Pré-traitement d'Images de Fruits

Fruits!

OPENCLASSROOMS

Eloi Le Quilleuc, 04/01/2021

Introduction

Entreprise *Fruits!* : propose des solutions innovantes de récolte de fruits

Objectif : rendre possible l'identification de fruits via un smartphone

Données : images de fruits (pommes, carottes, …)

Mission : effectuer un pré-traitement d'images de fruits au sein d'une architecture Big Data

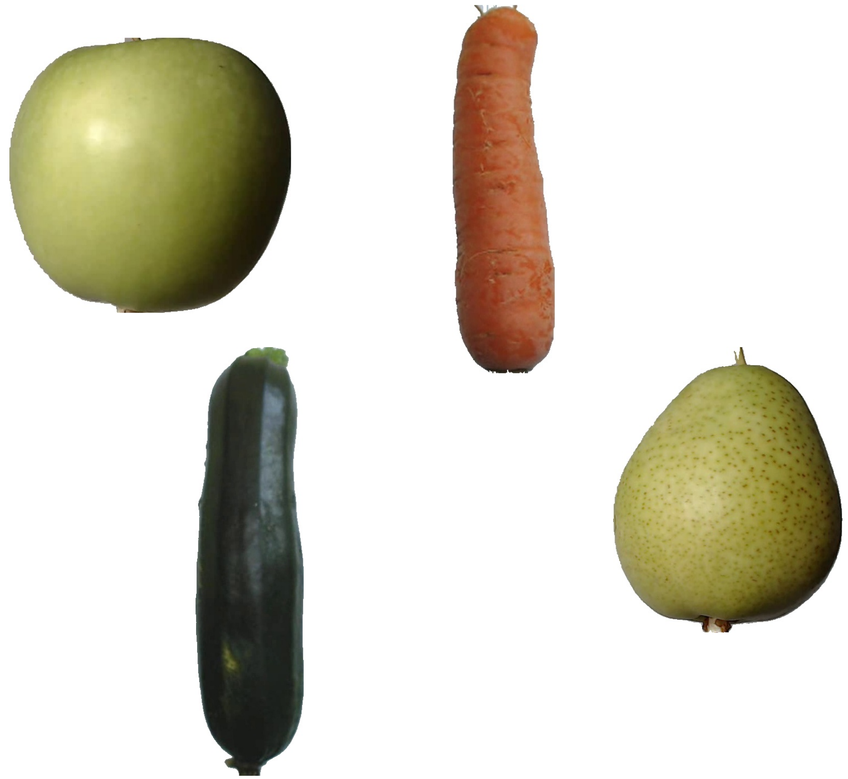
Les Données

Type : photos de fruits

Nombre d’images : ~ 6000 (~300 Mo)

Nombre de fruits : 23

Nombre d’images par fruit : 270 en moyenne



Quelques échantillons de photos de fruits et légumes

Pré-traitement des Images

1. **Featurisation** : image -> vecteur à partir du CNN pré-entraîné Resnet50 (dimension des vecteurs = 100k)

2. **Réduction de dimensions** : réduire les vecteurs à une dimension n ~ O(10) avec une PCA

Environnement Big Data

Fournisseur : AWS

Espace de stockage : S3

Calcul : EC2

Interface : Databricks

Ressources de Calcul

Worker/driver type : i3.xlarge, 30.5 GB, 4 cores

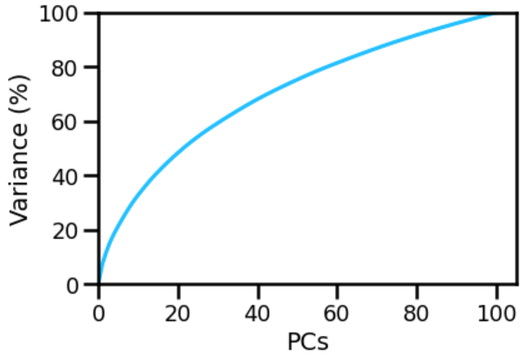
Min - Max workers : 2 - 8

Software : Apaches Spark 3.1.2, Scala 2.12, + bibliothèques ML

Réduction PCA

Étant données les ressources limitées, on choisit un nombre de composantes principales n = 100 (16 heures de calcul, ~40 €)

% de variance cumulée vs nb de PC



Sélection de PCs : variance > 1%, soit les 28 premières CP

Données Pré-traitées

Out[39]:

	path	label	pca[0]	pca[1]	pca[2]	pca[3]	pca[4]	pca[5]	pca[6]	pca[7]	pca[8]	pca[9]	pca[10]	pca[11]	pca[12]	pca[13]	pca[14]	pca[15]	pca[16]	pca[17]	pca[18]	pca[19]	pca[20]	pca[21]	pca[22]	pca[23]	pca[24]	pca[25]	pca[26]	pca[27]
0	dbfs:/mnt/fruit-360/test-compute/apple_hit_1/r...	apple_hit_1	-4.244428	-1.297873	-19.306000	1.027980	-12.189369	28.808328	33.051261	4.241512	65.718174	-20.346723	-12.926818	12.600773	10.726280	13.537684	-63.182794	17.919787	69.851866	13.306892	4.491259	-21.578765	-4.950536	11.870880	14.009452	-28.178817	28.235327	-4.757267	13.596134	-24.964799
1	dbfs:/mnt/fruit-360/test-compute/apple_hit_1/r...	apple_hit_1	-7.018301	-0.920443	-18.315981	1.381388	-12.419158	26.816941	32.507030	2.377980	62.481883	-16.071861	-12.668102	14.036984	10.910419	12.110568	-55.533819	18.842450	62.320071	9.863960	2.468955	-23.098668	-4.261786	11.004788	14.919112	-26.287162	31.100628	-4.804088	14.505976	-28.283948
2	dbfs:/mnt/fruit-360/test-compute/apple_hit_1/r...	apple_hit_1	-1.693223	-3.599726	-22.868818	-0.686607	-15.410438	26.981728	33.973652	1.065745	69.836426	-18.650492	-16.401048	19.044008	14.997612	14.152865	-64.210602	24.998482	69.357034	9.929316	5.393910	-26.230093	-2.139088	11.916052	18.734644	-33.314726	32.256622	-2.012242	19.683679	-36.692740
3	dbfs:/mnt/fruit-360/test-compute/apple_hit_1/r...	apple_hit_1	-6.182198	-1.590132	-16.349514	1.237405	-11.328907	24.901114	30.628064	4.463446	58.449516	-15.328397	-9.752992	14.119545	5.955996	9.052385	-46.039637	12.803706	55.353541	10.548072	-3.253482	-19.452759	-5.032876	9.753706	12.224153	-23.583799	29.677810	-2.058196	13.138348	-23.052202
4	dbfs:/mnt/fruit-360/test-compute/apple_hit_1/r...	apple_hit_1	-4.675239	-3.264141	-18.014367	3.049207	-16.580817	28.431690	36.584868	11.506236	69.417906	-11.865460	-12.744217	15.198413	8.637235	13.304050	-62.250095	13.321413	73.712128	14.221983	2.847803	-17.779351	-7.483926	12.544003	15.957012	-28.257754	27.994669	0.832953	14.178282	-27.772353

Dataframe
des 5 premières images et 28 PCS

Location : S3 bucket

Taille : 2.6 Mo (réduction : facteur 120)

Adresse :

Conclusion

Des images de fruits :

- pré-traitées

- disponibles sur le cloud (S3 AWS)

Une étape de pré-traitement :

- disponible sur le cloud (DataBricks)

- scalable (Pyspark, EC2 AWS)

Merci