T-DEV-810
Image recognition

Brice TOSSIM
Hugo LEVEBVRE
Stephane OLIVIEIRA DOS SANTOS
Sébastien KELLER
Maximilien DESNOS

1. INTRODUCTION

1.1 Contexte

Ce ne sont pas des médicaments mais leur rôle en médecine est crucial. Appareils auditifs, prothèses, pacemakers, tests de grossesse, pompes à perfusion :tous font partie des dispositifs médicaux, des produits de santé dont l'action fonctionne par un moyen mécanique. Comme dans de nombreux autres domaines, là aussi, l'intelligence artificielle a permis de nouvelles avancées, si bien que de nombreux nouveaux dispositifs médicaux fonctionnent grâce à un algorithme. Du logiciel de radiologie qui arrive à détecter les tumeurs sur une imagerie médicale aux logiciels compagnons sur application mobile, qui permettent aux patients de suivre leur maladie au jour le jour, les utilisations de l'intelligence artificielle sont variées.

1.2 Le machine learning

Le Machine Learning peut être défini comme étant une technologie d’intelligence artificielle permettant aux machines d’apprendre sans avoir été au préalablement programmées spécifiquement à cet effet. Le Machine Learning est explicitement lié au Big Data, étant donné que pour apprendre et se développer, les ordinateurs ont besoin de flux de données à analyser, sur lesquelles s’entraîner.

De ce fait, le Machine Learning, issu par essence du Big Data, a précisément besoin de ce dernier pour fonctionner. Le Machine Learning et le Big Data sont donc interdépendants.

1.3 Notre projet

Ici, des médecins nous ont demandé d'utiliser le machine learning afin de les aider à détecter une pneumonie sur une radio. Pour ce faire les médecins nous ont donnée accès à 3 jeux de données correspondant à des radiologies de poumons sain et atteint de pneumonie.

1.4 Nos moyens

Ce projet de machine learning à été réalisé en Python 3 avec notamment les librairies Numpy, Skimage ou encore Sklearn.

2 Traitement des données

Dans un premier temps nous avons besoin d'importer les différentes librairies nécessaire aux traitement des données:

- Numpy: destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.
- Matplotlib : Sert à la création de graphique
- os : Ce module fournit une manière portable d'utiliser les fonctionnalités dépendantes du système d'exploitation.
- PIL : Python Imaging Library est une bibliothèque de traitement d'images pour le langage de programmation Python. Elle permet d'ouvrir, de manipuler, et de sauver différents formats de fichiers graphiques.
- Sklearn : Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique.

```
import numpy as np
import matplotlib.pyplot as plt
import os
from numpy import asarray
from PIL import Image
from sklearn.ensemble import RandomForestClassifier

from sklearn import metrics
from sklearn import tree
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Ensuite nous allons définir les différents dossiers qui seront utilisés pour les jeux de données

```
14 mainDIR = './chest_xray/'
15
16 train_folder= mainDIR + '/train/'
17 val_folder = mainDIR + '/val/'
18 test_folder = mainDIR + '/test/'
19
```

Nous allons maintenant charger, stocker et transformer les images stocker dans les dossiers vu précédemment. Pour se faire on va utilisé la fonction suivante, loadDataset, en y passant la variable correspondant au chemin du dossier contenant les images.

```
75 train = loadDataset(train_folder)
76 test = loadDataset(test_folder)
77 val = loadDataset(val_folder)
```

Cette fonction génère un tableau contenant toutes les nuances de gris des images, pixel par pixel. Le tableau x_array contiendra tous les tableaux de pixels pour chaque image et y_array est le tableau qui contiendra tous les labels qui serviront à déterminer si les tableaux seront issus d'une pneumonie ou non.

Dans un premier temps nous allons charger le contenu du dossier dans la variable listDir (L33), elle contiendra deux autres dossiers PNEUMONIA et NORMAL.

Nous allons boucler sur le contenu de ces dossiers en omettant les fichiers cachés et traiter les images.

Dans la variable img (L38) nous allons charger le contenue de l'image. Nous nous retrouvons avec un tableau contenant lui même un tableau correspondant aux informations RGB du du pixel.
La variable img contient donc tout les pixels de l'image charger.

Mais le format obtenue n'est pas celui que l'on souhaite.

Tout d'abord nous ne pouvons pas comparer chaque pixel de chaque image pour savoir si le patient à une pneumonie.

Il va nous falloir réduire le nombre de pixel, nous allons ici réduire l'image à 50 sur 50 (L39) et stocker dans la variable img.

Le format de la variable img ne correspondant pas à nos besoin nous allons utilisés la fonction asarray (L40) pour avoir le tableau souhaité.

```
30 def loadDataset(folder):
31     x_array = []
32     y_array = []
33     listDir = os.listdir(folder)
34     for dir in listDir:
35         if dir[0] != ".":
36             for file in os.listdir(folder + dir):
37                 if(file[0] != "."):
38                     img = Image.open(folder + dir + '/' + file).convert('L')
39                     resize_img = img.resize((50, 50))
40                     img = asarray(resize_img)
41                     img_array = getImageValue(img)
42                     x_array.append(img_array)
43                     y_array.append(dir)
44     return {'samples': x_array, 'labels': y_array}
45
```

(L41) nous allons utiliser la fonction suivante qui prend en paramètre l'image précédemment traité. Cette fonction prend le tableau de valeur de l'image et le reformate pour qu'il puisse être lu par le modèle de l'arbre.

```
21 def getImageValue(img):
22     return_tab = []
23     for line_pixel in img:
24         for pixel in line_pixel:
25             return_tab.append(pixel)
26     return return_tab
27
```

Après avoir remplie les tableau x_array avec les tableaux de pixels et y_array avec les labels, nous allons générer un classificateur de forêt aléatoire (L62). Une forêt aléatoire est un méta-estimateur qui ajuste un certain nombre de classificateurs d'arbre de décision sur divers sous-échantillon de l'ensemble de données et utilise la moyenne pour améliorer la précision prédictive et contrôler le sur-ajustement.

```
62 clf = RandomForestClassifier()
63 clf.fit(train["samples"], train["labels"])
64
65 predicted = clf.predict(test["samples"])
66
67 print(accuracy_score(test["labels"], predicted))
```

La dernière étape est de passer (L63) nos tableaux samples qui correspond aux images et labels à partir de train à l'arbre qui se chargera de faire le travail à partir de ces données.

Pour finir nous utilisons l'apprentissage fait à la machine pour prédire l'existence d'une pneumonie dans les images contenue dans le tableau test.

Nous pouvons enfin d'afficher le taux de réussite de reconnaissance qu'à notre machine à reconnaître des poumons sein et des poumons avec une pneumonie.