Self-Supervised Learning for Automated Species Detection from Passively Recorded Soundscapes in Avian Diversity Monitoring

1st Dario Dematties
Mathematics and Computer Science Division
Argonne National Laboratory
Lemont IL, USA
ddematties@anl.gov

2nd Bhupendra A. Raut
Environmental Sciences Division
Argonne National Laboratory
Lemont IL, USA
braut@anl.gov

3rd Rajesh Sankaran
Mathematics and Computer Science Division
Argonne National Laboratory
Lemont IL, USA
rajesh@mcs.anl.gov

4th Nicola J. Ferrier
Mathematics and Computer Science Division
Argonne National Laboratory
Lemont IL, USA
nferrier@anl.gov

I. Introduction

By detecting different animal species reliably at scale we can protect biodiversity. Yet, traditionally, biodiversity data has been collected by expert observers which is prohibitively expensive, not reliable neither scalable. On the other hand, automated species detection via machine-learning is promising, but it is constrained by the necessity of large training data sets all labeled by human experts [4], [6].

Here, we propose to use Self-Supervised Learning for studying semantic features in spectrograms extracted from passively collected acoustic data. We recorded audio using the recording devices shown in Fig. 1, from the The Morton Arboretum natural reserve.

Next, we split the audio files into 6 secs segments and converted them into mel spectrograms [5]. We utilized a joint embedding configuration called DIstillation NO labels (DINO) [2] to acquire features from the spectrograms in the data set.

We processed 230,259 (1024x1024) spectrogram images from recordings from Jun 28 to Jul 06 2021 ($\sim$190 hours of audio). In order to process these volumes of data we utilized ThetaGPU which is a HPC cluster provided by the Argonne Leadership Computing Facility.

Finally we analyzed the output space from a trained backbone using KMEANS on 100 clusters and found that the clusters retain important semantic attributes of the spectrograms.

We envisage these preliminary results as compelling for future automatic assistance of biologist as a pre-processing stage for labeling very big data sets.

National Science Foundation (NSF). Award Number: 1935984

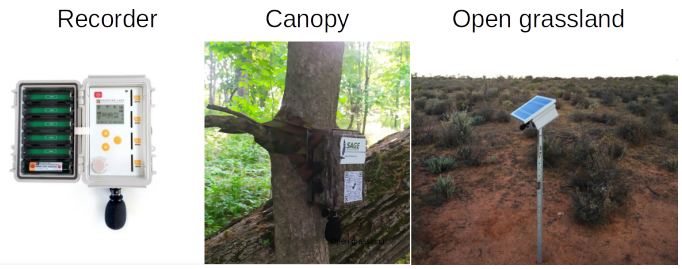

Fig. 1. Left: The BAR-LT is a battery-powered device. Center: Its Canopy installation. Right: The Solar BAR has a built in solar panel and battery charger to 24/7 recording capabilities over long periods of time.

II. Methods

A. Passively Recorded Soundscapes and HPC Resources

We collected approximately 2.3TB of audio data by using nine recorders (2 open grassland and 7 canopy in Fig. 1). Recordings were conducted at Morton Arboretum, located in Illinois, US. from May 24th to September 1st 2021. We utilized ThetaGPU which is comprised of 24 NVIDIA DGX A100 nodes. Each DGX A100 node comprises eight NVIDIA A100 Tensor Core GPUs and two AMD Rome CPUs. For the experiments conducted for this presentation we trained a Vision Transformer (ViT) [3] on $\sim$230k images during 50 epochs using one node (8 GPUs) in the machine.

B. Self-Supervised Learning and Joint Embedding

Self-supervised learning is a machine learning approach—mostly utilized in the deep learning workflows—in which supervision does not come from human assigned labels but from pretext tasks assigned on an unlabeled data set.

Self-supervised learning has brought superlative progress in the last years, from the breakthroughs produced by big lan-