氏　　名	TUERSUNJIANG YIMAMU
授 与 学 位	博士（工学）
学 位 記 番 号	博甲第 208 号
学位授与年月日	令和 5 年 3 月 17 日
学位授与の要件	学位規則第 4 条第 1 項
学位論文題目	Feature Extraction for Single Shot Multibox Object Detector （単画像多矩形物体検出のための特徴抽出）
論文審査委員	主査　教　授　榮坂　俊雄 准教授　吉澤　真吾 教　授　升井　洋志 教　授　黒河　賢二 教　授　平山　浩一

学位論文内容の要旨

The Object detection is one of the momentous research parts in computer vision due to its wide application. The essence of the object detection is not only one that classify the particular object categories but also one that we need to find out location information using bounding boxes. In recent years, object detector made a lot of progress with the help of deep convolutional neural networks (CNNs). Recent CNN based object detection algorithm can be categorized into two types of frameworks including the one stage detector, such as Faster R-CNN, R-FCN, SPP-Net, and the two-stage detector, e.g., YOLO, SSD and RetinaNet etc.

However, scale variation in object detection is still a challenging problem for both methods. Researchers have proposed many approaches to handle the multi-scale objects detection problem. CNN based object detector such as SSD, Faster RCNN extract features from the input images using various backbone networks such as VGG, ResNet. Backbone networks of object detector are usually pre trained by the ImageNet classification dataset. Many novel image classification networks are designed to get higher accuracy for ImageNet. VGG consists of convolution layer with different kernel size and max pooling layer to build a deeper network without residual connection. GoogLeNet, a deeper and wider neural network, consists of inception modules which enhance feature extraction at different scales using convolution layers with different kernel size. ResNet architecture consists of “bottleneck” design which is using skip connections to jump over some layers with residual sum operation in each stage. DenseNet densely connects several layers, through dense blocks, where we connect all layers with each other. To mitigate scale variation problem, feature fusion strategies have been proposed such as Image pyramids, FPN, DSSD and FSSD. Image pyramids generate different scale feature maps using CNN with different scale images, which is computationally expensive. Some researcher proposes single scale feature map to produce anchors with different scales. This kind of method has a limitation to detect various size of objects due to the fixed receptive fields. Different scale features structure has been applied by FPN, DSSD, SharpMask methods which fuse different scale features using element wise summation, while DSSD combines lower-level features to higher level features through an encoder–decoder structure. Feature pyramid is adopted by SSD to detect objects with different scales. In SSD, Conv 4-3 and Conv 8 layers are used to extract features for small objects and large objects, respectively.

In object detection network, the main features of small objects are generated by shallower layers without semantic rich information, which causes lower accuracy on small objects for a deeper network. Specifically, the extracted features in deep layer not only contain semantic rich information with lower resolutions, but also have larger receptive fields, which are useful for object classification. The extracted features in shallow layer mainly contain spatial-rich information with higher resolution and smaller receptive fields, which are beneficial for object localization or vector regression. It is vital that high resolution representations are made available to small object detection. There are poor high-resolution features available in the deeper layer, which made model hard to detect small objects.

We can expand kernel size with original weights using a dilated convolution which samples sparsely at different locations and increases receptive field with same computational cost. Dilated convolution avoids the negative effect of down sampling operation, and we can still use the high-resolution features even in the deeper layers. However, large object detection without enough receptive fields is difficult. DetNet uses dilated convolution to design a specific detection backbone.

In this research, to mitigate problems pointed out above, we propose Multi-path Feature Fusion Single Shot Multi-Box Detector (MF-SSD) by adding an efficient feature fusion module which embed two newly designed modules to fuse features with different scales. We employ dilated convolution with different dilation ratios to design Two-branch Residual dilated Convolution Module (TRDCM) and Two-branch Residual dilated Add Convolution Module (TRDACM), which enlarges receptive field without extra computational cost. We have conducted numerous experiments on MS COCO， PASCAL VOC2007, and PASCAL VOC2012 datasets to explicate the efficacy of our proposed detector. In the proposed feature fusion module, we not only extract features with sufficient boundary information, which is beneficial for vector regression, but also extract features with contextual information which is beneficial for object classification. Using the proposed feature fusion module, we improve a lot of performance compared with original SSD, especially for small objects. In summary, we highlighted our main contributions as follows:

1. A new object detection framework, MF-SSD, is proposed to handle multiscale problem, especially for small objects.

2. We newly designed an efficient feature fusion module to fuse different features with different scales to improve better performance for object detection compared with the conventional SSD. The proposed feature fusion module contains two newly designed modules including the TRDCM and TRDACM. Through these two modules, we improved contextual information without losing original resolution of the feature map.

3. We have explicated the efficacy of our proposed MF-SSD through numerous experiments.

審査結果の要旨

要　　旨

近年、自動運転やセキュリティ、医療検査など、様々な分野で、AI、特に深層学習を活用した画像認識の要望が高まっている。AI画像認識には目的に応じて様々なタイプがあり、分類classification（犬か猫か）より、検出detection（チワワがどこにいるか）に高い精度が必要となる。本研究は、重なった物体の正確な位置判定、極めて小さいサイズの物体の検出という特徴を持つAI画像物体検出手法を新たに提案し、従来の方法と比較検討した。

従来の深層学習による物体検出は、大きく分けて、大まかな領域を抽出した後に正確な位置判定と内容分類を行う2段階の検出法と、後者を同時に行う１段階の検出法があり、１段階検出法は計算時間が少ないものの精度が低いという問題があった。本研究では１段階検出法の代表的方法であるSDD（Single Shot Multibox Detector）をベースにし、16層のCNN（Convolutional Neural Network）と内容分類のための畳込み層の間に新たに独自に開発した特徴抽出モジュールを挿入することで、精度を向上させた。さらに標準的なテスト画像を用いて、提案法と従来法：複数の改良されたSSD法、同様に複数のCNN法、YOLOなどを比較し、提案法の有効性を検証した。

これを要するに、申請者は、AI画像認識分野において、新たに精度が高い物体検出が可能な方法を考案したものであり、画像認識および深層学習分野に貢献するところ大である。

よって、申請者は、北見工業大学博士（工学）の学位を授与される資格があるものと認める。