SUPERVISED PCA VS LDA AS PREPROCESSING FOR IMAGE CLASSIFICATION

Dhruv Jayesh Dholakia, Mohammed Saquib Suglatwala, Najeebuddin Mohammed

Department of Electrical and Computer Engineering,
University of Waterloo

ABSTRACT

One of the biggest issues when working with images is the high dimensionality of the data, commonly in the order of tens of thousands. In this paper, we present an analytical study comparing supervised Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) as dimensionality reduction methods for image classification using a simple deep neural network (DNN). We first explain the theoretical foundation of the two methods followed by an experimental comparison on the Modified National Institute of Standards and Technology (MNIST) dataset.

***Index Terms*—** supervised PCA, LDA, Dimensionality Reduction, Image preprocessing

1. INTRODUCTION

The rapid growth in the field of visual communication over the past decade has led to images and videos becoming a primary form of interaction on the Internet. This has in-turn led to a wide range of applications and services being provided that either serve or use these visual objects for a variety of applications. From the recommendation algorithms behind Instagram and TikTok, the reCAPTCHA security service that can distinguish humans from bots, biomedical image segmentation to weather forecasting, all of these have been made possible by the exponential growth in deep learning for image and video processing.

Image classification is a fundamental process that is used by various applications in the visual communication space, which involves categorising images received by the system to a predefined set of classes or labels. It is used both as a standalone system in fields like medical imaging as well as a part of a more complicated pipeline in fields like autonomous driving. Due to versatility and indispensable nature, making the classification process faster and more efficient can have a significant and multiplying effect on the performance of applications. A very common problem in the visual space is the high dimensionality of the data, which in turns leads to need of larger amount of data in order to obtain usable results along with a greater amount of computational power.

A now widely accepted solution that has been proposed to speed-en up the process of image classification is to perform dimensionality reduction on the images before they are fed to any model as natural images are sparse in nature. Dimensionality reduction is the transformation of data from a high-dimensional space to a relatively low-dimensional space such that only meaningful properties are retained [1]. There are various methods that have been proposed under this, and are widely classified into 2 subcategories; feature selection and feature extraction. Feature selection involves finding a subset of the original features that meaningfully represent the distribution of the input data. Feature extraction involves the mathematical transformation of the original features into new features (usually much fewer in number) while retaining meaningful information. The respective methods that fall under these subcategories have their own advantages and disadvantages, but the latter has proven to perform better for a wider variety of datasets.

In the following sections, we present a objective comparison between supervised PCA vs LDA. Both of these as supervised linear feature extraction methods but differ in the objective and mathematical foundation on which they are based on. Section-2 provides literature review while section-3 will go over the theoretical aspects and methodology followed by the experimentation details of our comparative study in section-4. Section-5 and 6 will present our results and conclusions respectively.

2. LITERATURE REVIEW

Various methods have been proposed over the years to perform dimensionality reduction on images. The most commonly used is the standard unsupervised version of Principal Component Analysis (PCA) [2]. The objective of PCA is to find the optimal projection matrix using eigen value decomposition. PCA computes vectors called principal components, and choosing a smaller subset of these vectors based on the descending order of their eigen values results in a smaller dimension representation of the input while preserving the variance on the distribution. PCA was proposed as a general method but has since been proven to work well on various different modes of input including images. Another linear method that has been shown to work for images is Multidimensional Scaling [3]. For nonlinear transformations, Kernel PCA [4] was proposed as an improvement on PCA. It uses

kernel functions to map the nonlinear data to a linear latent space and then compute the principal components. Kernel PCA is a little trickier to implement for images due the selection of the right kernel depending on the dataset. Despite this, it has been proven to work for face recognition problems ([5], [6]) and for hyperspectral images ([7], [8]). Local Linear Embedding [9] and Isomap [10] are two other nonlinear methods have been proposed for dimensionality reduction, but kernel PCA is still the most widely used one.

Various papers go in detail comparing two or more of the methods stated above in terms of performance and efficiency for image processing. Supervised PCA is a comparatively newer method that has not gained much traction despite the popularity of PCA. Since supervised PCA aims to compute principal components that maximise the dependence to the target, it has a semantic similarity to LDA with regards to the objective of the two methods. This is the motivation behind our comparative study proposed on this paper.

3. METHODOLOGY

3.1. LDA

Linear Discriminant Analysis was derived from the discriminant function originally proposed by R.A Fischer [11], with one of the key differences being that LDA makes a few assumptions regarding the data like the classes being normally distributed [12]. LDA aims to transform that data into a lower dimensional space such that the distance between samples of different classes is maximum while the separation between samples of the same class is minimum. This is shown in the Figure 1. As a result, here both categorical dependant variable (class label) and continuous dependant variables are required [12]. LDA reduces dimensions from number of features to the number of classes. The concept behind LDA can itself be used as a classifier but here this is not significant.

Now to project data, we need to have a measure that display data points from different classes as far as possible, but keeping data points close if they are from a same class and to achieve this a criteria is defined as:

$$(\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)$$

The above criteria was proposed by Fischer in [11] where μ_1 and μ_2 shows the mean of classes 1 and 2 respectively and σ_1 and σ_2 is the variance of the above mentioned classes. LDA looks for maximizing the above criteria. Moreover, the projection is shown in the Fig. 1 where the right one has a good separation between classes after reducing dimensions.

LDA suffers from one major disadvantage which is the assumption of Gaussian distribution in the data and this assumption is impractical. LDA as a dimensionality reduction works by arranging dataset in an another space by trying to achieve maximum linear separability. There will be still overlapping scenarios because of non-linearity.

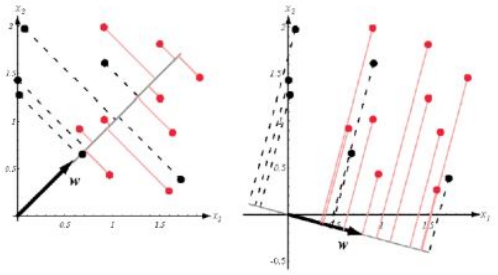

Fig. 1. Projection of LDA [13]

3.2. Supervised PCA

Principal Component Analysis (PCA) is the most popular dimensionality reduction methods but its effectiveness is limited by the unsupervised nature of its computation. Since the principal components are calculated to maximise the variance of the dataset, no information from the class labels is used to drive the transformation towards the modes of variability that can allow for better classification. To remedy this, supervised PCA [14] has been proposed. Instead of trying to maximise the variance, the principal components computed by supervised PCA aim to maximise the dependence on the response. In terms of dimension transformation, supervised PCA aims to find a lower dimensional space where the statistical dependence between the samples and response is maximised. The original authors of supervised PCA state that one of the biggest advantage of their method over other supervised dimensionality reduction methods is that it can be used for regression problems, but that is not of importance for our case.

As stated, the statistical dependence between the features and the response is to be maximised in the lower dimension subspace. This statistical dependence is defined by the Hilbert-Schmidt Independence criterion [15]. This metric measures the dependence between two random variables by computing the Hilbert-Schmidt norm of the cross-covariance operator associated with the reproducing kernel Hilbert Spaces (RKHSs) [16] of the variables. Mathematically. it is defined as follows.

$$\begin{aligned} HSIC(P_{X,Y}, F, G) = E_{x,x`,y,y`}[k(x,x`)l(y,y`)] \\ +E_{x,x`}[k(x,x`)]E_{y,y`}[l(y,y`)] \\ -2E_{x,y}[E_{x`}[k(x,x`)]E_{y`}[l(y,y`)]] \end{aligned}$$

Here, X and Y are the two random variables and x and y are samples from the respective variables. Let F and G are the RKHS of the two random variables while k and l are the kernel functions associated with the two RKHS respectively. Higher the value of the HSIC, greater is the dependence between the two variables. So, in supervised PCA, the subspace to which the input data is transformed has the highest HSIC between the features and the response.

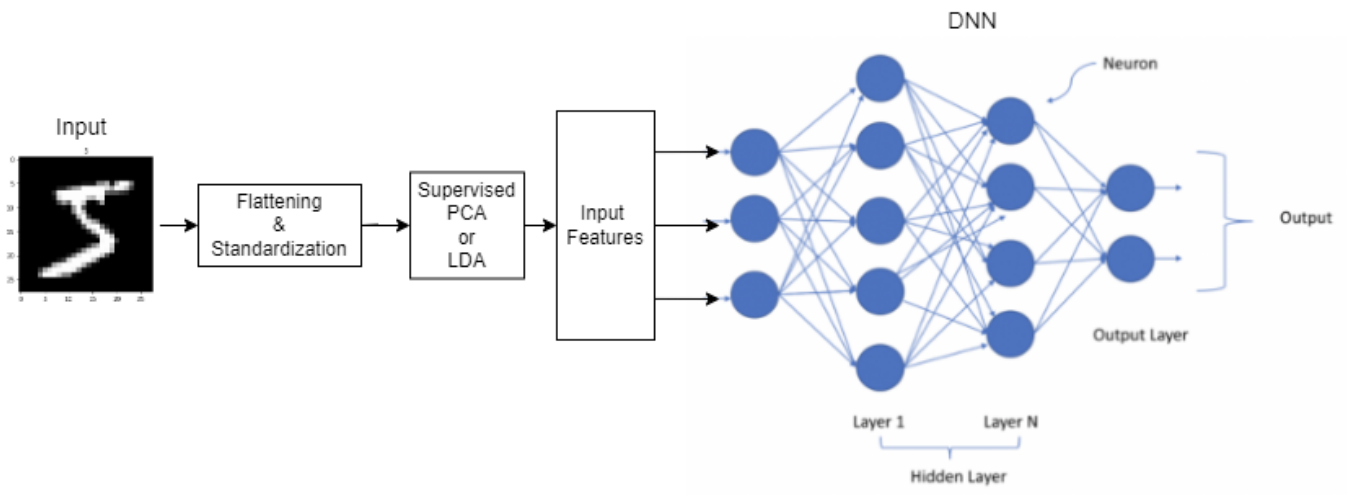

Fig. 2. Proposed Architecture with DNN

While supervised PCA does achieve its objective, it fails to prove that the lower dimension space found actually contains or retains the intrinsic information from the features of the data. This indicates the principal components computed do not provide a lower dimensional representation of the data, rather a representation that is directed towards the supervised task at hand.

4. EXPERIMENTATION

4.1. Data

This study uses the MNIST digit classification [17] dataset, specifically the implementation from Keras API for performance comparison. The original train dataset contains 60,000 images of 10 classes with a size of 28x28, but we used 30,000 images for training due to limited computational resources. For testing, we used 10,000 images. The image is first flattened into array of size 1x784 making the training dataset of size 30000x784 and test dataset of size 10^3x784 and then standardization is applied to each individual images with a vector size of 784.

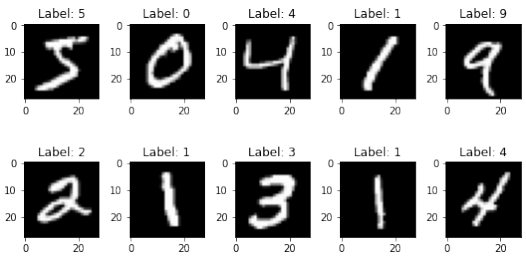

Fig. 3. MNIST Digit Classification Dataset

4.2. Metrics

The performance metric is accuracy and since we are working on classification problem, the loss metric used for training is categorical cross-entropy. Accuracy is a metric which gives the fraction of predictions our model got right. It is defined by number of correct predictions over total number of predictions. Categorical cross-entropy shows how distinguishable the given classes are by their discrete probability distribution. With this loss function, the model learns to give a higher probability of estimation to the correct digit compared to probability for the other digits.

4.3. Architecture

The DNN has 5 hidden layers with nodes 50, 150, 300, 200, 70 consecutively. The number of input nodes is dependent on the features chosen by our respective method. The output layer consists of 10 nodes with softmax activation function for classification of digits. All intermediate layers has Rectified Linear Unit (ReLU) as their activation function. For training the neural network, ADAM optimizer is used with its default learning rate. The number of epochs and batch size is set to 20 and 50 respectively.

Figure 2 shows our entire pipeline from the input image to the final prediction of the digit.

4.4. Software

All programming was done using **Python3** and **TensorFlow** v2 in Google Colab. For array computation and image processing, we use *NumPy, SciPy and Scikit-Learn* libraries. The python variant of the *Matplotlib* library was used for plotting of sample images and the final results.

5. RESULTS

Figure 4 and 5 shows the comparisons between the improvement of training accuracy and loss as the two models are trained. As observed, both models reach similar values for both loss and accuracy as the curves approach stability. The training metrics are obviously not the a true representation of how the models compare but are one of the first observations from which differences can be drawn.

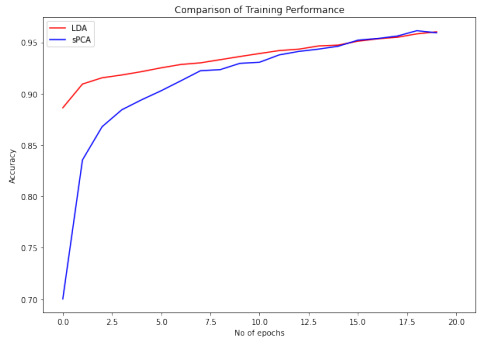

Fig. 4. Supervised PCA vs LDA Training Accuracy

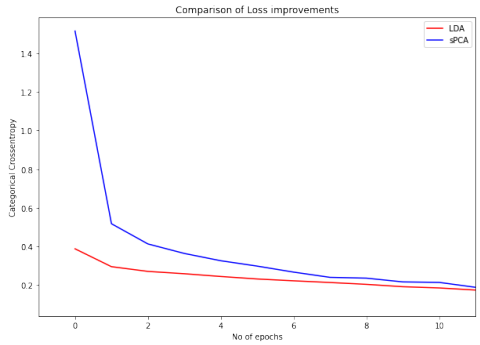

Fig. 5. Supervised PCA vs LDA Training Loss

In terms of test evaluation, LDA performs better achieving a test accuracy of 91% when compared to supervised PCA with 80%. Hence, for our MNIST dataset with 30,000 images, using LDA as the chosen method of dimensionality reduction is the better option. Another important observation is that the training and test accuracy of the model that uses LDA is very close while that of the model using supervised PCA is not. This indicates that the latter is slightly overfitting when compared to the former.

Our main comparison is between two linear methods using the same DNN to allow a level playing field. However, a modification of supervised PCA is also proposed in [14] that allows for non-linearity and is expected to perform better in most cases. This is done using non-linear kernels and is hence called Kernel supervised PCA (similar to kernel PCA). We test out our pipeline with kernel supervised PCA using sigmoid function in order to observe if an improvement is achievable. This produced a test accuracy of 90%, which is much closer to the LDA model and a significant improvement over the linear supervised PCA. This shows that despite the linearity constraint of LDA, it can perform on par with a non-linear method in our case.

6. CONCLUSION

We successfully present a comparative study between LDA and supervised PCA as dimensionality reduction methods for image classification problems. It is important to note that our study was only conducted using one dataset and results can vary based on the data. The computational resources used for experimentation were unique to us, so a subjective statement comparing training and prediction times can be presented but objective values have no real-world meaning.

6.1. Advantages and Disadvantages

LDA performs well for our case giving better test accuracy while using lower computational resources like time and space as compared to supervised PCA. One of the reason for this result can be the disparity in number of components that both method uses to reduce the dimensions of the input image. During our experimentation, we observed that when the amount of training data is less (10,000), supervised PCA performs on par with LDA.

6.2. Future Work

The code used for supervised PCA was an unofficial implementation while LDA was implemented using an industry-standard library. This could have added complexity and speed issues to our comparison, so an obvious future improvement would be to implement an optimised and industry-ready version of supervised PCA to allow for a more objective comparison in terms of computation. Our work can be progressed upon by including more non-linear methods in the comparison to provide a survey of the best performing dimensionality reduction methods for image preprocessing. We touch upon this slightly in our results with some experimentation using the kernel version of supervised PCA, but a more extensive study can be proposed. Our comparison is only on one dataset with gray scale images and comparatively fewer features than natural images. Implementing the same on larger and more complex datasets in the future should give much more generalised observations about the performance of the two methods.

7. REFERENCES

[1] Wikipedia contributors, "Dimensionality reduction — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Dimensionality_reduction&oldid=1072944127, 2022, [Online; accessed 9-April-2022].

[2] Harold Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, pp. 417, 1933.

[3] Quan Wang and Kim L. Boyer, "Feature learning by multidimensional scaling and its applications in object recognition," in *2013 XXVI Conference on Graphics, Patterns and Images*, 2013, pp. 8–15.

[4] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[5] Quan Wang, "Kernel principal component analysis and its applications in face recognition and active shape models," 2012.

[6] Hala M Ebied, "Feature extraction using pca and kernel-pca for face recognition," in *2012 8th International Conference on Informatics and Systems (INFOS)*. IEEE, 2012, pp. MM–72.

[7] Aloke Datta, Susmita Ghosh, and Ashish Ghosh, "Pca, kernel pca and dimensionality reduction in hyperspectral images," in *Advances in Principal Component Analysis*, pp. 19–46. Springer, 2018.

[8] David Ruiz, Bladimir Bacca, and Eduardo Caicedo, "Hyperspectral images classification based on inception network and kernel pca," *IEEE Latin America Transactions*, vol. 17, no. 12, pp. 1995–2004, 2019.

[9] Sam T Roweis and Lawrence K Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[10] Joshua B Tenenbaum, Vin de Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[11] Ronald A Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[12] Wikipedia contributors, "Linear discriminant analysis — Wikipedia, the free encyclopedia," 2022, [Online; accessed 10-April-2022].

[13] Krishan, "Dimensionality reduction via linear discriminant analysis," Dec 2018.

[14] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357–1371, 2011.

[15] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *International conference on algorithmic learning theory*. Springer, 2005, pp. 63–77.

[16] Saburou Saitoh, "Theory of reproducing kernels," in *Analysis and Applications—ISAAC 2001*, pp. 135–150. Springer, 2003.

[17] Yann LeCun, Corinna Cortes, and CJ Burges, "Mnist handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.