Object recognition report

Lloyd Garrett - 2015825

December 2022

Abstract

In this paper, the performance of two popular machine learning algorithms are compared, Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs), on a task of object recognition in images. We present experimental results on the CIFAR-10 dataset, showing that accuracy of CNNs outperform SVMs. Additionally, the strengths and limitations of each approach are discussed.

1 Introduction

Upon encountering an object they have never seen before, a human will begin to create a category in their memory for objects that look similar to it. When a human has seen similar objects many times they become excellent at correctly categorising unseen objects that they are familiar with.

We want to replicate the behaviour of a human learning and classifying objects using learning algorithms. The CIFAR-10 dataset consists of 60,000 32x32 colour images which are each a member of one of ten mutually exclusive classes such as airplanes, dogs and cats [1]. This dataset represents the finite visual data that would be collcted by the eye. Two methods have been selected to perform the image processing and learning that the human brain is capable of. These methods are support vector machines(SVM) and convolutional neural networks(CNN).

The dataset of 60,000 images has been split into 50,000 training and 10,000 testing for use in both the SVM and CNN.

2 Method

2.1 SVM

SVM's are a type of supervised machine learning algorithm that can be used for classification tasks. In the case of the CIFAR data sets, the SVM will be trained to find the best hyperplane that can linearly separate the images into their respective classes. This helps to ensure that unseen images that are not part of the training set will be classified correctly. The particular SVC model that will be used is C-support vector classification.

We can use SVM as the classes in the dataset are linearly separable. Maximising the boundary gap between values is useful for this problem as some classes such as cats and dogs are somewhat similar and if the hyperplane boundary is too close to one of these classes there is likely to be a large preference.

The sklearn SVM model that requires an input dimensionality of 2 (samples, data vector). Because CIFAR is 4 dimensional, we extract the histogram of oriented gradients for each image to represent the variance of image width, height and colour channel as a vector. The HOG features are extracted from both the training and testing set.

The HOG features in the train and test set are then standardised so that the train set does not dominate the distance metric when fitting.

The rbf kernel for SVM has showed good performance in previous research on classifying objects with SVM's so this is used for fitting the data [3]. The max number of iterations has been set to 100 as it is not computationally feasible to wait for the result to converge.

2.2 CNN

CNN's are a type of deep learning algorithm that are particularly well-suited for image classification tasks [4]. A CNN can learn to extract important features from the images in the CIFAR-10 dataset and use them to make predictions.

A CNN has been chosen for this problem as the CIFAR-10 dataset has very high dimensionality that convolutions are well-suited to proccess. As well as, this CNN's are a deep-learning model which is the type of learning model that is most comparable to the process of how a human might learn to recognise images as it is inspired by neurons in the brain.

As classifying low-resoltion images of objects is a moderately-complex problem, a moderately complex CNN model of layers has been built. In this model there are two convolutional layers that use small kernels to reduce

the chance of underfitting. Each convolutional layer has a pooling layer for downsampling. The multi-dimensional tensor output of the final pooling layer is then flattened so that dense layers can be used for classification. One large dense layer is used for more learning which feeds into a dense layer with 10 outputs which corresponds to the 10 different possible classifications of the input image. ReLu is used as the activation function for the layers because it is computationally inexpensive and has good performance with CNN's.

For the training of the model the "Adam" optimiser is used as it gave the best performance when trying out hyperparameters. Sparse categorical crossentropy is used to calculate the loss between label predictions and ground truths.

In the first dense layer, there is a 30% chance for neuron to drop out. This means that sometimes in training neurons in this layer will become completely disconnected. This has been done as a measure to reduce overfitting as during training the model does not become reliant on certain neurons that would trip up the model when tested on unseen data [2].

3 Results

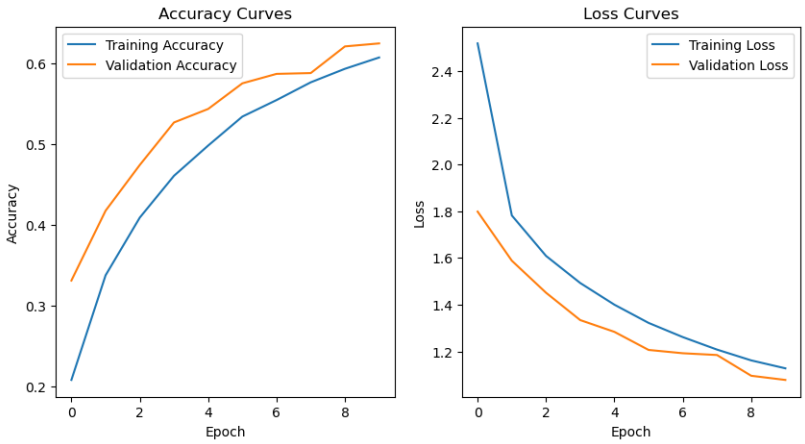
3.1 SVM

The final accuracy of the SVM model on unseen objects is 32.6%. This is only around 20 better than guessing classes at random. Overall it is difficult to interpret why the SVM performed poorly but it is possible that the low accuracy is a result of the image pre-processing not being appropriate.

It is also worth mentioning that this model will only be able to make predictions on the HOG features of images as that is what it is trained on.

3.2 CNN

The final accuracy of the CNN on unseen objects is 62.4%. Considering the relatively small amount of training performed, this is a good result.



The accuracy of classification for both the training and testing set increased over the whole training process so the CNN was successfully learning. The improvement in validation accuracy is greater than the training accuracy across all 10 epochs so the CNN is well-fit at this scale. You can also see that the loss and accuracy curves for the training and validation sets are getting closer in later epochs which suggests the loss and accuracy of the training set may overtake the validation set. This means that without stopping early this CNN may overfit.

4 Conclusion

From observing the accuracy produced by SVM compared to CNN, it is fair to conclude that for this problem a CNN is more well suited than SVM. However, it is worth noting that the performance of the SVM could have been improved with further tuning of hyperparameters.

While the results of the best performing method of CNN were good, the accuracy of correct object classification in the cifar-10 dataset is significantly less than the accuracy of human classification on the same set [5]. In order to increase the accuracy of my method, a more carefully considered model must be made with further measures to reduce overfitting so that the model can be run for more epochs and continue to increase its accuracy.

5 References

References

[1] Alex Krizhevsky. *CIFAR-10 and CIFAR-100 datasets*. URL: https://www.cs.toronto.edu/~kriz/cifar.html (visited on 12/18/2022).

[2] Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. *Survey of Dropout Methods for Deep Neural Networks*. arXiv:1904.13310 [cs]. Oct. 2019. DOI: 10.48550/arXiv.1904.13310. URL: http://arxiv.org/abs/1904.13310 (visited on 12/18/2022).

[3] Meera M .K. and Mohan Shajee B.S. "Object recognition in images". In: *2016 International Conference on Information Science (ICIS)*. Aug. 2016, pp. 126–130. DOI: 10.1109/INFOSCI.2016.7845313.

[4] Keiron O'Shea and Ryan Nash. *An Introduction to Convolutional Neural Networks*. arXiv:1511.08458 [cs]. Dec. 2015. DOI: 10.48550/arXiv.1511.08458. URL: http://arxiv.org/abs/1511.08458 (visited on 12/17/2022).

[5] Tien Ho-Phuoc. *CIFAR10 to Compare Visual Recognition Performance between Deep Neural Networks and Humans*. arXiv:1811.07270 [cs]. Aug. 2019. DOI: 10.48550/arXiv.1811.07270. URL: http://arxiv.org/abs/1811.07270 (visited on 12/18/2022).