Heterogeneous Graph Neural Networks for Species Distribution Modeling

Lauren Harrell[1], Christine Kaiser-Chen[2]
{laurenharrell, christinech}@google.com

Burcu Karagol Ayan[2], Keith Anderson[2], Michelangelo Conserva[1], Elise Kleeman[1], Maxim Neumann[2], Matt Overlan[2], Melissa Chapman[1], Drew Purves[2]

[1]Google Research, [2]Google DeepMind

Google Research

Introduction

Motivation

Understanding where species are distributed across the globe is essential not only for advancing our scientific understanding of ecological processes, but also to enable the strategic implementation of conservation measures and policies. Species distribution models (SDMs) are used to model the relationship between species detections and their associated environmental variables, but often are limited due to constraints with the data design, single-species model construction, and other factors. We introduce a novel presence-only SDM with graph neural networks (GNN) as a way to capture more complex and nuanced information from species detection data to predict species occurrence or habitat suitability.

Key aspects of the GNN approach

- In our model, species and locations are treated as two distinct nodesets with their own unique feature sets and are connected through edges that represent observed detections of a given species at a given location.
- Model is based on the Interaction Network architecture (Battaglia et al, 2016) and trained with a JAX-based implementation of heterogeneous GNNs used in GraphCast (Lam et al., 2023). First step of the model is embedding all nodes and edges into the same latent dimensional space with multi-layer perceptrons (MLPs). Then for each number of message passing steps, edges and nodes are updated through MLPs that aggregate the information from sender and receiver node pairs. We decode all potential edges in the predicted bipartite graph using the dot-product of the resulting species and location node embeddings for each <species, location> pair
- Learning task is edge "prediction" for detection edges, where we infer whether a link should exist between a species location pair given the graph embeddings.

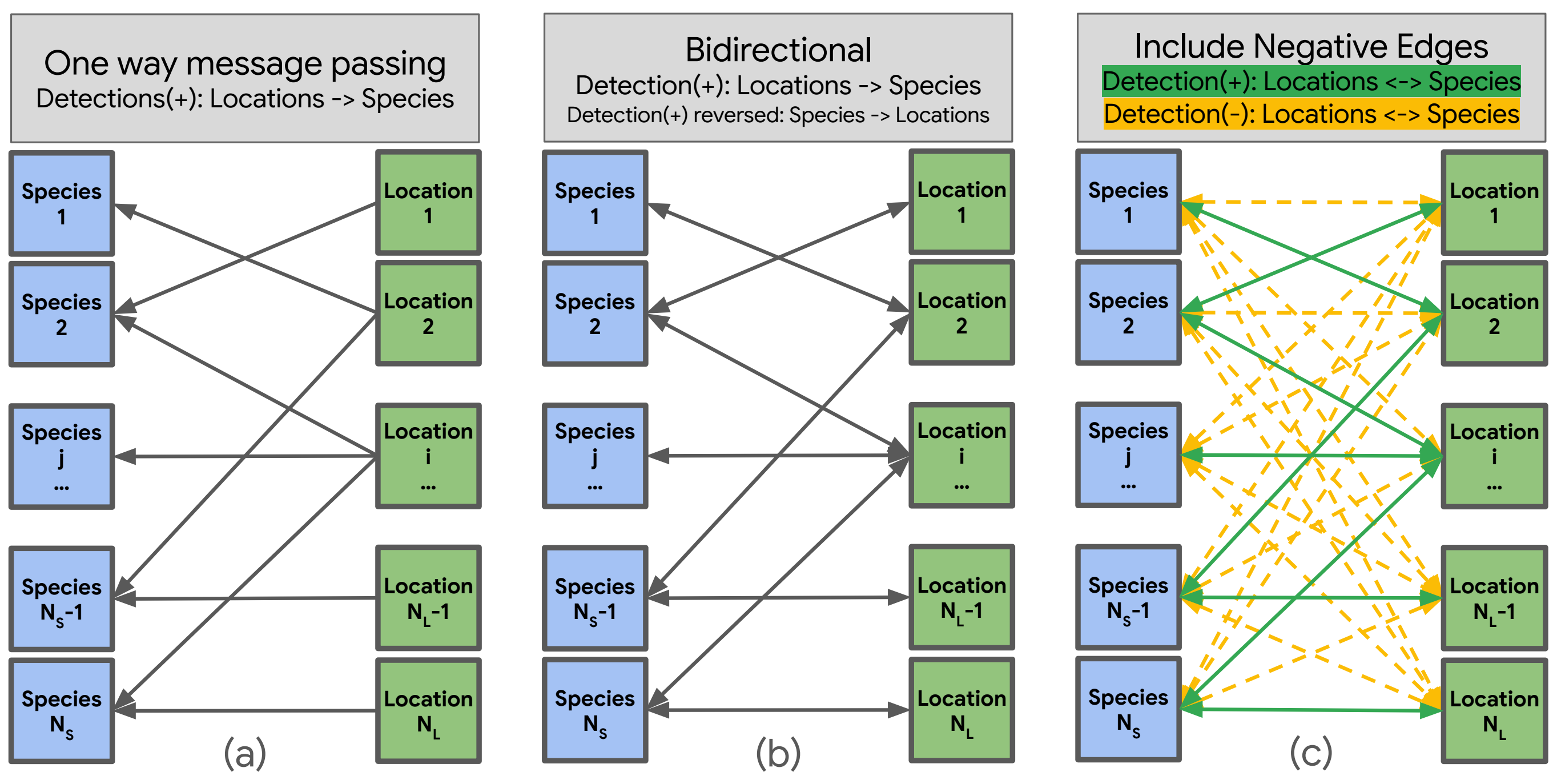


Figure 1: Example of the construction of edge sets in the bipartite graph connecting species to locations through detections of species *j* at location *i*. (a) One-way message passing of information from location nodes to species nodes through positive detections. (b) Bidirectional message passing between location nodes and species nodes through positive detections. (c) Bidirectional message passing through positive detection edges (solid green) and (pseudo)-negative edges (dashed yellow).

Case Study on NCEAS Benchmarking Data

United States National Center for Ecological Analysis and Synthesis (NCEAS) data description

Dataset published by Elith et al. (2020) for the purposes of benchmarking species distribution models. Training data consists of presence-only detection records with associated environmental variables in six regions, while test data includes presence-absence data from surveys for the selected species per region. All species names are de-identified, and thus species-specific information is limited.

Region Code	Region	Species set	Number of location variables	Training Locations (detections)	Unique Test Locations
AWT	Australian Wet Tropics, Queensland, Australia	20 birds 20 vascular plants	13	3806	442
CAN	Ontario, Canada	30 birds	11	5063	14571
NSW	North-east New South Wales, Australia	7 bats; 8 diurnal birds; 2 nocturnal birds; 8 open-forest trees; 8 open-forest understorey vascular plants; 7 rainforest trees; 6 rainforest understory vascular plants; 8 small reptiles	13	3323	8746
NZ	New Zealand	52 vascular plants	13	3088	19120
SA	Continental Brazil, Ecuador, Colombia, Bolivia, and Peru, South America	30 vascular plants	11	1178	152
SWI	Switzerland	30 trees	13	11429	10013

Application of GNN to NCEAS data

- Separate graphs and models were constructed, trained, and evaluated for each region; location node features include all available environmental features; species node features include indicators for each unique (obfuscated) species ID, and taxonomic group (where available).
- Experiments included one-way and bidirectional message passing between species and location nodes in addition to inclusion of negative edges (locations with detections only) for message passing
- Sigmoid cross-entropy loss function on the output logits
- Sampling of "negative" edges for loss function was part of active experiments, particularly the ratio of sampling background locations versus only locations in the detection data for loss computation

Baseline MLP

- Feed-forward MLP model as a baseline, using the same environmental features from the NCEAS dataset as input and multi-class classification head as the output. Similar loss function and negative sampling as the GNN approach.
- We ablate the hidden layer size, the number of hidden layers, and the ratio of sampled negative data per training batch.

Evaluation

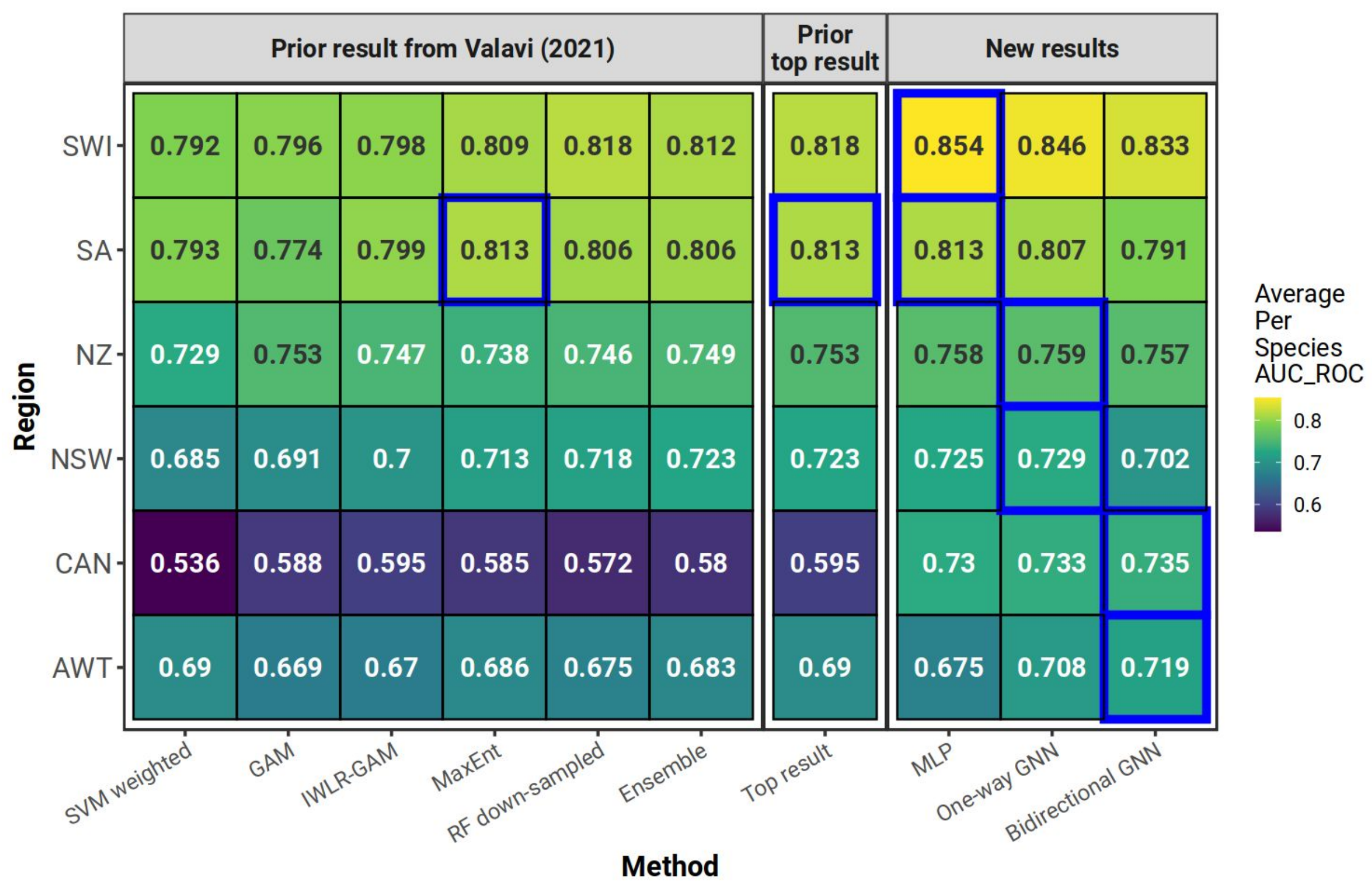
- Consistent with prior studies that benchmarked single-species SDM methods using the NCEAS data (Valavi et al. 2022), we evaluate the models on the averaged per-species Area Under the Receiver Operating Characteristic Curve (AUC_{ROC}).

Results

Key Findings

1. Compared to prior benchmarking results from Valavi et al. (2022) on single-species models, GNN has better performance on 3 regions (SWI, CAN, AWT) and comparable performance on 2 regions (NZ, NSW), with a relative increase of 23.5% over prior methods in CAN. The GNN has lower performance only on the SA region.

2. Compared to baseline MLP models, GNN is significantly better on AWT, comparable on 3 regions (NZ, NSW, CAN), and slightly worse on 2 regions (SWI, SA).



Region	Prior result from Valavi (2021)						Prior top result	New results		
	SVM weighted	GAM	IWLR-GAM	MaxEnt	RF down-sampled	Ensemble	Top result	MLP	One-way GNN	Bidirectional GNN
SWI	0.792	0.796	0.798	0.809	0.818	0.812	0.818	0.854	0.846	0.833
SA	0.793	0.774	0.799	0.813	0.806	0.806	0.813	0.813	0.807	0.791
NZ	0.729	0.753	0.747	0.738	0.746	0.749	0.753	0.758	0.759	0.757
NSW	0.685	0.691	0.7	0.713	0.718	0.723	0.723	0.725	0.729	0.702
CAN	0.536	0.588	0.595	0.585	0.572	0.58	0.595	0.73	0.733	0.735
AWT	0.69	0.669	0.67	0.686	0.675	0.683	0.69	0.675	0.708	0.719

Figure 2: AUC_{ROC} averaged across species per site by region and model methodology. The values of prior results were taken from the top scoring models in Valavi et al. (2022). Top result per region highlighted in blue.

Conclusions

We provide a basic example of using heterogeneous graph neural networks for species distribution modeling through a bipartite graph as a proof-of-concept on publicly-available benchmarking data.

- Performance is on par with or exceeds other SDM methods for presence-only data, demonstrating the GNN methodology has promise for further exploration.
- For both the MLP and GNN approaches, inclusion of background locations in loss computation showed no benefit over computing loss from pseudo-negatives sampled from locations with detections of other species, echoing findings from other deep-learning methods for SDMs (e.g. Cole et al., 2023)
- We are continuing to develop the approach further for more complex, feature-rich data and algorithms that include:
  - Training on blended presence-only and presence-absence data,
  - Informative species traits and remote-sensing-derived environmental features,
  - More complex graph structures and model architectures that include message-passing edges from species-to-species and location-to-location,
  - Alternative decoding functions,
  - Weighted loss functions that can vary the relative contribution of positives to different pseudo-negative strategies

Abstract

Species distribution models (SDMs) are necessary for measuring and predicting occurrence and habitat suitability of species and their relationship with environmental factors. We introduce a novel presence-only SDM with graph neural net-works (GNN). In our model, species and locations are treated as two distinct nodesets, and the learning task is predicting detection records as the edges that connect locations to species. Using GNN for SDM allows us to model fine-grained interactions between species and the environment. We evaluate the potential of this methodology on the six-region dataset compiled by National Center for Eco-logical Analysis and Synthesis (NCEAS) for benchmarking SDMs. For each of the regions, the heterogeneous GNN model is comparable to or outperforms previously-benchmarked single-species SDMs as well as a feed-forward neural network baseline model.

Selected References

Elith J, Graham C, Valavi R, Abegg M, Bruce C, Ferrier S, Ford A, Guisan A, Hijmans RJ, Huettmann F, Lohmann L. Presence-only and presence-absence data for comparing species distribution modeling methods. Biodiversity informatics. 2020 Jul 22;15(2):69-80.

Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological monographs. 2022 Feb;92(1):e01486.

Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S. and Mac Aodha, O., 2023, July. Spatial implicit neural representations for global-scale species mapping. In International conference on machine learning (pp. 6320-6342). PMLR.

Battaglia, P., Pascanu, R., Lai, M. and Jimenez Rezende, D., 2016. Interaction networks for learning about objects, relations and physics. Advances in neural information processing systems, 29.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W. and Merose, A., 2023. Learning skillful medium-range global weather forecasting. Science, 382(6677), pp.1416-1421.

Google Research

Paper link: