Titre du Sujet 1 : Optimisation du Stockages d'Images pour le Machine Learning

Objectif :

Comment optimiser le stockage des images dans le cadre de l'IA et la computer vision en utilisant les méthodes de clustering ?

- Quelle méthode de clustering choisir pour la compression avec :
  - Les méthodes de Set Redundancy Compression
  - La compression vidéo

Contexte :

Pour entraîner une IA à reconnaître un objet, il faut constituer un dataset avec beaucoup d'images. Si l'on regarde ces datasets sur Kaggle, on peut remarquer que ces datasets sont lourds et qu'il y a beaucoup d'images similaires. Par exemple pour le challenge de classification ImageNet qui contient plus d'un million d'images avec 1000 classes, il y a des centaines de Gigaoctets de données à télécharger et stocker. Pour créer le fichier de téléchargement, les images de chaque classe sont compressées dans un zip, puis l'ensemble des zips sont compressés dans un autre zip. En utilisant un algorithme de compression mieux adapté à l'exploitation des redondances inter-image, la taille de ce fichier de téléchargement peut être réduit. Y a-t-il un moyen de réduire la taille du dataset en gardant le même niveau d'information pour entrainer nos algos ? On veut faire mieux en utilisant un algorithme de clustering puis de compression d'ensemble d'images. Le Set Redundancy Compression et la compression vidéo sont deux méthodes déjà utilisées pour la compression d'ensemble d'image. Ces deux méthodes utilisent des méthodes de clustering en amont pour découper l'ensemble d'images en sous-ensembles. Et donc on se pose la question de quel algorithme de clustering est-ce qu'on doit utiliser pour avoir les meilleurs résultats.

Descriptions des Activités :

- État de l'art des méthodes de clustering pour les images, le Set Redundancy Compression, et la compression vidéo.
- Recherche des méthodes de clustering et librairies Python
- Implémentation de la méthode Set Redundancy Compression en Python
- Recherche des librairies libres de compression vidéo Python
- Application de plusieurs méthodes de clustering pour les deux méthodes de compression et déduire comment choisir la méthode
- Comparaison des tailles des datasets compressés avec les méthodes clustering + SRC/compression vidéo contre le taux de compression en utilisant des zips.

Datasets : MNIST, CIFAR-10, CIFAR-100, ImageNet

Profil: Data Scientist/IA

Outils : Scikit Learn, Python, OpenCV

Encadrant :

Kaveena PERSAND
SII Sud-Ouest
kaveena.persand@sii.fr