Pl@ntNet-300K: a new plant image dataset for the evaluation of set-valued classification (supplementary material)

Camille Garcin[*1], Alexis Joly[†2], Pierre Bonnet[‡3], Antoine Affouard[§2,3], Jean-Christophe Lombardo[¶2,3], Maximilien Servajean[‖4], and Joseph Salmon[**5]

[1]IMAG, Univ Montpellier, Inria, CNRS, Montpellier, France
[2]Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France
[3]CIRAD, AMAP
[4]LIRMM, AMIS, UPVM, Univ Montpellier, CNRS, Montpellier
[5]IMAG, Univ Montpellier, CNRS, Montpellier, France

1 URL to download the dataset

https://doi.org/10.5281/zenodo.4726653

2 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
  Pl@ntNet-300k dataset was created to evaluate set-valued classification, in particular for plant identification. Unlike previous datasets, Pl@ntNet-300k is designed so as to preserve the high level of ambiguity across classes of the initial real-world dataset (Pl@ntNet) as well as its long tail distribution.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  The dataset was created by Pl@ntNet team, Pl@ntNet being a consortium composed of four French research organisms (Inria, INRAE, CIRAD and IRD).
- Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
  The creation of the dataset was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 863463 (Cos4Cloud project) and by the French national research agency under the grant agreement ANR-20-CHIA-0001-01 (CaMeLOt project). Pl@ntNet has also received the support of Agropolis Fondation for the platform creation.

[*]camille.garcin@inria.fr
[†]alexis.joly@inria.fr
[‡]pierre.bonnet@cirad.fr
[§]antoine.affouard@cirad.fr
[¶]jean-christophe.lombardo@inria.fr
[‖]servajean@lirmm.fr
[**]joseph.salmon@umontpellier.fr

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

3 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
  The dataset is composed of pictures of plants. We are in the multi-class classification setting: there is a single plant species per image.
- How many instances are there in total (of each type, if appropriate)?
  There are 306,146 plant images : 243,916 in the training set, 31,118 in the validation set and 31,112 in the test set.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
  The dataset is sampled from a larger set such that two particular features are preserved. These features are inherent to the way the images are acquired and to the intrinsic diversity of plants morphology: i) The dataset exhibits a strong class imbalance, meaning that a few species represent most of the images. ii) Many species are visually similar, making identification difficult even for the expert eye. More details about these properties are available in Section 3.
- What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
  Each instance is an image of a single plant.
- Is there a label or target associated with each instance? If so, please provide a description.
  Each instance is associated to its species.
- Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
  There is no missing information.
- Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
  There is no particular relationships between our instances.
- Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
  The dataset already provides a train/validation/test. For more detail see section 3.1.
- Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
  Only Pl@ntNet observations with a valid species name were included the dataset. The species name validation is based on a Bayesian inference taking as input the names proposed by Pl@ntNet users with a principle of adaptive weights depending on the user's expertise.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
  The dataset is self contained.
- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
  No protected data are available in the paper.

- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
  No, the dataset only contains plant pictures.
- Does the dataset relate to people? If not, you may skip the remaining questions in this section.
  No.
- Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
  Irrelevant.
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
  Irrelevant.
- Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
  Irrelevant.
- Any other comments?

4 Collection Process

- How was the data associated with each instance acquired?

  Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? **Each image comes from the picture of a plant taken by a user of the Pl@ntNet application**

  If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
  Only Pl@ntNet observations with a valid species name were included the dataset. The species name validation is based on a Bayesian inference taking as input the names proposed by Pl@ntNet users with a principle of adaptive weights depending on the user's expertise.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
  The data was collected through Pl@ntNet mobile application and curated through crowdsourcing (by Pl@ntNet users) in addition to the automated filtering (CNN-based) of unappropriated or irrelevant content (faces, humans, animals, buildings, etc.).

  How were these mechanisms or procedures validated?
  The mechanisms were validated by Pl@ntNet curators (expert botanists) and by the scientific and technical committee of Pl@ntNet.
- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? **The sampling is done at the genus level : 10% of the genus are randomly sampled, and all images that belong to these genera are kept. As a last step, we only retained species with at least 4 images.**
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? **The data is collected by users of the Pl@ntNet application (which has more than 10 millions users). Pl@ntNet users are citizen scientist who gracefully participate to the project.**

Their reward is the acclaimed performance of the application which enables them to identify plant species.

- Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. **The dataset was created with images collected by the Plantnet application from 2011 up to November 2020.**
- Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. **An ethical review was processed by CIRAD's institutional review board. The main outcome was the terms of use of Pl@ntNet application (https://api.plantnet.org/views/terms_of_use?lang=en)**
- Does the dataset relate to people? **No**. If not, you may skip the remainder of the questions in this section.
- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? **not applicable**
- Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. **not applicable**
- Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. **not applicable**
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). **not applicable**
- Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. **not applicable**

5 Preprocessing/cleaning/labeling

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
  No pre-preprocessing was applied (appart from the curation process, see previous section).
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
  not applicable
- Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
  No.

6 Uses

- Has the dataset been used for any tasks already? **Not this specific Pl@ntNet subset.** If so, please provide a description.
- Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. **The list all or some papers that use our**

dataset will be displayed and updated at the following address: https://github.com/plantnet/PlantNet-300K/

- What (other) tasks could the dataset be used for?
  The dataset can be used for any supervised or unsupervised classification tasks.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms? **No**
- Are there tasks for which the dataset should not be used? If so, please provide a description.
  No

7 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. **the dataset will be publicly available**
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? **the dataset will be distributed through zenodo under doi: https://doi.org/10.5281/zenodo.4726653**
- When will the dataset be distributed? **the dataset will be distributed after acceptance of the paper**
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
  The dataset and all images composing it will be distributed under Creative-Common Attribution-ShareAlike 2.0 license.
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
  No.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
  No.

8 Maintenance

- Who is supporting/hosting/maintaining the dataset?
  The Pl@ntnet team will maintain the dataset and provide support. The dataset is hosted by http://zenodo.org.
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  The owner/manager of the dataset can be contacted by mail at plantnet-300k@inria.fr.
- Is there an erratum? If so, please provide a link or other access point.
  Zenodo will provide a versioning of any correction of the dataset. We will keep the users informed at https://github.com/plantnet/PlantNet-300K/
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
  The dataset will be updated if errors are spotted. The update will be performed by the

Plantnet team, and these modifications will be listed at https://github.com/plantnet/PlantNet-300K/.

- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
  Irrelevant.
- Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
  Zenodo will provide a versioning of any correction of the dataset.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
  No.
- Any other comments?

9 Author statement

The authors confirm that all data in Pl@ntNet-300K dataset are under a Creative-Common Attribution-ShareAlike 2.0 license (see terms of use here) and bear responsibility in case of violation of copyrights.