DEEP LEARNING-BASED IMAGE RECOGNITION FOR AUTONOMOUS DRIVING

NAME: OLADIRAN LORETTA

ID: 2019380065

DIGITAL IMAGE PROCESSING

U10M12021

CONTENTS

(Please do not edit the space; it will affect the placement of the images. Thank you)

• ABSTRACT

Various image recognition tasks were handled in the image recognition field prior to 2010 by combining image local features manually designed by researchers (called handcrafted features) and machine learning method. After entering the 2010's, many image recognition methods that use deep learninghave been proposed. Hence, the paper we analyzed explains how deep learning is applied to the field of image recognition, and also explains the latest trends of deep learning-based autonomous driving.

Keywords: Image processing, Image recognition, Computervision, Deep learning, Convolutional Neural Network

• INTRODUCTION

In the late 1990s, it became possible to process a large amount of data at high speed with the evolution of general-purpose computers.The mainstream method was to extract a feature vector (called the image local features) from the image and apply a machine learning method to perform image recognition. Supervised machine learning requires a large amount of class-labeled training samples, but it doesnot require researchers to design some rules as in the case of rule- based methods. Hence, versatile image recognition can be realized.

In the 2000 era, handcrafted features such as scale-invariant featuretransform (SIFT) and histogram of oriented gradients (HOG) as image local features, designed based on the knowledge of researchers, havebeen actively researched. By combining the image local features withmachine learning, practical applications of image recognition technology have advanced, as represented by face detection.

Next, in the late 2010s, deep learning to perform featureextraction process through learning has come under the spotlight. A handcrafted feature is not necessarily optimalbecause it extracts and expresses feature values using a designed algorithm based on the knowledge of researchers.

Deep learning is an approach that can automate the feature extraction process and is effective for image recognition. Deep learning has accomplished impressive results in the general object recognition competitions, and the use of image recognition required for autonomousdriving (such as object detection and semantic segmentation) is in progress.

- ## AUTONOMOUS VEHICLE

An autonomous vehicle, also known as a self-driving car, isa vehicle that is capable of sensing its environment and moving safely with little or no human input. Self-driving cars combine a variety of sensors to perceive their surroundings, such as radar, sonar, GPS and inertial measurement units. Advanced control systems interpret sensory information to identify appropriate navigation paths, as well as obstacles. Possible implementations of the technology include personal self-driving vehicles, connected vehicle platoons and long-distance trucking.

Experiments have been conducted on automated driving systems (ADS) since at least the 1920s, however the trials began in the 1950s. The first semi-automated car was developed in 1977, by Japan's Tsukuba Mechanical Engineering Laboratory, which required specially marked streets thatwere interpreted by two cameras on the vehicle and an analog computer.

By 1985, the ALV had demonstrated self-driving speeds on two-lane roads of 31 kilometres perhour, with obstacle avoidance added in 1986, and off-road driving in day and nighttime conditions by 1987.

A major milestone was achieved in 1995, with CMU's NavLab 5 completing the first autonomous coast-to-coast drive of the United States. Of the 4,585 km between Pittsburgh, Pennsylvania andSan Diego, California, 4,501 km were autonomous, completed with an average speed of 102.7 km/h. The Carnegie Mellon University Navlab drove 4,584 kilometres across America in 1995, 4,501 kilometres. Navlab's record achievement stood unmatched for two decades until 2015, when Delphi improved it by piloting an Audi, augmented with Delphi technology, over 5,472 kilometres through 15 states while remaining in self-driving mode 99% of the time.

In 2015, the US states of Nevada, Florida, California, Virginia, and Michigan, together with Washington, DC, allowed the testing of automated cars on public roads. From 2016 to 2018, theEuropean Commission funded an innovation strategy development for connected and automateddriving.

In November 2017, Waymo announced that it had begun testing driverless cars without a safety driver in the driver position, however, there was still an employee in the car. An October 2017 report by the Brookings Institute found that the $80 billion had been reported as invested in all facets of self driving technology up to that point, but that it was "reasonable to presume that total global investment in autonomous vehicle technology is significantly more than this." In December 2018, Waymo was the first to commercialize a fully autonomous taxi service in the US, in Phoenix, Arizona. In October 2020, Waymo launched a geo-fenced driverless ride hailing service in Phoenix. The cars are being monitored in real-time by a team of remote engineers, and there are cases where the remote engineers need to intervene.

In March 2019, ahead of the autonomous racing series Roborace, Robocar set the Guinness WorldRecord for being the fastest autonomous car in the world. In pushing the limits of self-driving vehicles, Robocar reached 282.42 km/h.

On 5 March 2021, Honda began leasing in Japan a limited edition of 100 Legend Hybrid EX sedansequipped with the newly approved Level 3 automated driving equipment which had been granted the safety certification by Japanese government to their autonomous "Traffic Jam Pilot"driving technology, and legally allow drivers to take their eyes off the road.

In China two publicly accessible trials of robo-taxis have been launched, in 2020 in Shenzhen by Chinese firm AutoX and in 2021 in Beijing by Baidu.

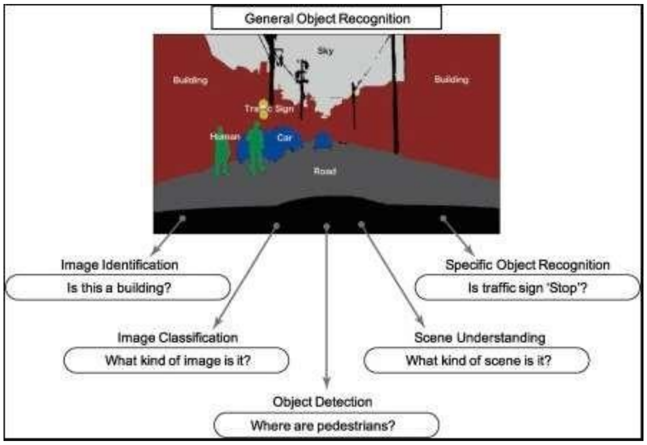
• MOTIVATION

The goal for this work is to explain the imagerecognition obstacles autonomous driving faces, and how we could apply deep learningto solve these aforementioned obstacles.

• PROBLEM SETTING IN IMAGE RECOGNITION

So first, we need to talk about the problems we will face and try to solve them using deep learning whileusing image recognition for autonomous driving.

In conventional machine learning (we define as a method prior to the time when deep learning gained attention), it is difficult to directly solve general object recognition tasks from the input image. This problem can be solved by distinguishing the tasks of **image identification, image classification,object detection, scene understanding, and specific object recognition**.



A. Image Verification

Image verification is a problem to check whether the object in the image is the same as the reference. For example fingerprint, face, and person identification relates to tasks in which it is required to determine whether anactual person is another person. In image verification, the distance between the feature vector of the reference pattern and the feature vector of the input image is calculated. If the distance value is less than a certain value, the images are determined as identical, and if the value is more, it is determined otherwise.

In deep learning, the problem of person identification is solved by designing aloss function (triplet loss function) that calculates the value of distance between two images of the same person as small, and the value of distance with another person's image as large.

B. Object Detection

Object detection is the problem of finding the location of an object of a certain category in the image. Practical face detection and pedestrian detection are included in this task. Face detection uses a combination of Haar-like features and AdaBoost, and pedestrian detection uses HOG featuresand support vector machine (SVM).

In conventional machine learning, object detection is achieved by training 2- class classifiers corresponding to a certain category and raster scanning in theimage. However, in deep learning-based object detection, multiclass object detection targeting several categories can be achieved with one network.

C. Image Classification

Image classification is the problem to find out the category to which an

object in an image belongs to, among predefined categories.

In the conventional machine learning, an approach called bag-of-features(BoF) has been used: a vector quantifies the image local features and expresses the features of the whole image as a histogram.

Yet, deep learning is well-suited to the image classification task, and becamepopular in 2015 by achieving an accuracy exceeding human recognition performance in the 1000-class image classification task.

D. Scene Understanding

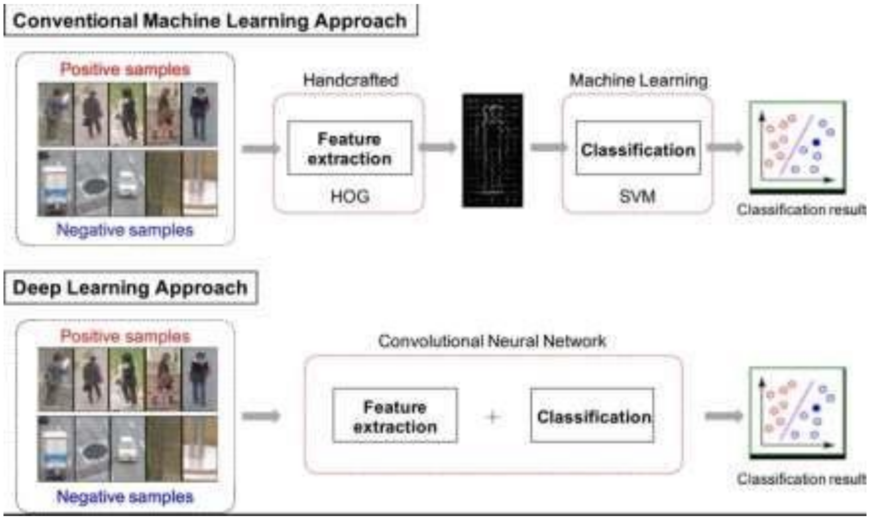
Scene understanding is the problem of understanding the scene structure in an image. Above all, semantic segmentation that finds object categories in each pixel in an image has been considered difficult to solve using conventional machine learning. Therefore, it has been regarded as one of theultimate problems of computer vision, but it has been shown that it is a problem that can be solved by applying deep learning.

E. Specific Object Recognition

Specific object recognition is the problem of finding a specific object. By giving attributes to objects with proper nouns, specific object recognition isdefined as a subtask of the general object recognition problem. Specific object recognition is achieved by detecting feature points using SIFT from images, and a voting process based on the calculation of distance from feature points of reference patterns.

- **DEEP LEARNING-BASED IMAGE RECOGNITION**

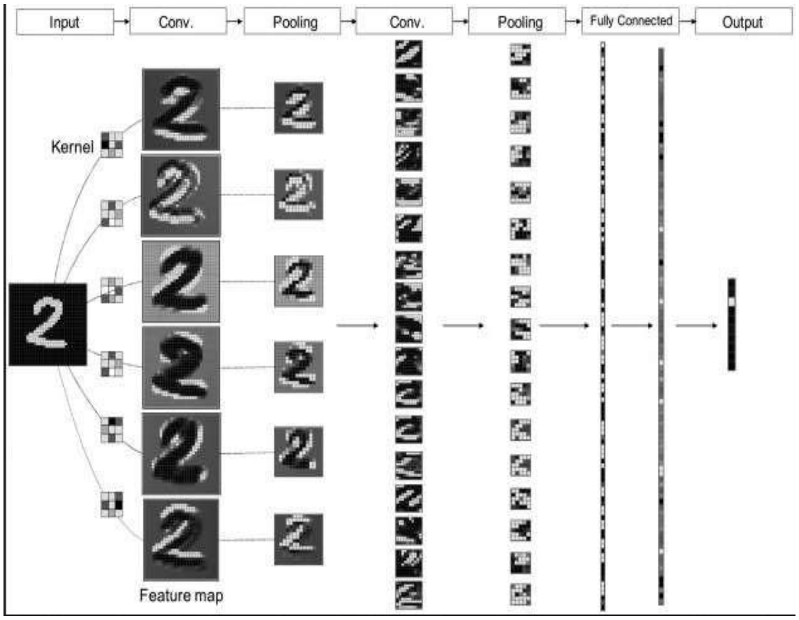
This work uses the convolutional neural network (CNN), which is onetype of deep learning, as the approach for learning classificationand feature extraction from training samples, as shown in the figure on the left.



1. CONVOLUTIONAL NEURAL NETWORK (CNN)

A CNN is a class of deep neural network, most commonly applied to analyze visual imagery. Asshown in the figure, CNN computes the feature map corresponding to the kernel by convolutingthe kernel (weight filter) on the input image.

Feature maps corresponding to the kernel types can be computed as there are multiple kernels. Next, the size of the feature map is reduced by the pooling feature map. As a result, it is possible to absorb geometrical variations such asslight translation and rotation of the input image. The convolution process and the pooling process are applied repeatedly to extract the feature map. The extracted feature map is inputto fully-connected layers, and the probability of each class is finally output. In this case, the input layer and the output layer have a network structure that has units for the image and the number of classes.

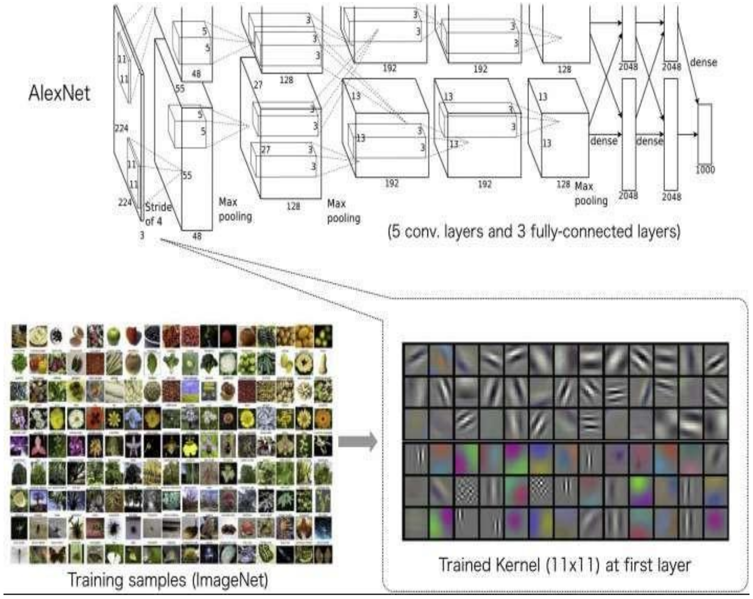


Training of CNN is achieved by updating the parameters of the network by thebackpropagation method. The parameters in CNN refer to the kernel of the convolutional layer and the weights of all coupled layers. The process flow ofthe backpropagation method is shown in the aforementioned figure. First, training data is input to the network using the current parameters to obtain the predictions (forward propagation). The error is calculated from the predictions and the training label; the update amount of each parameter is obtained from the error, and each parameter in the network is updated from the output layer toward the input layer (back propagation). Training of CNN refers to repeating these processes to acquire good parameters that can recognize the images correctly.

i. Advantages of CNN compared to conventional machine learning

This figure below shows some visualization examples of kernels at the first convolution layer of the AlexNet, which isdesigned for 1000 object class classification task at ILSVRC(ImageNet Large Scale Visual Recognition Challenge)2012. AlexNet consists of five convolutionlayers and three fully-connected layers, whose output layer has 10,000 units corresponding to the number of classes.

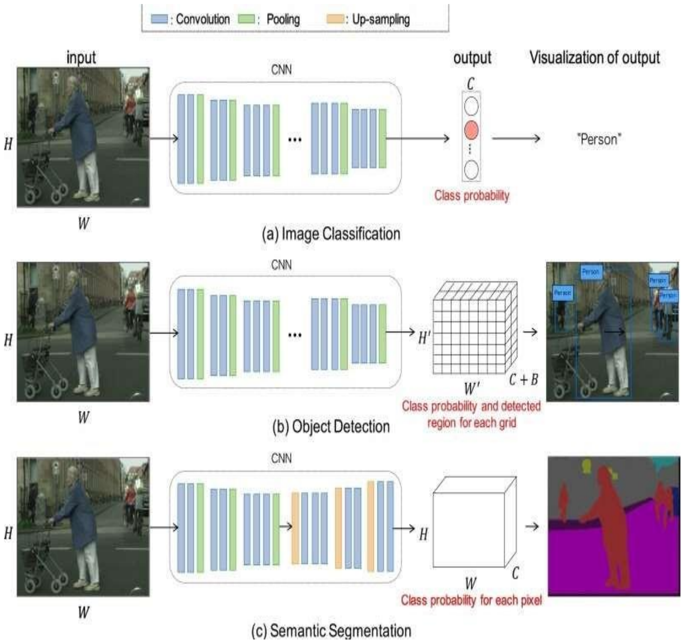
We see that the AlexNet has automaticallyacquired various filters that extract edge,texture, and color information with directional components. The detection miss rate for CNN filters is 3%, while the HOG is 8%. Although the CNN kernels of the AlexNet not trained for the human detection task, the detection accuracy improved over the HOG feature that is thetraditional handcrafted feature.


AlexNet
11
11
224
224
Stride of 4
3
55
48
5
5
27
128
Max pooling
Max pooling
3
192
13
128
dense
2048
1000
(5 conv. layers and 3 fully-connected layers)
Training samples (ImageNet)
Trained Kernel (11x11) at first layer

As shown in this figure below, CNN can perform not only image classification but also object detection and semantic segmentation by designing the output layer according to each task of image recognition. For example, if theoutput layer is designed to output class probability and detection region for each grid,it will become a network structure that can perform object detection. In semantic segmentation, the output layer should be designed to output the class probability for each pixel. Convolution and pooling layers canbe used as common modules for these tasks.

On the other hand, in the conventional

machine learning method, it was necessary todesign image local features for each task and combine it with machine learning. CNN has the flexibility to be applied to various tasks by changing the network structure, and this property is a great advantage in achieving image recognition.


🟦: Convolution 🟩: Pooling 🟧: Up-sampling
input
CNN
output
Visualization of output
C
"Person"
Class probability
H
W
H'
W'
C + B
Person
Class probability and detected region for each grid
H
W
C
Class probability for each pixel

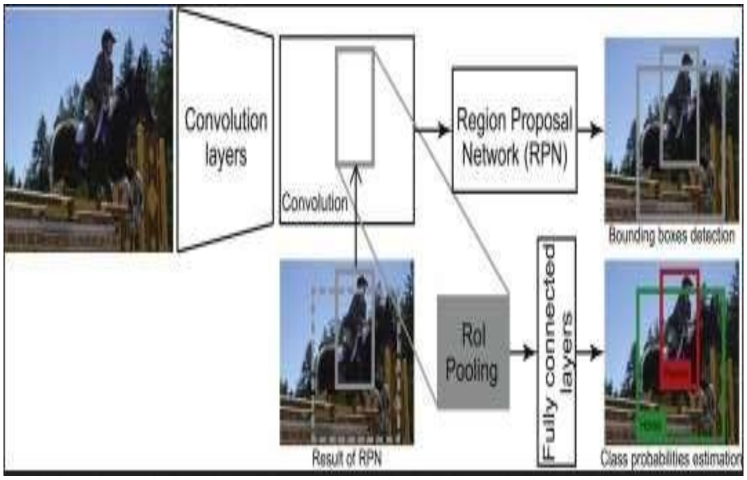
(a) Image Classification

(b) Object Detection

(c) Semantic Segmentation

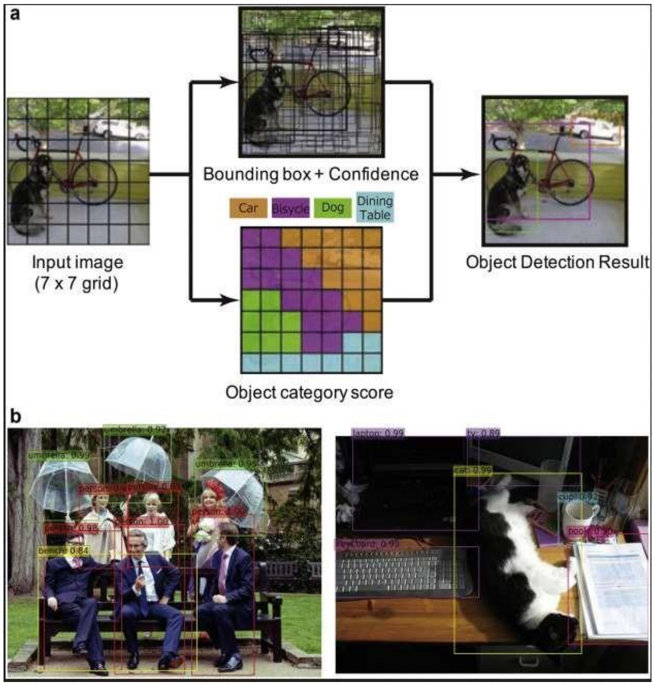
ii. Application of CNN to object detection task

Conventional machine learning-based object detection is an approach that raster scans two class classifiers. In this case, because theaspect ratio of the object to be detected is constant, it will be object detection of only acertain category learned as a positive sample.On the other hand, in object detection using CNN, object proposal regions with different aspect are detected by CNN, and multiclass object detection is possible using the Region Proposal approach that performs multiclass classification with CNN for each detected region. Faster R-CNN introduces Region Proposal Network (RPN) as shown in the figure, and simultaneously detects object candidate regions and recognizes object classes in those regions.



First, convolution processing is performed on the entire input image to obtain a feature map. In RPN, an object is detected by raster scanning the detection window on the obtained feature map. In raster scanning, detection windows inthe form of k number of shapes are applied centered on focused areas known as anchor. The region specified by the anchor is input to RPN, and the score ofobject likeness and the detected coordinates on the input image are output. Inaddition, the region specified by the anchor is also input to another all- connected network, and object recognition is performed when it is determinedto be an object by RPN. Therefore, the unit of the output layer is the number obtained by adding the number of classes and ((x, y, w, h) × number of classes)to one rectangle. These Region Proposal methods have made it possible to detect multiple classes of objects with different aspect ratios.

In 2016, the single-shot method was proposed as a new multiclass object detection approach. This is a method to detect multiple objects only by giving the whole image to CNN without raster scanning the image. YOLO (You Only Look Once) is a representative method in which an object rectangle and an object category is output for each local region divided by a 7 × 7 grid, as shown in the previous figure. First, feature maps are generated through convolution and pooling of input images. The position (i, j) of each channel of the obtained feature map (7 × 7 × 1024) is a structure that becomes a region feature corresponding to the grid (i, j)of the input image, and this feature map is input to fullyconnected layers. The output values obtained through fully connected layers are the score (20 categories) of the object category at each grid position and the position, size, and reliability of the two object rectangles. Therefore, the unit of the output layer is the number (1470) in which the position, size, and reliability ((x, y, w, h, reliability) × 2) of two object rectangles is added to the number of categories (20 categories) and multiplied with the number of grids (7 × 7). In YOLO, it is not necessary to detect object region candidates such as Faster R-CNN; therefore, object detection can be performed in real time. This figure shows an example of YOLO-based multiclass object detection.

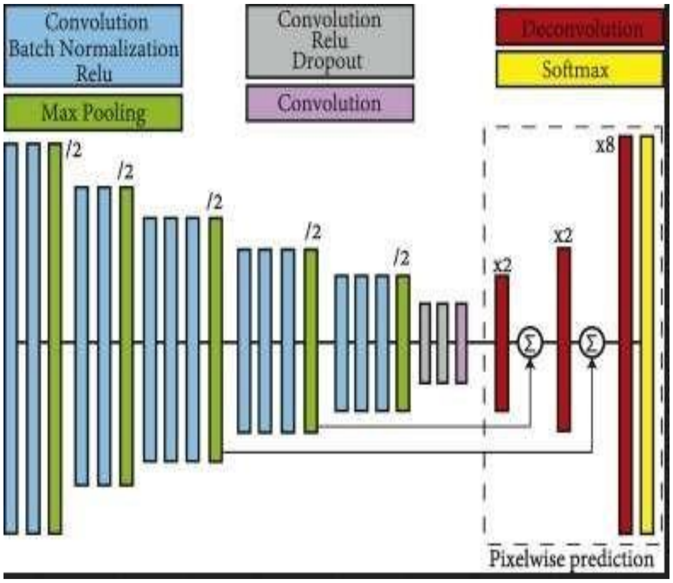


iii. Application of CNN to semanticsegmentation

Scene understanding is a difficult task. However, as with other tasks,deep learning-based methods have been proposed and achieved muchhigher performance than conventional machine learning methods.

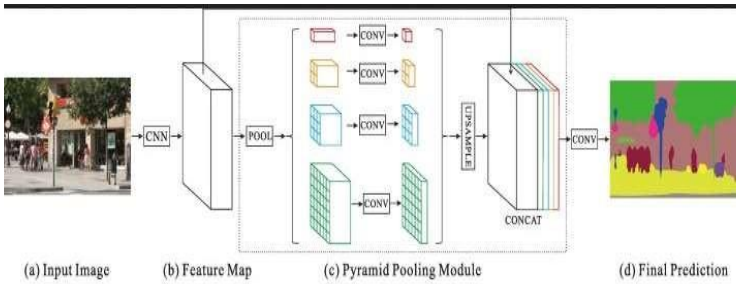
Fully convolutional network (FCN) is a method that enables end-to-endlearning and can obtain segmentation results using only CNN. The FCN has a network structure that does not have a fully-connected layer.

The size of the generated feature map is reduced by repeatedly performing the convolutional layer and the pooling layer on the input image. To make it the same size as the original image, the feature map is enlarged 32 times in the final layer, and convolution processingis performed. This is called deconvolution. The final layer outputs theprobability map of each class. The probability map is trained so that the probability of the class in each pixel is obtained, and the unit of output of the end-to-end segmentation model is (w × h × number of classes). Generally, the feature map of the middle layer of CNN captures more detailed information as it is closer to the input layer, and the pooling process integrates these pieces of information, resulting in the loss of detailed information. When this feature map isexpanded, coarse segmentation results are obtained. Therefore, high accuracy is achieved by integrating and using the feature map of the middle layer. Additionally, FCN performs processing to integrate feature maps in the middle of the network. Convolution process is performed by connecting mid-feature maps in the channel direction, and segmentation results of the same size as the original image are output.



When expanding the feature map obtained on the encoder side, PSPNet can capture information of different scales by using the Pyramid Pooling Module, which expands at multiple scales. The Pyramid Pooling Module is used to pool feature maps with 1 × 1, 2 × 2, 3 × 3, 3 × 6 × 6 in which the vertical and horizontal sizes of the original image are reduced to 1/8, respectively, on the encoder side. Then, convolution process is performed on each feature map. Next, the convolution process is performed and probability maps of each class are output after expanding and linking feature maps to the same size.

PSPNet is the method that won in the “Scene parsing” category of ILSVRC held in 2016. Also, high accuracy has been achieved with the Cityscapes Dataset taken with a dashboard camera. This figure shows the result of PSPNet-based semantic segmentation.



IV. CNN for ADAS application

The machine learning technique is applicable to use for system intelligence implementation in ADAS(Advanced Driving Assistance System). In ADAS, it is to facilitate the driver with the latest surrounding information obtained by sonar, radar, and cameras. Although ADAS typically utilizes radar and sonar for long-range detection, CNN-based systems can recently play a significant role inpedestrian detection, lane detection, and redundant object detection at moderate distances.

For autonomous driving, the core component can be categorized into three categories, namely perception, planning, and control. Perception refers to the understanding of the environment, suchas where obstacles located, detection of road signs/marking, and categorizing objects by their semantic labels such as pedestrians, bikes, and vehicles. Localization refers to the ability of the autonomous vehicle to determine its position in the environment. Planning refers to the process of making decisions in order to achieve the vehicle's goals, typically to bring the vehicle from a start location to a goal location while avoiding obstacles and optimizing the trajectory. Finally, the control refers to the vehicle's ability to execute the planned actions. CNN-based object detection issuitable for the perception because it can handle the multi-class objects. Also, semantic segmentation is useful information for making decisions in planning to avoid the obstacles by referring to pixels categorized as road.

- **DEEP LEARNING-BASED AUTONOMOUSDRIVING**

Now we will talk about end-to-end learning that can infer the control value ofthe vehicle directly from the input image as the use of deep learning for autonomous driving, and describes visual explanation of judgment grounds that is the problem of deep learning models and future challenges.

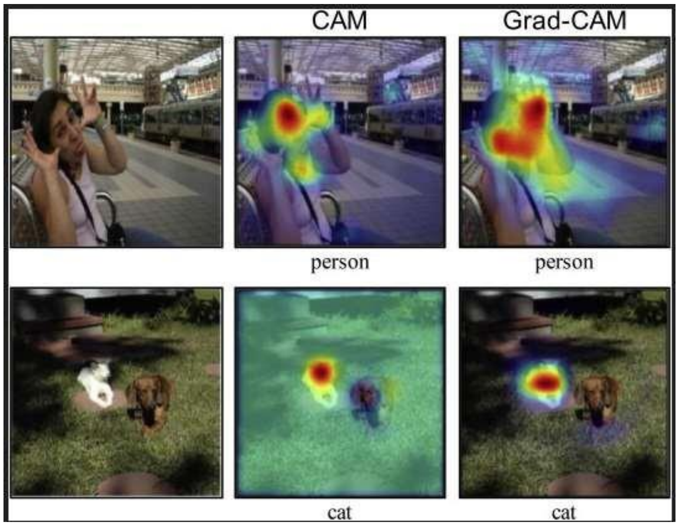
I. END-TO-END LEARNING-BASED AUTONOMOUS DRIVING

In most of the research on autonomous driving, the environment around the vehicle is understood using a dashboard camera and Light Detection and Ranging (LiDAR), appropriate traveling position is determined by motion planning, and thecontrol value of the vehicle is determined . Autonomous driving based on these three processes is common, and deep learning-based object detection and semantic segmentation are beginning to be used to understand the surrounding environment. On the other hand, with the progress in CNN research, end-to-end learning-based method has been proposed that can infer the control value of the vehicle directly from the input image . In these methods, network is trained by using the images of the dashboard camera when driven by a person, and the vehicle control value corresponding to each frame as learning data. End-to-end learning-based autonomous driving control has the advantage that the system configuration is simplified because CNN learns automatically and consistently without explicit understanding of the surrounding environment and motion planning.

II. VISUAL EXPLANATION OF END-TO-ENDLEARNING

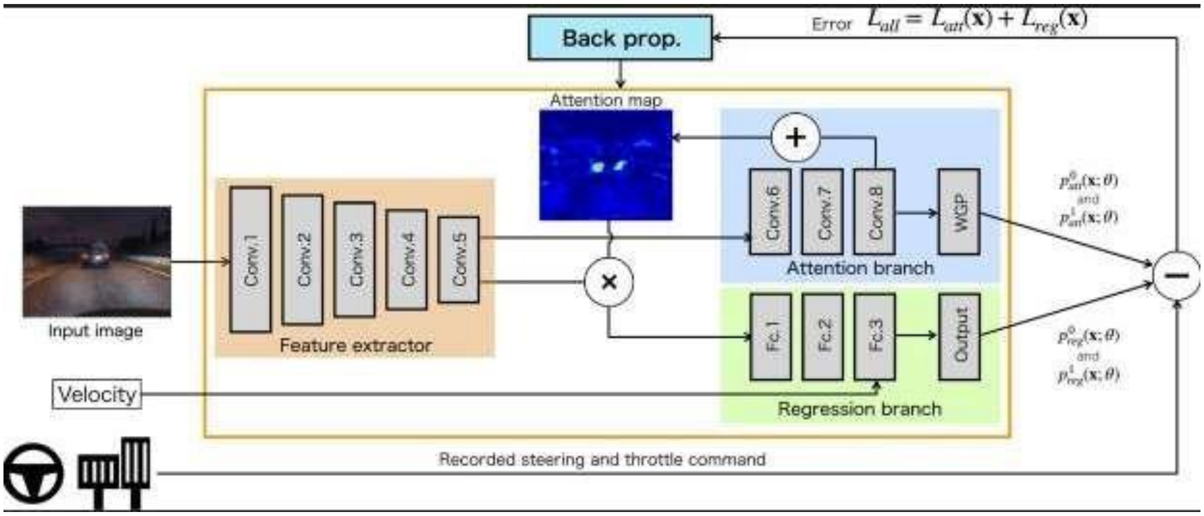
CNN-based end-to-end learning has a problem where the basis of output control value is not known. To address thisproblem, research is being conducted on an approach on the judgment grounds (such as turning steering wheel to the left or right and stepping on brakes) that can be understood by humans.

The common approach to clarify the reason of the network decision-making is a visual explanation. Visual explanation method outputs an attention map that visualizes the region in which the network focused as a heat map. Based on the obtained attention map, we can analyze and understand the reason of the decision- making. To obtain more explainable and clearer attentionmap for efficient visual explanation, a number of methods have been proposed in the computer vision field. Class activation mapping (CAM) generates attentionmaps by weighting the feature maps obtained from the last convolutional layer in a network. A gradient-weightedclass activation mapping (Grad-CAM) is another common method, which generates an attention map by using gradient values calculated at backpropagation process.
This method is widely used for a general analysis of CNNsbecause it can be applied to any networks. This figure shows example attention maps of CAM and Grad-CAM.



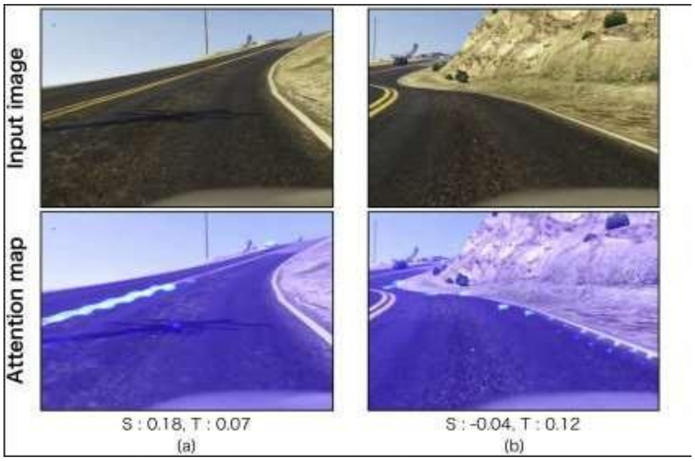
Visual explanation methods have been developed for general image recognition tasks while visual explanation for autonomous driving has been also proposed . Visualbackprop is developed for visualize the intermediate values in a CNN, which accumulates feature maps for each convolutional layer to a single map. This enables us to understand where the network highly responds to the input image.

Reference proposes a Regression-type Attention Branch Network in which a CNN is divided into a feature extractor and a regression branch, as shown in the figure, with an attention branch inserted that outputs an attention map that serves as a visual explanation. By providing vehicle speed in fully connected layers and through end-to-end learning of each branch of Regression-type Attention Branch Network, control values for steering and throttle for various scenes can be output, and also output the attention map that describes the location in which the control value was output on the input image.



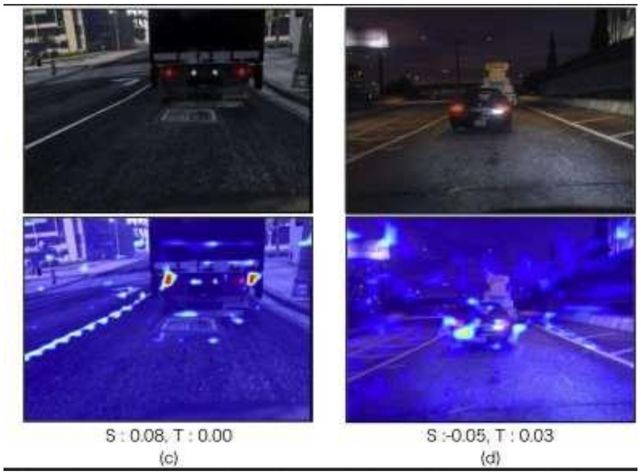
Back prop.
Error $L_{all} = L_{att}(\mathbf{x}) + L_{reg}(\mathbf{x})$
Attention map
Input image
Conv.1
Conv.2
Conv.3
Conv.4
Conv.5
Feature extractor
+
×
Conv.6
Conv.7
Conv.8
WGP
Attention branch
$p^0_{att}(\mathbf{x};\theta)$ and $p^1_{att}(\mathbf{x};\theta)$
Fc.1
Fc.2
Fc.3
Output
Regression branch
$p^0_{reg}(\mathbf{x};\theta)$ and $p^1_{reg}(\mathbf{x};\theta)$
−
Velocity
Recorded steering and throttle command

This figure below shows an example of visualization of attention map during Regression-type Attention Branch Network-based autonomous driving. S and T in the figure is the steering value and throttle value, respectively. (a) shows a scene where the road curves to the right where there is a strong response to the center line of the road, and the steering output value is a positive value indicating the right direction. On the other hand, (b) is a scene where the road curves to the left, the steering output valueis a negative value indicating the left direction, and the attention map responds strongly to the white lineon the right. By visualizing the attention map in this way, it can be said that the center line of the road and the position of the lane are observed for estimation of the steering value. Also, in the scene where the car stops as shown in (c), the attention map strongly responds to the brake lamp of the vehicle ahead. The throttle output is 0, which indicates that the accelerator and the brake are not pressed. Therefore, it is understood that the condition of the vehicle ahead is closely watched in the determination of the throttle. In addition, the night travel scenario in (d) shows a scene of following a car ahead, and it can be seen that the attention map strongly responds to the car ahead because the road shape ahead is unknown. It is possible to visually explain the judgment grounds through output of attention map in this way.



S : 0.18, T : 0.07
(a)

S : -0.04, T : 0.12
(b)



S : 0.08, T : 0.00
(c)

S :-0.05, T : 0.03
(d)

- **FUTURE CHALLENGES**

The visual explanations enable us to analyze and understand the internal state of deep neural networks. One of the future challenges is explanation for end users, i.e., passengers on a self-driving vehicle. In case of fully autonomous driving, for instance, when lanes are suddenly changed even when there are no vehicles ahead or on the side, the passenger in thecar may be concerned as to why the lanes were changed. In such cases, the attention map visualization technology enables people to understand the reason for changing lanes.

However, visualizing the attention map in a fully automated vehicle does not make sense unless a person on the autonomous vehicle always sees it. A person in an autonomous car, that is, a person who receives the full benefit of AI, needs to be informed of the judgment grounds in the form of text or voice stating, “Changing to left lane as a vehicle from the rearis approaching with speed.” Transitioning from recognition results and visual explanation to verbal explanation will be the challenges to confront in the future. In spite of the fact that several attempts have been conducted for this purpose, it does not still achieve sufficient accuracy and flexible verbal explanations.

Also, in the more distant future, such verbal explanation functions will eventually not be used. At first, people who receive the full benefit of autonomous driving find it difficult to accept, but a sense of trust will be gradually created by repeating the verbal explanations. Thus, if confidence is established between autonomous driving AI and the person, the verbalexplanation functions will not be required, and it can be expected that AI-based autonomousdriving will be widely and generally accepted.

- **CONCLUSION**

This work explains how deep learning is applied in image recognition tasks and introduces the latest image recognition technology using deep learning. Image recognition technology using deep learning is the problem of finding anappropriate mapping function from a large amount of data and teacher labels. Further, it is possible to solve several problems simultaneously by usingmultitask learning. Future prospects not only include “recognition” for input images, but also high expectations for the development of end-to-end learning and deep reinforcement learning technologies for “judgment” and “control” of autonomous vehicles. Moreover, citing judgment grounds for output of deep learning and deep reinforcement learning is a major challengein practical application, and it is desirable to expand from visual explanation to verbal explanation through integration with natural language processing.

- **REFERENCES**

1) D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng
Person re-identification by multi- channel parts-based cnn with improved triplet loss function
Proc. of IEEE Conference on Computer Vision and Pattern Recognition (27-30 June 2016)

2) P. Viola, M. Jones
Rapid object detection using a boosted cascade of simple features
Proc. of IEEE Conference on Computer Vision and Pattern Recognition (8-14 Dec. 2001)

3) N. Dalal, B. Triggs
Histograms of oriented gradients for human detection
Proc. of IEEE Conference on Computer Vision and Pattern Recognition (20-25 June 2005)

4) G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray
Visual categorization with bags of keypoints
Proc. of ECCV Workshop on Statistical Learning in Computer Vision (2004)

5) H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia
Pyramid scene parsing network
Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2017)

6) D.G. Lowe

Distinctive image features from scale-invariant keypoints
Int. J. Comput. Vis., 60 (2004), pp. 91-110

7) M. Bojarski, D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. Jackel, M. Monf ort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba
End to End Learning for Self-Driving Cars
arXiv preprint, arXiv:abs/1604.07316
(2016)

8) Q. Li, L. Chen, M. Li, S.L. Shaw, A. Nüchter
A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios
IEEE Transactions on Vehicular Technology, 63 (2) (2013), pp. 540-555

9) R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra
Grad-CAM: visual explanations from deep networks via gradient-based localization
International Conference on Computer Vision (2017), pp. 618-626

10) J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata
Textual explanations for self-driving vehicles
European Conference on Computer Vision (2018), pp. 563-578

11) Y. Mori, H. Fukui, T. Hirakawa, N. Jo, T. Yamashita, H. Fujiyoshi
Attention neural baby talk: captioning of risk factors while driving
IEEE International Conference on Intelligent Transportation Systems (2019)

12) Lassa, Todd (January 2013). "The Beginning of the End of Driving". Motor Trend.
13)] Kanade, Takeo (February 1986). Autonomous land vehicle project at CMU. CSC '86 Proceedings of the 1986 ACM Fourteenth Annual Conference on Computer Science. Csc '86. pp. 71–80.
14) "Navlab 5 Details". cs.cmu.edu.
15) "VisLab Intercontinental Autonomous Challenge: Inaugural Ceremony – Milan, Italy". (2013)
16) "Volvo Cars and AT&T Enter Multi-Year Agreement to Connect Future Models in U.S. and Canada" (Press release). AT&T Corporation. 16 April 2014.
17) "Automated Driving – Levels of Driving Automation are Defined in New SAE International Standard J3016" (PDF).
18) Elliott, Amy-Mae (25 February 2011). "The Future of the Connected Car". Mashable [12] Meola, Andrew. "Automotive Industry Trends: IoT Connected Smart Cars & Vehicles".
19) PwC Strategy& 2014. "In the fast lane. The bright future of connected cars".
20) Staff, CAAT. "Automated and Connected Vehicles". autocaat.org
21) J. Long, E. Shelhamer, T. Darrell
Fully convolutional networks for semantic segmentation

Proc. of IEEE Conference on Computer Vision and Pattern Recognition (7-12 June 2015)

22) S. Ren, K. He, R. Girshick, J. Sun
“Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence