Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution

Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, Joseph Salmon

Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution

Camille Garcin
IMAG, Univ Montpellier,
CNRS, Montpellier, France

Alexis Joly
Inria, LIRMM, Univ Montpellier,
CNRS, Montpellier, France

Pierre Bonnet
CIRAD, AMAP

Jean-Christophe Lombardo
Inria, LIRMM, Univ Montpellier,
CNRS, Montpellier, France

Antoine Affouard
Inria, LIRMM, Univ Montpellier,
CNRS, Montpellier, France

Mathias Chouet
Inria, LIRMM, Univ Montpellier,
CNRS, Montpellier, France

Maximilien Servajean
LIRMM, AMIS, UPVM,
Univ Montpellier, CNRS, Montpellier

Titouan Lorieul
Inria, LIRMM, Univ Montpellier,
CNRS, Montpellier, France

Joseph Salmon
IMAG, Univ Montpellier,
CNRS, Montpellier, France
Institut Universitaire de France (IUF)

Abstract

This paper presents a novel image dataset with high intrinsic ambiguity and a long-tailed distribution built from the database of Pl@ntNet citizen observatory. It consists of 306,146 plant images covering 1,081 species. We highlight two particular features of the dataset, inherent to the way the images are acquired and to the intrinsic diversity of plants morphology: (i) the dataset has a strong *class imbalance*, *i.e.,* a few species account for most of the images, and, (ii) many species are *visually similar*, rendering identification difficult even for the expert eye. These two characteristics make the present dataset well suited for the evaluation of set-valued classification methods and algorithms. Therefore, we recommend two set-valued evaluation metrics associated with the dataset (*macro-average top-k accuracy* and *macro-average average-k accuracy*) and we provide baseline results established by training deep neural networks using the cross-entropy loss.

1 Introduction

When classifying images, we are faced with two main types of uncertainties [Der Kiureghian and Ditlevsen, 2009]: (i) the *aleatoric* uncertainty that arises from the intrinsic randomness of the underlying process, which is considered irreducible, and (ii) the *epistemic* uncertainty that is caused by a lack of knowledge and is considered to be reducible with additional training data. In modern real-world applications, these two types of uncertainties are particularly difficult to handle. The ever-growing number of classes to distinguish tends to increase the class overlap (and thus the aleatoric uncertainty), and, on the other hand, the long-tailed distribution makes it difficult to learn the less populated classes (and thus increase the epistemic uncertainty). The presence of these two uncertainties is a central motivation for the use of set-valued classifiers, *i.e.,* classifiers returning a set of

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

candidate classes for each image [Chzhen et al., 2021]. Although there are several datasets in the literature that have visually similar classes [Nilsback and Zisserman, 2008, Maji et al., 2013, Yang et al., 2015, Russakovsky et al., 2015], most of them do not aim to retain both the epistemic and the aleatoric ambiguity present in real-world data.

In this paper, we propose a dataset designed to remain representative of real-life ambiguity, making it well suited for the evaluation of set-valued classification methods. This dataset is extracted from real-world images collected as part of the Pl@ntNet project [Affouard et al., 2017], a large-scale citizen observatory dedicated to the collection of plant occurrences data through image-based plant identification. The key feature of Pl@ntNet is a mobile application that allows citizens to send a picture of a plant they encounter and get a list of the most likely species for that photo in return. The application is used by more than 10 million users in about 170 countries and is one of the main data publishers of GBIF[1], an international platform funded by the governments of many countries around the world to provide free and open access to biodiversity data. Another essential feature of Pl@ntNet is that the training set used to train the classifier is collaboratively enriched and revised. Nowadays, Pl@ntNet covers over 35K species illustrated by nearly 12 million validated images.

The entire Pl@ntNet database would be an ideal candidate for the evaluation of set-valued classification methods. However, it is far too large to allow for widespread use by the machine learning community. Extracting a subsample from it must be done with care as we want to preserve the uncertainty naturally present in the whole database. The dataset presented in this paper is constructed by retaining only a subset of the genera of the entire Pl@ntNet database (sampled uniformly at random). All species belonging to the selected genera are then retained. Doing so maintains the original ambiguity as species in the same genus are likely to be visually similar and to share common visual features.

The rest of the paper is organized as follows. We first introduce the set-valued classification framework in Section 2, focusing on two special cases: top-k classification and average-k classification. In Section 3, we describe the construction procedure of the Pl@ntNet-300K dataset and show that it contains a large amount of ambiguity. Next, we present in Section 4 the metrics of interest for the dataset and propose benchmark results for these metrics, obtained by training several neural networks architectures. In Section 5, we compare Pl@ntNet-300K to several existing datasets. In Section 6, we discuss possible uses of Pl@ntNet-300K. Finally, we provide the link to the dataset in Section 7 before concluding.

2 Set-valued classification

We adopt the classical statistical setup of multi-class classification. Let L be the number of classes. We denote by $[L]$ the set $\{1, \ldots, L\}$ and by $\mathcal{X}$ the input space. Random couples of images and labels $(X, Y) \in \mathcal{X} \times [L]$ are assumed to be generated *i.i.d.* by an unknown joint distribution $\mathbb{P}$. Note that only one label is associated with each image, which differs from the multi-label setting [Zhang and Zhou, 2014]: P@ntNet-300K is composed of images containing a single specimen of a plant, so there is only one true label per image. For some plant images, predicting the correct class label (the correct species) does not present much difficulty (consider for instance a common species very distinctive from other species). For other images, however, classifying the photographed specimen with a high degree confidence is a much harder task, because some species differ only in subtle visual features (see Figure 4). In these cases, it is desirable to provide the user with a list of likely species corresponding to the image. We thus need a classifier able to produce sets of classes, also known as a set-valued classifier in the literature [Chzhen et al., 2021]. A set-valued classifier Γ is a function mapping the feature space $\mathcal{X}$ to the set of all subsets of $[L]$ (which we denote by $2^{[L]}$). Using these notations, we thus have $\Gamma : \mathcal{X} \to 2^{[L]}$ instead of $\Gamma : \mathcal{X} \to [L]$ for the classical setting in which the predictor can only predict a single class. Our goal is to build a classifier with low risk, defined as $\mathbb{P}(Y \notin \Gamma(X))$. However, it is not desirable to simply minimize the risk: a set-valued classifier that always returns all classes achieves zero risk but is useless. On the other hand, a classifier is most useful if it returns only the most likely classes given a query image. Therefore, a quantity of interest will be $|\Gamma(x)|$, the number of classes returned by the classifier Γ, given an image $x \in \mathcal{X}$.

[1] https://www.gbif.org/

In this section we will examine two optimization problems that lead to different set-valued classifiers. Both of them aim to minimize the risk, but they differ in the way they constrain the set cardinality: either pointwise or on average.

For $x \in \mathcal{X}$ and $l \in [L]$, we define the conditional probability $p_l(x) := \mathbb{P}(Y = l \mid X = x)$, and estimators of these quantities will be denoted by $\hat{p}_l(x)$. In the following, $k \in [L]$. Finally, for $x \in \mathcal{X}$, we define $\mathrm{top}_{\mathrm{p}}(x, k)$ as the set containing the k indexes corresponding to the k largest values of $\{p_l(x)\}_{l \in [L]}$.

The simplest constraint is to require that the number of returned classes is less than k for each input. This results in the following top-k error [Lapin et al., 2015] minimization problem:

$$\begin{aligned} \Gamma^*_{\text{top-k}} \in \arg\min_{\Gamma} \ & \mathbb{P}(Y \notin \Gamma(X)) \\ \text{s.t.} \ & |\Gamma(x)| \leq k, \ \forall x \in \mathcal{X} \ . \end{aligned} \tag{1}$$

The closed form solution to (1) exists and is equal to [Lapin et al., 2017]:

$$\Gamma^*_{\text{top-k}}(x) = \mathrm{top}_{\mathrm{p}}(x, k) \ . \tag{2}$$

Yet, this is not practical since we do not know the distribution $\mathbb{P}$ and thus $p_l(x)$. However, if we have an estimator $\hat{p}_l(x)$ of $p_l(x)$, we can naturally derive the plug-in estimator: $\widehat{\Gamma}_{\text{top-k}} = \mathrm{top}_{\hat{\mathrm{p}}}(x, k)$. While the *top-$k$ accuracy* is often reported in benchmarks, only a few works aim to directly optimize this metric [Lapin et al., 2015, 2016, 2017, Berrada et al., 2018]. An obvious limitation of top-k classification is that k classes are returned for every data sample, regardless of the difficulty of classifying that sample. Average-k classification allows for more adaptivity. In that setting, the constraint on the size of the predicted set is less restrictive and must be satisfied only on average, leading to:

$$\begin{aligned} \Gamma^*_{\text{average-k}} \in \arg\min_{\Gamma} \ & \mathbb{P}(Y \notin \Gamma(X)) \\ \text{s.t.} \ & \mathbb{E}_X |\Gamma(X)| \leq k \ . \end{aligned} \tag{3}$$

The closed form solution is derived in [Denis and Hebiri, 2017]:

$$\Gamma^*_{\text{average-k}}(x) = \{l \in [L] \, : \, p_l(x) \geq G^{-1}(k)\} \ , \tag{4}$$

where the function G is defined as: $\forall t \in [0, 1], \quad G(t) = \sum_{l=1}^{L} \mathbb{P}(p_l(X) \geq t)$, and G^{-1} refers to the generalized inverse of G, namely $G^{-1}(u) = \inf\{t \in [0, 1] \, : \, G(t) \leq u\}$.

Note that if we define the classifier Γ_t by: $\forall x \in \mathcal{X}$, $\Gamma_t(x) = \{l \in [L], p_l(x) \geq t\}$, then $G(t)$ is the average number of classes returned by Γ_t: $G(t) = \mathbb{E}_X |\Gamma_t(X)|$. From (4) we see that the optimal classifier corresponds to a thresholding operation: all classes having a conditional probability greater than $G^{-1}(k)$ are returned, with the threshold chosen so that k classes are returned on average. To compute a plug-in counterpart, we just need to estimate the threshold on a calibration set such that on average on that set, k classes are returned. For technical details, we refer the reader to Denis and Hebiri [2017].

3 Dataset

3.1 Label validation and data cleaning

Label validation is based on a weighted majority voting algorithm taking as input the labels proposed by Pl@ntNet users with an adaptive weighting principle according to the user's expertise and commitment. Thus, a single trusted annotator can be enough to validate an image label. On the other hand, images whose labels are proposed by several novice users may not be validated because they do not have sufficient weight. The technical details of this algorithm can be found in the supplementary material. At the time of the construction of Pl@ntNet-300K, the total number of annotators in the Pl@ntNet database was equal to 2,079,003. The average number of annotators per image is equal to 2.03.

In addition to the label validation procedure, Pl@ntNet's pipeline includes other data cleaning procedures: (i) automated filtering of inappropriate or irrelevant content (faces, humans, animals, buildings, etc.) using a CNN and user reports, and (ii) filtering on image quality (evaluated by users).

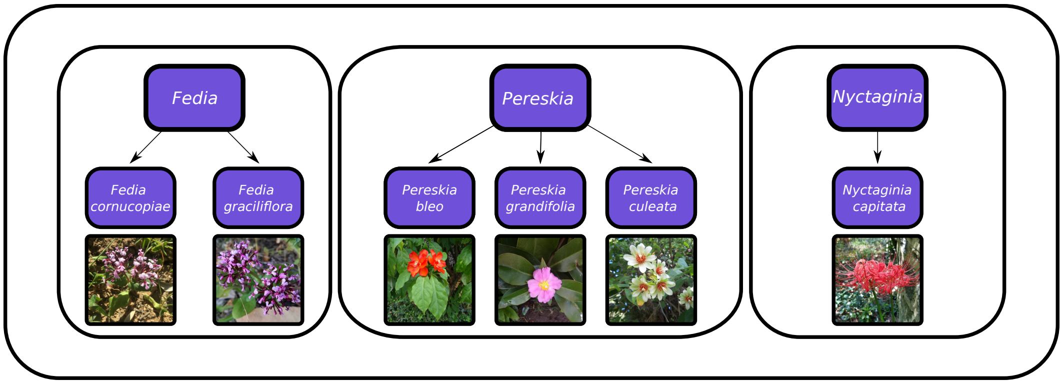


Figure 1: Genus taxonomy: we display three genera present in the proposed dataset—*Fedia*, *Pereskia* and *Nyctaginia*—which contain respectively two, three and one species.

3.2 Construction of Pl@ntNet-300K

In taxonomy, species are organized into genera, with each genus containing one or more species, and the different genera do not overlap, as illustrated in Figure 1.

Instead of retaining randomly selected species or images from the entire Pl@ntNet dataset, we choose to retain randomly selected genera and keep all species belonging to these genera. This choice aims to preserve the large amount of ambiguity present in the original database, as species belonging to the same genus tend to share visual features. The dataset presented in this paper is constructed by sampling uniformly at random only 10% of the genera of the whole Pl@ntNet database.

We then retain only species with more than 4 images, resulting in a total of 303 genera and $L = 1{,}081$ species. The images are divided into a training set, a validation set and a test set.[2] For each species, 80% of the images are placed in the training set ($n_{train} = 243{,}916$), 10% in the validation set ($n_{val} = 31{,}118$), and 10% in the test set ($n_{test} = 31{,}112$), with at least one image of each species in each set. More formally, given a class j containing n_j images, $n_{val,j} = \lceil 0.1 \times n_j \rceil$, $n_{test,j} = \lceil 0.1 \times n_j \rceil$ and $n_{train,j} = n_j - n_{val,j} - n_{test,j}$. This represents a total of $n_{tot} = n_{train} + n_{val} + n_{test} = 306{,}146$ color images. The average image size is (570, 570, 3), ranging from (180, 180, 3) to (900, 900, 3). The construction of the dataset preserves the class imbalance. To show this, we plot the Lorenz curves [Gastwirth, 1971] of the entire Pl@ntNet dataset and that of the Pl@ntNet-300K dataset in Figure 2.

3.3 Epistemic (model) uncertainty

Epistemic uncertainty refers mainly to the lack of data necessary to properly estimate the conditional probabilities. In Pl@ntNet, the most common species are easily observed by users in the wild and therefore represent a large fraction of the images, while the rarest species are harder to find and therefore less frequent in the database. In Figure 2, we see that 80% of the species (the ones with the lowest number of images) account for only 11% of the total number of images. Hence, training machine learning models is challenging for such a dataset, since for many classes the model only has a handful of images to train on, making identification difficult for these species.

In addition to the long-tailed distribution issue, epistemic uncertainty also arises from the high intra-species variability. Plants may take on different appearances depending on the season (*e.g.,* , flowering time). Furthermore, a user of the application may photograph only a part of the plant (for instance, the trunk and not the leaves). As a last example, flowers belonging to the same species can have different colors. Figure 3 shows some examples of these phenomena which contribute to high intra-class variability, making it more challenging to model the species.

[2]The division is performed at the species level due to the long-tailed distribution.

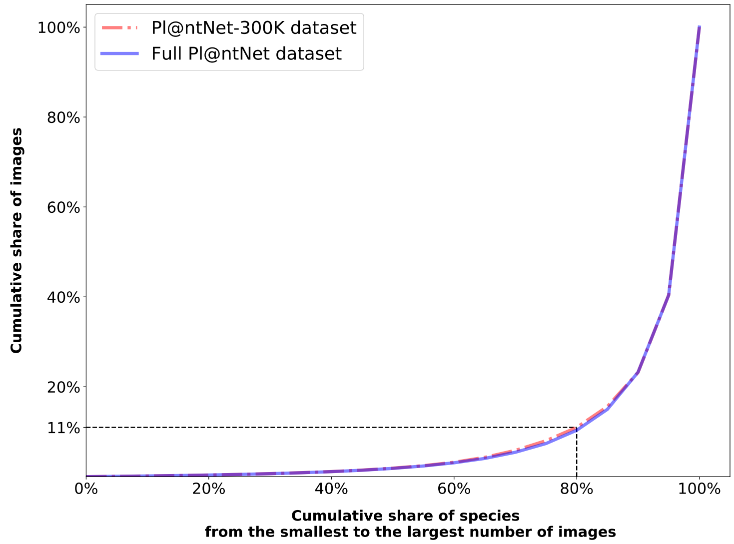


Figure 2: Lorenz curves of the Pl@ntNet database and the proposed dataset. Note that, for fair comparison, we discard species with less than 4 images in the Pl@ntNet database.

Guizotia abyssinica | *Diascia rigescens* | *Lapageria rosea* | *Casuarina cunninghamiana* | *Freesia alba*

Figure 3: Examples of visually different images belonging to the same class.

3.4 Aleatoric (data) uncertainty

In our case, the source of aleatoric uncertainty mostly resides in the limited information we are given to make a decision (assign a label to a plant). Some species, especially those belonging to the same genus, can be visually very similar. For example, consider the case where two species produce the same flowers but different leaves, typically because they have evolved differently from the same parent species. If a person photographs only the flower of a specimen of one of the two species, then it will be impossible, even for an expert, to know which species the flower belongs to. The discriminative information is not present in the image.

The combination of this irreducible ambiguity with images of non-optimal quality (non-adapted close-up, low-light conditions, etc.) results in pairs of images that belong to different species but are difficult or even impossible to distinguish, see Figure 4 for illustration. In this figure, we show the ambiguity between pairs of species, but we could find similar examples involving a larger number of species. Thus, even an expert botanist might fail to assign a label to such pictures with certainty. This is embodied by $p_l(x)$: given an image, multiple classes are possible.

Figure 4: Examples of visually similar images belonging to two different classes.

4 Evaluation

4.1 Metric

We consider two main metrics to evaluate set valued predictors on Pl@ntNet-300K: *top-k accuracy* and *average-k accuracy*. Let S denote a set of n (input, label) pairs: $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. *Top-k accuracy* [Lapin et al., 2016] is a widely used metric often reported in benchmarks. *Average-k accuracy* is a much less common metric that derives from average-k classification [Denis and Hebiri, 2017]:

$$\textit{average-k accuracy}(S) = \frac{1}{n} \sum_{(x_i, y_i) \in S} \mathbb{1}_{[y_i \in \widehat{\Gamma}_{\text{average-k}}(x_i)]} \text{ s.t. } \frac{1}{n} \sum_{(x_i, y_i) \in S} |\widehat{\Gamma}_{\text{average-k}}(x_i)| \leq k \ , \quad (5)$$

where $\widehat{\Gamma}_{\text{average-k}}$ is a set-valued classifier constructed using the training data.

For Pl@ntNet-300K, both *top-k accuracy* and *average-k accuracy* mainly reflect the performance of the set-valued classifier on the few classes which represent most of the images. If we wish to capture the ability of a set-valued classifier to return pertinent set of species for all classes, we will examine *macro-average top-k accuracy* and *macro-average average-k accuracy* which simply consist in computing respectively *top-k accuracy* and *average-k accuracy* for each class, and then computing the average over classes. For *macro-average average-k accuracy*, the constraint on the average size of the set must hold for the entire set S.

To derive both classifiers, one can first obtain an estimate of the conditional probabilities $\hat{p}_l(x)$ and then derive the plug-in classifiers, as explained in Section 2. Our hope is for the Pl@ntNet-300K dataset to encourage novel ways to derive the set-valued classifiers $\widehat{\Gamma}_{\text{top-k}}$ and $\widehat{\Gamma}_{\text{average-k}}$ to optimize respectively the *top-k accuracy* and the *average-k accuracy*. Notice that a few works already propose methods to optimize the *top-k accuracy* [Lapin et al., 2015, 2016, 2017, Berrada et al., 2018].

4.2 Baseline

This section provides a baseline evaluation of the plug-in classifiers. We train several deep neural networks with the cross-entropy loss: ResNets [He et al., 2016], DenseNets, [Huang et al., 2017], InceptionResNet-v2 [Szegedy et al., 2017], MobileNetV2 [Sandler et al., 2018], MobileNetV3 [Howard et al., 2019], EfficientNets [Tan and Le, 2019], Wide ResNets [Zagoruyko and Komodakis, 2016], AlexNet [Krizhevsky et al., 2012], Inception-v3 [Szegedy et al., 2016], Inception-v4 [Szegedy et al., 2017], ShuffleNet [Zhang et al., 2018], SqueezeNet [Iandola et al., 2016], VGG [Simonyan and Zisserman, 2015] and Vision Transformer [Dosovitskiy et al., 2021].

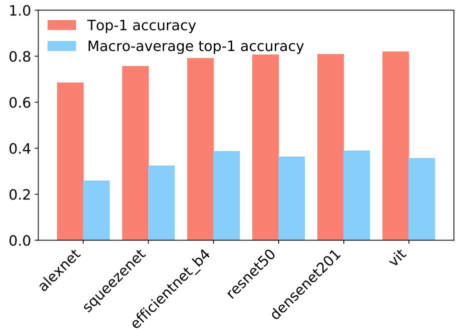


Figure 5: Pl@ntNet-300K test *top*-1 *accuracy* and *macro-average top*-1 *accuracy* for several neural networks.

Number of images	Mean bin accuracy
$0-10$	0.09
$10-50$	0.35
$50-500$	0.59
$500-2000$	0.79
>2000	0.93

Table 1: Test accuracy depending on the number of images per class in the training set. Obtained with a ResNet50.

All models are pre-trained on ImageNet. During training, images are resized to 256 and a random crop of size 224×224 is extracted. During test time, we take the centered crop.

The models are optimized with SGD with a momentum of 0.9 with the Nesterov acceleration [Ruder, 2016]. We use a batch size of 32 for all models and a weight decay of 1.10^{-4}. The number of epochs, initial learning rate and learning rate schedule used for each model can be found in the supplementary material. For the plug-in classifier $\widehat{\Gamma}_{\text{average-k, plug-in}}$, we compute the threshold λ_{val} on the validation set and use that same threshold to compute the *average-k accuracy* on the test set.

4.3 Difficulty of Pl@ntNet-300K

Figure 5 highlights the significant gap between Pl@ntNet-300K *top*-1 *accuracy* and *macro-average top*-1 *accuracy*. This is a consequence of the long-tailed distribution: the few classes that represent most of the images are easily identified, which results in high *top*-1 *accuracy*. However, this seemingly high *top*-1 *accuracy* is misleading, as models struggle with classes with few images (which are a majority, see Figure 2). This effect is illustrated in Table 1, which shows that the *top*-1 *accuracy* depends strongly on the number of images in the class.

Figure 8a shows the correlation between Pl@ntNet-300K *macro-average top*-1 *accuracy* and ImageNet *macro-average top*-1 *accuracy* (note that as the ImageNet test set is balanced, *top*-1 *accuracy* and *macro-average top*-1 *accuracy* coincide). As expected, the two metrics are positively correlated: deep networks allowing to model complex features work well both on ImageNet and Pl@ntNet-300K. Interestingly, due the difference between the two datasets (long-tailed distribution, class ambiguity, ...), some models which perform similarly on ImageNet yield very different on Pl@ntNet-300K (inception_v3, densenet201), and vice versa.

In Figure 8a we can notice that ImageNet *macro-average top*-1 *accuracy* and Pl@ntNet-300K *macro-average top*-1 *accuracy* vary at different scales: the former goes up to 80% while the latter does not exceed 40%, making Pl@ntNet-300K a challenging dataset with both epistemic and aleatoric uncertainty at play.

This can can also be seen in Figure 6: some models reach a *macro-average average*-5 *accuracy* of 97% for ImageNet, while that metric does not exceed 80% for Pl@ntNet-300K, which suggests that progress could be made with appropriate learning strategies.

To support that claim, we asked a botanist to label a mini dataset extracted from the Pl@ntNet-300K test set. The dataset is constructed as follows: we extract all species from two groups (*Crotalaria* and *Lupinus*), and select at most 5 images per species (randomly sampled). This results in 83 images. The botanist was asked to provide a set of possible species for each image. This results in an error rate of 20.5% for an average of 4,1 species returned (ranging from 1 to 10). We compare the botanist performance with that of several neural networks by calibrating the conditional probabilities' threshold to obtain on average 4,1 species on the mini-dataset. The results are reported in Figure 7, and show that the gap between the botanist error rate and the best performing model is

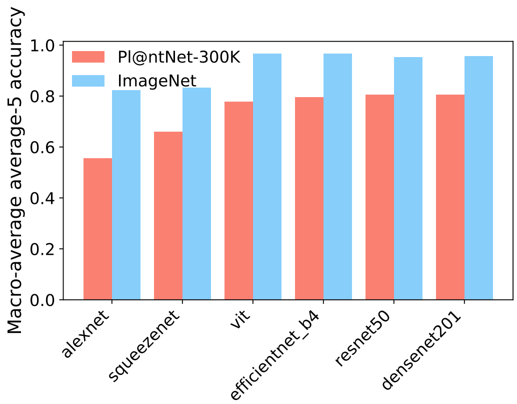


Figure 6: Pl@ntNet-300K vs ImageNet *macro-average average-5 accuracy* for several models (evaluated on the test set).

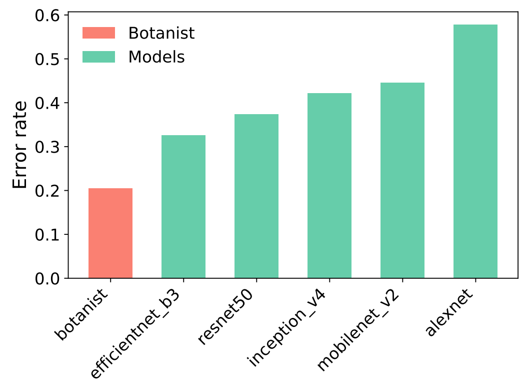


Figure 7: Error rate on the mini test set of an expert compared to several neural networks. All models and the expert return on average 4.1 species on the mini test set.

large (from 0.2 to 0.33), which suggests that there is room for improvement in the performance of average-k classifiers.

4.4 Top-k vs Average-k

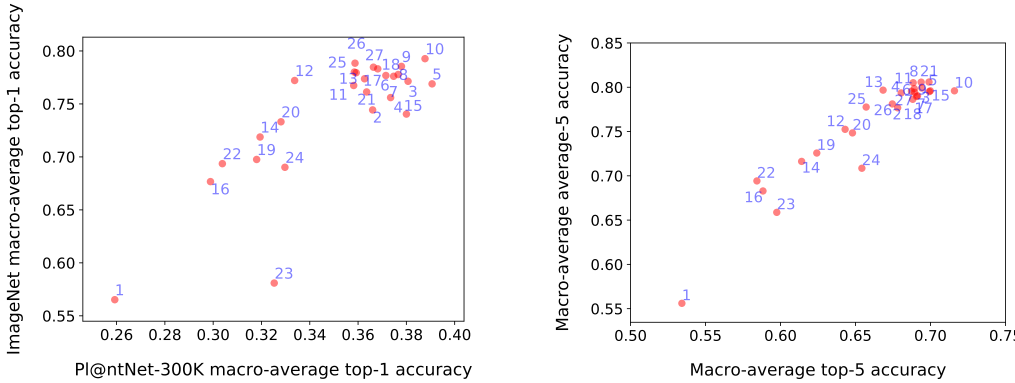
From Equation (1) and (3), it is clear that the Bayes average-k classifier has a lower risk than the Bayes top-k classifier. Therefore, a model that accurately estimates the conditional probabilities should yield a better *average-k accuracy* than *top-k accuracy*. This is what can be observed in Figure 8b which shows the correlation between *macro-average top*-5 *accuracy* and *macro-average average*-5 *accuracy*. As expected, the two metrics are positively correlated. However, the relationship does not appear to be trivial and Figure 8b shows models with similar *macro-average average*-5 *accuracies* having very different *macro-average top*-5 *accuracy* and vice versa. For an in-depth comparison of average-k classification and top-k classification, we refer the reader to Lorieul [2020]. Both metrics are of their own interest and deserve a specific treatment as they capture two different settings: *top-k accuracy* evaluates the performance of a classifier which systematically returns k classes, while *average-k accuracy* evaluates the performance of a classifier which returns sets of varying size (depending on the input), with the constraint to return k classes on average.

4.5 Evaluation of existing set-valued classification methods

To the best of our knowledge, there is no existing loss designed to specifically optimize *average-k accuracy*. For top-k classification, the most recent loss designed to optimize *top-k accuracy* is the one by Berrada et al. [2018]. We report the *top-5 accuracy* obtained by training this loss with $k = 5$ on Pl@ntNet-300K in the supplementary material. The results are close with what is obtained with the cross entropy loss. However, this topic is still open research and our hope in releasing Pl@ntNet-300K is precisely to encourage novel methods for optimizing such metrics.

5 Related work

Fined-Grained Visual Categorization (FGVC) is about discriminating visually similar classes. In order to better learn fine-grained classes, several approaches have been proposed by the FGVC community, including multi-stage metric learning [Qian et al., 2015], high order feature interaction [Lin et al., 2015, Cui et al., 2017], and different network architectures [Fu et al., 2017, Ge et al., 2016]. However, these approaches focus on optimizing *top-1 accuracy*. Set-valued classification, on the other hand, consists in returning more than a single class to reduce the error rate, with a constraint on the number of classes returned. Therefore, FGVC and set-valued classification methods are not mutually exclusive but rather complementary.



(a) ImageNet *macro-average top*-1 *accuracy* vs. Pl@ntNet-300K *macro-average top*-1 *accuracy* (evaluated on the test set).

(b) Pl@ntNet-300K *macro-average average*-5 *accuracy* vs. *macro-average top*-5 *accuracy* (evaluated on the test set).

Figure 8: Benchmark for several popular deep neural network architectures[3].

Several FGVC datasets, which exhibit visually similar classes, have been made publicly available by the community. They cover a variety of domains: aircraft [Maji et al., 2013], cars (Compcars [Yang et al., 2015], Census cars [Gebru et al., 2017]), birds (CUB200 [Welinder et al., 2010]), flowers (Oxford flower dataset [Nilsback and Zisserman, 2008]). However, most of these datasets focus exclusively on proposing visually similar classes (aleatoric uncertainty) with a limited amount of epistemic uncertainty. This is the case for balanced datasets which have approximately the same number of images per class, or with small intra-class variability such as aircraft and cars datasets, where most examples within a class are nearly the same except for angle, lightning, etc. ImageNet [Russakovsky et al., 2015] has several visually similar classes, organized in groups : it contains many bird species and dog breeds. However, these groups of classes are very different: dogs, vehicles, electronic devices, etc. Besides, ImageNet does not exhibit a strong class imbalance. Several of these datasets were constructed by web-scraping, which can be prone to noisy labels and low quality images. Most similar to our dataset is the iNat2017 dataset [Horn et al., 2018]. It contains images from the citizen science website iNaturalist. The images, posted by naturalists, are validated by multiple citizen scientists. The iNat2017 dataset contains over 5000 classes that are highly unbalanced. However, iNat2017 does not only focus on plants but proposes several other 'super-classes' such as *Fungi*, *Reptilia*, *Insecta*, etc. Moreover, the authors selected all classes with a number of observations greater than 20, whereas we choose to randomly sample 10% of the genera of the entire Pl@ntNet database and keep all species belonging to these groups with a number of observations greater than 4. We argue that keeping all species of the same genus maximizes aleatoric uncertainty, because species belonging to the same genus tend to share visual features. Finally, a plant disease dataset is introduced in [Sladojevic et al., 2016], containing 4483 images downloaded from the web spread across 15 classes. This is a very different scale than Pl@ntNet-300K. We summarize the properties of the mentioned datasets in Table 2.

6 Possible uses of Pl@ntNet-300K

Although we are convinced of the need to design new set-valued methods due to the ever increasing amount of classes to discriminate, the properties of Pl@ntNet-300K described in Section 3 make it an ideal candidate for various other tasks. The strong class imbalance can be used by researchers to evaluate new algorithms specifically designed for tackling class imbalance [Zhou et al., 2020, **?**]. Pl@ntNet-300K contains a large amount of aleatoric uncertainty resulting from many visually similar classes. It can therefore be used as a FGVC dataset to evaluate methods that aim to optimize

[3] The architectures chosen are: alexnet (1), densenet121 (2), densenet161 (3), densenet169 (4), densenet201 (5), efficientnet_b1 (6), efficientnet_b1 (7), efficientnet_b2 (8), efficientnet_b3 (9), efficientnet_b4 (10), inception_resnet_v2 (11), inception_v3 (12), inception_v4 (13), mobilenet_v2 (14), mobilenet_v3_large (15), mobilenet_v3_small (16), resnet101 (17), resnet152 (18), resnet18 (19), resnet34 (20), resnet50 (21), shufflenet (22), squeezenet (23), vgg11 (24), vit_base_patch16_224 (25), wide_resnet101_2 (26), wide_resnet50_2 (27).

Table 2: Comparison of several datasets with Pl@ntNet-300K. "Focused domain" indicates whether the dataset is made up of a single category (*i.e.,* cars) and "Ambiguity preserving sampling" indicates whether in the construction of the dataset, all classes belonging to the same parent in the class hierarchy were kept or not (in our case, the parent corresponds to the genus level).

	Human-in-the-loop labeling	Long-tailed distribution	Intra-class variability	Focused domain	Ambiguity preserving sampling
Plant disease dataset	✗	✗	✗	✓	✗
CUB200	✗	✗	✗	✓	✗
Oxford flower dataset	✗	✗	✓	✓	✗
Aircraft dataset	✓	✗	✗	✓	✗
Compcars	✗	✗	✗	✓	✓
Census cars	✗	✗	✗	✓	✓
ImageNet	✗	✗	✓	✗	✗
iNat2017	✓	✓	✓	✗	✗
Pl@ntNet-300K	✓	✓	✓	✓	✓

top-1 *accuracy* for such datasets, see for instance [Lin et al., 2015, Fu et al., 2017, Cui et al., 2017]. Finally, in this paper we do not use the genus information and thus do not exploit the hierarchical structure of the problem. In this sense, we adopt the flat classification approach described in [Silla and Freitas, 2011]. This is consistent with ImageNet [Russakovsky et al., 2015] or CIFAR-100 [Krizhevsky, 2009], where a hierarchy does exist but is rarely used in benchmarks. However, researchers are free to use the genus information to evaluate hierarchical classification methods on Pl@ntNet-300K.

7 Data access and additional resources

The Pl@ntNet-300K dataset can be found here:

`https://doi.org/10.5281/zenodo.5645731`.

It is organized in three folders named "`train`", "`val`" and "`test`". Each of these folders contains $L = 1{,}081$ subfolders. We provide the correspondence between the names of the subfolders and the names of the classes in the file "`plantnet300K_species_id_2_name.json`". We also provide a metadata file named "`plantnet300K_metadata.json`" containing for each image the following information: the species identifier (class), the organ of the plant (flower, leaf, bark, ...), the author's name, the license and the split (*i.e.,* train, validation or test set). A github repository containing the code to reproduce the experiments of this paper (where potential issues related to the dataset can be reported too) is available at: `https://github.com/plantnet/PlantNet-300K/`.

8 Conclusion

In this paper, we share and discuss a novel plant image dataset, called Pl@ntNet-300K, obtained as a subset of the entire Pl@ntNet database and intended primarily for evaluating set-valued classification methods. Unlike previous datasets, Pl@ntNet-300K is designed to preserve the high level of ambiguity across classes of the initial real-world dataset as well as its long-tailed distribution. To evaluate set-valued predictors on Pl@ntNet-300K, we investigate two different metrics: *macro-average top-k accuracy* and *macro-average average-k accuracy*, which is a more challenging task requiring to predict sets of various size but still equal to k on average. Our results suggest that there is room for new set-valued prediction methods that would improve the performance of average-k classifiers. We hope that Pl@ntNet-300K can serve as a reference dataset for this problem, which is our main motivation for releasing and sharing it with the community. We also stress that Pl@ntNet-300K can also be used to evaluate new methods for long-tailed classification and FGVC.

Acknowledgments

This work was partially funded by the ANR CaMeLOt ANR-20-CHIA-0001-01. It has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 863463 (Cos4Cloud project).

References

Antoine Affouard, Hervé Goëau, Pierre Bonnet, Jean-Christophe Lombardo, and Alexis Joly. Pl@ntnet app in the era of deep learning. In *ICLR - Workshop Track*, 2017.

Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Smooth loss functions for deep top-k classification. In *ICLR*, 2018.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Titouan Lorieul. Set-valued classification – overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge J. Belongie. Kernel pooling for convolutional neural networks. In *CVPR*, pages 3049–3058, 2017.

Christophe Denis and Mohamed Hebiri. Confidence sets with expected sizes for multiclass classification. *J. Mach. Learn. Res.*, 18:102:1–102:28, 2017.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 4476–4484, 2017.

Joseph L Gastwirth. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039, 1971.

ZongYuan Ge, Alex Bewley, Christopher McCool, Peter I. Corke, Ben Upcroft, and Conrad Sanderson. Fine-grained classification via mixture of deep convolutional neural networks. In *CVPR*, pages 1–6, 2016.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, volume 31, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.

Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. In *ICCV*, pages 1314–1324, 2019.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017.

Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5MB model size. *CoRR*, abs/1602.07360, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1106–1114, 2012.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass SVM. In *NeurIPS*, pages 325–333, 2015.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Loss functions for top-k error: Analysis and insights. In *CVPR*, pages 1468–1477, 2016.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(7): 1533–1554, 2017.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, pages 1449–1457. IEEE Computer Society, 2015.

Titouan Lorieul. *Uncertainty in predictions of deep learning models for fine-grained classification*. PhD thesis, Université de Montpellier, December 2020.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *CVPR*, pages 3716–3724, 2015.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

Carlos Silla and Alex Alves Freitas. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.*, 22(1-2):31–72, 2011.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, Dubravko Culibrk, and Darko Stefanovic. Deep neural networks based recognition of plant diseases by leaf image classification. *Comput. Intell. Neurosci.*, 2016:3289801:1–3289801:11, 2016.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284. AAAI Press, 2017.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, volume 97, pages 6105–6114, 2019.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9716–9725, 2020.

SUPPLEMENTARY MATERIAL

A Pl@ntNet label validation process

Only Pl@ntNet observations with a valid species name were included in the Pl@ntNet-300K dataset. The species name validation process of Pl@ntNet is based on a weighted majority voting algorithm taking as input the labels proposed by Pl@ntNet users with a principle of adaptive weights depending on the user's expertise and engagement. More precisely, the most probable label y_i of an image x_i is computed as:

$$y_i = \underset{k \in 1,\ldots,d}{\arg\max} \sum_{u \in U_i} w_u \mathbb{1}(y_i^u = k) \ ,$$

where U_i is the set of users who suggested a label for the image x_i and y_i^u is the label proposed by the user u. The weight w_u of a user u is computed as:

$$w_u = n_u^\alpha - n_u^\beta + b_0 \ ,$$

where n_u is the number of species observed by user u (*i.e.*, the number of species for which the user is author of at least one valid image), plus the number of distinct labels y_i^u proposed by a user u in the whole dataset. The constant power values $\alpha = 0.5$, $\beta = 0.2$ and b_0 are determined empirically. To be considered as valid, the label y_i of an image x_i must satisfy:

$$\sum_{u \in U_i} w_u \mathbb{1}(y_i^u = y_i) > \theta$$

where θ is a fixed threshold. Images with non-valid labels were discarded from Pl@ntNet-300K dataset.

B Hyperparameters

The hyperparameters used for the experiments in Section 4 of the paper can be found in Table 3.

Table 3: Learning rate, number of epochs and learning rate schedule for the different models. At each learning rate decay, the learning rate is divided by ten.

Models	Initial learning rate	Number of epochs	First decay	Second decay
mobilenet_v2, mobilenet_v3_large, resnet 18, 34, 50, 101, 152, densenet 121, 161, 169, 201, inception_v3, inception_v4, inception_resnet_v2, wide_resnet50_2, wide_resnet101_2, shufflenet	0.01	30	20	25
vgg11, alexnet, mobilenet_v3_small, squeezenet	0.001	30	20	25
efficientnet b0, b1, b2, b3, b4	0.01	20	10	15
vit	5e-4	20	15	-

C Evaluation of existing set-valued classification methods

Figure 9 compares the top-5 accuracy of several models when they are trained with either the cross entropy loss or the top-k loss by Berrada et al. [2018]. The hyperparameters used for training the top-k loss are the same as in Section 4. The smoothing parameter τ of the top-k loss is set to 1.0.

D 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.

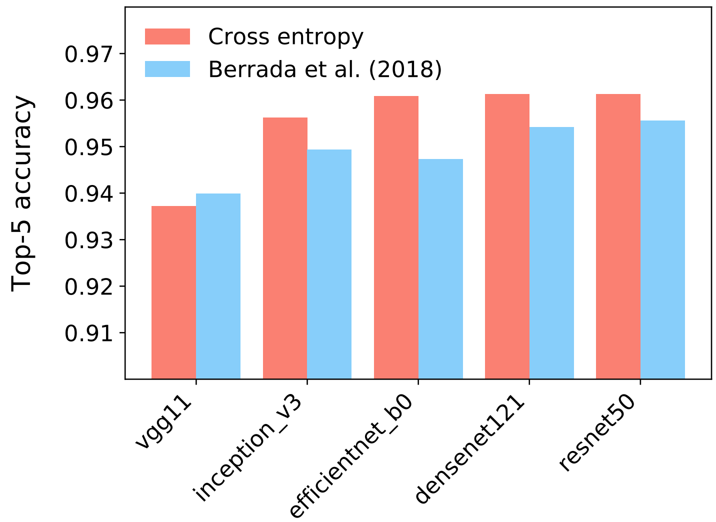


Figure 9: Comparison of top-5 accuracy of models trained with either the cross entropy loss or the loss by Berrada et al. [2018] (with $k = 5$)

- 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.

E 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?

F 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**

G 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.

H 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

I Distribution

- Will the dataset be distributed to third parties outside 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 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.

J 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?

K 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.