Location-Based Plant Species Prediction Using A CNN Model Trained On Several Kingdoms - Best Method Of GeoLifeCLEF 2019 Challenge

Mathilde Negri[3,4], Maximilien Servajean[2,3], Benjamin Deneu[1,3], and Alexis Joly[1,3]

[1] Inria, Montpellier, France
alexis.joly@inria.fr, benjamin.deneu@inria.fr
[2] LIRMM, Montpellier, France
mathilde.negri@lirmm.fr, maximilien.servajean@lirmm.fr
[3] Université Paul Valéry, Montpellier, France
[4] Université de Montpellier, France

Abstract. This technical report describes the model that achieved the best performance of the GeoLifeCLEF challenge, the objective of which was to evaluate methods for plant species prediction based on their geographical location. Our method is based on an adaptation of the Inception v3 architecture initially dedicated to the classification of RGB images. We modified the input layer of this architecture so as to process the spatialized environmental tensors as images with 77 distinct channels. Using this architecture, we did train several models that mainly differed in the used training data and in the predicted output classes. One of the main objective, in particular, was to compare the performance of a model trained with plant occurrences only to that obtained with a model trained on all available occurrences, including the species of other kingdoms. Our results show that the global model performs consistently better than the plant-specific model. This suggests that the convolutional neural network is able to capture some inter-dependencies among all species and that this information significantly improves the generalisation capacity of the model for any species.

1 Introduction

Predicting a list of the most likely species present at a given location can be very useful. First, it could improve species identification processes and tools by reducing the list of candidate species observable at a given location (whether automated, semi-automated or based on classical field guides or flora). More generally, it could facilitate biodiversity inventories through the development of location-based recommendation services (typically on mobile phones) as well as the involvement of non-expert nature observers. Finally, it could be used for educational purposes through biodiversity discovery applications providing innovative features such as contextualized educational pathways.

This challenge is related to the problem of **Species Distribution Modelling** (SDM) in ecology. SDM goal is to predict the spatial distribution of a species over a territory [1], in our case we used spatial positions and also environmental data. SDM have become increasingly important in the last few decades for the study of biodiversity, the ecology of conservation, landscape management, preservation of rare and/or endangered species, measurement of human impact or climate change on species, etc.

Concretely, the objective of SDM is to infer the spatial distribution of a given species, and they are often based on a set of geolocalized occurrences of that species (collected by naturalists, field ecologists, nature observers, citizen sciences project, etc.).

Recently, SDM based on deep neural networks have begun to appear [2]. These experiments have shown that they can have a good predictive power, potentially better than the models used conventionally in ecology such as MaxEnt [3]. Deep neural networks can learn complex non-linear transformations in a wide variety of fields. In addition, they provide an opportunity to learn an area of environmental representation common to various species, which stabilizes predictions from one species to another and improves them globally. Finally, spatial patterns of environmental variables often contain useful information for species distribution, but are generally not considered in conventional models. On the contrary, convolutional neural networks effectively use this information and improve their predictions.

In this paper, we present a study to evaluate a convolutional neural network to determine the ecological preferences of species through ranges of environmental image patches provided as input (temperature, soil type, etc.) as part of the GeoLifeCLEF challenge.

- Section 2 gives an overview of the various data we used to build our model.
- Section 3 provides the detailed description of our model.
- Section 4 presents the results of the experiments and their analysis.

2 Data

2.1 Occurrences

The data set protocol is explained in the challenge protocol [4].

- plantnet data set: composed of 2,367,145 plant species occurrences with uncertain identifications, because they come from the automatic identification of plant images, from the Pl@ntnet application.
- glc_18: Global Biodiversity Information Facility (GBIF) data set (same data set as last year challenge), it is composed of 291,392 occurrences of 3,336 plant species observed on French territory between 1835 and 2017.

– trusted data set: a sample of plantnet data set without uncertainty.
– no_plant: 10,618,839 species occurrences from the GBIF database, of other kingdoms (such as mammals, birds, amphibians, insects, fungus etc.). We have removed occurrences that didn't match any environmental rasters, for example birds in the middle of the sea.

Then, we filter the database to get only plant species that were given by the challenge.
We used different methods to structure our train data set from each data set.

– full_noplant: we train our model with all the data (plantnet, glc_18 and no_plant) we had including animal observations and species from other kingdoms.
– full_prediction: we only used the Pl@ntnet data set and glc_18 data set.
– filter_predictions: we try a random weighted selection scheme, the same as the one explained in the challenge notes for the test set. For each occurrence s_i in our data set, we compute a weight w_i which corresponds to the importance of this plant over an area of 2 kilometres. Then we did a random sampling on the z_i number between 0 and the maximum weight, and we kept the occurrence only if $z_i < w_i$. We did this sampling over 2 data set: the plantnet data set with uncertain identification set at 70%, we had a train set around 32,000 occurrences after sampling, and another with the data set pl_trusted, we had around 27,000 occurrences after sampling.

2.2 Environmental rasters

We have associated 33 rasters windows cropped into each corresponding global environmental rasters provided by the challenge as input features to each occurrence.[4] An extraction protocol is given by the challenge.[5]

These environmental rasters were constructed from various open data sets including Chelsea Climate, ESDB soil pedological data, Corine Land Cover 2012 soil occupation data, CGIAR-CSI evapotranspiration data, USGS Elevation data (Data available from the U.S. Geological Survey.) and BD Carthage hydrologic data. All these data try to best represent the environment where the plant is observed. We print them on a given area in the Figure 1, we can see the diversity of information we have for one occurrence near Montpellier.

In the following, we generally denote to x X an occurrence, each x being associated to a spatial position p(x) in the spatial domain D, a species label y(x) and an environmental tensor g(x) of size 64x64x33. We denote as P the set of all spatial positions p covered by X. It is important to note that a given spatial position $p_0 \in P$ usually corresponds to several occurrences $x_j \in X$, $p(x_j) = p_0$ observed at that location. In the training set, up to several hundreds of occurrences can be located in the same place.

The environmental data provided [4] is composed of tensors of 64x64X33 pixels. The corresponding 64x64 pixel matrices can be processed as classical

[5] http://www.github.com/maximiliense/GLC19

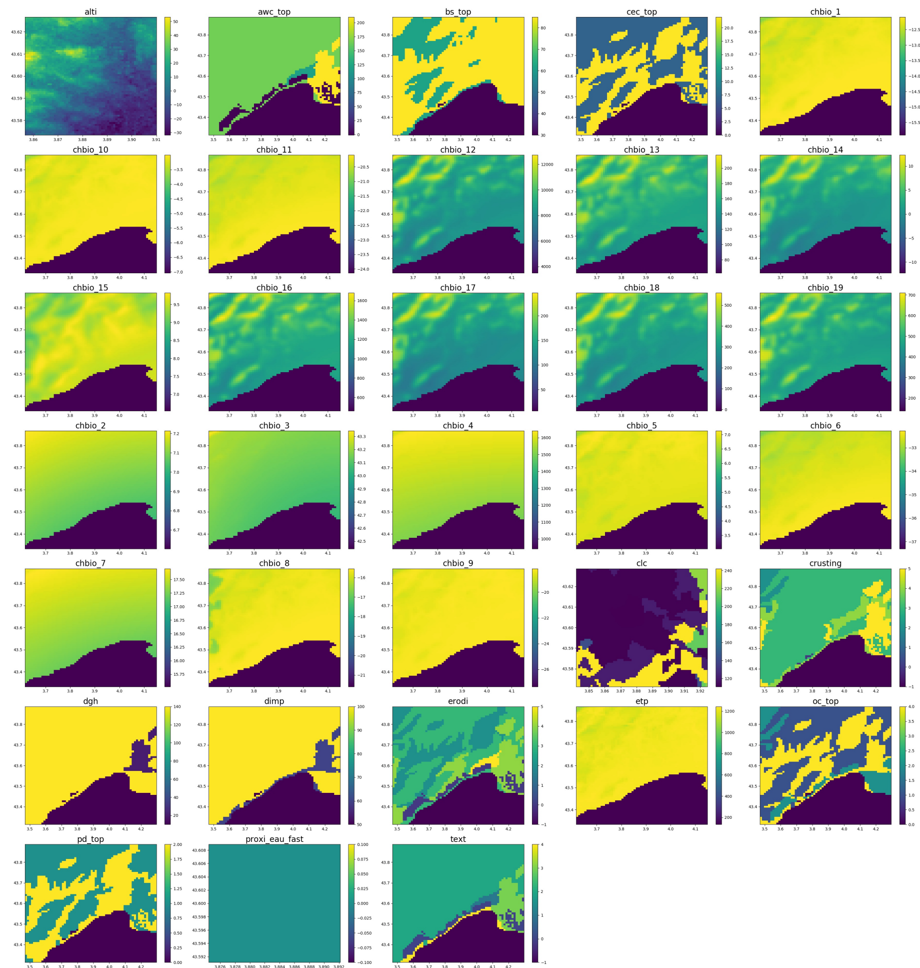

Fig. 1. Environmental rasters

image channels provided as CNN input. Most of them are continuous variables such as average temperature, altitude or distance to water as we can see them in the Figure 1. Thus, the corresponding 64x64 pixel matrices can be processed as classical image channels provided as CNN inputs. Some of the variables are ordinal type (such as ESDB v2), they can be considered as additional channels of the CNN in the sense that the order of pixel values is not significant.
For categorical variables like the Corine Land Cover soil type variable provided within GeoLifeCLEF. This variable can take up to 48 different categorical values, but 3 of them are not used therefore we have kept 45 categorical variables. The order of these values has no meaning. Consequently, we preferred unstacking the corresponding channel into 45 different binary channels, then concatenate it

with the other 32 tensors. Thus, instead of having a tensor of 64x64X33 pixels, we do have a tensor of 64x64x77 pixels.

3 Convolutional Neural Network

It has been previously shown in [2] that Convolutional Neural Networks (CNN) may reach better predictive performance than classical models used in ecology.

Our model attempts to predict the most likely species to be observed based on environmental features learned. Our structure is very different from last year challenge [5]. We used an Inception V3 convolutional neural network [6], it is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenates, dropouts, and fully connected layers. In addition to this model, we added a first layer of 77 layers, to match the size of the environmental raster and a dropout to avoid over fitting.

For the last layer we used a softmax and a cross entropy loss. The **softmax** layer is computed as:

$$p_j = \frac{e^{h_j}}{\sum_{k=1}^{K} e^{h_k}} \tag{1}$$

where y_k are the scores inferred by the net for each class, and K the number of different classes that means the number of species we trained our model with, the softmax activation for a class y_j depends on all the scores in y.
Then we used a **cross entropy loss** [7]:

$$Loss = -\sum_{k=1}^{K} y_k \log(p_k) \tag{2}$$

with y_k a binary indicator if class label k is the correct classification for the observation and p_k the probability of the observation is of class k.

Learning set up and parameters: all our experiments were conducted using PyTorch deep learning framework 4 and were run on a single computing node equipped with 4 Nvidia GTX 1080 ti GPU. We used the Stochastic Gradient Descent optimization algorithm with a learning rate of 0.1, a momentum of 0.9 and a batch size of 300.
We trained our model over multiple epochs, then to export our result we took the epoch where the model gets the highest Top 1 score in our validation set. For our best runs (full_noplant_predictions and full_predictions) we trained only over 20 epochs due to the number of data and for the two other runs we trained our model over 180 epochs.

4 Result

For GeoLifeCLEF 2019 they decided that the main evaluation criteria will be the accuracy based on the first 30 answers, also called Top 30, i.e. the function scoring 1 when the right specie is in the 30 first answers, and 0 if not. And as a

Rank	runId	top30	runName	participant
1	27007	0.1769	full_noplant_predictions	LIRMM
2	27086	0.1687	RUN2_GRINNELL_FULL_INITIAL	SaraSi
3	27087	0.1653	RUN3_GRINNELL_TRANS	SaraSi
4	27088	0.1646	RUN1_GRINNELL_TESTSPECIES_INITIAL	SaraSi
5	27006	0.1364	full_predictions	LIRMM
6	26997	0.1342	submit_xgb_spatial_4x4_all	SSN_CSE
7	26996	0.1288	submit_xgb_spatial_allnoclc	SSN_CSE
8	27013	0.1273	submit_xgb_dep3_1	SSN_CSE
9	27069	0.1268	submission_sel_4x4	SSN_CSE
10	27012	0.1263	submit_xgb_4x4_all_dep3	SSN_CSE
11	27070	0.1227	submission_sel_1	SSN_CSE
12	27064	0.1198	submission1x1	SSN_CSE
13	27067	0.1135	submission4x4	SSN_CSE
14	27124	0.1135	Lot_Of_Lof_2	Lot_of_Lof
15	27089	0.1110	RUN4_ELTON_TRANS	SaraSi
16	27082	0.1090	RUN0_ELTON_FULL_INITIAL_TESTSPECIES	SaraSi
17	26988	0.1063	submit_xgb_spatial	SSN_CSE
18	27123	0.0984	Lot_Of_Lof_3	Lot_of_Lof
19	27063	0.0864	Lot_Of_Lof_1	Lot_of_Lof
20	26875	0.0844	submission	SSN_CSE
21	27102	0.0834	rfspatial	SSN_CSE
22	26821	0.0570	submit	SSN_CSE
23	27004	0.0470	plcomplete_predictions	LIRMM
24	27005	0.0465	inception_glc19_filter_predictions	LIRMM
25	26968	0.0205	run_14	sergiu_atodiresei
26	26964	0.0191	run_10	sergiu_atodiresei
27	26961	0.0190	run_7	sergiu_atodiresei
28	26971	0.0184	run_17	sergiu_atodiresei
29	26967	0.0180	run_13	sergiu_atodiresei
30	26960	0.0168	run_6	sergiu_atodiresei
31	27062	0.0159	run_20	sergiu_atodiresei
32	26958	0.0146	run_3	sergiu_atodiresei
33	26970	0.0102	run_16	sergiu_atodiresei
34	26969	0.0099	run_15	sergiu_atodiresei
35	26972	0.0089	run_18	sergiu_atodiresei
36	26963	0.0079	run_9	sergiu_atodiresei
37	26965	0.0068	run_11	sergiu_atodiresei
38	26926	0.0067	run_4	sergiu_atodiresei
39	26973	0.0064	run_19	sergiu_atodiresei
40	26959	0.0063	run_5	sergiu_atodiresei
41	26962	0.0062	run_8	sergiu_atodiresei
42	26957	0.0061	run_2	sergiu_atodiresei
43	26966	0.0058	run_12	sergiu_atodiresei
44	26956	0.0058	run_1	sergiu_atodiresei

Table 1. Table of Result of GeoLifeCLEF 2019 challenge

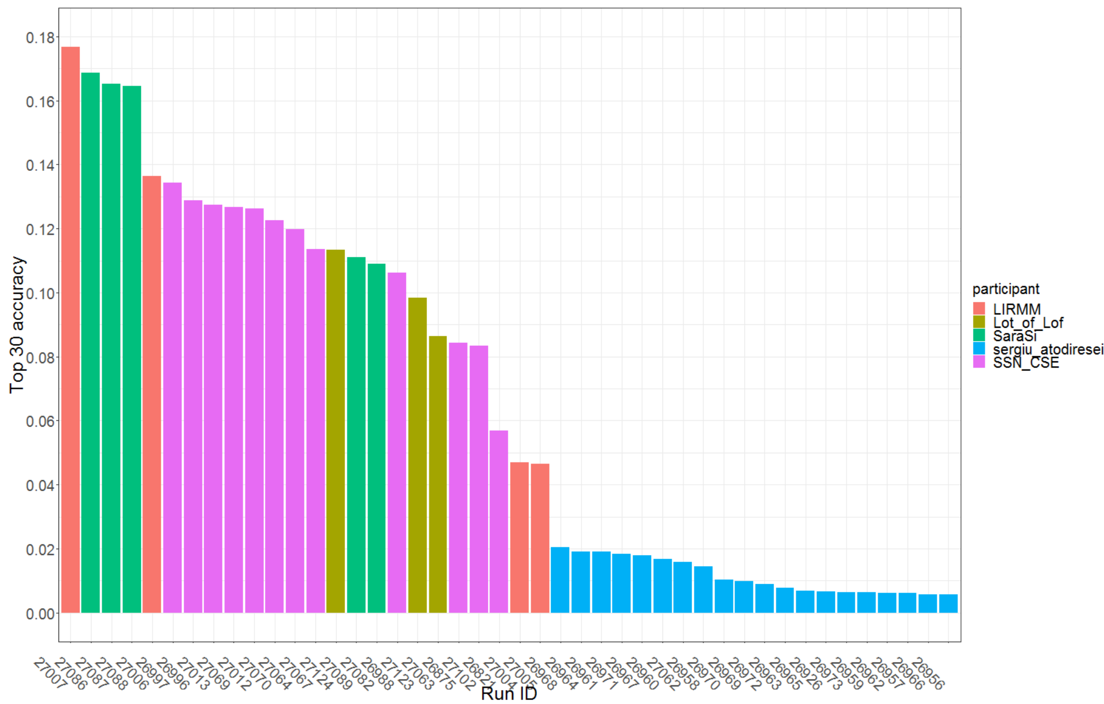

Fig. 2. Top30 accuracy score of every participant of GeoLifeCLEF 2019

second metric they use the Mean Reciprocal Rank (MRR) to compare with the GeoLifeCLEF 2018 challenge.
For this challenge we submitted 4 different runs. We can see the global result in Figure 2 and Table 1.

- The first one: full_noplant_predictions is a CNN trained with every data set (plantnet, no_plant, glc_18). We used pl@ntnet database without any filters (2,367,145 plant species occurrences), train and test occurrences from last year's challenge (file GLC_2018.csv) and occurrences from other kingdoms (such as mammals, birds, amphibians, insects, fungus etc.) from the GBIF database (file noPlant.csv). There are 34,375 different classes, so this CNN has a larger last layer output, but for the test we let it choose only the plant classes which are provided by the challenge. This method showed the best result with a **Top 30 of 0.1769** and a MRR of 0.031.
- The second submission, full_predictions is a CNN trained with plantnet data set and the data set of last year challenge. This means that we only have plants to train our model, we have 3,859 classes. It is the 5 th submission of the challenge, after SaraSi models with a **top 30 of 0.1364**.
- The two other submissions were not as good (plcomplete_predictions and inception_glc19_filter_predictions), we sampled our train set as explained in

the challenge protocol for the test set. Therefore, we have reduced our data, our model didn't capture as much information as the other models.

In this case, it seems that the more data available to the CNN, the better the results. It seems that our model is learning from the species from another kingdoms that live in the same area. Indeed, our model train with more than only plant occurrences showed a better result for predicting plants. Here, the CNN has more classes but can classify plants even better. We can deduct from this that our CNN network not only captures environmental information but also, the interaction between different species, we can imagine that it can also learn from the information of the species that live around the plants and perhaps from the ecological niche of an occurrence.

5 Conclusion

This paper describes our participation in GeoLifeCLEF challenge to evaluate location-based species prediction models. We compared Convolutional Neural Network trained with different data sets. The main conclusion of our study is that the convolutional neural network model is the most efficient model and it can learn information not only from environmental rasters but also from interaction between species of other kingdoms. Indeed, it achieved the best performance of the whole GeoLifeCLEF challenge when we trained our model with other kingdom observations.

In future work, we will attempt to better understand what information the CNN does capture from different data sets and how it could be improved.

References

1. Kearney, M., Porter, W.: Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters **12**(4) (2009) pp.334–350
2. Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F.: A deep learning approach to species distribution modelling. (2018)
3. Phillips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. Ecological Modelling **190**(3) (2006) pp. 231–259
4. Botella, C., Servajean, M., Bonnet, P., Joly, A.: Overview of geolifeclef 2019: plant species prediction using environment and animal occurrences. In: CLEF working notes 2019. (2019)
5. Deneu, B., Servajean, M., Botella, C., Joly, A.: Location-based species recommendation using co-occurrences and environment-GeoLifeCLEF 2018 challenge. In: CLEF: Conference and Labs of the Evaluation Forum. Volume CEUR Workshop Proceedings., Avignon, France (September 2018)
6. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR **abs/1512.00567** (2015)
7. de Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of Operations Research **134** (Feb 2005) pp. 19–67