How Feature Reduction Affects Image Classification of the CIFAR-10 Dataset With Neural Networks

Introduction

Machine learning is a rapidly growing field in computer science. Image classification is at the heart of machine learning, and CIFAR-10 dataset is widely known and used to test classification techniques. The dataset is a collection of 32 by 32 pixel images of airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks that have been specially collected for the purposes of using machine learning [1]. This report will compare the effects of feature reduction on the image classification accuracy when using neural networks to classify the CIFAR-10 dataset. A control neural network will be trained and tested using features extracted from the dataset. Next the extracted features of the dataset will be reduced using a dimensionality reduction technique and will be used to train and test a neural network with the same architecture as the control. Finally the confusion matrices and accuracies of both neural networks will be compared and analysed.

Method

Feature extraction

To solve the problem of image recognition of the CIFAR-10 dataset how to extract the image features must first be decided. The method chosen was to compute the histogram of oriented gradients (HOG) for each image. HOG is often used in edge detection and image processing, especially for the purpose of image or human recognition [2]. For this report training and testing data have already been chosen from the larger dataset. For each image in the dataset 324 HOG features will be extracted.

Control Neural Network

The extracted training data will then be input into a neural network with the hyperparameters of a single hidden layer of 128 neurons. A neural network with two or more hidden layer may be considered as deep, and is generally but not always better at classification [3]. Evaluating the effectiveness of deep neural networks is beyond the scope of this project so the networks will remain shallow. There is no strict rule on how many neurons to have in a hidden layer, however it is thought that the optimum number lies between the number of neurons in the input layer and the number of neurons in the output layer [4]. For this dataset there are 324 and 10 respectively. Any number between these two variables would be a reasonable choice but arbitrarily, a number of 128 neurons has been selected in this case.
The activation function for the hidden layer will be ReLU, where as the output layer will use a Softmax function for classification as this gives probabilities of the different output classes. The optimiser that will be used for the network is adam or Adaptive Moment Estimation because it has been shown to work well on the CIFAR dataset [5].
Since the classification problem has multiple classes and the labels are integers and are not one-hot encoded, loss will be calculated using the sparse categorical cross entropy [6]. 50 epochs will be used to train the networks, this amount was settled on empirically by

PCA feature reduction

Principal Component Analysis (PCA) is a form of feature reduction that can be used on the dataset to reduce it's dimensionality [7]. As defined in *Principal Component Analysis, Second Edition* "The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated" [8]. Sklearn's PCA algorithm will be used instead of creating a PCA algorithm from scratch because it is simple to use and is well documented [9].

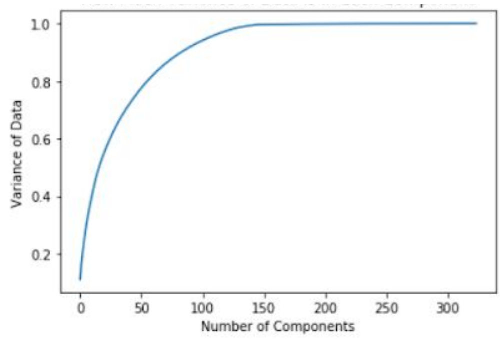

Figure 1 - The relation of variance of data to the number of components.

Sklearn's PCA implementation has a built in function to calculate "the amount of variance explained by each of the selected components". Using this Figure 1 shows that very close to 100% of the variance can be captured in the first 150 components. This is a great start is it means the number of features can be reduced from 324 to 150, half the size, without losing any important variance in data that could help in the classification.

Neural Network on varying amounts of principal components

To see how feature reduction can affect image classification multiple neural networks were trained each with an increasing amount of PCA components. For consistency a neural network with the same architecture as used on the control dataset will be used, despite it breaking the rule laid out previously about the number of neurons in the hidden layer. The task of running 150 neural networks with upto 150 inputs, is a task which is unable to be completed within reasonable time span with the available resources. Therefore the number of epochs were halved to make the computation simpler.

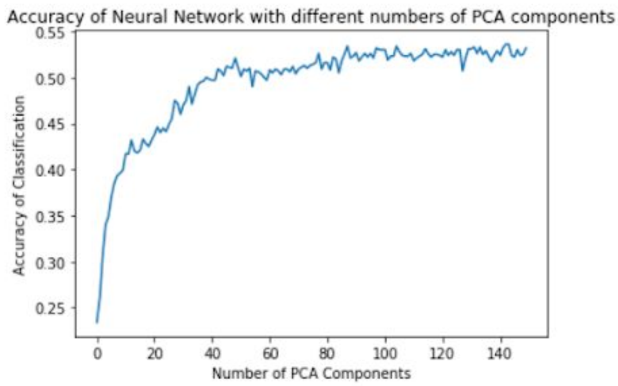

Figure 2 - How the accuracy of a neural network changes with more PCA components.

Despite Figure 1 showing 150 components were needed to capture 100% of the variance Figure 2 shows a neural network will stop becoming much notably more accurate after 100 components. From this it is concluded that around 100 features is the most the dataset can be reduced to so that a neural network can still meaningfully classify it.

Neural Network on reduced HOG features dataset

The first 106 principle components have 95% of the variance of data, therefore 106 components will be used as an input to a neural network with the same architecture as specified before. Reducing from 324 to 106 is a reduction of 67.3% to 3sf and so the input layer of the neural network can be made smaller, and thus it will require less calculations and time to train.

Results

The confusion matrices output the labels as integers, so to identify the different categories of the dataset each category has been assigned a unique number to identify them as shown in Table 1.

Table 1 - The relation between numerical labels and the CIFAR-10 categories

0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Airplane	*Automobile*	*Bird*	*Cat*	*Deer*	*Dog*	*Frog*	*Horse*	*Ship*	*Truck*

Neural Network on control data

After training with the extracted HOG features data the first neural networks has an accuracy of 54.8%, with a loss of 1.30 to 3sf. Figure 3 shows that the neural network generally performs well at classifying the majority of the data, being extremely competent at classifying ships, and poorly identifying birds and cats.

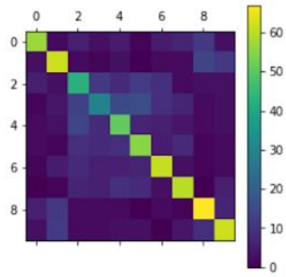

Figure 3 - The confusion matrix of the control dataset.

Neural Network on reduced features

After training with the reduced HOG features the neural networks has an accuracy of 56.3%, with a loss of 1.25 to 3sf. As shown in Figure 4, this network is now best at classifying frogs and airplanes and similarly has issues classifying birds and cats.

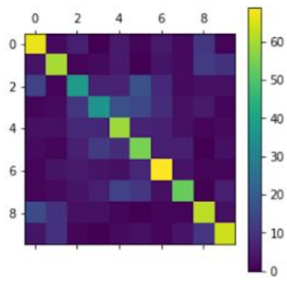

Figure 4 - The confusion matrix of the dataset with 106 reduced features.

Conclusion

This report shows feature reduction using PCA has a positive effect on the classification of the CIFAR-10 dataset using neural networks. If a large enough number of reduced components are defined then feature reduction can even outperform using no reduction. The slight increase of 2% accuracy is not the most notable, but is significant when the reduced number of features are taken into account. PCA can reduce the number of features, and thus the space required to store that data by 67.3% to 3sf, while still outperforming using no feature reduction.

A possible shortcoming of the method was having a relatively low number of epochs, a higher number of epochs could have perhaps increased accuracy or emphasised the difference between classification of the normal and reduced datasets, however since the number of epochs was kept the same for the control and the reduced features neural network the effect on the results will not be too pronounced.
When classifying images, deep convolutional neural networks are often used by researchers to get much better accuracy than is possible with a simpler neural network like the one used in this report. It is not uncommon for deep convolutional neural networks to be composed of up to 650000 neurons and many layers [10]. Compare this to the network used in this report with less than 500 neurons and only a single hidden layer, and the difference in complexities is obvious. However increasing the the final accuracy of the networks was not the ambition of this report, since the hyperparameters of the network was kept consistent the actual accuracy values should not be the primary focus of the results, rather the difference in accuracy of the two networks should be.

References

[1] *Learning Multiple Layers of Features from Tiny Images*, Krizhevsky, A., 2009.

[2] Deniz, O,. Bueno, G. Salido, J., De la Torre, F., (2011) *Face recognition using Histograms of Oriented Gradients.*

[3] Jimmy Ba, L,. Caruana Ba, R, (2013) *Do Deep Nets Really Need to be Deep?*

[4] Hagan, M., Demuth, H., Beale, M., Jesus, O., 1997, *Neural Network Design 2nd Edition*, PWS Publishing Co., Boston.

[5] Sebastian Ruder, *An overview of gradient descent optimization algorithms* (2016) viewed on 08/12/18, http://ruder.io/optimizing-gradient-descent/index.html#whichoptimizertochoose.

[6] Chengwi, *How to chose Last-layer activation and loss function* (2017) viewed on 08/12/18, https://www.dlology.com/blog/how-to-choose-last-layer-activation-and-loss-function/.

[7] Shlens, J., 2005, *A Tutorial on Principal Component Analysis Version 2*.

[8] Jolliffe, I., 2002, *Principal Component Analysis, Second Edition, Springer.*

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011) *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, **12**, pp. 2825-2830.

[10] Krizhevsky, A., Sutskever, I.,Hinton, G.,(2012) *ImageNet Classification with Deep Convolutional Neural Networks.*