Homework 3: CS498 AML (ONL)

Ramya Narayanaswamy (rpn2@illinois.edu)

Description of code : HW3Revised.R

The above code contains R code for part1 and part2 of HW3. The binary data of images is read into the code using readBin. While reading images, they are put into different categories. Every category has 6000 images with 3072 features, stored as a data matrix. Each row of the matrix corresponds to a single image (i.e 6000 rows). Overall, array of dimensions 6000x3072x10 was used to store all images of all 10 categories. Additionally, random images were displayed to test the read function.

Part 1:

For each category, mean of all images was calculated by summing over individual pixel values and dividing by number of images in every category. The mean images are displayed below

1) Airplane

2) Automobile

3) Bird

4) Cat

5) Deer

6) Dog

7) Frog

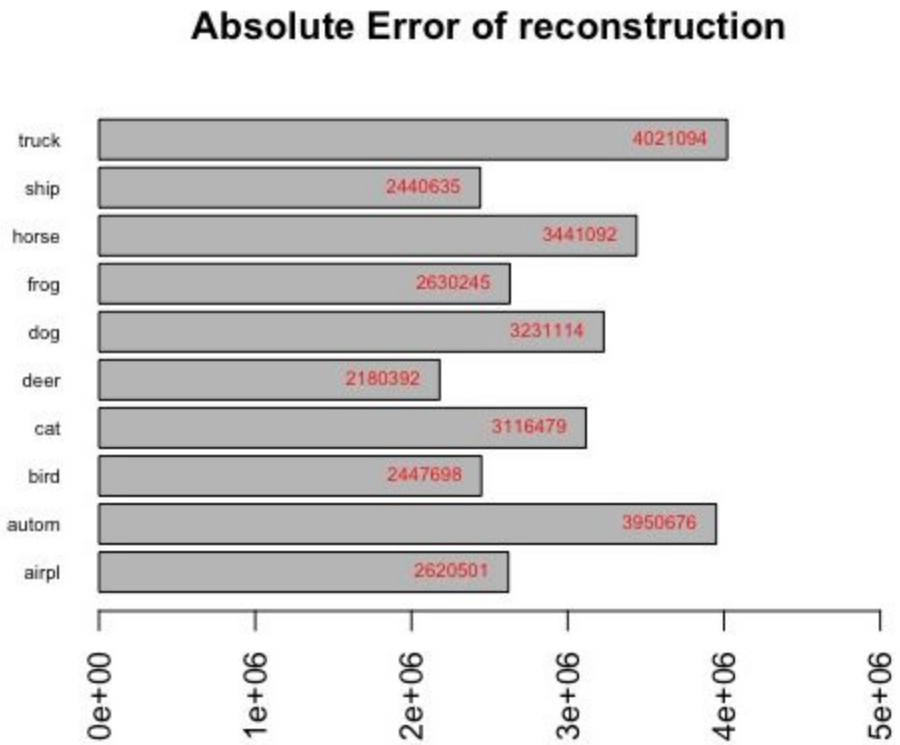
8) Horse

9) Ship

10) Truck

Principal component analysis and reconstruction:

The principal component analysis is performed using “svds()” function from the library “rARPACK” for every category. The images are centered and the number of principal components are limited to 20 using argument k in svds function. The images are reconstructed by matrix multiplication of left singular vector, diagonalized singular value matrix and right singular vector. The run time is less than a few mins for the entire code and using svds with limited principal components increased the performance when compared prcomp and fast.prcomp for PCA. Error for reconstructed image is calculated as sum of squared pixel differences. Absolute average error for each category is provided below. The error values are rounded for clean display purposes



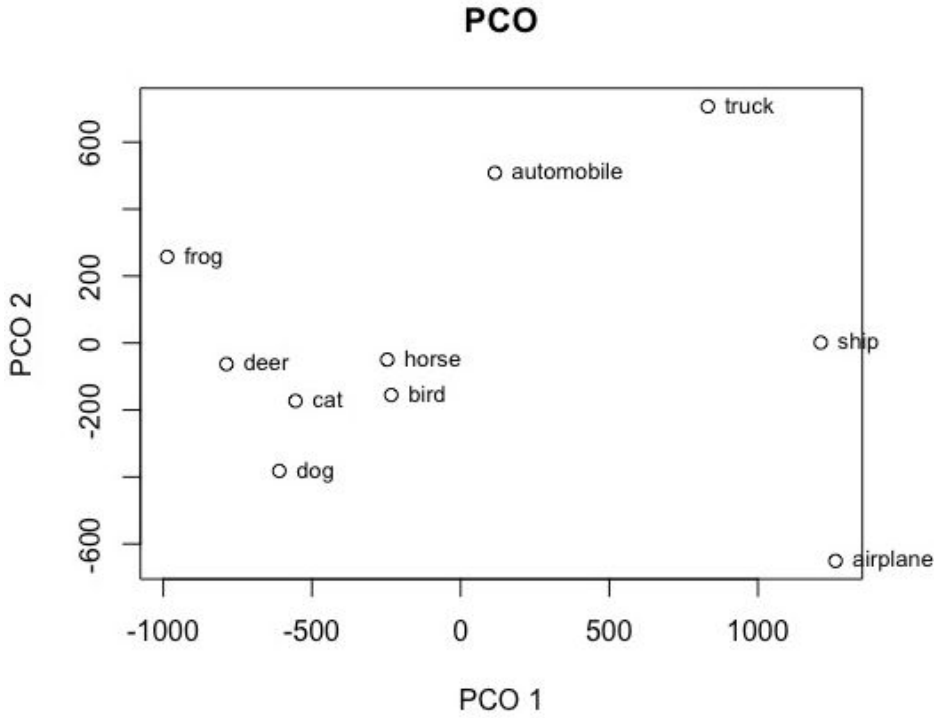
Part 2: Principal coordinate analysis for mean images

The distance matrix for mean images of each category was obtained using dist() function. The argument "method" was set to 'euclidean' to obtain L2 norm of distance. Principal coordinate analysis was performed using "cmdscale()" function. Distance matrix and 2D scatter plot of distances between mean images are provided below

Distance matrix: 10x10 matrix. Row and column order is (airplane,automobile,bird,cat,deer dog,frog,horse,ship,truck)

0	1683.635	1605.024	1905.535	2148.763	1965.221	2445.68	1663.646	945.5411	1449.095
1683.635	0	886.2367	1027.65	1143.081	1216.079	1191.192	950.7861	1303.467	949.9958
1605.024	886.2367	0	517.3115	601.2503	701.4682	913.7475	418.2763	1557.715	1416.675
1905.535	1027.65	517.3115	0	469.7917	412.1817	677.492	596.3767	1851.215	1676.468
2148.763	1143.081	601.2503	469.7917	0	617.6971	460.5109	684.3469	2065.622	1830.741
1965.221	1216.079	701.4682	412.1817	617.6971	0	828.5811	843.6721	1897.592	1880.244
2445.68	1191.192	913.7475	677.492	460.5109	828.5811	0	948.704	2249.2	1913.241
1663.646	950.7861	418.2763	596.3767	684.3469	843.6721	948.704	0	1660.268	1347.334
945.5411	1303.467	1557.715	1851.215	2065.622	1897.592	2249.2	1660.268	0	1066.942
1449.095	949.9958	1416.675	1676.468	1830.741	1880.244	1913.241	1347.334	1066.942	0

Mean Distance Plot:



Part 3: Principal coordinate analysis for the defined similarity matrix

The code is HW3Part3.R

The principal component analysis is performed using “svds()” function from the library “rARPACK” for every category. The images are centered and the number of principal components are limited to 20 using argument k in svds function. Each image is individually resconstructed as per HW instructions. The for-loops (LX’s) is explained below

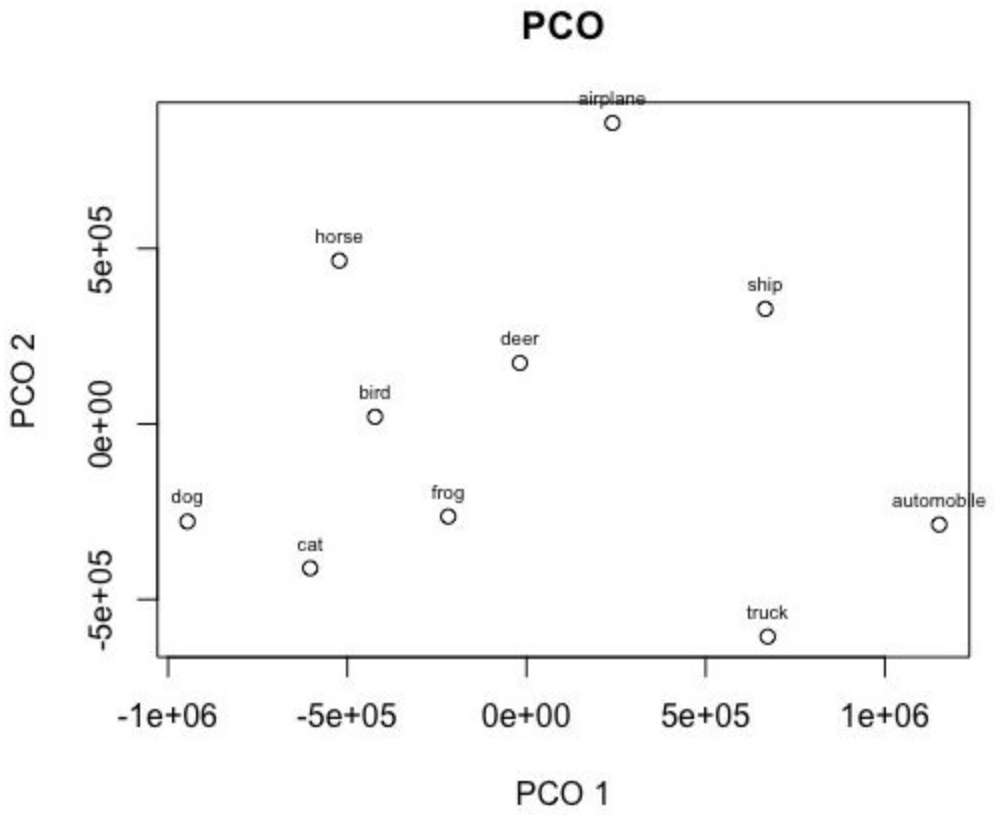
L1: For each category “catg”, corresponding eigenvectors are obtained using svds()
L2: For every category “crossg”, use eigenvectors from L1 and centered image from “crossg”, reconstruct the images. Calculated the image error as sum of squared pixel difference between reconstructed image and original image of “crossg” category.

A 10x10 Error Matrix was used to represent error for using PC’s of A to represent B, where A and B are categories. A similarity matrix is constructed from error matrix and it was verified that the matrix is triangular, with diagonal values matching the absolute error as in Part1. Principle coordinate analysis is performed on similarity matrix using cmdscale() function. The similarity matrix (rounded figures) and a 2-D scatter plot are shown below

Similarity Matrix: 10x10 matrix. Row and column order is (airplane,automobile,bird,cat,deer dog,frog,horse,ship,truck)

A	B	C	D	E	F	G	H	I	J
2620501	3723676	2794365	3306932	2554925	3399592	2950223	3394717	2715857	3794218
3723676	3950676	3699977	4039273	3450549	4211172	3685206	4230300	3489125	4134930
2794365	3699977	2447698	2939377	2423654	2967088	2685785	3199991	2813363	3634776
3306932	4039273	2939377	3116479	2889583	3262798	3028700	3546316	3202341	3897344
2554925	3450549	2423654	2889583	2180391	2954958	2553877	3041230	2543500	3441359
3399592	4211172	2967088	3262798	2954958	3231113	3100441	3610865	3386717	4075472
2950223	3685206	2685785	3028700	2553877	3100441	2630244	3341395	2873799	3648283
3394717	4230300	3199991	3546316	3041230	3610865	3341395	3441091	3386206	4139677
2715857	3489125	2813363	3202341	2543500	3386717	2873799	3386206	2440635	3541624
3794218	4134930	3634776	3897344	3441359	4075472	3648283	4139677	3541624	4021094

Similarity Matrix Plot



Comparison of scatter plots from Part 2 and Part 3:

The magnitude scale of plots from part 2 and part 3 are different because of ranges of values used for plotting. Part3 values are 1000 times more, and the axes absolute magnitude is not of concern. What we are interested is the distance between two categories in each of the plots.

From the mean distance plot of Part 2, it is inferred that mean images of (horse,bird), (cat,bird), (deer,cat) are closer than rest of the categories. (airplane, ship) are close in one axes because of the blue background predominant in sky and water pictures. The above hypothesis could also be explained by looking at the values of the mean distance matrix. If we try to draw imaginary non-spherical clusters on the mean distance plot, One might get (ship,airplane), (horse,bird,cat,deer,dog, frog) and (truck, automobile) as clusters. What we predominantly infer from mean distance plot is how close or far-apart the mean images of every category.

From the similarity or rather dissimilarity plot of Part 3, we try to infer is how big is the error in representing each image with principal components of other categories. For instance, from the plot, we can infer that airplane category when represented using principal components of truck or automobile produces high error values, as airplane location is far away from truck

and automobile location. The error values in the similarity matrix also confirm the above hypothesis.

By comparing Part 2 and Part 3, categories in the imaginary cluster as defined in previous section, when reconstructed using principal components of categories in same cluster, the error is small. However, if categories are reconstructed using principal components of categories outside their imaginary cluster, the error is huge. For example, truck's principle components could not be used for reconstruction of dog and vice-versa as error is huge. However, truck could be reconstructed from automobile's principal components with less errors.

References

1.http://pj.freefaculty.org/blog/?p=122

2. https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca

3. http://www.gastonsanchez.com/visually-enforced/how-to/2013/01/23/MDS-in-R/

4. https://statr.me/2016/02/large-scale-eigen-and-svd-with-rarpack/

5. Piazza posts : @551, @585, @539, @436, @565