CS 498 HW 3 Report

4.10 a

The mean image is computed by adding all the pixels value and divided by the number of images. It should be noted that the original data is ‘uint8’ type. To avoid overflow, the type should be converted to int or float, and return the averaged pixels as ‘uint8’. The mean image for each category is obtained by computing the mean of all images of that category from batch 1 to batch 5. The mean images in plotted in 4.10a.ipynb.

We use sklearn.decomposition.PCA to compute the first 20 principal components. The plot of error resulting from representing the images of each category using the first 20 principal components against the category is shown in Fig 1.

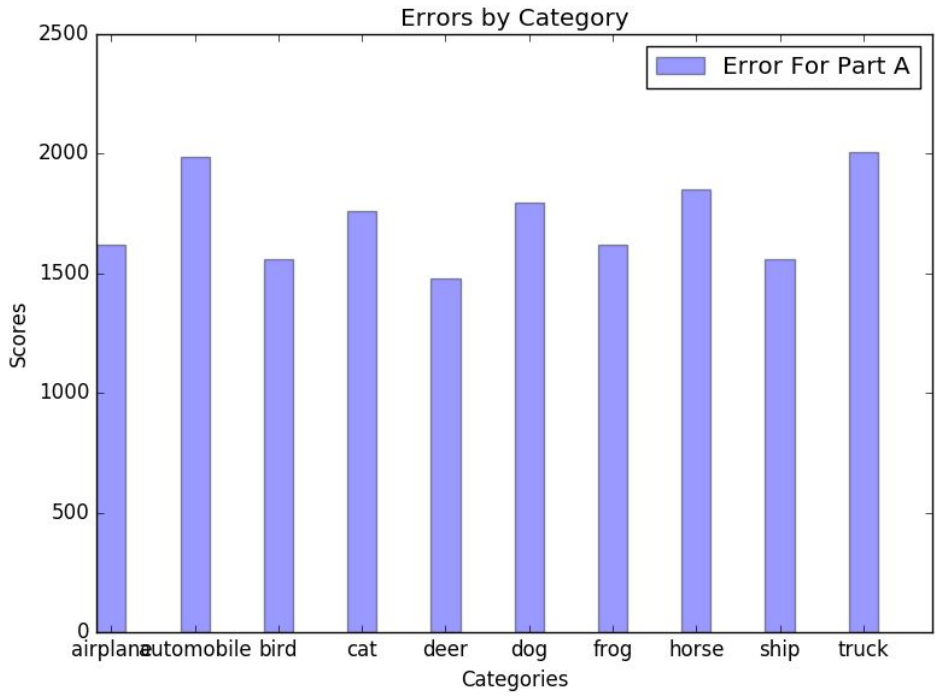

Fig 1.

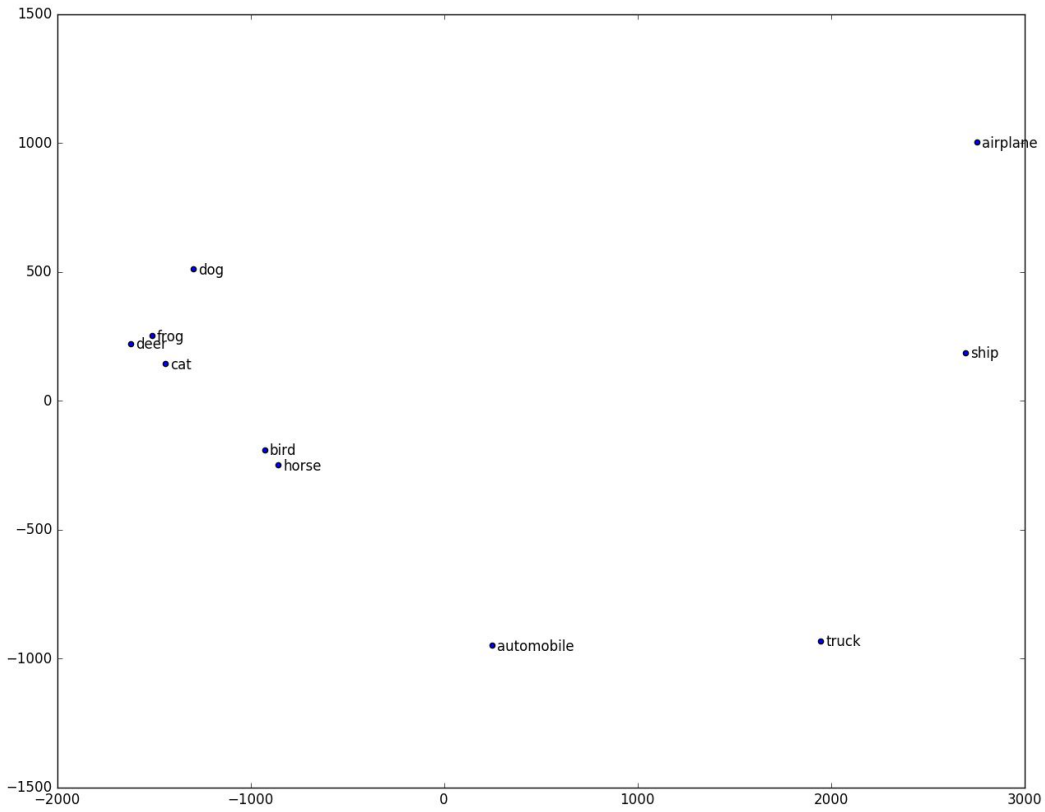

Fig 2.

PCA was applied to a distance matrix constructed using Euclidean distances, the result is plotted in Fig. 2.

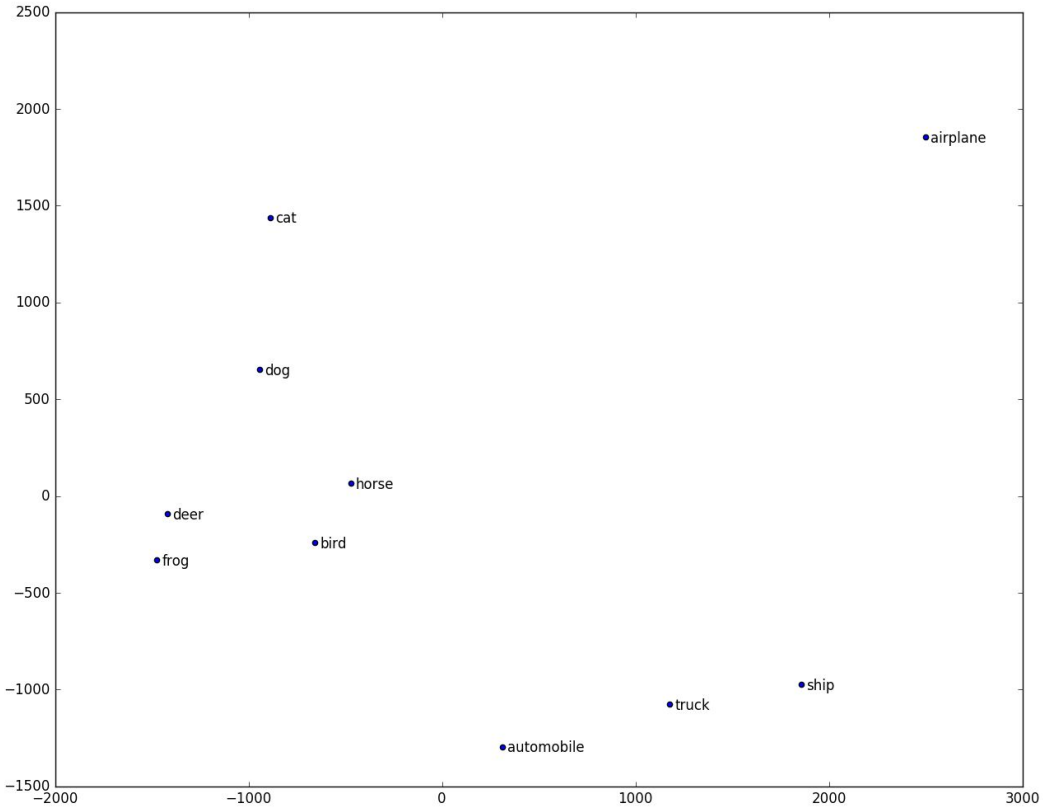

Fig 3

4.10c

PCA was applied to a distance matrix constructed using using a different distance definition, defined as the one half of the sum (the mean) of the average error from A to B and B to A of two given points.

The advantage in fig 3 vs fig 2 is that it's based on a better understanding in distance: instead of pixel wise distance, we find the best measure of distance in some category which can preserve their difference and similarity. Then we summarize this info and extract 2 dimension graph which can represent the similarity and difference to the best extent. So Fig 3 has a better cluster and scatter.

So while distance in Fig. 2 represents actual differences between the means of each category, distances in Fig. 3 should directly correlate to error rates when predicting labels based on those first 20 primary components. In other words, points may be close to one another in Fig. 2 because those averages are visually similar, but they might be further apart in Fig. 3 if the PCA model has few issues actually distinguishing them.

*It should also be noted that Euclidean distance doesn't account for the specific positions of individual pixels, ie the formula wouldn't be able to distinguish between two images with the same colored pixels, but in different positions. This is also probably a contributing factor as to why PCA was more effective than the raw Euclidean distances might have suggested.

. This one is different from the previous one because this one shows the similarity based on the similarity we defined. As we can see the cat and airplane are really hard to confuse with the other ones.