Convolution Neural Network Fine-tuning for Plant and Animal Species Distribution Modelling

Mélisande Teng[1,2], Sal. Elkafrawy[1,2]

[1]Université de Montréal
[2]Mila, Quebec AI institute

Abstract

Biodiversity is declining at an unprecedented rate, and understanding the current state of biodiversity is a first step towards tackling this challenge. In particular, it is important to understand which species can be found in a given location for biodiversity management and conservation. In these working notes, we present our submission to the 2022 edition of the GeoLifeCLEF challenge which aims at predicting species from remote sensing data, as well as the avenues we explored. Using altitude and land cover data along with remote sensing image RGB patches as input, our model performs better than the baselines and was ranked 5th in the challenge.

Keywords

Representation learning, Transfer learning, Self Supervision,

1. Introduction

Biodiversity is declining at an unprecedented rate, threatening the achievement of the Sustainable Development Goals (SDGs) worldwide. Understanding the current state of biodiversity and where species can be found is an important first step towards biodiversity conservation, and a way to do so is through species distribution modelling. Traditional methods in ecology leverage environmental data to predict species range in space and time but recent advances in computer vision and remote sensing offers new perspectives. Indeed, information derived from remote sensing, such as the NDVI (Normalized Difference Vegetation Index) which is computed using the Red and Near-Infrared bands of a remote sensing image, has been shown to be relevant variables to include in ecology species distribution models [1] to improve performance, and there is growing interest in aligning remote sensing products and essential biodiversity variables to establish frameworks for priority biodiversity metrics to observe from space [2].
In this context, the GeoLifeCLEF challenge[3] was set up and aims at predicting the list of plant and animal species that are the most likely to be observed at a given location from remote sensing data, as one of the components of the LifeCLEF challenge[4]. The organizers propose three baseline models : a random forest on environmental variables, and convolution neural networks on remote sensing imagery (R, G, B or R, G, near IR).

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
tengmeli@mila.quebec (M. Teng); sal.elkafrawy@gmail.com (Sal. Elkafrawy)
https://melisandeteng.github.io (M. Teng); https://www.linkedin.com/in/saraelkafrawy/ (Sal. Elkafrawy)

2. Data Description

The dataset is a cleaned-up version of the previous year's challenge dataset [5]. The dataset contains 1.6M observations of locations in the US and France. Each observation has the following information:

- Remote sensing imagery: 256mx256m RGB-IR patches centered at each observation with 1m pixel resolution
- Land cover data: 256mx256m patches centered at each observation with 1m pixel resolution
- Altitude data: 256mx256m patches centered at each observation with 1m pixel resolution
- Environmental variables: 19 low-resolution rasters of Bioclimatic data and 8 low-resolution rasters of Pedologic data

  Each location is associated with one of 17037 possible observed species, and in addition to the previous editions of the challenge, information about the genus and family of each species is provided.

3. Methodology

3.1. Building on the challenge baseline

The challenge organizers propose two CNN baseline models using a pre-trained ResNet50 architecture, with either R,G,B bands or R,G and near-IR bands as inputs, the latter performing better than the former.
Thus, after a first phase of replicating the challenge baselines, we decided to include the near-IR band in all the methods we considered.
Unless otherwise stated, we normalized the R,G,B and near-IR patches channel-wise using the means and standard deviations computed over the whole training set.

Adding input information

While the challenge baseline was using a pre-trained ResNet50 on R,G,near-IR bands, we used all four R, G, B and near-IR bands and initialize the R, G, B channels with the pretrained weights, and randomly initialize the input weights of the near-IR band. We trained the model with cross entropy loss, stochastic gradient descent optimizer without momentum and initial learning rate of 0.01, and used random horizontal and vertical flipping transformations for data augmentation. We will refer to this model as *Baseline++*.

We also added land cover and altitude information in the input by stacking extracted patches to the RGB-IR patches as two extra channels, matching the resolution to that of the aerial data. We also tried adding land cover only (and not altitude) but the model did not perform as well as the one having both land cover and altitude information.
We also tried training a model from scratch but the one using pre-trained weights (on ImageNet) performed better.

Satellite Imagery Pretrained Model

Instead of using a ResNet-50 pretrained on the ImageNet dataset we used a recently publicised pretrained model on satellite image dataset that was trained with self-supervision using MoCov2 [6] method called SeCo [7][1]. The dataset in SeCo has 1M images (to be comparable with ImageNet). We used the same setting as in our baseline with RGB-IR as input: batch size is 32, number of finetuning epochs is 66, start learning rate is 0.01 with SGD optimizer without momentum. we used a 'reduce on plateau' learning rate scheduler with patience value of 5. The top-k error rate on the validation set was close but not better than the baseline's top-k error rate. The model's name is SeCo in table(1).

One reason why SeCo is not better than a Resnet50 pretrained on ImageNet could be the difference in their type which led to different resolutions; SeCo is a satellite imagery dataset obtained from publicly available satellites, it has resolution of 10m, 20m and 60m. However, GeoLifeCLEF dataset is aerial imagery dataset which has 1m resolution as aerial images tend to have more details than satellite images. Also the heuristic involved to collect SeCo's dataset could be another reason; as the authors randomly sampled locations around the cities (within 50km radius).

3.2. Multimodal with Environmental variables

We also tried to incorporate the environmental variables as an additional learning signal to the extended baseline Resnet50 model. The environmental variables are the tabular data that were extracted at the point of occurrence. We passed the environmental variables tabular data through a small scale of a Resnet block and concatenated both outputs (from the RGB-IR patches and the environmental variables Resnet) in a multimodal way to predict the final target of 17k species. Figure(1) shows the model architecture. We used ResNet model, one of few models that are reported as SOTA methods for deep learning with tabular data[8][2]. The best model has a top-k error rate higher than the organizers baseline by a margin of 0.04 on the validation set. The model was trained for 70 epochs with no indication of overfitting in earlier epochs. The model's other hyperparameters are: batch size is 32, number of finetuning epochs is 66, start learning rate is 0.01 with SGD optimizer without momentum. we used a 'reduce on plateau' learning rate scheduler with patience value of 5.

3.3. Multitask learning

We explored a multitask learning setting where the tasks are :

- Species prediction (task S)
- Land cover classification (semantic segmentation task) (task LC)
- Country prediction (task C)

All tasks share the same encoder and we use different decoders for each of the tasks. For the encoder, we tried two architectures: Deeplabv2[9] and ResNet50 (up to the classification layer).

[1] https://github.com/ElementAI/seasonal-contrast/tree/main
[2] https://github.com/Yura52/rtdl#papers-and-projects

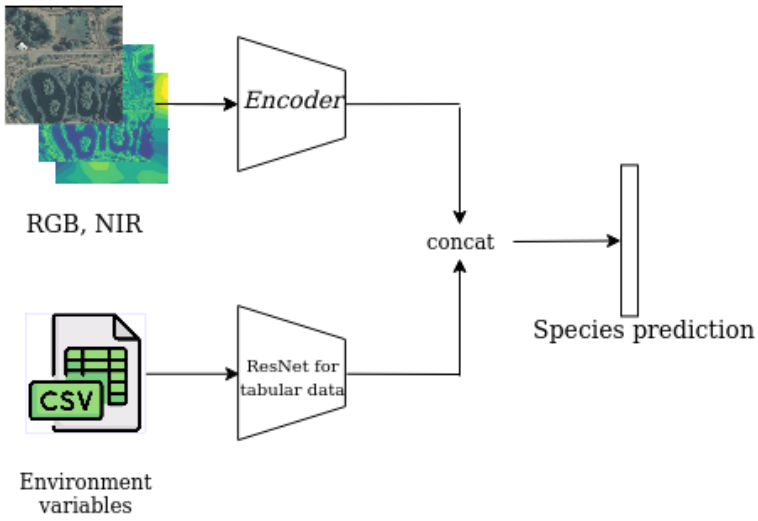


Figure 1: Multi-modal model architecture

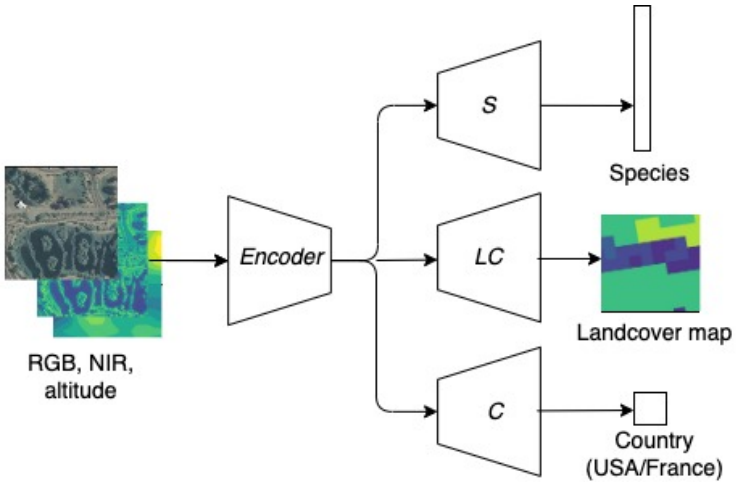


Figure 2: Multitask model architecture : all tasks share the same encoder and are trained jointly

We used a Deeplabv2 decoder for *LC*, a multi-layer perceptron for *S* and a linear layer for *C*. In both cases we used pretrained ResnNet-50 and Deeplabv2 on ImageNet weights for the encoder. The loss is the sum of the cross-entropy losses for the *LC* and *S* tasks and the binary cross-entropy for the *C* task.
However, our initial trials took too long to train, and due to time constraints, and need for optimization for faster training, we could not carry out these experiments fully. Hence we do not report our results in Table 1.

Model name	Validation top-30 error	Public test top-30 error
Challenge organizers Baseline	-	0.7059
Baseline++	0.7052	0.6889
SeCo	0.7191	0.7116
multimodal	0.7465	0.7416

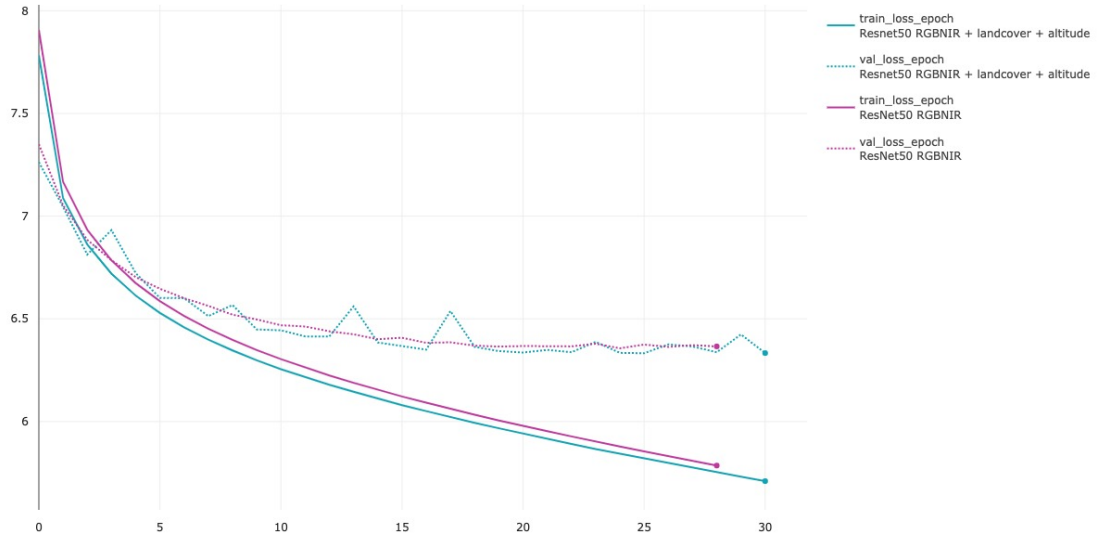
Table 1
Top-30 error values on the validation set and the public test set (computed over 10% of the entire test data)

4. Submission and discussion

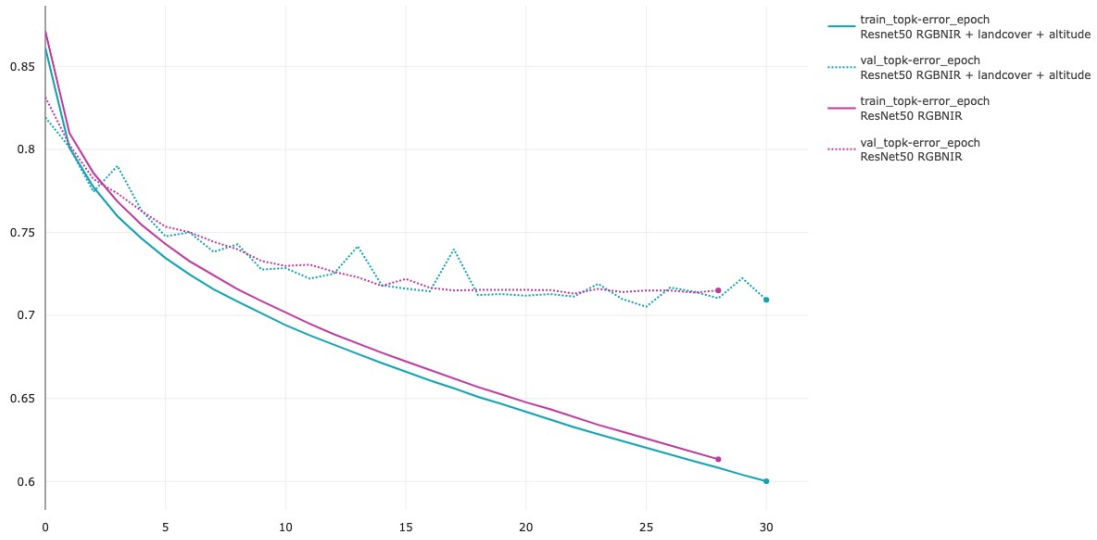
Our best model, which was chosen for submission to the challenge was the one built up on the baseline as described in 3.1 a ResNet50 with RGB, near-IR, altitude and landcover patches as input trained for 25 epochs. Its performance reported on the private leaderboard 0.7013 top-30 error, and ranked 5th among the submissions to the challenge. In Figure 3, we compare it with the same model using only RGB and near-IR as input and notice that adding altitude and land cover only reduces the validation top-k error by 0.01.

We also tried the same setting with a pretrained ResNet18 and Inceptionv3 architecture but the ResNet50 performed the best.

We were surprised that the other methods we tried did not improve much the challenge baseline, but we only did minimal hyperparameter tuning for each of the methods we tried. A main limitation we faced was the slow training of bigger models than ResNet50, especially in the multitask setting, as we trained our models on one GPU, but there is room for improvement of training speed. In the future, we would like to explore further the multitask learning setting, and adding data augmentations. We also would like to continue exploring models leveraging self-supervised learning as learning a good representation of remote sensing imagery could be meaningful for more tasks than the one defined by the challenge.



(a) Loss vs Epoch



(b) Top-K error vs Epoch

Figure 3: Comparison of ResNet50 with RGBNIR and RGBNIR+landcover+altitude as input. Training is in full lines and Validation in dotted lines. The addition of the landcover and altitude informations improves very slightly the top-k error.

References

[1] J. Pinto-Ledezma, J. Cavender-Bares, Predicting species distributions and community composition using satellite remote sensing predictors, Scientific Reports 11 (2021) 16448. doi:10.1038/s41598-021-96047-7.

[2] P. Stephenson, T. Brooks, S. Butchart, E. Fegraus, G. Geller, R. Hoft, J. Hutton, N. Kingston, B. Long, L. McRae, Priorities for big biodiversity data, Frontiers in Ecology and the Environment 15 (2017) 124–125. doi:10.1002/fee.1473.

[3] T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Joly, Overview of GeoLifeCLEF 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[4] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, M. Hruz, Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.

[5] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, A. Joly, The geolifeclef 2020 dataset, arXiv preprint arXiv: Arxiv-2004.04192 (2020).

[6] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, arXiv preprint arXiv:2003.04297 (2020).

[7] O. Mañas, A. Lacoste, X. Giro-i Nieto, D. Vazquez, P. Rodriguez, Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data, arXiv preprint arXiv:2103.16607 (2021).

[8] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data, Advances in Neural Information Processing Systems 34 (2021).

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062 (2014).