Global and Local Entailment Learning for Natural World Imagery

A. Proof for Lemma 1

Lemma 1 states that finer-grained concepts are progressively projected: 1) away from the entailment root and 2) into smaller subregions in a transitivity-enforced entailment. We begin with the definition of distance in an entailment configuration:

$$\Xi(T_j^i, T_l^k) = \arccos\left(\frac{\langle (T_j^i - T_0), (T_l^k - T_j^i)\rangle}{||T_j^i - T_0||.||T_l^k - T_j^i||}\right) \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is an inner product between the embeddings. The distance between two textual embeddings are computed with respect to the entailment root. In an entailment configuration with transitivity, the following property is satisfied [2] between a parent and its child:

$$\Xi(T_j^i, T_{j+1}^i) \leq \psi(T_j^i) \leq \pi/2 \tag{2}$$

This means that $\Xi(T_j^i, T_{j+1}^i) \in [0, \pi/2]$. It follows:

$$0 \leq \arccos\left(\frac{\langle (T_j^i - T_0), (T_{j+1}^i - T_j^i)\rangle}{||T_j^i - T_0||.||T_{j+1}^i - T_j^i||}\right) \leq \frac{\pi}{2} \tag{3}$$

$$0 \leq \left(\frac{\langle (T_j^i - T_0), (T_{j+1}^i - T_j^i)\rangle}{||T_j^i - T_0||.||T_{j+1}^i - T_j^i||}\right) \leq 1 \tag{4}$$

Simplifying the above equation, we get the following expressions:

$$0 \leq \langle (T_j^i - T_0), (T_{j+1}^i - T_j^i)\rangle \tag{5}$$

$$0 \leq \langle T_j^i, T_{j+1}^i\rangle + \langle T_j^i, T_0\rangle - \langle T_{j+1}^i, T_0\rangle - \langle T_j^i, T_j^i\rangle \tag{6}$$

Case 1: ***Radial Geometry*** In radial geometry, all textual embeddings lie on a unit hypersphere. As a result, the inner product between any two embeddings can never exceed the value of 1. As a result, we get the following expressions:

$$1 + \langle T_{j+1}^i, T_0\rangle - \langle T_j^i, T_0\rangle \leq \langle T_j^i, T_{j+1}^i\rangle \tag{7}$$

$$1 + \langle T_{j+1}^i, T_0\rangle - \langle T_j^i, T_0\rangle \leq 1 \tag{8}$$

$$\langle T_{j+1}^i, T_0\rangle - \langle T_j^i, T_0\rangle \leq 0 \tag{9}$$

$$\boxed{\langle T_{j+1}^i, T_0\rangle \leq \langle T_j^i, T_0\rangle} \tag{10}$$

As can be seen from equation 10, the cosine similarity between T_j^i and T_0 is always greater than that of T_{j+1}^i and T_0. This means the distance of T_{j+1}^i and T_0 is always greater than that of T_j^i and T_0.

Case 2: ***Euclidean Geometry*** In Euclidean geometry, the entailment root is considered to be the origin (a vector of zeros). This means $T_0 = 0$. Textual embeddings in this geometry are unnormalized and can have arbitrary norms. The distance of textual embeddings in this geometry is simply the L-2 norm. Using equation 6, we get the following expressions:

$$\langle T_j^i, T_j^i\rangle \leq \langle T_j^i, T_{j+1}^i\rangle \tag{11}$$

$$||T_j^i|| \leq ||T_{j+1}^i||.\cos\theta \tag{12}$$

$$\boxed{||T_j^i|| \leq ||T_{j+1}^i||} \tag{13}$$

In Euclidean geometry, the norms of the embeddings increase with increasing ranks.

In both geometries, we can conclude that *the distance of textual embeddings monotonically increase with increasing ranks*. This leads to the following expression for the distance of an embedding from the root:

$$r(T_{j+1}^i, T_0) \geq r(T_j^i, T_0) \tag{14}$$

$$r(T_j^i, T_0) = f(i, j; T_0) \tag{15}$$

where f is a monotonically increasing function with respect to the rank j and r is the distance function. Now let, the aperture angle of a cone defined at each textual embedding have the following expression (as done in [2]):

$$\psi(T_j^i) \propto \arcsin(1/r(T_j^i, T_0)) \tag{16}$$

The above expression establishes the relation between the aperture angle of a cone defined at some textual embedding T_j^i and its semantic granularity. From the expression, it is evident that the aperture angle monotonically decreases with increasing j which defines its semantic granularity. Hence, we can conclude that fine-grained concepts are progressively projected into smaller subregions.

The proof is complete.

Kingdom	# Samples	Phylum	Class	Family	Order	Genus	Species	Average
Fungi	3410	68.09	38.24	31.40	23.78	63.84	73.05	49.73
Plantae	42710	92.17	37.01	15.82	30.35	66.34	74.45	52.69
Animalia	53880	84.73	72.86	73.40	55.68	68.25	70.41	70.89

Table 1. Zero-shot classification performance for each distinct *kingdom* class present in the iNaturalist-2021 dataset.

B. Implementation Details

All our models are based on the ViT-B/16 architecture and use the OpenCLIP implementation in PyTorch. For training, we use a learning rate of $1e^{-7}$ with OneCycleLR scheduler and the Adam optimizer. We use a batch size of 32 and accumulate gradient batches of 2. We use 2 NVIDIA H100 GPUs with the Distributed Data Parallel training strategy. We train for a single epoch. We found training for larger number of epochs hindered the performance of the model especially in the fine-grained taxonomic ranks like *genus* and *species*. We fixed the value of β to 0.1 and 1.0 for the model trained from BioCLIP's [4] and OpenCLIP's [3] checkpoints respectively. For our global entailment objective, we set the margin α to $\pi/2$.

C. Experimental Setup

Below we describe the details of the experiments done in the main paper.

Ordering of taxonomic labels. We use the same setup as Alper *et al*. [1]. We sample 50 equally spaced points from the entailment root to the closest textual embedding to a given query image in the embedding space. At each point, we retrieve a textual embedding from a database which is closest to the given image embedding. We define a radius equivalent to the distance between the points for retrieving relevant embeddings at each level. We compute the Kendall's Correlation Coefficient (τ_d) to evaluate the quality of ordinal association among the retrieved embeddings. Similarly, we compute precision and recall metric relative to the seven ranks of ground-truth taxonomic labels.

Zero-shot image classification. For each evaluation dataset, we first create a database of unique textual embeddings for each rank of the taxonomy. For a given rank, we compute the top-1 recall/accuracy metric on image to text retrieval task. Unlike the ordering task, we compute the accuracy metric for each taxonomic rank independently. From the experiments, we notice that the performance of the models does not decrease monotonically with increasing ranks of the taxonomy. In Table 1, we present kingdom-wise performance of our model. We notice that classification performance of plants especially in the *family* and *order* ranks is abnormally low. We believe this is due to highly similar traits and mislabeling of plant species in these ranks. Note that in this experiment, we create independent database of textual embeddings for each kingdom.

Image-to-image retrieval. In this experiment, we retrieve images of species with a given taxonomic label at a given rank using a query image. For an evaluation dataset, we precompute the embeddings for each of the images. Subsequently, we retrieve images by calculating the cosine similarity between the query image embedding and the precomputed image embeddings. We compute the recall metric (R@1).

UMAP visualization. We show additional UMAP visualizations of textual embeddings from the models in Figure 1.

D. Additional Ablations

In Table 2, we show the performance of our global objective function with varying margins (see equation 6 in the main paper). We see that our objective function's performance improves with increasing margins.

α	Kendall's τ_d	Precision	Recall	F1
$\pi/2$	0.991	0.162	0.467	0.241
$\pi/4$	0.990	0.152	0.467	0.229
$\pi/8$	0.990	0.154	0.470	0.232
0	0.990	0.151	0.454	0.226

Table 2. Hierarchical retrieval metrics on HierarCaps dataset with varying margins (α) in our global entailment objective.

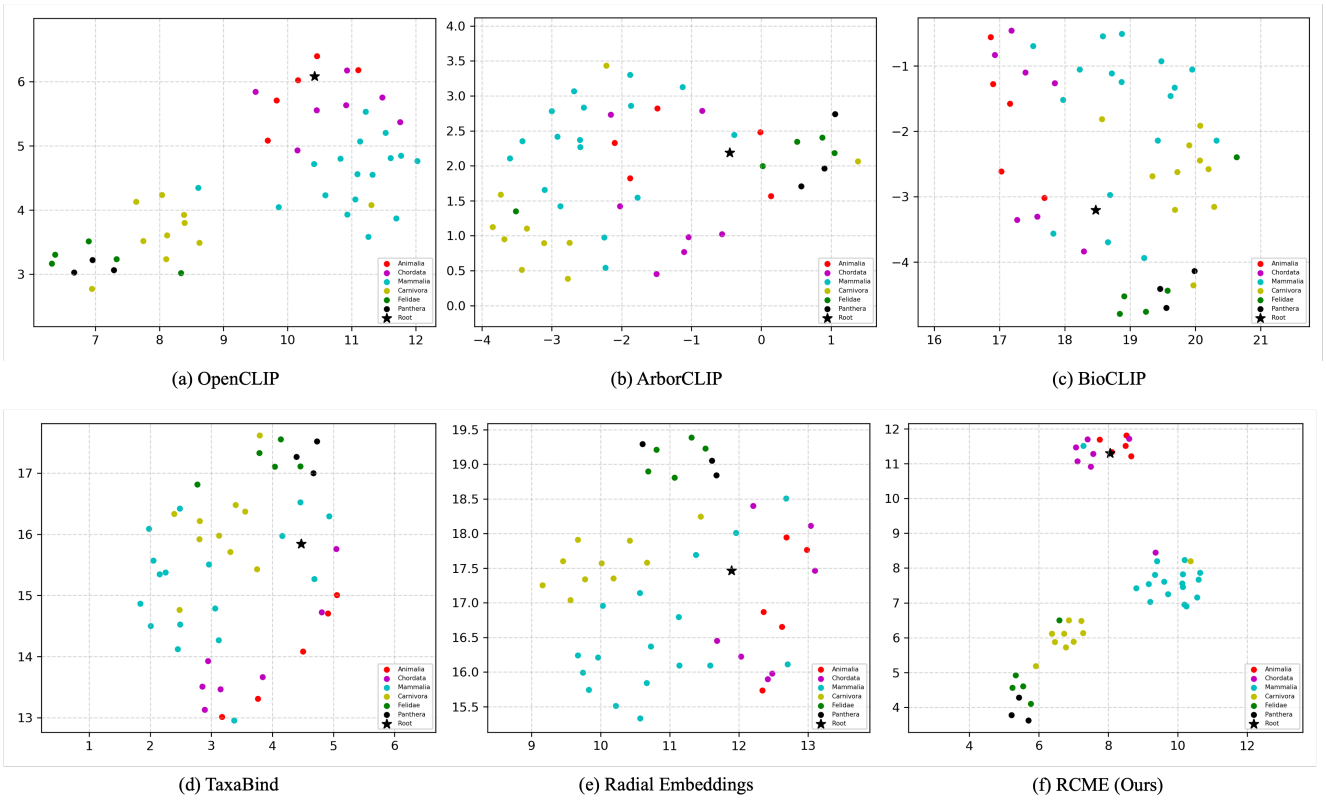
Additionally, we assess our model's performance in the ordering task by varying the number of retrieval steps in the embedding space. Tables 3 and 4 present the results. Reducing retrieval steps improves precision, but negatively affects recall. The ordering performance remains consistent, as expected.

Steps	Kendall's τ_d	Precision	Recall	F1
10	0.993	0.527	0.472	0.498
20	0.993	0.491	0.552	0.520
30	0.993	0.493	0.618	0.548
40	0.993	0.455	0.568	0.505
50	0.993	0.458	0.572	0.508

Table 3. Hierarchical retrieval metrics on iNaturalist-2021 dataset with varying number of retrieval steps.

Steps	Kendall's τ_d	Precision	Recall	F1
10	0.991	0.224	0.344	0.271
20	0.991	0.190	0.419	0.261
30	0.991	0.174	0.450	0.251
40	0.991	0.165	0.465	0.244
50	0.991	0.162	0.467	0.241

Table 4. Hierarchical retrieval metrics on HierarCaps dataset with varying number of retrieval steps.



(a) OpenCLIP (b) ArborCLIP (c) BioCLIP

(d) TaxaBind (e) Radial Embeddings (f) RCME (Ours)

Figure 1. **UMAP Visualization of Textual Embeddings**. The visualizations show our model has successfully imparted partial order in the textual embeddings.

References

[1] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. *European Conference on Computer Vision*, 2024. 2

[2] 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. 1

[3] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 2

[4] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 2