PRINCIPAL COMPONENT ANALYSIS (PCA) FOR IMAGE COMPRESSION AND EIGENVECTORS

YIWEI YU

OCT 2022

PCA (PRINCIPAL COMPONENT ANALYSIS)

- PCA (Principal Component Analysis) which is a popular unsupervised machine learning algorithm primarily used for dimensionality reduction of large dataset. We can use PCA for dimensionality reduction for images as well.

- We will compress image and extract characteristics of Lego bricks.

PREPROCESS ON IMAGE

- Pixels are features

- Convert colorful image to gray image – just deal with brightness on image,

- Image source:
  https://en.wikipedia.org/wiki/File:Lego_Color_Bricks.jpg

- Image shape: (857, 1280) -> (850, 1280): cut few pixels to align size with 10.
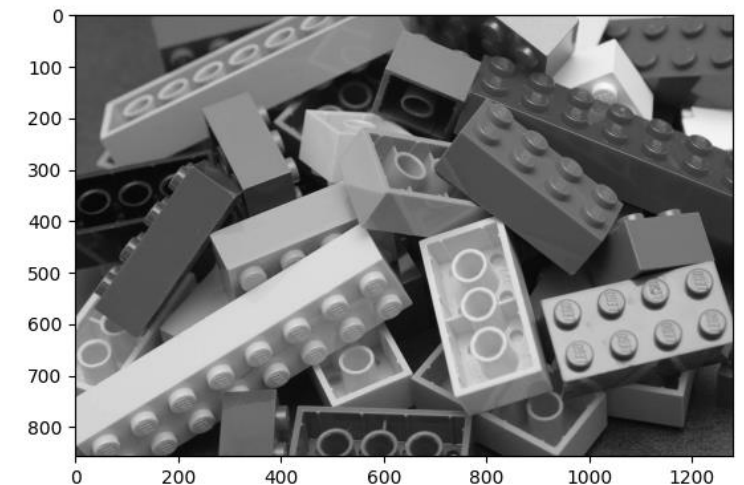
BUILD DATASET FROM SINGLE IMAGE

1	2	3	4	..	97	98	99	100
1	2	3	4	..	97	98	99	100

- Split image into 10x10 pixels sections

- We get dataset which size = 10880 (850x1280 / (10x10))

- Each data has 10x10 = 100 features

- Flatten sections to get array which shape = (10880, 100)

10 pixels

10 pixels
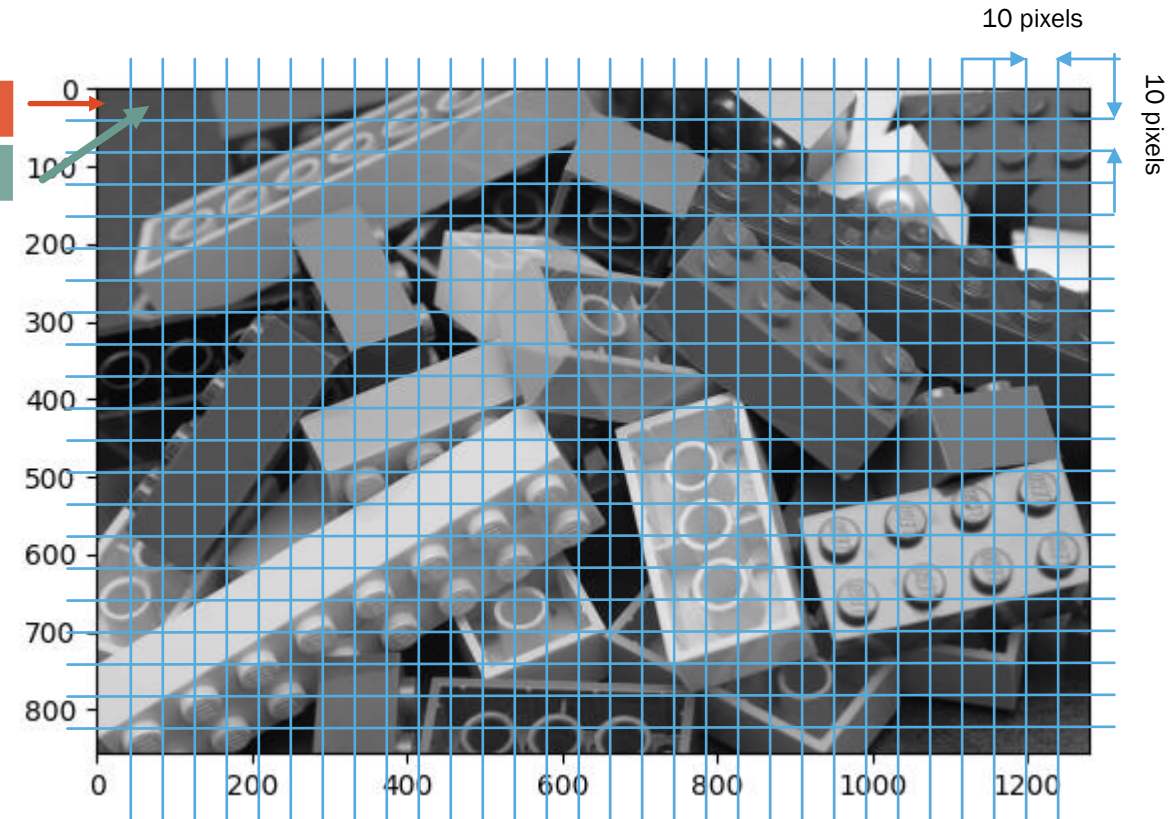
PRINCIPAL COMPONENT ANALYSIS

- If we summary percentage of variance explained by each of the selected components, get about 0.95

- np.sum(pca.explained_variance_ratio_[:2]) = 0.9531883098582244

- That means we can get most of context (95%) of this image just keep 2 components from PCA model. That is the power of PCA to reduce high dimensions.

100 (10x10) -> 4 (2x2)

components = 2 > 95%

100 (10x10) -> 25 (5x5)


100 (10x10) -> 1 (1x1)
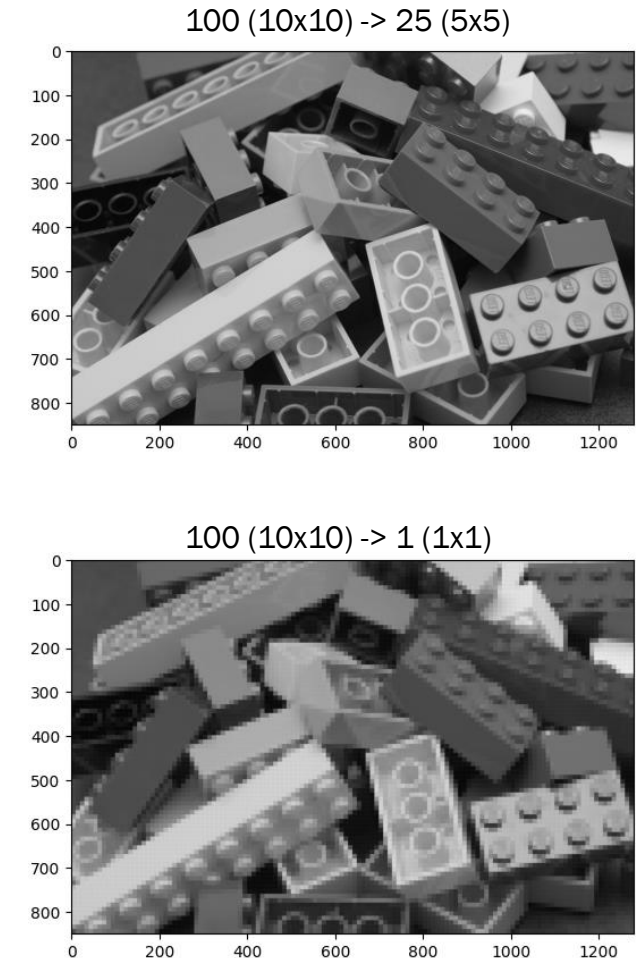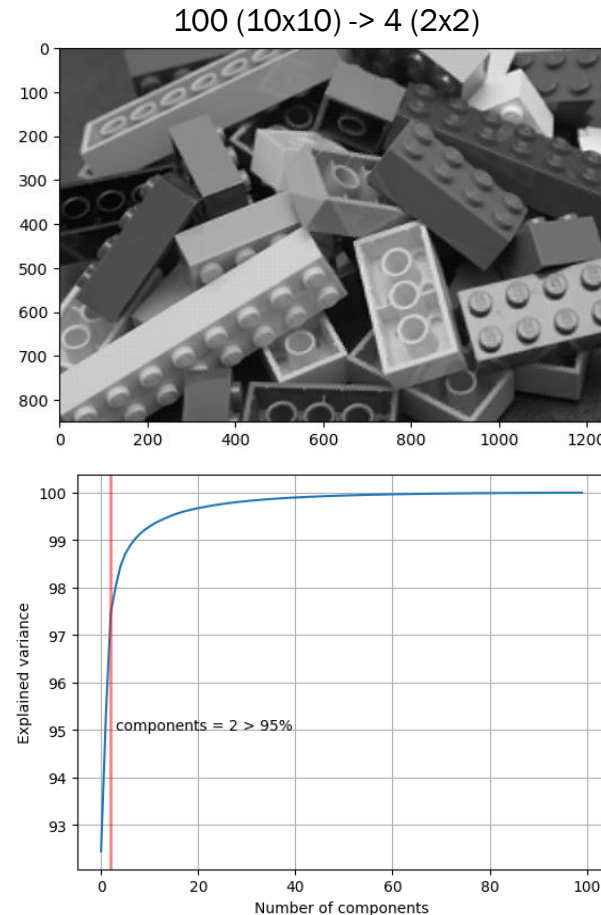
CHECK BY PAIRPLOT

- Let's look at the paired plot and notice that the correlations between variables that were present in the original data disappeared in the principal components.

- Therefore, PCA was able to significantly reduce the correlation. The distribution plot along the diagonal tells us that PCA also successfully transfers the variance related to compressibility.

Pair plot of Image original features
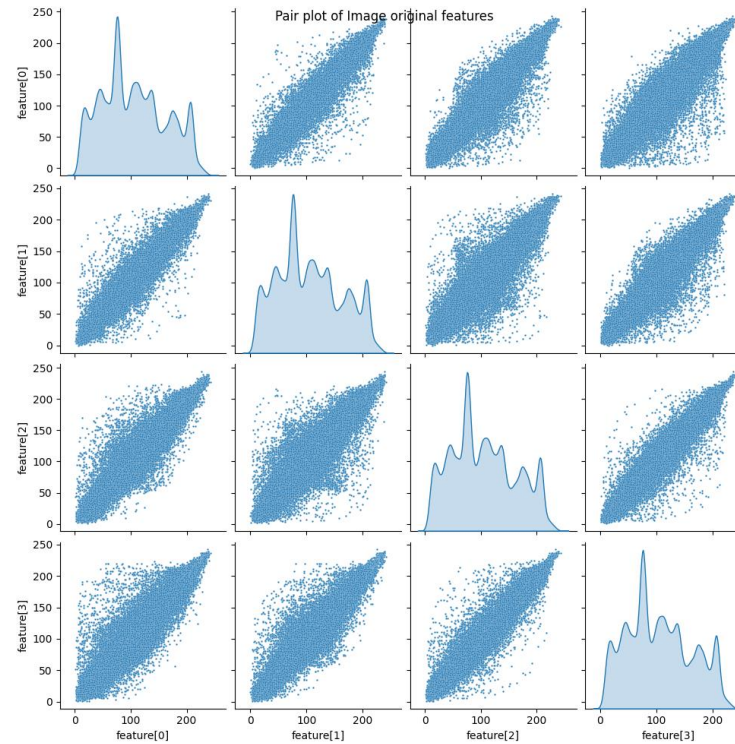
Pair plot of PCA (components=4)
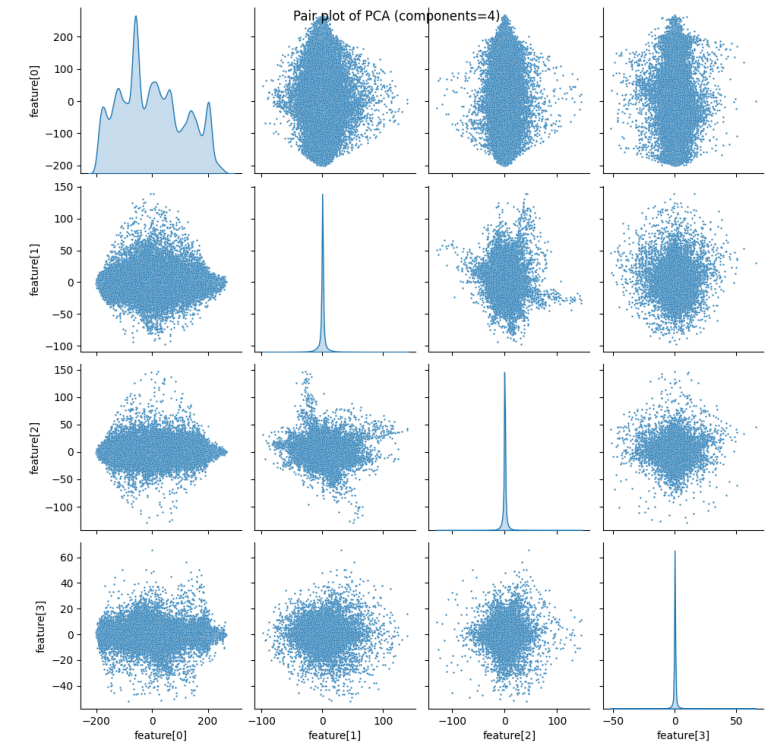
LOGO BRICKS CHARACTERISTICS

- We can find characteristics of Lego bricks from eigen values.

- Pick up some examples to filter characteristics.