

Projet E4

IA et Biodiversité



Tuteur de projet: **Abdelghani Chibani**

Sommaire

Introduction	3
Fonctionnement général	4
Présentation du dataset	5
Modèle CCN	5
Comparaison avec modèles existants	6
Site	8
Problèmes rencontrés	9
Ce que nous avons appris	10
Continuité	11
Conclusion	12
Bibliographie	13

Introduction

Les aéroports ne se limitent pas seulement à des bâtiments et des pistes de décollage. En réalité, une partie de leur terrain comprend des prairies et des terrains vagues qui favorisent fortement la biodiversité. Toutefois, cette biodiversité peut engendrer divers problèmes.

Par exemple, l'aéroport de Roissy Charles-de-Gaulle où l'on peut trouver de nombreux types d'insectes, mais également 170 espèces d'oiseaux différents ainsi que certains mammifères. D'après la direction de l'aéroport, les oiseaux et les mammifères représentent un danger qui peut entraîner des risques de retard et donc des pertes économiques.

De nos jours, on constate qu'il y a énormément d'aéroports dans le monde où, afin de lutter contre les invasions végétales et animales, les responsables sont contraints d'utiliser des pesticides ou de chasser les animaux des terrains sans égard pour la biodiversité.

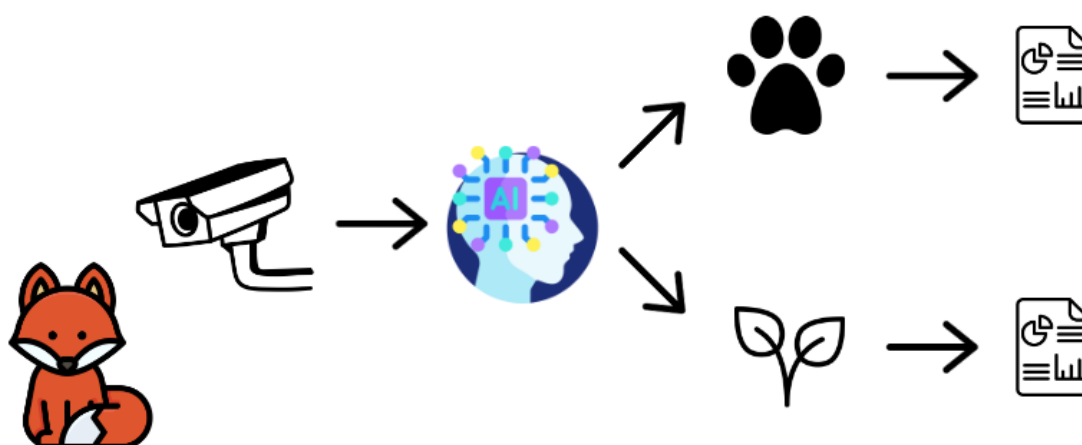
Afin de combiner une gestion efficace de l'aéroport et un respect de la nature, une bonne connaissance de la biodiversité présente sur les lieux est donc nécessaire.



L'objectif de notre projet est donc d'utiliser des images prises par des caméras de sécurité ou des drones et d'identifier l'espèce animale ou végétale affichée. Cette connaissance permettra une action ciblée afin de conjuguer respect de la biodiversité et efficacité.

Fonctionnement général

Pour ce qui est du fonctionnement général, nous allons créer une fonction qui prend comme entrée une photo d'une plante ou d'un animal. Elle va ensuite différencier les animaux des plantes puis va donner le nom de l'espèce correspondante et ceci à partir d'un modèle pré-entraîné sur un large jeu de données.



En raison de contraintes temporelles et techniques nous avons fait le choix de seulement nous concentrer sur la partie flore. Nous entraînons également notre modèle sur un dataset préexistant et ceci pour les mêmes raisons. L'idée est ensuite de créer un site web qui servira d'interface entre l'utilisateur et le modèle.

Présentation du dataset

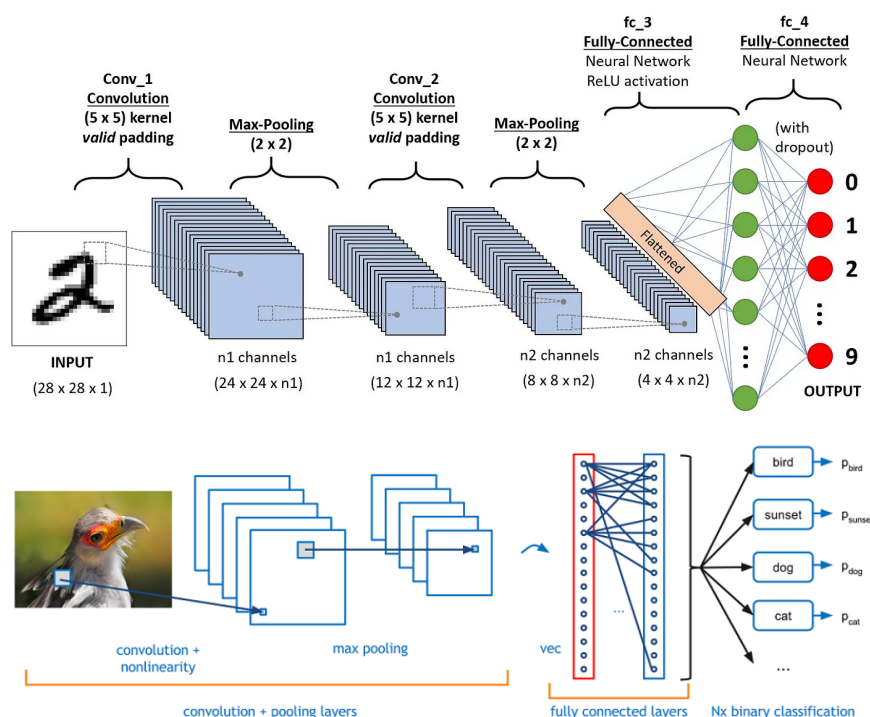
L'ensemble de notre projet a été réalisé sur le jeu de données du site plantnet. Ce jeu de données d'environ 30 Go comprend plus de 300 000 images réparties en 1081 classes. Ce dataset est divisé en trois parties, un train, un test et un val. A noter que toutes les classes ne disposent pas du même nombre d'images ce qui peut créer des soucis de performance du modèle en raison du déséquilibre de classe. Nous utilisons ce jeu de données car c'est l'un des plus complets en accès libre sur internet. Il est également déjà labellisé ce qui nous fait gagner un temps précieux.

Modèle CCN

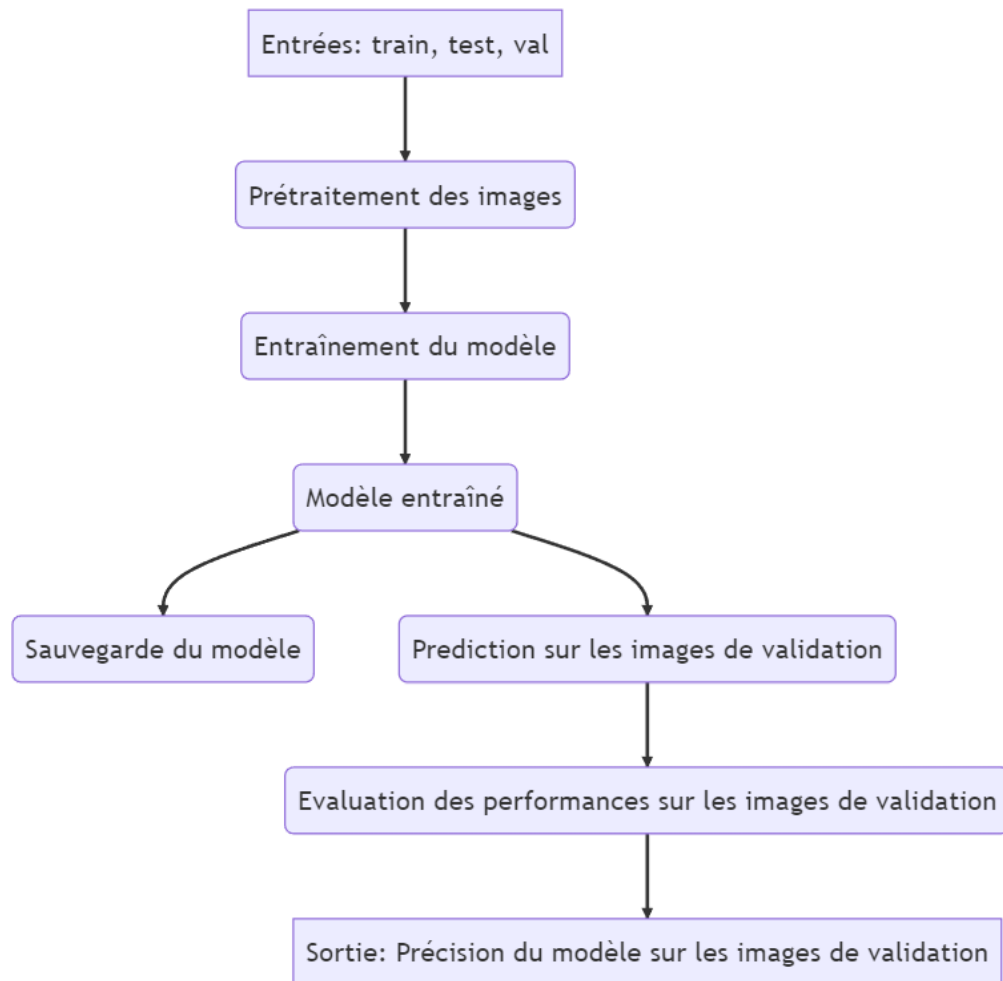
Pour réaliser la reconnaissance d'images, nous allons utiliser un réseau de neurones convolutifs (CNN). Les CNN utilisent des opérations de convolution pour extraire les caractéristiques des images en les parcourant avec des filtres. Ces filtres permettent de détecter des patrons et des traits spécifiques dans les images, tels que des bords, des contours, des textures. L'idée étant de repérer des motifs afin que le modèle puisse prédire la classe d'une image qu'il n'a jamais vu en s'appuyant sur les caractéristiques de cette espèce.

Ensuite, les résultats de la convolution sont suivis d'une opération de pooling, qui permet de réduire la taille des données en sélectionnant les valeurs les plus importantes. Cette opération a pour but de réduire la dimensionnalité des données afin de rendre le traitement plus rapide.

Les couches de neurones convolutifs et de pooling sont ensuite empilées les unes sur les autres pour former un modèle CNN complet. On peut s'appuyer sur le schéma suivant afin de mieux comprendre le fonctionnement général.



Pour réaliser notre modèle nous utilisons Keras ainsi que le jeu de données de plantnet. Concernant le programme en lui même voici son fonctionnement général:



La partie validation permet d'avoir une idée des performances du modèle et ainsi de voir s'il est satisfaisant ou pas. A noter qu'ici en raison de la complexité des données, il ne faut pas s'attendre à une performance élevée, le modèle fera beaucoup d'erreurs et répondra même des fois complètement à côté. Par ailleurs nous avons seulement réalisé les prédictions à partir d'un modèle CNN alors que dans le cas d'un modèle plus professionnel il y a plusieurs modèles qui entrent en jeu afin de réaliser la même prédiction et ainsi améliorer sa fiabilité.

Comparaison avec modèles existants

Afin d'avoir une idée du degré d'efficacité de notre modèle, nous effectuons une comparaison sur un nombre réduit d'image avec des modèles pré-entraînés qu'on peut trouver sur le github du projet plantnet.

L'idée est la suivante, nous prenons un modèle pré-entraîné ainsi qu'une image et nous demandons au modèle de prédire les 5 espèces les plus probables pour cette image. Voici un exemple réalisé avec torch et le modèle vgg11 qui est le modèle le plus performant disponible gratuitement sur le github.



	indice	values	proba	Nom
0	152	16.361076	0.998176	Pancratium_maritimum
1	686	8.150163	0.000271	Tradescantia_pallida
2	191	7.475308	0.000138	Tradescantia_pallida
3	868	7.365052	0.000124	Pancratium_maritimum
4	714	7.129431	0.000098	Dendrobium_crumenatum

La première image correspond à l'image dont il faut prédire l'espèce, les 5 autres images sont des illustrations des 5 espèces les plus probables afin d'avoir une confirmation visuelle. En effet, il peut arriver que le modèle se trompe de prédiction mais affiche une plante assez similaire. Le tableau donne les 5 espèces les plus probables avec leur probabilité.

Dans ce cas ci le modèle prédit une appartenance à l'espèce Pancratium Maritimum avec un taux de confiance de 99,8%. Ce qui est un résultat très satisfaisant.

Cependant il faut noter que ce modèle n'est pas parfait et qu'il lui arrive de se tromper. Lorsque les images sont bien nettes et que seule la plante concernée apparaît le modèle à des prédictions correctes, cependant dans le cas contraire il lui est beaucoup difficile d'arriver à un bon résultat. Ceci est notamment dû à l'inégalité du jeu de données, en effet certaines classes ne sont entraînées que sur quelques images ce qui les rend presque impossible à détecter par le modèle.


Il est également intéressant de comparer les performances avec le site de plantnet. Sur ce site, le modèle est bien plus efficace et il est combiné avec un système de zoom qui permet de prendre en compte des images mal cadrées. Avec une série de tests on se rend compte que le modèle du site est bien plus performant que celui que nous avons créé. Ceci s'explique par un plus grand nombre de classes enregistrées, un modèle mieux entraîné et un prétraitement des images.

Site


Afin de rendre notre projet un peu plus interactif nous avons créé un site en utilisant streamlit. Le principe est simple, l'utilisateur peut uploader une image et choisir le modèle qu'il veut utiliser et le site va donner les 5 espèces qui correspondent le plus en y ajoutant leur probabilité. Un bandeau est également affiché pour permettre à l'utilisateur de voir une photo de chacune de ces 5 espèces. Cela permet un examen visuel qui valide ou non la prédiction du modèle.

Classification d'espèce

Choisissez une image

 Drag and drop file here
Limit 200MB per file • JPEG, JPG

Browse files

 encoreuntest.jpg 84.6KB



Choisir le modèle

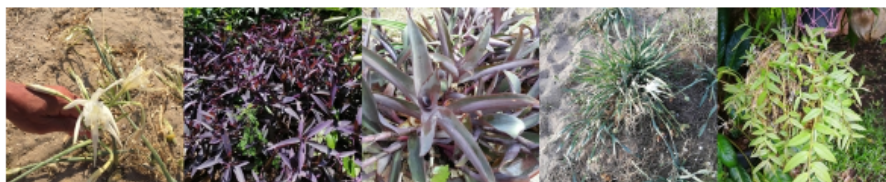
vgg11



Prédiction du modèle



Image originale



5 espèces possibles

	proba	Nom
0	0.998	Pancratium_maritimum
1	0.0003	Tradescantia_pallida
2	0.0002	Tradescantia_pallida
3	0.0001	Pancratium_maritimum

Problèmes rencontrés

Comme tout projet, nous avons dû pendant ces quelques semaines faire face à de nombreux soucis. La plus grande difficulté pour nous résidait dans le temps d'apprentissage du modèle. En effet, sur nos machines personnelles l'entraînement du modèle sur l'ensemble du jeu de données prenait environ 30h. Il n'était donc pas envisageable de s'attaquer directement à l'entraînement sur l'ensemble du jeu de données. Afin de pouvoir quand même avancer nous avons décidé de commencer par créer un modèle sur 10 classes. L'objectif étant de pouvoir l'entraîner rapidement et de réaliser différents tests d'optimisation.

Le deuxième problème a concerné l'hétérogénéité de nos spécialités. En effet dans notre groupe nous avons 2 personnes en DSIA, 2 en Informatique et 2 en Génie Industriel. Le projet étant majoritairement axé sur du machine learning et de la manipulation de données, les personnes en DSIA étaient donc les seules à avoir un bagage scolaire dans ces disciplines. Afin de garantir une charge de travail équilibrée entre tous les membres du groupe nous avons décidé de mettre en place des sessions de formation et d'auto-apprentissage pour permettre aux membres de notre groupe qui n'ont pas de connaissances en machine learning et manipulation de données d'acquérir au moins les compétences nécessaires à la compréhension du code.

Grâce à ces efforts, nous avons réussi à surmonter la bigarrure de nos spécialités et nous avons pu travailler efficacement ensemble.

Ce que nous avons appris

Pendant ce projet nous avons eu l'occasion de développer nos compétences dans deux domaines distincts. Tout d'abord dans le domaine technique et deuxièmement dans le domaine de la gestion de projet.

Concernant les apports de la partie gestion de projet, la première chose est bien sûr la collaboration et la communication entre les différents membres du groupe. En effet, collaborer et communiquer est le point central qui nous a permis de remplir nos objectifs. Grâce à cela nous avons pu exploiter au maximum les compétences et les qualités de chacun, compétences étant toutes différentes car nous venons de trois filières différentes. Nous avons aussi pu prendre en compte l'avis de tout le monde afin de régler les problèmes et d'atteindre l'objectif commun. Nous avons ensuite dû apprendre à bien organiser notre temps car nous n'avions qu'un seul jour de travail dans la semaine et nos cours à côté.

Tout cela pourra être très précieux dans notre futur, dans le monde professionnel, où la capacité à travailler en groupe est primordiale. Également la gestion en parallèle de ce projet et d'autres projets scolaires nous a permis de nous améliorer en multi-gestion. En effet, il est assez rare en entreprise de n'avoir qu'une seule mission pour un seul projet.

Dans le cadre de projet nous avons eu également la possibilité de développer nos compétences techniques. Bien que ces compétences ne soient pas forcément en accord avec nos filières, il est toujours intéressant et valorisant d'augmenter notre champ de savoirs. Premièrement, nous avons pu en apprendre davantage sur les modèles CNN et plus largement sur le fonctionnement du machine learning. Cette discipline prenant de plus en plus d'ampleur de nos jours, ce projet a été pour nous une véritable chance de s'intéresser sérieusement à la question. Dans un deuxième temps, nous avons pu approfondir nos connaissances sur le langage python en découvrant de nouveaux packages et nouveaux moyens de coder.

Tout ceci nous permet sans nul doute de dire que nos compétences dans ces domaines sont bien meilleures qu'au début du projet, la partie apprentissage est donc une franche réussite.

Continuité

Par manque de temps nous n'avons pas pu faire tout ce que nous voulions, cependant nous avons pour objectif initial de développer en parallèle de la reconnaissance de la flore une reconnaissance de la faune. Malheureusement, il est beaucoup plus complexe de créer un modèle d'identification d'espèces animales en raison du flou de mouvement à prendre en compte et de la taille de certains insectes.

Deuxièmement, il est possible d'optimiser notre modèle par exemple en modifiant certains paramètres comme le nombre de batch ou le nombre d'epochs. Un changement de modèle de machine learning pourrait également permettre une augmentation de la véracité des prédictions. Finalement, il est possible de fine-tuner notre modèle ainsi que ceux disponibles sur le github de plantnet en ajoutant des images et en rééquilibrant les classes. L'idée étant au final d'avoir un modèle précis. Par ailleurs, il faut noter que ces étapes peuvent prendre un temps considérable que ce soit en temps d'exécution ou en temps passé à regrouper de nouvelles images.

Troisièmement, il est envisageable de créer un algorithme de prétraitement qui s'appuie sur notre modèle pour traiter l'image et ceci afin que le modèle donne la meilleure réponse possible. On peut penser à un zoom dans l'image ou à un système de flou d'arrière-plan. Ce pré-traitement permettra non seulement d'améliorer les performances du modèles mais également de récolter de bonnes images pour fine-tuner le modèle.

Quatrièmement, à l'instar de plantnet il est possible d'encapsuler le modèle dans une application mobile afin que n'importe qui puisse prendre en photo une plante dans la nature et identifier son espèce. On peut également rajouter un système de suggestion ou l'utilisateur pourra rectifier la prédiction si jamais elle n'est pas correcte. Grâce à cela on accèdera à une image labellisée qui pourra ensuite être utilisée pour améliorer le modèle.

Pour finir, dans le cas d'une application dans un cas réel, par exemple dans un aéroport il serait judicieux de fabriquer un modèle prenant uniquement en compte la biodiversité locale. En effet, pas besoin de savoir reconnaître le palétuvier en Ile de France. Cette approche permettra non seulement d'avoir un modèle plus simple mais également de restreindre le nombre de classes ce qui fera automatiquement baisser le nombre de potentielles erreurs de classification.

Conclusion

Pour conclure, ce projet nous a permis de mieux approcher le domaine de l'IA appliquée à la reconnaissance d'image. Il est important de noter que ce domaine est encore en développement et que par conséquent il n'existe pas de solutions et de modèles parfaits. La réalisation de notre propre modèle nous a démontré la difficulté que cela représentait de construire un modèle fiable. Nous avons également pu comparer notre travail avec des modèles préexistants afin de juger nos accomplissements. Dans l'ensemble nous sommes très satisfait d'avoir eu la possibilité de réaliser ce projet. Nous avons beaucoup appris que ce soit d'un point de vue technique mais aussi d'un point de vue gestion de projet. Nous n'avons malheureusement pas pu aller au bout de ce que nous voulions faire mais nous espérons que ce projet aura la succession qu'il mérite.

Bibliographie

https://github.com/ClementTh/IA_biodiveristy

<https://plantnet.org/>

<https://identify.plantnet.org/fr>

<https://github.com/plantnet/PlantNet-300K/>

https://www.google.com/url?sa=i&url=https%3A%2F%2Farchzine.fr%2Flifestyle%2Fart%2Fpaysage-fleuri%2F&psig=AOvVaw1yHHmXD4-uCII6lSPpAMLj&ust=1681907894818000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCPjchKi5s_4CFQAAAAAdAAAAABAD

<https://www.flaticon.com/fr/>

Image Classification using CNN : Python Implementation - Analytics Vidhya