Running a Kaggle Competition Team

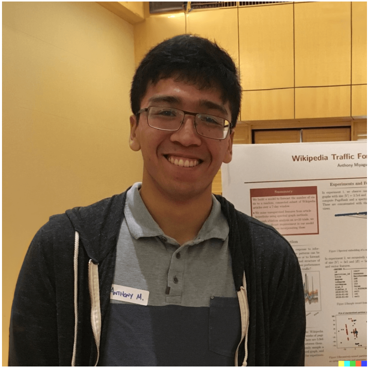
Insights and Lessons from BirdCLEF with Data Science @ Georgia Tech

Anthony Miyaguchi
acmiyaguchi@gatech.edu

Georgia Institute of Technology

2023-05-03

Who am I?

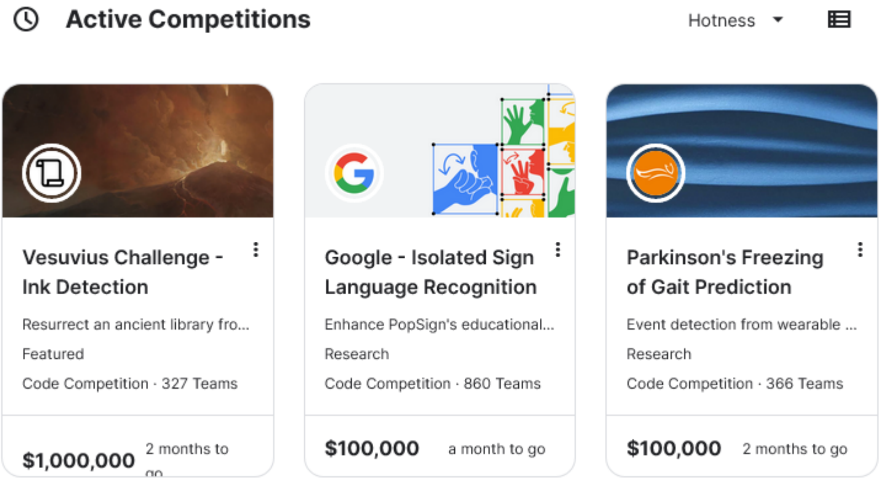


- A Software Engineer
  - 5 years as a Data Engineer at Mozilla
  - 1.5 years as a Software Engineer at Planet Labs
- OMSCS, matriculated Spring 2022
  - B.S. Computer Science and Engineering from UCLA 2016
  - Graduate Certificate from Stanford Center for Professional Development (SCPD) 2018
- Career focus on scalable data systems and machine learning

Overview

- What is Kaggle?
- Why Kaggle?
- BirdCLEF 2022
  - Recruitment and Technical Approach
- BirdCLEF 2023
  - Updated Recruitment and Technical Approach
- Thoughts and Advice
- Q&A

What is Kaggle?



Active Competitions

Hotness

Vesuvius Challenge - Ink Detection

Resurrect an ancient library fro...

Featured

Code Competition · 327 Teams

$1,000,000 2 months to go

Google - Isolated Sign Language Recognition

Enhance PopSign's educational...

Research

Code Competition · 860 Teams

$100,000 a month to go

Parkinson's Freezing of Gait Prediction

Event detection from wearable ...

Research

Code Competition · 366 Teams

$100,000 2 months to go

Figure 1: Kaggle: a platform for data science competitions

Kaggle is what got me into Graduate School

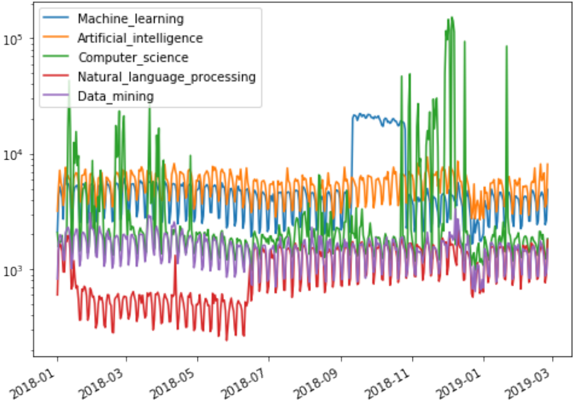
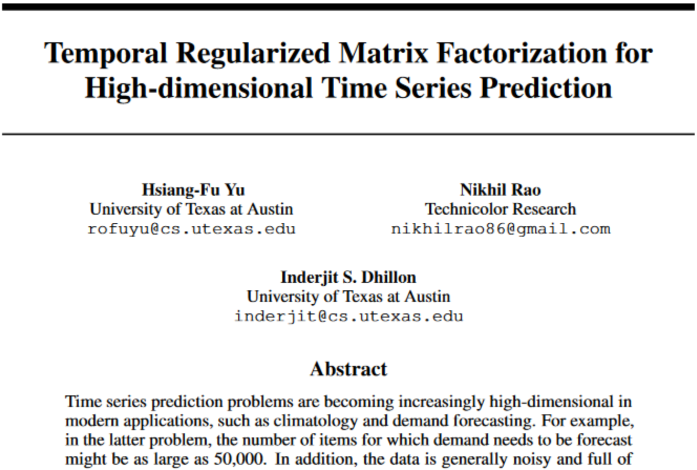


Figure 2: Wikipedia page view data in the style of the “Web Traffic Time Series Forecasting” Kaggle Competition

Reading and Implementing Research Papers



Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction

Hsiang-Fu Yu
University of Texas at Austin
rofuyu@cs.utexas.edu

Nikhil Rao
Technicolor Research
nikhilrao86@gmail.com

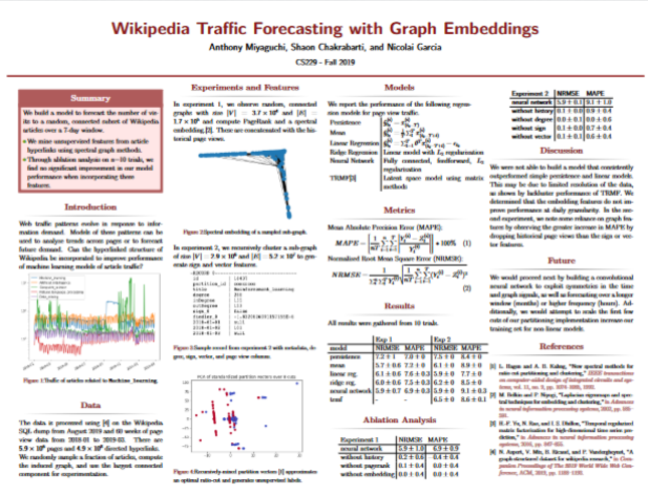
Inderjit S. Dhillon
University of Texas at Austin
inderjit@cs.utexas.edu

Abstract

Time series prediction problems are becoming increasingly high-dimensional in modern applications, such as climatology and demand forecasting. For example, in the latter problem, the number of items for which demand needs to be forecast might be as large as 50,000. In addition, the data is generally noisy and full of

Figure 3: The first piece of literature I implemented outside of school work as part of the “Web Traffic Time Series Forecasting” Kaggle Competition

Stanford CS229 Machine Learning Project



Wikipedia Traffic Forecasting with Graph Embeddings

Anthony Miyaguchi, Shaon Chakrabarti, and Nicolai Garcia

CS229 - Fall 2019

Summary

We build a model to forecast the number of visits to a random, connected subset of Wikipedia articles over a 7-day window.

- We mine unsupervised features from article hyperlinks using spectral graph methods.
- Through ablation analysis on n=10 trials, we find no significant improvement in our model performance when incorporating these features.

Introduction

Web traffic patterns evolve in response to information demand. Models of these patterns can be used to analyze trends across pages or to forecast future demand. Can the hyperlinked structure of Wikipedia be incorporated to improve performance of machine learning models of article traffic?

Figure 1:Traffic of articles related to Machine_learning.

Data

The data is processed using [4] on the Wikipedia SQL dump from August 2019 and 60 weeks of page view data from 2018-01 to 2019-03. There are 5.9×10^6 pages and 4.9×10^8 directed hyperlinks. We randomly sample a fraction of articles, compute the induced graph, and use the largest connected component for experimentation.

Experiments and Features

In experiment 1, we observe random, connected graphs with size $|V| = 3.7 \times 10^4$ and $|E| = 1.7 \times 10^5$ and compute PageRank and a spectral embedding [2]. These are concatenated with the historical page views.

Figure 2:Spectral embedding of a sampled sub-graph.

In experiment 2, we recursively cluster a sub-graph of size $|V| = 2.9 \times 10^6$ and $|E| = 5.2 \times 10^7$ to generate sign and vector features.

```
-RECORD 0-----------------------------
id            | 11431
partition_id  | [illegible]
title         | Reinforcement_learning
degree        | 210
indegree      | 111
outdegree     | 113
sign_0        | false
fiedler_0     | -1.8[illegible]E-4
2018-01-01    | null
2018-01-02    | 103
2018-01-03    | null
```

Figure 3:Sample record from experiment 2 with metadata, degree, sign, vector, and page view columns.

Figure 4:Recursively-mined partition vectors [1] approximates an optimal ratio-cut and generates unsupervised labels.

Models

We report the performance of the following regression models for page view traffic.

Persistence	$\hat{y}_{t_k}^{(i)} = x_{(t_k - T)}^{(i)}$
Mean	$\hat{y}_{t_k}^{(i)} = \frac{1}{T}\sum_{t}^{T} x_{(t_k - T + t)}^{(i)}$
Linear Regression	$\hat{y}_{t_k}^{(i)} = \sum_{t=1}^{T} \theta^T x_{(t_k - T + t)}^{(i)} + \epsilon_k$
Ridge Regression	Linear model with L_2 regularization
Neural Network	Fully connected, feedforward, L_2 regularization
TRMF[3]	Latent space model using matrix methods

Metrics

Mean Absolute Precision Error (MAPE):

$$MAPE = \left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left|\frac{Y_t^{(i)} - \hat{Y}_t^{(i)}}{Y_t^{(i)}}\right|\right] * 100\% \quad (1)$$

Normalized Root Mean Square Error (NRMSE):

$$NRMSE = \frac{1}{\sum_i^n \sum_t^T Y_t^{(i)}}\sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_t^{(i)} - \hat{Y}_t^{(i)})^2} \quad (2)$$

Results

All results were gathered from 10 trials.

	Exp 1		Exp 2	
model	NRMSE	MAPE	NRMSE	MAPE
persistence	7.2 ± 1	7.0 ± 0	7.5 ± 0	8.4 ± 0
mean	5.7 ± 0.6	7.2 ± 0	6.1 ± 0	8.9 ± 0
linear reg.	6.1 ± 0.6	7.6 ± 0.3	5.9 ± 0	7.7 ± 0
ridge reg.	6.0 ± 0.6	7.5 ± 0.3	6.2 ± 0	8.5 ± 0
neural network	5.9 ± 0.7	6.9 ± 0.3	5.9 ± 0	9.1 ± 0.3
trmf	-	-	6.5 ± 0	8.6 ± 0.1

Ablation Analysis

Experiment 1	NRMSE	MAPE
neural network	5.9 ± 1.0	6.9 ± 0.9
without history	0.2 ± 0.6	−0.4 ± 0.4
without pagerank	0.1 ± 0.4	0.0 ± 0.4
without embedding	0.0 ± 0.4	0.0 ± 0.4

Experiment 2	NRMSE	MAPE
neural network	5.9 ± 0.1	9.1 ± 1.0
without history	0.1 ± 0.0	0.9 ± 0.4
without degree	0.0 ± 0.1	0.0 ± 0.6
without sign	0.1 ± 0.0	0.7 ± 0.4
without vector	0.1 ± 0.1	0.6 ± 0.4

Discussion

We were not able to build a model that consistently outperformed simple persistence and linear models. This may be due to limited resolution of the data, as shown by lackluster performance of TRMF. We determined that the embedding features do not improve performance at daily granularity. In the second experiment, we note some reliance on graph features by observing the greater increase in MAPE by dropping historical page views than the sign or vector features.

Future

We would proceed next by building a convolutional neural network to exploit symmetries in the time and graph signals, as well as forecasting over a longer window (months) or higher frequency (hours). Additionally, we would attempt to scale the first few cuts of our partitioning implementation increase our training set for non linear models.

References

[1] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 11, no. 9, pp. 1074–1085, 1992.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *in Advances in neural information processing systems*, 2002, pp. 585–591.

[3] H.-F. Yu, N. Rao, and I. S. Dhillon, "Temporal regularized matrix factorization for high-dimensional time series prediction," *in Advances in neural information processing systems*, 2016, pp. 847–855.

[4] N. Aspert, V. Miz, B. Ricaud, and P. Vandergheynst, "A graph-structured dataset for wikipedia research," *in Companion Proceedings of The 2019 World Wide Web Conference*, ACM, 2019, pp. 1188–1193.

Figure 4: "Wikipedia Traffic Forecasting with Graph Embeddings" is a machine learning project during my time in SCPD inspired by the Kaggle competition.

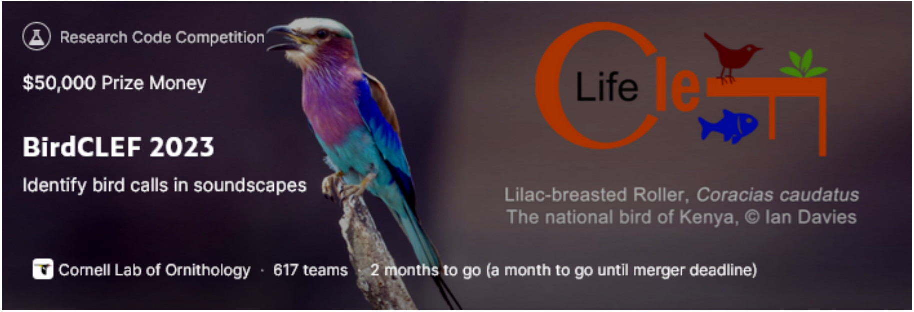
Why Kaggle?

Figure 5: The Titanic dataset is an iconic Kaggle challenge.

Reasons to consider Kaggle

- Good start to a **research project**
    - Access to impactful real-world problems
        - bird conservation, cancer detection, etc.
- **Structure and deadlines** to stay on track
- Anyone can participate, with a focus on **practical applications and the potential to win cash prizes**

BirdCLEF



Research Code Competition

$50,000 Prize Money

BirdCLEF 2023

Identify bird calls in soundscapes

Cornell Lab of Ornithology · 617 teams · 2 months to go (a month to go until merger deadline)

LifeCLEF
Lilac-breasted Roller, *Coracias caudatus*
The national bird of Kenya, © Ian Davies

Figure 6: BirdCLEF is a competition to help with bird conservation.

DS@GT Competition Team

Recruitment

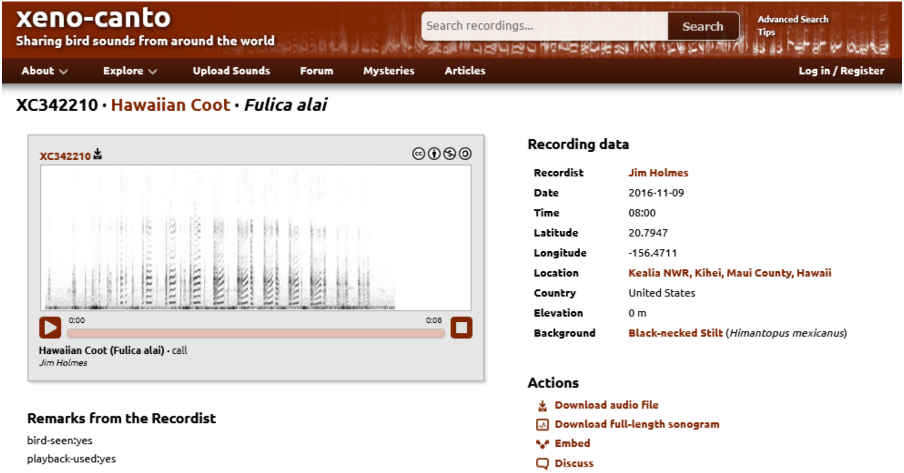
- Built a team of 5 people from DS@GT in Spring 2022
- 3 masters students and 2 undergraduates

Technical Approach

- **Motif Mining** with SiMPLe-Fast
  - Find salient sections of audio for classification
- **Unsupervised Representation Learning** via Triplet Loss
  - Train an embedding for downstream tasks

Figure 7: DS@GT: a student-run data science organization

Why is audio classification challenging?



XC342210 • Hawaiian Coot • *Fulica alai*

Hawaiian Coot (Fulica alai) · call
Jim Holmes

Recording data

Recordist	Jim Holmes
Date	2016-11-09
Time	08:00
Latitude	20.7947
Longitude	-156.4711
Location	Kealia NWR, Kihei, Maui County, Hawaii
Country	United States
Elevation	0 m
Background	Black-necked Stilt (*Himantopus mexicanus*)

Remarks from the Recordist

bird-seen:yes

playback-used:yes

Actions

- Download audio file
- Download full-length sonogram
- Embed
- Discuss

Figure 8: xeno-canto is a crowd sourced database of bird sounds.

Reading the literature

Motif Mining - SiMPLe-Fast

Silva, D. F., Yeh, C. C. M., Zhu, Y., Batista, G. E., & Keogh, E. (2018). Fast similarity matrix profile for music analysis and exploration. IEEE Transactions on Multimedia, 21(1), 29-38.

Spatial Embeddings - Tile2Vec

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., & Ermon, S. (2018). Tile2Vec: unsupervised representation learning for spatially distributed data. arXiv.

Motif Mining with SiMPLe-Fast

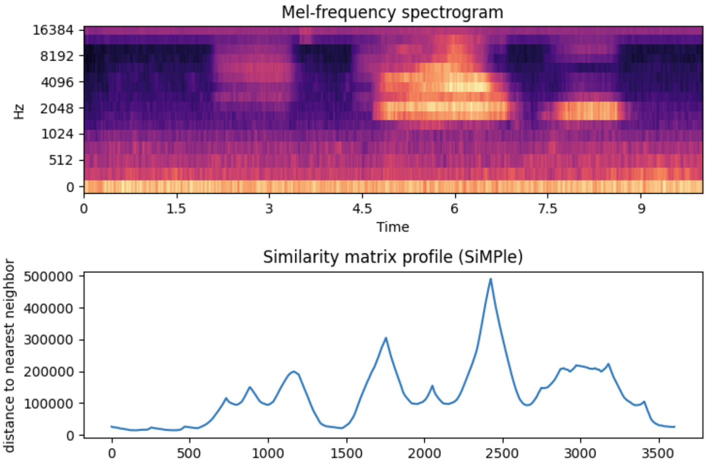


Figure 9: SiMPLe-Fast is a motif mining algorithm used to find all pairs similarity in a time-series.

Representation Learning via Triplet Loss

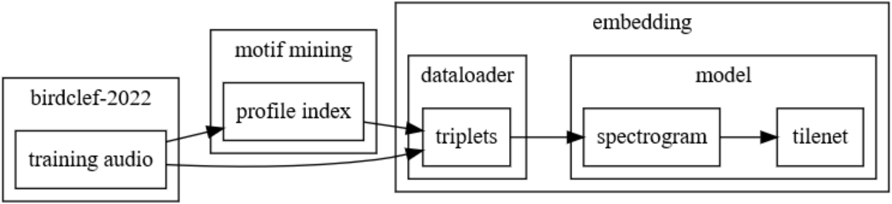
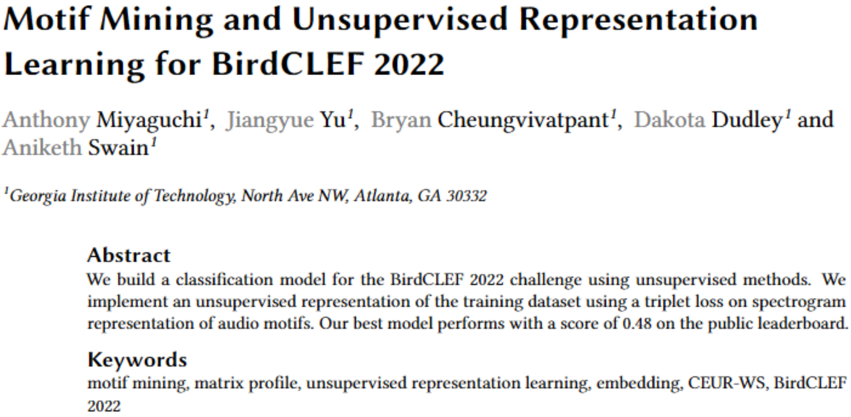


Figure 10: We train an embedding using mined motifs via a triplet loss

Working Notes



Motif Mining and Unsupervised Representation Learning for BirdCLEF 2022

Anthony Miyaguchi[1], Jiangyue Yu[1], Bryan Cheungvivatpant[1], Dakota Dudley[1] and Aniketh Swain[1]

[1] *Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332*

Abstract

We build a classification model for the BirdCLEF 2022 challenge using unsupervised methods. We implement an unsupervised representation of the training dataset using a triplet loss on spectrogram representation of audio motifs. Our best model performs with a score of 0.48 on the public leaderboard.

Keywords

motif mining, matrix profile, unsupervised representation learning, embedding, CEUR-WS, BirdCLEF 2022

Figure 11: Working notes submitted to the CLEF 2022 conference

Results of BirdCLEF 2022

Best Working Notes Award

We performed poorly, but we kept good notes and had a unique approach. We won $2,500 in Google Cloud Platform credits.

Personal Takeaways

- Building a team is worthwhile and a forcing function for progress
- Managing a team of 5 people is challenging
- 6 weeks is not enough time for meaningful contributions to a large codebase

Approaching BirdCLEF 2023

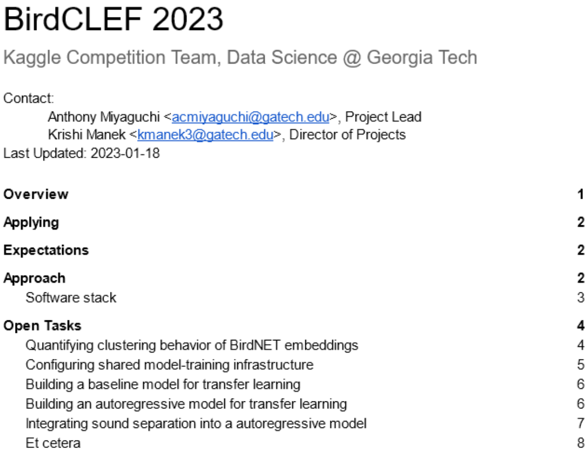
New year, new team

- Building the team from the get-go
- Proposals, assessments, and interviews
- Reaching out to Slack, EdStem, and DS@GT

New technical approach

- Retrain using embeddings from older models
- Build a process for machine-assisted dataset annotation
- Toy with sequence models (RNNs, Transformers, etc.)

Recruitment: Proposal



BirdCLEF 2023

Kaggle Competition Team, Data Science @ Georgia Tech

Contact:
Anthony Miyaguchi <acmiyaguchi@gatech.edu>, Project Lead
Krishi Manek <kmanek3@gatech.edu>, Director of Projects
Last Updated: 2023-01-18

Figure 12: A proposal document is a useful way to document intent and expectations.

Recruitment: Assessment

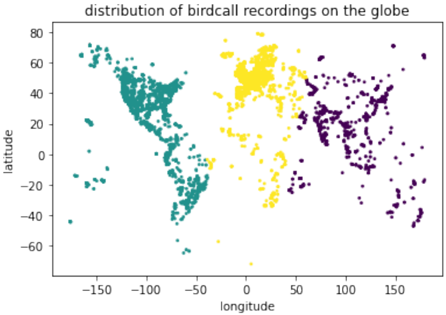
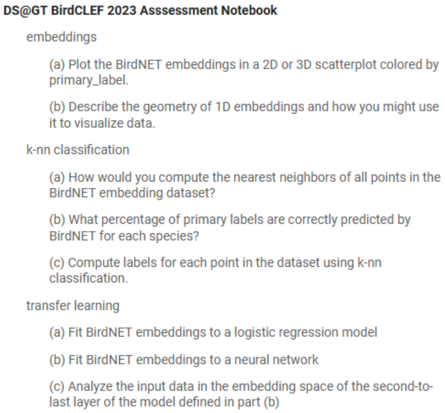


Figure 13: Assessment in a Fall 2022 project group, using BirdCLEF metadata

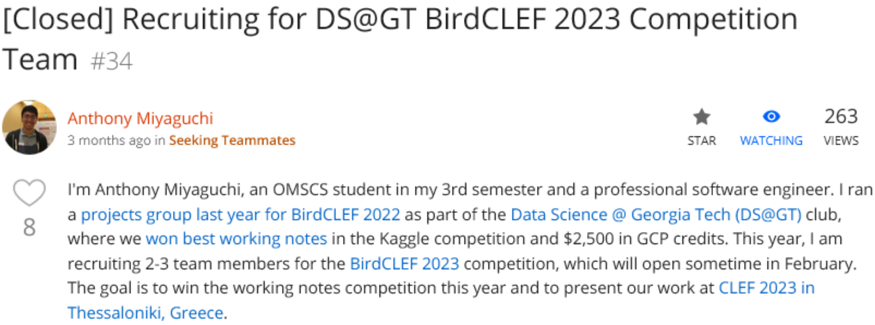


DS@GT BirdCLEF 2023 Asssessment Notebook

- embeddings
  - (a) Plot the BirdNET embeddings in a 2D or 3D scatterplot colored by primary_label.
  - (b) Describe the geometry of 1D embeddings and how you might use it to visualize data.
- k-nn classification
  - (a) How would you compute the nearest neighbors of all points in the BirdNET embedding dataset?
  - (b) What percentage of primary labels are correctly predicted by BirdNET for each species?
  - (c) Compute labels for each point in the dataset using k-nn classification.
- transfer learning
  - (a) Fit BirdNET embeddings to a logistic regression model
  - (b) Fit BirdNET embeddings to a neural network
  - (c) Analyze the input data in the embedding space of the second-to-last layer of the model defined in part (b)

Figure 14: Table of Contents from the BirdCLEF 2023 assessment.

Recruitment: Outreach



[Closed] Recruiting for DS@GT BirdCLEF 2023 Competition Team #34

Anthony Miyaguchi
3 months ago in Seeking Teammates

I'm Anthony Miyaguchi, an OMSCS student in my 3rd semester and a professional software engineer. I ran a projects group last year for BirdCLEF 2022 as part of the Data Science @ Georgia Tech (DS@GT) club, where we won best working notes in the Kaggle competition and $2,500 in GCP credits. This year, I am recruiting 2-3 team members for the BirdCLEF 2023 competition, which will open sometime in February. The goal is to win the working notes competition this year and to present our work at CLEF 2023 in Thessaloniki, Greece.

Figure 15: A post on the OMSCS Research EdStem board.

Reach out to other students!

- Slack: OMSCS Study Group and OMSA Study Group
- EdStem: OMSCS Research
- Clubs: Data Science @ Georgia Tech

Organization

Team Structure

- Weekly meetings over Zoom once a week for 45 minutes
- Slack channel for communication
- Shared GitHub and Google Cloud Project

Being a Project Lead

Leading a team is challenging. Every team is different, and every leader has their own style.

Technical approach

Outline

- Building data pipelines with Luigi
- BirdNET embeddings
- Sound Separation with MixIT
- Automated dataset annotation
- Sequence models with embeddings

Reading the literature, yet again

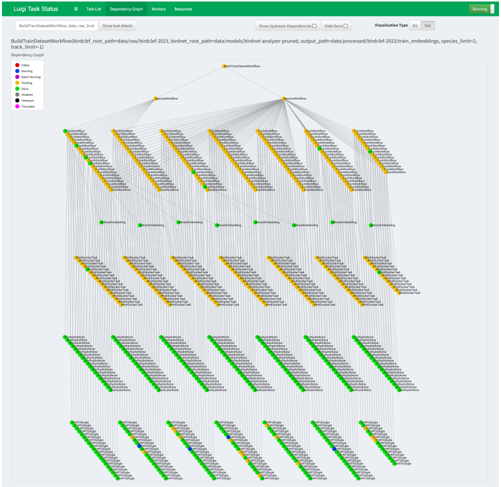
Domain specific deep learning model - BirdNET

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, 61, 101236.

Sound separation - MixIT

Denton, T., Wisdom, S., & Hershey, J. R. (2022, May). Improving bird classification with unsupervised sound separation. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 636-640). IEEE.

Building data pipelines with Luigi



BuildTrainDatasetWorkflow(birdclef_root_path=data/raw/birdclef-2023, birdnet_root_path=data/models/birdnet-analyzer-pruned, output_path=data/processed/birdclef-2023/train_embeddings, species_limit=2, track_limit=-1)

Figure 16: Luigi is a Python library for building data pipelines.

BirdNET embeddings

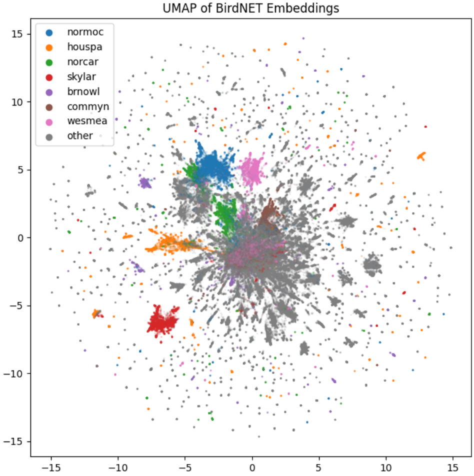


Figure 17: We can use the BirdNET embedding space for search and nearest neighbor queries.

BirdNET predictions for annotation

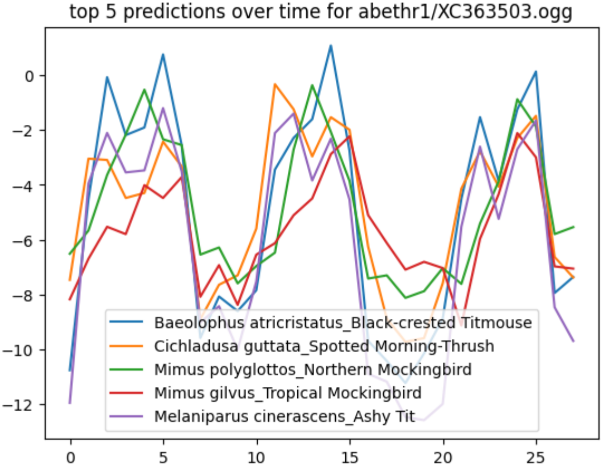


Figure 18: The BirdNET predictions can help with data annotation.

Sound Separation with MixIT

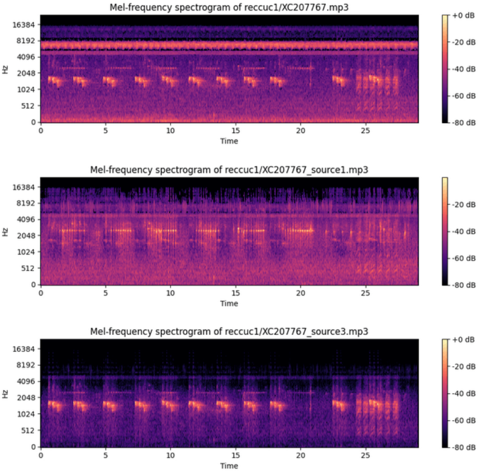


Figure 19: MixIT is a sound separation algorithm.

Automated dataset annotation

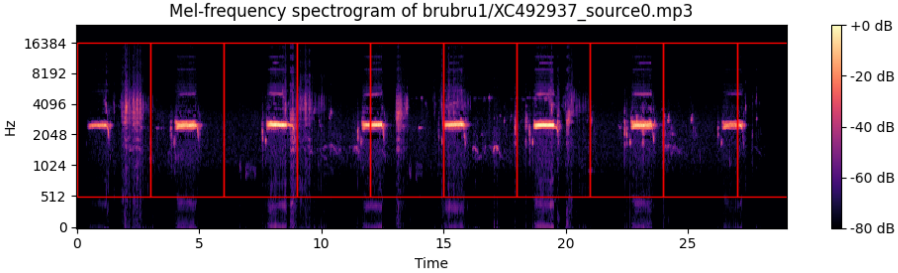


Figure 20: Chunked spectrogram of a bird call.

Sequence models with embeddings

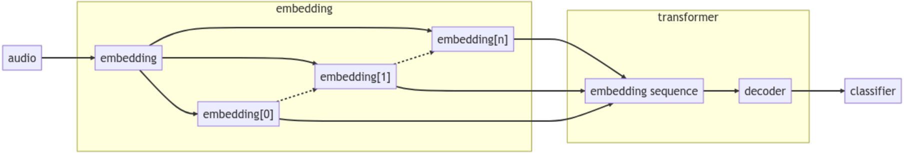


Figure 21: We experiment with embeddings in a sequence model (e.g. Transformers) to imbue temporal context.

BirdCLEF 2023 is ongoing!

Figure 22: Bird conservation is a worthy cause and a great opportunity to learn.

Interested in joining?

Talk to me if you're interesting in jumping in last minute.

Advice for myself one year ago

Building a team is worthwhile

- A strong team can help you achieve more than you could on your own. It's also an opportunity to connect with other students.

Be prepared to learn how to lead a team

- Effective communication and clear timelines are key to keeping the team on track
- Remember that everyone on the team is capable and valuable, and make an effort to recognize and appreciate their contributions

Reach out to OMSCS and OMSA early

- Working professionals have *a lot* to bring to the table.

Be on the lookout for opportunities

Figure 23: Be on the lookout!

There's an abundance of opportunities for OMSCS students to collaborate with other students.

Thank you to everyone involved

DS@GT Leadership

- Pulak Agarwal
- Krishi Manek

BirdCLEF 2022

- Jiangyue Yu
- Bryan Cheungvivatpant
- Dakota Dudley
- Aniketh Swain

BirdCLEF F22 EDA

- Jinsong Zhen
- Kien Tran
- Siying Liu
- Muskaan Gupta
- Xinjin Li

BirdCLEF 2023

- Chris Hayduk
- Erin Middlemas
- Grant Williams
- Nathan Zhong

Links and Resources

- Working Notes, “Motif Mining and Unsupervised Representation Learning for BirdCLEF 2022”
- DS@GT, Kaggle Competition Team Proposal, BirdCLEF 2022
- DS@GT, Project Group Proposal, BirdCLEF EDA Fall 2022
- DS@GT, Kaggle Competition Team Proposal, BirdCLEF 2023
- DS@GT, Assessment, BirdCLEF EDA Fall 2022
- DS@GT, Assessment, BirdCLEF 2023
- BirdCLEF Motif Viewer, Barn Owl, XC138041
- BirdCLEF 2023 MixIT Exploration, Red-chested Cuckoo, 2FXC207767

Thank you!

Time for Questions and Answers

Figure 24: Q&A