Deep Approximation SSD for Pedestrian Detection

Abstract

Pedestrian detection is a important task in computer version and has promising prospect, such as surveillance and automobile. However, it’s accuracy and speed are also challenging at the same time. We surveyed stat-of-the art methods for object detection and chose one to improve. SSD, one of the best, has high accuracy and faster than real-time speed. Above all first, this paper adapted SSD to pedestrian detection task. Then we explored the different performance and reason. To get more efficiency, we applied Deep Approximation on SSD to make it faster without accuracy sharp dropped.

# Introduction

A Pedestrian is a person walking on road or pavement. And Pedestrian detection is a computer vision task on detecting walking people.

Suppose we are given the task of counting the number of pedestrian in a crossroad (see Figure 1.1 as an example). A computer vision approach to solving the problem might start with detecting and identifying objects in the scene that can be classiﬁed as pedestrian. Once those objects are detected, the counting process is straightforward. In general, the problem of detecting pedestrian occurs in many applications of computer vision, including image and video content management, video surveillance, assisted automotive, etc. Pedestrian detection is an active research topic in computer vision and the problem can be stated simply as: given an image or video sequence, localise all objects that are pedestrian. This problem corresponds to determining the regions within an image or video sequence containing pedestrian. The usual representation of such regions is a rectangular bounding box. Figure 1.1 shows some examples of pedestrian detection results.

//TODO general pipeline: traditional pipeline& DNN pipeline

Pedestrian detection could be widely used, such as surveillance and automobile. Both surveillance and manless driving could save human labor and improve safety. There is no doubt that pedestrian detection has promising prospect.

However, two key factor requirements, accuracy and speed, for practical scenes are challenging and paradoxical to be met at the same time.

Recently, Deep Neural Network has achived the most accurate result of object detection for now.

SSD, single shot multi box detector, is one of state of the art DNN for object detection. It is composed of feature extractor DNN model, VGG 16, and own detector. Due to its’ accuracy and speed, we decide to adapt it on pedestrian detection task. First of all, we prune SSD full connection layer to one label pedestrian, then we use pedestrian dataset to retrain SSD to improve its accuracy on single task.

Meanwhile SSD’s feature extractor DNN model, VGG 16, consumes much time in computation and results in low speed. In our test using Nvidia 1060, SSD detection speed is 30fps on 300\*300 RGB images dataset. This is not fast enough because modern surveillance image is over 1080P and the hardware price is expensive.

To solve these problems, we propose a method Deep Approximation to accelerate DNN model. Deep approximation has two steps: firstly, we prune the redundant convolution and full connect connections on DNN model without accuracy sharp drawn, then we make quantization on the pruned model to decrease computation operation time. After all, we apply Deep Approximation on SSD to make it further efficient.

# Related Work

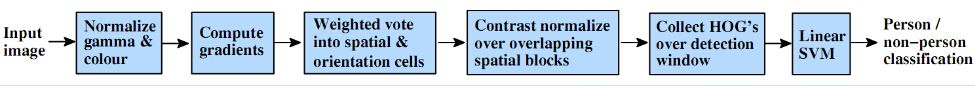


Figure 1 HOG+SVM

Traditional method for pedestrian detection is handcraft feature with classifier, e.g. HOG+SVM, as presented in Figure 1, Haar like+adaboost etc. //1st page

HOG[1], Histogram of Oriented Gradien, may be the most popular feature. It statistics the gradient of several adjacent bins of image as a histogram and uses the distribution of histogram as feature of image.//2nd page

SVM[2], support vector machine, performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.//3rd page

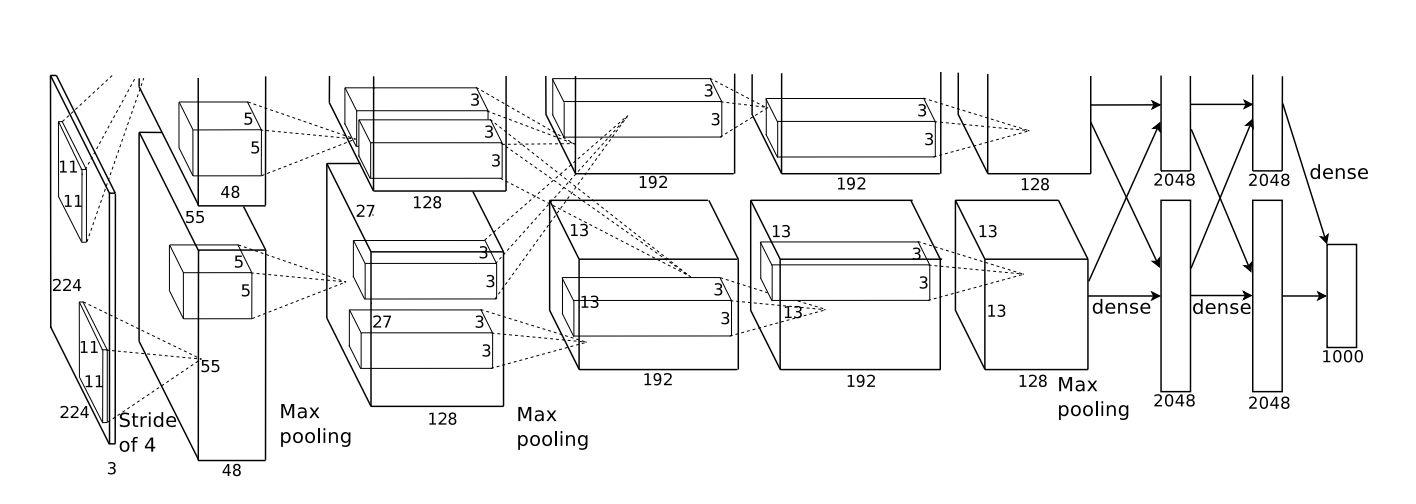


Figure 2 AlexNet Architecture

Recently, DNN, Deep Neural Network, has achived the best performance on object detection task. Science Alexnet[3], as presented in Figure[2], have achived the champion of ISLVRC 2012. Several DNN models have been proposed, such as Alexnet[3], VGG[4], Faster-Rcnn[5], YOLO[6], SDD[7].//4th page

//TODO intro VGGNet 5th page

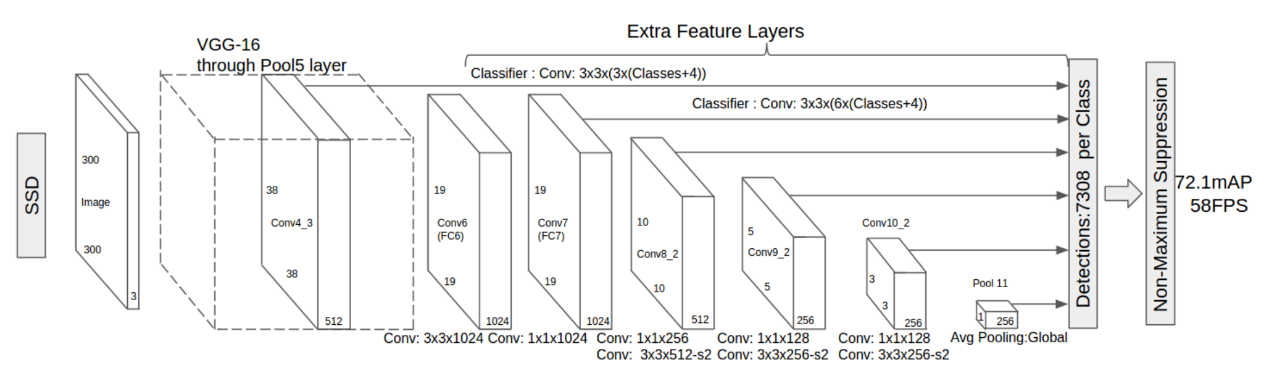


Figure 3 SSD Architecture

SSD[7], one of state-of-the-art, is a method for detecting objects in images using a single deep neural network. It discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape.//6~7th page

Through SSD have a speed faster than real-time, the real world need faster speed and less size.

Nowadays, several methods to compress and accelerate DNN models have been presented and gained good results. There are three principle directions among them, i.e. pruning[8], quantization[9] and distilling[10].//8th page

//TODO introduce what is pruning, quantization and distilling //9~11th page

This paper adopts the method in Deep Compress to compress and accelerate the SSD network.//TODO brief intro Deep Compression// 12th page

# SSD for pedestrian detection

Why SSD for pedestrian detection?

Advantage: high mAp in object detection. Rather high speed.

However,disadvantage: or not meet real requirement, ie need spped up.so compress:prune and quntization.

SSD is the state of art for object detection. However, is it good enough for just pedestrian detection? This chapter will explore it.

To adapt SSD to pedestrian detection, we prune the full connect layer to one label, i.e. pedestrian. Then we compare the accuracy and speed with original model and try to explain the reasons cause difference. Furthermore，we would find some solutions to enhance the benefit brought by simple full connect layer.

# Deep Compression SSD

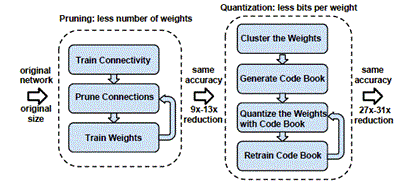


Figure 4 Deep Compression pipeline

To achive more efficiency, like faster speed and less space，we apply two measures on SSD. Firstly, we prune weights below a threshold and retrain the model. This process will recycle several times until the sparsity reach our requirement. So we could get a less weights model. And Less weights result less inference time and less size. Secondly , we quantization the weights from 32bit float to 8bit integer. In the inference process, most operation is matrix computation. So this step could further reduce inference time by reducing matrix computation time due to 8bit integer computation consumes less time than 32 bit float in computer.

# Conclusion

We made a improve on SSD by applying Deep Compression and cutting detection labels. The Deep Compression SSD is faster, less size and less energy consuming without sharp accuracy drop. In other word, the model becomes more efficient. In fact, this method is also suitable for other DNN object detection models and other specific labels. In the future, we would consider apply cascade strategy to the model. So the easy instance could be detected earlier and reduce the whole time for plenty instances tasks.

# References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection.  
CVPR,2005.

[2] Corinna Cortes, Vladimir Vapnik, Support-Vector Networks[J], Machine Learning, 1995,20(3),273-297.

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, ImageNet Classification with Deep Convolutional Neural Networks,NIPS,2012

[4] Karen Simonyan, Andrew Zisserman,Very Deep Convolutional Networks for Large-Scale Image Recognition,ICLR,2015.

[5] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. (2015)

[6] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. (2016)

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single Shot MultiBox Detector. ECCV,2016.

[8] Song Han, Huizi Mao, William J Dally, Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, ICLR,2016.

[9] Jiaxiang Wu, C. Leng, Y. Wang, Q. Hu, J. Cheng, Quantized Convolutional Neural Networks for Mobile Devices, CVPR, 2016.

[10] G. E. Hinton, O. Vinyals, J. Dean, "Distilling the knowledge in a neural network", NIPS Deep Learning Workshop, 2014.