Week 14 Report - NFHM

Romouald Dombrovski

November 22, 2024

1 Time Log

What progress did you make in the last week?

- Ran Training for clusters of size 1000, 2000, 3000, 5000, 10000
- Set up eval for different cluster-sized trained UNICOM models vs Bioclip vs ViT-H-14
- Reviewed Bioclip paper
- Met with Thomas to discuss next step plans, Hypergator, Bioclip paper and diverging evaluations
- Created up to date flow chart of my evaluation script
- Went through Hypergator Coursera training course
- Set up and ssh'd into Hypergator
- Reviewed zero-shot and few-shot scripts on Bioclip repo
- Met with Thomas, Bree, Dr. Porto and Moritz to discuss current results and publication conferences

2 Abstract

Sastry S., et al. "TaxaBind: A Unified Embedding Space for Ecological Applications." arXiv:2411.00683. https://doi.org/10.48550/arXiv.2411.00683

Abstract We present TaxaBind, a unified embedding space for characterizing any species of interest. TaxaBind is a multimodal embedding space across six modalities: ground-level

images of species, geographic location, satellite image, text, audio, and environmental features, useful for solving ecological problems. To learn this joint embedding space, we leverage ground-level images of species as a binding modality. We propose multimodal patching, a technique for effectively distilling the knowledge from various modalities into the binding modality. We construct two large datasets for pretraining: iSatNat with species images and satellite images, and iSoundNat with species images and audio. Additionally, we introduce TaxaBench-8k, a diverse multimodal dataset with six paired modalities for evaluating deep learning models on ecological tasks. Experiments with TaxaBind demonstrate its strong zero-shot and emergent capabilities on a range of tasks including species classification, cross-model retrieval, and audio classification

Summary (GPT-4o) TaxaBind introduces a multimodal embedding space that integrates six modalities—ground-level images, geographic location, satellite imagery, text, audio, and environmental features—to address ecological challenges such as species classification and distribution mapping. Utilizing a novel "multimodal patching" technique, TaxaBind aligns all modalities with ground-level images, preserving unique modality-specific information while enhancing the performance of binding modality encoders. The researchers curated two datasets, iSatNat and iSoundNat, to pretrain the models, and TaxaBench-8k for evaluation, showcasing strong zero-shot and cross-modal retrieval capabilities. Experiments demonstrated TaxaBind's superiority over state-of-the-art methods like ImageBind and BioCLIP in tasks ranging from audio classification to species mapping. This unified approach has broad potential for ecology-related applications, including climate prediction and biodiversity monitoring. The study emphasizes ethical considerations and acknowledges spatial biases in datasets, highlighting the framework's role as a foundation for future ecological research.

3 Scripts and Code Blocks

There was no new scripting work done this week, most of it was running former training loops and evaluations for the remaining cluster sizes, then evaluating the models. Next week will involve learning how to work with SLURM scripting on the Hypergator

4 Visualization

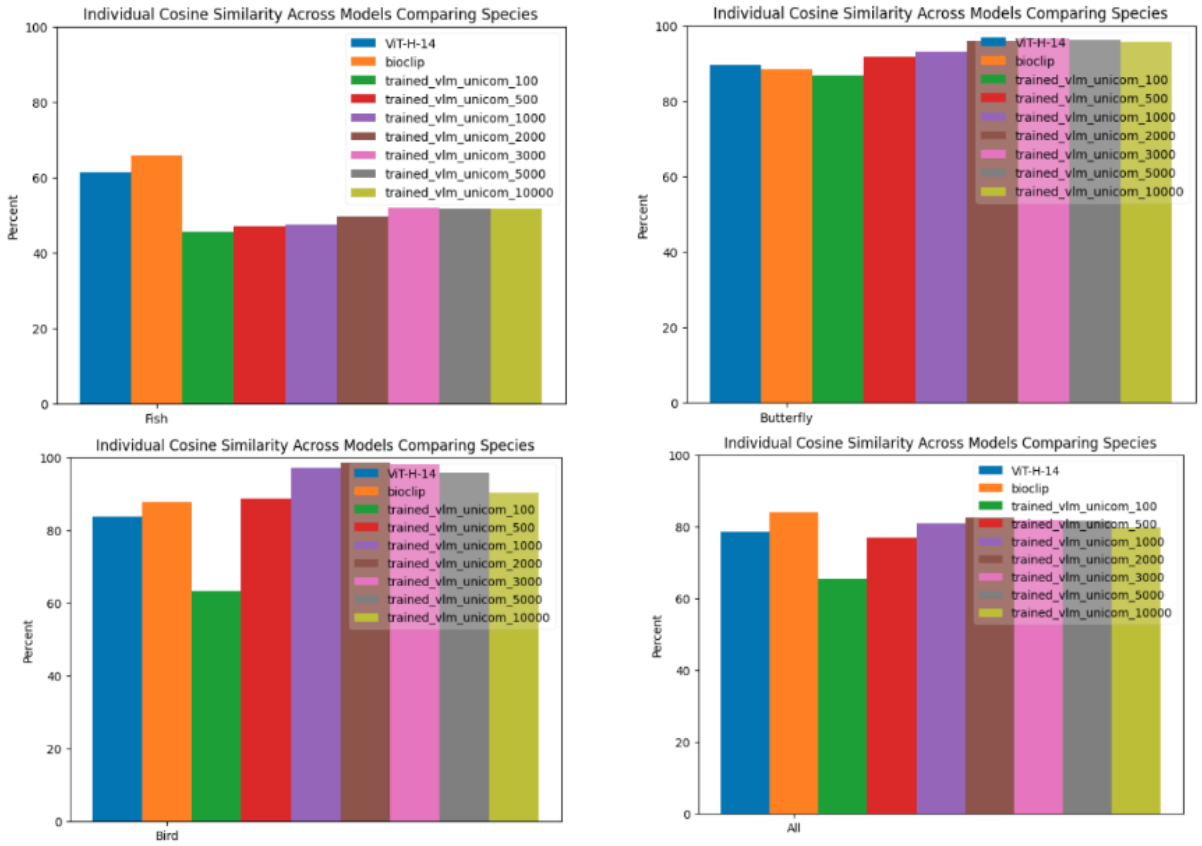

Figure 1: Cosine Similarity comparison among different models for each taxa category

So this week I finished running my training loops for all interesting cluster sizes (100, 500, 1000, 2000, 3000, 5000, 10000). This was to understand how the number of clusters affects the accuracy of the model in its representation of the data.

The VLM4Bio dataset consists of the taxa Fish, Butterfly and Birds consisting of 742 individual species, with uneven distribution of images on each. In order of even distribution the categories went Bird ¿ Butterfly ¿¿ Fish, with many of the individual classes within Fish having only 1 associated image. Figure 1 demonstrates that Fish was on average the worst performing category due to this uneven representation of classes.

Looking across the different cluster sizes used in the UNICOM training method, it seemed as though the 2000 & 3000 cluster sizes yielded the best accuracy, and came closest to Bioclip's level of accuracy. Figure two emphasizes this by directly comparing the best performing trained UNICOM models (organized by cluster size) to the pure Openclip ViT-H-14 model and the Bioclip model.

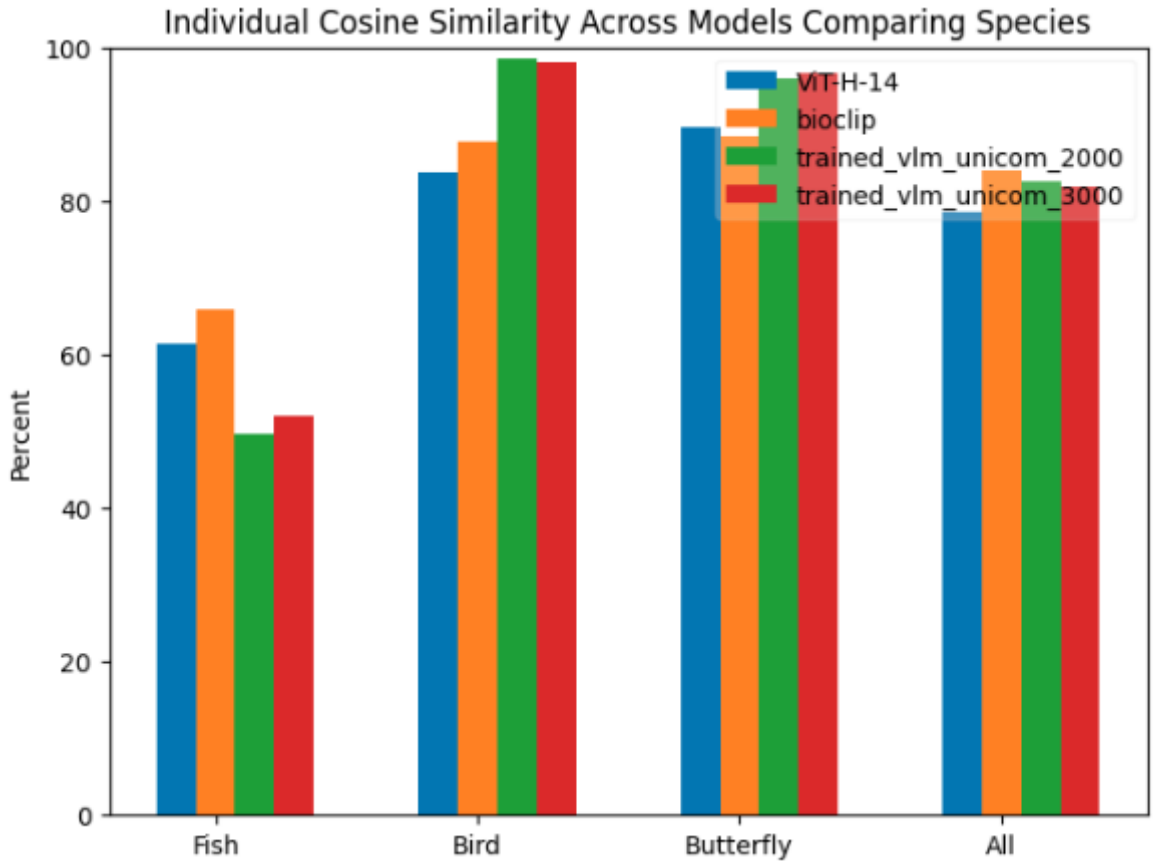

Figure 2: Cosine Similarity comparison across taxa for different models

This is highly relevant to our work, as we will be using Bioclip as a benchmark to beat with our cluster training method. The fact that the 2000/3000 cluster size models perform better fits in with our general hypothesis, that there should be several clusters per class to adequately represent the contextual information within each image without overfitting the model itself.

5 Next Week Proposal

- Learn how to write SLURM scripts
- Download VLM4Bio dataset using SLURM on Hypergator
- Meet with Thomas to discuss plans on Tree-of-Life 10M dataset download
- Rewrite Zero-shot script from Bioclip to only focus on Visual model
- Rewrite Few-shot script from Bioclip to only focus on Visual model
- Fill out milestone report for NFHM project for Vy