Assignment Task 1: Project Requirements and Specification

Project Direction: Object Detection

Team 4

Weilin Sun 13462383
Yangyang Jin 13647716
Junting Song 13833936
Qibai Chen 13757467
Tianyang Zhang 13653112

Table of Contents

1. Analysis of the problem & Availability of datasets

Object detection refers to using the theory and method in the fields of image processing and pattern recognition to detect target objects in an image. It can determine the category of these target objects and mark the position of the target object in the image. Object detection is the prerequisite for object recognition. Only when the object is detected can the object be recognized.

The input of object detection is an image or video sequence frame, and the underlying features of the image can be described by a feature extraction algorithm. Select representative features from feature vectors and reduce the dimensionality of features. Then, a specific classifier is used to classify the features and determine whether the candidate area contains the target and its category. Finally, the intersecting candidate areas of the same category are judged, and the bounding box of each target is calculated to complete the object detection.

Object detection has a wide range of applications in the fields of face recognition, medical imaging, intelligent video surveillance, robot navigation, image editing, and augmented reality. Using theories and methods in the fields of image processing and pattern recognition, we can separate meaningful entities such as people, cars, buildings and other objects from images or videos.

In this project, the dataset we selected is a video sequence frame on a highway (refer to Figure 1). The data set includes a training set and a test set. The training set has 1000 sequence frames, and the test set has 176 sequence frames. It can detect whether there is a vehicle in the image and mark the detected vehicle.

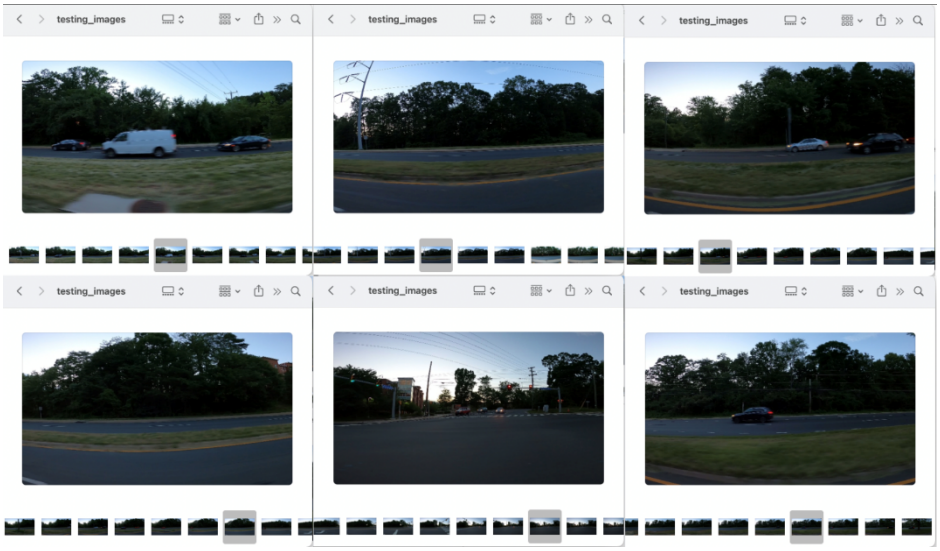

Figure 1 # A screenshot from the testing_image

In this dataset, there are also some variability and difficulty of the dataset. For example, the weather is approaching the evening, which causes the image to be too dark, and the excessively fast vehicle speed increases the blurring of the edges of the vehicle, which may reduce the accuracy of vehicle recognition.

2. Technology that is expected to solve this problem

For the object detection problem, we searched two solutions, RCNN and YOLO, the specific description is as follows:

2.1 R-CNN

R-CNNs (Region-based CNN) is a kind of machine learning algorithm used in computer vision, especially object detection. It has been developed mainly based on linear regression, and support vector machines (SVM), especially on convolutional neural networks (CNN).

A convolutional neural network (CNN) is a type of artificial neural network that has been widely used in deep learning for image processing, classification, segmentation, and other auto-correlated data (Saha, 2018). Thomas (2019) stated that CNN can be simply considered as sliding a filter across the input which can be looking at the surrounds of a function to create better/accurate predictions of its outcome. It may be more efficient to look at tiny sections of an image to discover certain traits than looking at the full image at once. By using the framework of CNN, the RCNN has been derived and first launched in 2013, it contains four steps in working progress which has been displayed below.

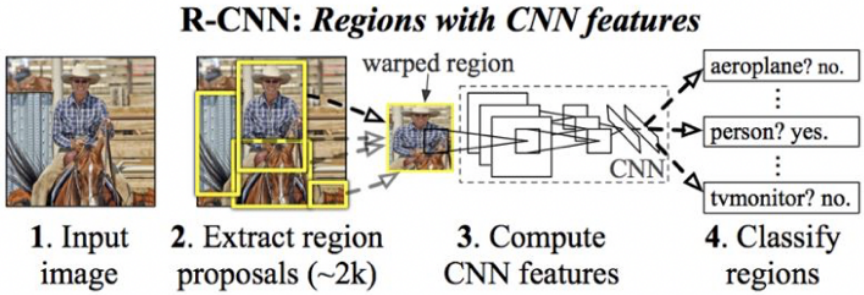

Figure 2 # R-CNN Progress

Based on a selective search strategy, the R-CNN yields around 2000 candidate areas. After that, each candidate region is shrunk to a predetermined size and put into a CNN model, yielding a feature vector. This feature vector is then input into a multi-class SVM classifier, which forecasts the likelihood that the items in the candidate areas belong to each class. For each class, one SVM classifier is trained, and the feature vector is used to infer the likely magnitude of belonging to that class. Finally, the R-CNN trained another bounding box regression model to compensate

for the precise position of the box by the border regression model to enhance localisation accuracy.
Moreover, RCNN has several versions. In this case, Fast RCNN and Faster RCNN are used widely in the objective detection field.

- Fast RCNN: The Fast RCNN can be explained as a faster object detection algorithm than the normal RCNN which is not required to input the CNN with 2000 area suggestions every time. By contrast, each picture is convolved once only, and a feature map is created as a result (Gandhi, 2018).
- Faster RCNN: Compare with Fast RCNN and Faster RCNN, the separate network has been used to predict the region proposals by the latter which will receive a higher efficiency, which makes Faster RCNN ca even be used for real-time object detection.

2.2 YOLO

YOLO (You Only Look Once) is a real-time object recognition and location algorithm that based on deep convolutional neural network. The features of it are the high speed and real-time prediction. It is published by Joseph Redmon et al in 2015 (ODSC, 2018). This part will explain three versions of YOLO and the process of YOLOv3.

Three versions of YOLO:

- YOLOv1: It only uses a single convolutional neural network to achieve the purpose of detecting object end to end. The backbone of YOLOv1 is modelled on GoogLeNet. The output layer is a fully connected layer.
- YOLOv2: Compare with YOLOv1, YOLOv2 had used DarkNet-19 as the backbone the output layer is changed to fully convolutional layers. Then, removing the dropout layer, YOLOv2 introduced batch normalization to increase the mean average precision of the network. It had fixed the problems of the detection of small objects in groups and the localization accuracy of YOLOv1.
- YOLOv3: Compare with YOLOv2, the main change is that YOLOv3 had used the DarkNet-53 as the model backbone. This backbone has 106 later neural network complete with residual blocks and upsampling networks. It makes that YOLOv3 can predict at three different scales. Therefore, YOLOv3 is easier to make the detection of smaller objects than YOLOv1 and YOLOv2.

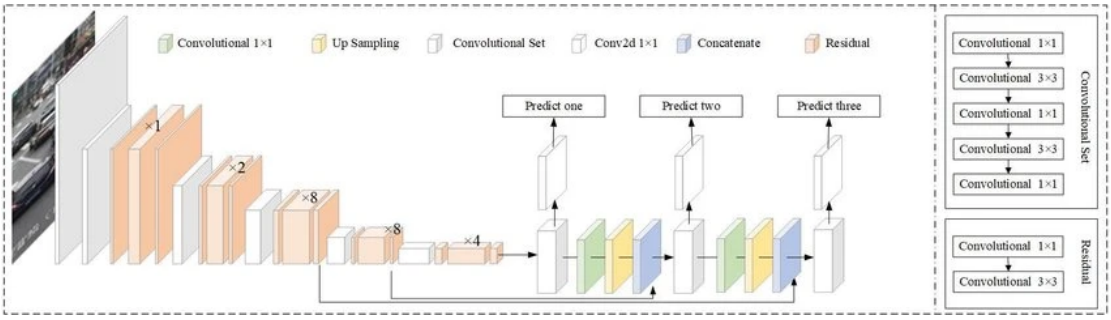

Figure 3 # The Structure detail of YOLOv3

YOLOv3 Working process:

Firstly, it will scale the input image to the standard size and divide the input image into 13*13, 26*26, 52*52 grids of three scales, predicting the object if the center point of an object falls in the grid unit. Using k -means clustering to determine the bounding box priors on each grid unit. Each grid unit has three clusters. The image had been dividing into three scales. Therefore, there are 9 clusters per grid unit into input image. Then, it will extract the feature from the input image through the network and produce a feature map of 13*13 on a small scale. this 13*13 small-scale feature map will be subjected to convolutional set and 2 times up-sampling, then connected to the 26*26 feature map and output the prediction result. After that, let 26*26 feature map output do same steps with the 13*13 small-scale feature map, then connected to the 52*52 feature map and output the prediction result. Last, fusing the features of three scale predictive outputs, using a probability score as a threshold to filter out most anchors with low scores. Then using Non-Maximum Suppression (NMS) for post-processing, providing more accurate results.

3. Motivation for the approach based on the knowledge acquired during lectures and independent readings

In our lecture, we learned some knowledge that can be applied to object detection technology, such as Convolutional Neural Network (which can efficiently find and identify objects in images, as well as non-image data such as signal data), Semantic Segmentation (intensive reasoning for each pixel, so that each object is marked with its category), Edge Detection (which can find the set of pixels with sharp changes of highlights in the image), ROI Detection Of Regions Of Interest (image processing by delineating the regions that need to be processed), Filtering Algorithm, filtering (by removing the pixels in the image that do not want to be processed, leaving only the pixels that do want to be processed), Hough transform (a very important method to detect the boundary shape of discontinuity points, by transforming the coordinate space of the image into parameter space to achieve line and curve fitting).

The project we chose was about the dataset on the highway, which could detect and tag the vehicle changes in the image. However, there are many factors that affect the data set. For example, in the evening, due to light problems, the image color is too dark. At the same time, the fast speed may make the vehicles in the image too fuzzy, which may lead to image recognition errors.

Due to the above possible influence on image clarity, we found that convolutional neural network is a better method for our project. First of all, convolutional neural network is a network architecture that can carry out deep learning directly from data without manual extraction. Secondly, CNN is good at identifying objects and scenes in images, and the self-driving function of Tesla that we are familiar with relies heavily on CNN. Through the tutor's explanation in class and independent reading after class, I found two object detection methods based on the convolutional neural network framework, R-CNN and YOLO.

4. Performance evaluation

For the performance evaluation of these two methods, specific experiments are analyzed from the following aspects:

- Accuracy
  Divide the number of pairs of samples at the time of detection by the total number of samples. The accuracy rate is generally used to evaluate the global accuracy of the detection model, and the information contained is limited, and the performance of a model cannot be fully evaluated.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision
  The precision is the ratio of true positives in the identified pictures.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{N}$$

- Recall
  Recall indicates how many positive examples in the sample are predicted correctly.

$$R = \frac{TP}{TP + FN}$$

.

- Detection speed
  It is how many pictures can be detected in one second. Different target detection technologies often have different mAP and detection speed.

R-CNN and YOLO are good object detection methods. According to Wang et al. (2019), R-CNN constructively neural the RPN structure. After the convolutional neural network, RPN is added as a branch network to realize the extraction of the anchor box and merge it into the deep network. R-CNN proposed RPN as a region selection network. It still stays in the "two-step" thinking and still precisely reflects the region proposal process, even if it realizes the function of a neural network to select the monitoring area.

For YOLO, through continuous improvement and upgrading YOLO-VI and YOLO-V2, the current YOLO-V3 network is more advanced than the previous two generations YOLO-V1 directly regresses the position and category of the bounding box in the output layer, making the detection speed very fast. However, the accuracy of identifying the position of the object is not high. When each grid contains multiple objects, only one of them can be detected. In order to improve the accuracy of object positioning, YOLO-V2 introduces the idea of "anchor box" in R-CNN, and uses k-means clustering algorithm to generate appropriate prior bounding boxes. In addition, YOLO-V2 improves the design of the network structure, using a convolutional layer in the output layer to replace the fully connected layer of YOLO-V1. YOLO-V3 uses multi-scale prediction to detect the final object based on YOLO-V2. The network structure is more complex and the detection performance is stronger.

To sum up, it seems that R-CNN has better performance in detection accuracy, while the detection speed of the YOLO-V3 model is faster.

5. Reference

Bandyopadhyay, H. (n.d). *YOLO: Real-Time Object Detection Explained.* https://www.v7labs.com/blog/yolo-object-detection#yolo-versions

Gandhi, R. (2018, July 10). *R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms.* towardsdatascience. https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e

ODSC. (2018). *Overview of the YOLO Object Detection Algorithm.* https://medium.com/@ODSC/overview-of-the-yolo-object-detection-algorithm-7b52a745d3e0

Q. -C. Mao, H. -M. Sun, Y. -B. Liu and R. -S. Jia. (2019). *Mini-YOLOv3: Real-Time Object Detector for Embedded Applications.* https://ieeexplore.ieee.org/document/8839032

Sumit, S. (2018, Dec 16). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. towardsdatascience. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Thomas, C. (2019, May 27). *An introduction to Convolutional Neural Networks.* towardsdatascience.

https://towardsdatascience.com/an-introduction-to-convolutional-neural-networks-eb0b60b58fd7

Wang, H., Yang, G., Li, E., Tian, Y., Zhao, M., & Liang, Z. (2019). *High-Voltage Power Transmission Tower Detection Based on Faster R-CNN and YOLO-V3.* 2019 Chinese Control Conference (CCC), 8750–8755. https://doi.org/10.23919/ChiCC.2019.8866322