Week 14 Report

Thomas Deatherage
NFHM/BioCosmos
22 November 2024

Time Slot

1) What progress did you make in the last week?

- Weekly catchup
- Performance Evaluation on VLM4Bio Unicom-trained CLIP model
    - Compared evaluations on training vs testing splits
- Successfully logged into HiPerGator
- Read/watched the HiPerGator training material

2) What are you planning on working on next?

- Explore HiPerGator
- Download VLM4Bio dataset to HiPerGator storage via Slurm as a test-run for TreeOfLife

3) Is anything blocking you from getting work done?

- Nope.

Abstracts & Summaries

SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning

Abstract: Few-shot learners aim to recognize new object classes based on a small number of labeled training examples. To prevent overfitting, state-of-the-art few-shot learners use meta-learning on convolutional-network features and perform classification using a nearest-neighbor classifier. This paper studies the accuracy of nearest-neighbor baselines without meta-learning. Surprisingly, we find simple feature transformations suffice to obtain competitive few-shot learning accuracies. For example, we find that a nearest-neighbor classifier used in combination with mean-subtraction and L2-normalization outperforms prior results in three out of five settings on the miniImageNet dataset.

Summary: The paper, SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning, investigates the performance of nearest-neighbor classifiers in few-shot learning tasks.

It challenges the assumption that complex meta-learning methods are essential, proposing instead that simple feature transformations, such as mean subtraction and L2-normalization, can achieve competitive or even superior results.

Key contributions and findings include:

- Feature Transformations: The study demonstrates that centering (mean subtraction) and L2-normalization significantly improve the accuracy of nearest-neighbor classifiers.
- Competitive Performance: The proposed SimpleShot approach, leveraging these transformations, outperforms or matches state-of-the-art methods in multiple few-shot learning benchmarks like miniImageNet, tieredImageNet, and CIFAR-100.
- Baseline Importance: The paper reestablishes nearest-neighbor classification as a strong baseline for future few-shot learning research, emphasizing simplicity and efficiency.

The study validates its conclusions through experiments across various network architectures, highlighting the robustness and generalizability of its approach.

Relevance: This paper was referenced by the BioClip paper when describing few shot evaluation.

Scripts and Code Blocks

This week I compared my testing vs training datasets. For the most part, it was running code I had already written on PACE. The relevant code being the test/train dataloader:

```python
class TestDataset(Dataset):
    def __init__(self, csv_path, base_path="../Image-Captioning",
transform=None):
        df = pd.read_csv(csv_path)
        self.df = df[df['split'] == 'train'].reset_index(drop=True)
        self.df = self.df.dropna(subset=['scientific_name'])

        self.base_path = Path(base_path)
        self.transform = transform

        # Normalize scientific names and extract genus
        self.df['scientific_name'] =
```

```python
self.df['scientific_name'].apply(normalize_text)
        self.df['genus'] = self.df['scientific_name'].str.split().str[0]
        self.df = self.df.dropna(subset=['genus'])

        # Create label mappings with normalized names
        self.categories = sorted(self.df['category'].unique())
        self.species = sorted(self.df['scientific_name'].unique())
#         print(self.df['genus'].unique())
        self.genera = sorted(self.df['genus'].unique())

        self.category_to_idx = {cat: idx for idx, cat in
enumerate(self.categories)}
        self.species_to_idx = {sp: idx for idx, sp in
enumerate(self.species)}
        self.genus_to_idx = {gen: idx for idx, gen in
enumerate(self.genera)}
```
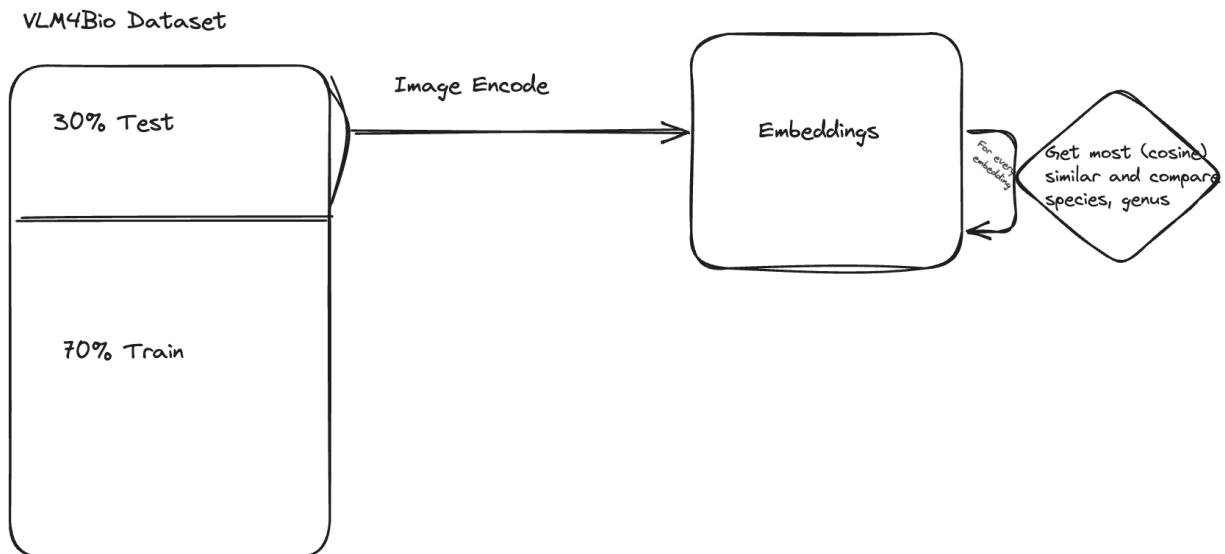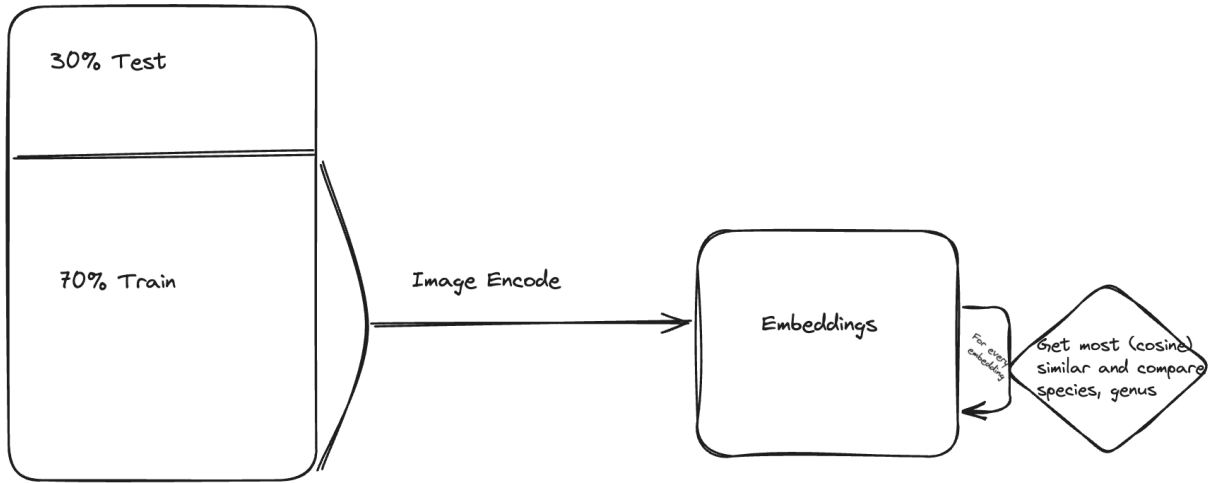
Flow Charts/Diagrams

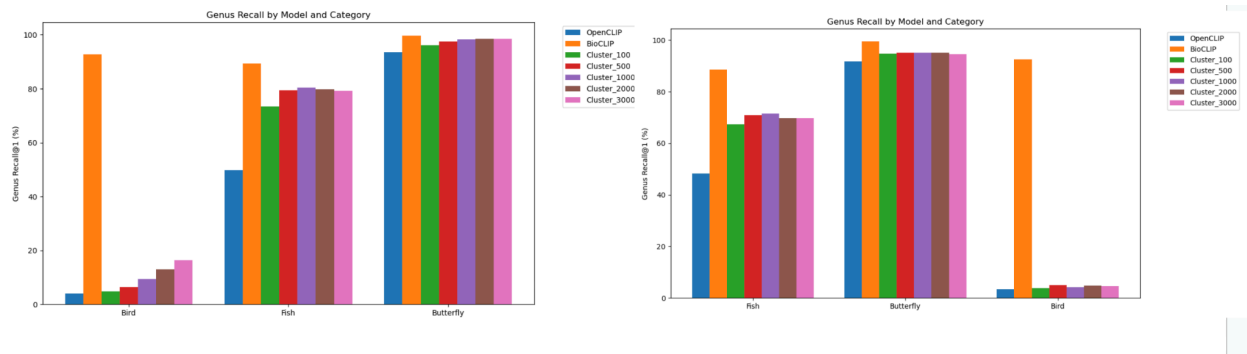Test evaluations on training/testing splits of dataset

VLM4Bio Dataset

30% Test

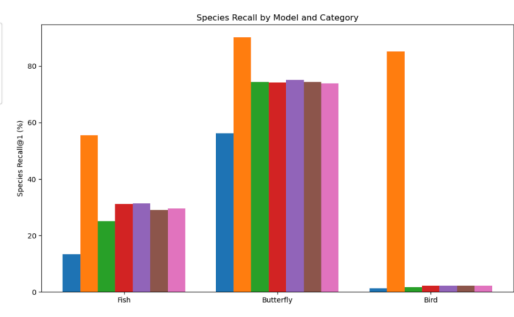70% Train

Image Encode

Embeddings

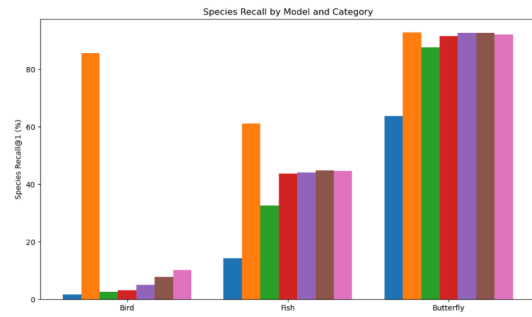For every embedding

Get most (cosine) similar and compare species, genus

Documentation

No documentation to add really just yet

Results Visualization + Proof of Work

Genus recall

Species recall

Species Recall by Model and Category

Next Week's Proposal

- Test run downloading VLM4Bio in HiPerGator
- Work on ideas for cross-comparing bioclip with our UNICOM-ified model