Analyzing and studying Dimensionality Reduction Techniques for High-Dimensional Data

RutujaP. Deshmukh[1], ShwetaA. Jagdale[2], Satish S. Banait[3], Akash R. Kadlag[4], Pranjali A. Baviskar[5]
[1,2,4,5] *Student, Department of Computer Engineering*
[3]*Research Guide, Department of Computer Engineering*

***Abstract*—In the fast-moving world, data is accumulating at an unprecedented speed from vivid sectors across the sphere such as micro-array gene expression data, medical data, ECG and MEG data research, satellite images, IoT devices, etc. is considered as high dimensional data. This data has a lot of features and thus directly affects the output of machine learning algorithms at an exponential rate. Thus, dimensionality reduction (DR) helps to solve the problem of the curse of dimensionality by extracting the relevant features without forfeiting the useful data. The purpose of this research is to compare and analyze different dimensionality reduction techniques namely Principal Component Analysis (PCA), Independent Component Analysis (ICA), Singular Value Decomposition (SVD), Truncated-SVD and Non-negative Matrix Factorization (NMF) on Imagenet dataset (unsupervised dataset) for five different values of components - 40, 45, 50, 55 and 60 each. These algorithms are examined on the basis of execution time, accuracy of dimensionality reduction techniques and load analysis, that is, Mean Squared Error (MSE). The algorithm with the least execution time and number of components giving the most information is concluded as a suitable algorithm for dimensionally reducing high-dimensional data.**

***Index Terms*—Big Data, Dimensionality Reduction, High-dimensional data, ICA, Mean-Squared Error, NMF, PCA, SVD, Truncated-SVD**

I. INTRODUCTION

In the current world, statisticians have mentioned the fact that by 2025, 465 exabytes of data will be generated each day. This data is very large, complex, high dimensional and prominently exists in unstructured form. Such type of data is not suitable as it affects the processing speed and conquers storage space. To handle this, there is an urgent need to find effective ways for storing and retrieving data by preserving the salient features of the data.

This data contains useful information along with some unwanted data such as noise, incomplete, redundant data, etc. If such high number of input features are passed to ML algorithms, it affects the performance – processing speed and increases the complexity of the model which is also referred to as the curse of dimensionality[6]. Dimensionality reduction is necessary to select or extract features from the present feature space of the data which are pertinent to our domain of application[23].

The main aim of this paper is to take high-dimensional images as inputand dimensionally reduce them using five different DR techniques by varying the number of components to the values 40, 45, 50 55 and 60. The dataset used is unsupervised and hence the techniques used for dimensionality reduction are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Singular Value Decomposition (SVD), Truncated-SVD and Non-negative Matrix Factorization (NMF). Time analysis measures the time taken by a technique to transform the data and project it to a lower-dimensional subspace. This can also be interpreted as the load handling capacity of a DR technique. Accuracy analysis gives a concise summary about the information retention after applying dimensionality reduction. This gives an inference about the error handling capacity of a DR technique. The major observations and results are as follows:

- Presenting a detailed and structured survey ofliterature and background relating to existing unsupervised dimensionality reduction techniques.
- Analyzing the performance of PCA, ICA, SVD, Truncated-SVD and NMF with respect to MSE score, fit and transform time and total execution time.

- Propounding each dimensionality reduction technique by varying the number of components and analyzing the error rate after DR.
- Proving that dimensionality reduction on high-dimensional data preserves maximum information in order to upgrade the performance of ML algorithms especially in terms of time and memory space.

The rest of the paper is organized as follows - Section II gives the literature review. Section III describes the proposed methodology for performing the analysis. Section IV gives a detailed analysis of the results of all DR techniques. Section V concludes the entire results and mentions the future scope.

II. LITERATURE REVIEW

Neelam Agarwal, et al. [1], this paper concludes that classification accuracy is inversely proportional to dimensions. PCA outperforms all the methods with an accuracy of 95% for the first two principal components. Thus, proper selection of the number of components is significant in the spectral-spatial feature classification process for the quality and integrity of the data set.

Raji Ramachandran et al. [3], proposed step by step explanation of how dimensions are reduced using supervised approach (i.e., LDA) and unsupervised approach (i.e., PCA, KPCA, SVD, ICA). This paper also proposes comparison of each technique with different parameters which are complexity of execution, efficiency, ability of handling the linear and non-linear dataset. In comparison, it can be concluded that PCA, LDA, ICA, SVD have linear structure while KPCA have non-linear structure. Error handling in PCA and kPCA is high while in others error handling is low.

Wei Wei, Member, IEEE et al. [6], in their work has enabled users to view different views of data distribution information. This paper proposes a dimensionality reduction method for fusing multiple clustering results obtained by K-means algorithm. Thus, this method achieves a significant improvement over some representative and current unsupervised dimensionality reduction methods.

Yousef Jaradat et al. [15] proposed a detailed study on SVD and mathematical overview of the same. Singular Value Decomposition provides a base for all data analysis techniques, for example, Image compression and dimension reduction. The proposed paper states that dimensionality reduction of images using SVD assures guaranteed results.

Shereena V. B. I et al. [20], the main concept of PCA is to transform the high dimensional input space onto the feature space where the maximal variance is displayed. The performance of these techniques has been evaluated by precision and recall measures. This paper concludes that PCA gives better results than LDA when compared in terms of recall and precision.

Analysis of Dimensionality reduction techniques on Big Data [24], the paper concludes that ML algorithms with PCA produce better results when dimensionality of the datasets is high and ML algorithms without dimensionality reduction yields better results when dimensionality is low. The accuracy of the model is calculated on the basis of precision, recall, F1-score, sensitivity, specificity.

III. PROPOSED METHODOLOGY

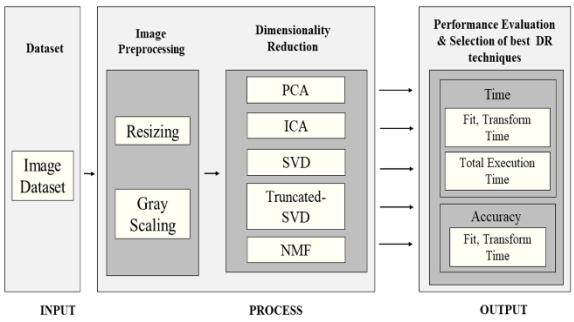


Fig. 1. Architecture Diagram

The basic idea behind this research is to apply and analyze dimensionality reduction techniques - PCA, ICA, SVD, Truncated-SVD and NMF on preprocessed images. After this, the output will be analyzed on the basis of fit-time, transform-time, total-execution time and Mean-Squared Error (MSE). Fig 1 shows the proposed model for the analysis.

A. INPUT

The dataset given as input to the system is the Imagenet dataset which is unsupervised and the size of the dataset is 6.71 GB. The dataset is split into 89% training and 11% testing images.

B. PROCESS

This step consists of two sub-steps: Preprocessing and Dimensionality reduction. For preprocessing, resizing is a critical step, as smaller sized images can train machine learning models faster. Hence, each image is resized to 256 * 256 dimensions. The

dataset contains images in RGB format which implies that the images have three channels - Red, Green and Blue. This data can be reduced by converting the images into gray scale format - shades of black and white and thus reducing the number of channels from 3 to 1. This helps in performing algorithms in a shorter time as it passes only a single channel instead of three channels of a particular image.

The second sub-step is to apply DR technique on these preprocessed images for five values of components - 40, 45, 50, 55 and 60. These values are selected on the basis of a Scree Plot which is plotted for a few images in the data. From the Scree plot shown below, we can infer that considering the 40 to 60 range of components is capable of retrieving maximum useful information of an image.

1. PRINCIPAL COMPONENT ANALYSIS (PCA):

PCA is used to compute the principal components in an image and apply changes on the data and selecting the first few components which preserve the maximum information in an image [20]. In mathematical terms, the variance-covariance shape of an image for variables is explained using linear combinations of these variables.

2. INDEPENDENT COMPONENT ANALYSIS (ICA):

This algorithm is generally used to separate multivariate signals into additive components. In mathematical terms, the independent mean of data of x does not provide sufficient information about y and vice versa.

3. SINGULARVALUEDECOMPOSITION (SVD):

SVD guarantees good results after applying dimensionality reduction [15]. In mathematical terms, it is a method that generalizes the Eigen decomposition of a (m*n) matrix to a left triangular unitary matrix (m*m), diagonal matrix (m*n) and a right triangular unitary matrix (n*n). Splitting the original matrix into three component matrices helps in extracting relevant and interesting information from an image.

4. TRUNCATED-SVD:

Truncated SVD is applied on preprocessed images. Truncated SVD belongs to the Scikit-learn module of Python. This algorithm factorizes a matrix, wherein the number of columns is equal to truncation. It is efficient for sparse matrices as they are not data centric estimators.

5. NON-NEGATIVE MATRIX FACTORIZATION (NMF):

NMF effectively projects high-dimensional image data to low-dimensional spaces by splitting the features of an image into independent features, which if added will reinvigorate the original image.

C. OUTPUT

For PCA, ICA, Truncated-SVD and NMF, the fit-transform-time and the total execution time for each of the five components is recorded for further analysis. For SVD, total time of execution is calculated for each value of component. Along with time, MSE is taken as a parameter for judging the accuracy of each algorithm. It is calculated for every component and a box plot is plotted for analysis by taking the natural logarithm of MSE.

IV. RESULTS AND DISCUSSIONS

This experimentation is performed on the Imagenet Dataset downloaded from Kaggle. A laptop with Windows 10 Operating System powered by GPU Processor of NVIDIA, Radeon, etc. and minimum RAM of 4 GB with an i5 Processor is required.

1. DATASET DESCRIPTION:

The size of Imagenet dataset is 6.71 GB. The dataset has a total of 50,000 images which are divided into two parts for training and testing. 89% of the dataset is divided for training (45,000 images) and 11% for testing (5,000 images). The dataset is unlabeled and has 1,000 classes in total. Fig. 2. Shown below is an image from the dataset.

Fig. 2. Sample Image from the dataset

2. DESCRIPTION OF EVALUATION METRICS:

The performance evaluation of dimensionality reduction is done on the basis of two aspects: time and accuracy.

Fit time means time required to perform calculations on the values of the input data. Transform time is the time required to apply the results of the fit method to every data point. There exists a combination of these two methods which is called as fit-transform. The time required to perform this operation i.e., fit and transform on input data at the same timealong with conversion of the data points is fit-transform time. After performing fit-transform operation on the dataset for dimensionality reduction, the total time required is calculated and the results are analyzed below.

3. PERFORMANCE EVALUATION OF DR BASED ON TIME:

For fit-transform time, Fig. 3. shows that Truncated-SVD and PCA takes the least time, 2.75 seconds and 3.57 secondsrespectively,followed by ICA and NMF which takes 48.90 seconds and 240.71 seconds.For fit-transform time, Fig. 3. shows that Truncated-SVD and PCA takes the least time, 2.75 seconds and 3.57 seconds respectively,followed by ICA and NMF which takes 48.90 seconds and 240.71 seconds.

For total execution time, Fig. 4. shows thatTruncated-SVD and PCA takes the least time, 4.75 seconds and 5.60 seconds respectively, followed by SVD, ICA and NMF which takes 7.13 seconds, 244.21 seconds and seconds.

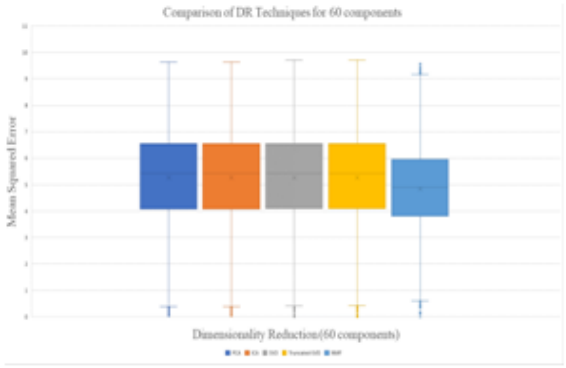
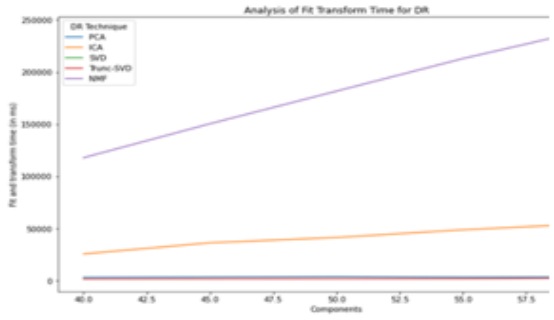


Fig. 3. Analysis of Fit-transform time



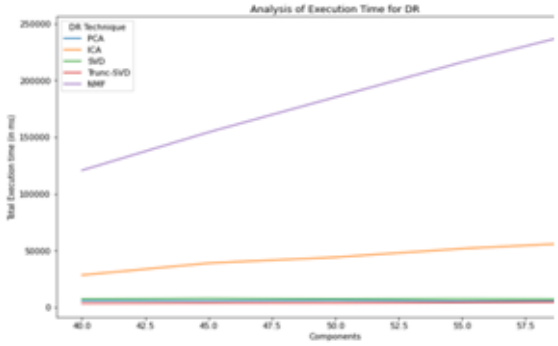


Fig. 4. Analysis of Execution Time

4. PERFORMANCE EVALUATION OF DR BASED ON ACCURACY:

Accuracy of the algorithms is calculated on the basis of Mean Squared Error (MSE). It calculates the difference between the preprocessed image before dimensionalityreduction and after performing dimensionality reduction as shown in Fig. 5. The formula to calculate MSE is:

MSE =

where, MSE = Mean Squared Error, n = no. of data points, x_i = Observed values, y_i = predicted values.

$$\sum_{i=1}^{n} (x_i - y_i)^2$$

Percent Accuracy of DR Technique	Number of Components				
	40	45	50	55	60
PCA	85.7	87.8	90.4	93.2	95.9
ICA	85.7	87.8	90.4	93.2	95.9
SVD	84.3	87.5	90.1	93.0	95.7
Truncated-SVD	84.2	87.3	90.3	92.9	95.6
NMF	83.0	87.2	87.8	89.5	93.4

Table. 1. Percentage Accuracy Analysis of DR Techniques with varying components

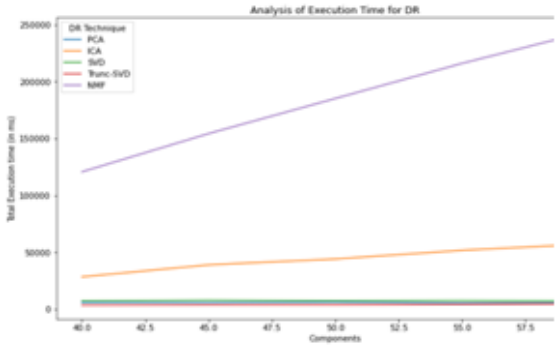


Fig. 5. Accuracy analysis of DR Techniques for 60 components

V. CONCLUSION AND FUTURE SCOPE

In this work, the effect of dimensionality reduction techniques, namely PCA, ICA, SVD, Truncated-SVD and NMF have been applied on ImageNet dataset. On the basis of time requiredforexecuting a particular algorithm and60 components of each image, Truncated-SVD has lowest execution time of 4.75 seconds followed by PCA having time 5.60 seconds. Thus, when considering load handling as the basis of dimensionality reduction, truncated-SVD will give the fastest result. By applying dimensionality reduction using above algorithms, we get highest accuracy of PCA and ICA with accuracy of 95.88% for 60 components, whereas NMF gives the least accuracy of 93.40% and accuracy of SVD and Truncated SVD is 95.71% and 95.64% respectively. PCA and ICA has same accuracy as the co-variance matrix of the image can explain the redundancies present in an image.To conclude, in terms of error-handling and load-handling, PCA and Truncated SVD will give fastest output and retain maximum information. Fig. 6. depicts analysis plot.

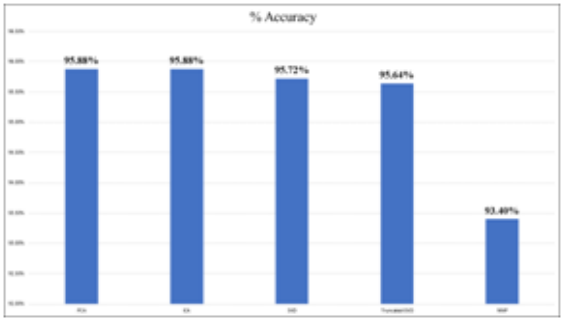


Fig. 6. Accuracy Analysis Plot

This project can be further extended by using other techniques of dimensionality reduction on the same dataset. These unsupervised DR techniques can also be analyzed for different forms of high dimensional data such as video, audio, micro gene expression data, etc.

REFERENCES

[1] Neelam Agrawal, and Kesri Verma, "Dimensionality Reduction on Hyperspectral Data Set," 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), pp. 20 April 2020.

[2] Md. Golam Sarowar, Arthy Anjum Jamal, Aniksaha, AbirSaha, "Performance Evaluation of Feature Extraction and Dimensionality Reduction Techniques on Various machine learning classifiers," 9th International Conference on Advanced Computing (IACC).

[3] Raji Ramachandran, Gopika Ravichandran and Aswathi Raveendran, "Evaluation of Dimensionality Reduction Techniques for Big Data," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 23 April 2020.

[4] Remya Rajesh, Shaji, C. P., Kaimal, M. R., "Singular Value Decomposition – A Revisit on A CUDA Platform," 2014.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[6] Wei Wei, Member, IEEE, Qin Yue, Kai Feng, Junbiao Cui, and Jiye Liang, Senior Member, IEEE, "Unsupervised Dimensionality Reduction based on fusing multiple clustering results," IEEE Transactions on Knowledge and Data Engineering, 21 September 2021.

[7] S. Velliangiria, S. Alagumuthukrishnanb, S. IwinThankumar Joseph, "A Review of Dimensionality Reduction Techniques for Efficient," International Conference on Recent Trends in Advanced Computing 2019, ICRTAC 2019.

[8] Alireza Sarveniazi, "An Actual Survey of Dimensionality Reduction," American Journal of Computational Mathematics, 2014.

[9] L.J.P. van der Maaten, E.O. Postma, H.J. van den Herik, "Dimensionality Reduction: A Comparative Review," Journal of Machine Learning Research January 2007.

[10] Prof Rasendu Mishra, Dr PritiSajja, "Experimental Survey of Various Dimensionality Reduction Techniques," International Journal of Pure and Applied Mathematics Volume 119 No. 12 2018

[11] Xiao-Dong Wang, Rung-Ching Chen, Zhi-Qiang Zeng, Chao-Qun Hong, and Fei Yan, "Robust Dimension Reduction for Clustering with Local Adaptive Learning," IEEE Transactions on Neural Networks and Learning Systems, Vol. 30, No. 3, March 2019.

[12] K. Swanthana, K. Swapnika, Dr. Y. Vijayalatha, "Dimensionality Reduction in Big Data using Unsupervised Learning an Overview," International journal of Innovations in Engineering and Technology (IJIET)

[13] A. Eftekhari, R. A. Hauser, and A. Grammenos, "MOSES: A streaming algorithm for linear dimensionality reduction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 11, pp. 2901–2911, 2020.

[14] K. Raghavan, V. Samaranayake, and J. Sarangapani, "A multi-step nonlinear dimension-reduction approach with applications to big data," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2249–2261, 2019.

[15] Yousef Jaradat, Mohammad Masoud, Ismael Jannoud, Ahmad Manasrah, "A Tutorial on Singular Value Decomposition with Applications on Image Compression and Dimensionality Reduction," 2021 International Conference on Information Technology (ICIT).

[16] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1-3, pp. 37–52, 1987.

[17] Prof. Sumit Sharma, Omprakash Saini, "A Review on Dimension Reduction Techniques in Data Mining," Bhopal, India, 2018.

[18] Joshi Snehal, K. SahistaMachchhar, "An Evaluation of Dimensionality Reduction Techniques – A Comparative Study," Rajkot, India, 2006.

[19] Aamir Khan, Hasan Farooq, "Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition," Punjab, India 2011.

[20] Shereena V. B. and Julie M. David, "Significance of Dimensionality Reduction in Image Processing," Signal & Image Processing: An International Journal (SIPIJ) Vol.6, No.3, June 2015.

[21] M. Dash, H. Liu, Yao, "Dimensionality reduction of unsupervised data," National University of Singapore.

[22] W. Zhao and S. Du, "Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach," IEEE Trans. Geosci. Remote Sens. 54(8): 4544–4554 2016.

[23] Thippa Reddy Gadekallu, Praveen Kumar Reddy, Kuruva Lakshman, Rajesh Kaluri, "Analysis of Dimensionality Reduction Techniques on Big Data," IEEE Access (Vol: 8) 16 March 2020.