# Big Data & Machine Learning

Coursework: Ellis Thompson 951536

## Introduction

Using a subset of the CIFAR-10 dataset, one which contains a train set of 10000 images over 10 categories for training and 1000 images over the same 10 categories for testing, a system is to be produced which, through machine learning techniques, can be used to identify similarly formatted images. This was to be done using learned algorithms from the course and within python.

To do this, solutions using K-means, Gaussian Mixture Modelling, Linear Discriminant Analysis, Support Vector Machines and Neural Networks have been generated. Focus has been paid to the Neural Network approach as, due to the higher number of changeable variables, it is expected to be able to be the most accurate. Additionally, through further examination, the field of Neural Networks can be expanded on and, using Convolutional neural networks in partnership with other methods, it may be possible to achieve higher accuracy [1].

Of these it has been found that the Neural Network approach was found to yield the best result, being able to correctly classify images in the test data set 52.2% of the time.

## Method

Many of the algorithms and methods were adapted from material available on the course with code and implementations being iterations of the supplied material.

### *Overview*

K-means, Gaussian Mixture Modelling, Linear Discriminant Analysis, Support Vector Machines and Neural Networks, have been selected as methods to focus on for a solution. This provides a broad spectrum of methods to compare, contrast and analyse, and so allowing for a more accurate method to be selected. Each system took training data as a Histogram of Oriented Gradients (HOG) array. The HOG reduces the image data, removing unnecessary features, and, in this case, keeping useful information, the magnitude of x and y derivatives (gradients) which we expect to be larger around corners. As the corners of an object contain larger gradients and thus more information about an object, useful for image recognition.

Data and labels are then shuffled to introduce entropy into the system and remove any orders present in the data, and then normalised. Normalising the data brings it onto a common scale without distorting differences. This reduces the chances of larger values influencing the systems more.

### *K-means and GMM*

Both a form of unsupervised learning the systems are set to contain 10 clusters/components. Both systems perform with an average accuracy on the test data <15%. However, as these systems are unsupervised their accuracy varies wildly, with no backpropagation, the machine cannot accurately learn. This makes it a bad choice for image classification due to the ambiguity and unreliability for accurate results.

### *Linear Discriminant Analysis*

LDA is a method used to find a linear combination of features that characterises/separates two or more classes [2]. As LDA is supervised it is expected to perform consistently better than GMM or K-means. Averaging 49.4% accuracy LDA performed better than initially expected, being able to classify with an accuracy of nearly 50%. Using Principal Component Analysis to reduce the dimensionality down to 100 components, the LDA method performs better, correctly classifying 50.9% of the test images. This dimensionality reduction reduces noise in the input features allowing the system to be more accurate.

### *Support Vector Machine and Neural Network*

An SVM, using PCA to reduce the input dimensionality was able to achieve an average accuracy of around 50.6%.

On average the Neural Network could achieve an accuracy of ≥53% on the test dataset. The final architecture for the Neural Network which gave the best results, were two hidden (dense) layers with 22 neurons in each, using the `ELU`

(Exponential Linear Unit) and default Keras values, with a bias of 0.1. Additionally, the model split the data into batch sizes of 238 and ran for 700 epochs. It was found that accuracy only improved on the training data with more than 600 epochs. Smaller improvements were seen in higher epochs on the test data but none noticeable enough to require the additional computational time. So was not used for the task. Switching the activation function to `SELU` was considered as 'It is generally found that SELUs on their own learn faster and better than RELUs' [3], However in practice, for the task, the SELU performed with less accuracy, as did RELUs on average.

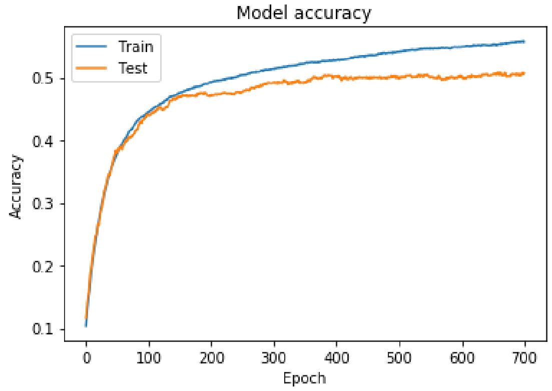
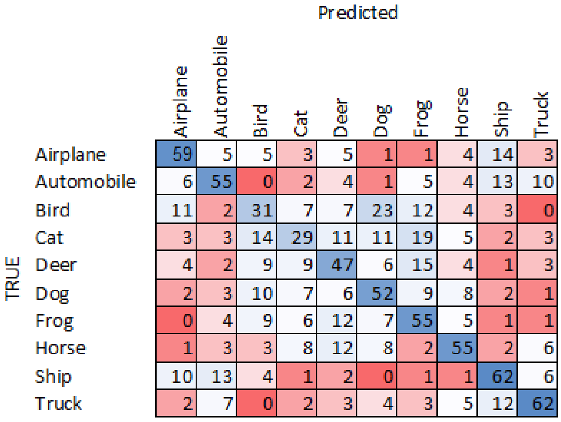


*Figure 1: Showing the Train vs Test accuracy on the neural network of $f_1$ accuracy = 51%.*


Predicted
TRUE

| | Airplane | Automobile | Bird | Cat | Deer | Dog | Frog | Horse | Ship | Truck |
|---|---|---|---|---|---|---|---|---|---|---|
| Airplane | 59 | 5 | 5 | 3 | 5 | 1 | 1 | 4 | 14 | 3 |
| Automobile | 6 | 55 | 0 | 2 | 4 | 1 | 5 | 4 | 13 | 10 |
| Bird | 11 | 2 | 31 | 7 | 7 | 23 | 12 | 4 | 3 | 0 |
| Cat | 3 | 3 | 14 | 29 | 11 | 11 | 19 | 5 | 2 | 3 |
| Deer | 4 | 2 | 9 | 9 | 47 | 6 | 15 | 4 | 1 | 3 |
| Dog | 2 | 3 | 10 | 7 | 6 | 52 | 9 | 8 | 2 | 1 |
| Frog | 0 | 4 | 9 | 6 | 12 | 7 | 55 | 5 | 1 | 1 |
| Horse | 1 | 3 | 3 | 8 | 12 | 8 | 2 | 55 | 2 | 6 |
| Ship | 10 | 13 | 4 | 1 | 2 | 0 | 1 | 1 | 62 | 6 |
| Truck | 2 | 7 | 0 | 2 | 3 | 4 | 3 | 5 | 12 | 62 |

*Figure 2: Showing the confusion matrix for a NN solution of $f_1$ accuracy = 51%.*

The network was unable to confidently identify birds and cats, this may be due to image quality or too much noise in the images in the data set causing the accuracy drop.

## Result

The SVM, LDA and NN performed the best each achieving close to 50% on average accuracy. The NN had the most consistent results, however, all accuracies seemed to drop when performing on a lower powered machine. Additionally, all methods attempted were un able to classify the cat more than 50% of the time and the bird and deer's accuracy fluctuated, which is expected as these were only minor fluctuations.

## Conclusion

The methods were run on two separate machines, a dedicated system with high volumes of processing power and taking advantage of Nvidia's CUDA modules and a smaller personal laptop. It was found that results were lower across all methods when using a laptop.

Over all the Neural network solution appeared the most promising and with further research into Convolutional Neural Networks [1] could achieve higher, more accurate results. On accuracy SVMs, Neural Networks and LDA are all closely ranked in accuracy with them achieving near to 50% each on the test data.

However, it is possible that the dataset used is too small to draw any sound conclusions. Using the full CIFAR dataset could improve the accuracy of the overall solution.

## References

[1] A. Kumar, "Achieving 90% accuracy in Object Recognition Task on CIFAR-10 Dataset with Keras: Convolutional Neural Networks," appliedmachinelearning.blog, 24 March 2018. [Online]. Available: https://appliedmachinelearning.blog/2018/03/24/achieving-90-accuracy-in-object-recognition-task-on-cifar-10-dataset-with-keras-convolutional-neural-networks/. [Accessed 04 December 2019].

[2] jcgonzalez, "Using Linear Discriminant Analysis (LDA) for data Explore: Step by Step.," apsl, 2017 July 2017. [Online]. Available: https://www.apsl.net/blog/2017/07/18/using-linear-discriminant-analysis-lda-data-explore-step-step/. [Accessed 5 December 2019].

[3] T. Böhm, “A first Introduction to SELUs and why you should start using them as your Activation Functions,” Towards Data Science, 28 August 2018. [Online]. Available: https://towardsdatascience.com/gentle-introduction-to-selus-b19943068cd9. [Accessed 07 December 2019].