HyperVLM: Hyperbolic Space Guided Vision Language Modeling for Hierarchical Multi-Modal Understanding

Sarthak Srivastava*
Amazon
sarthasr@amazon.com

Kathy Wu*
Amazon
rhaow@amazon.com

Abstract

State-of-the-art performance has been achieved in recent years on tasks such as search, recommendation and classification using Visuo-Lingual Multi-Modal models. While the pretrained Vision-Language models like Contrastive Language-Image Pre-training (CLIP) have achieved promising zero-shot performance on several generalized tasks by learning vision-language concepts in a common space, the natural hierarchical relationship between them remains unexplored. In this work we propose HyperVLM: a hyperbolic Poincaré geometry based vision-language model that learns joint text-image representation considering the hierarchical relation between the two. We compare the performance of HyperVLM with CLIP model for zero-shot image classification and retrieval tasks to demonstrate the efficacy of the proposed method. We also demonstrate the effectiveness of proposed method for retrieval task when applied to BLIP architecture's ITC loss module. Proposed method holds immense value for recommendation and search tasks.

1. Introduction

Vision Language Models Large vision-language models like CLIP [30] and ALIGN [15] learn visual concepts from their natural language description via multi-modal contrastive learning. In contrastive learning [16], an anchor item representation is compared with a similar and a dissimilar item with the aim of bringing similar item representation together and pushing different ones away. The effectiveness [35] of these models results from their pretraining over a diverse large-scale image-text dataset sources from the web, allowing them to learn diverse concept from real world resulting in their impressive generalizability over a variety of tasks in zero-shot setting like classification and retrieval. These models assume the geometry of the higher dimensional representation space as affine Euclidean [12, 25], making it harder to capture the visual-text hierarchical concepts. The entity containing more *general* concepts should be located close to the root of the hierarchy tree than the entity encapsulating a more *specific* and complex information. Hyperbolic spaces [3, 31] are natural candidate for capturing this hierarchical information about data points as their volume grows exponentially away from the origin, against polynomial growth in case of Euclidean space. Hyperbolic space can be thought of as a continuous version of a tree with it's root at the origin. **Vision Language Hierarchy** The saying "A picture is worth a thousand words" conveys the information difference between an image and words describing them. For example, in Figure 1, the picture can be broken down into individual concepts consisting of "kitty" and "doggo", which might be transformed in different manner to generate caption, for e.g. 'my dog's innocence brings smile to my face', 'a dog and a cat having fun in field', etc. Following equivalence, a many words can be put together encapsulating complex concept to build an informative image. Injecting these inductive biases in the training of multi-modal models [30, 32] will allow them to learn a more generalizable and interpretable representation. **Hyperbolic Space Representation with *HyperVLM*** In this work we project the image-text concepts onto a Poincaré ball model of hyperbolic space while following the state of the art contrastive methodology, to help capture the hierarchical information about the image-text pair, in addition to their semantic similarity. The contribution of this work can be described as: i. We introduce HyperVLM, a Poincaré ball based hyperbolic representation model trained using ViTs and Transformer encoder based contrastive loss using RedCap dataset containing 12M image-text pairs. ii. We introduce an embedding entropy based entailment loss to enforce the hierarchy between image-text in the Poincaré space. We compare the performance of the proposed method with strong baseline CLIP and MERU to demonstrate it's competitiveness.

*Equal contribution.

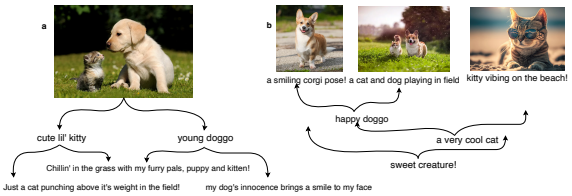

Figure 1. A picture is worth a thousand words. **Left:** Given an informative image it is possible to generate several textual concepts leveraging the visuo-lingual hierarchy. **Right:** Likewise, beginning from a simple text concept, it is possible to come up with complex visuo-lingual concepts by leveraging their hierarchical relation.

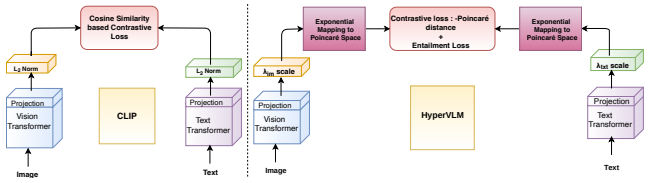

Figure 2. Overall Model Architecture. **Left:** Describes the baseline CLIP architecture based on which we have defined HyperVLM. Image and text are encoded by Vision and Text Transformers respectively before being normalized and compared for contrastive loss calculation **Right:** Describes HyperVLM architecture. It differs from CLIP in aspect that encoder output is scaled and projected onto Poincaré space before computing contrastive loss and entailment loss for optimization.

2. Related Work

The idea of hyperbolic space to better represent multimodal entities is very recent and there are few related work in this field. MERU [9] attempts to capture the image-text semantics using Lorentz hyperboloid space. However, Lorentz manifold has less representation capacity compared to Poincaré ball as described in [29]. [10] discusses application of hyperbolic space based approach to learn hierarchical information between different image samples.

3. Hyperbolic Geometry

In this section, we will walk through some key concepts of hyperbolic geometry which are relevant to our approach. Hyperbolic geometry, also known as Lobachevskian geometry is a non-Euclidean geometry where the Euclid's fifth postulate of parallels don't hold true and the space has a constant negative curvature. Hyperbolic spaces can be thought of as a continuous versions of tree data structure where the number of nodes until level h grow exponentially with the value l as $((b+1)b^h - 2)/(b-1)$ where b is the branching factor. This tree grows from origin where h is 0 and it grows in terms of nodes exponentially away from origin. Such a structural arrangement is not possible in $\mathbb{R}^2$ Euclidean space as the area and circumference of the hypercircle only grows quadratically and linearly respectively against an exponential growth in case of hyperbolic space. A brief introduction to the concept of Manifold, Curvature and hyperbolic space is discussed in supplementary material.

3.1. Poincaré Disk

A Poincaré disk is a hyperbolic geometric model in which we represent a line as an arc of a circle whose ends are perpendicular to the disk's diameter. It's a useful model that uses hyperbolic geometry to discover continuous hierarchical relations among data pairs by embedding them into n dimensional Poincaré hypersphere. Mathematically, we can define an n-dimensional Poincaré ball in constant negative curvature value of $K = -1$ as: $\mathbb{P}^n_{K=-1} = \{x \in \mathbb{R}^n : ||x||^2 < 1\}$ (1) where $||.||$ represents the Euclidean norm of a data point. The metric tensor for a Poincaré ball is represented as $g_x^{\mathbb{P}_{K=-1}} = (\gamma_x^{K=-1})^2 g_x^{\mathbb{E}}$ where $\gamma_x^{K=-1} = \frac{1}{1-||x||^2}$ is the conformity factor and $g_x^{\mathbb{E}}$ is the metric tensor for Euclidean space represented as $g_x^{\mathbb{E}} = diag([1, 1, ...1])$. The distance $d_h(p_1, p_2)$ between two samples p_1 and p_2 in the Poincaré space $\mathbb{P}^n_{K=-k}$ is calculated as:

$$d_h(p_1, p_2) = \frac{2}{\sqrt{k}} \tanh^{-1}(\sqrt{k}\|(-p_1) \oplus_k p_2\|_2) \tag{2}$$

Where $||.||$ represents the Euclidean norm of a data point. We map Euclidean feature into hyperbolic Poincaré ball manifold via $h_i = exp_0^{K=-1}(x_i^{Euc})$ where h_i represents the transformed x_i value in the hyperbolic space. The exponential map value exp_x^k for a vector p in a space having curvature value K is calculated as:

$$exp_x^K(p) = x \oplus_K \left(tanh\left(\frac{\sqrt{-K}\gamma_x^K ||p||}{2} \right) \frac{p}{\sqrt{-K}||p||} \right) \tag{3}$$

To reverse map a vector p from hyperbolic space of curvature value K to Euclidean space, we apply logarithmic mapping as following:

$$log_x^K(p) = \frac{2}{\sqrt{-K}\gamma_x^K} arctanh\left(\sqrt{-K}||v||\right) \frac{v}{||v||} \tag{4}$$

Where v is calculated as $-x \oplus_K p$ and $\oplus_K$ represents the Möbius addition defined as follow:

$$x \oplus_K y = \frac{\left(1 - 2K\langle x, y\rangle - K||y||^2\right) x + \left(1 + K||x||^2\right) y}{1 - 2K\langle x, y\rangle + K^2||x||^2||y||^2} \tag{5}$$

Where $\langle x, y\rangle$ represents the inner product between x and y in hyperbolic space.

4. Methodology

In this section we discuss the learning objective and modelling details of HyperVLM to learn the hierarchy aware representations for input text and images. HyperVLM is based on CLIP methodology consisting of a vision transformer based image encoder and a text transformer based text encoder using byte pair encoding. Both encoders generate image and text representations for input image and text respectively, which are then passed into a projection layer to obtain embeddings of a fixed size n. Additionally, we:

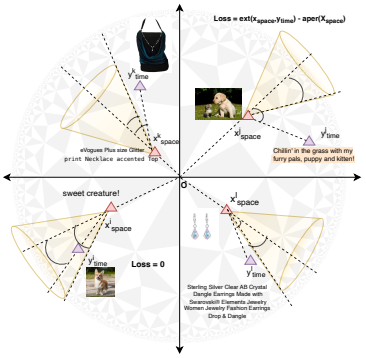

Figure 3. Entailment Cone (projection from Poincaré space on Euclidean Space). Loss pushes y_{time} embedding inside an entailment cone projected by embedding x and is defined as the difference between exterior angle $\angle OXY$, and half aperture of the cone. Loss is zero if the y_{time} is already inside the cone. Indices i and j in superscripts represent two different instances of image-text pairs.

Transfer of embeddings onto the Poincaré Space While training, the image and text samples are passed to ViT and Text Transformer encoders respectively followed by a projection layer as shown in Figure 2. This is followed by transformation of the embeddings (ν_{im}, ν_{txt}) from Euclidean geometry to hyperbolic Poincaré geometry as (h_{im}, h_{txt}) following the eq. 3 w.r.t the origin.

Numerical Overflow Prevention Since transfer from Euclidean space to hyperbolic space to calculate (h_{im}, h_{txt}) requires an exponential operation, the norm of embeddings changes from order of $\sqrt{n}$ to $e^{\sqrt{n}}$, potentially causing numerical overflow. To fix this, embedding scaling is applied before exponential mapping via two learnable parameters λ_{im} and λ_{txt} initialized to $1/\sqrt{n}$ to prevent the norm of the embedding from numerical overflow in the Poincaré space.

Training Objectives Our training objective is to enforce semantic similarity and structural partial order relation between given image-text pairs to improve the generalization capability of vision-language models. To this end, we optimize for image-text contrastive loss and entailment loss.

4.1. Contrastive Loss

We have implemented same multi-class N-pair version of the contrastive loss as used in CLIP [30] with an important difference that we calculate the similarity via distances in Poincaré space from eq.3 instead of cosine similarity. For a given batch size N we use the negative Poincaré space distance to compute contrastive loss between 1 positive and $N-1$ negative pair per image and per text. The average of image wise and text wise loss is used as overall contrastive loss $\mathcal{L}_{cont}$ to enforce image-text semantic similarity.

4.2. Entailment Loss

We apply an additional entailment loss from [9] with modification to enforce partial order relationship between image-text pairs. In [9], the assumption is that text always entails the image within the entailment cone. In contrast, we adopt an entropy based strategy to determine correct entailment order between text and image per instance. In Physics, the structure of space-time is knitted together by the causal connections represented by the causal graph, the analog of entailment cone. An entailment cone is essentially a structure representing the "time evolution" from a particular initial condition [37]. Keeping this view in perspective and given that image-text embeddings from respective transformers are learned in same latent space, we can determine the relative position in entailment cone comparing the entropy of embeddings with the assumption that entropy increases with evolution of time along the entailment cone. For a given image-text pair, the simpler concept with lower entropy should be entailing more complex concept with higher entropy with time. We calculate the information entropy [33] of embeddings as: $H(x_{emb}) = -\sum_{i=1}^{n} x_i \log_2 x_i$ where H is the entropy of embedding x_{emb} and x_i represents the content of size n embedding for i^{th}dimension. We define $x = x_{img}$, the image embedding if $H(x_{img}) < H(x_{txt})$ else, $x = x_{txt}$. Similarly define $y = x_{txt}$, the image embedding if $H(x_{img}) < H(x_{txt})$ else, $x = x_{txt}$. Please refer supplementary material for more theoretical insights.
Figure 3 gives an overview of the entailment loss as projected in Euclidean space. Exterior angle $\angle Oxy$ is defined as:

$$ext(\angle Oxy) = arcos\left(\frac{\langle x,y\rangle\left(1+||x||^2\right) - ||x||^2\left(\left(1+||y||^2\right)\right)}{||x||.||x-y||\sqrt{1+||x||^2||y||^2-2\langle x,y\rangle}}\right) \tag{6}$$

While the aperture of the entailment cone is defined as:

$$aper(x) = arcsin\left(K\frac{1-||x||^2}{||x||}\right) \tag{7}$$

We calculate the entailment loss as: $\mathcal{L}_{entail}(x,y) = max(0, ext(\angle Oxy) - aper(x)) - \lambda_{reg} ext(\angle Oxy)$ (8) where λ_{reg} is the regularization coefficient. Hence, the overall loss to be optimized becomes $\mathcal{L} = \mathcal{L}_{cont} + \lambda \mathcal{L}_{entail}$ where λ is entailment regularization factor.

5. Experiments

To establish the competitiveness of Poincaré hyperbolic representations of HyperVLM compared to Euclidean representations obtained from CLIP-style models, we compare zero shot classification and retreival performances of HyperVLM, MERU and CLIP. We train HyperVLM on public RedCaps dataset [8] consisting of 12M image-text pairs for 120k iterations 8xV100 GPUs. **Model** We use different size versions of Vision Transformers (S/B/L) as vision encoder using patch size of 16, freezing the positional encoding layer of the model. Text encoder is same as that of CLIP with 12 layer 512 dimensional Transformer with 77 maximum length byte pair encoding. Poincaré ball of 512 dimensions and learnable curvature K is used for Poincaré

		CIFAR-10[17]	CIFAR-100[17]	CUB[38]	SUN397[39]	Aircraft[22]	DTD[5]	Pets[28]	Flowers[27]	STL-10[6]	EuroSAT[14]	RESISC45[4]	Country211[30]	MNIST[18]	PCAM[36]	SST2[30]
ViT-S/16	**CLIP**	**60.1**	24.4	33.8	27.5	1.4	15.0	**73.7**	47.0	88.2	18.6	31.4	**5.2**	10.0	50.2	50.1
	MERU	52.0	24.7	33.7	**28.0**	1.3	16.2	72.3	**49.2**	**91.1**	30.4	32.0	4.8	7.5	51.0	50.0
	HyperVLM	53.6	**27.7**	**35.1**	27.6	**1.6**	**17.6**	71.9	47.9	90.9	**30.8**	**32.1**	5.1	**10.4**	**53.8**	**50.8**
ViT-B/16	**CLIP**	65.5	33.4	33.3	29.8	1.4	17.0	77.9	50.9	92.2	25.6	31.0	**5.8**	10.4	**54.1**	**51.5**
	MERU	67.7	32.7	34.8	30.9	1.7	17.2	**79.3**	**52.1**	**92.5**	30.2	**34.5**	5.6	**13.0**	49.8	49.9
	HyperVLM	**70.4**	**35.4**	**34.9**	**31.3**	**2.1**	**17.9**	78.5	51.3	91.9	**31.7**	33.5	5.5	12.1	49.6	50.0
ViT-L/16	**CLIP**	72.0	36.4	36.3	32.0	1.1	16.5	78.8	48.6	93.7	26.7	35.4	6.1	**14.8**	51.2	**51.1**
	MERU	68.7	35.5	37.2	33.0	2.2	17.2	80.0	**52.1**	93.7	**28.1**	36.5	6.2	11.8	52.7	49.3
	HyperVLM	**74.3**	**38.8**	**37.5**	**33.3**	**2.6**	**18.5**	**80.1**	51.3	**93.8**	27.9	**37.2**	**6.5**	12.0	**55.7**	50.0

Table 1. Comparison of Proposed Method HyperVLM vs Baseline Methods on different datasets. The metrics in color represent the best performance metric for particular dataset. We observe that HyperVLM outperforms all methods in 13 out of 18 datasets.

		$text \rightarrow image$		$image \rightarrow text$	
		R5	R10	R5	R10
ViT-S/16	**CLIP**	29.9	40.1	37.5	48.1
	MERU	**30.5**	**40.9**	39.0	50.5
	HyperVLM	**30.5**	40.2	**40.4**	**50.7**
ViT-B/16	**CLIP**	32.9	43.3	41.4	52.7
	MERU	33.2	**44.0**	41.8	52.9
	HyperVLM	**33.3**	43.7	**42.1**	**53.4**
ViT-L/16	**CLIP**	31.7	42.2	40.6	51.3
	MERU	**32.6**	**43.0**	41.9	53.3
	HyperVLM	**32.6**	42.7	**43.2**	**53.8**

Table 2. Zero Shot Image and Text Retrieval on COCO Dataset. Metric in color represent best performance for the task.

space transformation post embedding scaling. **Optimizer** We use AdamW Optimizer [21] with weight decay of 0.2 and (β_1, β_2) = (0.9, 0.98). Weight decay is disabled for all gains, biases, and learnable scalars. model is trained for 120K iterations with batch size 1024 ($\approx$ 10 epochs). The maximum learning rate is 5×10^{-4}, which increases linearly for first 4K iterations, followed by cosine decay to 0 [20]. We evaluate the performance the HyperVLM with CLIP and MERU on 18 datasets for zero shot classification and on COCO dataset for retrieval task.

Additionally, we evaluate image-text and text-image retrieval accuracy using BLIP [19] architecture by implementing the ITC loss calculation module in Poincarè hyperbolic space with entropy based image-text entailment order and we compare it with Euclidean space BLIP architecture for COCO dataset in Table 5 in **supplementary material**.

5.1. Results

From Table 1, we compare HyperVLM's performance for zero shot classification and observe it performing better than the Euclidean space CLIP for 14 out of 18 datasets and than Lorentz model based MERU for 15 out of 18 datasets and on 13 out of 18 datasets overall. Comparing Top N retrieval recall for COCO dataset in Table 2 we see that HyperVLM performs better than CLIP on 4 out of 4 tasks while it performs better than MERU on 3 out of 4 tasks. Overall, HyperVLM performs better than all methods on 3 out of 4 tasks demonstrating the competitveness of the proposed method. Ablation results for regularization terms λ_{reg} and λ will be shared in supplementary material. In Table 5 we observe that the proposed method outperforms the euclidean space representation in BLIP architecture for image-text and text-image retrieval task.

6. Discussion

We obtain better performance for HyperVLM over Euclidean space CLIP owing to hyperbolic nature of Poincaré geometry which allows the capture of partial order relation between image and text, in addition to the semantic relation for learning representation. The incremental benefit over MERU can be attributed to 2 reasons: 1. Use of Poincaré space over Lorentz space: As per the work done in [24] Poincare geometry has a relatively larger capacity than the Lorentz model for correctly representing points 2. Entailment loss based on entropy derived relative hierarchy between image and text at instance level.

7. Conclusion

In this work we discussed Poincaré geometry based large scale image-text model that learns image-text partial order hierarchical relation, in order to capturing their semantic similarity. The main contribution of this work can be summarised as: 1.Poincaré Hyperbolic Geometry based Image-Text model capturing image-text semantics along with their hierarchical-relation. 2. Embedding entropy based method to decide the entailment order of image-text when enforcing partial order relationship. We demonstrate the efficacy of the proposed method via experiments comparing accuracy for zero shot classification and recall for zero shot retrieval.

References

[1] Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016. 2

[2] Guillaumin M. Van Gool L. Bossard, L. Food-101 – mining discriminative components with random forests. *In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8694. Springer, Cham. https://doi.org/10.1007/978-3-319-10599-4*$_2$9, 2014. 3

[3] Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076, 2020. 1

[4] Han J. Cheng, G. and X Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017. 4, 3

[5] Maji S. Kokkinos I. Mohamed S. Cimpoi, M. and A. Vedaldi. Describing textures in the wild. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4, 3

[6] Ng A. Coates, A. and H Lee. An analysis of single layer networks in unsupervised feature learning. *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. 4, 3

[7] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1999. 2

[8] 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. 3

[9] 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. 2, 3

[10] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning, 2022. 2

[11] Fergus R. Fei-Fei, L. and Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVPR Workshop*, 2004. 3

[12] Nóra Frankl, Andrey Kupavskii, and Konrad J Swanepoel. Embedding graphs in euclidean space. *Journal of Combinatorial Theory, Series A*, 171:105146, 2020. 1

[13] Mikhael Gromov. Hyperbolic groups. *Essays in group theory*, pages 75–263, 1987. 2

[14] Bischke B. Dengel A. R. Helber, P. and D Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 4, 3

[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. 1

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 1

[17] A Krizhevsky. Learning multiple layers of features from tiny images. *URL https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.*, 2009. 4, 3

[18] Cortes C. LeCun, Y. and C Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2010. 4, 3

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 4

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 4

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 4

[22] Rahtu E. Kannala J. Blaschko M. B. Maji, S. and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4, 3

[23] Jiří Matoušek. Geometric discrepancy: An illustrated guide. *Algorithms and Combinatorics*, 18, 1999. 2

[24] Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning, 2023. 4, 1

[25] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 1

[26] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[27] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. *ICVGIP*, 2008. 4, 3

[28] Vedaldi A. Zisserman A. Parkhi, O. and C. V. Jawahar. Cats and dogs. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 4, 3

[29] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44(12):10023–10044, 2021. 2, 1

[30] 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, 3, 4

[31] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018. 1

[32] 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

[33] C.E. Shannon. Claude shannon introduced the concept of information entropy in his 1948 paper, "a mathematical theory of communication. *A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x*, 1948. 3

[34] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. 2

[35] 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. 1

[36] Linmans J. Winkens J. Cohen T. Veeling, B. S. and M. Welling. Cnns for digital pathology. *arXiv preprint arXiv:1806.03962*, 2018. 4, 3

[37] Athanasios Vlontzos, Henrique Bergallo Rocha, Daniel Rueckert, and Bernhard Kainz. Causal future prediction in a minkowski space-time. *arXiv preprint arXiv:2008.09154*, 2020. 3

[38] Branson S. Welinder P. Perona P. Wah, C. and S. J. Belongie. The caltech-ucsd birds-200-2011 dataset. 4, 3

[39] Hays J. Ehinger K. A. Oliva A. Xiao, J. and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 4, 3

[40] Puigcerver J. Kolesnikov A. Ruyssen P. Riquelme C. Lucic M. Djolonga J. Pinto A. S. Neumann M. Dosovitskiy A. Beyer L. Bachem O. Tschannen M. Michalski M. Bousquet O. Gelly S. and Houlsby N. Zhai, X. A large-scale study of representation learning with the visual task adaptation benchmark. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

HyperVLM: Hyperbolic Space Guided Vision Language Modeling for Hierarchical Multi-Modal Understanding

Supplementary Material

A. Hyperbolic Geometry

A.1. Manifold

A manifold is a topological space that locally resembles Euclidean space. A precise definition from topology is that an n-dimensional manifold M is a topological Hausdorff space with a countable base which is locally homeomorphic to $\mathbb{R}^n$. For every point p in M, there exists an open neighbourhood U and a homeomorphism h: $U \rightarrow V$ which maps the set U onto an open set $V \subset \mathbf{R}^n$. Thus the point is either an isolated point (when $n = 0$), or it has a neighborhood which is homeomorphic to the open ball

$$\mathbf{D}^n = \{(x_1, x_2, ..., x_n) \in \mathbb{R}^n : x_1^2 + x_2^2 + ... + x_n^2 < 1\}$$

Riemannian manifold refers to real and smooth manifold with Riemannian tensor, which is metric tensor and can be defined by a family a inner products as follow:
Suppose p is a point on the curve of manifold M with $p \in M$ and denote the tangent space by $T_p(M) \in \mathbb{R}^n$, for any two tangent vectors $X(p)$ and $Y(p)$, $\quad q : T_pM \times T_pM \rightarrow \mathbf{R}$ defines a smooth function for the point $p \in M$

A.2. Curvature

In simple terms, curvature of a curve is its measure of deviation from a straight line and that of a surface is the measure of its deviation from a plane. In terms of space, a curved space refers to spatial geometry which shows some finite curvature w.r.t a plane surface.

A.3. Hyperbolic Space

Hyperbolic n-space, denoted $\mathbb{H}^n$, is the unique simply connected, n-dimensional Riemannian manifold which has constantly negative sectional curvature. Let (H, d) denote a metric space, it is said to be a hyperbolic metric space if the following conditions are satisfied: 1) for any points $p, q \in H$ that are the endpoints of a unique metric segment is isometric to the interval of real line $[0, d(p, q))$ 2) let the unique point $t = \alpha p \oplus (1 - \alpha)q$ where $\alpha \in [0, 1]$, it satisfies $dpt = 1 - \alpha dpq, dtq = \alpha dpq$ 3) for all $x, y, p, y \in H$ and $\beta \in [0, 1]$ we have $d\beta x \oplus 1 - \beta p, \beta y \oplus 1 - \beta q \leq \beta dxy + 1 - \beta dpq$

B. Ablation Study

In Table 3 and 4, we observe the difference between the proposed Poincaré embedding with the entropy inferred text-image order entailment loss compared with Poincaré embedding without entropy inferred text-image order entailment where text always entails image in entailment loss. The zero shot classification and retrieval experiments have been conducted for ViT S/16 model for 120000 iterations using RedCaps dataset, same optimizer and learning rate as the proposed method. As can be observed, the addition of entailment leads to improvement in 14 out of 18 datasets in zero shot classification setting while improving performance in all 4 zero shot retrieval tasks for COCO dataset.

In Table 5 we run the ablation study for λ_{reg} and λ by training ViT-S/16 HyperVLM model for 1 epoch (6k iterations) and compare the average zero shot retrieval accuracy for COCO dataset. We find that $\lambda_{reg} = 0.1$ and $\lambda = 0.1$ provides the best performance and was chosen as the value for our experiments. The entailment loss described in eq. 9 depends on λ_{reg} and λ for calculation of the overall entailment loss.

C. Advantages of Poincaré Ball over Lorentz Hyperboloid

The choice of Poincaré ball model over Lorentz hyperboloid for vision-language representation offers several theoretical and practical advantages [29]. In the Poincaré ball model, the representation capacity scales more effectively with dimension compared to the Lorentz model $\mathbb{L}^n = \{x \in \mathbb{R}^{n+1} : \langle x, x \rangle_{\mathbb{L}} = -1, x_0 > 0\}$ [31]. This superior scaling property emerges from three key aspects:
Geometric Properties The Poincaré ball model provides conformal mapping that preserves angles, leading to more stable optimization. The metric tensor at point x is given by $g_x^{\mathbb{D}} = (\frac{2}{1-\|x\|^2})^2 g^E$ where $g_x^{\mathbb{E}}$ is the metric tensor for euclidean space represented as $g_x^{\mathbb{E}} = diag([1, 1, ...1])$, which naturally adapts to the hierarchical structure of the data [26]. In contrast, the Lorentz model's metric tensor $g_x^{\mathbb{L}} = \text{diag}(-1, 1, ..., 1)$ remains constant, potentially limiting its adaptability to complex hierarchical relationships.
Numerical Stability The bounded nature of the Poincaré ball ($\|x\| < 1$) provides inherent numerical stability during optimization [24]. The gradients in Poincaré space are naturally scaled by the conformal factor, preventing exponential explosion or vanishing issues common in Lorentz space where coordinates can grow unboundedly. This leads to more stable training dynamics:$\|\nabla_{\mathbb{D}} f(x)\| \leq \frac{2}{1-\|x\|^2} \|\nabla_E f(x)\|$
Representation Efficiency For hierarchical structures of depth d and branching factor b, the Poincaré ball achieves distortion $O(\log(d))$ compared to $O(\sqrt{d})$ in Lorentz space [31]. This leads to more efficient embedding of hi-

erarchical structures, particularly for deep hierarchies: $\text{Distortion}_{\mathbb{D}}(T) < c\log(d) << c\sqrt{d} < \text{Distortion}_{\mathbb{L}}(T)$ where T represents a tree structure and c is a constant. This efficiency translates to 1. Better preservation of hierarchical relationships, 2. More accurate representation of fine-grained semantic differences and 3. Improved gradient flow during optimization

These advantages make the Poincaré ball particularly suitable for vision-language modeling where preserving both hierarchical structure and semantic similarity is crucial.

D. Theoretical Insight

Motivation Vision-language representation learning inherently involves hierarchical structures in both modalities. For instance, visual concepts form natural hierarchies (e.g., animal → mammal → dog → breed), and textual descriptions similarly exhibit hierarchical relationships. Traditional Euclidean spaces, with their polynomial volume growth [23], are suboptimal for representing such hierarchical structures. In contrast, hyperbolic geometry, characterized by exponential volume growth [13], naturally accommodates tree-like hierarchical structures.

Information-Theoretic Hierarchy and Compositional Entailment

Information Content and Hierarchical Structure in Shared Space The fundamental connection between embedding complexity and hierarchical relationships can be established through Shannon's information theory [34]. For embeddings of different modalities projected into a common space through encoders $f_{\theta_{img}}$ and $f_{\theta_{txt}}$, the shared representation ensures that information content comparison is meaningful. This is because:

1. The encoders map inputs to a common manifold where geometric and information-theoretic properties are preserved
2. The contrastive learning objective ensures semantic alignment in this shared space
3. The hyperbolic nature of the space maintains consistent hierarchical relationships across modalities

For an embedding vector x in this shared space, the information entropy:

$$H(x) = -\sum_{i=1}^{n} p_i \log p_i, \quad p_i = \frac{|x_i|}{\sum_{j=1}^{n} |x_j|}$$

represents the evolved information content of the concept in the common space [7]. This measure provides theoretical justification for hierarchical relationships because:

1. **Common Information Currency**: The shared space acts as a "common currency" for information across modalities, making entropy comparisons meaningful [1]
2. **Information Evolution**: The embedding entropy reflects how information evolves from general to specific concepts in the shared manifold:

$$H_{shared}(x) = H_{modal}(x) + I_{alignment}(x)$$

where $I_{alignment}$ represents the additional information gained through cross-modal alignment
3. **Information Content Principle**: More specific concepts require additional information to be fully specified beyond their parent concepts, leading to higher entropy values:

$$\Delta H = H(child) - H(parent) \geq 0$$

		text → *image*		*image* → *text*	
		R5	R10	R5	R10
ViT-S/16	**Poincaré**	30.1	**40.2**	39.0	50.2
	HyperVLM	**30.5**	**40.2**	**40.4**	**50.7**

Table 3. Zero Shot Image and Text Retrieval on COCO Dataset. Metric in color represent best performance for the task. Row corresponding to Poincaré represents the case where no entropy derived entailment order is enforced in the entailment loss and we assume that text always entail the image as assumed in MERU. The row corresponding to HyperVLM represent the case where embedding entropy derived image-text entailment order is applied in entailment loss.

		λ				
		0	**0.01**	**0.1**	**0.5**	**1**
λ_{reg}	**0**	20.2	17.5	18.6	16.5	18.7
	0.01	20.2	15.4	22.1	16.7	16.0
	0.1	20.2	21.1	**22.3**	18.9	19.0
	0.5	20.2	19.2	18.6	16.8	15.7
	1	20.2	18.6	18.1	15.4	19.5

Table 4. To select proper values of λ and λ_{reg} we run a grid search for different values and compare the average of average zero shot retrieval accuracy for different retrieval tasks for COCO dataset and zero shot classification accuracy for CIFAR 100 dataset by training ViT-S/16 model for 1 epoch (6K iterations). We find the best performance at $\lambda = 0.1$ and $\lambda_{reg} = 0.1$. The best performance metric in color

Model	Text				Image			
	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean
BLIP	77.60	94.10	97.20	89.63	61.00	84.50	90.70	78.73
HyperVLM(BLIP)	**80.24**	**94.52**	**97.32**	**90.69**	**62.32**	**85.12**	**91.32**	**79.58**

Table 5. Comparison of HyperVLM(BLIP) and BLIP Models for COCO Text-Image and Image-Text Retrieval

		Food-101[2]	CIFAR-10[17]	CIFAR-100[17]	CUB[38]	SUN397[39]	Aircraft[22]	DTD[5]	Pets[28]	Caltech-101[11]	Flowers[27]	STL-10[6]	EuroSAT[14]	RESISC45[4]	Country211[30]	MNIST[18]	CLEVR[40]	PCAM[36]	SST2[30]
ViT-S/16	**Poincaré**	74.9	**55.3**	27.5	34.1	**28.3**	1.5	16.4	**72.9**	60.0	**48.4**	90.7	28.3	30.6	4.9	8.3	14.4	48.9	50.2
	HyperVLM	**75.1**	53.6	**27.7**	**35.1**	27.6	**1.6**	**17.6**	71.9	**62.1**	47.9	**90.9**	**30.8**	**32.1**	**5.1**	**10.4**	**14.8**	**53.8**	**50.8**

Table 6. Comparison of proposed method HyperVLM implementing entropy inferred image-text entailment order, with HyperVLM without entropy inferred image-text entailment order where text always entail image on different datasets. The metrics in color represent best performance metric for particular dataset. We observe that HyperVLM outperforms the Poincaré method where we always assume text to be entailing image, in 14 out of 18 datasets.