Rapport de séance semaine du 17-12-2018

Cette semaine je me suis concentré sur une seule chose, réussir à importer les données de la base de l'EMNIST dans deux tableau python.

Cette étape semble facile, mais ce n'est pas du tout le cas. Il y a très peu de documentation sur l'EMNIST car la plupart des exemples sont fait avec l'MNIST (une version bien moins complète).

La base de l'EMNIST est une ressource divisée en plusieurs fichiers et qui existe en plusieurs versions, d'un poids de 700 mo, elle ne comporte pas moins de 700 000 caractères (chiffres et lettres).

On le charge sous forme de deux tableau à 3 dimensions :

Le premier contient l'index du caractère et les pixels.

Le deuxième contient le caractère écrit en version ordinateur.

Au bout d'environ 3 heures, j'ai réussi à mixer plusieurs codes pour pouvoir charger les données (il faut environ 1 minutes à un ordinateur bien au-dessus de la moyenne pour pouvoir les charger (exécution en single core), et nécessite au minimum 5.5 go de mémoire vive non utilisé).

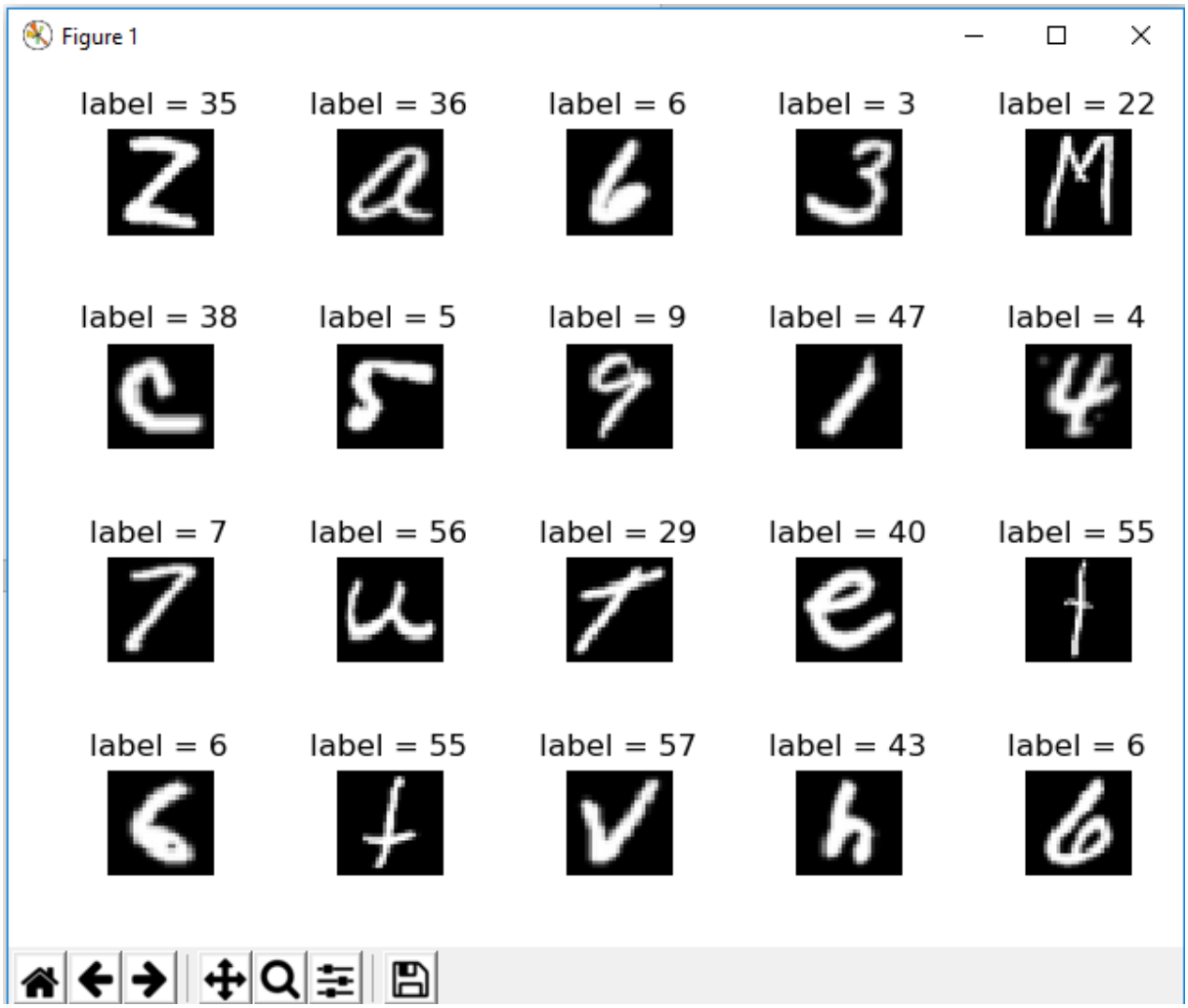
Le code est au final minuscule car il j'ai trouvé une classe (MNIST) déjà toute faite qui s'occupe de tout, sinon le code aurait était très compliqué (manipulation de gros tableaux avec beaucoup de transformations).

J'ai écrit ce tableau alliant label et correspondance

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A
11	B
12	C
13	D
14	E
15	F
16	G
17	H
18	I
19	J
20	K
21	L
22	M
23	N
24	O
25	P
26	Q
27	R
28	S
29	T

30	U
31	V
32	W
33	X
34	Y
35	Z
36	a
37	b
38	c
39	d
40	e
41	f
42	g
43	h
45	i
46	j
47	k
48	l
49	m
50	o
51	p
52	k
53	r
54	s
55	t
56	u
57	v
58	w
59	z
60	y
61	z

Après avoir finalement créer le tableau, j'ai tester, et voici quelques résultats (une infime partie) :



On peut maintenant utiliser ces données pour créer le premier réseau de neurones capable de reconnaitre les caractère .

Information : Le fichier de données est trop lourd pour être portée sur GitHub