Project4 Design

Application of Neural Networks for Object Detection

Team members：

11711416 韩铭基

11712510 李浩南

11711335 张艺凡

Introduction

Object detection deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Methods for object detection generally fall into either machine learning-based approaches or deep learning-based approaches. We mainly look at the deep learning approaches, includes but not restricted to: Region Proposals, Single Shot MultiBox Detector, YOLO, Single-Shot Refinement Neural Network for Object Detection. The following sections introduces our main model candidates, data sets performance goal and possible application.

Data Selection

VOC

The project is required to test on The PASCAL Visual Object Classes Chanllenge 2007, so we our test data includes the VOC2007 official test set. Also possibly we would use the PASCAL VOC 2012 official test set.

Common Objects in COntext

COCO was one of the first large scale datasets to annotate objects with more than just bounding boxes, and it became a popular benchmark to use when testing out new detection models.

Visual Genome

Visual Genome is a dataset, a knowledge base, and it can connect structured image concepts to language. It has 108077 photos, 5.4 million region descriptions, 2.3 million relationships etc. and its labels includes objects, attributes and relationships whithin the photo.

Model Selection

Convolutional Neural Networks

Convolutional networks are a class of deep neural networks, they are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers. CNNs are most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks, based on their shared-weights architecture and translation invariance characteristics.

A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of a series of convolutional layers that convolve with a multiplication or other dot product. The activation function is commonly a RELU layer, and is subsequently followed by additional convolutions such as pooling layers, fully connected layers and normalization layers, referred to as hidden layers because their inputs and outputs are masked by the activation function and final convolution. The final convolution, in turn, often involves backpropagation in order to more accurately weight the end product.

Region-based Convolutional Neural Network

The region-based convolutional network is one of the main contemporary approach to object detection. With R-CNN, Features are then extracted with convolutional neural network from these proposals and classified with SVM classifier.

Region-based Convolutional Neural Network mainly contains three steps:

1. CNN is pre-trained on ImageNet for image classification. It scans input image and

generates about two thousands regions of interest from a Selective Search algorithm. Such methods can provide high recall rate but very low precision. Thus they cannot be used as object detectors themselves, but they can be used as a first step in the detection pipeline.
2. Forwards every region proposal through Convolutional Neural Networks.CNN is fine-tuned for object detection on limited object detection data set.
3. Linear classifier and bounding box regressors are trained on top of CNN features extracted from object proposals.

Experimental evolution has demonstrated that R-CNN outperforms previous object detection methods, even if only pre-treated CNN is used for feature extraction. Fine-tuning of convolutional neural network on turgid data set and application of bounding box regression, both improves the performance significantly. One very important advantage of R-CNN method is that it can use any convolutional neural network for feature extraction. You can improve the performance of R-CNN by using more advanced convolutional architecture.

Fast R-CNN

R-CNN was a major breakthrough, combining region proposals with a CNN. However, it had several problems: it was slow, hard to train and has large memory requirement. So Fast R-CNN was introduced, by combining the three different parts in the R-CNN system: CNN, SVM, Bounding Box Regressor into one architecture, it tired to solve the above problems of R-CNN.

Comparing to Region-based Convolutional Neural Network, Fast R-CNN has a tremendous improvement in detection speed, mainly because of two enhancement methods:

1. It first Process the whole image with the CNN instead of the 2000+ CNNs of R-CNN. The result is a feature map of the image.
2. Instead of SVMs, it uses a softmax classifier for future prediction.

Performance Goal

Mean of Average Precision

We anticipated that when our neutral network model is trained on VOC2007, it can achieve a mAP value on the official test set higher than 55. Hopefully it can get a mAP value higher than 70.

Detection Speed

	R-CNN	Fast R-CNN
Training Time	84 hours	9.5 hours
Speedup	1x	8.8x
Test time per image	50 seconds	2 seconds
Speedup	1x	25x
mAP(VOC 2007)	66.0	66.9

According to the table, Fast R-CNN has a 25 times speed up comparing to R-CNN, so it has a huge advantage when it comes to the detection speed.

Memory Efficiency

In R-CNN, we have to save every feature map of each region proposal and it needs a lot of memory. But in Fast R-CNN, we don't have to. Since it process the whole image with the CNN, we only need to save this one feature map of the whole image.

Provided functions

Object Detection

We provide a neural network model that aims to detect instances of semantic objects of certain classes (such as humans, buildings, or cars) in digital images and videos. After processing the input data, it can recognize objects such as specific animals, humans, airplanes, cars, bicycles, and so on, and the bounding boxes and specific labels would appear around these objects. When trained on VOC2007, the mAP value on the official test set can be higher than 55.

Application of Object detection in traffic

We found that object detection can be well applied to public traffic. For example, there are a lot of traffic accidents every day due to the lack of attention of both drivers and pedestrians. If we apply this object detection technology to cameras of crossroads, it can detect whether there are cars or oncoming cars on the zebra crossing, and can send warnings to pedestrians. This can primarily benefit the disabled like the blind. Before, they could at best gain information about the color of the traffic lights by particular sound, with zero knowledge about the situation on the road. But now they can be significantly benefited by object detection and have a clear idea of whether it's safe for them to cross the street now because it can be dangerous even when it's green light without additional information. Object detection can also give information about the flow of traffic.

References

[1] https://towardsdatascience.com/part-2-fast-r-cnn-object-detection-7303e1988464

[2] https://github.com/jwyang/faster-rcnn.pytorch