MSBD5002 KNOWLEDGE DISCOVERY AND DATA MINING, FALL 2018
ASSIGNMENT 3

Zhe WANG
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
zwangec@connect.ust.hk
STU NO.:20550960

1 FEATURE SELECTION

If we want to collect a list of number from the picture, we have to use CNN, but we didn't have train set for this assignment. So I used Transfer Learning to manage this problem.

With the pre-trained model, we can easily capture the features from the pictures by using tensorflow. In this problem, I used the Resnet model, which is the champion Net in the ImageNet competition in the year 2015.

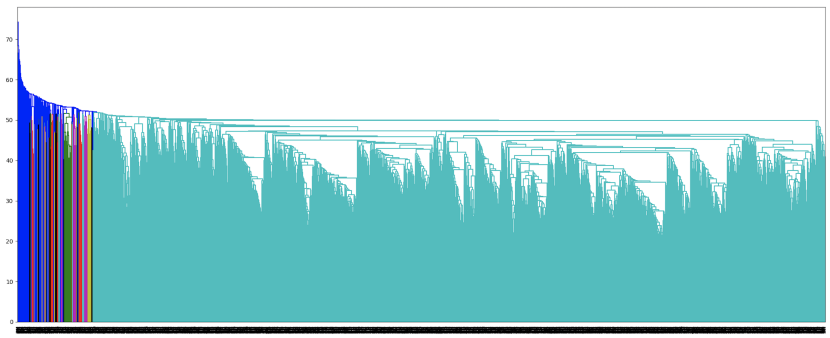
I just resized each picture and put them into the Resnet model, then the Net gave me a 1*2048 dimension feature for each picture. After getting the feature, the problem turned out to be a traditional clustering problem.

The feature is showing as below:

	0	1	2	3	4	5	6	7	8	9	...	2038	2039	2040	2041
images/00000.jpg	0.034156	0.003976	0.668483	0.008688	0.415430	1.047398	0.616220	0.442628	0.575453	1.396779	...	0.146114	0.077588	0.169070	0.007827
images/00001.jpg	0.052353	0.390571	0.000000	0.077936	2.478270	0.298646	0.000790	1.685311	0.000000	1.108239	...	0.196073	0.057446	0.199812	0.000000
images/00002.jpg	0.234393	0.000000	0.162660	0.012463	0.060977	0.099239	1.429605	0.786047	1.174776	0.706574	...	0.206025	0.203504	0.850234	3.374629
images/00003.jpg	0.174163	5.169977	0.201850	0.141075	0.028669	0.000000	0.634073	0.259189	0.006404	0.086920	...	0.743725	0.569133	0.382348	1.948218
images/00004.jpg	0.695237	0.257804	0.071102	0.538792	0.012170	0.315565	0.305934	0.077250	0.450681	0.102810	...	4.197408	0.486907	0.541301	1.700132
images/00005.jpg	0.423428	0.992589	1.623664	1.522760	0.348125	0.697367	0.637526	0.274832	1.236069	5.106035	...	0.581807	2.534561	0.780105	0.991068
images/00006.jpg	0.813323	1.557508	0.518885	0.155341	0.049335	0.352050	1.000118	1.900895	1.308205	0.454984	...	0.248917	1.822501	1.029561	0.384946
images/00007.jpg	2.509067	0.978396	0.169228	0.665242	0.000000	0.457232	0.290595	0.075880	4.313491	0.000000	...	0.000000	2.087213	0.023586	0.291290

2 FIND K

In this problem, the number of clusters K is not given us, so we have to find the K by ourselves, we can use hierarchy clustering to draw a picture and then find the proper K according to the picture's result, the picture is showing as below: The feature is showing as below:

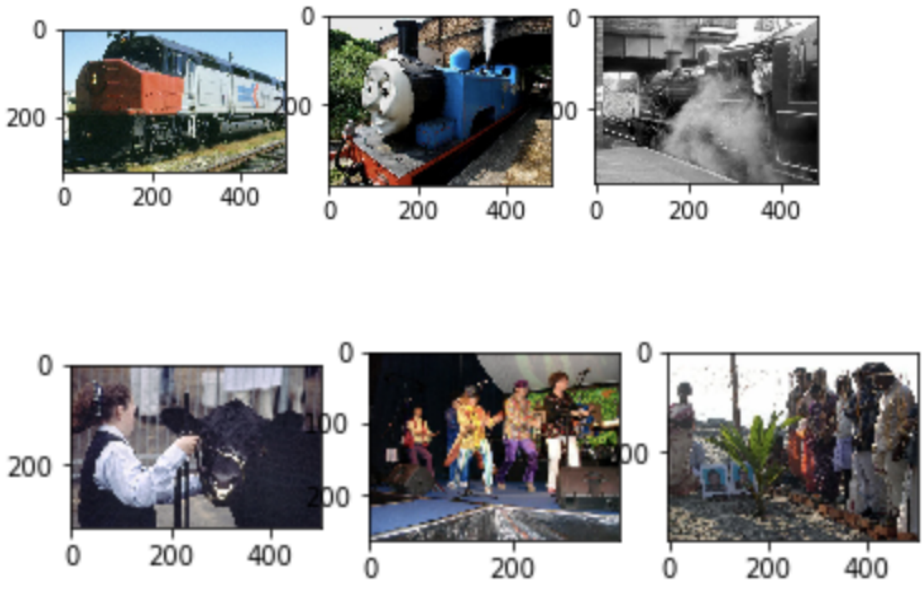


3 CREATE THE CLUSTERS

After drew the picture by using hierarchy clustering, I make the K 18 with a distance threshold 60. Then I tried to cluster the picture by using the hierarchy clustering, but there were a lot of clusters only had one or two pictures, the result is showing like below:

```
[array([ 744, 2398, 2560, 3915]),
 array([838]),
 array([807]),
 array([2008, 3821]),
 array([ 489, 1319, 2760]),
 array([2224, 4611]),
 array([970]),
 array([4594]),
 array([3745]),
 array([4631]),
 array([3580, 4669]),
 array([ 847, 4915]),
 array([ 389, 1503, 3874, 4922]),
 array([2077, 3349]),
 array([2242, 2744]),
 array([1254, 3217]),
 array([2161, 3238]),
 array([ 154, 2182, 2705]),
 array([3149, 3697]),
 array([3892, 4063, 4438]),
 array([2414, 2521, 3163]),
 array([ 921, 1784, 4017, 4335]),
 array([ 584,  716, 1030, 2448, 2586, 2608, 2817, 3210, 3616, 3660, 4638,
        4876]),
 array([ 132,  214, 1074, 1213, 1228, 1453, 1629, 1772, 1891, 2001, 2267,
        2326, 2725, 3212, 3367, 3695, 4430, 4456, 4873, 4914]),
 array([1923, 4044]),
 array([   0,    1,    2, ..., 5008, 5009, 5010]),
 array([870]),
 array([97]),
```

So I changed the method to Kmeans. Continue using the K = 18, I got the final result, there is no single or double pictures in the result. After checking the result, we can find that the train is in the same cluster and the people is in another cluster.



From the result we can judge the result is ideal.

4 CONCLUSION

I used Resnet pre-trained model to create the feature of each picture. Then used hierarchy clustering to find the extra value of K, but the classification result by hierarchy clustering is not ideal. Finally I used the Kmeans with K = 18 to calculate the final result.