Pl@ntNet: towards the recognition of the world’s flora

Alexis Joly et al.
Challenge

• More than **369K species** of flowering plants in the world

• Increasing our knowledge of them is of crucial importance
  – Health
  – Food crisis
  – Biodiversity crisis

• However, the *taxonomic gap* is penalizing the aggregation of new data and knowledge
  – Only specialists can identify plants
  – Specialists cannot carry the burden of all routine identifications
  – Particularly in south countries with the richest biodiversity
An innovative **citizen science** platform making use of **machine learning** to help people **identify plants** through their mobile phone.
Image Recognition Technology: Convolutional Neural Networks
Image Recognition Technology: Similarity search

Similarity search index (high-dimensional hashing)

Visual features extraction

Species 1
Species 2
Species N
More than 8M downloads
More than 60k - 100K users / day
11 languages
17K species (illustrated by 1M revised images)
22 projects & micro-projects
35M raw plant images / 55M users sessions
12K followers on social networks

In 2018: 3,352,788 users in 235 countries

<table>
<thead>
<tr>
<th>Sessions</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than 5 sessions</td>
<td>1,866,423</td>
</tr>
<tr>
<td>More than 10 sessions</td>
<td>1,293,698</td>
</tr>
<tr>
<td>More than 25 sessions</td>
<td>735,666</td>
</tr>
<tr>
<td>More than 100 sessions</td>
<td>96,167</td>
</tr>
</tbody>
</table>

1. France 641,569 (19.19%)
2. Germany 345,933 (10.30%)
3. United States 345,880 (10.34%)
4. Italy 282,842 (8.46%)
5. Spain 180,291 (5.39%)
6. Brazil 172,949 (5.17%)
7. Netherlands 101,057 (3.02%)
8. India 96,576 (2.89%)
9. United Kingdom 86,670 (2.59%)
10. Belgium 79,050 (2.36%)
22 projects around the world

Based on a wide international partnership:

- Univ. TEC (Costa-Rica),
- Univ. Los Andes (Bolivie),
- Univ. Bobo-Dioulasso (Burkina F.),
- Univ. Nat. Maurice (Maurice),
- National herbarium of Comores,
- Botanical Garden Geneva (Switzerland)
- National parks
- NGOs (Tela Botanica, iScanTree, Endémia, ..)
Pl@ntNet Mobile App Usage

- Professional usage: 12%
- Personal usage: 88%
Pl@ntNet Mobile App Usage

Personal usage (88%)

- Houseplants
- Gardening
- Phytotherapy, eatable
- Walk, jannie, trekking
- Fun, delusional

12% Professional usage
88% Personal usage
Pl@ntNet Mobile App Usage

Agriculture & Agri-food industry (4.8%)
Pl@ntNet Mobile App Usage

Professional usage: 12%
Personal usage: 88%

Education & animation (3.2%)
Pl@ntNet Mobile App Usage

Other professional usage (4%)

- Professional botanists, consulting, expertise
- Merchants
- Natural area management
- Tourism
Infrastructure

Images Cluster
Hadoop / HBase / HDFS

Node

Node

Node

Node

Front API

Backup
ArangoDB

DataBase
ArangoDB

Data Access Layer
Node.js

Public API
Node.js

Private API
Node.js

Partner API
Node.js

Identification service

Server 1
Server 2

Partners

Pro API
Node.js

Android
iOS
Web

Public fronts

Private fronts

Customers
Infrastructure

Contextualized Projects

Large regional or thematic projects: e.g. "western europe", "Hawai", "Useful plants"

Micro-project: a very specific project dedicated to a small region, or a dedicated flora
All screens are contextualized with the project’s species of interest.
Infrastructure: Micro-projects

- Currently 3 micro-projects, several others in discussion
Infrastructure: Pro API

• Currently experimented by 15 beta-testers (app developers)
  – Start-ups: BiodivGo, NaturalSolutions, ecoBalade, Garden-answers, Jardin Imaginaire, etc.
  – Universities, public bodies, associations, student projects
Infrastructure

Research projects driven by PI@ntNet data

Images Cluster
Hadoop / HBase / HDFS
Research projects in plant sciences

- Two examples of projects centrally driven by Pl@ntNet data

### Invasive species distribution models

<table>
<thead>
<tr>
<th>Species names</th>
<th>Species distribution computed from Pl@ntNet data</th>
<th>New invasion occurrences</th>
<th>Risks of invasion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acer negundo L.</td>
<td>Invasibility: 0.8</td>
<td>Species risk: 0.72</td>
<td><img src="image1" alt="New invasion occurrences" /></td>
</tr>
<tr>
<td>Cercidium obtusum (L.) Torr. &amp; A. Gray</td>
<td>Invasibility: 0.7</td>
<td>Species risk: 0.69</td>
<td><img src="image3" alt="New invasion occurrences" /></td>
</tr>
<tr>
<td>Eriogonum lanatula DC.</td>
<td>Invasibility: 0.5</td>
<td>Species risk: 0.67</td>
<td><img src="image5" alt="New invasion occurrences" /></td>
</tr>
<tr>
<td>Opuntia fasciculata (L.) Papp</td>
<td>Invasibility: 0.5</td>
<td>Species risk: 0.65</td>
<td><img src="image7" alt="New invasion occurrences" /></td>
</tr>
<tr>
<td>Rhamnus venosa Host.</td>
<td>Invasibility: 0.5</td>
<td>Species risk: 0.64</td>
<td><img src="image9" alt="New invasion occurrences" /></td>
</tr>
</tbody>
</table>

### Pl@ntHealth: automated plant epidemiology

- Chesnut gall

![Chesnut gall image](image11)
Biodiversity informatics research within CLEF forum

- Pl@ntNet organizes a world-wide challenge since 2011
- Tens of research teams working on Pl@ntNet data
- **System-oriented** benchmarks/competitions
Yearly frontier between **training data (public groundtruth)** vs. **test data (private groundtruth)**

Year of delivery:
- 2017
- 2016
- 2015
- 2014
- 2013
- 2012
- 2011

- 71 sp. 1.5K im.
- 126 sp. 2.2K im.
- 250 sp. 11K im.
- 500 sp. 60K im.
- 1000 sp. 113K im.
- 1000 sp. 121K im.
- 10,000 sp. 1.2M im.
## PlantCLEF

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Espèces</td>
<td>71</td>
<td>126</td>
<td>250</td>
<td>500</td>
<td>1,000</td>
<td>1,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Images</td>
<td>5,400</td>
<td>11,500</td>
<td>26,077</td>
<td>60,962</td>
<td>113,205</td>
<td>121,205</td>
<td>1.2 M</td>
</tr>
<tr>
<td>Nb. of particip.</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>22</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Best perf.</td>
<td>0,209</td>
<td>0,38</td>
<td>0,393</td>
<td>0,456</td>
<td>0,652</td>
<td>0,742</td>
<td>0,92 !</td>
</tr>
</tbody>
</table>

### Chart Description

- **Shape descriptors**
- **Bags of Words & Support Vector Machine**
- **Fisher vectors**
- **Arrival of deep learning**
- **Convolutional neural networks**

The chart illustrates the progression of performance metrics over the years, with notable advancements in 2015 and 2016.
PlantCLEF 2018: Experts vs. Machines plant images identification
- 9 of the best of the best experts of the French flora
- 100 obs. including very difficult taxonomic groups
Is the problem solved? Not really...

World: 369K species

LifeCLEF: 10K species

Pl@ntNet: 17K species
Is the problem solved? Not really...

#images

95% accuracy

50% accuracy

#species
We did query Bing and Google image with 300K species names
- Using ThePlantList: the first effort to list all plants on earth

We collected 12 million images of 294K plant species (1.5 Tb):
- Expert data (Encyclopedia of Life, 350K images) + Citizen science data (Pl@ntNet data, 400K images) + Web data (11 M images)

Highly imbalanced distribution: only 50K species with more than 10 images, 50% with 1 images)
- Noise: depends on the species

“Arnica montana”
Challenges/questions

Scalability to hundreds of thousands of classes
- Which hardware?
  - Memory usage: last layer is 300 times larger than state-of-the-art models
  - To distribute or not to distribute?: communication cost, large batch size
  - CPU vs GPU?
- Which network architecture?
  - Convergence of state-of-the-art models? No guaranty
  - Do we need a new dedicated architecture?
  - Acceptable training time?
- Quality of the learned models?
  - Top-1, top-5, top-30 accuracy? On average? In the long tail?
  - Robustness to noise in the training data?
Evaluation methodology: test set

- **30K never published images** of expert botanists
  - Stored on their local disks or on slides
  - Complex groups in the long tail of the distribution
    - 342 Orchids species
    - 1K Guyana species
    - 469 Alpine species
    - 75 Grass species

- **PlantCLEF 2017 test set (25K Pl@ntNet images)**
  - 1K species living in America and Europe (including common ones)
  - Never published labels
GENCI proposed us to be beta-tester of prototype platforms

- **Ouessant: GPU cluster** hosted by IDRIS (IBM OpenPOWER platform)
  - **12 nodes** IBM Power Systems x 4 **GPU** Nvidia P100 + Infiniband
  - IBM *powerAI* framework v4:
    - Caffe-DLL & TensorFlow-DLL
    - Stochastic gradient

- **Irene: CPU cluster** hosted by CEA (Intel skylakes platform)
  - **1600 nodes** x 48 Intel Skylakes
  - Intel-CAFFE library
Ouessant/GPU experiments (1/3)

By Hervé Goëau, data scientist Pl@ntNet (CIRAD / Inria)

- Encountered difficulties: feedback from a data scientist without experience in HPC or distributed deep learning
  - File systems / inodes issues: quota exceeded notifications, file creation errors, etc.
  - No internet access: no wget, no curl to download pre-trained models, tests, etc.
  - Lack of documentation
  - Limitation of the installed frameworks: old versions, no data augmentation, no shuffling, etc.
  - Jobs limitation (20h00 & 4 nodes)
  - Within the allocated time: No efficiency gain observed in multi-nodes

- Succeeded in training a model at the scale of the world’s flora using transfer learning
  - Inception v2 model pre-trained on ImageNet and fine-tuned on 294K species during about 7 epochs
  - About 60h of training on 1 node with 4 P100 GPUs
Ouessant/GPU experiments (2/3)

By Hervé Goëau, data scientist Pl@ntNet (CIRAD / Inria)

- The model works! state-of-the-art performance on PlantCLEF 2017 dataset (without using ensembles)

Our world’s flora model (with different testing configurations: data augmentation, post-filtering, duplicates removal, multi-image)
Performance in the long tail is low but fair with regard to 294K classes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Top1 accuracy (single image)</th>
<th>Top1 accuracy + multi-image</th>
<th>Top5 accuracy + multi-image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orchids</td>
<td>0.04</td>
<td>0.12</td>
<td>0.22</td>
</tr>
<tr>
<td>Alpine</td>
<td>0.19</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>Guyana</td>
<td>0.07</td>
<td>0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>Grasses</td>
<td>0.37</td>
<td>0.57</td>
<td>0.71</td>
</tr>
<tr>
<td>Random</td>
<td>0.000003</td>
<td>0.000003</td>
<td>0.000015</td>
</tr>
</tbody>
</table>
Irene/CPU experiments (1/3)

- Team
  - Valeriu Codreanu & Damian Podareanu (Research engineers at SURFsara, state-of-the-art results on 1K Intel Skylake)
  - Jean-Christophe Lombardo (Research engineer at Inria - Pl@ntNet)
  - Gabriel Hautreux (HPC engineer, CINES/GENCI)
  - Vikram A Saletore (Principal Engineer for Artificial Intelligence Products at Intel)

- Preparatory phase on Occigen & Frioul CPU cluster from CINES
  - Occigen: 3306 nodes x 2 Intel processors (12-14 cores)
  - Frioul: 48 nodes x Intel KNL processor (68 cores)
Irene/CPU experiments (2/3)

Encountered difficulties
- Intel-CAFFE (MLSL library) requires a password less ssh connexion for initialization (only possible to run in interactive mode)
- Protobuf library is limited to 2Gb files: impossible to serialize ResNet-50 model with 275K classes → dimensionality reduction trick

ResNet-50
- Last layer size: 2.3GB

ResNet-50
- Last layer size: 1.8GB
<table>
<thead>
<tr>
<th>Experiment</th>
<th>Top1 accuracy (all world flora test sets)</th>
<th>Top5 accuracy (all world flora test sets)</th>
<th>Training time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ouessant: 1 node - 4 x P100 Inception v2 fine-tuned 10 epochs</td>
<td>0.356</td>
<td>0.454</td>
<td>60 hours 6 hours/epoch</td>
</tr>
<tr>
<td>Irene: 512 skylake nodes ResNet-50 from scratch 50 epochs</td>
<td>0.375</td>
<td>0.463</td>
<td>10 hours 12 minutes/epoch</td>
</tr>
<tr>
<td>Irene: 1320 skylake nodes ResNet-50 from scratch 82 epochs</td>
<td>0.362</td>
<td>0.451</td>
<td>9 hours 9 minutes/epoch</td>
</tr>
</tbody>
</table>
Conclusions

- Data deficiency in the long tail remains the core problem: 50% of species illustrated by only 1 image on the web
- State-of-the-art CNNs scale to 300K classes (without much modifications)
- Synchronous SGD on hundreds of nodes provides high scaling efficiency but this requires significant know-how

Perspectives

- Integrate The Big One in Pl@ntNet platform
- Sustain the platform to continue the aggregation of data and knowledge about plants
Thank you