International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 1958–1966

Object Detection Related to Construction Activity using Deep Learning

Sudip Mahajan[1] , Geetanjali Kale[2], Adwait Bhave[3]

[1,2]*Department of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, India.*

[3]*Algoanalytics Private Limited Pune, India*

Abstract

Object detection problem has seen lot of applications recently with growing requirement to make detection real time. Present work focuses upon detection of objects related to construction activity that may be encountered unexpectedly on roads and cause traffic disruption. Recent developments in deep learning networks have made detection of complicated objects feasible within timing constraints suitable for Real time applications. This work uses Yolo algorithm in the form of darknet architecture to perform detection on a custom built dataset of traffic signs and cones. Yolo as an object detection algorithm has proved its applicability to real time tasks. The train and test data sets were compiled using standard open source data sets with labels in the form of bounding boxes around interesting objects. This work also tells importance of Yolo in video based object detection and highlights its limitations. The training and testing process resulted in average loss of 0.1010 with mAP stabilizing at 91.21%

Keywords: *Object Detection, Deep Learning, Video Analytics, Neural Network*

Introduction

There are wide varieties of civil activities happening on roads which cause disruption of traffic. They may include road construction, bridge construction, over-head bridge construction, building and repair work of monuments, work due earth-moving machinery etc. Such work should be constantly supervised to avoid accidents and troublesome effects on surrounding traffic. It is likely that such activities can lead to unwanted and unexpected obstacles of varying dimensions on the road. Keeping in mind public safety at large and with a view to help governing bodies in administering various regions, this project will serve helpful.

Object detection is an important problem of computer vision area. It involves locating semantically meaningful objects from an image. Localizing objects can be done in various ways including bounding box based methods. By this logic, object detection differs from allied problems like semantic segmentation, instance segmentation, classification and localization. Object detection has seen lot of activity throughout computer vision literature with modern approaches based upon Deep Learning algorithms. It is helpful for problems like image annotation, pedestrian detection, face detection and various surveillance objectives.

The future belongs to unmanned and autonomous vehicles such as self-driving cars, drones etc powered by technological advancements in the field of AI. Such systems are equipped with various sensors for better decision making and maneuvering ability through dense traffic and unfamiliar scenarios. Such sensors like cameras and optical recording devices can aid in detection of objects related to construction activity on road. They form eyes for concerned governing bodies and give insights into decision making and governance of region. Autonomous vehicle help to perform projects on massive scales and make

ISSN: 2233-7857 IJFGCN

large scale sensing possible. With the growth of processing power and availability of sensor data grows the need for better detection and processing algorithms making the work on such projects possible.

Object detection as a computer vision problem can be based upon images as well as videos. Videos are a collection of image frames recorded and played together at certain frame rates for end result of continuity. However, the challenges of processing videos are high due to high frames per second (fps) requirement. Video based object detection requires detection logic to be faster than average video frame rates of today's technology. There is growing requirement to make detection real time so as to extract maximum benefit from object detection decisions. Classical image based detection algorithms do not perform at the required speed with videos.

This work tries to apply object detection strategies based upon deep learning techniques to detect objects of particular interest. Generic object detection approaches do not readily apply to construction related objects. The data set is compiled using standard data sets and an implementation of deep architecture is chosen for experimentation. The data set contains traffic sign and traffic cone images due to scarcity of such material data sets. This work will highlight the importance and challenges that can be addressed by future works on this front even when an array of efforts have gone into object detection.

Literature Review

Several efforts have taken place to perform object detection from images. Recently many standard surveys have been published on the problem of Object Detection [11], [13], [25], [24], [19], [2]. Here, the backbone networks are responsible for success of object detection networks are also mentioned with object detection efforts. Broadly, object detection efforts can be divided into one stage as well as two stage approaches.

A. Backbone Networks

Backbone networks are responsible for robust feature extraction techniques that allow object detection systems to achieve competitive performance. They are primarily used for classification tasks with slight modifications leading to successful use in object detection tasks. Often last fully-connected layers are removed as a standard modification of such classification based networks to make them compliant for feature extraction. Some famous backbone networks are ResNet [8], ResNeXt [20], AmoebaNet [5], MobileNet [9], ShuffleNet [22], SqueezeNet [10], Xception [3].

B. Two Stage Detectors

This category of approaches contains networks that work upon generating region proposals. Traffic sign detection is an area with over two decades of efforts [15]. Recently, deep learning has been employed to address this problem as well. Two stage approaches make use of region proposals which are regions of interest where required objects can reside. Region proposals are first stage upon which object detection stage can be applied upon. R-CNN [7] (region based CNN) for object detection task which used the Selective Search (SS) algorithm and gave a performance of 66% on PASCAL VOC (2007) data set. This was a slow algorithm as the number of region proposals generated to classify was 2000. The same authors proposed Fast R-CNN [6] which used convolution operation to generate region proposals. Fast R-CNN gave mAP of 66.9% on PASCAL VOC (2007) data set. Yet, Fast R-CNN was a slow technique due to SS algorithm. Faster R-CNN [18] used separate network for generating region proposals and was a considerable improvement over previous techniques. Fast R-CNN presented mAP of 69.9% on PASCAL VOC 2007 data set with running time reduced to a tenth of the time for Fast R-CNN. The frame rate of Faster R-CNN was 5 fps. Yet these algorithms were slow to perform detection as required by video based frame rates.

ISSN: 2233-7857 IJFGCN

C. One Stage Detectors

Two stage object detectors were known to be slow to complex algorithms and stages in the process. Hence, one stage detectors were explored which performed entire object detection process (region proposals and detection) in one stage. These approaches are based upon regression/classification techniques on a grid of image. Initial Yolo algorithm was significant improvement over Fast R-CNN and Faster R-CNN with a frame rate of 45 fps and mAP of 63.4% on PASCAL VOC. Further improvements were required to the architecture to achieve better localization, higher accuracy and precision.
Yolov3 [17] (average precision of 33% on MS COCO data set), Yolov4 [1] were important developments towards real time object detection to suit video frame rates. Yolov3 gave rise to a new backbone architecture called as darknet, which evolved from 19 layers, to 56 and now 102 layers.

Aside from Yolo based works, several developments have taken place that are competitive in bounding box detection. Single Shot Multibox Detector (SSD) [14] algorithm detects pre-defined bounding boxes of different scales at each location corresponding to multiple categories. SSD showed reasonable performance with VGG16 backbone network giving mAP of 81.6% on PASCAL VOC 2007 and mAP 80% on PASCAl VOC 2012. Further developments on these lines are Deconvolutional SSD (DSSD) [4], RetinaNet [12] (gave average precision of 40.8% on MS COO test-dev data set with ResNeXt-101-FPN backbone), M2Det [23] (AP 44.2% using VGG-16 backbone on MS COCO), RefineDet [21].

In all above works, standard object detection datasets like MS-COCO, PASCAL-VOC 2007 & 2012 were used. They contain object categories like bird, bicycle, dog, aeroplane, thermometer, person, animals etc. There is lack of dedicated work on construction activity related object detection towards road safety.

3. Contributions

This work presents the following contributions:

a. Compilation of a data set suitable for construction activity related object detection
b. Training the darknet architecture (Yolov3) on the compiled data set.
c. Testing the architecture on a different data set than train data.

4. Data Set Overview

The details of the individual data sets used to compile together the large data set are presented in table I. These data sets were compiled into one data set with two category labels. The traffic sign images 1 are 'in the wild' images without any filtering effects applied to them. The images of traffic cones 2 have been processed to have a white background. All images are annotated (with bounding box coordinates) in Yolo format as required by darknet.

Dataset	Object	# objects	Comments
Kaggle Traffic Signs in Yolo format	Traffic Signs: prohibitory, danger, mandatory, other	More than 1000	All traffic signs considered as one class. The images are in non-standard conditions with random light effects.
Traffic Cone	Traffic Cones	600	Standard illumination condition , white background, pre-processing done

TABLE 1. Relevant Data Sets

5. Main Results

In this work, results are presented in the form of Mean Average Precision (mAP) on the compiled data set. Also, average loss according to training epochs is reported.

A. Experimental Setup

The brief methodology steps taken are presented in Fig. 1. Our experiments on the compiled data set reveal promise in applying deep learning techniques for object detection tasks. We used the Yolo darknet architecture [1] for this task. The implementation 3 used for this work is a variant of standard darknet implementation launched by original authors.

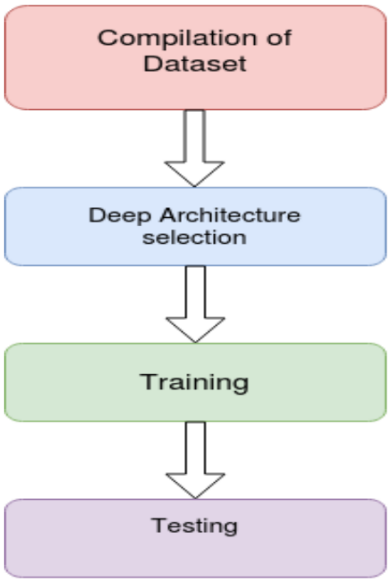

Fig. 1. Proposed methodology

The broad steps used to perform experimentation are as follows:

a. Search of available data sets on object detection related to civil activity on roads.
b. Two data sets chosen: traffic signs (from Kaggle) and traffic cones
c. Pre-processing data sets to generate only one data set with target labels (Bounding Boxes) in Yolo format.
d. Splitting of data set into train and test sets. The train set contains 913 images(692 traffic signal images and 221 traffic cone images) whereas test set contains 99 images (49 traffic signal images and 50 traffic cone images).
e. Preparation of Yolo configuration files for training phase.
f. Select pre-trained darknet architecture (Yolo) weights for object detection task. The starting point for weights of the network is the weights obtained over ImageNet data set.
g. Training on GPU machines for better training times. The GPU machine is Intel(R) Core(TM) i5-8400 CPU with 2.8 GHZ capacity. It is equipped with 6 CPU cores and 32GB main memory.
h. Testing phase on GPU machines to generate accuracy counts.
i. Prediction on unprocessed, ’in the wild’ video.

ISSN: 2233-7857 IJFGCN

B. Training and Testing

Training as well as testing was done on high-end general purpose GPU machine which have advantages over normal processors. The data set is divided into train and test sets before training begins. The ImageNet weights are used as pre-trained network weights to start the training process. The implementation dumps weights after regular intervals and average loss with mAP are computed on those dumped weights. Testing is done using best weights (according to average loss) found as a result of training phase.

C. Dataset Results

Results on the compiled dataset show converging trends in initial epochs. Table II shows the results during initial 10,000 epochs.

Epoch Number	MAP using stored weights (in %)
4000	85.57
5000	87.53
6000	87.34
7000	90.29
9000	89.36

TABLE 2. Training Progress

After considerable number of epochs the best **average loss** could be seen to converge at **0.101005**. The best weights found by the implementation gave **mAP of 91.21**%. Fig. 2, Fig. 3 and Fig. 4 are sample predictions generated by the best weights found by our algorithm.

Fig. 2. Sample prediction 1

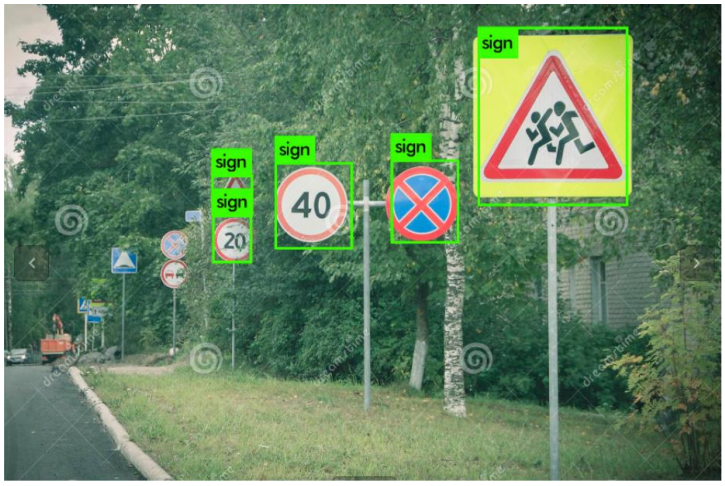

Fig. 3. Sample prediction 2

Fig. 4. Sample Prediction 3

6. Challenges

Object detection has been a long standing mountain in the field of computer vision. There are various aspects that change when moving from generic object detection to specific object detection. In regard to generic object detection, the challenges faced are similar to the following issues [13].

a. Many different instances in a class
b. Object instance diversities
c. Image noise
d. Inter-class ambiguities
e. Thousands of real-world object classes
f. Large scale image/video data

These challenges are applicable to present problem as well. Following are some challenges posed by the detection algorithm used.

a. Difficult to detect smaller objects due to spatial constraints of the algorithm. Object need to be of distinguishable size for Yolo to detect them. This is an inherent limitation for which features extracted from image need to be of robust nature. This can be observed in Fig. 3, the traffic signs further away from the observer are not detected by the algorithm.
b. Proximity (closeness) of objects makes it difficult for object detection. Yolo divides the image into grid cells and it requires that two objects be certain distance apart from each other. This limitation is observed in Fig. 2, only those cones are detected which are certain distance apart from each other whereas the cones close to each are not detected.
c. Yolo anchor boxes: Yolo does not handle the case like two anchor boxes (shape templates) and three objects inside one grid cell. In present work, anchor boxes were not used.
d. Static anchor boxes also present challenges like pre-defined scales and aspect ratios of objects to detect are made for specific training data sets. They do not perform well on further data sets.

Traffic sign detection has attracted lot of attention owing to self driving cars and autonomous vehicles. Detection of traffic signs/ traffic lights is full of different set of challenges
as follows:

a. Illumination variation
b. Motion Blur
c. Bad weather
d. Real-time detection

Object Detection problem is also prone to imbalance problems in data set. Some of the typical imbalances that need to be addressed are class, scale, spatial, objective imbalances as detailed by [16].

Conclusion and Future Work

The problem of object detection has seen many advances. It has varied applications one of which can be to aid safety of autonomous driving vehicles by detection of unfamiliar objects left on the road. In this project, a data set of traffic cones and signs was compiled with successful training of darknet architecture on the compiled data set. We hope that such studies will prove useful for decision making, policy making, safety audits of public infrastructure and overall better management of construction activities for public safety. In the future, dynamic anchor box generation to avoid problems related to pre-defining anchor boxes can be explored. Better feature extraction from images for detection of small and close
objects is required.

Acknowledgments

This work is supported by Algoanalytics Pvt. Ltd.

References

[1] Alexey Bochkovskiy, C. Wang and Hong-Yuan Mark Liao, “Yolov4: Optimal speed and accuracy of object detection”, arXiv preprint arXiv:2004.10934, 2020.
[2] Ali Borji, M. Cheng, Q. Hou“Salient object detection: A survey”, Computational Visual Media, vol. 5, no. 2,(2014),pp.117-150
[3] F. Chollet, “Xception: Deep learning with depthwise separable convolutions”, 2017 Proceedings of the 2017 IEEE conference on computer vision and pattern recognition, 1251-1258.

[4] C. Yang, W. Liu, A. Ranga, A. Tyagi, A. Berg, "Dssd: Deconvolutional single shot detector",2017 arXiv preprint arXiv:1701.06659,
[5] G. Ghiasi, T. Lin, Q. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection",2019 Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 7036-7045.
[6] R. Girshick, "Fast r-cnn", 2015 Proceedings of the IEEE international conference on computer vision, pp. 1440-1448.
[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", 2014 Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587.
[8] K. He, G. Gkioxari, P. Dollar, R. Girshick, " Mask r-cnn", 2017 Proceedings of the IEEE international conference on computer vision, pp. 2961-2969.
[9] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "Mobilenets: Efficient convolutional networks for mobile vision applications", 2017 arXiv preprint arXiv:1704.04861.
[10] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and o.5 mb model size",2016 arXive preprint arXiv:1602.07360.
[11] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, R. Qu, "A survey of deep learning-based object detection",2019 IEEE Access, 7:128837–128868.
[12] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal loss for dense object detection",2017 Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.
[13] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikainen, "Deep learning for generic object detection: A survey",2020 International journal of computer vision,128(2):261–318.
[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg, "Ssd: Single shot multibox
[15] detector",2016 In European conference on computer vision, pp. 21–37.
[16] A. Mogelmose, M. Manubhai Trivedi, T. Moeslund. "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey",2012 IEEE Transactions on Intelligent Transportation Systems, 13(4):1484–1497.
[17] K. Oksuz, B. Cam, S. Kalkan, E. Akbas, "Imbalance problems in object detection: A review",2020 IEEE Transactions on Pattern Analysis and Machine Intelligence.
[18] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement", 2018 arXiv preprint arXiv:1804.02767.
[19] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn:Towards real-time object detection with region proposal networks",2015 In Advances in neural information processing systems, pp. 91–99.
[20] X. Wu, D. Sahoo, S. Hoi, "Recent advances in deep learning for object detection",2020 Neurocomputing.
[21] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, "Aggregated residual transformations for deep neural networks", 2017 In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.
[22] S. Zhang, L. Wen, X. Bian, Z. Lei, S. Li, "Single-shot refinement neural network for object detection", 2018 In Proceedings of the IEEE conference on computer vision and pattern
[23] recognition, pp. 4203–4212.
[24] X. Zhang, X. Zhou, M. Lin, J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices", 2018 In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848–6856.
[25] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, H. Ling, "M2det: A single-shot object detector based on multilevel feature pyramid network",2019 In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9259–9266.
[26] Z. Zhao, P. Zheng, S. Xu, X. Wu, "Object detection with deep learning: A review", 2019 IEEE transactions on neural networks and learning systems, 30(11):3212–3232.

[27] Z. Zou, Z. Shi, Y. Guo, J. Ye, "Object detection in 20 years: A survey",2019 arXiv preprint arXiv:1905.05055.

ISSN: 2233-7857 IJFGCN