CSCM45 Big Data and Machine Learning

Coursework 2: Object Recognition

**Introduction**

We have been instructed to use an image dataset set for this experiment, which is a subset of CIFAR-10 [1]. The aim, of the experiment, is to try and find out what machine learning technique or combination of techniques, creates the best prediction accuracies. The following techniques will be using: Principle component analysis (PCA), linear discriminant analysis (LDA), support vector machine (SVM) and a neural network (NN). We have a benchmark of 44.68% provided. We will also feed the data directly into a NN for a direct comparison along with SVM. Another metric that we used for comparison is the time it takes to compute, as we can assume that using a form of dimension reduction will lower the accuracy. However, the possible faster time it takes to train the model might be a justifiable trade-off. With experimenting with different optimisers and hidden layers within the NN, the SVM method, using a poly kernel, came out the best with an average of 55.90% accuracy which took 47 seconds while PCA into a NN come out the worst with an average of 21.89%. We found that using both PCA and LDA created the worst results.

**Method**

**Extracting the data from the Dataset**

We first need to extract the data from the two provided 4D arrays, one of size 32x32x3x10000, the training data, and the other 32,32,3,1000, the testing data. In order to be able to effectively use the data a provided function, from the SKLearn library, SKImage was used to extra the features. The features become a numpy array shaped 324, 10000 with the same done to the testing set, which created a numpy array shaped 324, 1000.

**Initial Benchmarks**

We first fed the extracted features into a NN. Doing this was to provide an initial benchmark for comparison. The NN settings were changed using a different number of layers, optimisers and activations to try and find the best outcome for the data. Feeding the data directly into SVM was also done. Again changing the kernel from linear, sigmoid and poly as well as changing the penalty of error.

**Dimensionality Reduction Techniques**

We will use both PCA and LDA; these are ranked highly as some of the best tools for this purpose [4]. Dimensionality reduction will help with the curse of dimensionality [5,6]. First by themselves, fitting and then testing the data. We will then feed the reduced data then into a NN and into SVM, to see what generates the best results. Both forms of dimensionality reduction will also be used together to see what kind of impact this has on the outcomes. For PCA it was decided to only use 100 components (fig 0).

**Supervised learning Techniques**

As we have provided labels for the data, we will use supervised learning to do the predicting. As we now have initial benchmarks, we will use a series of combinations to see what creates the best results. The experiments are PCA into NN; PCA into SVM; LDA into NN; LDA into SVM; PCA into LDA into NN; PCA into LDA into SVM. In order to inform where the training might get stuck, we will use a confusion matrix.

**Optimisation**

To find out the best outcomes, we tried using different optimisers for the NN and SVM. Optimisations included changing the activation on the hidden layers, the number of hidden layers as well as the number of nodes on each layer. As well with the SVM, multiple kernels were changed to see what generated the best results.

**Results**

When feeding the data directly into the NN, the best results found was using just one hidden layer with 200 nodes. This method, on average, took about 126 seconds to complete with a 52% accuracy (fig 1). However, the data directly into SVM took 47s and had an accuracy of 55.90% (fig 2).

When we fed PCA's data into the NN, at the NN optimised settings, one layer and 200 nodes, it took 280s and only had an accuracy of 21.80% (fig 3). The results show that not only is using PCA with a NN less accurate but it also takes longer to compute it. However, feeding the PCA into SVM did create faster speeds, 17s, it also produces worse prediction results, 18.70% (fig 4).

LDA into the NN optimised settings generated results around the 37% mark and took 105s (fig 5). As the results show, this is defiantly quick that just the NN but quite a considerable drop off with accuracy. However, when we optimised the NN for the LDA data, one layer of 4 nodes, the results were better. 54% accuracy and 100s. So using this method would defiantly be useful, but the model did struggle with predicting cats and deer (fig 6). While feeding it through the SVM, the speed was incredible. It was completed in 1.5s and had 35.90% accuracy. So the accuracy was not as good, but for the speed, it could be a justifiable the choice (fig 7).

For PCA into LDA into NN or SVM, we found the results were quite poor but feeding the data into the NN done better than using SVM. NN had 29% and 118s while SVM had 25.20% and took 1.6s.

**Conclusion**

After using a range of different dimensionality techniques and then feeding them into a NN or an SVM, we found some great insights. The processes included: PCA into a NN and SVM; LDA into a NN and SVM; PCA into LDA into NN and PCA into LDA into SVM. As we expected, the dimensionality reduction did lower accuracy, but the speed gains were not enough to warrant using them. Nevertheless, the method that does the best, in terms of accuracy and speed, was using just the SVM with an accuracy of 55.90% and taking 47s. A close contender behind is using LDA into a NN with the NN with it optimised for the LDA data. This method achieved Both these method not only beat the benchmark score, but it also was the quickest. The quickest again being the SVM. On the other hand, using both PCA and LDA together into either a NN or SVM created poor results but the SVM method only took a staggering 1.5s to compute.

An unexpected outcome found was that overall, all of the methods struggled with differentiating between the animals. Some do better than others, but these were the lowest scores in the confusion matrix.

Through optimising, there was no best one fits all. In order to get good results with LDA into the NN, the best results came from having one layer and four nodes, but this would give poor results for the data directly into a NN. Directly into a NN tended to do better when there were fewer layers but more nodes within the layers. However, these methods alone are not suitable for this task. Research suggest that using a Convolutional Neural Network (CNN) would be a better NN for completing this task [2]. Also, most research suggests using PCA for this task along with a CNN, so more research needs to be done into this as PCA was the weakest option within these experiments.

**References**

[1] subset reference CIFAR-10 <https://www.cs.toronto.edu/~kriz/cifar.html>

[2] <https://towardsdatascience.com/cifar-10-image-classification-in-tensorflow-5b501f7dc77c>

[3] <https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9>

[4] <https://thenewstack.io/3-new-techniques-for-data-dimensionality-reduction-in-machine-learning/>

[5] <https://www.quora.com/Is-PCA-supervised-or-unsupervised>

[6] <https://en.wikipedia.org/wiki/Curse_of_dimensionality>

**Appendix**

Fig 0: Fig 1: Fig 2:

A picture containing screenshot

Description automatically generated A screenshot of a cell phone

Description automatically generated A close up of a piece of paper

Description automatically generated

Fig 3: Fig 4: Fig 5:

A picture containing electronics

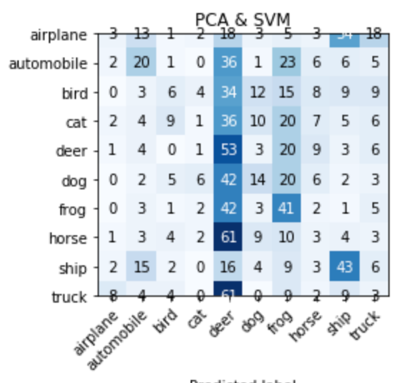
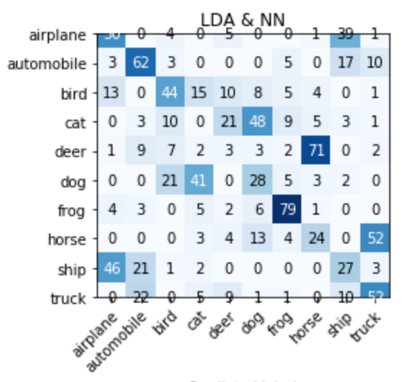
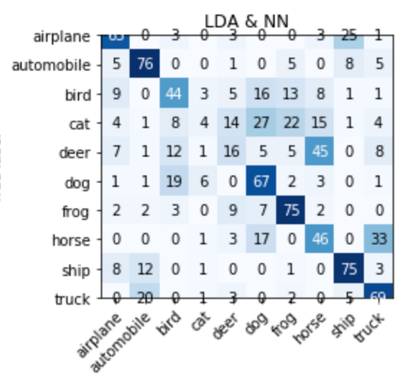
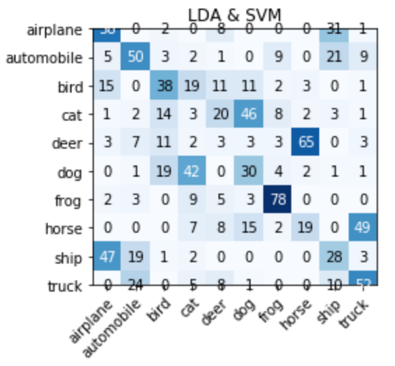
Description automatically generated  

Fig 6: Fig 7:

Submission

Submit your work electronically to Blackboard. Your report should be in PDF format only. Compress your Python source code and report into a Single Zip file. The deadline for this coursework is 11AM Monday 9th December.

Notes:

With the provided dataset being so big, with many features, it is potentially going to take a long time to process and train the Neural Networks. With using a dimensionality reduction techniques along side a NN, it will be compared to see if the reduction in potential accuracy is a valid trade off for processing time.

The aim of the project is to see what method creates the best prediction accuracy form the provided dataset. This will be using LDA, PCA, SVM and a Neural Network to see what provides the best results. The accuracy, as a percentage, and how long the algorithms take to run will be used as factors/benchmarks for comparison.

The first task will be using the provided function to extract the features from the dataset and then add them to a numpy array. The same will be done to the training data set provided. Training and testing data labels will be used for classifying the data and testing the accuracy.

**Supervised learning Techniques**

As we were provided with labels for the data, we will use supervised learning to do the predicting. The data will also be fed through a NN without any reduction to see a comparison, along with the running time of the task processing. As well as SVM, to have a comparison of non-dimensionality techniques. In order to see where the training might get stuck, a confusion matrix will be generated.

Present your experimental results in this section. Explain the evaluation metric(s) you use and present the quantitative results (including the confusion matrix).

PCA into SVM or NN does not produce good results. Along with using it into a LDA (PCA-> LDA-> ???)

All methods tend to confuse the animals up.