Toward Self-Supervised Avian Diversity Monitoring

**ABSTRACT**

- **Monitoring biodiversity** is more important than ever in human history, but it is expensive, not reliable neither scalable.
- Machine Learning could bring a solution but it still requires big data sets all labeled by human experts [4,5].
- Here, we propose **Self-Supervised learning** for studying biodiversity through Avian Diversity Monitoring.
- **High Performance Computing** resources are demanded since we need to process great volumes of data on computationally demanding models.
- We use a joint embedding configuration [2] to acquire features from ~190 hours of audio.
- These algorithms are able to **discern different birdsongs**, as well as identify noise in various spectrograms at scale with **limited human intervention**.

**METHODS**

- We applied a Joint Embedding architecture with a ViT backbone [2].
- The cloud of points from the Embeddings [2] is obtained by running inference with a trained model.
- A KMEANS clustering with 100 clusters is conducted on a random sub-sample of 50k images.
- The dimensionality in the central figure has been reduced to 2 for visualization reasons. When using all the dimensions (384), birds can be separated with more accuracy.

**RESULTS**

- It distributes the spectrograms in a multidimensional output space.
- Similar acoustic conditions are mapped to neighboring regions in the output space.
- For instance, some clusters contain silent audio spectrograms with similar noise conditions.
- Remarkably, there are some clusters populated with spectrograms containing similar bird songs.

**CONCLUSIONS**

- **Self-Supervised learning is well suited for analyzing audio data**, making the labeling process for biologists and ornithologists tractable and scalable.
- Ornithologists—for instance—could easily **tease apart millions of spectrograms with different noise characteristics**, detecting anomalies as outliers in the output space.
- Ideally experts can **detect different birdsongs** in different regions of the output space, and label millions of examples per hour.

**NEXT STEPS**

- Process all the 2.3TB of audio and also test different ViTs sizes as well as different joint embedding strategies such as Variance Invariance Covariance Regularization [1].
- Compare the clusterization performance of several output spaces utilizing pseudo labels generated by trained supervised algorithms such as BirdNet [4].
- Use labeled datasets as a mean of validation of the clustering performance as well as testing the segmentation produced from attentional maps in order to detect informative features from background noise in spectrograms.

**REFERENCES**


Dataset: https://zenodo.org/record/6964885
Code: https://zenodo.org/record/6964966