International Journal of Advanced Culture Technology Vol.8 No.4 119-128 (2020)
DOI https://doi.org/10.17703/IJACT.2020.8.4.119

IJACT 20-12-16

Object detection technology trend and development direction using deep learning

[1]NaeJoung Kwak, [2]DongJu Kim

[1]Prof. Dept. of Cyber and Security , Baejae Univ., Korea
[2]Prof. Postech Institute of Artificial Intelligence, POSTECH, Korea
E-mail: knj0125@pcu.ac.kr, kkb0320@postech.ac.kr

Abstract

Object detection is an important field of computer vision and is applied to applications such as security, autonomous driving, and face recognition. Recently, as the application of artificial intelligence technology including deep learning has been applied in various fields, it has become a more powerful tool that can learn meaningful high-level, deeper features, solving difficult problems that have not been solved. Therefore, deep learning techniques are also being studied in the field of object detection, and algorithms with excellent performance are being introduced.

In this paper, a deep learning-based object detection algorithm used to detect multiple objects in an image is investigated, and future development directions are presented.

Keywords: *Deep-learning, Object-detection, Image processing, Classification, Computer vision,*

1. INTRODUCTION

Object detection is a computer technology related to computer vision and image processing. It detects semantic objects of a specific class (eg, people, buildings, or cars) in digital images and videos, and is applied to various fields. In recent years, with the development of GPUs to speed up computation, object detection technology using deep learning is developed rapidly. Deep neural network, called deep learning, is a detailed technology in the field of artificial intelligence, and object detection uses Convolutional Neural Network (CNN)[1] suitable for image processing among various deep learning techniques. Deep learning methods for object detection based on the CNN technique are divided into two types. One detects objects in images in two stages, and the most representative one is Faster R-CNN [2]. This method detects the object bounding box of the candidate object called RPN (Region Proposal Network) in the first step, and performs image classification and optimization of the object bounding box using the features of each bounding box in the second step. The other processes object detection in one stage, such as YOLO (You only look once) [3] and SSD (Single shot multibox detector) [4]. YOLO divides the image into N×N grids and calculates the confidence of the accuracy of object recognition in the grid. The bounding box with the highest object recognition accuracy is detected using this confidence. The SSD extracts a feature map from the result of a CNN on the input image. By making the extracted feature map into several sizes, it is possible to detect objects of various scales by making small objects detection in large maps and large objects in small maps.

Deep learning-based object recognition technology has been continuously developed by deceiting these techniques. In this paper, we attempt to find out how to detect various objects based on deep learning and find out the future prospects.

Manuscript received: November 17, 2020 / revised: December 07, 2020 / accepted : December 13, 2020
Corresponding Author: kkb0320@postech.ac.kr
Tel:+82-2-3277-2793
Prof. Postech Institute of Artificial Intelligence, POSTECH, Korea

2. OBJECT DETECTION METHOD

The object detection technology based on deep learning is divided into a two-stage method and a one-stage method. The two-stage method is to find an object candidate region (ROI: Region of Interest) in the image using Selective Search[5], Region Proposal Network[2], etc., and to process class classification and bounding box regression for the found candidates for object detection. However, the object detector's execution speed is degraded because of the operation principle of extracting the ROI object and performing object classification and bounding box regression for each object of the ROI. The representative technology is Faster R-CNN[2]. The one-stage method detects an object with a deep learning model that performs unified detection, which simultaneously performs feature classification on the object of the input image and prediction of the bounding box.

2.1 R-CNN

R-CNN[6] is an object detection method proposed by Girshick et al. and showed for the first time that object detection performance can be significantly improved in the PASCAL VOC data set.

The R-CNN detector consists of four modules: candidate region generation, feature vector extraction, image classification, and bounding box regression. The first module generates 2000 class-independent object candidate regions (Region Proposal) by applying selective search method. The second module extracts a 4096-dimensional feature vector by processing the generated object candidate region as an input through the pre-trained CNN module. The CNN network consists of 5 convolutional layers and 2 fully connected layers, and the selected region is made into an image of a fixed size of 227x227 and used as an input. The third module classifies objects into one class using linear SVM. The last module, which is not essential, is bounding box regression for accurate bounding box prediction.

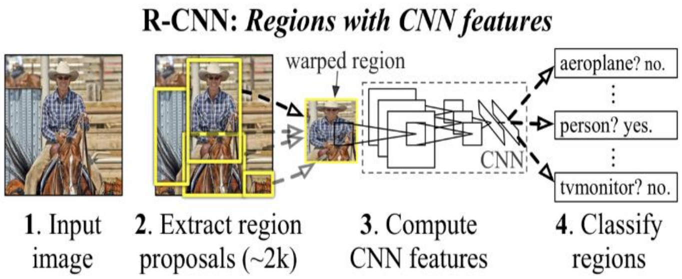

Figure 1. R-CNN object detection system

2.2 Fast R-CNN

Ross Girshick proposes an algorithm, called Fast R-CNN[7], that improves the speed and performance of R-CNN. R-CNN separates the image feature extraction (CNN) and classification model (SVM), and the regressor of the bounding box. Fast R-CNN processes everything with one network. After extracting features from the entire input image, it passes through a region of interest (RoI) pooling layer to generate fixed-sized inputs for classification and bounding box regression. In addition, by placing a softmax layer for classification, the CNN result generate a class, and the bounding box coordinates by placing the bounding box regression layer parallel to the softmax layer.

As a result of the experiment, in the PASCAL VOC 2007 data set, Fast R-CNN improved mAP by 0.9% compared to R-CNN, and training time was reduced by 9 times.

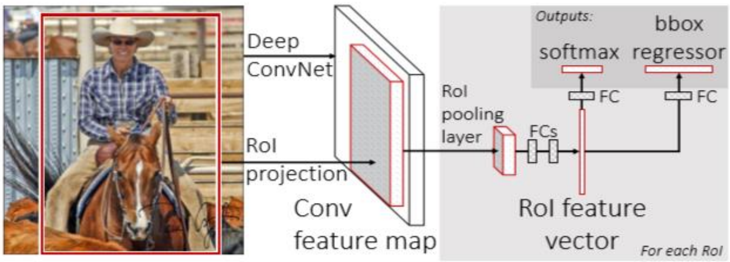

Figure 2. Fast R-CNN structure

2.3 Faster R-CNN

Although Fast R-CNN improves the shortcomings of R-CNN, it is slow because it proposes RoI using selective search. Faster R-CNN[2] proposes a region proposal network (RPN) to improve the prediction accuracy of region proposals with various sizes and aspect ratios. RPN receives an image and outputs a rectangular object proposal and objectness score, and shares convolutional layers with Fast R-CNN. To generate region proposals, slide an n x n window (usually 3 x 3) on the feature map extracted from the input image. For each sliding-window location, several region proposals are predicted at once. There are anchors that are used as candidates for the Bounding Box at each position of the sliding window. Usually, there are 9 anchors at each sliding window, 3 different aspect ratios and 3 different scales. Since multiple anchors or multiple region proposals (bounding boxes) may be duplicated per object, the number of region proposals is reduced by using a non-maximum suppression (NMS) algorithm.

Compared to Fast R-CNN, Faster R-CNN improved 69.9% mAP in the PASCAL VOC 2007 test set, when using VGG-net[8], the processing speed is improved almost 10 times.

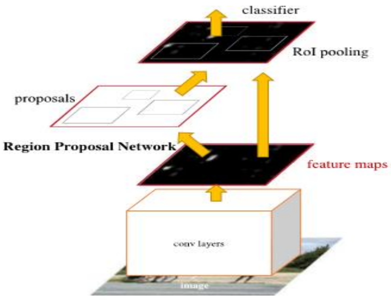

Figure 3. Faster R-CNN network

2.4 Mask R-CNN

Mask R-CNN[9] is a method applied to instance segmentation by extending Faster R-CNN. In Faster R-CNN's image classification and bounding box regression method, a binary mask that determines whether each pixel corresponds to an object was added to elaborately segment the object. By introducing FCN (Feature Pyramid Network) [10], performance was improved in accuracy and processing speed, In addition, by replacing the existing RoI pooling with RoIAlign, a more accurate pixel location can be extracted and accuracy is improved by reducing the decimal point error occurring in the location of the RoI pool area using bilinear interpolation.

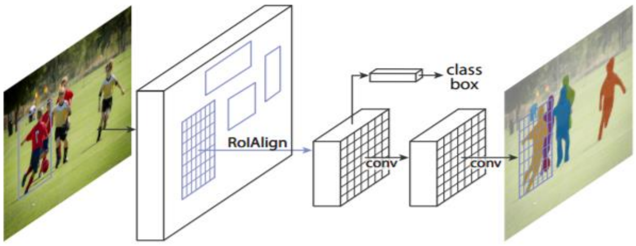

Figure 4. Mask R-CNN framework

2.5 YOLO

1) YOLOv1

YOLO[11] is an object detection method proposed by Redmon et al. that detects objects in real time, displays bounding boxes, and classifies objects. Unlike the two-stage object detection method, YOLO uses only one network to extract features, find bounding boxes, and classify objects. At the output stage, the location of the bounding box and image classification are performed at the same time.

YOLO first divides the input image into SxS grid cells and recognizes one object for each cell. A grid cell has a bounding box composed of four points (x, y, w, h) and a confidence score of the bounding box. The x and y of the bounding box mean the center point of the bounding box, and w and h are the relative values for the width and height of the entire image.

Because YOLO predict one bounding box per grid cell and one class for each bounding box, it is difficult to predict a large object or an object close to the boundary of several cells. In addition, detection is poor when there are multiple objects around one object, such as small objects are gathered like a flock. YOLO achieved 63.4% mAP at 45 fps compared to Fast R-CNN (70.0% mAP, 0.5fps) and Faster R-CNN (73.2% mAP, 7fps) for the PASCAL VOC dataset.

2) YOLOv2

YOLOv2[12] improves the speed and precision of YOLO. In order to maintain the processing speed, the network is simplified, and an easy-to-learn expression is used. The following are the improvements of YOLOv2.

Table 1. The improvements of YOLOv2

Batch Normalization.	• Addition of a batch normalization (BN) [13] layer in front of each convolutional layer
High Resolution Classifier	• The classification network is pre-trained, it works well for high-resolution images.
Convolutional with Anchor Boxes.	• The fully connected layer was removed, replaced with a convolutional layer, and the bounding box was predicted using anchor box. The class and object probabilities were separately predicted for each anchor box
Dimension clusters.	• Instead of randomly selecting the size and aspect ratio of the anchor box, it is determined using K-means clustering in the bounding box of the training data set, and d(box, centroid) = 1 - IOU(box,centroid) used.
Direct location prediction	• bounding box's center point is restricted not to deviate from grid cell
Fine-Grained Features	• The size of the final map was set to 13x13, and a method of importing the features of the immediately preceding 26x26 feature map was used to detect detailed features for small object detection.
Multi-Scale Training.	• In order to make it possible to process images of various resolutions well with one network, various resolutions are evenly trained. For this, every 10 batches {320,352, ⋯ , 608} randomly selects a new image size.
Darknet-19	• Darknet-19 with 19 convolutional layers and 5 max pooling layers using YOLOv2 is proposed

As mentioned above, YOLOv2 can achieve high computation and object detection performance through 8 major improvements and new backbone.

3) YOLOv3

YOLOv3[14] uses anchor boxes like YOLOv2 and predicts the objectness score (objectness score: whether there is an object in the bounding box) for each bounding box. In addition, since YOLOv3 can have multi-labels, it does not use softmax for class prediction, but uses independent logistic classifiers, and accordingly, binary cross-entropy is also used as loss in training. YOLOv3 uses a total of 9 anchor boxes, which are boxes 3 bounding boxes and 3 different scale feature maps and determined through k-means clustering. It also proposes a deeper and more powerful feature extractor called Darknet-53. Darknet-53 is applied the skip connection concept proposed by ResNet[15] to Darknet-19, and a much more layers were stacked. Compared to YOLOv2, performance is much improved. Due to the advantage of multi-scale prediction, YOLOv3 can detect much more small objects, but the performance of medium and large objects is relatively poor.

4) YOLOv4

YOLOv4[16] applies various methods by dividing into Bag of Freebies and Bag of Specials to improve performance. Bag of Freebies uses methods such as data augmentation, loss function, and regularization during learning and improves accuracy by increasing training cost. Bag of Specials is made of architecture techniques mainly and includes post processing, and refers to techniques that improve accuracy by increasing only inference cost. The Bag of Freebies is applied during training, and Bag of Specials only affects the forward pass in training, and is applied when inference is performed on the learned model. Also, it should be possible to use only a single GPU.

After applying various techniques to Bag of Freebies and Bag of Specials, YOLOv4 selects and applies the best techniques in evaluating the performance. Techniques applied with improved performance are as follows.

Table 2. VariousTechniques to Bag of Freebies and Bag of Specials

Bag of Freebies (BoF) for backbone	• augmentation : CutMix, Mosaic • regularization : DropBlock • etc : class label smoothing
Bag of Specials (Bos) for backbone	• activation : Mish • network : CSP, MiWRC
Bag of Freebies (BoF) for detector	• augmentation : Mosaic • regularization : DropBlock • loss : CIoU • layer : CmBN • lr scheduler : cosine annealing • etc : SAT, eliminate grid sensitivity, multiple anchors for a single gt
Bag of Specials (Bos) for detector	• activation : Mish • module : SPP, SAM, PAN • loss : DIoU-NMS

YOLO is vulnerable to small object detection. In YOLOv4, to detect various small objects well the input resolution has been increased. In addition, the number of layers was increased to physically increase the receptive field. In YOLOv4, CSPNet[17]-based CSPDarkNet53 was proposed, and although the number of parameters and FLOPS were large, the actual Inference Time (Throughput) showed the best results. The architecture of YOLOv4 was based on YOLOv3, and the backbone was changed to CSPDarkNet53, the SPP and Path Aggregation Network (PAN)[18] were applied to the neck, and the Bag of Freebies and Bag of Specials described in table 2 were applied.

YOLOv4 has improved accuracy (AP) by almost 10% compared to YOLOv3, and has the advantage of being able to train, test, and distribute models with a single GPU.

2.6 SSD

SSD[4] detects the bounding box of an object and recognizes the class using a single network like YOLO. It was used as an image feature extractor with some modifications based on VGG-16. YOLO has bounding box and class information only in the final feature map, whereas SSD information is distributed across several feature maps. Figure 5 shows the network structure of SSD and YOLO. As shown in the figure, there are six layers of different sizes, conv4_3, conv7, conv8_2, conv9_2, conv10_2, and conv11_2, corresponding to the last feature map of YOLO in SSD. In YOLO, each grid cell of the final feature map generates two bounding box candidates, whereas ,by introducing the concept of default boxes, the SSD is 4 or 6 bounding box candidates are generated with different aspect ratios in six multi-scale feature maps. As the result, object detection performance is improved without location estimation and size transformation.

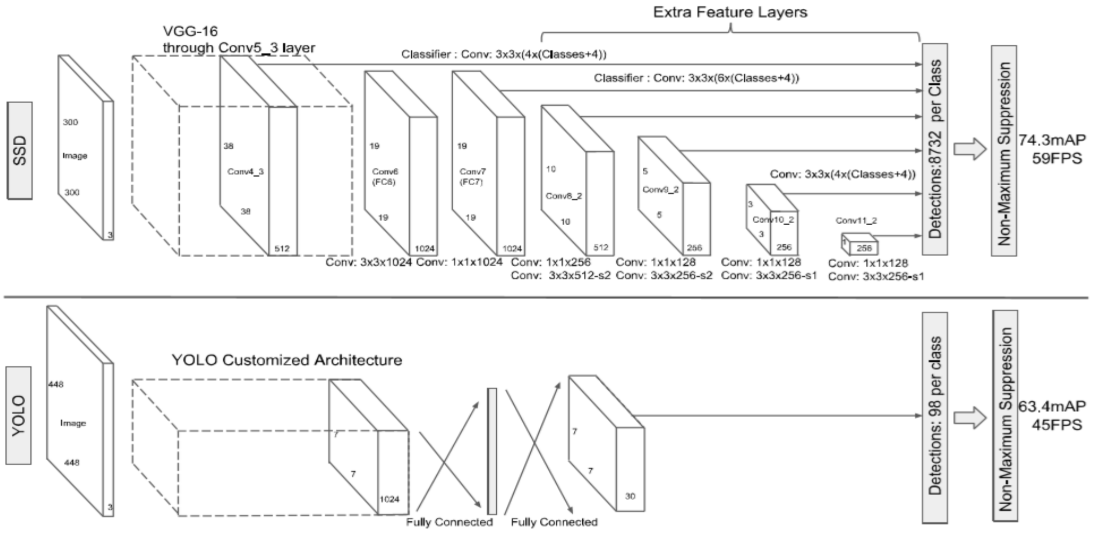

Figure 5. Structure comparison between SSD and YOLO

In general classification problems, it is common to attach an FC layer at the end of the CNN like YOLO, but SSD uses a convolutional layer as a classifier, and achieves speed improvement and model weight reduction by reducing the amount of computation and the number of model parameters. In addition, SSD learns only those with large reliability errors for objects in the direction of reducing errors, making them faster and more stable.

As a result of the experiment, SSD512 showed 81.6% mAP in PASCAL VOC 2007 test set and 80.0% mAP in PASCAL VOC 2012 test set, and showed better results than Faster R-CNN and YOLO.

2.7 RefineDet

RefineDet [19] is composed of two interconnected modules, the Anchor Refinement Module (ARM) and the Object Detection Module (ODM). ARM filters the negative anchors to reduce the search space of the classifier, and adjusts the position and size of the anchors to provide better initialization to the subsequent regressor. ODM uses the modified anchors as input to the previous ARM to regress accurate the object position and size and predicts the corresponding multiclass label. In this way, the functions of ARM have been transferred and improved to make the object more predictable in ODM, and these two modules are connected by a transport connection block. RefineDet shows the improved performance among the one-stage object detectors for PASCAL VOC 2007, PASCAL VOC 2012 and MS COCO data sets.

2.8 Relation Networks for Object Detection

Hu et al. [20] proposes an attention module for object detection called an object relation module (ORM) that considers interactions between other objects in an image including geometric information along with appearance characteristics. ORM is based on the attention module applied to NLT, etc. The proposed attention module extends the weight of the 1D's attention model to two components, the original weight and the new geometric weight. The latter models the spatial relationship between objects and extracts characteristics suitable for object recognition without changing the module transformation by considering only the relative geometric positions between objects. ORM is added to the detector head before two fully connected layers, applied to image classification and bounding box regression. In the COCO test-dev data set, accuracy increases by 0.2, 0.6, and 0.2, respectively, when the relationship module is added using Faster R-CNN, FPN, and DCN as backbone networks.

2.9 CenterNet

CenterNet[21] is similar to approaches to the existing one-stage Detectors (eg. SSD, YOLO) that use an anchor box, but there are significant differences.

Difference 1. CenterNet assigns anchor only by position, not box overlap.
Difference 2. CenterNet uses only one anchor.
Difference 3. CenterNet has a larger output resolution (output stride of 4).

The exiting one-stage detectors mostly used a large number of anchor boxes to predict the final bounding boxes to ensure that the assigned anchor boxes could sufficiently overlap with the ground truth box (box overlap).

CenterNet detects objects by using key point estimation. Only one keypoint(center point) per object is estimated, and each object is represented by the estimated one. Therefore, there is no need for grouping or post-processing processes (ex. NMS), and only one anchor exits.

Also CenterNet regresses various information such as object size, dimension, 3D extent, orientation, pose from the predicted center point, and can be easily extended to 3D Object Detection and Multi-person Human Pose Estimation as well as Object Detection.

2.10 EfficientDet

EfficientDet[22] applies the Weighted bidirectional feature network (BiFPN) and Compound Scaling and achieves the highest accuracy in the COCO dataset, and show that similar accuracy can be achieved with a very small amount of computation (FLOPS) compared to previous studies.

Weighted BiFPN is a method that gives weights to input features and learns weights through learning. The Compound Scaling technique is a complex scaling method that uniformly scales the resolution, depth, and width of all backbones, feature networks, and box/class prediction networks simultaneously by considering width, depth, and resolution, which are factors that determine the size and computation of the model. This idea was applied to EfficientDet's backbone, feature network, and box/class prediction network.

ImageNet-pretrained EfficientNet[23] was used as the backbone of EfficientDet, BiFPN was used as feature network, and applied to level 3-7 features. In addition, top-down and bottom-up bidirectional feature fusion were repeatedly used.

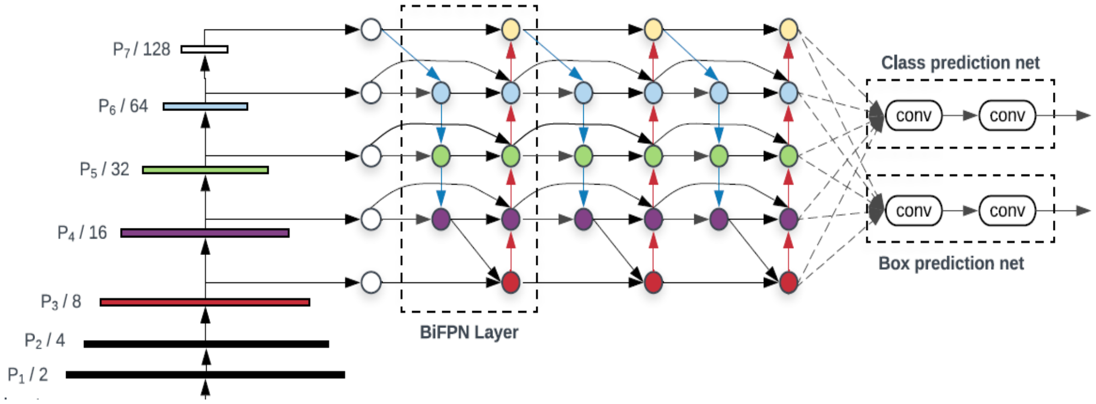

Figure 6. EfficientDet architecture

EfficientDet has a much smaller number of parameters and computations compared to object detection models with similar mAP performance. Therefore, the inference speed is about twice as fast as that of other models with the same performance.

3. Research direction of object detection

Deep learning-based object detection technology has rapidly developed based on the development of processors such as computers and GPUs that can rapidly process complex operations. However, the demand for a high-precision real-time system is increasingly required for more accurate application programs. Therefore, research should be studied in the direction of improving detection accuracy and speed.

3.1 Combined one-stage detector and two-stage detector in object detection networks

The two-stage object detection network is time consuming and has a dense process to obtain a bounding box. The one-stage detection network achieves fast processing speed applicable to real-time applications, but has low accuracy. Therefore, efficiency can be achieved the improved accuracy and high processing speed by creating a network to combine two-stage detector and one-stage detector.

3.2 Efficient post-processing method

In mostly object detection networks, only the best prediction result of one object is sent to the final classification layer to calculate the accuracy score. Post-processing methods, such as NMS, and subsequent improvements can improve the accuracy of detection and classification. Therefore, using a more efficient and accurate post-processing method should be studied.

3.3 Multi-domain object detection

In the existing object detection networks, the model is built to produce optimal performance in a single domain. However, if a general-purpose network to work in various image domains is implemented, the number of parameters can be reduced by learning by sharing features, and generalized features can be detected without prior knowledge of a new domain. Bilen et al. [24] add the batch normalized layer of a specific domain to a multi-domain shared network. Wang et al. [25] propose a universal object detector using a new domain attention module in various image areas (human faces, traffic signs and medical CT images) without prior knowledge of the area of interest.

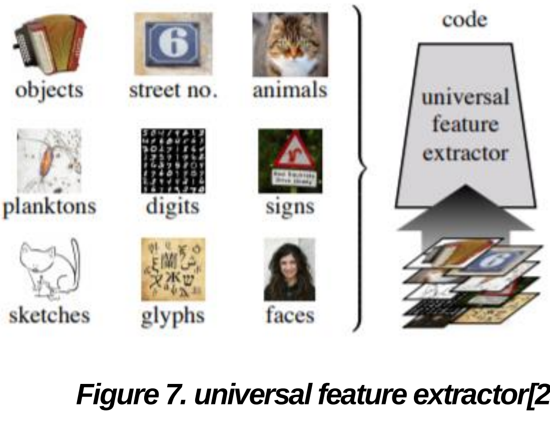

Figure 7. universal feature extractor[24]

3.4 Unsupervised object detection

Supervised learning takes a long time in the learning process and requires annotated data sets. In a large data set, annotating the bounding box for each object can be time consuming and laborious. Therefore, the development of automatic annotation processing technology and the detection of unsupervised learning objects are the future research directions.

3.5 Multi-source information support

With the advancement of social media and big data processing technology, multi-source information can be easily accessed. Rich social media information can provide both images such as photographs and texts that describe them as data, and processing fused multi-source information is a new research direction.

3.6 Object detection system for portable terminals

Cloud systems and portable terminals are playing a big role in helping people quickly solve any problems regardless of location. With the advent of lightweight networks, detectors in terminals have a wide range of applications. Therefore, it is necessary to develop a more efficient and stable object detection network.

3.7 GAN-based object detection system

GAN (Generative Adversarial Network) is a network that can generate fake images, and it can create and supply training data to deep learning networks that require large amounts of image data for training. Therefore, if the object detection network is trained by mixing the actual image and the data generated by the GAN, the object detection network can be more efficient and more efficient in generalization.

4. Conclusion

In this paper, deep learning-based technologies for detecting objects in images were addressed and a research direction was proposed. Key algorithms and differences between the recent Faster R-CNN, which improved performance from R-CNN, an initial object detection method based on deep learning, and object detection methods such as YOLO, SSD, RefineDet, CenterNet and EfficientDet are explained. In recent years, with the development of deep learning technology, object detection technology has also undergone a revolutionary development, but accuracy and speed need to be improved to be applied to various fields such as security, military, transportation, and life, and technologies for application to portable devices. These research directions are summarized and the future direction of object detection methods.

5. ACKNOWLEDGEMENT

This work was partly supported by NIPA grant funded by the Korea government(MSIT,MOJ) (NO. A0201-20-1011, Tracking system through abnormal behavior and face recognition with artificial intelligence analysis)

REFERENCES

[1] A.Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012.

[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, pp. 1137-1149, June 2017.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, June 2016.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in Computer Vision , pp. 21-37, 2016.

[5] Uijlings, J. RR et al.,“Selective search for object recognition,” *Int. journal of computer vision,* pp.154-171., 2013

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-

587, June 2014.

[7] R. Girshick, “Fast r-cnn,” in IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, Dec 2015.

[8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[9] K. He, G. Gkioxari, P. Dollr, and R. Girshick, “Mask r-cnn,” in IEEE International Conference on Computer Vision (ICCV), pp. 2980- 2988, Oct 2017.

[10] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936- 944, July 2017.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, June 2016.

[12] J. Redmon and A. Farhadi, “Yolo9000: Better, faster, stronger,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525, July 2017.

[13] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[14] J. Redmon et al, “YOLOv3: An Incremental Improvement,” arXiv 1804.02767

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, June 2016.

[16] A. Bochkovskiy, W. Chien-Yao, M. L. Hong-Yuan, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv:2004.10934v1, 2020.

[17] W. Chien-Yao Wang, M. L. Hong-Yuan, W. Yueh-Hua, C. Ping-Yang, H. Jun-Wei, and Y. I-Hau, “CSPNet:A new backbone that can enhance learning capability of cnn,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2020.

[18] S. Liu, L. Qi, Q. Haifang, S. Jianping, and J. Jiaya, “Path aggregation network for instance segmentation,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8759-8768, 2018.

[19] S.Zhang, L.Wen, X.Bian, Z.Lei, and S.Z.Li, “Single-shot refinement neural network for object detection,” in Proceedings of the IEEE Conference on Computer Vision, 2018

[20] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3588-3597, 2018.

[21] X. Zhou, D. Wang, P. Krahenb, “Objects as Points,” arXiv:1904.07850, 2019.

[22] M. Tan, R. Pang, Q. V. Le, “EfficientDet: Scalable and Efficient Object Detection,” arXiv preprint arXiv:1911.09070v4, 2020.

[23] M. Tan, Q.V. Le, “EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks, arXiv:1905.11946v5, 2019.

[24] H. Bilen, A. Vedaldi, “Universal representations : The missing link between faces, text, planktons, and cat breeds,” arXiv:1701.07275 , 2017.

[25] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, “Towards universal object detection by domain attention,” arXiv:1904.04402, 2019.