CMPUT-503 Survey Paper: A Survey in Detection Strategies and Techniques for Autonomous Driving Systems

Jihoon Og [1]

1. Introduction

The use of autonomous robots are rising in industries worldwide (Goel & Gupta, 2020). One such industry is the field of automobiles, and the development of autonomous driving systems (ADSs) which promise accidents, and emission reduction, transportation the mobility-impaired and the reduction of driving related stress (Crayton & Meier, 2017). However, such a system is complex and is made up with numerous different systems and sub-systems in order to handle the complex environments of driving. One such system is perception, a robot needs to perceive its surrounding environment and extract information that may be critical for safe navigation. A variety of tasks, using different sensing modalities, fall under the category of perception. These include normal 2D cameras which are the most used in perception with 3D vision being a strong alternative or supplement. This paper attempts to provide a highlight on a subset of technologies and strategies used in perception systems for autonomous driving systems (ADS). Moreover, this is an in-depth summary of some of the technologies found in this survey paper (Yurtsever et al., 2020).

2. Image-based object detection

The goal of image-based object detection is to identifying the location and size of objects of interest within a 2D image. These include both static objects like traffic lights, road signs, road markings, and dynamic objects like other vehicles, pedestrians, and cyclists. Generalized object detection has an open problem in the field of computer vision for a long time. The goal is to determine if objects of specific classes are present in an image, and then determines the size of the object by producing a rectangular bounding box. This section mainly discusses a few methods that are commonly used in current ADS detection systems.

2.1. AlexNet

Previous implementations of object recognition were too slow for autonomous driving. This all changed when AlexNet set new boundaries for the ImageNet image recognition challenge (Deng et al., 2009). AlexNet is a deep convolutional neural network (DCNN) that is used to classify over a million high-resolution images in the 2010 ImageNet LSVRC contest (Krizhevsky et al., 2012). Within the dataset there were 1000 different classes that a model needed to correctly classify. On the test data, the authors were able to achieve top-1 and top-5 error rates of 37.5% and 17.0% respectively, which is significantly better than the previous state-of-the-art at the time. To further improve performance and reduce overfitting, some of the neurons do not propagate their values to the next neuron as a form of regularization called ”dropout”. This proved effective as more neurons can learn important features without being influenced by other neurons. Previous methods of object detection involved carefully crafted, hand-designed solutions that involved detailed understanding of the task in hand. This limited the flexibility to other task and made it practically impossible for these methods to be deployed in ADS. With a data-driving training approach, a DCNN can automatically learn the features necessary to accuracy classify objects in images without manual interventions from researchers. Moreover, with the use of constantly improving, commodity-grade parallel compute processors like GPUs training these kinds of models will only get better and faster. Therefore, the focus of research into object detection has mostly shifted towards supervised learning and in particular deep learning.

2.2. YOLO

While state-of-the-art methods all rely on DCNNs like AlexNet for object recognition, there seems to be a clear distinction between them. A single stage detection model uses a single network to produce the objects location and classification simultaneously. They are generally faster than region proposal models which uses two distinct stages where the location and classification is done sequentially. Therefore, a single stage detection model is generally preferred for object detection in ADS due to their faster inference time. One popular single stage detector is YOLO (You Only Look Once) (Redmon et al., 2016) which has been constantly improving with YOLOv7 (Wang et al., 2022) being their latest iteration. YOLO is a new approach to object detection by treating it as a regression problem to spatially separated bounding boxes and associated class probabilities. As the name suggest, YOLO can predict bounding boxes and class

probabilities from full images in a single evaluation. This makes YOLO very quick at processing images with their full model processing at 45 frames per second with a smaller model processing at 155 frames per second with a small loss in accuracy. YOLO works by first resizing the image into 448 by 448 pixels and dividing them into a square grid. If the center of an object falls into a grid cell, then that cell is responsible for detecting that object. Each cell predicts a certain number of bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. Additionally, each cell also predicts the class probabilities of what it thinks is in the cell. Note, they only predict a single set set of class probabilities per cell, regardless of the number of bounding boxes. Finally they combine the two results together to produce a score that encodes both he probability of that class appearing the box and how well the predicted box fits the object. Compared to Fast R-CNN, YOLO struggles to localize objects correctly. Localization errors account for more of YOLO's errors than all other sources combined. While Fast R-CNN makes fewer localization errors but far more background errors.

2.3. Faster R-CNN

While single stage detection methods have the advantage of being quick they often perform worst compared to their region proposal counterparts when comparing their accuracy. Faster R-CNN (Ren et al., 2015) which is an improvement upon Fast R-CNN (Girshick, 2015) is one of those networks. Faster R-CNN is composed of two modules. The first module is a DCNN that proposes regions, the second is the Fast R-CNN detector (Girshick, 2015) that uses the proposed regions. Both modules are used in a single, unified network for object detection. A feature map is generated when an image is processed by the first module. This feature map is then processed by a sliding window. The sliding window is mapped to a lower dimensional feature layer where it is fed into two fully connected layers that are siblings. i.e., they both take in the same feature vector from the mapped sliding window. One sibling layer is responsible for predicting what reference box best fit the object and the other sibling layer is responsible for predicting the class probability for each proposal. At each sliding window location, the network simultaneously predict up to k max region proposals and estimate the probability of object or not object for each proposal. The k proposals are the user-defined bounding boxes that the network should find the best fit for the object within the sliding window. Because these are two different kind of problems they require two different loss functions that are aggregated together to help train the network. Compared to YOLO, Faster R-CNN performs better with a mean average precision of 70.4 vs 57.9 on the VOC 2012 test, and a higher per-class average on precision for all classes using the same benchmark. However, the average inference time for a single image is about 200ms, meaning the average throughput of the network is about 5 frames per second or 9 times slower than the full model of YOLO and over 17 times slower for the smaller version. Therefore, without massive improvements in inference performance region proposal networks are less likely going to be used in ADS applications even though they are superior in object detection and classification compared to their single stage counterparts.

3. Conclusion

Object detection is an important system for ADS as the robot needs to perceive its environment accurately and quickly in order to navigate safely. This paper summarized 3 papers on image-based object detection and how they relate to ADS. With AlexNet (Krizhevsky et al., 2012) being a pioneer in machine learning-based object detection, to real-time object detection with YOLO (Redmon et al., 2016) and to a lesser extent Faster R-CNN (Ren et al., 2015). Active research is being done to improve these methods so that they are more accurate, faster, and take less computational energy.

References

Crayton, T. J. and Meier, B. M. Autonomous vehicles: Developing a public health research agenda to frame the future of transportation policy. *Journal of Transport & Health*, 6:245–252, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Girshick, R. Fast r-cnn, 2015. URL https://arxiv.org/abs/1504.08083.

Goel, R. and Gupta, P. Robotics and industry 4.0. *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*, pp. 157–169, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), pp. 779–788, 2016. doi: 10.1109/CVPR.2016.91.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020. doi: 10.1109/ACCESS.2020.2983149.