Etat des Lieux PRE

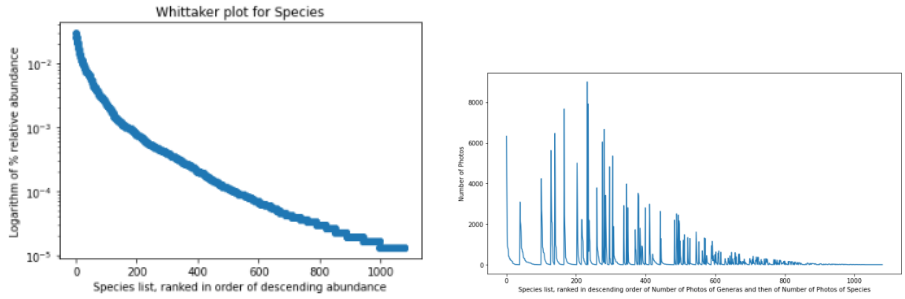
Reynaud Nils

IMAG, LIRMM, Université de Montpellier

Ce stage de recherche s'intéresse aux jeux de données présentant un fort déséquilibre entre les classes. De nombreux jeux de données synthétiques tels que CIFAR100 sont bien équilibrés avec le même nombre d'images par classe, diffèrant ainsi des données du monde réel au contraire souvent déséquilibrées et suivant par exemple des Long-Tail distribution. Les algorithmes de Deep Learning peuvent donner de mauvais résultats lorsqu'ils sont utilisés sur ces ensembles de données plus réalistes, puisque pour de nombreuses classes, le modèle ne dispose que d'une poignée d'images sur lesquelles s'entraîner.
Plus précisément, nous nous concentrerons sur le jeu de données Plantnet300k. Cet ensemble de données est extrait d'images du monde réel collectées dans le cadre du projet Plantnet. Ce jeu de données présente un fort déséquilibre entre ses classes, 80% des espèces ne représentant que 11 % des images, ce qui est logique puisque les quelques espèces les plus courantes sont facilement observées par les utilisateurs dans la nature, tandis que les nombreuses espèces rares sont plus difficiles à trouver. De plus, de nombreuses classes sont visuellement similaires dans ce jeu de données. Nous souhaitons l'étudier afin de modéliser la distribution du nombre d'images en fonction des différentes espèces de plantes(classes), des différents genres (métaclasses) ainsi qu'à l'intérieur de chaque genre. L'idée est d'ensuite créer des subdatasets réalistes de CIFAR100 basés sur ces résultats de modélisation.

1 Etude de Plantnet300K

Nous avons visualisé le jeu de données Plantnet300k en s'intéressant successivement à l'ensemble des espèces, puis à l'ensemble des genres, et enfin aux espèces à l'intérieur de chaque genre. Nous avons observé ces données à travers plusieurs graphiques de distribution d'abondance relative : les courbes de Lorentz (la part cumulée des images en fonction de la part cumulée des espèces), le graphique de Whittaker (le logarithme de l'abondance relative, par ordre d'abondance décroissante) ou encore le graphique de Preston (un histogramme de fréquence). Le nombre d'espèces par genre ainsi que d'autres éléments ont également été étudiés. Voici deux exemples de courbes obtenues:



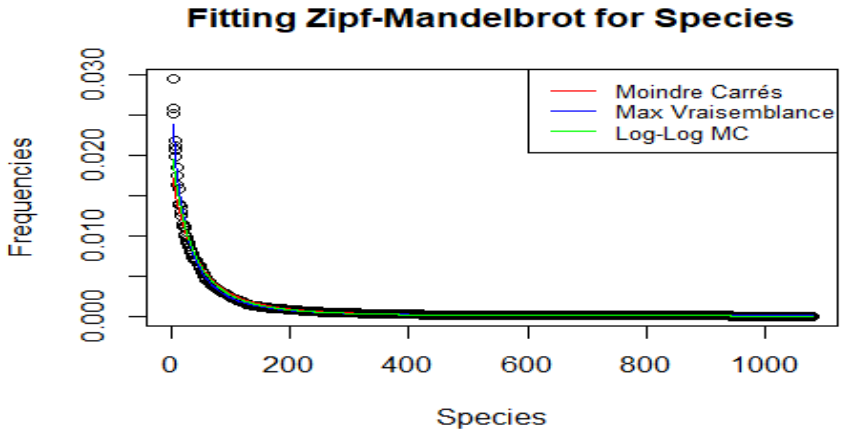
(a) Fréquence des espèces dans le jeu de données

(b) Nombre d'images des espèces, ordonnés par genres

Figure 1: Visualisation du jeu de données PlantNet300K

2 Modélisation

L'étape suivante consiste à modéliser nos données, et plus particulièrement nos diagrammes de Whittaker. Divers modèles mathématiques ont été envisagé ; sans ordre particulier, on y trouve des distributions continues telles que l'exponentielle, la lognormal, Pareto, mais aussi des distributions discrètes telles que Zipf ou Zipf-Mandelbrot. Différentes méthodes ont été utilisées pour les ajuster aux données, et de jolis difficultés sont apparus. On peut citer, avec divers niveaux de réussite, le package powerlaw de Python, des package R environnementalistes tels que vegan, ainsi qu'une méthode d'ajustement à Zipf-Mandelbrot utilisant les moindres carrés et des régression linéaire ou encore la méthode du maximum de vraisemblance. C'est Zipf-Mandelbrot qui est finalement privilégié, avec ci-dessous un exemple de fitting pour les espèces avec différentes méthodes.

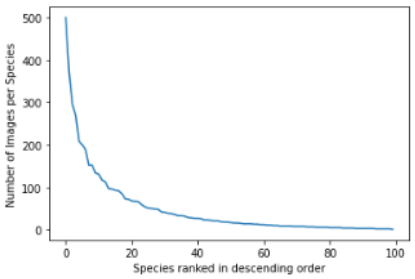


Cette modélisation nous a fourni un modèle mathématique que nous allons ensuite utiliser pour créer nos jeux de données synthétiques à partir de CIFAR100. Nous connaissons ainsi en effet la distribution mathématique de l'abondance relative, à la fois sur l'ensemble des espèces, ainsi qu'à l'intérieur de chaque genre.

3 Création de Datasets synthétiques

L'étape suivante consiste à créer plusieurs sous-ensembles de données du jeu de données CIFAR100 en suivant le modèle mathématique trouvé précédemment. On souhaite que les ensembles de données aient plusieurs points communs : ils doivent avoir le même nombre total d'images, chaque classe de chaque ensemble de données doit avoir au moins une image, et le graphique de Whittaker pour les espèces doit être le même : l'abondance relative des espèces classées par ordre décroissant doit suivre la même distribution. Dans un premier ensemble de données, les espèces sont distribuées de manière aléatoire au sein de cette même distribution ; dans un second, elles sont distribuées genre après genre, ce qui signifie que les x espèces ayant l'abondance relative la plus élevée appartiennent au même genre, puis les y espèces suivantes appartiennent à un autre genre, et ainsi de suite. Dans un troisième ensemble de données, la distribution des images sur l'ensemble des espèces ainsi qu'à l'intérieur de chaque genre suit une loi Zipf-Mandelbrot.

(a) Distribution du nombre d'images par espèces, ordonnés par genres, dans un des jeu de données



(b) Distribution du nombre d'images par espèces commune aux différents datasets

Figure 2: Création de datasets synthétiques

4 Déroulement du stage et Perspectives

Le stage se passe parfaitement bien outre les difficultés de modélisation rencontrés, j'ai la chance d'avoir de nombreux chercheurs prêts à m'aider en plus de mon tuteur de stage. Le stage se déroule en effet à la fois à l'IMAG

(Institut Montpelliérain Alexander Grothendieck) et au LIRMM (Laboratoire d'Informatique, de Robotique et d'Electronique de Montpellier), ce qui m'a permis, lorsqu'un des labo se vidait au plus fort des vacances, de me rendre à l'autre, et ainsi de toujours trouver de l'aide.
Pour la suite et fin du stage, outre la rédaction du rapport, ces jeux de données, "plus réalistes" que le CIFAR100 original, doivent désormais être entraînés par un même réseau neuronal, afin de comparer les résultats, la précision, le temps d'entraînement, non seulement entre eux mais également avec CIFAR100. Un problème à résoudre est la faible taille des datasets crées, trop faible pour y appliquer nombre de réseaux neuronaux classiques. L'utilisation d'une autre machine que la mienne est nécessaire afin de pouvoir utiliser un GPU.