Week 3 Report

Romouald Dombrovski

September 06, 2024

1. Abstract

Maruf, M. et al. (2024) “VLM4Bio: A Benchmark Dataset to Evaluate Pretrained Vision-Language Models for Trait Discovery from Biological Images”. *arXiv preprint arXiv:2408.16176v1*. https://doi.org/10.48550/arXiv. 2408.16176v1

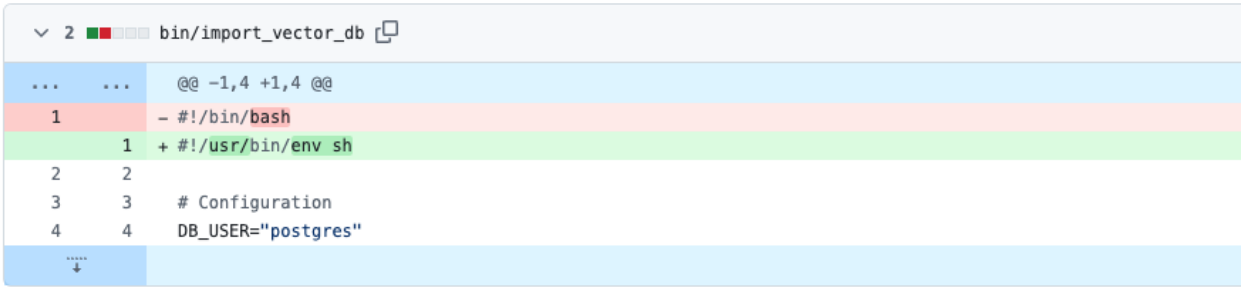
Abstract: Images are increasingly becoming the currency for documenting biodiversity on the planet, providing novel opportunities for accelerating scientific discoveries in the field of organismal biology, especially with the advent of large vision-language models (VLMs). We ask if pre-trained VLMs can aid scientists in answering a range of biologically relevant questions without any additional fine-tuning. In this paper, we evaluate the effectiveness of 12 state-of-the-art (SOTA) VLMs in the field of organismal biology using a novel dataset, VLM4Bio, consisting of 469K question-answer pairs involving 30K images from three groups of organisms: fishes, birds, and butterflies, covering five biologically relevant tasks. We also explore the effects of applying prompting techniques and tests for reasoning hallucination on the performance of VLMs, shedding new light on the capabilities of current SOTA VLMs in answering biologically relevant questions using images.

Summary (GPT-4o)**:** A Benchmark Dataset to Evaluate Pretrained Vision-Language Models for Trait Discovery from Biological Images" introduces a new dataset, VLM4Bio, designed to assess the capabilities of 12 state-of-the-art (SOTA) vision-language models (VLMs) in performing biologically relevant tasks using images of organisms such as fish, birds, and butterflies. This dataset contains 469,000 question-answer pairs derived from 30,000 images and focuses on tasks like species classification, trait identification, trait grounding, and trait counting. The paper evaluates the zero-shot capabilities of these models, without additional fine-tuning, and explores the impact of various prompting techniques and reasoning hallucination tests. Results reveal that while VLMs perform well on certain tasks like multiple-choice species classification, they struggle with more complex tasks like trait grounding and reasoning about images. The paper highlights gaps in current VLMs' ability to handle domain-specific tasks in organismal biology and provides suggestions for future work, such as fine-tuning models with biological data.

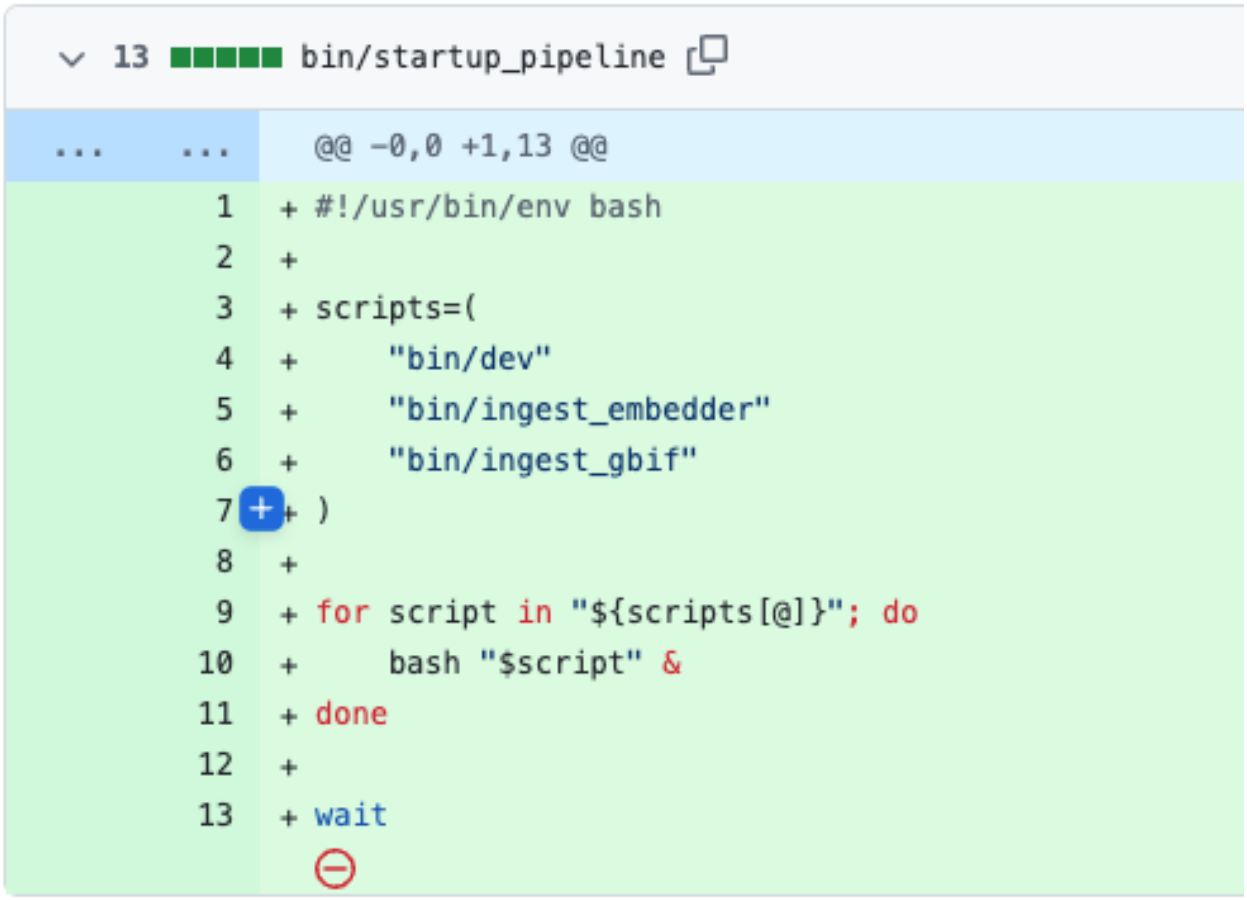
Relevance: This paper describes a dataset that includes relevant species images and data, which we can use to determine the validity of several models that we're examining for our tool. Once we have this dataset running on a script, we can determine some quantifiable metrics for which to compare the BioClip, OpenClip and Florence-2 models.

2. Scripts and Code Blocks

Some of the work involved standardization of our bin/ scripts:



Instead of running each pipeline separately, they can all be run together through one script now:



From last week's updates, some bugs were noticed, specifically for images without media_url not showing up as results on the search page:

10 frontend/static/models/Result.js

```
@@ -2,8 +2,14 @@ export default class Result {
    constructor(search_results) {
      this.id = search_results.id;
      this.media_url = search_results.media_url;
-     const url = new URL(this.media_url);
-     this.image_source_name = url.host;
+     try {
+       const url = new URL(this.media_url);
+       this.image_source_name = url.host;
+     }
+     catch (e) {
+       const url = null;
+       this.image_source_name = 'NA';
+     }
      this.latitude = search_results.latitude;
      this.longitude = search_results.longitude;
      this.map_url = `https://maps.google.com/?q=${this.latitude}%2C${this.longitude}`;
```

5 frontend/static/components/ResultsDisplay.js

```
@@ -8,7 +8,7 @@ export default {
          <div class="img-column" v-for="column in resultColumns" :key="column">
              <template v-for="result in column">
                  <img
-                     :src="result.media_url"
+                     :src="result.image_source_name != 'NA' ? result.media_url : 'static/unavailable-image.jpg'"
                      onerror="this.onerror=''; this.src='static/unavailable-image.jpg';"
                      class="img-fluid"
                      :id="result.id"
@@ -21,11 +21,12 @@ export default {
    data() {
      return {
        resultColumns: [[], [], [], []],
-       display: false
+       display: false,
      }
    },
    watch: {
      async apiResult(value) {
+       this.resultColumns = [[], [], [], []];
        if (value.length != 0) {
          await this.addImages(this.apiResult);
        }
```

1 frontend/static/components/SearchComponent.js

```
@@ -41,6 +41,7 @@ export default {
              </button>
            </div>
          </div>
+       </div>
      `,
      data() {
        return {
```

The rest of the work this week involved experimenting with models and datasets from HuggingFace. None of these changes have been pushed and are currently in progress.

3. Documentation

Along with the changes specified above, the project README was updated to reflect the new usage of the ingestor pipeline. Example:

Generate Embeddings

Once we've imported raw-form data into Mongo, we'll want to generate vector embeddings for the data and store them to Postgres. This is where the web api serves query results from.

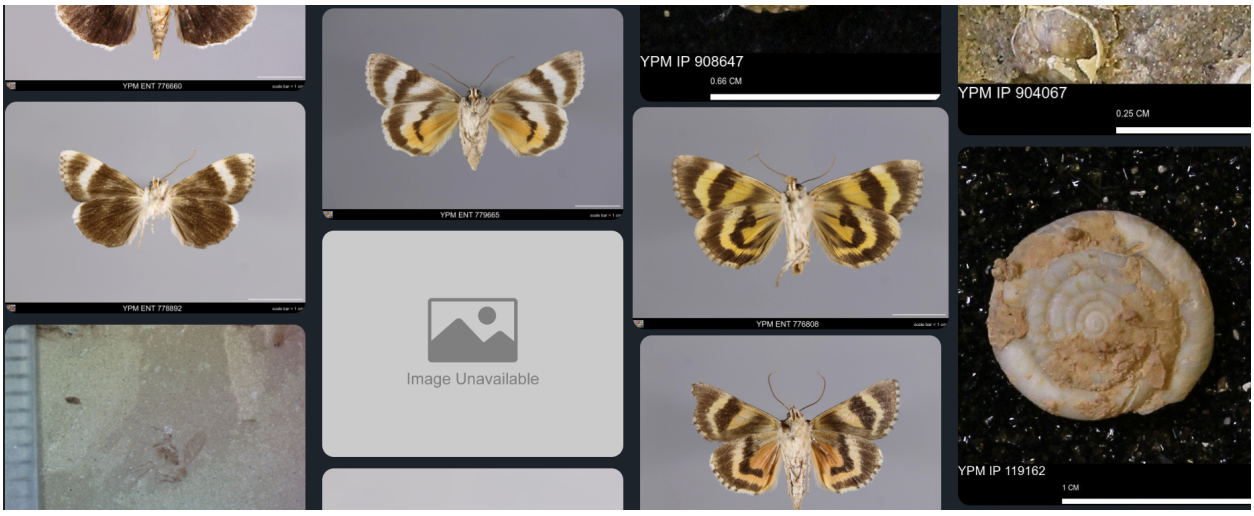
The process is very similar to importing data into Mongo. Again, if you've just started up the dev container, make sure to open a new terminal tab (assuming you're using VSCode) so that conda will init. Similarly, we can run a script to activate the embedder, or run it ourselves:

```
$ bin/ingest_embedder
```

As this ingestor is running, it is waiting a signal from the Redis queue to begin the embedding process. This will work very similarly to the `gbif` and `idigbio` queues above: From the workbench of Redis Insight, pass a simple search string to the `embedder` queue: `LPUSH embedder '{}'`

The new readme can be seen on the NFHM repo: https://github.com/Human-Augment-Analytics/NFHM/tree/main

4. Proof of work



5. Next Week Proposal

After having another meeting with Dr. Porto, we had discussed some potential directions we can take to evaluate ML models, which is a higher priority right now for the project. To that end, the goal for next week is to try and get them set up and running on sample scripts:

- create scripts to run Florence-2 and BioClip
- run models on VLM4Bio dataset
- begin porting components to .vue modules
- discuss and plan potential UX upgrades