Advances in Social Science, Education and Humanities Research, volume 638
Proceedings of the 2021 International Conference on Public Art and Human Development (ICPAHD 2021)

Deep Learning in Perception of Autonomous Vehicles

Yunxiang Jiang[1*], Tengyu Hsiao[1]

[1]*Zhengzhou University*
[1]*Wake Forest University*
**corresponding author. Email: hsiat18@wfu.edu*

ABSTRACT
With the development in deep learning and sensor technologies in recent years, the ultimate goal to build full autonomous vehicles (AV) has come closer to practicality. Autonomous vehicles need to be able to percept the environment in order to make the correct decision in controlling the vehicles under different situations. In addition, the process needs to be as accurate as possible since operation under safe conditions is one of the most important issues. Moreover, the efficiency of methods is another crucial factor since complicated traffic requires vehicles to be flexible and react accordingly. Besides, if an AV cannot operate properly and timely, it would be pointless to consider it as an alternative way to the human-control car. While the progress in sensors contributes to the development of AV, the data obtained by sensors is still possible to fail due to environmental or weather conditions. This article first gives a review of the concepts in AV, especially focusing on environmental perception, then concludes and compares several deep learning methods on environment perceptions in the field of AV.

Keywords: *deep learning, autonomous vehicles, self-driving cars; perception, localization, object detection.*

1. INTRODUCTION

With the improvement of human life automation and intelligence level, autonomous driving technology is becoming an excellent solution to improve traffic safety and efficiency.

According to the U.S. Department of Transportation's Fatality Analysis Reporting System (FARS), there were 33,244 fatal vehicle accidents in the United States in 2019 and caused 36,096 deaths. In addition, 65 percent of fatally injured passenger vehicle drivers have a blood alcohol concentration (BAC) higher than 0.08 percent[1]. There is no doubt that human factors play an important role in the cause of road traffic accidents, every person involved in road traffic can have a huge impact on road traffic safety, and there is no doubt that we are impossible to guarantee that every driver on the road is reliable.

In this case, autonomous vehicles can be a great solution in preventing human-caused accidents like drunken drinking or drivers' distractions during driving. In addition, AVs with high accuracy and efficiency is more likely to solve the problem of traffic jam and save people's time.

Autonomous vehicles have been attached great importance to by governments of all countries. Research institutions around the world have invested a lot of manpower and material resources, and major automobile enterprises, technology companies, auto parts suppliers and autonomous vehicle start-ups have also entered this field of investment and research and development. For example, Amazon announced at CES 2018 that it is jumping into the development of self-driving cars through a partnership with Toyota. Their demonstration car, called Electronic Palette, had been scheduled to debut at the Tokyo 2020 Summer Olympics.

To achieve autonomous driving, the computer should be able to do the following: perception, localization and mapping, path planning, and vehicle control. These four elements also help build the architecture of AVs: Inputs of data, for example, data retrieved from the camera, LiDAR, or RADAR, are processed through Neural Network. After interpreting data through a different method like Clustering, Regression, and finally, produce output in Environment perception. The generated information helps the computer to make the decision and finally control the AV the operate properly.

In terms of the level of autonomous driving, the Society of Automotive Engineers (SAE) clearly defines 6 levels of driving automation from 0 to 5[2]

Level 0: The vehicle is fully controlled and supervised by the human driver. No automation is involved.

Level 1: Vehicle assists driver with either longitudinal or lateral steering.

Level 2: Vehicle assists driver with both longitudinal and lateral steering.

Level 3: The vehicle is in full control under some situations and will inform the driver to take over if needed.

Level 4: Vehicles takes control of the full trip under some situation and drivers do not need to take over.

Level 5: Vehicle takes control of everything under all conditions.

Many cars in large production are now at the level between level 2 and level 3. Many autonomous features like automatic braking, automatic parking, and lane-keeping aid are already prevalent in many commercial cars. Companies also have different choices of sensing hardware in collecting environmental information. For example, the well-known debate between cameras and LiDAR. In this article, we review different deep learning methods as well as the hardware used in Autonomous Vehicles. The authors especially focus on the environment perception and finally discuss some challenges and possible improvement and development. For vehicle control, a survey by Kuutii provides a comprehensive review of the current state of autonomous vehicle control[3].

2. LITERATURE REVIEW OF BASIC DEEP LEARNING

2.1. Supervised

Statistical learning includes supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Supervised learning is a task model of learning that can be well predicted for any given corresponding input and output. The basic operation of a computer is to produce output in the case of a given input, so supervised learning is an extremely important branch of statistical learning, and it is also the most abundant and widely used part of that. In the scenarios of automatic driving problems, the recognition of radar images and camera images often needs to use manual annotation assistance and supervised learning algorithms.

2.1.1.Definition of supervised learning:

Supervised learning is defined as machine learning problems that learn predictive models from annotated data. Its essence is to learn the statistical law of input-output mapping. Annotated data represents the correspondence between inputs and outputs, and the predictive model generates outputs for a given input.

2.1.2. Basic elements of supervised learning

Feature space: Each concrete input is an instance, usually represented by feature vectors. Each dimension of the eigenvector corresponds to a feature, and the space in which all the eigenvectors exist is called the eigenspace.

Input space: The set of all possible values that will be entered

Output space: The set of output all possible values

2.1.3. Three major processes of supervised learning

(1) approximation

Question: How much hypothesis space do we have?

Belong to "approximate theory" and "harmonic analysis" category

(2) optimization

Question: How to find or approach an approximation?

It includes design of large scale optimization algorithm, convergence analysis and effective implementation

(3) the generalization

Question: Can we generalize to unknown data?

This involves a fundamental interaction between the size of the dataset and the complexity of the hypothesis space.

2.2. CNN

Neural network is a part of artificial intelligence research field. At present, the most popular neural network in image processing, feature extraction and other fields is deep convolutional neural network. Although convolutional networks also have shallow structures, they are rarely used due to accuracy and expressiveness. At present, when it comes to CNN and convolutional neural networks, academia and industry no longer distinguish between them deliberately. They usually refer to convolutional neural networks with deep structures ranging from "several layers" to "tens or hundreds of layers". CNN can automatically learn features from (often large-scale) data and generalize the results to the same type of unknown data.

Pattern recognition can generally be thought of as having four stages:

Data acquisition: e.g. digitized images.

Preprocessing: such as image denoising and image geometry correction.

Feature extraction: Searches for attributes recognized by the computer that describe how the current image differs from other images.

Data classification: Classify the input images into specific categories.

At present, CNN is the best method to extract image domain features, which greatly improves the accuracy of data classification. It has achieved great success in many research fields such as speech recognition, image recognition, image segmentation and natural language processing.

2.2.1. The network structure of CNN

The basic CNN consists of input layer, convolution layer, activation layer, pooling layer and full connection layer.

Input layer: Used to data input.

Convolution layer: feature extraction and feature mapping using convolution kernel.

Activation layer: Since convolution is also a linear operation, nonlinear mappings need to be added.

Pooling layer: the feature graph is down-sampled and sparsely processed to reduce the amount of data calculation.

Full connection layer: usually reinstalled at the tail of CNN to reduce the loss of characteristic information

Output layer: Used to data output.

Other functional layers can be used in between, such as: Normalization layer, Split layer, Fusion layer

Normalization layer: normalization of CNN functions

Split layer: An individual study of some (picture) data by region

Fusion layer: a branch that fuses independent learning features

The output result of CNN is the specific feature space of each image. When dealing with the task of image classification, we take feature space CNN as the full connection layer or the full connection neural network of input and output, and use the complete connection layer to map the tag set from the input image, namely, classification. The most important work in the whole process is the back propagation algorithm, that is, how to adjust the weight of the network through training data iteration. At present, mainstream convolutional neural networks, such as Xception[4] and ResNeXt[5] , are improved on the basis of simple CNN.

3. DEEP LEARNING IN ENVIRONMENT PERCEPTION

In the following, we give an overview of environment perceptions in terms of how they are performed in assisting AVs to make the decision. In addition, different sensing hardware is introduced and compared. Finally, several deep learning methods are concluded and compared.

3.1 Object detection: Classification and Localization

Object detection is crucial in environment perceptions. In order to learn about the surrounding situation, AVs need to be able to produce useful information from the data collected by the sensors. As a result, CNN is a more accurate deep learning model in detecting objects from images. Two main issues are solved through object detection: Classification and Localization. Classification is to assign the class to objects in the image. Therefore, CNN is a good option. Localization requires the computer to locate the exact position of an object by drawing a bounding box to the object. In higher automation, no bounding box is needed to localize an object. Object detection is the combination of object classification and object localization. AVs should be able to identify several objects in one single image. To evaluate how well a model localize objects, Intersection over Union (IOU) is an easy understanding measure[6]. IOU is the Area of Intersection of the bounding box predicted by the model and the true bounding box divided by the area of the union of them:

$$IOU = \frac{Area\ of\ Intersection\ of\ two\ boxes}{Area\ of\ Union\ of\ two\ boxes}$$

A larger IOU implies better object localization of the method.

3.2. Sensing hardware

As mentioned above, there are different ways to collect input for AVs. In order to detect objects and percept environments at the high speed of cars, sensing hardware should be able to collect a wide range of environmental conditions. There is two main sensing hardware for the AVs industry: Camera and LiDAR. Cameras collect 2D images while LiDARs generate 3D point clouds as input for the deep learning method. There is a continuous debate between this two hardware as they have both advantages and disadvantages. LiDARs are favored due to their high resolution. In addition, unlike a camera, LiDARs can still precisely collect information when there is no sufficient light. On the other hand, the camera is a well-known choice of Tesla, one of the largest car manufacturers. The low cost of the camera allows large production. However, both cameras and LiDARs may not operate normally under bad weather conditions,

for example, fog. Bad weather conditions still need to be solved in order to achieve level 5 automation. One more option is to combine camera and LiDARs to increase the accuracy of the deep learning model by fusing the Point Cloud of LiDARs into CNN and thus increase the IOU[5]. Besides, other minor sensing hardware, for example, radar, are also used to assist and enhance the LiDAR and Camera's perception.

3.3. CNN based algorithm

Since LiDARs, Cameras, and other sensing hardware produce visual data for the environment perception, convolutional networks (CNN) based algorithms are frequently used since they work well with visual data like images. During the development of CNN-based algorithms, two parallel paths evolved: A two-stage object detector and a Single-stage object detector. Two-stage detectors first deal with the regional proposal before predicting while Single-stage detectors directly do the prediction[7]. In the following, different CNN-based object detection algorithms are introduced in the order of development of CNN algorithms.

3.3.1. Regional-Based CNN (R-CNN)

Regional-Based CNN (R-CNN) is the first two-stage detector. The idea is to select 2000 proposal regions from test images. Since it does not cover every region of the image, the detection speed is improved. R-CNN network first extracts feature from selected regions, then a Support Vector Machine (SVM) classify the object. However, the training of R-CNN still has a high time-complexity and space-complexity to be used in the practical application of AV.

3.3.2. Fast R-CNN

Fast R-CNN is different from R-CNN in the way of processing the region proposals. R-CNN uses a CNN network while fast R-CNN produces a convolutional feature map using the whole image and generates Regions of interest (ROI) and finally feed to the fully connected layer. Two layers are generated as output branches. One is the probability of the class predicted by softmax layer while the other outputs four real-valued numbers of the class[8]. Fast R-CNN has faster training and testing speed as well as higher detection accuracy.

3.3.3. Faster R-CNN

Faster R-CNN is very similar to Fast R-CNN. The difference is that it uses a separate neural network called the region proposal network (RPN) when predicting the bounding boxes. CNN and RPN layers are merged and finally, the ROI layer classifies objects in each region.

3.3.4. You Only Look Once (YOLO)

Unlike previous processes, YOLO is a Single-Stage detector. In YOLO, it treats the problem as a one-step regression. The entire image is input to a single neural network to predict the bounding box and class probability for the boxes. Despite its high detection speed, YOLO has a high localization error and low detection accuracy if the size of the object is small. But when dealing with a large object like vehicles, YOLO is still a nice choice as it is fast enough to be applied in the real-time practice of AVs.

3.3.5. Single-shot Multibox Detector (SSD)

Unlike previous methods, Single-shot Multibox Detector (SSD) is allowed to have different sizes and ratios of bounding boxes, thus it can be applied to detect objects of different sizes in one image. The entire image is input to the VGG-16 network to produce the feature maps. Then Convolutional layers use feature maps to do the classification and produce the bounding boxes. Then train the network by matching the boxes to the ground truth detection. SSD is not only faster but also more accurate than faster R-CNN and YOLO[9].

4. CONCLUSION

In this article, we review the concept of autonomous driving and focus on the environment perception branch. To achieve comprehensive environment perception, efforts in developing the best hardware and software are necessary. It can be noticed that the field of Autonomous Vehicle is an interdisciplinary field. There are still many aspects of this field that can be explored, for example, more precise sensors, more efficient algorithms, credible data collection, and so on.

Moreover, there are still many challenges in this field to overcome in order to achieve full automation[10]. Bad weather condition is still an unsolved challenge. The sensor would fail to collect precise input and the model may produce misleading executions. More needs to be invested to develop sensors that can also work under harsh weather conditions.

In addition, the system needs to provide an alternate solution when the sensor is not working properly, for example, give the control back to the human driver. Security is another aspect that needs to be further researched. Cyber security should be ensured so that AVs can operate safely under drivers' will. Moreover, sensors should be protected from being attacked by other sources, for example, invisible lights and radar interference[11].

It is no doubt that our paper has some limitations. One of the major limitations is that the authors did not browse enough papers and did not make a detailed discussion in the study due to the limited time. Besides, this study lacks practical practice to validate and deepen our understanding of the literature, which the authors read.

These limitation are what we are trying to improve in future research

REFERENCES

[1] Fatality Facts 2019 (State by state) doi: https://www.iihs.org/topics/fatality-statistics/detail/state-by-state#yearly-snapshot

[2] SAE Levels of Driving Automation™ doi:https://www.sae.org/blog/sae-j3016-update

[3] S. Kuutti, R. Bowden, Y. Jin, P. Barber and S. Fallah, "A Survey of Deep Learning Applications to Autonomous Vehicle Control," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 2, pp. 712-733, Feb. 2021, doi: 10.1109/TITS.2019.2962338.

[4] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", in Cornell university, Tue, 4 Apr 2017, doi: arXiv:1610.02357

[5] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, "Aggregated Residual Transformations for Deep Neural Networks" in CVPR 2017 doi:arXiv:1611.05431

[6] Intersection over Union (IOU) doi:https://medium.com/analytics-vidhya/iou-intersection-over-union-705a39e7acef7

[7] X. Du, M. H. Ang and D. Rus, "Car detection for autonomous vehicle: LIDAR and vision fusion approach through deep learning framework," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 749-754, doi: 10.1109/IROS.2017.8202234.

[8] Girshick, Ross, "Fast R-CNN", in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Dec.2015, dio:https://openaccess.thecvf.com/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg,"SSD: Single Shot MultiBox Detector" in ECCV 2016, 29 Dec 2016, doi:10.1007/978-3-319-46448-0_2

[10] Bagloee, S.A., Tavana, M., Asadi, M. et al. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. J. Mod. Transport. 24, 284–303 (2016). https://doi.org/10.1007/s40534-016-0117-3

[11] Wei Wang,Yao Yao,Xin Liu,Xiang Li, Pei Hao,Ting Zhu in CCS '21: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications SecurityNovember 2021 Pages 1930–1944
https://doi.org/10.1145/3460120.3484766