International Journal of
Advanced Science and Convergence

ISSN : 2619-8150
Volume 2, Number 4, December 2020
https://doi.org/10.22662/IJASC.2020.2.4.017

Comparison Analysis and Case Study for Deep Learning-based Object Detection Algorithm

Min-hye Lee[1] and Hyung-Jin Mun[2]

[1] Professor, Convergence College of Liberal Arts, Wonkwang University, South Korea

[2]Professor, Dept. of Information & Communication Engineering, Sungkyul University, South Korea

Abstract

Background/Objectives Deep learning which main technology in AI has high growth with being applied to field of speech recognition and Image classification. Especially, Deep learning technology in the field of Image classification is being applied as a core technology to Self-driving and crime prevention monitoring system that is recently emerging as the future industry. **Methods/Statistical analysis**: Various algorithm which is improved and developed CNN being able to do image process is suggested as Deep learning model in image recognition field. In this paper, we introduce various object detection algorithm including CNN. And explore most representative algorithms just R-CNN, Fast R-CNN, Faster R-CNN and difference between versions of YOLO devised to detect and track in real time. **Findings**: This paper evaluates deep learning algorithm's performance by comparative analysis about mAP (mean average precision) and FPS (frames per second). In result of performance evaluation, YOLO algorithm is confirmed as that It shows excellent result in speed that detects and recognizes object and accuracy in real time system environment. Finally, we search cases in field of autonomous driving and access control system and home anti-crime system. **Improvements/Applications**: In this research, we can understand object detection algorithm among speech recognition technologies and proper field in each algorithm, apply security service based on image, recommend proper algorithm in various environment just like autonomous driving and security work, etc.

Index Terms

Deep Learning, Object Detection, Image Process, YOLO, CNN, Image Recognition Technology

Corresponding author : Hyung-Jin Mun
mun@sungkyul.ac.kr

• Manuscript received November 5, 2020.
• Revised December 4, 2020 ; Accepted December 20, 2020.
• Date of publication December 31, 2020.

I. INTRODUCTION

The development of Information Communication and Technology makes innovation around society convergence with fields of various industry. The AI(Artificial intelligence)technology is widely researched as tool that can solve and respond to various problems using big data and excellent deep learning algorithm in various industries such as security, transportation, finance, and medical care[1]. The AI is a computer-implemented human high-level information processing such as learning, reasoning, and cognition, and today's artificial intelligence technology is based on deep learning algorithms.

The deep learning is machine learning technology built on the basis of an artificial neural network for that computers can learn by themselves using multiple data. The deep learning has the advantage that obtains the optimal result from only given data through learning because Feature Extraction process is included in learning process, not the classic way to extract and change features in kind of problem manually[2]. The deep learning algorithm brought the breakthrough development in field of voice recognition and image extraction by that it can recognize data have complex shape and much number of possibility with high quality.

In the field of speech recognition, it is possible to improve the quality of speech recognition services by reducing errors and improving accuracy by learning vast amounts of speech data of various people with deep learning.

In the field of image detection, CNN(Convolution Neural Network)-based deep learning algorithms have been introduced and have become an important model for image classification and object detection. [3-5]

In this paper, we introduce about speech recognition that can be applied in intelligent object detection method and object detection deep learning algorithm and explain related cases that applied this algorithm. Section 2 introduces the kinds of deep learning algorithms about voice and image recognition, and section 3 explains object detection deep learning algorithms for object tracking. In Section 4 we analyze comparation about object detection algorithm, and cases applied by object detection.

II. RELATED WORKS

As the calculation performance of the computer improved, the overfitting problem in ANN(Artificial Neural Network) and slow speed of learning time resolves and then DNN(Deep Neural Network) that improves result of learning by increasing hidden layers in model into two or more is appeared(Fig. 1).

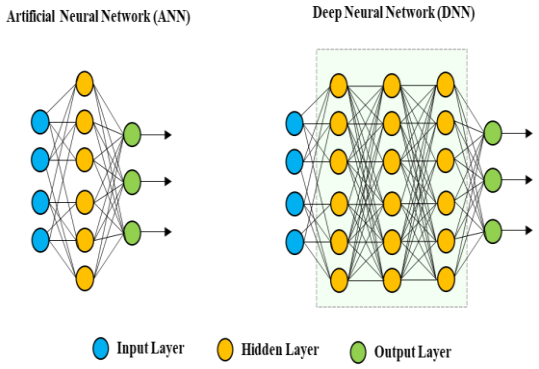

Fig 1. Structure of ANN and DNN

This is a representative deep learning model in which a computer generates a classification label by itself, distorts space, and repeats the process of classifying data to derive optimal result[6]. The DNN is widely used algorithm by much data and repeated learning and previous learning and error back propagation method, has become a technology that proposes RNN(Recurrent Neural Network), LSTM(Short-Term Memory models) [4,7,8], GRU(Gated Recurrent Units)[9] that is used in speech recognition and CNN[3] that is representatively used algorithm in image recognition.

This section introduces a deep learning model that is representatively used for speech recognition and image recognition.

A. Speech Recognition Technology

Speech recognition means that a series of processes that extract features of a speech signal, analyze it, and convert it into words or sentences.

The speech recognition technology can be divided into 4 generations according to the development process. In the 1960s, the speech recognition equipment of first generation that can recognize simple numbers or syllables was developed by IBM and a neural network-based language model was applied through the second generation using DTW technology, in the 1980s. In the 1990s, the speech recognition technology was first commercialized as the large vocabulary speech recognition system was developed along with discrimination learning techniques such as MCE (Minimum Classification Error) and MMI (Multi Media Interface) for minimizing speech recognition errors.

As the internet diffusion has become active, it is repeatedly developed into dialogue system that access information service and speech recognition system that can transfer, understand, and summarize information by speech[10]. With the proposal of Deep speech in 2013, research on deep learning using large amounts of speech data has gradually developed, and research on speech recognition based on CNN, RNN, and LSTM is still in progress.

Speech recognition-based interface is widely used in various industries because communication is more natural and free from visual restrictions than text or graphic interfaces. Voice recognition assistant platforms such as Google Assistant of Google, Siri of Apple, Bixby of Samsung, and Genie of LG U+ and NUGU of SKY which are combined with speech recognition technology in various fields are representative examples[4, 5].

The CNN model that extracts and classifies features of data with deep learning was used to recognize speech, but CNN has a limitation that it can't take into account the voice information according to the time flow of time series data. To overcome these limitations, RNN-based deep learning models that learn considering the time properties of signals have begun to be applied[11,12].

RNN (Recurrent Neural Network)[7] is an artificial neural network model that learns sequential data and performs classification or prediction. It is mainly used for speech recognition, language processing, and sequence-to-sequence. As shown in Fig. 2, the RNN has an internal cyclical structure and reflects the past learning to the current learning through weight using the cyclical structure. In the conventional DNN, parameters were independent for each layer, but since the RNN shares these parameters, the current output result is affected by the previous result, which is dependent on time.

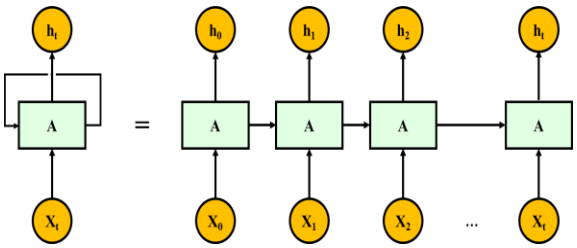

Fig 2. Structure of RNN[4]

LSTM (long short-term memory) [8] is a model that appeared to solve the vanishing gradient problem and long-term dependency problem of RNN. LSTM implements 3 gates(input/forget/output) and control state information of current node(Fig 3).

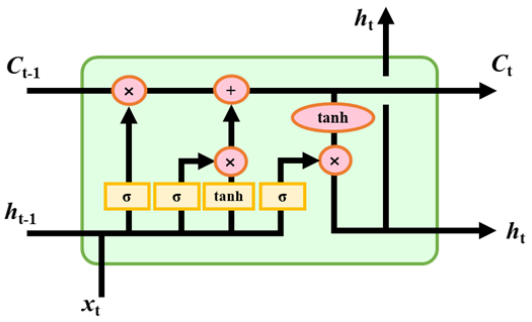

Fig 3. Structure of LSTM [4]

Forget gate makes decision previous state information, input gate makes decision whether new information inputted is saved or not, output gate control outputs of updated cell to learn long-term dependency.

B. Image Recognition Technology

CNN (Convolution Neural Network)[3] is a technology that mimics the structure of the human optic nerve, and automatically learns features necessary for recognition in image processing to character recognition, image recognition, and object recognition. CNN was proposed to solve the problems of training time, network size, and number of variables[14].

We introduce LeNet, AlexNet, VGGNet, GoogLeNet, ResNet, and DenseNet, which are representative CNN models used for image recognition.

LeNet

After LeNet-5 was proposed by LeCun in 1998, the CNN algorithm has shown great prominence in the field of image classification[13]. In LeNet, LeNet-1 is proposed first in 1990 for the purpose of handwriting recognition of postal codes and checks it forms LeNet-5 by development. In classification algorithm which is a fully-connected multi-layer network (MLP), parameters increase exponentially according to the number of pixels in the input data. In addition, it is vulnerable to local distortion.

LeNet expands the perspective of looking at input data in one dimension to two dimensions and We tried to solve the existing problem by applying the concept of a spatial characteristic (local receptive field) through a convolutional layer using a 5×5 size filter. Also, apply the same kernel to the image within the network and Sub sampling was reflected through shared weight and average pooling[14].

Fig. 4 shows the structure of LeNet for 32×32 input data.

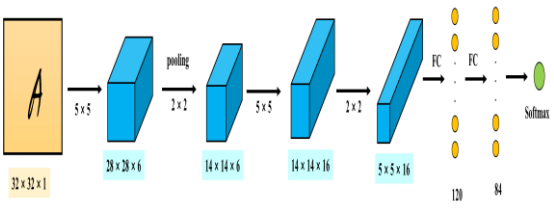

Fig 4. Structure of LeNet

AlexNet

AlexNet[3] is a CNN structure that showed excellent performance in the 2012 ILSVRC (ImageNet Large Scale Visual Recognition Challenge) competition and received attention in the

image classification field. Previous neural networks showed an accuracy of about 75% and the limitation of performance in the field of image classification, but AlexNet showed an accuracy of about 84% and prepared the foundation for image classification.

AlexNet is consisted of 8 layers (Fig. 5), 5 convolutional layers and 3 fully connected layers. AlexNet reduces the calculation speed of neural networks by using ReLU, an activation function and pooling was performed with a 3×3 filter size at 2 stride intervals to improve the overfitting problem. Fig. 5 shows the structure of AlexNet for input data [15].

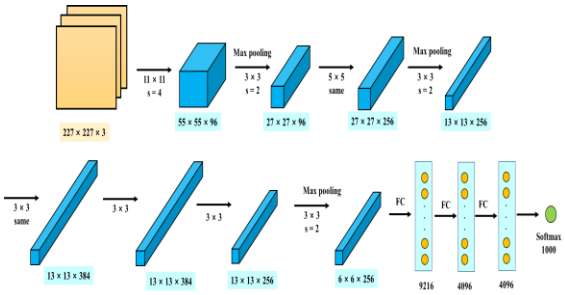

Fig. 5. Structure of AlexNet

VGGNet

VGGNet[16], which is developed by VGG search team in oxford university is suggested to confirm relationship between feature and depth of network. Different from that most previous models before VGGNet-16 are consisted of 8 layers. However, layer of network is deeper and performance is improved after VGGNet-16. VGGNet-16 is model that is consisted of 16 layers, this is a model in which the size of the convolution filter is fixed at 3×3 to check the effect of the network depth on the performance. In VGGNet, instead of using the 5×5 convolution filter once, the 3×3 convolution filter is used twice. One 5×5 convolution filter and two 3×3 convolution filters result in a feature map of the same size.

However, when applying the 3×3 convolution filter twice, It has advantage that learn decision function better because ReLU which is non-linear function included and shows performance improvement in speed because number of parameters is decreased. VGGNet is a CNN model that is widely used because it is easy to implement and has good performance with a simple structure in which layers are stacked sequentially.

Fig. 6 shows the structure of VGGNet about input data[17,18].

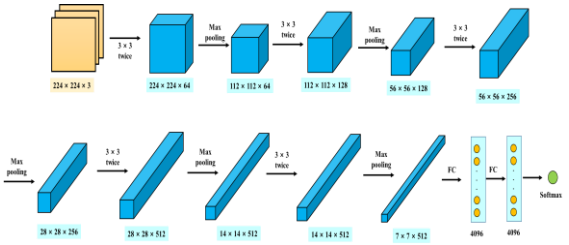

Fig. 6. Structure of VGGNet

GoogLeNet

In order to improve the object recognition performance, as the CNN model is composed of a complex and deep layer of the network, a method to solve the problem of overfitting and computational load caused by an increase in the number of parameters was sought. GoogLeNet[19], which ranked first in the 2014 ILSVRC, is a model consisting of a total of 22 long and complex layers. For solving this problem, It introduces concept of inception to design effective network to solve this problem of calculation amount with keeping network deep.

The inception module reduces the dimension by using several filters of 1×1, 3×3, and 5×5 sizes and analyzes information on the height and width of the image variously with decreasing calculation amount. And GoogLeNet solves the problem of increasing the calculation amount by reducing the number of weights by using a global average pooling method, unlike fully connected (fc) used by many other representative models (AlexNet, VGGNet, etc.). To overcome the vanishing gradient problem, two auxiliary classifiers were added in the middle of the network. Fig. 7 shows the structure of GoogLeNet[18,20].

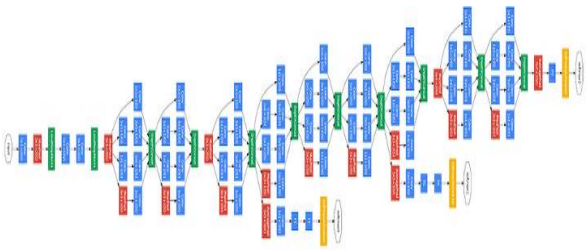

Fig. 7. Structure of GoogLeNet

ResNet

In the CNN structure, when the network layer is deepened beyond a certain level, there is a problem that vanishing gradient occurs, so that learning is not performed properly. ResNet[21] proposed the residual block method as a solution for the problem, and as a result, it showed excellent performance with a 3.6% classification error rate.

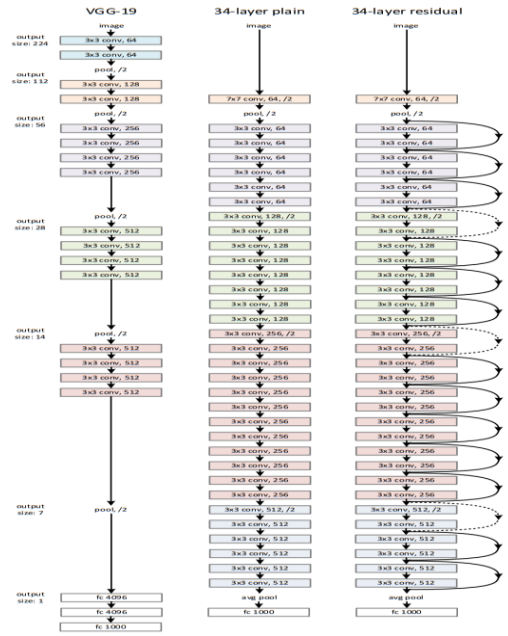

Fig. 8. Structure of ResNet

ResNet is basically similar to VGGNet structure and has 152 layers network structure that is 8 times deeper than VGGNet (Fig. 8). Residual Block which is main technology in ResNet make skip connection so that gradient flows. It has ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152 labeled by number of layers, by using method of Residual Block[22].

DenseNet

DenseNet[23] is suggested as neural network in 2016 to improve Gradient Vanishing, added the Dense connectivity stacking the layers in the front despite deepening network. If the Residual learning of the ResNet model was connected with the input and output in a single layer (function) and In the DenseNet model, all outputs of dense connectivity are connected with the inputs of all layers.

Dense connectivity connects with feature map of early layer is near the first input image to the late layer so that keeps information of layer, Prevents backpropagation signals from fading.

By that, the problem that as the neural network gets deeper, the feature of the input signal become faded. and Gradient Vanishing is resolved.

Fig. 9 shows this DenseNet's core structure[18].

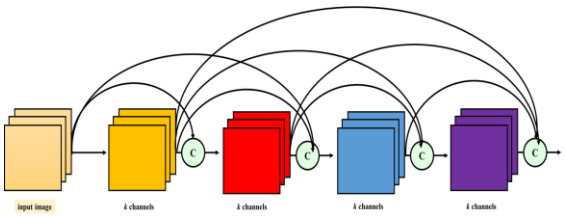

Fig. 9. Structure of DenseNet

III. OBJECT DETECTION TECHNOLOGY

The object detection technology is classified as traditional methods using local features, contours, and statistical transformation techniques and as a deep learning-based approach that Even if a person does not directly extract the features of the image, it can process from feature extraction to classification in a batch. The representative examples of the traditional method are Haar-like feature[26] method that recognizes faces by measuring changes according to contrast and SIFT (Scale Invariant Feature Transform)[25] that extracts feature vectors as center around feature points in an image and HOG (Histogram of Oriented Gradients)[26] that uses the distribution of the object's gradient direction as a feature vector.

Object detection models using deep learning are CNN-based models described in section 2. For example, There are R-CNN, fast R-CNN, faster R-CNN, Mask R-CNN that extracts a region of interest and classifies objects within the region and YOLO(You Only Look Once) that divides one image into a grid to distinguish objects at a time. In this section, we will explain these algorithms and models.

A. Two-Stage Detector

Region-based CNN (R-CNN)

CNN can express local information around pixels through calculation of convolution neural network, but it shows limitations in recognizing many objects and detecting the location of objects.

R-CNN (Region-based Convolutional Neural Networks)[27] applies Selective Search and region proposal using algorithm, feature extraction, SVM Classifier, and Linear Regression. It is an object detection model that solves the problem by that. R-CNN finds a box where there are likely 2,000 objects by applying the Selective Search algorithm to the input image(Fig. 10). Area of 2000 box that is found by Selective Search extracts feature vector by passing CNN as size of 227×227 after warp.

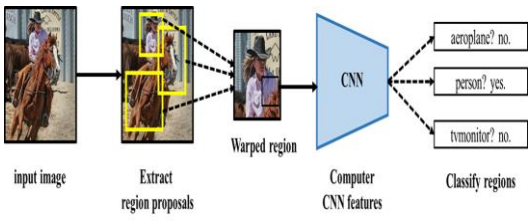

Fig. 10. Structure of R-CNN

Vectors extracted from CNN classifies objects in candidate area by using SVM Classifier and passes Non-Maximum Suppression and Bounding Box

Regression to increase the accuracy of the detected object and adjust position of bounding box which points position of object.

Fast R- CNN

R-CNN has problem that its calculation takes too much time by performing with separating CNN, SVM, and regression learning.

Fast R-CNN[28] suggested end-to-end method that makes CNN, classification, bounding box regression to learn on only one network and resolve the problem of R-CNN. All of input image is being passed CNN that is already learned (Fig.11). and we extract feature map and process RoI Pooling then get feature vector of fixed size. The feature vector passes fully connected layer and softmax. After that, It is divided classifier that classifies RoI(Region Of Interest) as something and part that settles position of box by Bounding box regression.

Fast R-CNN is not indifferent from R-CNN. But it uses end to end method to have feature that its learning speed and inference speed and accuracy are improved.

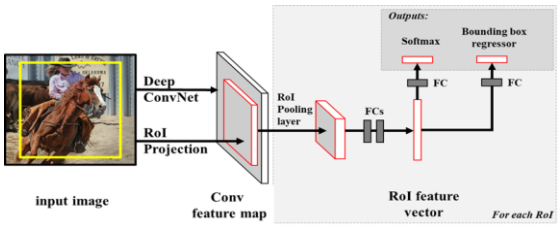

Fig. 11. Structure of Fast R-CNN

Faster R- CNN

Faster R-CNN[29] is model that recognizes an object by applying a separated Region Proposal Network (RPN) instead of a selective search algorithm as a candidate region generation method. Faster R-CNN solves the problem of inefficient learning and execution speed to perform a selective search algorithm, which is a part that generates candidate regions in Fast R-CNN model, independently of CNN by using RPN.

The structure of the Faster R-CNN model calculates RoI by passing the feature map extracted through CNN from the input image to the RPN. After that, it processes ROI obtained by that, and perform classification and regressor after RoI Pooling (Fig. 12).

The difference from R-CNN and Fast R-CNN models is that the feature map extraction process and candidate region generation are performed in a series of networks, and the calculation speed is faster than the previous model.

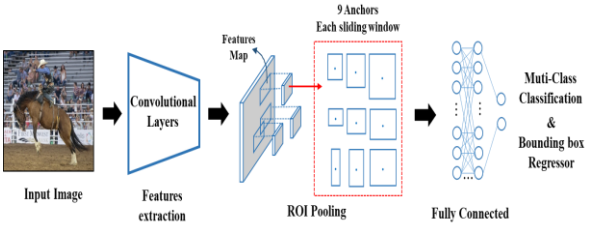

Fig. 12. Structure of Faster R-CNN

B. One-Stage Detector

SSD

SSD(Single Shot Multibox Detector) [46] is object extractor that trains and extracts without change of input image. YOLO divides input image into grid of 7×7 size and predict bounding box each grid. So it has difficulty to predict object smaller than size of gird. The problem of decreasing accuracy existed because use only feature map of last passage while passing network. SSD uses Multi-scale feature maps for solving this problem. In the feature map for each step, it applies method that performs all of object detection to detail of object and introduces anchor of Fast R-CNN to extract various shapes of object. We present this structure of SSD in Fig. 13[30].

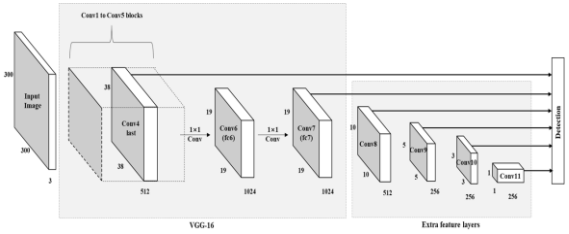

Fig 13. Structure of SSD

YOLO

The object detection deep learning model, YOLO (You Only Look Once)[30] is algorithm that detects input image once and predicts position of object in that image. R-CNN performs classification with classifier that is in Bounding box from using region proposal to detect object. It is slow for calculation and difficult for optimization because each process that mediates bounding box remove overlapped detection should be trained independently.

YOLO is the model that redefines a series of procedures that separates bounding box multidimensionally from image pixels finds class probabilities into a single regression problem. It divides the input image into a grid, each grid cell calculates confidence score about each kind and whether object exists or not to recognizes object of the local

-YOLO v1

YOLO is an object detection deep learning model proposed by Jeseph Redmon in 2016. Its feature is that uses a model integrated into a single network. (Fig. 14)

This method is less accurate than the existing R-CNN model, it has an advantage of being suitable for real-time object detection because it shows high performance in terms of calculation speed.

Given the input image, YOLO divides it into S×S (7×7) grid cells, and then each cell predicts the bounding box and the Confidence Score.

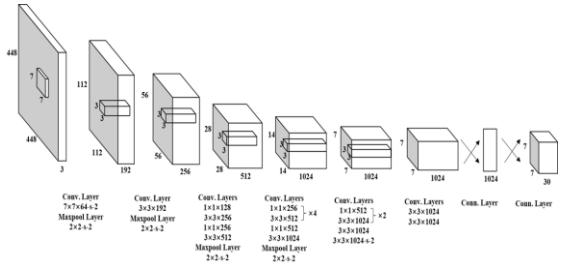

Fig. 14. Structure of YOLOv1

-YOLO v2

YOLO v2[31], released in 2017, designed to utilize large amounts of classification data is later model of YOLO v1, suggests better accuracy and baster speed and Stronger prediction of more classes. First, it added Batch Normalization layer for better speed and accuracy, and attempted stabilization of initial learning by setting an anchor box.

The classification network was fine-tuned to improve performance in high-resolution images. As a method for speed faster, many object detection models use VGGNet as the network at the front end.

YOLO v2 uses the independent Darknet based on GoogLeNet to significantly reduce the amount of calculation. In addition, in YOLO v1, there was a limitation of detection of 20 classes, but in YOLO v2, it was improved to detect in 9000 classes.

Fig. 15 is an image showing the network structure used in this paper.

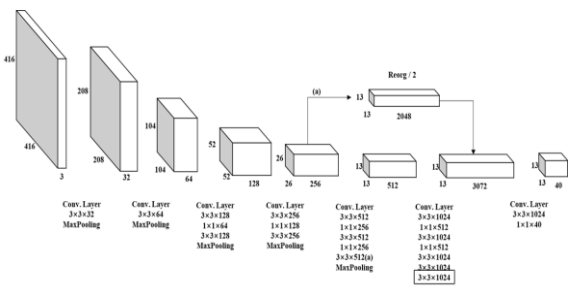

Fig. 15. Structure of YOLOv2

-YOLO v3

YOLO v3[31] is a deep learning model that improves the YOLO v2 model by using ResNet's Residual Block.

It is predicted situation with using the Objectness Score which calculates the presence or absence of an object in each bounding box by applying logistic regression. YOLO v3 extracts feature from various scales with using three features pyramid network.

YOLO v3 improved bounding box prediction, class prediction, feature detector and repetitive detection prevention through this process. Fig. 16 is an image showing the structure of YOLO v3 [32,33].

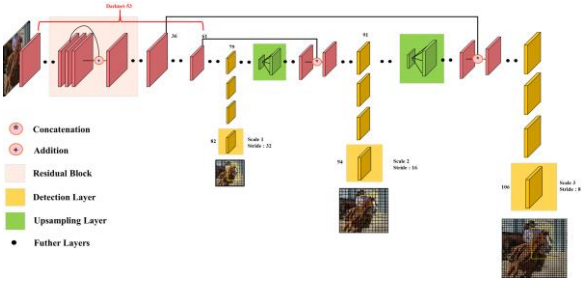

Fig. 16. Structure of YOLOv3

-YOLO v4

Recent neural network has high accuracy but has advantage that it needs a number of GPU for effective learning.

For solving this problem, YOLO v4[34] model applies method of BOF(Bag of Freebies), BOS(Bag of Specials) and suggests method of high accuracy is high in common learning environment and can perform quick object detection learning.

YOLO v4 model improves performance by applying various latest deep learning techniques just like WRC (Weighted-Residual-Connections), CSP (Cross-Stage-Partial-Connections), CmBN (Cross mini-Batch Normalizations), SAT (Self-Adversarial-Training), Mish Activation, Drop Block Regularization, CIOU Loss.

Fig. 17 is the basic structure of the YOLO v4 model[35].

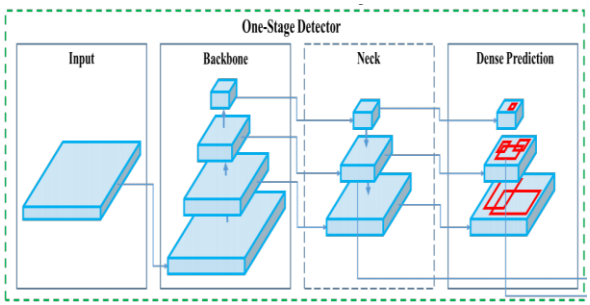

Fig. 17. Structure of YOLOv4

IV. ALGORITHM AND APPLICATION CASES

A. Algorithm comparison analysis

This section compares and analyzes mAP (mean average precision) and FPS (frames per second) about representative deep learning models in image processing.

In the field of computer vision, the performance of the CNN algorithm for object detection can be evaluated usually with using AP (Average Precision) and mAP (mean average precision). In the case of multiple object classes, the AP per class is obtained, summed, and then divided by the number of classes of objects to obtain average.

PASCAL VOC and MS COCO are used as datasets for performance evaluation of object detection deep learning models. PASCAL VOC is a dataset commonly used in the field of object detection, and most deep learning models use this dataset to evaluate performance, and recently mAP performance evaluation index about MS COCO have been added as a result.

FPS (frames per second) and inference time are used as index to evaluate the speed of the deep learning model. In object detection, FPS refers to the detection rate per second, and inference time refers to the time taken to detect one frame.

In general, for real-time video, the mAP index above 30 FPS is evaluated as performance. By the type of training dataset, the network size, and the performance of the graphics card, the mAP and FPS values has change, so the performance comparison result about the algorithm may be different.

Fig. 19 is a figure comparing performance of representative algorithms.

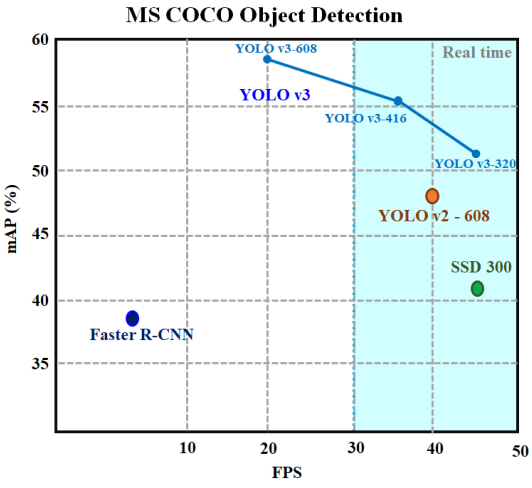

Fig 19. Comparison analysis of speed and accuracy

B. Application cases and discussion

Object detection using deep learning models is being applied in the field of autonomous driving, access control systems, and home security management.

The semantic segment technology that has attracted attention in the field of autonomous driving for a long time and classifies images obtained from cameras into major object units such as roads, lanes, structures, people, and vehicles.

It detects and recognizes a specific object required for driving in a complex photo in which several objects are distributed.

For smooth driving, it divides lanes in the road area and recognizes people and vehicles, which are dynamic objects, to prevent dangerous situations.
Also, by recognizing the traffic lights, it is possible to determine whether to drive in a specific situation.

As research cases for such autonomous driving, researches for detecting vehicle light heads in images by using R-CNN models and YOLO models and methods for identifying and tracking pedestrian behavior patterns are being searched actively[56-59].

Recently research on autonomous driving by using drones is also being processed as a method for detecting disasters in buildings outside the road and in water environments.

Face recognition technology by using deep learning is being used as the method for recognizing a visitor in an access management system.

It can be applied to a monitoring system that can check whether person wears mask or not when he enters the building and generate an alarm if he wears mask incorrectly and can be used as a face-based user authentication that checks whether people are entering or not. And then, object detection technology by using such deep learning can be applied to security management systems.

It can be applied to a system that recognizes the license plate of a vehicle to calculate parking fees and controls vehicle access, or the tracking stranger system that recognizes an object through a mobile CCTV and tracks the object by CCTV in the movement path of the object. In addition, after recognizing a person, the physical condition of the person is checked, and based on this, it is possible to detect a person's behavior by using a skeleton model or to recognize the number of visitors by recognizing a person in the screen. In addition, it can confirm whether fire is happened or not by object detection with cases may occur in areas where it is difficult for security personnel to directly identify fires or crimes in all places and checks whether a person is a criminal or not [62-64].

V. CONCLUSION

As image processing technology develops, various deep learning-based algorithms have been proposed.

In particularly, there are R-CNN, Fast R-CNN, and Faster R-CNN, which extend CNN. This is an object detection algorithm with high accuracy but slow processing speed. There are technologies that need to detect objects in real time, such as autonomous driving. For this, yolo was proposed as an algorithm with low accuracy but fast processing speed. Accuracy can be improved through network size.

CNN has a good object identification rate, but it is difficult to detect in real time, so it is used to recognize objects in a number of classes, Although the object identification rate of YOLO is lower than that of R-CNN, but real-time detection is possible, so it can be used for fast detection in a small number of classes. It is an algorithm suitable for CCTV that needs to recognize people or for autonomous vehicles that need to detect obstacles.

It is necessary to select an appropriate algorithm in various fields requiring object detection technology.

This research introduced the object detection algorithm in image processing, and especially compared and analyzed YOLO, a real-time object detection algorithm.

For future research, a research on the design and development of a system that can detect visitors in real time based on YOLO and track visitors that are difficult to be identified will appear.

ACKNOWLEDGMENT

This work is supported by Research Institute of 4th Industrial Revolution Technology (RI4IRT)

REFERENCES

[1] T. B. Yoon. (2020). Weekly Technology Trends Artificial Intelligence Trends and Technology Service Cases, *Weekly ICT Trends*, ITFind, 1938, 2-12, Retrieved from https://www.itfind.or.kr/WZIN/jugidong/1938/file6848593148610205154-1938(2020.03.18)-10.pdf

[2] B. H. Kim, & B. T. Zhang. (2017). Deep learning: the cutting-edge technologies driving artificial intelligence, Technical Report: BI-17-001, Retrieved from https://bi.snu.ac.kr/Publications/tech-report/bhkim170416.pdf

[3] A. Krizhevsky, I. Sutskever and G. E.Hinton. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *In Advances in neural information processing systems*, 1097-1105.

[4] C. U. Chun, H. J. Lee, & K. B. Kim. (2020. April). Voice AI market trends and business opportunities, ISSUE MONITOR, KPMG, 126, 4 Retrieved from https://home.kpmg/kr/ko/home/insights/2020/04/issue-monitor-126.html

[5] Y. H. LEE. (2017). Speech/Audio Processing based on Deep Learning. *Broadcasting and Media Magazine*, 22(1), 47-58

[6] S. H. Park, & D. S. Choi. (2017). Artificial intelligence security issues, *Journal of the Korean Society for Information Security*, 27(3), 27-32

[7] K. W. Kug. (2019). Artificial intelligence technology and application examples by industry, *Weekly Technology Trend*, 20, 15-27 Retrieved from https://www.itfind.or.kr/WZIN/jugidong/1888/file6111801471006205940-188802.pdf .

[8] http://www.irobotnews.com/news/articleView.html?idxno=22194

[9] Lee. J. G, Jun. S. Cho, Y. W. Lee, H. Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: general overview. Korean journal of radiology, 18(4), 570-584. DOI : 10.3348/kjr.2017.18.4.570

[10] Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.

[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780. DOI : 10.1162/neco.1997.9.8.1735

[12] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[13] http://www.kocca.kr/knowledge/publication/ct/__icsFiles/afieldfile/2011/12/07/87NEmyIcVWMc.pdf

[14] http://infosec.pusan.ac.kr/wp-content/uploads/2017/11/CNN-and-RNN-%EC%9D%B4%EB%A1%A0.pdf

[15] Feasibility of Deep Learning Algorithms for Binary Classification Problems (2017) DOI : https://doi.org/10.13088/jiis.2017.23.1.095

[16] http://166.104.231.121/ysmoon/mip2017/lecture_note/%EC%A0%9C9%EC%9E%A5.pdf

[17] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[18] https://kjhov195.github.io/2020-02-10-CNN_architecture_1/

[19] Hong. J. Y, & Jung. Y. J. (2020). Evaluation of Deep-Learning Feature Based COVID-19 Classifier in Various Neural Network. *Journal of radiological science and technology*, 43(5), 397-404.

[20] Simonyan. K, & Zisserman. A, (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[21] https://bskyvision.com/504

[22] Chung. S, & Chung. M. G. (2019). Pedestrian Classification using CNN's Deep Features and Transfer Learning. *Journal of Internet Computing and Services*, 20(4), 91-102.

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1-9.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in Thirty-First AAAI Conference on Artificial Intelligence, 2017. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14806
[25] He. K. Zhang, X. Ren. S, & Sun. J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.
[26] Rhyou. S.-Y, Kim. H.-J, & Cha. K.-A. (2019). Development of Access Management System based on Face Recognition using ResNet. *Journal of Korea Multimedia Society* , 22(8), 823–831. Retrieved from https://doi.org/10.9717/KMMS.2019.22.8.823
[27] Huang G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition , 4700-4708.
[28] Viola. P, & Jones. M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, I-I). IEEE.
[29] D. G. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, 2, 1150-1157, 1999.
[30] N. Dalal and B. Triggs.(2005). Histograms of Oriented Gradients for Human Detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, 886-893. Retrieved from https://doi.org/10.1109/cvpr.2005.177
[31] Girshick. R. Donahue, J. Darrell. T, & Malik. J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (580-587).
[32] Girshick. R. (2015). Fast r-cnn. *In Proceedings of the IEEE international conference on computer vision* ,1440-1448.
[33] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6), 1137-1149.
[34] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *In Proceedings of the IEEE international conference on computer vision*, 2961-2969.
[35] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 779-788. Retrieved from: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html