PRINCIPAL COMPONENT ANALYSIS (PCA)

By Ashwin Mittal

CONCEPT

PCA is one of the unsupervised learning techniques used for dimensionality reduction. It computes eigenvectors from the covariance matrix, which I called principal axes, and sorted decreasingly by the eigenvalues called the explained variance percentage. The dataset is then centered and projected to the principal axes, which form principal components (or scores). To reduce the data dimension, I only keep a certain number of principal components n to explain the variance of the original dataset and ignore the rest.

Let's say I have a dataset X_original with m observations and n features. Subtract the mean for each row. I get the centered data X. Then, PCA will compute k eigenvectors for each feature, yielding a matrix V with shape n×k. The PCA projection or score will be given as Z=XV, where the dimension of Z is now m×k.

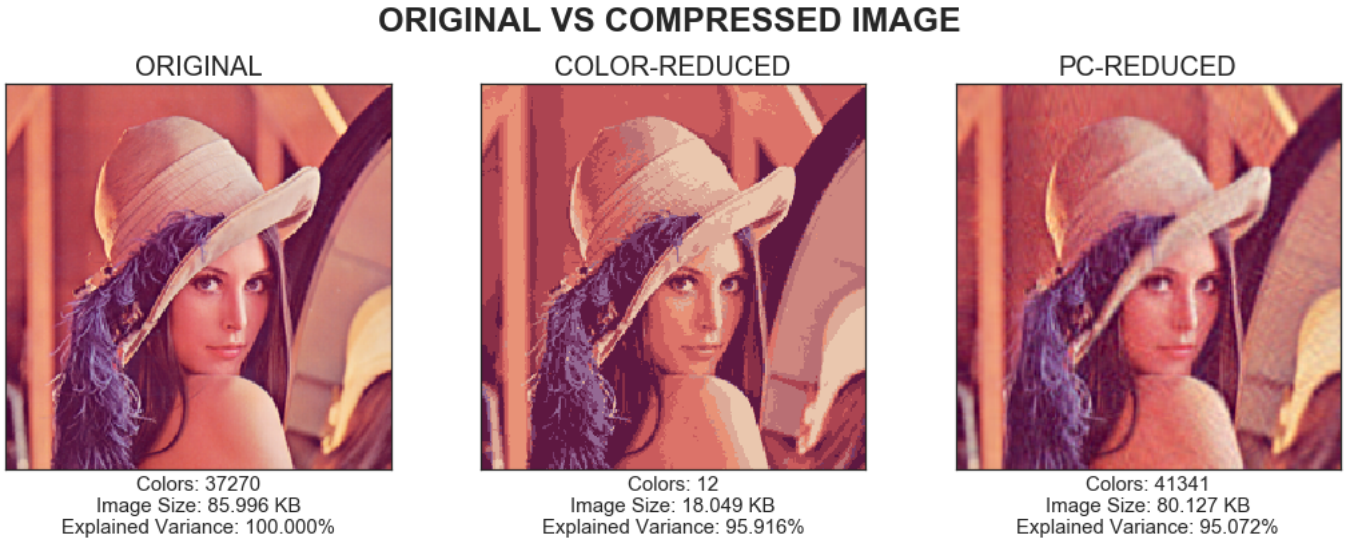
IDEA

I will utilize PCA to reduce the image size by selecting a certain number of principal components n_select only to store the necessary pixels to preserve the original image's variance — making it more efficient in the storage. Our original image consists of three color channels: red, green, and blue. I treat the pixels as a 2D matrix with (height) observations and (width) features for each color channel.

PRINCIPAL COMPONENT OF RGB CHANNEL

PCA is performed on each color channel, resulting in a PCA projection (or scores) and principal components (axes), which both will be in the form of a matrix with the shape height×width.

The visualization of PCs is not informative enough. It's quite random. I should introduce a metric called the percentage of explained variance to evaluate the PC performance. The value ranges from 0 to 100 percent, indicating the similarity between the original image with the compressed image.



CONCLUSION

I successfully perform image compression using the Unsupervised Learning algorithm, such as K-Means Clustering and Dimensionality Reduction using Principal Component Analysis (PCA).

In K-Means, the selection of an optimal number of clusters k is usually made subjectively through visualization.

In PCA, determining the number of PCs used starts from the target explained variance. It also considered reducing image size and number of colors to analyze their similarity with the original image.

Using K-Means, image size reduction reaches 79.012% and can explain 95.916% variance of the original image with only 12 colors. Using PCA, image size reduction is only 6.825% and explains 95,072% variance according to our target. More colors are present in the PC-reduced image than the original image indicating the presence of noise. It can be seen subjectively that the colors in the color-reduced image are coarser than the PC-reduced image. Here image in consideration is **lena.png**.

K-Means is more recommended to reduce image size than PCA, but if you want to keep the original image's overall color, use PCA.

Thank you!