LIF PROJET

RC4. Kaggle Challenge

Rapport sur le Projet effectué du 14 /09/2018 au 20/12/2018 par :
PACCALIN Axel et INAYA Victor

Étudiants en :

LICENCE 3 D'INFORMATIQUE

2018-2019

A l'université :

Université Claude Bernard Lyon 1

PLAN

I. SUJET

II. NOS PROJETS

III. DIFFICULTÉS RENCONTREES

IV. CONCLUSION

I. SUJET

Kaggle est un site internet basé sur l'intelligence artificiel qui permet à des entreprises ou organismes de poster des projets qui sont appelé challenges. Ils vont fournir des données et explicité un problème et ainsi n'importe quelle équipe ou personne inscrite sur le site peut soumettre une solution. Sachant que certain projets propose un cash prize ainsi les équipe ayant obtenu les meilleurs résultats peuvent être rémunéré.

II. NOS PROJETS

A. Challenge 1: Digit Recognizer

Pour notre premier challenge nous avons décider de réaliser un challenge déjà complété, l'idée étant d'étudier comment ce déroule un challenge afin de voir les différentes étapes de conception d'un algorithme de Machine Learning. Dans ce challenge l'objectif est d'identifier correctement les chiffres d'un ensemble de données contenant des dizaines de milliers d'images manuscrites. Ce challenge est un peux un exemple type du fonctionnement de kaggle, c'est-à-dire qu'une grande partie des équipes qui ont travaillé dessus ont mis à disposition leurs travail dans la section kernel. Ce qui fait office de tutorial et ce fut donc l'idéal pour commencer notre projet.

Dans un premier temps il a fallu nous familiariser avec python ainsi que keras (tensorflow), numpy, pandas, scikit learn qui sont des bibliothèques majeures dans le domaine de l'apprentissage automatique. Donc dans un premier temps nous avons étudier une solution simple qui utilisait Scikit-learn, puis afin d'obtenir un meilleur résultat nous avons décidé de mettre en place un réseau neuronal convolutif qui nous a permis d'atteindre un pourcentage de succès de 99,98%. Ce réseau nous a servit ensuite de base pour nos autres challenges.

B. Challenge 2: A-Z Handwritten

Fort de notre expérience sur les chiffres manuscrits nous avons décidé de nous attaquer aux lettres manuscrites. Ce challenge a été assez vite complété car le type de donné est sensiblement le même que pour notre premier challenge ainsi nous n'avons eux que quelque modification à faire pour adapter notre algorithme à ce challenge.

Malheureuses il n'était possible de soumettre nos résultats sur kaggle ainsi nous avons réalisé nous même une estimation de la précision de notre programme en prenant soin de découper au préalable les données en deux sous-catégories une pour l'entrainement et une pour le tester. Ainsi le score obtenu pour ce challenge est de 99,65%.

C. Challenge 3: QuikDraw Doodle Recognition

Pour notre troisième challenge nous avons décidé de nous intéresser à un challenge non complété, c'est-à-dire en cours de compétions. Ce challenge a été publié en tant que jeu expérimental pour éduquer le public de manière ludique sur le fonctionnement de l'intelligence artificielle. Le jeu invite les utilisateurs à dessiner une image représentant une certaine catégorie, telle que « banane », « table», etc. Le jeu a généré plus de 1 milliard de dessins, dont un sous-ensemble a été rendu public comme base de l'ensemble d'entraînement de cette compétition. Ce sous-ensemble contient 50 millions de dessins comprenant 340 catégories d'étiquettes. Les données d'entraînement proviennent du jeu lui-même, les dessins peuvent être incomplets ou ne pas correspondre au label. Nous avions donc besoin de créer un outil de reconnaissance capable de tirer efficacement parti de ces données bruyantes et de bien fonctionner sur un ensemble de tests étiquetés manuellement à partir d'une distribution différente.

Notre tâche consistait à créer un meilleur classificateur pour le logiciel existant Quick, Draw. En faisant progresser les modèles sur cet ensemble de données pour améliorer les solutions de reconnaissance de modèles plus largement.

Cette fois-ci nous avions aussi des données manuscrites mais aussi une autre information très intéressante qui est l'ordre des traits. Donc tout l'enjeux de départ à été de trouver comment organiser nos données d'entrées, c'est-à-dire trouver la bonne taille pour nos images afin de réduire leurs poids sans perdre trop

d’information mais aussi d’intégrer l’ordre des traits qui est souvent déterminant pour prédire la catégorie du dessin. Ensuite s’en ai suivi de longue série de teste afin d’améliorer nos résultats, pour arriver à un pourcentage final de 65,34%

D.Challenge 4: Human Protein Atlas Image Classification

En parallèle nous avons réalisé un autre challenge dont le but est de prévoir des étiquettes de localisation des organites de protéines pour chaque échantillon. Au total, 28 étiquettes différentes sont présentes dans l'ensemble de données. L'ensemble de données est acquis de manière hautement normalisée en utilisant une modalité d'imagerie (microscopie confocale). Cependant, l'ensemble de données comprend 27 types de cellules de morphologie très différente, qui affectent les profils protéiques des différents organites. Tous les échantillons d'images sont représentés par quatre filtres (stockés sous forme de fichiers individuels), la protéine d'intérêt (vert) et trois repères cellulaires : noyau (bleu), microtubules (rouge), réticulum endoplasmique (jaune). Le filtre vert devrait donc être utilisé pour prédire l'étiquette, et les autres filtres sont utilisés comme références. Les différentes classes sont : Nucléoplasme, Membrane nucléaire, Nucléole, Centre fibrillaire nucléole, Taches nucléaires, Corps nucléaires, Réticulum endoplasmique, Appareil de Golgi, Peroxysomes, Endosomes, Lysosomes, Filaments intermédiaires, Filaments d'actine, Sites d'adhésion focaux, Microtubules, Microtubules se termine, Pont cytokinique, Fuseau mitotique, Centre d'organisation des microtubules, Centrosome, Gouttelettes lipidiques, Membrane plasma, Jonctions cellulaires, Mitochondries, Agresome, Cytosol, Corps cytoplasmiques, Baguettes et bagues.

Les deux principales spécificités de ce challenge sont la particularité des données ainsi nous n’avons pas ici une seule photo mais 4 images représentant les différents filtres d’écrit plus haut.

La deuxième particularité étant le multi classe ainsi une image peut avoir une ou plusieurs classes. Vu qu’on parle ici de cellule de nombreuse classe sont présente dans une grande partie des photos.

Il a donc fallu au préalable faire un travail de préparation des données avant de les injecter dans notre réseau. Ensuite après entrainement du réseau sur le jeu de donné on à pu lancer notre réseau sur les données test afin d’obtenir un résultat. Le principal enjeu du challenge a été de définir un seuil d’acceptation c'est-à-dire à partir de quel taux de reconnaissance dans la phase de teste on décide que l’image appartient à la classe en question. Taux qui a finalement été établi à 0,5, c'est-à-dire que si le programme juge que l’image A appartient à 0,65 à une classe on admet que le la classe est présente sur l’image.

Après soumissions de nos résultats sur kaggle nous obtenons un score de 0,258% ce qui est assez faible mais du fait de la complexité des données cela se comprend.

III. DIFFICULTÉS RENCONTREES

La principale difficulté à laquelle nous avons été confronté a été la longueur des cycles d'édition et de test. En effet le Machine Learnings demande énormément de calcul et donc une grande puissance, ainsi nos machines était souvent un frein à l'obtention de résultat rapide.

C'est notamment la raison pour laquelle nous avons mit en place un système de « feed » des données qui permet de mettre en place une sorte de sous traitement des données par petit paquet afin d'optimiser leurs traitements.

Mais même en utilisant cette technique il est difficile de rivaliser avec les grands laboratoires de recherche du domaine qui eux dispose notamment d'une grande maitrise du sujet mais aussi de machine extrêmement performant et optimisé pour le Machine Learning.

IV.CONCLUSION

Le bilan est très positif, pour nous ce projet nous a permit pour la première fois de nous intéresser à l'intelligence artificiel qui est sans doute une des grandes technologies de demain. Ça aussi été l'occasion pour nous de réaliser pour la première fois un projet exclusivement en python, ainsi de nous initier aux concepts et structure basique des réseaux neuronaux convolutif.