Machine Learning Engineer Nanodegree

Object Detection on Command

Seyeon Lee
July 24, 2019

Proposal

Domain Background

Computer Vision is one of the greatest applications of machine learning. Convolutional Neural Network, also known as CNN, is a state-of-the-art algorithm for computer vision that was inspired by animals' visual cortex. CNN works with image pixels and the data they inherit. By detecting edges, depth, colors, and many other features of images, CNN can identify shapes and motions.

Object detection is the technology that detects objects, classify, and locate them in an image. Many ideas have been proposed for this task. Region proposals is a method that allows CNN to detect objects among many cropped regions with help of selective search (Uijlings et al, "Selective Search for Object Recognition", IJCV 2013)[1]. Models like RCNN (Region-based Convolutional Neural Networks)[2] and YOLO (You Only Look Once)[3] make use of these algorithms and accomplish high accuracy on object detection tasks.

Problem Statement

The goal is to build a program that intakes users' real-time video data from the webcam on a user's computer and do the following tasks:

1. Identify commands like "zoom in on" and "zoom out" along with the labels of the objects in the PASCAL VOC (Visual Object Classes) dataset.
2. Find the objects in the video and follow the according command: Zooming in on the object ("zoom in" command), Zooming out

This is an object detection task. It will classify 20 objects from PASCAL VOC datset: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining

[1] J. R. Uijlings , K. E. Sande , T. Gevers , A. W. Smeulders, Selective Search for Object Recognition, International Journal of Computer Vision, v.104 n.2, p.154-171, September 2013 [doi>10.1007/s11263-013-0620-5]

[2] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2014.81

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.91

table, potted plant, sofa, and tv/monitor. Then, the bounding boxes of each object in a test image will be predicted and output.

Datasets and Inputs

For the real-time recognition of humans and objects, PASCAL VOC (Visual Object Classes) 2012 dataset (source link: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/) will be in use. This dataset contains large number of images for classification and detection task with 20 different classes: bird, cat, cow, dog, horse, sheep, airplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv/monitor, and person.

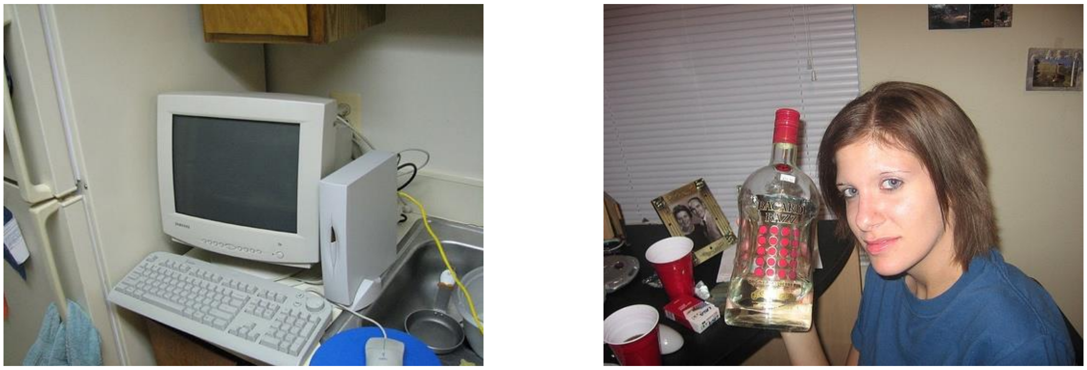


Figure 1 Images from the PASCAL VOC 2012 dataset

Fundamentally a supervised-learning, the training set of labelled images is provided. Below is the table that includes statistics of the image sets for each 20 classes. You can see it is an imbalanced dataset.

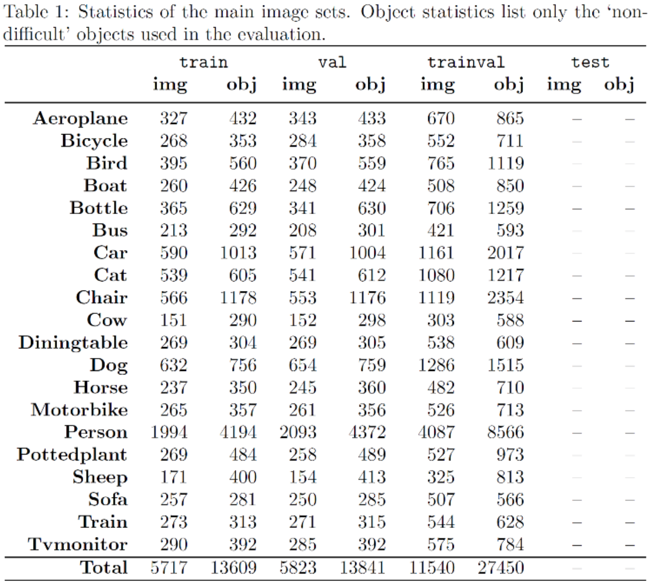


Table 1: Statistics of the main image sets. Object statistics list only the 'non-difficult' objects used in the evaluation.

	train		val		trainval		test	
	img	obj	img	obj	img	obj	img	obj
Aeroplane	327	432	343	433	670	865	–	–
Bicycle	268	353	284	358	552	711	–	–
Bird	395	560	370	559	765	1119	–	–
Boat	260	426	248	424	508	850	–	–
Bottle	365	629	341	630	706	1259	–	–
Bus	213	292	208	301	421	593	–	–
Car	590	1013	571	1004	1161	2017	–	–
Cat	539	605	541	612	1080	1217	–	–
Chair	566	1178	553	1176	1119	2354	–	–
Cow	151	290	152	298	303	588	–	–
Diningtable	269	304	269	305	538	609	–	–
Dog	632	756	654	759	1286	1515	–	–
Horse	237	350	245	360	482	710	–	–
Motorbike	265	357	261	356	526	713	–	–
Person	1994	4194	2093	4372	4087	8566	–	–
Pottedplant	269	484	258	489	527	973	–	–
Sheep	171	400	154	413	325	813	–	–
Sofa	257	281	250	285	507	566	–	–
Train	273	313	271	315	544	628	–	–
Tvmonitor	290	392	285	392	575	784	–	–
Total	5717	13609	5823	13841	11540	27450	–	–

The annotations for each image are consisted of the following fields:

- width: width of the frame
- height: height of the frame
- Nobj : The number of objects in the frame
- fileID: The png file name
- For each frame, there are at most 56 objects in one frame. The infomation of the ith object is:
  - bbx_i_nm: The type of object inside of the bounding box i e.g.,
  - bbx_i_xmin: The x coordinate of the minimum corner in bounding box i
  - bbx_i_ymin: The y coordinate of the minimum corner in bounding box i
  - bbx_i_xmax: The x coordinate of the maximum corner in bounding box i
  - bbx_i_ymax: The y coordinate of the maximum corner in bounding box i

Solution Statement

I would like to explore Convolutional Neural Networks (CNN) for object detection. I will use transfer learning with CNN, exploring models like ResNet50, VGG, Inception, Xception. By differentiating activation function, layer structure, and optimizer, I will observe how the model gets improved.

Benchmark Model

For the benchmark model, there is an example detector that is provided by the offical PASCAL VOC 2012 website, which is trained on PASCAL VOC 2012 training set. It is then applied to validation set, and the result is also provided in the development kit of the dataset.

"The file example detector.m contains a complete implementation of the detection task. For each VOC object class a simple (and not very successful!) detector is trained on the train set; the detector is then applied to the val set and the output saved to a results file in the format required by the challenge; a precision/recall curve is plotted and the 'average precision' (AP) measure displayed."

Table 1 Result for the benchmark model -- AP for each class (mean Average Precision (mAP) = 0.00105)

Class	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
AP	0.006	0.000	0.000	0.000	0.000	0.001	0.000	0.005	0.000	0.000
Class	Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tvmonitor
AP	0.001	0.003	0.001	0.001	0.002	0.000	0.000	0.001	0.001	0.000

Evaluation Metrics

Being an unbalanced supervised learning dataset, this object detection model can be measured in terms of precision, recall, and f1 score. Parameters are True Positives, False Positives, True Negatives, and False Negatives.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

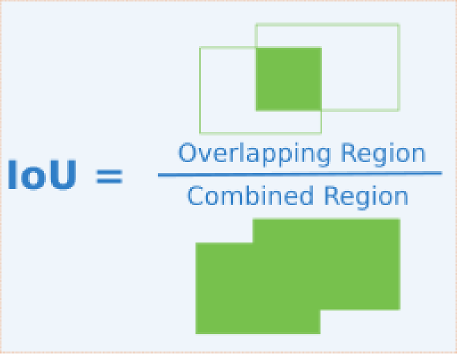
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1\ score = \frac{2 * precison * recall}{precison + recall}$$

PASCAL VOC documentation suggests to use mAP (mean Average Precision) method for object detection models.

> The AP summarizes the shape of the precision/recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels [0,0.1,...,1]

Intersection over Union (IoU) metric is a well-known metric system when it comes to evaluating the performance of bounding boxes. You can find the diagram below brought from this website (http://ronny.rest/tutorials/module/localization_001/iou/):



Project Design

The project will initially deal with object detection. Importation of datasets would be the first step to be made. The model will be trained on PASCAL VOC 2012, for the convenience in comparing to our benchmark models, and this will be done by downloading the dataset to my AWS GPU instance through this website: http://host.robots.ox.ac.uk/pascal/VOC/ . This is about 2GB. After downloading each training and testing dataset, these pictures will be preprocessed to gray scale images, each image sharing the same size.

To understand the data, preliminary analysis has to be preceded. Histograms of the number of objects per image will be drawn as below (Both Figure 2 and 3 below are from Yumi's Blog about Object Detection[4]):

[4] https://fairyonice.github.io/Object_detection_with_PASCAL_VOC2012_data_preparation_and_understanding.html

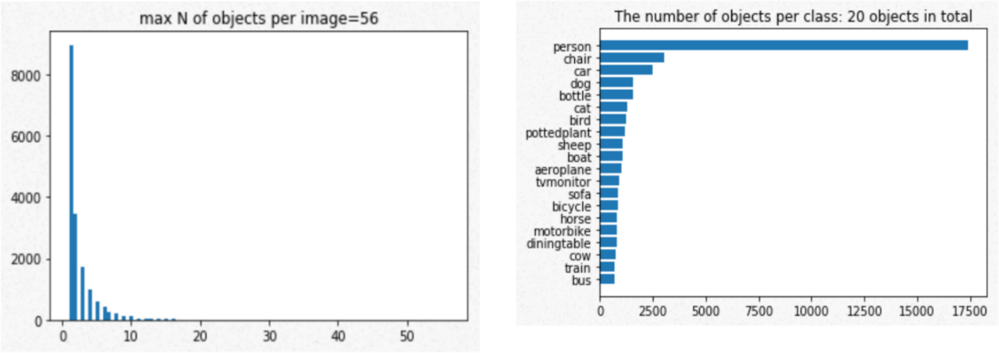


Figure 2 Histograms for the PASCAL VOC 2012 dataset

The Selective Search Algorithm, one of the most successful region proposal algorithms, will choose regions from multiple partitions of images (a.k.a. "initial regions") sorted out by fast segmentation method of Felzenswalb and Huttenlocher.[5] Then, it will group regions based on various criteria. After the search, selected regions have to undergo CNN feature extraction. The bounded images will be warped to match the CNN's input size.

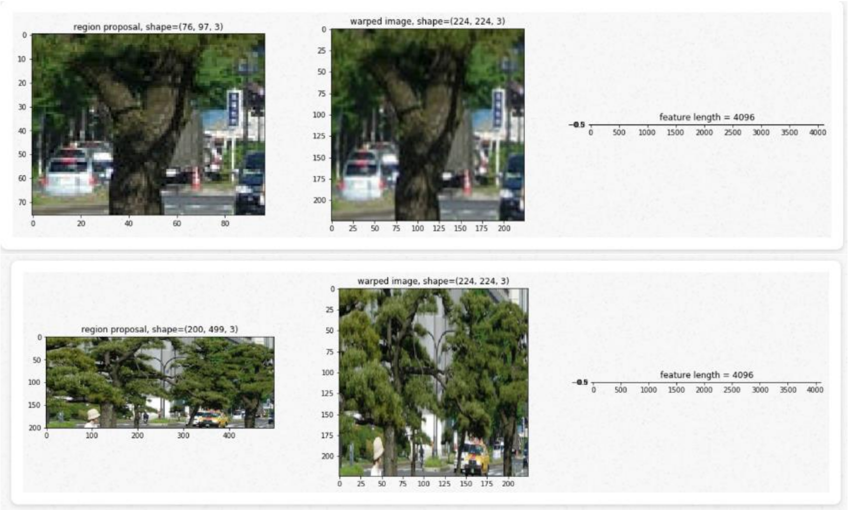

region proposal, shape=(76, 97, 3)
warped image, shape=(224, 224, 3)
feature length = 4096
region proposal, shape=(200, 499, 3)
warped image, shape=(224, 224, 3)
feature length = 4096

Figure 3 Warping the selected regions

The CNN model will be explored via the utilization of transfer learning on the basis of ResNet50, VGG, Inception, and Xception. For more accurate and precise model, experiments on different structures will be done. Each model will be evaluted based on the evaluation metrics mentioned above. Each model will be compared by the mAP value.

[5] Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision,59*(2), 167-181. doi:10.1023/b:visi.0000022288.19776.77