PCA - Result explanation report

Matteo Antonio Senese

November 2018

1 Principal Component Visualization

1.1 PCA for image reconstruction

By applying PCA on the entire dataset the result is a very poor reconstruction. The reason is that the samples images belonging to 4 different classes: guitar, dog, person and house. All these classes share irrelevant similarities and so the PCs introduced wrong information inside the reconstructed image. In particular we can notice that the first PCs bring with them infos about the shape of a head. This happens because of the misproportion among the number of samples belonging to each class, indeed the number of samples of the class *person* represents almost the 40% of the entire dataset. The last 6 PCs are the ones with the lowest variance and in fact they contains very little information for image reconstruction.

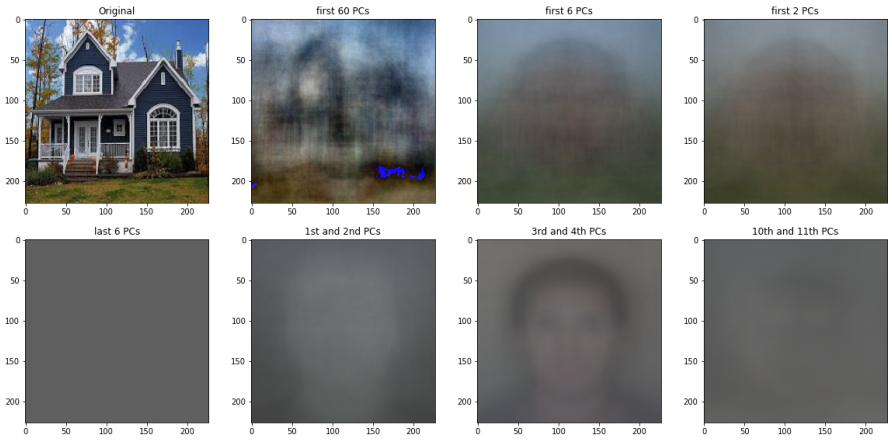

Figure 1: Experiments for *house* image reconstruction

A more reasonable way to reconstruct the original image is to compute PCA within the same class. In some images there are noises that I personally reconduct to some cast from floating point to integer.
Here some results.

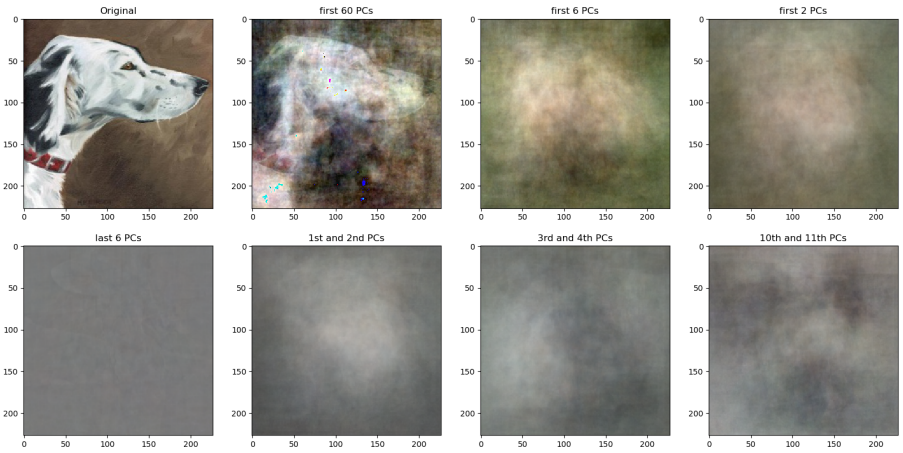

Figure 2: Reconstruction using *dog* class only

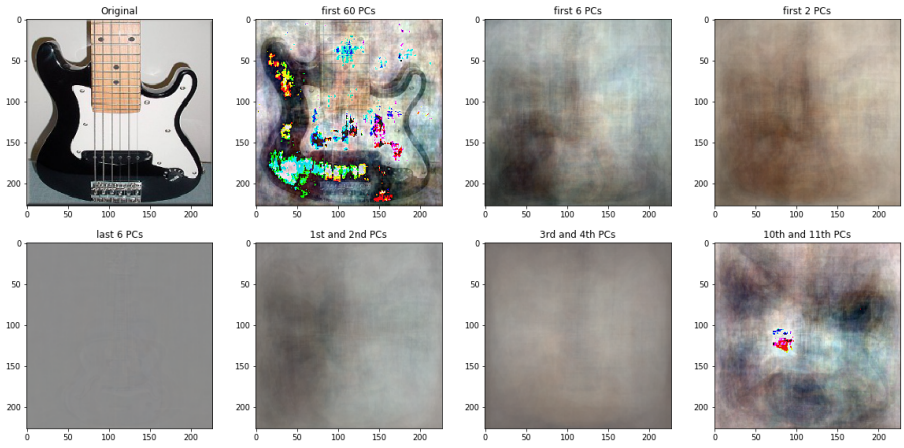

Figure 3: Reconstruction using *guitar* class only

In the following set of images you can notice how the results on the house reconstruction are sensibly changed in the respect of the previous experiment.

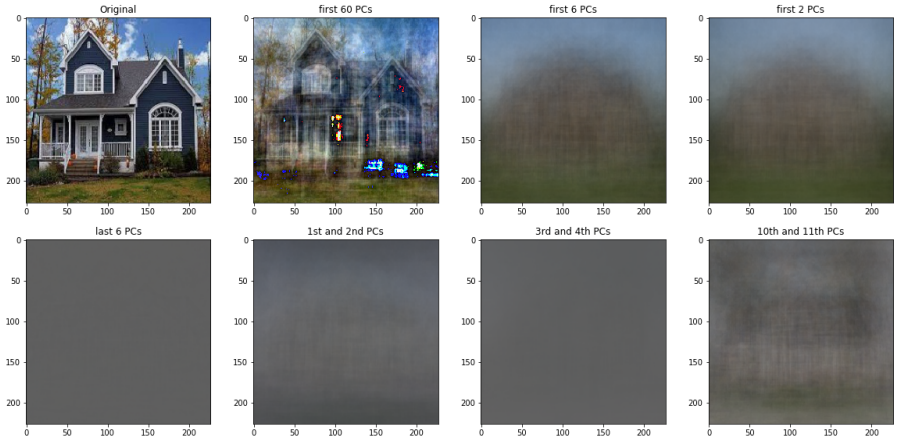

Figure 4: Reconstruction using *house* class only

In particular I want to analyze the PCA on the class *person*. As you can see the 3rd and 4th PCs outline a face with soft features that can be considered belonging to a woman. By analyzing data it can be discovered that the woman represents the 40% of all the samples of the class. Another interesting aspect that deserve to be mentioned is the presence of beard starting from the sixth PC (*first 6 PCs, 10th and 11th PCs*). The percentage of man with a beard is almost 18%.

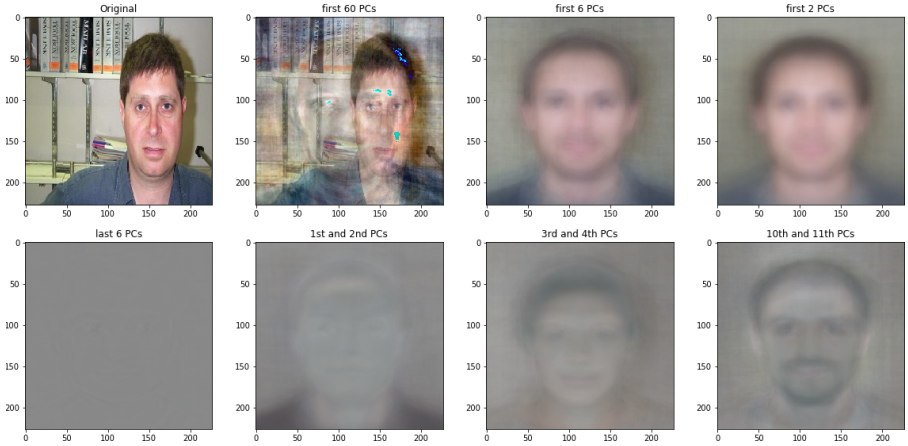

Figure 5: Reconstruction using *person* class only

1.2 How to choose the number of PCs?

There are actually a lot of different techniques to choose the minimum number of PCs needed for a decent image reconstruction. I want to list three of this:

1. *Eigenvalues one criterion:* by following this criterion we will choose the first eigenvectors for which the correspondant eigenvalues is greater than 1

2. *Amount of explained variance:* we choose the number of components which summed up can cover from 70% to 80% of the total amount of variance

3. *Scree plot:* this is a method in which you choose the number of PCs by search for the elbow of the graph

I used the *scree plot* method to choose the number of PCs to choose. On the y axis the cumulative sum of variance of eigenvectors is plotted, while on the x axis the number of principal components is plotted.

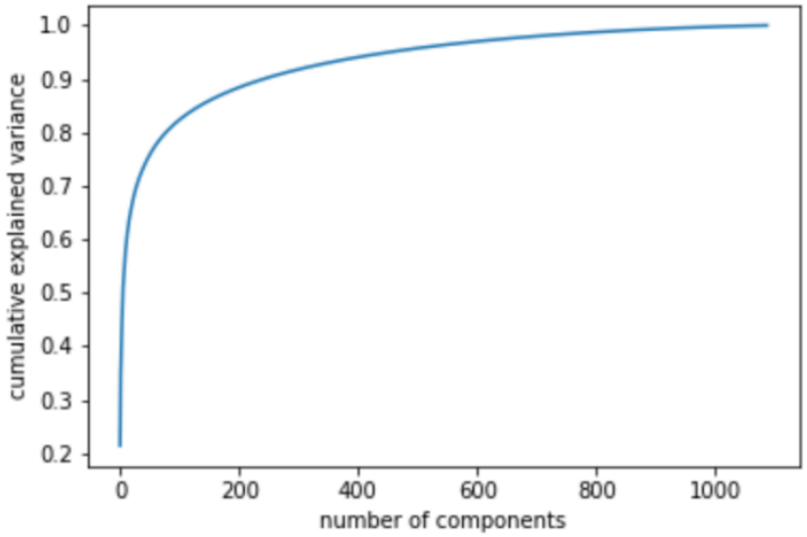

Figure 6: Scree plot

By looking at this plot we notice that the greatest variance contribute is brought by almost the first 200 PCs. So an acceptable number of PCs should be 210-230.

1.3 Scatter plot

The scatter plot represents on the axis the two principal components I choosed. In particular we deduced the greater sparsity of the plot representing the first 2 principal components than the ones representing 10th and 11th principal components. This is caused by the fact that the first PCs are the ones which have the greatest variance and then they sorted in descending order. So the more we move towards the last components, the more we loose variance and then information.

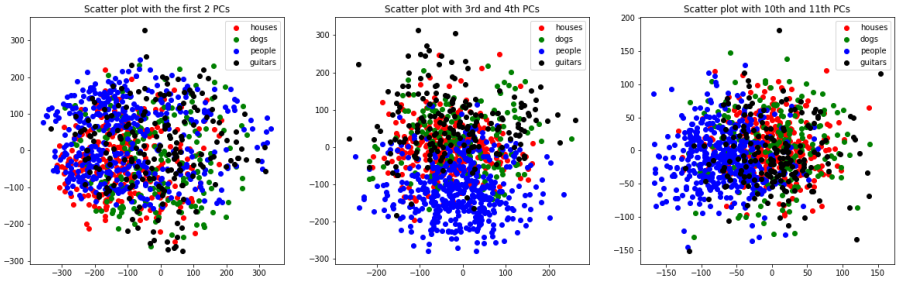

Figure 7: Scatterplots using different PCs

2 Classification

For what regard the classification task I notice a great improvement when standardized data are used. I compute a very simple statistics on 10 experiments per part and I notice that the score of the classifier based on standardized data is $78.39\% \pm 0.59\%$ on the train data and $75.62\% \pm 1.53\%$. Without standardize data the scores are $78.09\% \pm 0.84\%$ for train data and $74.03\% \pm 2.14\%$ on test data.

By appling PCA before the classification the perfomance changes based on the type of PCs we have choosed. If we choose the first PCs we can obtain quite good results (anyway very bad compared to the previous ones), the more you move up by choosing components, the more you obtain very poor results because of those PCs have a very low variance and so they cannot represents well the main characteristics of each class.
Some results:

- *2PCs:* SCORE ON TRAIN SET: 62.64%
- *2PCs:* SCORE ON TEST SET: 63.23%
- *3th-4thPCs:* SCORE ON TRAIN SET: 50.00%
- *3th-4thPCs:* SCORE ON TEST SET: 47.91%