Hackathon 2018

Décembre en Amérique du Sud

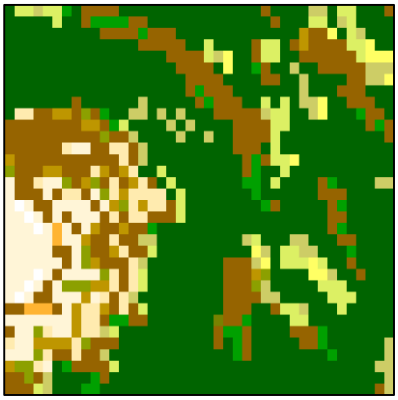
Contexte:

Le programme Sentinel de l'ESA vise à fournir gratuitement des images satellite des zones immergées de la planète à partir de différents capteurs. Le but premier de cette initiative est de permettre un suivi continu de la condition climatique à l'échelle planétaire. L'ouverture de cette océan de données a des répercutions aussi sur la recherche et l'innovation.

En effet, tout comme la mise à disposition des bases de données ImageNet, Mnist ou encore VOC a permis d'initier la vague de progrès en traitement d'images « naturelles », l'amoncellement de quantité importante d'images satellite permet aujourd'hui le développement de nouvelle solutions pour les applications spatiale.

Votre but lors de ces différents jours, sera de mettre à profits ces bases de données.

Sujet :



Votre tâche pendant ce Hackathon est d'entrainer un algorithme de classification des sols à partir des jeux de données qui vous sont fournis. Ces jeux de données dont issus des 4 canaux R, G, B et I d'images Sentinel 2 prise au-dessus du Brésil pendant le mois de décembre 2017. Afin de faciliter le développement de vos solutions, ces images ont été découpées en tuiles 16*16*4 pixels, associées à une classification des sols existante faite par photo-interprétation en 2009 et rassemblées dans différentes bases fichiers hdf5. Bien que les jeux de données fournis soient déjà conséquents, toute source de données complémentaire pourra être utilisée pour répondre à la problématique.

Afin d'évaluer vos algorithmes, l'équipe encadrante aura à disposition des jeux de données provenant des mêmes zones que celles des jeux d'entrainement mais prises un autre jour.

Données :

Les données qui vous sont fournies sont organisées en deux ensembles, les données d'entrainement et les données d'évaluations:

- **Les données d'entrainement :**

Chacun des fichiers contient deux datasets. Le dataset 'S2' contient une matrice de taille N*16*16*4 correspondant aux patchs d'images satellite et le dataset 'TOP_LANDCOVER' contient une matrice N*1 correspondant à la nature des différents patchs comme indiqué dans le relevé des sols de 2009.

Trois jeux de données de tailles différentes vous sont fournis afin d'effectuer vos entrainements. Les deux plus petits (1GB et 10GB) contiennent respectivement 1/40^{e} et 1/4 des échantillons du plus gros (40GB). Ces échantillons ont été extraits dans l'ordre du fichier. Un dernier fichier sera mis à disposition. Il contient 75GB de données annotées et pourra servir de ressources complémentaires pour vos solutions.

Indication :

Les jeux de données qui vous sont fournis sont des jeux de données réels, aussi bien au niveau de l'entrainement que de l'évaluation. Pensez à prendre en compte la répartition des classes.

- **Les données d'évaluations :**

Afin d'évaluer les performances de vos solutions, plusieurs jeu de données serviront de cas de tests. Ils sont organisés en trois catégories. Les jeux de données pred_full et pred_half ont été extrait de façon aléatoire à partir des zones couvertes par les jeux de données de 75GB et de 40GB (respectivement). Les jeux de données pred_samples vérifieront la robustesse de vos jeux de données aux changements de répartitions (les classes y sont présentes en proportions différentes).

Le tableau de la Figure 1 présente la correspondance des classes dans le jeu de données.

Class	Label	Red	Green	Blue
C0	Post-flooding or irrigated croplands (or aquatic)	170	240	240
C1	Rainfed croplands	255	255	100
C2	Mosaic cropland (50-70%) / vegetation (grassland/shrubland/forest) (20-50%)	220	240	100
C3	Mosaic vegetation (grassland/shrubland/forest) (50-70%) / cropland (20-50%)	205	205	102
C4	Closed to open (>15%) broadleaved evergreen or semi-deciduous forest (>5m)	0	100	0
C5	Closed (>40%) broadleaved deciduous forest (>5m)	0	160	0
C6	Open (15-40%) broadleaved deciduous forest/woodland (>5m)	170	200	0
C7	Closed (>40%) needleleaved evergreen forest (>5m)	0	60	0
C8	Open (15-40%) needleleaved deciduous or evergreen forest (>5m)	40	100	0
C9	Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m)	120	130	0
C10	Mosaic forest or shrubland (50-70%) / grassland (20-50%)	140	160	0
C11	Mosaic grassland (50-70%) / forest or shrubland (20-50%)	190	150	0
C12	Closed to open (>15%) (broadleaved or needleleaved, evergreen or deciduous) shrubland (<5m)	150	100	0
C13	Closed to open (>15%) herbaceous vegetation (grassland, savannas or lichens/mosses)	255	180	50
C14	Sparse (<15%) vegetation	255	235	175
C15	Closed to open (>15%) broadleaved forest regularly flooded (semi-permanently or temporarily) - Fresh or brackish water	0	120	90
C16	Closed (>40%) broadleaved forest or shrubland permanently flooded - Saline or brackish water	0	150	120
C17	Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil - Fresh, brackish or saline water	0	220	130
C18	Artificial surfaces and associated areas (Urban areas >50%)	195	20	0
C19	Bare areas	255	245	215
C20	Water bodies	0	70	200
C21	Permanent snow and ice	255	255	255
C22	No data (burnt areas, clouds,…)	0	0	0

Figure 1:Correspondance des classes dans le jeu de donées

Bagages informatiques

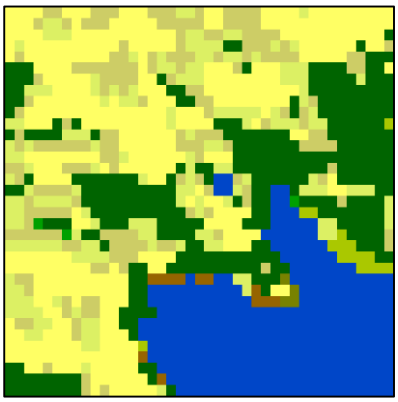
Afin de vous faire gagner du temps au démarrage, un notebook Hackathon.ipynb vous est fournis. Il contient quelques lignes de codes vous permettant de charger les données, d'entrainer un réseau de classification et d'effectuer une soumission sur un jeu de données d'évaluation. Ces lignes de codes ne sont là qu'à titre informatif.

Infrastructure et installations :

Afin de répondre à la problématique, il est conseillé de disposer de ressources de calcul suffisantes (GPU, en local, en cluster ou accès d'un cloud public), sur laquelle seront installés python, jupyter, ainsi qu'une librairie de calculs tensoriels pour l'apprentissage profond telles que tensorflow, theano, ou pytorch (pour le code fourni nous utilisons keras avec tensorflow en backend). La seule contrainte étant la nature des soumissions, les frameworks utilisés sont au choix des équipes (Spark…).

Déroulement de la compétition :

Le but de la compétition est d'obtenir les meilleurs performances sur les différents jeux de tests quelques soit la solution envisagée. A la fin de la compétition, il vous sera demandé d'effectuer une présentation de ce que vous avez entrepris durant les différentes séances.



Les jeux de données resteront disponibles pendant plusieurs semaines après la compétition afin de vous permettre de continuer vos recherches.

Bonne chance !