# A Survey on Deep Learning Models Based Road Object Detection Inference

Omar BOUAZIZI ENSA, Abdelmalek Essaadi
University Data Engineering & Systems TEAM Tangier
Morocco
{bouaziziom@gmail.com}

Aimad EL MOURABIT ENSA
Abdelmalek Essaadi University Data Engineering & Systems TEAM Tangier
Morocco
{elmourabit_aimad@yahoo.fr}

***ABSTRACT:*** *In this paper, we present a comparative study of the object detection accuracy and speed of various state-of-the-art models for the road scene context. Commensurate with the model training method, the algorithms can be divided into two types: one-stage models and two-stage models. We focused on the road context in order to detect all occurrences of objects in the road such as, car, person, traffic signs, etc. Accordingly, we find that one-stage detectors are stronger in terms of prediction speed, while two-stage models are stronger in terms of accuracy. To train deep neural networks with platform GPU type on a large amount of image data that required time, Because the computational cost of computer vision is very high, so we are focused to transfer learning technique, where a model trained on one task is reused on another related task, gives better results.*

**Keywords:** Object Detection, Deep Learning, CNN, Transfer Learning, GPU

**Received:** 30 June 2021, Revised 29 August 2021, Accepted 9 September 2021

**DOI:** 10.6025/dspaial/2022/1/1/37-41



## 1. Introduction

During the last recent years, ADAS systems come for reducing road accidents, so there are various systems have already been deployed in high-end vehicles. Therefore, the challenge is to detect all object principal in road with better precision and lower cost by using the new Deep Learning architectures. The image analyst is in charge of extracting information from the volume of data collected by the embedded cameras in the vehicle. So deep learning models, more specifically convolutional neural networks (CNN), are increasingly being used as the core enabling technology for detecting and classifying objects from within the images. In this paper, we compare the detection accuracy and speed measurements of various models by using the transfer learning.

## 2. Related Work

Object detection plays a very important role in computer vision, namely that it has a distinctly big challenge in detecting, locating and estimating the class. More and more the object detection process has evolved in a very practical and efficient way through the application of convolutional neural networks (CNN). We have focused on studying and analyzing these object detection algorithms and their application to the road context in order to detect cars, pedestrians and traffic signs that are the objects of interest via the vehicle's front camera[1].

Several researchers have designed and developed robust object detection algorithms thanks to advances in technology and the availability of very powerful graphics processing units (GPUs), these algorithms can be divided in two categories: one-stage and two-stage detectors. For the two-stage models that detect firstly regions of the images where the object might be present and then apply a classifier to these regions, while one-stage models provide an estimate of the position and classes in one step.

Scientists, researchers and experts have proposed and developed object detectors. Therefore, we have a review for both types of these algorithms.

### 2.1. The Convolutional Neural Networks

The methods of detecting objects on the road have been studied for several years. While the object detection mechanism has evolved by applying convolutional neural network while improving performance. Therefore, the key point for all detection algorithms is the CNN structure which is based on four large parts : The 1st is the input part, followed by the feature extraction part which constitutes various layers of convolution and pooling, the 3rd part is dedicated to the classification based on Fully connected layers then the last layer is the output layer [2].

### 2.2. The Two-stage Object Detection Models

• **R-CNN:** In 2014, Ross Girshick proposed the R-CNN algorithm, which is the first real target detection model based on convolutional neural networks specifically on Selective region. R-CNN [3] achieves a mean average precision (mAP) of 53.3% on PASCAL COV [3]. Convolutional neural networks based on the selective search for regions of objects and then classification are performed. Compared with the traditional detection method, the accuracy of R-CNN [4] does improve a lot, but the amount of calculation is very large, and the efficiency of the calculation is too low. Secondly, direct scaling of the region proposal to a fixed-length feature vector may cause object distortion, which leads to an extremely slow detection speed which is 14 s per image with GPU [5].

• **SPPNet:** In 2015, K. He et al proposed the Spatial Pyramid Pooling (SPP) model which solves the problems of fixed input size image in R-CNN and low detection efficiency, The Spp-Net algorithm [3], extracts at the same time the features of the regions proposal of the input image which has passed through the convolution layer and doing all the convolution calculations, then after the last convolutional layer the addition of the FC layer which will be given the feature vector of fixed size. This algorithm is 20 times more efficient and faster than the R-CNN. SPP-net achieved an average mean precision (mAP) of 59.2% on the COV-2007 [6].

• **Fast R-CNN:** In 2015, R. Girshick et al [3] proposed the Fast R-CNN model which is an improvement of SPPNet [6] and R-CNN. Compared to the previous models, Fast R-CNN has made changes, it replaced the SVM of the classification with the softmax function, then the change of the last pooling layer in the convolutional by the layer of the region of interest pooling(RoI). In order to transform the feature from the ready box into a feature map with with fixed size for access the full connection layer. Finally, the replacement of the last classification layer by two parallel FC layers. However, Fast R CNN achieves 70.0% mAP accuracy in VOC2007 and VOC2012 [3].

• **Faster R-CNN:** In 2016, S. Ren et al proposed a new method for object detection, so Faster R-CNN [4] introduced a layer called RPN (Network Proposal Region) instead of using selective search, this new model is divided in two parts, one of which is an entire CNN block used to generate RPN and the other is the Fast R-CNN algorithm. However, there is a sharing of layers between these two phases. This algorithm achieves a mAP greater than 70% on VOC2007 and VOC2012 [7].

• **Mask R-CNN:** In 2017, K. He et al. Developed the Mask R-CNN model, which works for the detection of bounding boxes, and It produces three outputs which are an object mask, a bounding box coordinates and a class label. So Mask R-CNN generates a segmentation mask and simultaneously detects objects more efficiently in the input image or video. Also to get the best results in terms of accuracy and speed, Mask R-CNN uses ResNet-FPN [6] as the base model for feature extraction despite a computational overhead problem and attains a speed of nearly 5 Fps.

### 2.3. The one-stage object detection models

• **YOLO:** In 2015, R. Joseph and al. developed a framework dedicated to detection called YOLO [6] and which supports real-time predictions. The basic idea of YOLO is totally different compared to other algorithms. YOLO applies a single CNN to the entire

image, the latter is divided into regions, and for each region it predicts bounding boxes and class probabilities. Therefore, the confidence score is defined by the product of the detection probability and the value of IoU. Consequently, YOLO is extremely fast and achieves more than 53% of mAP on the VOC dataset, Finally the improved versions (YOLOv2, YOLOv3, YOLOv4) work at 45 fps. YOLO has great difficulty in processing small objects in groups, which is imposed by the spatial constraints (predictions, the bounding box) [5].

• **SSD:** In 2016, Liu et al proposed the SSD (Single Shot MultiBox Detector) model [3]. This algorithm uses the idea of regression like YOLO and is inspired by the concept of the anchor box to improve the effect of multi-scale object detection. The network merges the predictions of several feature maps with different resolutions for managing the different sizes of objects. VGG16 is the backbone of the SSD architecture with the replacement of the last two FC layers by convolution layers [6] . The SSD achieves 76.8% mAP on the VOC dataset and at 63 FPS on Nvidia Quadro RTX 8000. However, the main difference between SSD and the other older detectors is that SSD execute detection tests just on the deeper layers while the older algorithm execute detection test on different layers [8].

• **RetinaNet**: In 2017, Lin et al. implemented another one-stage model called RetinaNet [7], the introduction of this model comes to overcome the problem of the SSD algorithm which creates a class imbalance as suddenly the object classes present in these locations are not detected. RetinaNet uses a new loss function called Focal Loss to remove the gradients of negative samples instead of rejecting them. This function achieves an accuracy of 59.1% mAP on the MSCOCO dataset.

### 2.4. Kitti Dataset

To train the SSD-MobileNet algorithm a labeled data set is required. There are many road context datasets available online that can be used for training and testing, such as the COCO dataset [9], Pascal VOC [10], Cityscapes [11], Berkeley DeepDrive [12] and the 'google open image dataset [13].

However, Kitti's dataset [14] which consists of 7,481 training images with seven classes labeled: cars, vans, streetcars, trucks, pedestrians, seated people, and cyclists.

## 3. Results and Discussion

In this section, we worked with Transfer Learning [15] which consists of using the knowledge acquired during a task before solving and improving the learning of a new task. Therefore, the Transfer Learning is used to remove the fully connected layers from the pre-trained model and just keep the convolution blocks and add a new fully connected layer, namely just the upper layer parameters that are being trained[16].

To train the SSD-MobileNet [17] model on the KITTI dataset [18], we follow the steps below:

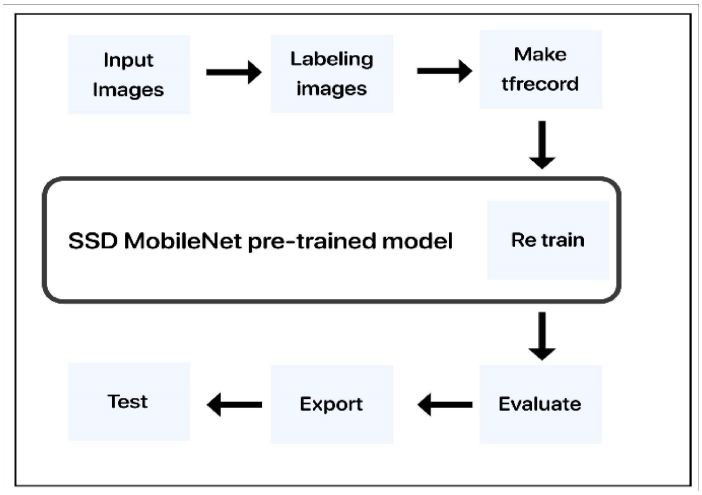


Figure 1. The steps for detecting objects using SSD MobileNet

After download the KITTI dataset, we move 20% of the images to the test directory, and 80% to the train directory. Then we labeled manually the desired objects in every picture, and by the use of the Nvidia Quadro RTX 8000 [19] graphics card and the Machine Learning Library particularly Tensorflow [20], we have trained the object detection model, the inference of the model on an image us follow:

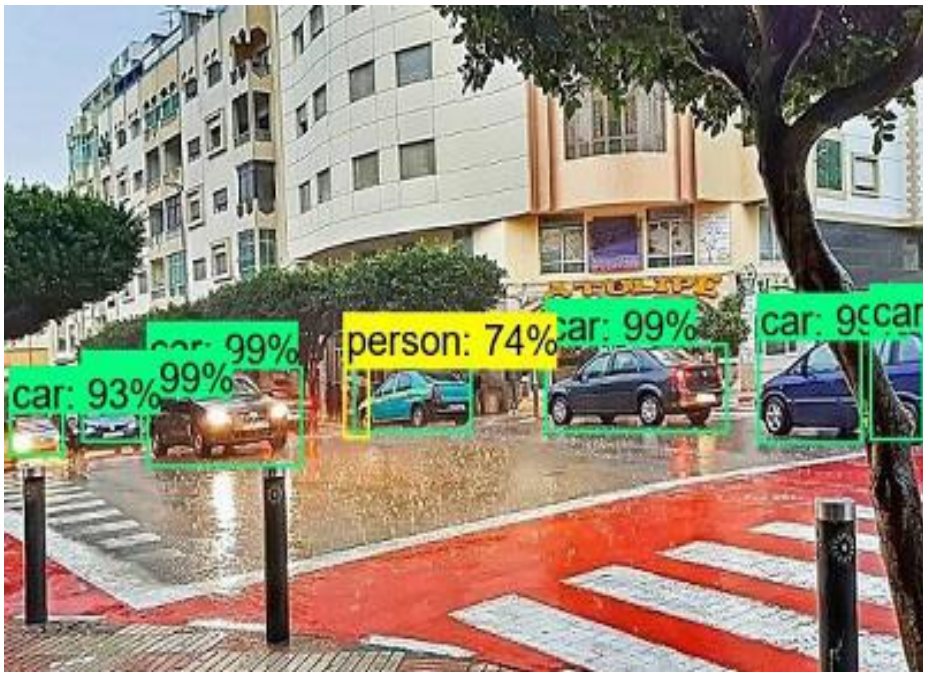


Figure 2. SSD MobileNet Inference Result

As it is seen in Figure 2, the model names the objects and indicates the probability that this object is correctly detected and recognized.

## 4. Conclusion

In recent years, object detection based on deep learning has received a great deal of attention. It is very difficult to have a fair comparison between the different object detectors. Because there is no direct answer to the question of which is the best model for road scene applications, we make choices to balance accuracy and speed. This article provides a summary of object detection models as well as a detection experiment in the road environment using SSD-MobileNet on the KITTI dataset based on a GPU platform. We have achieved better performance, but the notion of real time is even further. Finally, we point out that this area of research is still active and it opens the door to several future directions.

## Acknowledgement

We acknowledge financial support for this research from the National Scientific Research and Technology Center (CNRST), Morocco.

## References

[1] (PDF) Review of advanced driver assistance systems (ADAS). https://www.researchgate.net/publication/321364551_Review_of_advanced_driver_assistance_systems_ADAS (accessed Aug. 25, 2021).

[2] Prabhu. (2020). Understanding of Convolutional Neural Network (CNN) — Deep Learning, Medium, Nov. 21, 2019. https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148 (accessed Oct. 10, 2020).

[3] Deng, J., Xuan, X., Wang, W., Li, Z., Yao, H., Wang, Z. (2020). A review of research on object detection based on deep learning, *J. Phys*. Conf. Ser., 1684, 012028, November 2020, doi: 10.1088/1742-6596/1684/1/012028.

[4] Mittal, U., Srivastava, S., Chawla, P. (2019). Review of different techniques for object detection using deep learning, *In*: Proceedings of the Third International Conference on Advanced Informatics for Computing Research, New York, NY, USA, June 2019, 1–8. doi: 10.1145/3339311.3339357.

[5] A comprehensive and systematic look up into deep learning based object detection techniques: A review - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S1574013720304019 (accessed Jun. 20, 2021).

[6] Murthy, C. B., Hashmi, M. F., Bokde, N. D., Geem, Z. W. (2020). Investigations of Object Detection in Images/Videos Using Various Deep Learning Techniques and Embedded Platforms—A Comprehensive Review, *Applied Sciences*, 10 (9), 9, January 2020, doi: 10.3390/app10093280.

[7] Groener, A., Chern, G., Pritt, M. (2019). A Comparison of Deep Learning Object Detection Models for Satellite Imagery," 2019 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–10, October 2019, doi: 10.1109/AIPR47015.2019.9174593.

[8] Liu, W., et al. (2016). SSD: Single Shot MultiBox Detector, in Computer Vision – ECCV 2016, Cham, 2016, 21–37. doi: 10.1007/978-3-319-46448-0_2.

[9] Lin, T.-Y., et al. (2021). Microsoft COCO: Common Objects in Context," arXiv:1405.0312 [cs], Feb. 2015, Accessed: Aug. 25, 2021. [Online]. Available: http://arxiv.org/abs/1405.0312

[10] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge, *Int J Comput Vis*, 88 (2) 303–338, June, doi: 10.1007/s11263-009-0275-4.

[11] Cordts, M. et al. (2020). The Cityscapes Dataset for Semantic Urban Scene Understanding," arXiv:1604.01685 [cs], Apr. 2016, Accessed: October 10. [Online]. Available: http://arxiv.org/abs/1604.01685

[12] BDD100K: A Large-scale Diverse Driving Video Database – The Berkeley Artificial Intelligence Research Blog. https://bair.berkeley.edu/blog/2018/05/30/bdd/ (accessed Aug. 25, 2021).

[13] Kuznetsova, A., et al. (2020). The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale, *Int J Comput Vis*, 128 (7) 1956–1981, July 2020, doi: 10.1007/s11263-020-01316-z.

[14] Al-refai, G., Al-refai, M. (2020). Road Object Detection using Yolov3 and Kitti Dataset, *IJACSA*, 11 (8) 2020, doi: 10.14569/IJACSA.2020.0110807.

[15] Pan, S. J., Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng*., 22 (10), 1345–1359, Oct. 2010, doi: 10.1109/TKDE. 2009.191.

[16] Athanasiadis, I., Mousouliotis, P., Petrou, L. (2020). A Framework of Transfer Learning in Object Detection for Embedded Systems, arXiv:1811.04863 [cs], Nov. 2018, Accessed: June 30, 2020. [Online]. Available: http://arxiv.org/abs/1811.04863

[17] Sanjay Kumar, K. K. R., Subramani, G., Thangavel, S. K., Parameswaran, L. (2021). A Mobile-Based Framework for Detecting Objects Using SSD-MobileNet in Indoor Environment, In Intelligence in Big Data Technologies—Beyond the Hype, Singapore, 2021, p 65–76. doi: 10.1007/978-981-15-5285-4_6.

[18] Geiger, A., Lenz, P., Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite, in 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3354–3361. doi: 10.1109/CVPR.2012.6248074.

[19] Carte graphique NVIDIA Quadro RTX 8000, NVIDIA. https://www.nvidia.com/fr-fr/design-visualization/quadro/rtx-8000/ (accessed Aug. 25, 2021).

[20] TensorFlow Core. https://www.tensorflow.org/tutorials?hl=fr (accessed Oct. 10, 2020).