First, we show that $d_{i,j}(\theta_{i,j})$ is a concave function over $\theta_{i,j}$ (since r is fixed, we can treat $\sinh(r) = a$ as a constant). Taking the second derivative of $d_{i,j}(\theta_{i,j})$ we have,

$$\frac{d^2}{d\theta_{i,j}^2} d_{i,j}(\theta_{i,j}) = - \frac{a^3 \sin^3(\theta_{i,j}) + \left(a^3 \cos^2(\theta_{i,j}) + a\right)\sin(\theta_{i,j})}{\left(a^2 \sin^2(\theta_{i,j}) + 1\right)^{\frac{3}{2}}},$$

Which is negative for $\theta_{i,j} < \pi$. Thus, $d_{i,j}(\theta_{i,j})$ is concave for $0 < \theta_{i,j} < \pi$. Therefore, applying Jensen's inequality for concave functions, we have

$$\frac{\sum_{j=1}^{n} \operatorname{arcsinh}(\sin(\theta_{i,j})a)}{n} \leq \operatorname{arcsinh}(\sin(\frac{\sum_{j=1}^{n}\theta_{i,j}}{n})a).$$

One can see that the equality is achieved for equiangular $\theta_{i,j}$, i.e., at $\theta_{i,j} = \frac{\pi}{n}$, $\sum_{j=1}^{n} d_{i,j}$ is maximized. On the other hand, [11] showed that the angle of hyperbolic entailment cones is less than π. Therefore, for $n > 1$, at least one point lies outside the cone.

Proof for proposition 2: We consider a hyperbolic entailment cone eminating from a point $\mathbf{x} \in \mathbb{H}^2$. We consider the area η within the cone where $\eta = \{\mathbf{u} \in \mathbb{H}^2 | d_{\mathbb{H}}(\mathbf{u}, \mathbf{x}) \leq d_{max}\}$. Let the angle of the cone be θ. Now, we divide the cone in to n equiangular hyperbolic triangles;see Fig. 8. Invoking the trigonometric relationship in the hyperbolic space, we get,

$$\sin(\frac{\theta}{2n}) = \frac{\sinh(\frac{p}{2n})}{\sinh(d_{max})}$$

.

$$n\sinh(d_{max})\sin(\frac{\theta}{2n}) = n\sinh(\frac{p}{2n})$$

.

Then, we invoke the following Lemma.

Lemma 1. *If the function f is differentiable at 0 and $k \neq 0$, then $nf(k/n) \to kf'(0)$ as $n \to \infty$.*

Consider the limit $n \to \infty$. Then, $p \to C$. Therefore we have,

$$\frac{\theta}{2}\sinh(d_{max}) = C/2$$

$$C = \theta \sinh(d_{max})$$

Now, we invoke the following lemma.

Lemma 2. *If a hyperbolic triangle ABC has a right angle at A, and $d(A, B) = c$, $d(B, C) = a$, $d(C, A) = b$, then its hyperbolic area τ is given by $\sin(\tau) = \frac{\sinh(b)\sinh(c)}{(\cosh(a)+1)}$.*

Let the sum of the areas of the triangles be m. Then using the above result, we get,

$$\sin(m/2n) = \frac{\sinh(a)\sinh(p/2n)}{(\cosh(r)+1)}$$

$$2n\sin(m/2n) = \frac{\sinh(a)2n\sinh(p/2n)}{(cosh(r)+1)}$$

Now let $n \to \infty$. Then $m \to \eta$ and $p \to C$. Again, by Lemma 1, we get,

$$\eta = \frac{\sinh(d_{max})C}{(\cosh(d_{max})+1)} = \frac{\theta\sinh(d_{max})\sinh(d_{max})}{(\cosh(d_{max})+1)}$$

By applying hyperbolic trigonometric relationships, we get,

$$\eta = 2\theta \sinh^2(\frac{d_{max}}{2})$$

,

Now, since $\sinh(x) = \frac{e^x - e^{-x}}{2}$, consider,

$$\lim_{d_{max} \to \infty} \frac{\eta}{e^{d_{max}}}$$

.

Clearly, the above limit is larger than 1 at the infinity. Therefore, η decreases exponentially with d_{max}, which completes the proof.

B. Hyperparameters and training details

Our training setups closely resemble those of MERU. We employ the AdamW optimizer, setting the parameters (β_1, β_2) to $(0.9, 0.98)$ and applying a weight decay of 0.2, except for biases and learnable scalars where weight decay is not applied. Our models undergo training across $120,000$ iterations, each with a batch size of 2048. The peak learning rate is set to 5×10^{-4}, which initially increases linearly over the first 4000 iterations and then undergoes a cosine decay down to zero. For data agumentation, we randomly crop 50–100% area of images and resize them to 224×224.

C. Experiments on a larger model

To validate if our empirical findings extend to larger architectures, we conduct experiments with a model using ViT B/16 as the base architecture. The results are illustrated in Table 6, 7 8, 9, and 10. As shown, our results hold for larger architectures.

D. Suitability for curved spaces

We noticed that when the curvature of the models are trainable, MERU's loss tend to suppress the curvature until it gets clamped at 0.01. Intuitively, when the curvature gets lowered, the hyperbolic space converges towards an Euclidean space. On the other hand, with our loss, the curvature tends to increase; see Fig. 13. This can probably be attributed to the fact that our loss is more suitable for curved spaces. MERU's inability to converge in higher curvatures provides further evidence for this hypothesis.

		Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	MNIST	CLEVR	PCAM	SST2
ViT S/16	CLIP	83.8	89.0	71.4	68.0	58.8	44.1	68.3	89.8	84.5	96.4	95.0	95.9	87.7	13.9	98.5	83.9	56.4	55.3
	MERU	83.6	89.1	71.1	67.5	58.2	41.4	67.9	88.4	83.9	94.9	94.9	95.6	86.5	13.7	98.3	84.4	57.6	55.3
	Ours	83.9	88.8	70.9	68.4	57.8	39.8	67.3	87.6	82.9	95.0	94.6	95.6	87.1	13.5	98.2	82.9	54.1	54.8
ViT B/16	CLIP	86.1	91.2	74.2	70.4	61.3	49.2	70.5	90.6	86.0	96.5	95.7	96.5	89.0	15.2	99.0	86.4	55.7	56.6
	MERU	85.5	90.9	72.8	68.8	59.1	47.3	68.9	89.0	83.5	95.9	95.2	96.3	87.8	14.5	98.8	84.3	55.2	56.9
	Ours	85.9	90.4	73.1	67.1	60.4	48.2	67.4	89.1	83.2	96.2	95.4	95.9	88.4	16.1	98.8	85.2	55.9	56.7

Table 6. **Linear probe evaluation.** We train a logistic regression classifier on embeddings extracted from the image encoders of CLIP, MERU and our model.

	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	30.9	73.0	57.9	27.7	30.4	23.2	1.4	10.1	64.8	58.6	49.7	88.0	26.7	26.0	4.3	7.7	16.0	50.5	50.0
MERU	30.6	74.3	63.2	28.1	30.9	24.3	1.8	11.2	70.5	59.0	48.1	87.6	17.4	22.2	4.0	10.2	14.1	50.1	50.8
Ours	31.1	74.7	64.1	28.9	30.3	24.7	1.1	14.3	71.1	59.2	48.3	88.3	22.9	23.4	6.1	11.1	10.9	50.8	51.1

Table 7. **Zero shot image classification performance with ViT B/16 architecture.** We show overall better performance over both MERU and CLIP.

	text → image				*image → text*			
	COCO		Flickr		COCO		Flickr	
	R5	R10	R5	R10	R5	R10	R5	R10
CLIP	25.1	33.9	34.3	45.0	28.0	36.9	36.3	45.3
MERU	25.1	34.0	34.3	44.5	28.3	37.4	36.8	46.4
Ours	25.1	34.1	34.6	44.9	28.7	38.4	38.9	47.2

Table 8. **Zero-shot image and text retrieval with ViT B/16 architecture**. We show overall better performance over both MERU and CLIP.

			Curvatures					
			0.1	0.2	0.5	1.0	2.0	3.0
Car Parts	depth-1	MERU	21.1	18.1	-	5.1	-	-
		Ours	94.6	93.3	93.3	83.7	93.7	88.8
	depth-2	MERU	0.0	0.0	-	0.6	-	-
		Ours	33.5	28.6	28.6	28.6	32.9	29.8
Open Images	depth-1	MERU	31.1	34.6	-	26.3	-	-
		Ours	67.9	67.8	67.8	64.7	65.7	68.2
	depth-2	MERU	10.2	12.8	-	9.1	-	-
		Ours	33.2	34.2	34.0	30.7	31.0	33.6

Table 9. **Image hierarchy accuracy (%) with ViT B/16 architecture.** Our method significantly outperforms MERU.

Figure 9. **Qualitative results showing visual hierarchy as a measure of uncertainty in image retrieval**. As illustrated, 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.

Figure 10. **Qualitative examples 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.

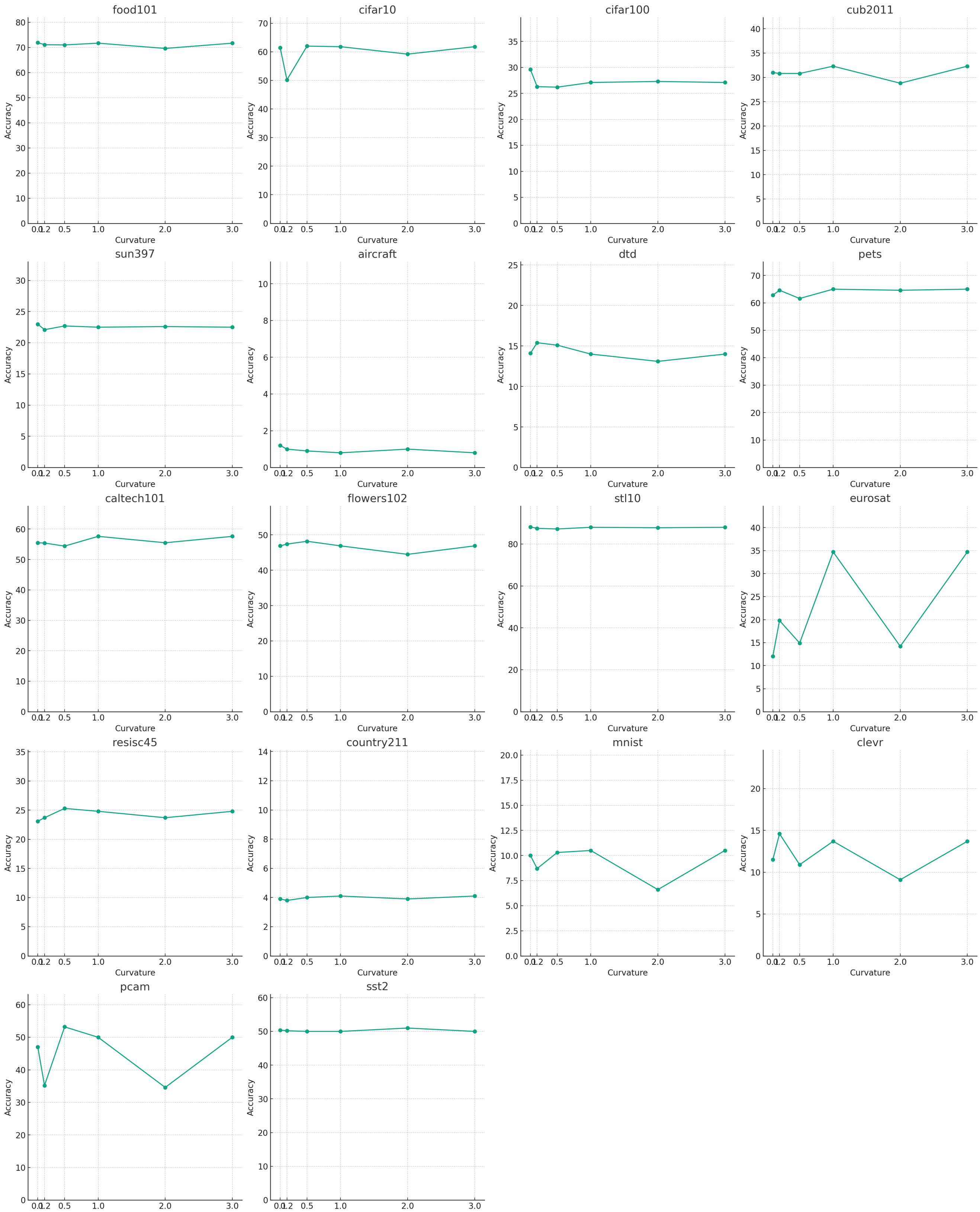

Figure 11. **Zero shot classification performance over curvature on different datasets.** Our model is able to maintain a an approximately consistent performance over varying curvature. In contrast, MERU did not converge for curvatures larger than 0.2.

		Curvatures					
		0.1	0.2	0.5	1.0	2.0	3.0
depth-1	MERU	83.1	81.9	-	80.9	-	
	Ours	**90.3**	**92.2**	92.7	**93.3**	88.8	91.7
depth-2	MERU	57.7	50.4	-	54.4	-	-
	Ours	**67.1**	**67.3**	69.0	**66.6**	63.2	68.6

Table 10. **Text hierarchy accuracy (%) with ViT B/16 architecture.** Our method further improves text hierarchies.

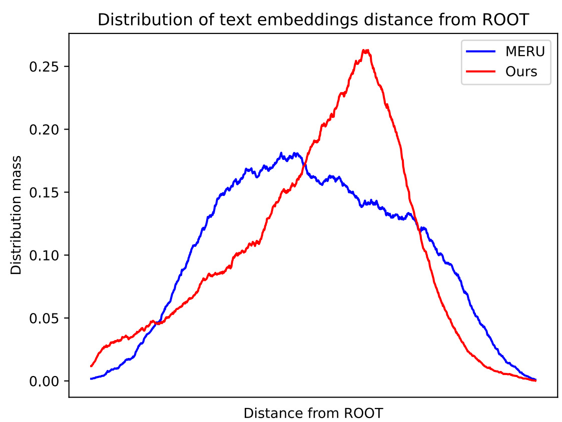

Figure 12. **Distribution of the text embeddings of our approach and MERU.** The skew appearance of our approach aligns with a hierarchical structure where the taxonomy of concepts generally expand such that high-level concepts populate the areas closer to `[ROOT]` and low-level details further away.

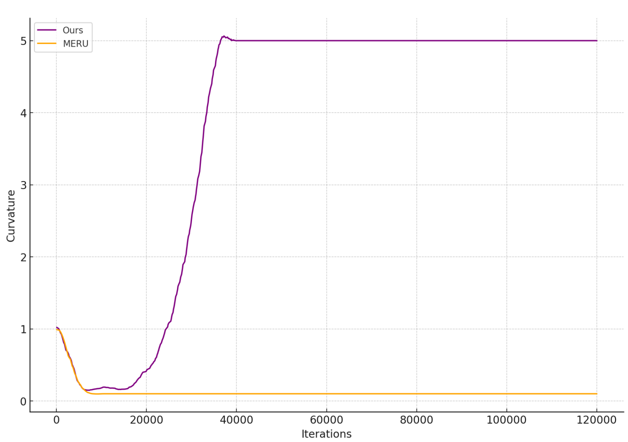

Figure 13. **Behavior of the curvature while training.** When the curvature is trainable, MERU suppresses the curvature until it gets clamped at 0.1. In contrast, our model increases the curvature. This maybe an indication of the better suitability of our loss for curved spaces.

	ImageNet	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	MNIST	CLEVR	PCAM	SST2
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
Ours (w/o Einstein reg)	29.5	71.3	61.8	27.0	32.1	22.7	0.9	15.1	64.8	57.8	46.5	87.5	34.3	24.9	4.9	11.8	15.5	50.0	50.2

Table 11. **Effect of the Einstein regularization loss in zero shot image classification.** We noticed that having the regularization marginally improves the results.

Figure 14. An illustration of the OpenImage hierarchies used to evaluate the models.

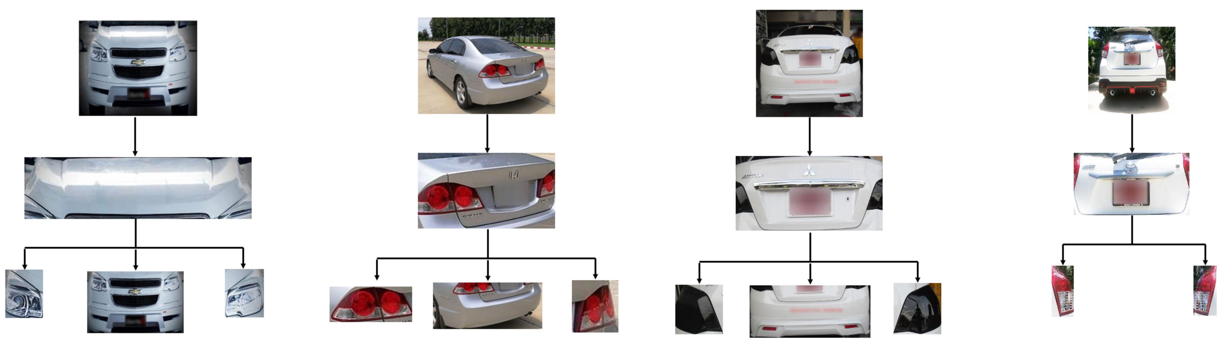

Figure 15. An illustration of the car parts hierarchies used to evaluate the models.