Journal of KIIT. Vol. 17, No. 5, pp. 91-97, May 31, 2019. pISSN 1598-8619, eISSN 2093-7571 91
http://dx.doi.org/10.14801/jkiit.2019.17.5.91

Real Time Object Detection Based on YOLO with Feature Filter Bank

Md Foysal Haque*[1], Hye-Youn Lim*[2], and Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (NO.2017R1D1A1B04030870).

Abstract

Real-time object detection is one of the most important and challenging tasks in current object detection and classification fields. Lots of research work have contributed to improving the ability of detection and classification accuracy in the last decades. Also, recent research on computer vision is prevailing to improve the accuracy of video surveillance, robotic vision, self-driving cars, and many applications. YOLO (You Only Look Once) is one of the fastest CNN (Convolutional Neural Network) and, it is one of state of the art techniques fo performing real-time object detection tasks. However, there are still some localization problems. In this paper, we propose a network with feature filter bank to improve the performance of YOLO network. Experiments have shown that the object detection performance of the proposed network is improved.

요 약

실시간 물체 탐지는 현재 물체 탐지 및 분류 분야에서 가장 중요하고 도전적인 과제 중 하나이다. 탐지 및 분류 정확성을 향상시키기 위해 지난 수십 년 동안 많은 연구가 이루어졌다. 비디오 감시, 로봇 비젼, 자율주행 차량 및 많은 어플리케이션의 정확성을 향상시키기 위해 컴퓨터 비젼에 대한 연구가 증가하고 있다. YOLO (You Only Look Once) 는 가장 빠른 CNN (Convolutional Neural Network) 중 하나이며 실시간 객체 탐지 분야에 최신 기술로 자리 잡았다. 그러나 여전히 약간의 지역화 문제점이 존재한다. 본 논문에서는 YOLO Network의 성능을 향상 시키기 위해 feature filter bank를 추가한 network를 제안한다. 실험을 통해 제안하는 network의 객체 검출 성능이 향상됨을 증명하였다.

Keywords
convolutional neural network, real-time object detection, YOLO, feature filter bank

* Dong-A University Electronic Engineering
- ORCID[1]: https://orcid.org/0000-0003-0634-9783
- ORCID[2]: https://orcid.org/0000-0002-5189-461X
** Dong-A University Electronic Engineering professor
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Jan. 08, 2019, Revised: Apr. 29, 2019, Accepted: May 02, 2019

· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University, 37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. Introduction

Deep learning exhibits its remarkable performance by the convolutional neural network (CNN) on machine learning and computer vision tasks [1]. The recent experiments of deep learning task show remarkable results on labeled large datasets like MS COCO [2], ImageNet [1], PASCAL VOC [3]. The deep convolutional neural network [4] achieved enormous performance on computer vision task but the structure of the network is very complex, and they need to be labeled large dataset for training. That involves the network high computational complexity for detecting and classifying the image. The main challenge of the deep convolutional neural network reduces the layers complexity with less computational complexity. To create a deep neural network for real-time detection system such a challenging work. Nowadays, real-time object detection becomes more popular because of the increasing demand for self-driving cars, robotic vision, and image classification for the survey of security surveillance system.

Deploying the deep neural network for developing the advanced driver-assistance system (ADAS) is one of the most challenging tasks because the system required most accurate object detector with classifying the system. To develop a perfect real-time system most challenging task is to extract fine-grained feature maps with less computational calculation. Moreover, labelling image and feature maps detection task is expensive than labeling for classification task [5]. YOLO (You Only Look Once) [5] suffers some localization error for classifying and detect objects that the main drawback to achieving the state-of-the-art detection and classification system. To compare with other convolutional neural networks like R-CNN [6], and Fast R-CNN [7], YOLO network face the significant number of localization errors. Thus, considering the structure of YOLO network with Fast R-CNN [7], the structure of Faster R-CNN is very complex. Moreover, conventional YOLO network performance is not good when it deals with the blurred image [8]. This problem mainly occurs for localization error and low recall compared capability. To improve the detection accuracy of YOLO, we focus to reduce the localization error and recall problem. In this paper, we aim to improve the structure to reduce complexity. Furthermore, to improve the localization and feature extraction of YOLO network we introduce feature filter banks. The feature filter bank [9] is a unique solution for extracting fine-grained feature extraction and reduce computational calculation with satisfactory response time. These improved characteristics help proposed network to achieve state-of-the-art accuracy for detecting objects for real-time object detection task.

Ⅱ. Related Algorithm

Computer vision is gradually increasing its popularity because of its making human life more manageable by different useful applications. A variety of steps can be followed to improve the detection accuracy of deep convolutional neural network such as increase the convolutional kernels, reduce the localization errors, and resemble each pixel for extracting feature. We follow these approaches to improve the network.

2.1 Feature Filter Bank

Extracting feature maps the most crucial and challenging task in a convolutional neural network. Moreover, after extracting the feature maps, the next challenge is to preserve the feature maps. Preserving feature maps is very expensive in object detection task. Considering this problem, we designed filter banks to preserve the feature maps. The feature filter bank [9] collect feature maps for different convolutional layers and preserves them for the further classification task. The feature filter bank constructed with a set of simple convolutional layers. These layers are all interconnected with each other. To construct the feature filter bank we used 1×1 and 3×3 convolutional filters.

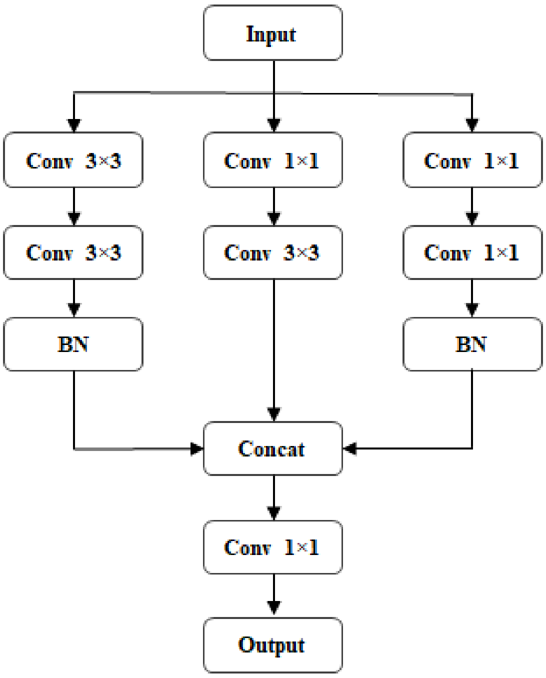


Fig. 1. Architecture of feature filter banks

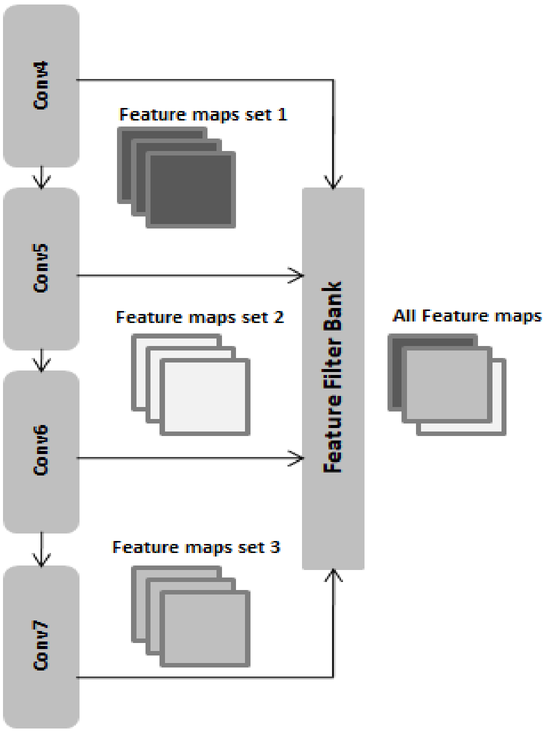


Fig. 2. Feature extracting process

To considering the network complexity, we introduced a simple network to preserve the feature maps. Furthermore, we also consider constructing a suitable network that can preserve feature maps in a convenient way because preserving feature maps is the most challenging and expensive task. The proposed architecture of feature banks shown in Fig. 1, and the feature extraction process shown in Fig. 2.

2.2 Batch Normalization

Batch normalization [10] conducts to eliminate the regularization error to improve the network classification accuracy. It also leads to label the feature maps as network requirement. In our network, batch normalizer used to remove dropout. It removes dropout from the network without creating overfitting and without affecting on network speed. Moreover, it also speeds up the network training process with a higher learning rate. Batch normalizer enhances learning accuracy by using mini-batch. It discrete learning data into small part of mini-batch and increase the learning process by eliminating regularization error.

2.3 YOLO(You Only Look Once)

YOLO [5] is mostly known as a real-time object detector. It predicts multiple bounding boxes simultaneously form the source image as like regional proposal network. YOLO network is relatively faster than other convolutional neural networks. Moreover, it examines the source image by sliding window to predict bounding boxes to detect the object. However, the performance of the network is not satisfactory when it deals with blurred objects. It also has localization error with a low recall. To overcome the problem, we modified the YOLO network structure. The improved YOLO network shows significant accuracy to detect objects. We also improve the feature extracting performance of the network to extract fine-grained feature maps. These fine-grained feature maps help the classifier to classify exact object location for the detection task.

2.4 Anchor Boxes

Predicting bounding boxes one of the most crucial tasks in an object detection system. The grid cells can predict only one object at a time. The grid cells are unable to detect multiple objects from the source image. In this experiment, for detecting multiple objects, we used anchor box technic. It examines source image by creating sliding windows.

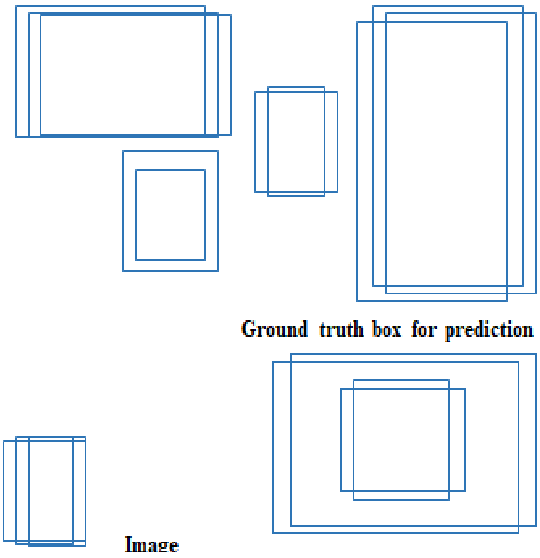


Fig. 3. Predicting anchor boxes

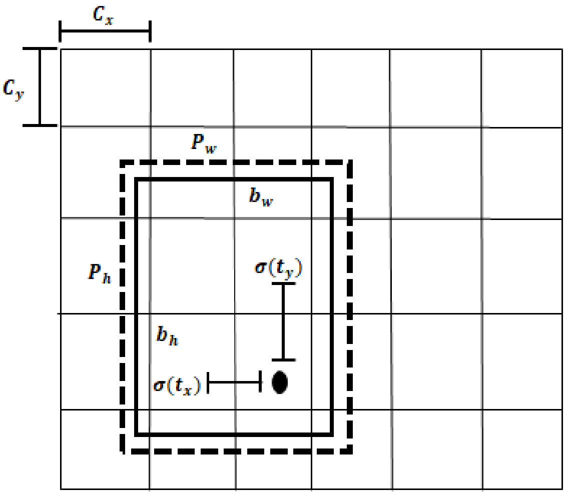


Fig. 4. Process of predicting bounding box

These sliding windows have a specific value after investigating the whole image the all values compared by IoU (Intersection over Union). It randomly produces five ground-truth boxes for one object to confidence about the location of the object. Each of the ground-truth boxes has the different value that carries exact object location. After that, all confidence value compared by IoU. The anchor box process examines each of the content from the source image and collects exact location to detect the object. The prediction of anchor box shown in Fig. 3. The all location value of the source image preserved in feature filter bank with feature maps. The following function uses for extracting object location.

$$d(box, centroid) = 1 - 10U(box, centroid) \quad (1)$$

III. Proposed Algorithm

This section, describe the proposed YOLO network with feature filter bank network. An improved network architecture developed that can able to extract fine-grained feature maps to detect real-time objects. However, to improve the performance of the convolutional neural network, different steps can be followed like increase the network's depth and width, improve the learning process and reduce cost complexity.

Furthermore, train the network with a large labeled dataset that it can extract more feature maps to train about the exact object location. To carry the experiment, we aim to improve the feature extraction part. Introduced a modified YOLO network with improved feature extractor and after extracting feature maps, preserve them in feature bank with object location. The modified YOLO network extract fine-grained feature maps that help to confidence about the object location. The architecture of YOLO network with feature filter banks shown in Fig. 5. The top part of the network (conv1, conv2, and conv3) mainly down-sample the image and extract primitive feature maps.

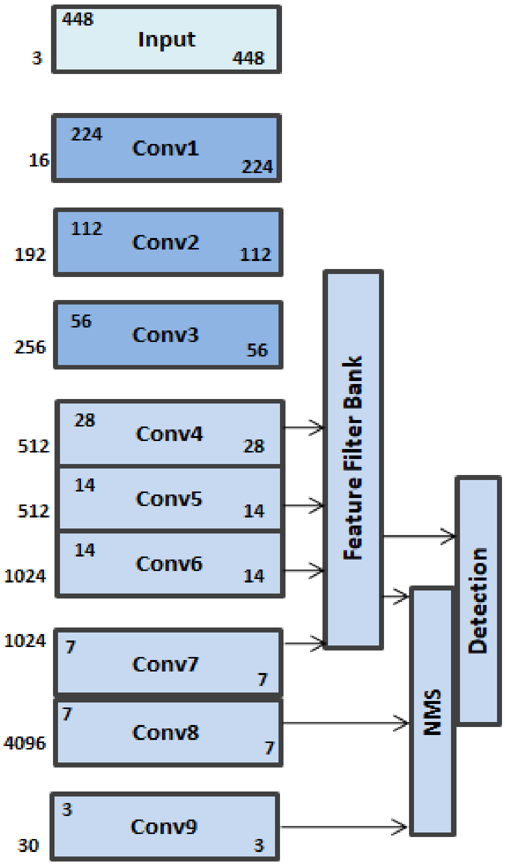


Fig. 5. Architecture of YOLO with feature filter bank

The output of conv3 passed to conv4 to examine and extract more small information of the object. Moreover, the main resources feature maps extract by the conv4, conv5, and conv6. The main resources for locating object exact location conducted by these three convolutional layers. To scale the feature maps maxpooling removed from conv5 for extract 14×14 feature maps. That mostly covers whole object location confidently.

Maxpooling removed from conv5 to scale the output and extract fine-grained feature maps. The proposed feature filter bank connected with the middle portion of the YOLO network. The feature filter bank constructed with a set of convolutional filters. The convolutional filters interconnected with each other. The feature filter bank first reduces the volume of feature maps next it expends all data again to reduce the cost calculation of the network. The feature filter bank conducts three different steps to preserve the feature maps in three different feature bank. The feature filter bank not only preserves the feature data it also reduces the computational cost and removes the localization error. To classify the objects and detect the objects confidently in our network we used non-maximum suppression (NMS). The feature filter bank and NMS work together to detecting objects. The last layers of the YOLO network also connected with NMS. It collects object location values from feature bank and fine-grained feature layers to examine the features and detect the object confidently in a short time. NMS examines feature maps and compares the object location by objects location higher value. It examines and takes the higher location value for final detection.

Ⅳ. Experiments

In this experiment, we developed a modified YOLO network with feature filter bank and compared the performance of the network with traditional YOLO network. The modified YOLO network classifies the data with regression and recognition process. To conduct this experiment we aim to reduce the complexity of the network. We designed the proposed network with simple experimental parameters. The YOLO network constructed for extracting fine-grained feature maps and the objects information preserved on proposed feature filter bank. To classify the objects information and locate the exact location of the object for real-time detection we used NMS. To train and test the network we used MS COCO [2] dataset. Table 1 shows the experimental parameters of the experiment.

Table 1. Experimental parameters

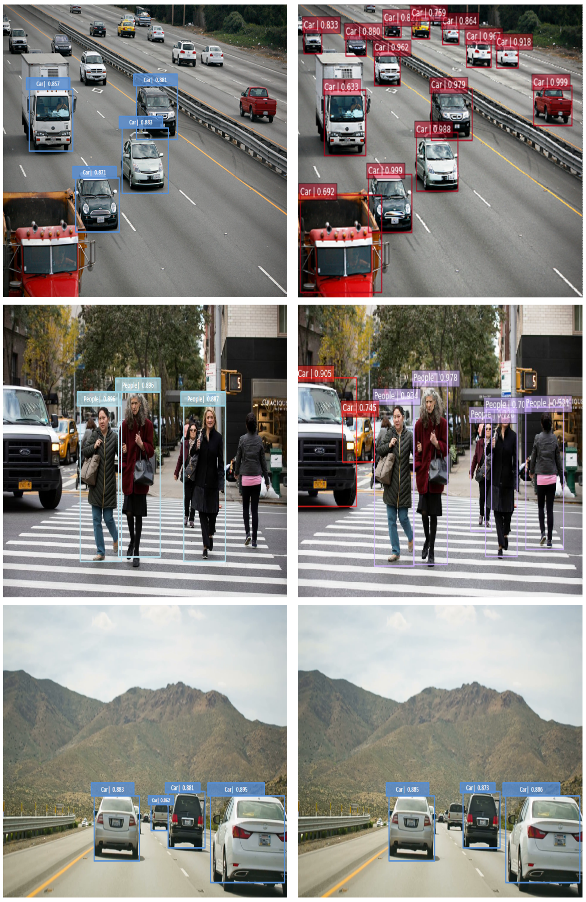
Experimental Parameters	Set Value
Learning rate	0.001
Batch size	64
Weight decay	0.0006
Input size	448×448

4.1 MS COCO Dataset

MS COCO [2] (Common Objects in Context) is a database that presented by microsoft corporation. The database mainly aims to image instance segmentation, object detection, person keypoints localization, and image captioning. The database contains 328,000 images and 91 different object types.

Ⅴ. Results

The detection performance of proposed YOLO with feature filter bank shown in Fig. 6. The improved network achieved state-of-the-art detection accuracy to detect the objects. The proposed architecture achieved significant detection accuracy on MS COCO test dataset.



(i) YOLO (ii) Proposed network

Fig. 6. Detection results

From Fig. 6 the improved detection results have shown. To compare the results with traditional YOLO network the proposed network achieved significant accuracy for detecting objects. It can detect multiple objects. The detection accuracy on MS COCO dataset shown in Table 2. The proposed method achieved higher performance on MS COCO. The feature filter bank helps the network to predict the exact bounding box for each object for detections.

Table 2. Detection results of MS COCO test

Method Average detection accuracy Class	YOLO with feature filter bank (Proposed Method)	Fast R-CNN [7]	YOLO [5]
	mAP		
	86.4	84.1	83.6
People	90.4	88.1	83.7
Bike	89.3	92.1	87.6
Car	89.2	93.9	88.4
Train	90.6	89.2	87.5
Bicycle	90.4	90.6	89.2
Cat	92.9	92.4	88.3
Table	80.7	80.1	78.9
Horse	91.1	91.8	88.9
Airplane	90.4	89.7	83.6

Ⅵ. Conclusions

The proposed network able to reduce the localization error and low recall. The feature filter bank achieved enormous performance to preserve the fine-grained feature maps with low computational complexity. The proposed network detection accuracy improved by 4.8% than the conventional YOLO network. Moreover, the network achieved an average value of 86.4% mAP to detecting the objects.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks", NIPS, pp. 1097-1105, 2012.

[2] T. Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", ECCV, pp. 740-755, 2014.

[3] M. Everingham, S. M. A. Eslami, L.J.V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective", International Journal of Computer Vision, Vol. 111, No. 1, pp. 98-136, 2015.

[4] K. Simonayan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv preprint arXiv:1409.1556, April, 2015.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger, CVPR, pp. 6517-6525, 2017.

[6] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", NIPS, pp. 379-387, 2016.

[7] R. Girshick, "Fast R-CNN", ICCV, pp. 1440-1448, Apr. 2015.

[8] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", CVPR, pp. 779-788, 2016.

[9] M. F. Haque and D. S. Kang, "Object Detection System using YOLO-based Feature Filter Banks", KIIT 4th Industrial Revolution and Artificial Intelligence Conference, pp. 188-190, Nov. 2018.

[10] S. Loffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", ICML, pp. 448-456, 2015.

Authors

Md Foysal Haque

2015 : B.Sc. in Electrical & Electronic Engineering, University of Information Technology & Sciences(UITS).

2018 ~ Present : M.Sc. in Electronic Engineering, Dong-A University.

Research Interests: Digital Image Processing, Computer Vision and Pattern Recognition

Hye-Youn Lim

2013 : Ph.D. in Electrical Engineering, Dong-A University.

2013 ~ 2018 : Assistant professor, Dong-A University.

2019 ~ present : part-time instructor, Dong-A University.

Research Interests: Computer Vision, Tracking.

Dae-Seong Kang

1994 : Ph.D. in Electrical Engineering, Texas A&M University.

1995 ~ Present : Professor at Dong-A University.

Research Interests: Image Processing and Compression.