A survey: Comparison between Convolutional Neural Network and YOLO in image identification

Richeng Cheng
RuiJingXiaoQu, SongMen, WenLing, ZheJiang, China.
chengr5@uni.coventry.ac.uk

Abstract: The main purpose of this paper is discussing Convolutional Neural Network (CNN) family and You Only Look Once (YOLO) family, comparing with the structure of the frame, speed of calculating and efficiency of identifying objects. There are two main summaries. The first summary is that training Faster Region-based Convolutional Neural Networks (Faster R-CNN) can achieve excellent detection effect, which not only reduces the time cost but also improves the quality of the proposal. Therefore, the method of alternating training Region Proposal Network (RPN) + Faster R-CNN in the Faster R-CNN is more advanced than the original SelectiveSearch + Faster R-CNN. Another summary is that YOLO is a convolutional neural network that supports end-to-end training and testing and can detect and recognize multiple targets in images with certain accuracy.

1. Introduction
This paper will discuss several image process methods which have been invented and applied since 1980s, such as CNN and YOLO. In the real world, image processes have been used in various aspect of daily life and demands of technology of image process have increased considerably. Recently, CNN has been used in image processes of clinical medical treatment frequently and deeply [1]. CNN families can be identified such as CNN, Region-based Convolutional Neural Networks (R-CNN), Fast Region-based Convolutional Neural Networks (Fast R-CNN), Faster R-CNN and Mask Region-based Convolutional Neural Networks (Mask R-CNN). Also, YOLO was invented lately as well for solving image identifying problems. YOLO has a large family as well and it can be identified such as You Only Look Once version 1 (YOLOv1), You Only Look Once version 2 (YOLOv2), and You Only Look Once version 3 (YOLOv3). In this paper, the main content includes several stages such as summarizing the development of CNN family and YOLO family, introducing the methodology of them, and discussing the conclusion of advantages and disadvantages of each other

2. Background
Artificial Intelligence (AI) in the 1980s tried to solve various computing problems by simulating the cognitive mechanism of the brain. Artificial Neural Network (ANN), one of the representatives of AI, is a network system that consists of large plenty of neural nodes. Namely, ANN simplifies and mimic the neural network system of human brain. As a result, ANN has some fundamental features which are similar to the brain [2]. However, CNN was invented as ANN cannot always satisfy requirements of in image process area [7]. It is well known that CNN is one of ANN structures based on the mechanism of the creature of natural vision (feature extraction) [3] and CNN can recognize the rule of vision directly from the original pixels, with little pre-processing because of the function of extracting the features of the image [8].
In 1986, Rumelhart, Hinton and Williams published the famous back-propagation algorithm for training neural networks in nature, which is still widely used today [17]. In the 1990s, modern CNN structures were confirmed, applying the backpropagation algorithm [4]. However, due to various reasons, most scholars gave up neural network for a long time. The neural network has a large number of parameters, and the problem of overfitting often occurs, that is, the accuracy rate is very high in the training set and the performance is poor in the test set. This is partly due to the small size of the training data set at that time. And with limited computing resources, even training a smaller network can take a long time [5]. The second reason for the poor performance was mainly due to the insufficient optimization of neural network algorithm network framework and the limited computing power of computers at that time. Until the 2000s, the computing abilities of computer has been increased dramatically so the processing time of recognizing image was decreased [6]. As a result, the importance of CNN in image recognition came back to the people’s attention again. In 2006, Geoffrey Hinton proposed deep learning. Since then, deep learning has achieved great success in many fields and attracted wide attention [18]. The most influential breakthrough of deep learning in the field of computer vision occurred in 2012, when Hinton’s research group won the image classification competition of ImageNet [2] by using deep learning.

3. CNN family

3.1. R-CNN

R-CNN was proposed in 2014 for improving the method of CNN in the image recognition [10]. R-CNN can choose a large number of bounding-boxes which could be final targets by optional searching and then divide the different independent area of the image by using feature extraction of CNN [9]. The fact is that, a frame of R-CNN consists of four steps, including inputting images, extracting region proposals, computing CNN features and classifying regions. Algorithm pattern of R-CNN can be identified to some several processes. Firstly, CNN can be trained by classifying images and target areas can be searched by the selective search. Furthermore, these target areas require re-size to a default size. Secondly, for each image area, a feature vector is generated through the forward propagation of CNN, and then the feature vector needs to be input into the binary SVM. Finally, the regression model is used to reduce positioning errors and correct the boundary box. To sum up, CNN, SVM and the regression model are included in R-CNN. As a result, R-CNN is very computational [10].

Based on the above, the main advantage of R-CNN is that it has lower error rate than conventional CNN, by dividing images into individual regions. Also, another advantage is that CNN network can extract features of image automatically. On the other hand, the disadvantage includes several points. Firstly, the training is staged and the steps are complicated. Secondly, each region proposal needs to extract features through the CNN network, generating a large number of feature files and occupying too much physical memory. Thirdly, each region proposal needs to extract features through the CNN network, resulting in a slow running speed. Finally, because SVM is used for classification, end-to-end training cannot be realized [6].

3.2. Fast R-CNN

In order to improve the image processing efficiency of R-CNN, Fast R-CNN was introduced in 2015. Three of the independent models were merged into a joint training framework, sharing computational results. Specifically, trainers no longer set independent feature vectors for each feature, but each image uses a CNN forward channel and shares the feature matrix. The same eigenmatrix is used to construct the classifier and the boundary regression matrix [11]. There is a fact that many of the steps in Fast R-CNN are the same as those in R-CNN. Also, Fast R-CNN consist of some several processes. Firstly, A pre-trained CNN was applied to the classification task as well as choosing the searching area. Then, the max-pooling layer which is the final layer of pre-trained CNN will be exchanged by Rol pooling layer. Secondly, Fast R-CNN algorithms replace the last full connection layer and the last SoftMax of k class with a full connection layer and SoftMax of k+1 class. Finally, Fast R-CNN output discrete probability
distribution for each RoI and predict bounding box regression model of offsets relative to the original RoI for each K class [11].

There are several advantages of Fast R-CNN compared with R-CNN. Firstly, Fast R-CNN solves the multi-stage training method of R-CNN and SPP-Net and trains the whole network by multi-task. Secondly, it uses SoftMax to replace SVM for multi-classification prediction to achieve end-to-end training. Finally, it realizes weight sharing of a convolutional network and adopts RoI pooling to maintain multi-scale input. Also, the disadvantage is the significant time-consuming, since the region proposal extracted by selective search can be only run in CPU [7].

3.3. Faster R-CNN

However, Fast R-CNN is still not fast enough. As a consequence, Faster R-CNN was introduced in 2016. Firstly, many of the steps in Faster R-CNN are the same as those in Fast R-CNN. Secondly, in order to propose distribution, the fine-tune of Faster R-CNN is an end-to-end RPN by using a pre-trained image classification. Thirdly, a Fast R-CNN object checking model is proposed through the proposed distribution of the current generation of RPN. Then, Fast R-CNN is used to initiate RPN training. While retaining the shared convolution layer, only the RPN layer is adjusted. At this stage, RPN and detection network convolution layer are shared. Finally, the unique layer of Fast R-CNN can be fine-tuned and Fast R-CNN can be repeated by 4-5 steps [12]. Also, Faster R-CNN can increase processing speed by integrating the regional proposal distribution into the CNN model and build a unified model that consists of regional proposal network (RPN) and Fast R-CNN with shared convolution feature layers [12].

The Faster R-CNN, as the most improved version, has several identified advantages. To begin with, Faster R-CNN optimizes the generation of the proposed area, so that it truly realizes an end to end training. Then, it achieves real-time on GPU and high accuracy. However, object proposal requires a lot of time, and the performance of the system depends on the performance of the previous system [8].

After this round of improvement, R-CNN series has evolved to complete basic structure. First of all, the feature of candidate region is extracted by the method of shared convolution. Secondly, in order to suit different candidate area sizes, Fast R-CNN have adopted the ROI pooling dimension with immobilized. Finally, it will make for multi-stage way of training task of training methods so that the model is simpler and more convenient [8].

3.4. Mask R-CNN

There is a new update in Faster R-CNN in 2017, Mask R-CNN. It applies faster R-CNN to pixel image segmentation. The key point is to classify and generate pixel-level masks. Mask R-CNN add a third branch such as classification, location, and mask based on Faster R-CNN. The mask branch generates a small full-link network for each RoIAlign and generates a pixel-to-pixel mask. Additionally, as pixel cutting requires more accurate boundary box, it has also modified RoI pooling [13].

From the perspective of function, Mask R-CNN needs classification, box regression and Mask regression. From the perspective of algorithm development, parallel operation is a trend, because it is simple and efficient. In Mask R-CNN three tasks were trained in parallel in the training phase. In the test phase, classification and box regression were carried out first, and then mask regression was carried out to obtain a more accurate mask, which could also reduce the amount of calculation.

4. YOLO family

4.1. Mask R-CNN

At the same time, in order to identify objects faster selective search for object recognition is used to choose different objects in processing images [17]. Also, first version of YOLO, YOLOv1, is an object detection system for real-time processing was invented and it can divide an image into different boxes to identify objects. The main idea of YOLOv1 is to build a CNN network to predict tensors [14]. Also, there is a rule of this method which is each divided grid cell predicts a fixed number of bounding boxes. YOLOv1 converts the object detection into regression problems. YOLOv1 processes an inference to
obtain the position of all objects in one image, their categories, and the corresponding confidence probability. Therefore, it can better identify the background with a comprehensive understanding of the background. Moreover, each grid cell predicts bounding boxes so that each of grid cell has a confidence score for the box which detects one object regardless of the number of boxes and predicts the probability of conditional classes [14].

On the other hand, because of the fast speed of YOLOv1, the advantages of YOLOv1 stand out. Firstly, predictions including object locations and classes consist of a single network and method of YOLOv1 can train end-to-end to improve accuracy. Secondly, YOLOv1 is more prevalent. It is superior to other methods when extending from natural images to other fields, such as art. Finally, the region proposal method restricts classifiers to specific regions. YOLOv1 accesses the entire image while predicting boundaries. With additional context, YOLOv1 shows fewer false positives in the background area [9].

4.2. YOLOv2
Afterward, YOLOv2 was created for improving YOLOv1 which consists of two phases (object categories and object locations). Accuracy of YOLOv2 improved significantly, which made YOLOv2 faster. Firstly, classifier networks that are like VGG16 were trained. Then replace the fully connected layer was replaced by the convolution layer and it was retrained end-to-end for object detection. Secondly, YOLOv2 use images to train classifiers and use images for object detection. This makes it easier to train the detector and improve mAP [15].

Compared to area-based detectors, YOLOv1 has a higher positioning error and a lower recall. YOLOv2 is the second version of YOLOv1, whose goal is to increase speed and accuracy significantly [10].

4.3. YOLOv3
Then, new version of YOLOV2, YOLOv3 was invented, which can use logistic regression to predict the objectiveness score of each bounding box and changes the way of calculating cost functions as well. YOLOv3 uses a separate logical classifier to replace the SoftMax function to calculate the probability that the input belongs to a particular label. In the calculation of classification loss, YOLOv3 does not use mean square error but uses binary cross entropy loss for each label. This can decrease computational complexity by avoiding SoftMax features [16].

In YOLOv3, it can be three times faster. But calculating average position (AP) of YOLOv3 still lags behind, because YOLOv3 has a higher positioning error. YOLOv3 also shows significant improvements in detecting small objects [11].

5. Conclusion
In this essay, there is a discussion between CNN family and YOLO family. Firstly, there is a common part that they need unified networks. But YOLO does not show the process of region proposal extraction. Although RPN and fast R-CNN share convolution layer, RPN network and fast RCNN network need to be trained repeatedly in the process of model training. However, YOLO is unified as a regression problem, and R-CNN solves the detection results in two parts: object categories (classification problems) and object locations (bounding boxes).

References
[1] Milletari,F., Navab, N., Ahmadi, S.A. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA. pp. 2-8.
[2] Yao, X. (1999) Evolving Artificial Neural Networks. Proceedings of the IEEE., 9: 1423-1447.
[3] Fukushima, K. (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernetics., 36: 193-202.
[4] LeCun, Y., Bengio, Y. (1995) Convolutional networks for images, speech, and time-series.
http://yann.lecun.com/exdb/publis/pdf/lecun-bengio-95a.pdf.

[5] Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S. (2014) CNN Features off-the-shelf: an Astounding Baseline for Recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, OH, USA. pp.806-813.

[6] Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A. (2016) XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. https://pjreddie.com/media/files.

[7] Li, H.X., Lin, Z., Shen, X.H., Brandt, J., Hua, G. (2015) A Convolutional Neural Network Cascade for Face Detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. pp.1-10.

[8] Zhang, X.Y., Zhou, X.Y., Lin, M.X., Sun, J. (2018) ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. pp.1-9.

[9] Liang, M., Hu, X.L. (2015) Recurrent Convolutional Neural Network for Object Recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. pp.3367-3375.

[10] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. pp.1-8.

[11] Girshick, R. (2015) Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). In: Santiago, Chile. pp. 1440-1448.

[12] Shaqing Ren, Kaiming He, Ross Girshick, Jian Sun, (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence., 6: 1137 – 1149.

[13] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017) Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. pp. 1-10.

[14] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. pp. 1-10.

[15] Redmon, J., Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. pp. 1-23.

[16] Uijlings, J.R.R., Sande K.E.A., Smeulders A.W.M., Selective Search for Object Recognition. International Journal of Computer Vision., 104:154-171.

[17] Rodan,A., Faris, H., Alqatawna, J. (1986) Optimizing Feedforward Neural Networks Using Biogeography Based Optimization for E-Mail Spam Identification. International Journal of Communications, Network and System Sciences., 9: 19-28.

[18] Hinton, G.E., Osindero, S., The,T.W. (2006) A Fast Learning Algorithm for Deep Belief Nets. Neural Computation., 7: 1527 – 1554.