CCE, IISc. 2022

Course name : AIML 2022 Project Report

Title: Image compression using SVD, PCA, K-mean algorithm

Submitted by

Parinita Bora
Srishti Singh

Date of Submission: 29/11/2022

1.1 Data compression in machine learning application

For training a machine learning model when there is large amount of unlabelled data, several unsupervised learning algorithms can helps in the understanding of the data.

- Unsupervised learning also can help in dimensionality reduction.
- Dimensionality reduction again can help in data visualization
- When the data is reduced, the complexity of the model can be reduced, so as the training time.

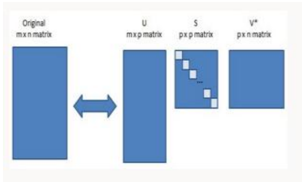
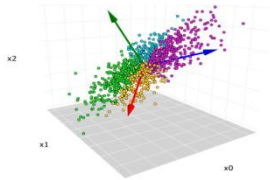
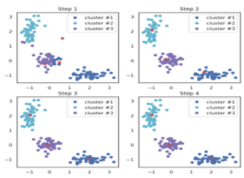
1.2 The Techniques, data and scope

Three unsupervised algorithms namely Singular value decomposition (SDV), Principal component analysis (PCA) and K-mean are experimented as a part of this work.
The algorithms are applied as a part of pre-processing of task, with goal for experimental study on data reduction or compression for high resolution image.

The data file is with dimension 570X 985 x 3 , image of Cosmic object, Captured by James Webb Space Telescope (publicly available in NASA website)

Figure 1 Sample data file of image dimension 570X 985 x 3

1.2 A Brief overview of the methods

Method	Inventor	Purpose	General overview
Singular Value Decomposition(SVD) - https://en.wikipedia.org/wiki/Singular_value_decomposition	Independently Eugenio Beltrami, Camille Jordon over 100 yrs back	To predict a set of optimal factors.	
Principal comonent Analysis(PCA) - https://en.wikipedia.org/wiki/Principal_component_analysis reduction	Karl Pearson in 1901, later in 1930, developped by Harold Hotelling |	Dimnetionality reduction	
K-Means clustering - https://en.wikipedia.org/wiki/K-means_clustering	First used by James MacQueen in 1967 ,used by Steinhaus in 1956	In pulse code modulation(by Steinhaus)	

The Advantages:

1. **SVD** : SDV simplifies data, can remove noise also it can be used for coloured image to segregation components for computational efficiency

2. **PCA** : Dimensionality reduction is the biggest advantage preserving most significant data. PCA can also be used in data exploratory analysis and visualization

3. **K-Mean**: Simplicity and guarantees convergence. It provides good representation of reduced features/ data.

1.3 The implementation and results

The algorithms are implemented in python. The Python libraries scikitlearn, matplotlib libraries are used for visualizations. The details about the project is stored in the readme.md file in Github repository https://github.coom/Gitpabora/Data_reduction_compression.The source code and results are shared in google colab bnotepads links in table 2.

Table 1 Detailed implemenatation

PCA	https://drive.google.com/file/d/1_pBJL6v9sRRetdD0tLqvmihOVvtvivf8/view?usp=share_link
SVD	https://colab.research.google.com/drive/1eG843MHVTwohPAqRmsQa8JToxPNJZR1M?usp=share_link
K-Mean	https://drive.google.com/file/d/1VFxHAb34riaiYDiqaN0uYt8Jw4hqJbUk/view?usp=sharing

Table 2 The Algorithms flow

PCA	SVD	K-Mean
Step1. Calculate the covariance matrix of the data **step2.** Extract the eigenvectors and the eigenvalues of that matrix **Step3**. Select the number of desired dimensions and filter the eigenvectors to match it, sorting them by their associated eigenvalue **Step4**. Multiply the original space by the feature vector generated in the previous step.	**Step1**. getting three component matrices with Red , Blue and green constituents **Step2**. Applying SVD on each of the three components to generate three vectors for each of the matrices **Step3**. Preserving only K i.e. Selecting k columns from U matrix and k rows from VT matrix, and resetting rest to zero **Step4**. Reconstructing the coloured components from U and V **Step5**. Final image is formed by oncatenating the three components	**Step 1**. An optimal number of clusters (K) is chosen. **Step 2**. k number of points "centroids" are initialized randomly within the data area. **Step 3**. Each data or observation is attributed to own closest centroid. **Step 4.** Updating is done for the centroids to hold the value corresponding to the centre of its all attributed observations. **Step 5**. Steps 3-4 are repeated a number of times / until all of the centroids are prominent.

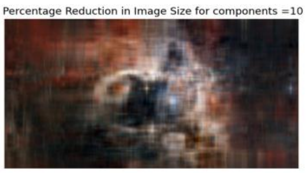
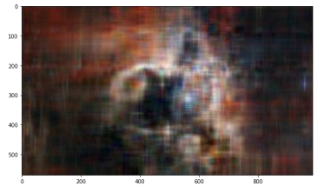
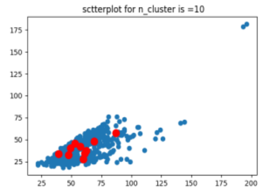
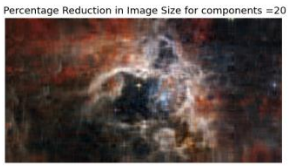
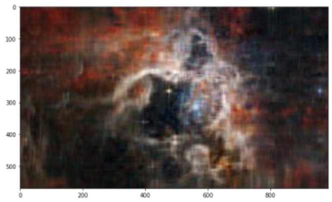
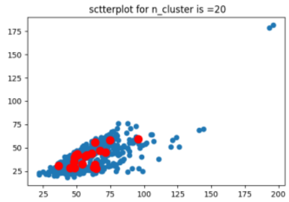
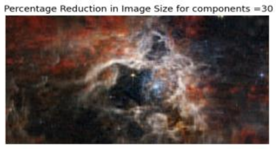
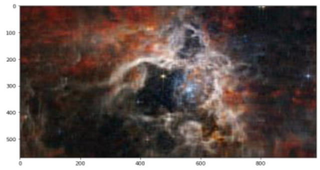
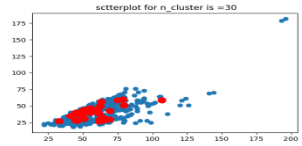
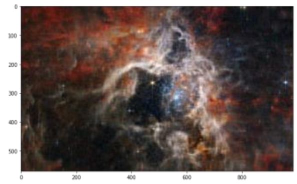
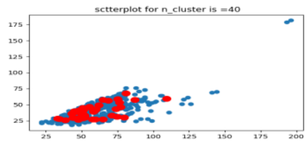
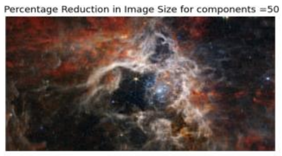
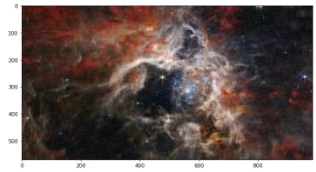
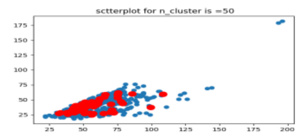
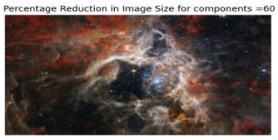
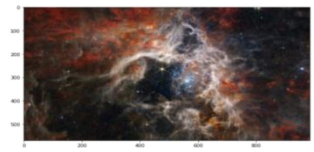
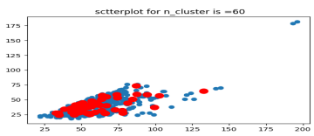
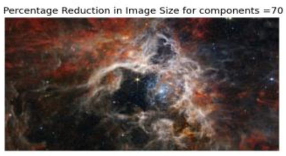
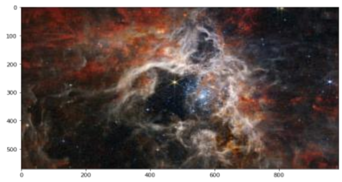
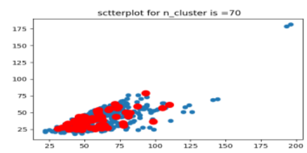
The measurement for compression or data reduction:

- The compression ratio is calculated using the below formula:
  Compression ratio = ((original_number_of_image_element -new_number_of_values after applying the algorithm)/original_number_of_image element)*100.
- The same is experimented for varying parameters like number of principal components in case of PCA, number of component selected in case of SVD and number of clusters in case of K-mean respectively.

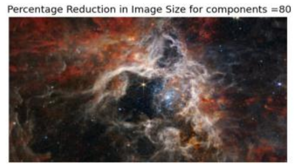
Table 4 The results & Observations (Note:all numeric results are rounded to 2decimal places)

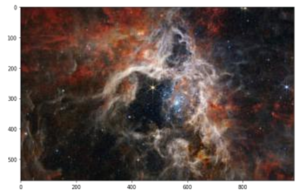
#components(Principal component) /component SVD/ cluster for K-mean	Compression ratio (%) PCA	Compression ratio (%) SVD	Compression ratio (%) K-Mean
10	99.08	97.23	98.25
20	98.15	94.46	96.49
30	97.23	91.69	94.74
40	96.30	88.91	92.98
50	95.38	86.14	91.23
60	94.46	83.37	89.47
70	93.53	80.60	87.72
80	92.61	77.83	85.96
90	91.69	75.06	84.21
100	90.76	72.29	82.46

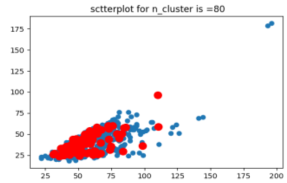
Table 5 Reconstructed Images for PCA & SVD , Kmean clustering

#components(Principal component) /component SVD/ cluster for K-mean	PCA reconstructed image	Reconstructed Image after SVD	K-mean Scatter plot
10			
20			
30			
40			
50			
60			
70			

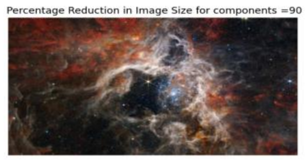
80

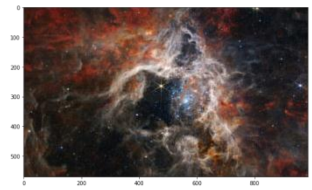


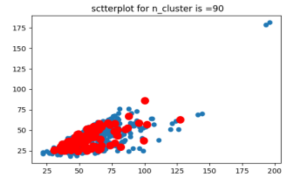




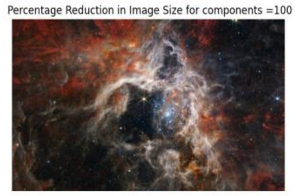
90

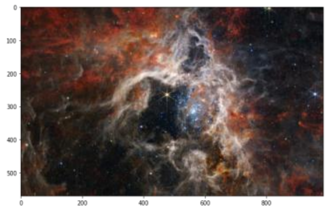


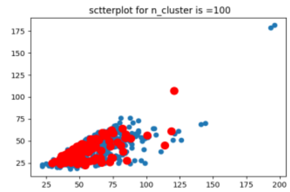




100







Observations:

1. Note: The image for K mean clustering is placed only showing the cluster formation, not comparable in terms of reconstruction.
2. In both the algorithms for PCA and SVD as the Number of principal component or K the compression ratio decreases.
2. Reconstruction for PCA is better at a lower value of number of principal components
3. The compression ratio higher in PCA for the same value of component in PCA and K value in SVD

2.1 Future scope:

1. Exploring other data reduction techniques for Machine learning.

2. Most importantly
(a) Experimenting with large dataset and setting up github CI
(b) test for the measures of these algorithms in terms of the impact on the model performance (c) when which algorithm is suitable.

The applicability which algorithm is most appropriate can only be experimented after evaluating accuracy of the model for the pre-treated data by these algorithms

Other References: 1) https://arxiv.org/pdf/1608.05148.pdf
2) https://bair.berkeley.edu/blog/2019/09/19/bit-swap/

Omar H.D et.al. (2020). Algeria Image Compression using PCA, 2021, 1–11. 2020 International Conference on Mathematics and Information Technology
Blog https://towardsdatascience.com/pca-102-should-you-use-pca-how-many-components-to-use-how-to-interpret-them-da0c8e3b11f0
https://courses.grainger.illinois.edu/cs357/fa2020/assets/lectures/complete-slides/18-PCA.pdf

https://scikit-learn.org/stable/modules/clustering.html
James Fowler, , Qian Du , IEEE member , Hyper spectral Image Compression using JPEG2000 and PCA
Stewart, W. G. (1993)On the early History of the Singular Value Decomposition, SIAM Review, Volume 35, Issue 4. Dec, 1993 551-556 , JTHOR, available at
https://www.math.ucdavis.edu/~saito/courses/229A/stewart-svd.pdf
Toderici, G. (2017), Full Resolution Image Compression wit Recurrent Neural Network ,
https://arxiv.org/pdf/1608.05148.pdf
Vieira, M.N.C.S Vasco (2012)Permutation tests to estimate significance on Principal component analysis ., Technical university of Lisbon , available at
https://www.researchgate.net/publication/255728363