Pl@ntNet

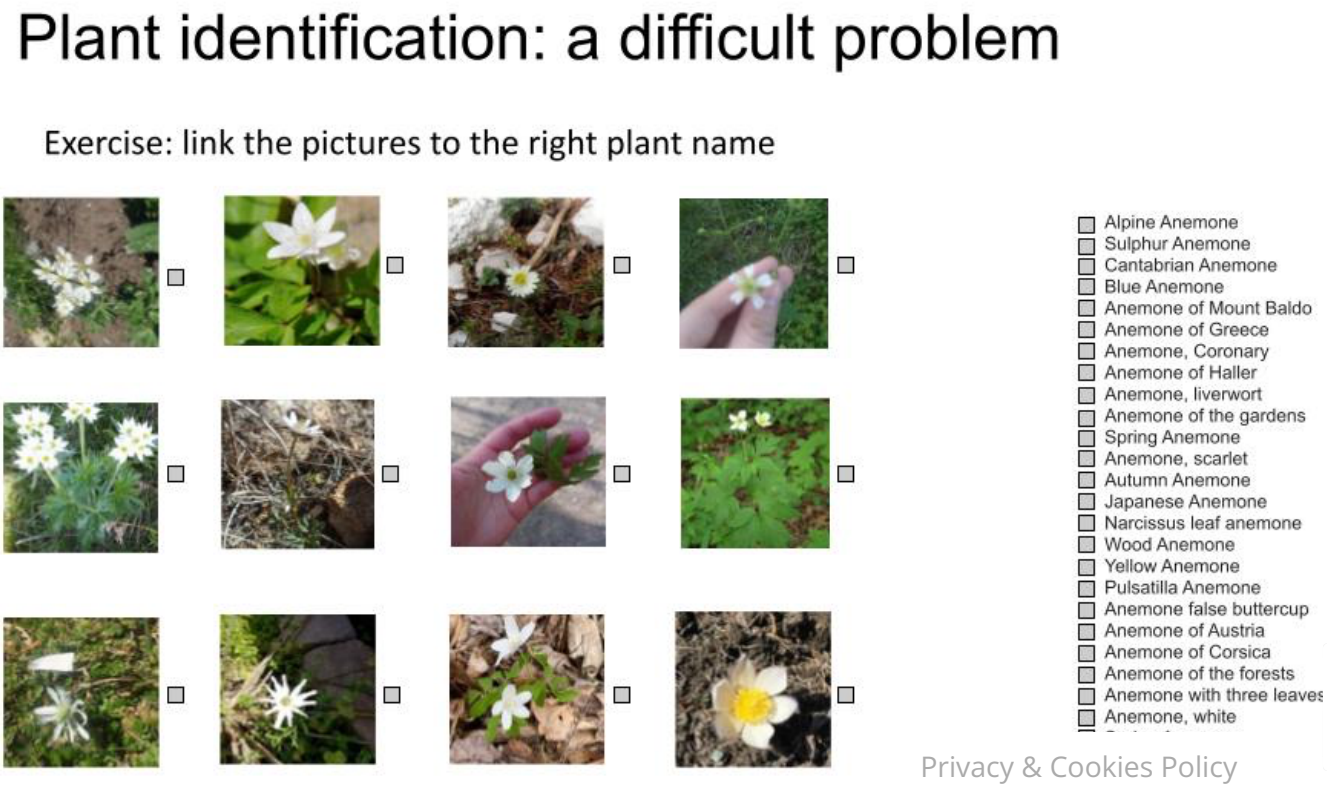
USE UNDERSTAND PARTICIPATE RESOURCES RESEARCH CONTACT

A Pl@ntNet dataset for machine learning researchers

POSTED ON 30 MARCH 2021 BY PL@NTNET

Direct link to the dataset: https://zenodo.org/record/4726653#.YhNbAOjMJPY

Plant identification is a difficult problem. Indeed, there are about 400K species of plants on earth and the problem is that the morphological characteristics that distinguish them are on the one hand very varied but also sometimes very subtle. We see for example on the image bellow that there is a very large number of anemones and that they are often distinguished only by details such as the shape of the petal or the way the leaves are inserted on the stem. We also see that the photographs made by the users can sometimes be of average quality, which complicates the identification.



From a more fundamental point of view, we distinguish two types of uncertainty in the classification process. On the one hand, there is what is called aleatoric uncertainty, which is an irreducible uncertainty because it comes from the fact that the image we have contains only partial information about the plant. Typically, if two species have the same leaves but different flowers and you only photograph the leaf, you will have an intrinsic ambiguity that you will never be able to remove unless you go back to see the plant itself. On the other hand, there is what is called epistemic uncertainty which is more related to the error that the recognition algorithm makes, especially when it is trained on too few images which is unfortunately the case for most species (long-tail distribution).

To enable machine learning researchers to study these challenging problems, the Pl@ntNet team built a public dataset that retains the same ambiguity and unbalanced distribution characteristics as the full Pl@ntNet database but that is much smaller to facilitate its analysis. To do so, the full database was not randomly sampled at the image level as usually done. It rather split at the genus level, i.e. the level upon the species level in the taxonomy. This allows preserving the ambiguity of species belonging to a same genus while conserving a long tail distribution (most genus are also poorly populated). As a result, the shared dataset contains 306,293 images of 1081 species that retain a level of ambiguity and a distribution close to the complete database.

POSTED IN RESEARCH