Cooperative learning for biodiversity monitoring: what's new and what's next in Pl@ntNet?

PART I
Pl@ntNet overview
A citizen science platform that uses machine learning to help people identify plants with their mobile phones
25 Million users
200+ countries
Up to 2M identifications per day

Personal Usage
- Nature, walks
- Gardening
- Phytotherapy

Professional Usage
- Agro-ecology
- Natural Areas Management
- Education, animation
- Tourism
- Trade
Key concept of Pl@ntNet: Cooperative Learning

User

Pl@ntNet app

AI model

Collaborative data revision & enrichment

Biodiversity data portals

GBIF

INPN

training data

recognized species

predictions

Plant

obs
PlantNet  AI model

Model trained with the cross-entropy loss on the set of valid observations (Jean Zay, a few days of training)

Production version: Convolutional Neural Network (IV3)  →  Top1 accuracy = 0.70
Beta version: Vision transformer (BEIT)  →  Top1 accuracy = 0.73

46K species (+ reject classes)
5M training images (undersampling for classes > 1000 images)
A difficult problem: uncertainty

Aleatoric uncertainty
Ambiguity (irreducible)

Epistemic uncertainty
Long-tail distribution

Images

Likely classes

Nymphaea noouchali

Hypericum graniflorum

Hypericum hircinum L.

Trifolium repens L.

Trifolium subteraneum L.

Sphaerocephalus umbrosus

Medicago arabica

Top1 accuracy > Macro-average Top1 accuracy

0.73 > 0.59

Nb of training images

45K species
Returned results: set-valued

Pointwise error control
Threshold the accumulated probability
\[
\sum_i \sigma_i(f(x)) > \theta'
\]

Average set size control
Threshold the probability so as to return \(K\) classes on average
\[
\sigma_i(f(x)) > \theta
\]

\begin{align*}
\text{Papaver rhoeas L.} & \quad 0.63 \\
\text{Papaver somniferum L.} & \quad 0.76 \\
\text{Papaver californicum A.} & \quad 0.87 \\
\hline
\text{Glaucium corniculatum L.} & \quad 0.94 \\
\text{Glaucium flavum L.} & \quad 0.98 \\
\end{align*}

\begin{align*}
\hline
\text{Glaucium corniculatum L.} & \quad 0.94 \\
\text{Glaucium flavum L.} & \quad 0.98 \\
\end{align*}

\(0.1\)

\(0.95\)

\(0.1\)

\(0.04\)

\text{→ Average-K classification (proof of consistency)}
Use of regional or thematic floras

Restricting the hypothesis space to a particular flora allows improving the identification accuracy

\[ p(y| x, \text{flora}) \geq p(y| x) \]

Thematic floras
- Useful plants
- Useful plants
- Useful plants

Regional floras
- Central America
- Brazil
- Europe Central
- Europe SW

Backbone (all species)
Use of regional or thematic floras

Query

Identify in World flora

**Schismus arabicus Nees**
Arabian grass
Poaceae  74.23%

**Schismus barbatus (L.) Thell.**
Arabian grass
Poaceae  17.16%

Use of regional or thematic floras

Identify in West Europe

Schismus arabicus Nees
Arabian grass

Schismus barbatus (L.) Thell.
Arabian grass

Query
PlantNet Similarity search

User’s visual control = uncertainty reduction

convolutional neural network

hash-based Index

query results = similar pictures

Papaver rhoeas L.

similarity search engine

9M images

→ Sub-linear algorithm based on locality sensitive hashing

Contribution

Users can contribute their observations
Users can revise observations of other users.
Cooperative learning

The weight of a user in the decision process depends on his estimated expertise

Most probable species \( y = \arg \max_j \hat{\eta}_j(x) \)

Validation decision (valid → used by AI) \( \hat{\eta}_y(x) > \theta \)
750M raw observations (=queries)

Not Geo-localized (52%)
Geo-localized (48%)

7.9M valid observations
16.6M shared observations

Correct species name + sufficient confidence score + sufficient quality score

User opted in to share it publicly (GDPR compliant)

→ Used to train the AI

ANONYMOUS (cc0)

AUTHENTICATED (cc-by-sa)
- Top-4 data provider to GBIF (world’s largest infrastructure for biodiversity data)
- Valid observations + trusted queries identified by the AI (AI score > 0.9)
- Additional quality filters: potted & cultivated plants removal, region-based filtering (Kew POWO)

13,856,500 occurrences
(87% identified by AI, 13% by humans)

https://doi.org/10.15468/mma2ec
A secured API providing developers programmatic access to Pl@ntNet engine

6K developer accounts (researchers, companies, citizen observatories)

Integrated in European Open Science Cloud (EOSC)
Pl@ntNet offline: identify plants without connection

User

Pl@ntNet frontend (mobile app)

Plant

Embedded AI model (compressed)

obs

recognized species

Local storage

Pl@ntNet backend (cloud)

AI model

predictions

Pl@ntNet database

Latest major developments
PART II

Latest cooperative learning algorithm
Cooperative Learning algorithm in detail

The most probable label of an observation is determined with a weighted majority voting rule:

$$y_i = \arg \max_k \sum_{u \in U_i} w_u \mathbf{1}(y_i^u = k)$$

$$U_i = \text{Set of users who provided a label } y_i^u \text{ for the observation } x_i$$
Cooperative Learning algorithm in detail

Unlike most state-of-the-art crowdsourcing approaches, the weight of a user is not determined by his estimated probability of success.

\[ \pi(u) = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.7 \end{pmatrix} \]

Problems:
- Not tractable for 45K classes
- Very sparse data for most users and species
- A user might be highly successful but only on a few very common species
- User scores interpretability (people love leaderboards)
Cooperative Learning algorithm in detail

Rather, the weight of a user in Pl@ntNet is a function of the estimated number of species he is able to identify

\[ w_u = g(n_u) \quad n_u = \left| \left\{ j : \exists i \ y_i^u = y_i \right\} \right| \]
Cooperative Learning algorithm in detail

Rather, the weight of a user in Pl@ntNet is a function of the estimated number of species he is able to identify:

\[ w_u = g(n_u) \]

\[ n_u = \left| \left\{ j : \exists i y^u_i = y_i \right\} \right| \]

New user \( n_u = 0 \), \( w_u = 0.3 \)

Early career contributor \( n_u = 5 \), \( w_u = 1.2 \)

Alexis Joly \( n_u = 321 \), \( w_u = 15.0 \)

Pierre Bonnet (botanist) \( n_u = 1145 \), \( w_u = 30.0 \)
Cooperative Learning algorithm in detail

Practically, $n_u$ is estimated from the set of valid observations for which the user has suggested the correct species first

$$n_u = |\{j : \exists i \ y^u_i = \hat{y}_i \ | \ v(x_i) = 1\}|$$

Where $v(x_i)$ is a function that determines if an observation is valid or not:

$$v(x_i) = \begin{cases} 1 & \text{if } s_{y_i}(x_i) > \theta, \eta_{y_i}(x_i) > \theta_\eta \\ 0 & \text{otherwise} \end{cases}$$

Confidence score (~ quantity of votes)

$$s_{y_i}(x_i) = \sum_{u \in U_i} w_u 1(y^u_i = y_i)$$

Agreement score (~ species proba)

$$\eta_{y_i}(x_i) = \frac{w_{y_i}(x_i)}{\sum_k w_k(x_i)}$$
Cooperative Learning algorithm in detail

Parameters are estimated through an iterative algorithm similar to expectation-maximisation:

**Initialization:**

\[ w_u = w_0 \quad \text{for all users} \]

**Repeat until convergence:**

\[ y_i = \arg \max_k \sum_{u \in U_i} w_u 1(y_i^u = k) \quad \text{Most likely labels} \]

\[ s_{y_i}(x_i) = \sum_{u \in U_i} w_u 1(y_i^u = y_i) \]

\[ \eta_{y_i}(x_i) = \frac{w_{y_i}(x_i)}{\sum_k w_k(x_i)} \quad \text{Confidence and agreement scores} \]

\[ v(x_i) = \begin{cases} 1 & \text{if } s_{y_i}(x_i) > \theta, \eta_{y_i}(x_i) > \theta_\eta \\ 0 & \text{otherwise} \end{cases} \]

\[ n_u = |\{ j : \exists i \ y_i^u = \hat{y}_i \ | \ v(x_i) = 1 \}| \quad w_u = g(n_u) \quad \text{Update user weights} \]
Cooperative Learning algorithm in detail

A new iteration is ran each night but only on new incremental data:

1 - Update user weights for
   - users who voted since last iteration
   - users who created new observation(s) since last iteration
   - users whose observations received a vote since last iteration

2 - Compute validity score for
   - new observations created since last iteration
   - updated observations since last iteration (including the ones with new votes)
   - observations having a vote whose author has had its weight modified since last iteration

Computation time: from 2 to 3 hours depending on the volume of new data (e.g. longer the week-end)
Cooperative Learning algorithm in detail

Valid observations (i.e. $v(x_i) = 1$) are the only ones:
- used for training the AI
- appearing in Pl@ntNet galleries
- appearing in the identification results (visual similarity search)

Papaver argemone

A valid observation can be revised at any time within the application so that the label noise can be reduced afterwards
Cooperative Learning algorithm in detail

New observations

Appear only once in the contribution stream → they can be revised/confirmed on the fly (low rate)

They can be directly valid if the author has a sufficient weight

\[ w_u > \theta \quad \rightarrow \quad s_{y_i}(x_i) = \sum w_u 1(y_i^u = y_i) > \theta \]

Such users are said self-validating (\( \theta = 2.0 \))

Obs of self-validating users can be unvalidated by a user with similar weight:

\[ \frac{w_u}{w_u + w_{u'}} < \theta' \quad (\theta' = 0.7) \]
Contributors

4M users accounts, 1M active contributors

**Top 10 contributors**

<table>
<thead>
<tr>
<th>#</th>
<th>Weight</th>
<th>Species count</th>
<th>Observations</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78.14</td>
<td>6932</td>
<td>17627</td>
<td>Diego Alex</td>
</tr>
<tr>
<td>2</td>
<td>65.43</td>
<td>4923</td>
<td>16408</td>
<td>Daniel Barthelemy</td>
</tr>
<tr>
<td>3</td>
<td>60.76</td>
<td>4269</td>
<td>15868</td>
<td>Lilane Roubaud</td>
</tr>
<tr>
<td>4</td>
<td>53.81</td>
<td>3381</td>
<td>13653</td>
<td>Maarten Vanhove</td>
</tr>
<tr>
<td>5</td>
<td>52.45</td>
<td>3219</td>
<td>11567</td>
<td>Yoan Martin</td>
</tr>
<tr>
<td>6</td>
<td>51.35</td>
<td>3091</td>
<td>11209</td>
<td>Dieter Albrecht</td>
</tr>
<tr>
<td>7</td>
<td>49.3</td>
<td>2859</td>
<td>10463</td>
<td>Michal Svit</td>
</tr>
<tr>
<td>8</td>
<td>49.06</td>
<td>2832</td>
<td>9964</td>
<td>William Coville</td>
</tr>
<tr>
<td>9</td>
<td>46.46</td>
<td>2552</td>
<td>9210</td>
<td>Martin Bishop</td>
</tr>
<tr>
<td>10</td>
<td>46.25</td>
<td>2530</td>
<td>8757</td>
<td>Sylvain Piry</td>
</tr>
</tbody>
</table>

**Typical contributor**

Weight = 9.0

Rossen Vassilev

**Stats**

- **Rank**: 14,062
- **Observations**
  - Observed species 134
  - Contributions 143
  - Images 463
- **Queries**
  - Identification requests 520
  - Images 1,005

**Votes**

- Votes 54
Active learning

*Corydalis cava* (L.) Schweigg. & Körte
Hollowroot, Hollow Root, Hollow Wort, Holewort, Brebenea
Active learning

*Corydalis cava* (L.) Schweigg. & Körte
Hollowroot, Hollow Root, Hollow Wort, Holewort, Brebenea

We believe that the determination of these images may be wrong.
Active learning

Pl@ntNet

Ask Pl@ntNet AI

species A: 49%

species B: 35%
Other collaborative tools

User page

Rank 985

Top contributors

Else Nolden 33 observations
Dražen Vrančićević 8 observations
Dieter Wagner 6 observations
Tela Botanica 16 observations
Gradwohl Markus 6 observations

Top identifiers

Else Nolden 63 votes
marie p. Ruf 29 votes
Sánchez García Juan... 24 votes

Messages

davidhocken il y a 2 heures
English : Hacquetia epipactis (Scop.) DC. is a synonym of Sanicula epipactis

davidhocken il y a 2 heures
English : powo.science.kew.org/taxon/urn:lsid:ipni.org:names:847830-1
PART III
Deep Species Distribution Modeling
Objective: which species are present in a given location and why?

Raw species occurrence data needs to be interpolated in space and time:

Many plant occurrences at world scale

But very few locally for most species

Viola canina L.
Species Distribution Models (SDM)

- Data
  - Environment
  - Plant observations

- Model
  - Modelled environmental distribution

- Projection
  - Predicted distribution
  - Geographical space
Species Distribution Models (SDM)

Motivations

- Help conservation plans
- Invasive plant monitoring
- Learn about species preferences
- Simulation under climate change
A deep learning approach to species distribution modelling


- NN can model complex relationships from heterogeneous data sources
- Learn a joint representation space $f(x)$ of the environment for all species ($\approx$ latent variables)
- Capturing multi-scale spatial information thanks to convolutional layers (CNN)

10x10 km quadrats
Understanding Deep Convolutional SDMs

Benjamin Deneu et al., "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment", *PLOS Computational Biology*

- Better knowledge transfer to least frequent species

**Model**
- **Architecture**: Inception v3
- **Loss**: categorical loss

**Data**
- **Source**: GBIF
- **Type**: occurrences
- **Nb of occurrences**: 97 683
- **Nb of species**: 4520
- **Environmental data**: 33 geographic rasters (19 bioclimatic, 1 evapotranspiration, 10 pedologic, altitude, 1 hydro, Corine Land Cover)
Understanding Deep Convolutional SDMs

Benjamin Deneu et al., "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment", PLOS Computational Biology

- Better knowledge transfer to least frequent species

**Senecio cacaliaster Lam.**

**Ulva lactuca L.**
Spatial structure of the local environment plays an important role in species distribution (landscape, barriers, relief, etc.)

- Deriving knowledge from Deep SDMs

- Complete information (structure+values)
- Only structural information (standardized values)
- Only values (no structure)
- Mean value (no structure)
- No spatial information (central value)
How to train Deep SDM models?

Input data: \( x \)

- **Abundance data** (very hard to produce)
  
  Task: predict \( \hat{y} = f_\theta(x) \in \mathbb{R}^d \)

- **Presence / absence data** (hard to produce)
  
  Task: predict \( \hat{y} = f_\theta(x) \in [0, 1]^d \)

- **Presence only data** (more data available)
  
  Task: predict \( \hat{y} = f_\theta(x) \in \{1, \ldots, d\} \)

<table>
<thead>
<tr>
<th>Input data: ( x )</th>
<th>Target: ( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundance data</td>
<td>Presence only data</td>
</tr>
<tr>
<td>Presence / absence data</td>
<td>Presence only data</td>
</tr>
</tbody>
</table>

Abundance data:

- 0
- 12
- 0
- 4
- 0
- 0
- 32
- 0

Presence / absence data:

- 0
- 1
- 0
- 1
- 0
- 0
- 1
- 0

Presence only data:

- 1
Predicting species assemblages from presence only data

Given presence-only occurrences 
\[(x_1, y_1), \ldots, (x_{n_t}, y_{n_t})\] sampled from \(P_{X,Y}\)

The assemblage of species likely to be present conditionally to \(x\) can be defined as:

\[S^*_\lambda(x) := \{k \in \mathcal{Y} : P_{X,Y}(Y = k|X = x) \geq \lambda\}\]

If we have an estimator \(\hat{n}_k(x)\) of \(P_{X,Y}(Y = k|X = x)\)

We can define the following plug-in estimator of the assemblage:

\[S_\lambda(x) := \{k \in \mathcal{Y} : \hat{n}_k(x) > \lambda\}\]
Predicting species assemblages from presence only data

How to get a good estimator $\hat{\eta}_k(x)$ of the conditional probability $\eta_k(x)$?

→ Train a model using the **negative log-likelihood** = a **strictly proper loss**, i.e.
it is minimized only when the model predicts the true

$$\eta_k(x) = \mathbb{P}_{X,Y}(Y = k \mid X = x)$$

$$\arg\min_\theta \sum_i -\log \hat{\eta}_{y_i}(x_i) \quad \text{e.g. with} \quad \hat{\eta}_k(x) = \frac{\exp(f^k_\theta(x))}{\sum_j \exp(f^j_\theta(x))} = \text{neural network output}$$

In brief:
- Our plug-in predictor simply consists in **thresholding the softmax output**
of a neural network trained with the so-called **cross-entropy** loss

$$S_\lambda(x) := \{ k \in \mathcal{Y} : \hat{\eta}_k(x) > \lambda \}$$

- It is proved that $S_\lambda(x)$ assymptotically converges towards $S^*_\lambda(x)$
GeoPl@ntNet
Discover plant biodiversity around you and help protect it better
Mapping biodiversity conservation indicators

From the species assemblage

\[ S_\lambda(x) := \{ k \in \mathcal{Y} : \hat{\eta}_k(x) > \lambda \} \]

We can compute indicators such as:
- The proportion of endangered species (e.g. on IUCN red list)
- The proportion is woody species
- The diversity of species (e.g. Shanon index)

We can construct maps of such indicators at very high resolution by computing \( S_\lambda(x) \) for all \( x_i \) on a dense spatial grid
Proportion of endangered species (Orchid Family, 14K species)

1x1 km resolution (view online)  PhD of Joaquim Estopinan
PART IV
Other ongoing stuff
GeoLifeCLEF challenge 2023

Presence / absence of 10K plant species

1 0 0 0 0 1 0 1 0 0 1 1 0 1 0 1 0 0 1 0

Model

5 Millions training samples

INPUT PREDICTORS
- Satellite image (sentinel 2)
- Multi-spectral time series (Landsat)
- Climatic time series (Chelsea)
- Environmental rasters (land use, human footprint, bioclim, soil)

OUTPUT PREDICTIONS
New biodiversity monitoring approaches

- Car views for the monitoring of invasive species (human vector)
- Quadrat images for the monitoring of vulnerable habitats or fields biodiversity
- Drones for the monitoring of forest canopies
New biodiversity monitoring approaches

- Car views for the monitoring of invasive species (human vector)
- Quadrat images for the monitoring of vulnerable habitats or fields biodiversity
- Drones for the monitoring of forest canopies
Habitats mapping and future trajectories prediction

PhD thesis of Cesar Leblanc

Input data = tabular data
- abundance
- presence/absence

Species-to-habitat classifier

Habitat N14
Mediterranean shifting coastal dune
Plant disease identification
- Collaborative epidemiology surveillance
- Reduction of phytosanitary products
- Jointly with Phytia

Identification of infra-specific taxa
- Crop varieties, horticol varieties, cultivar, hybrids, etc.
- Towards a selection more respectful of the environment
Handling uncertainty and bias of species identification

**Advanced optimization** techniques
- Uncertainty: top-K loss function
- Imbalance: shifting of the decision frontier

<table>
<thead>
<tr>
<th>K</th>
<th>$\ell_{CE}$</th>
<th>$\ell_{K,0.01,5,\max m_\gamma=0.2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.3±0.3 (12.6/42.9/71.7)</td>
<td>42.4±0.3 (23.9/46.3/72.1)</td>
</tr>
<tr>
<td>3</td>
<td>58.8±0.4 (32.4/75.3/92.0)</td>
<td>64.9±0.4 (44.8/74.5/92.1)</td>
</tr>
<tr>
<td>5</td>
<td>68.7±0.2 (45.1/86.3/95.4)</td>
<td>73.2±0.5 (55.3/84.2/95.3)</td>
</tr>
</tbody>
</table>

Statlearn poster today:

Thank you