The Road Is Life: Segmentation and Object Detection for Autonomous Driving

Dapeng Liu, *Student, Chalmers,* and Tian Xia, *Student, Chalmers,*

***Abstract*—The main content of our project is to use modern computer vision techniques, i.e., different neuron networks to perform perception work for autonomous driving. Two major tasks were carried out, semantic segmentation and object detection in different road scenarios. To do pixel-wise image segmentation, a deep neural network architecture, ENet (efficient neural network) is used, which is designed specifically for tasks requiring low latency operation and it can achieve similar or better accuracy comparing to existing models. The second task, object detection, we choose a paper addressing to speed/accuracy trade off, in which several meta-architectures, Faster R-CNN, R-FCN, and SSD, are implemented to evaluate the trade-off between speed and accuracy.In the experiment, to test the models' performance, we design road scenarios both similar to the training set, Gothenburg, and dissimilar, Beijing.**

***Keywords*—perception, ENet, segmentation, Faster R-CNN, SSD, object detection.**

I. Introduction

With the rapid development of autonomous driving in recent years, significant progress has been made in perception, which is playing a more and more important role. Strict requirements are raised in order to realize a higher level autonomous driving. With computer vision thriving in artificial intelligent field, camera images are now considered as a key component of sensing because it's low power consumption, informative and low cost.

Nowadays people have a strong need for real time semantic-segmentation as well as object detection algorithms with low latency, low computational requirement and high accuracy. In this paper, a image segmentation algorithm names ENet is implemented. Meanwhile, several object detection algorithms are tested and the trade-off between calculating speed and accuracy is evaluated. Finally, image segmentation based on ENet and one of the object detection algorithm are performed together under a vehicle driving circumstance. The algorithms were implemented under the environment of two cities, Peking and Gothenburg which have different city styles to test the robustness and other performance of these algorithms.

A. ENet

Pixel-wise algorithms which can label every single pixel with corresponding classification are able to have great performance with the hardwares with powerful computationally ability.

Traditionally, the convolutional neural networks(CNN) is widely used to realize image processing algorithms, however, CNN has a non-ideal performance when it comes to image segmentation in pixel level. There also has already been several segmentation algorithms based on the VGG16 architecture, for instance, SegNet and fully convolutional networks. However, the VGG 16 architecture is a large model which means it costs too much computation time. Aa new neural network architecture optimized for fast inference and high accuracy named ENet is introduced in[10]. The algorithm is realized in this project under a vehicle driving circumstance and shows a ideal performance.

B. A trade-off between speed and accuracy

Object detection models such as Faster R-CNN, R-FCN, Multibox, SSD, and YOLO which already had a terrific performance are very popular in recent years. The Single Shot Detector(SSD) is a method for detecting objects in images using a single deep neural network which is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detection.
Faster R-CNN is a kind of region-based convolutional neural networks which can achieve nearly real-time rates using very deep networks. The architecture is composed of two modules, a deep fully convolutional network that proposes regions and a Fast R-CNN detector that uses the proposed regions.
The trade off between speed and accuracy is discussed in[7]. In conclusion, for Faster R-CNN, using fewer proposals can significantly speed it up with only a small decrease in accuracy. SSD is less sensitive when the quality of the feature extractor than Faster R-CNN. In order to obtain a balance performance, some speed must be sacrificed to achieve a higher accuracy. In this project, the performance of both SSD and Faster R-CNN is evaluated and compared.

II. Related Work

In the state-of-the-art scene-parsing CNNs, an encoder and a decoder are combained together. A novel deep architecture named SegNet which is the first deep learning method to learn to map low resolution encoder feature maps to semantic labels is proposed in[1]. In the SegNet, a four layer SegNet architecture is used, the encoder is a vanilla CNN to classify the input and each encoder performs dense convolutions, ReLU non-linearity, a non-overlapping max pooling with a$\times$2 window and finally down-sampling. Each decoder upsamples its input using the memorized pooled indices and convolves it with a trainable filter bank. However, due to the large number of parameters and huge architecture of the model, the net work is too slow to satisfy our requirements. [9] shows us a

fully convolutional network (FCN) trained end-to-end, pixels-to-pixels on semantic segmentation exceeds the state-of-the-art without further machinery. [5] tried to directly adopt deep architectures designed for category prediction to pixel-wise labelling and made great succeed.

There are also a lot of locating class-specific or class agnostic bounding boxes methods in recent years. A integrated approach to object detection, recognition, and localization with a single ConvNet named Overfeat is proposed in[11]. The network is trained on the ImageNet 2012 training set, and a fully-connected layer is trained to predict the box boundary. A detector named DeepMultiBox which can generate a small number of bounding boxes as object candidates is introduced in[4]. With the detector, these boxes are generated by a single Deep Neural Network (DNN) in a class agnostic manner.

III. Methods

A. ENet

Table I shows the architecture of ENet, the initial stage contains a single block where the input is parallel followed by a 3×3, stride 2 convolution layer and a maxpooling layer after which the outputs are concatenated together. Stage 1 consists of 5 bottleneck blocks, while stage 2 and 3 have the same structure, with the exception that stage 3 does not downsample the input at the beginning. These three first stages are the encoder. Stage 4 and 5 belong to the decoder.

Table I: ENet architecture

Name	Type	Output size
initial		16×256×256
bottleneck1.0	downsampling	64×128×128
4× bottleneck1.x		64×128×128
bottleneck2.0	downsampling	128×64×64
bottleneck2.1		128×64×64
bottleneck2.2	dilated 2	128×64×64
bottleneck2.3	asymmetric 5	128×64×64
bottleneck2.4	dilated 4	128×64×64
bottleneck2.5		128×64×64
bottleneck2.6	dilated 8	128×64×64
bottleneck2.7	asymmetric 5	128×64×64
bottleneck2.8	dilated 16	128×64×64
Repeat section 2, without bottleneck 2.0		
bottleneck4.0	upsampling	64×128×128
bottleneck4.1		64×128×128
bottleneck4.2		64×128×128
bottleneck5.0	upsampling	16×256×256
bottleneck5.1		16×256×256
fullconv		C×512×512

B. SSD

The architecture of a SSD network is shown in Figure 1. In the SSD architecture, a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections[8]. Standard architectures used for image classification are used to build the bottom network layers. Auxiliary structures are added to the architecture to realize detections. Convolutional feature layers are added to the top of the network in order to predict detections at multiple scales.

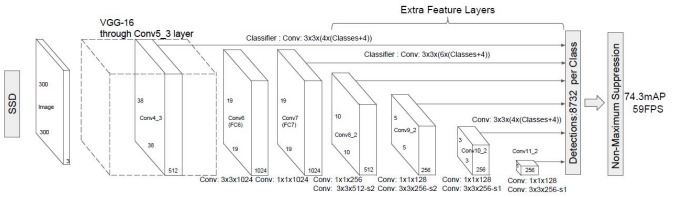

Figure 1: SSD architecture

C. Faster R-CNN

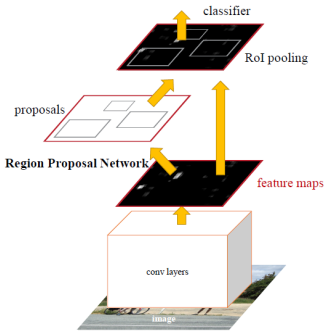

Figure 2: Faster R-CNN architecture

The illustration of Faster R-CNN is shown in 2. The Faster R-CNN is consist of a deep fully convolutional network that proposes regions and the Fast R-CNN detector that uses the proposed regions. The Region Proposal Networks, which takes an image as input and outputs a set of rectangular object proposals, each with an objectness score, can tell the Faster R-CNN where to look.

IV. Experiments and Results

In this section, we illustrate how we carry out our experiments, test the performance of neuron networks, and discuss about the results and insights we achieve.

A. Segmentation

Our Enet model has been pre-trained on Cityscape data set [3], which focuses on semantic understanding of city road scenarios. Two training data sets are provided, fine and coarse annotated. Our training is done using only the fine annotated images, in total 5000 including training, validation, and test sets. The images are taken from many different Germen cities and areas, covering vast variety of road scenes and conditions, mostly urban, but also country roads and highways. The camera is mounted on the top centre of the wind shield in a sedan. The annotated classes of the data sets are highly oriented for autonomous driving tasks, concentrated on the objects that are common and important for perception and control of the cars.

We used trained model to performs various tasks in different designed scenarios, i.e., the scenes, textures, and users of roads, layout of structures, buildings styles, terrain, and vegetation, also light conditions. The cases are intentionally designed to test the robustness and performance, both in accuracy and speed, of the algorithm. We have defined a few check points,

mentioned above, which are helping us to eventually identify what test scenarios to collect.

Given the limitations in time and resource, we decide to perform the tests using urban scenarios in Gothenburg, Sweden and Beijing, China. Gothenburg is similar to many Germen cities in the training data set in many ways. However, the road scenarios in Beijing we picked are much more alien to north west Europe. We are curious to how well generic the trained model can be.

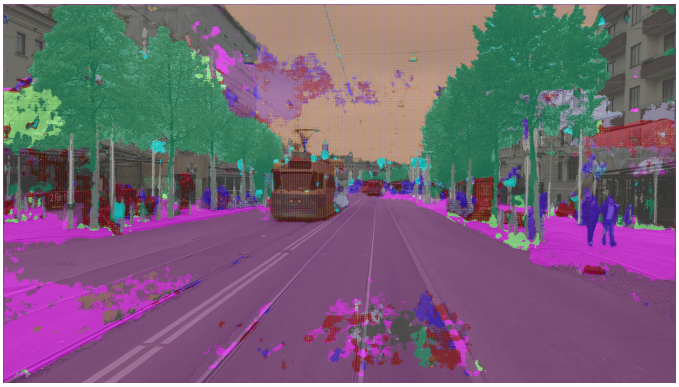

Figure 3: Segmentation result: Gothenburg

B. Results and discussions

The images of Gothenburg are extracted from a video shot by an iPhone and standing in the front of a public bus. The position and focal length of the camera are very different from the data set, as standing on a bus is much higher than the mounted position in a sedan. Also, the focal length is wider, so it sees more as well. These differences result in the surrounding areas, close to the edges, of the images, are not as good as the original testing data set.

For example, in Figure 3, the up left part of the image, where the model thinks it is a road area, probably due to the electricity lines for the tram are mistaken as the lane marking of roads. Similarly, the sidewalk colour on the building in that area is due to high correlation of the road and sidewalks, if road in the centre, the sidewalk is likely to be attached close by and to the outside. Meanwhile, the lower parts of the picture, where mixed colour in the centre and intensive of the pink for sidewalks in left down corner, are most likely to be caused by the strong bias from not using diversity of the camera positions. Namely, in most of the data the model has been exposed to, that area has high probability to be sidewalks.

However, if we focus on the central part of the image, the model has done a fairly good job. Tram, cars, bicycles and pedestrians are clearly marked. Overall, it is highly possible that we can augment the training images and pre-process the testing images to have better results in this testing sets.

The results are more interesting when applying to alien environment in Beijing, where streets are much more chaos, e.g., wider roads, intensive traffics, construction scenes on the sides of roads, various vehicle types. Once again, besides the issues mentioned above, most of the problems are due to lacking variety to break the correlations, so that model do not have clear understand of difference in classes, which results in poor robust performance.

C. Object detection

To address the issue of poor generalizability, in object detection, we choose a much more various data set, COCO [2], a data set for object segmentation, which has much richer object classes and for the same object class much richer contexts and poses. In total, 330K images and 200K are labelled, more than 800k instances.

Practically, we find object detection networks pre-trained on COCO, in the degree that it is already in the good shape. So here the potential improvement can be made by doing transfer learning on KITTI dataset [6], which is a specific data set for object detection in autonomous driving. However, we find the results from COCO are satisfactory. If given more time, we may explore the performance before and after transfer learning, it would be interesting as a future work.

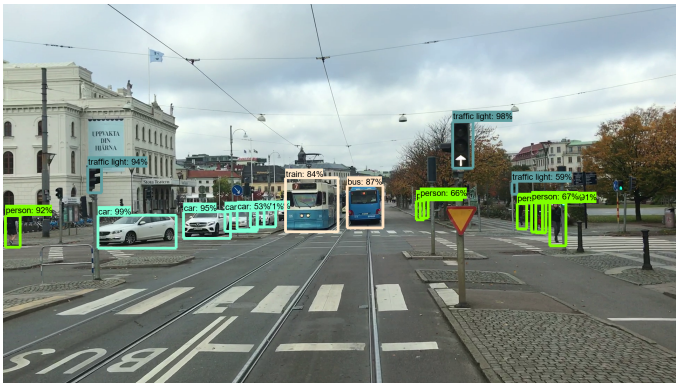

Figure 4: Object detection result: Gothenburg

D. Results and discussions

From Figure 4, the performance can be seen as relatively robust, though there are still objects omitted by the algorithm, but capable for human to identify, for example, bicycles in the central left, and traffic lights in central right.

Since the paper we adopt has implemented comparison of three major meta-architectures, faster R-CNN, R-FCN, and SSD. We have explored accuracy/speed trade-off in our testing sets. From speed perspective, SSD are considerably faster than faster R-CNN, however, from time to time it fails to label important objects which could have key impact on driving policy making. With safety requirements of self-driving cars several orders of magnitudes higher than human, the performance of our model with this training set are hardly capable to meet.

V. Conclusion

Given quantity, data variety is the key for robust performance. As deep neuron networks are profoundly data driven,

furthermore, object detection and segmentation are typical supervise learning tasks so far, in order to have good performance, not only large quantity of data is needed, but also it has to come in various ways. The power of networks is yet to unleash, the restrains are more likely from the data sets. A rough analogy would be, human build their visual experience not only on the road, but at everywhere on everything we gazed upon. It may be a trivial yet important take away from this project, for autonomous driving perception to work, the data collection for vision may go beyond the road.

REFERENCES

[1] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling". In: *arXiv preprint arXiv:1505.07293* (2015).

[2] Xinlei Chen et al. "Microsoft COCO captions: Data collection and evaluation server". In: *arXiv preprint arXiv:1504.00325* (2015).

[3] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[4] Dumitru Erhan et al. "Scalable object detection using deep neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2147–2154.

[5] Clement Farabet et al. "Learning hierarchical features for scene labeling". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1915–1929.

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[7] Jonathan Huang et al. "Speed/accuracy trade-offs for modern convolutional object detectors". In: *arXiv preprint arXiv:1611.10012* (2016).

[8] Wei Liu et al. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

[10] Adam Paszke et al. "Enet: A deep neural network architecture for real-time semantic segmentation". In: *arXiv preprint arXiv:1606.02147* (2016).

[11] Pierre Sermanet et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks". In: *arXiv preprint arXiv:1312.6229* (2013).

ACKNOWLEDGMENT

We would like to thank Prof. Lennart Svensson and TA Juliano Pinto, for their hard work and dedication.