Research Article

A Survey on Human Detection using Reinforcement Learning

Susmita Goswami[1], Spandana Bajpai[2], Ushasukhanya[3]

[1] Student, Department of Computer Science and Engineering, SRMIST, Chennai
[2] Student, Department of Computer Science and Engineering, Kattankulathur
[3]Assistant Professor, Department of Computer Science and Engineering, Kattankulathur

Article History:Received:11 november 2020; Accepted: 27 December 2020; Published online: 05 April 2021

Human Detection - technology related to computer vision and image processing work by finding people in digital photos and videos and surveillance videos that are part of the observation. Single Shot Detector (SSD) is a deep learning method and is one of the fastest algorithms that use a single convolutional neural network to detect objects involving humans, cats, dogs, etc., and extract feature maps to classify the candidate object in the respective images. The advantage that SSD has is that it is quick to detect and has high accuracy in a given situation compared to regional suggested networks with smaller resolution images and smaller objects. However, it is still somewhat lagging in detecting large objects in larger images as compared to other algorithms that have been used to achieve better accuracy. It is a simple, end-to-end solution for a single network, and detection and extraction are done with one step forward single pass. The proposed system is to use the Optimized-SSD algorithm to detect human accuracy in the proposed database with good accuracy which will be the task of learning to increase SSD capacity as a detection system.

Keywords: Human detection, Object detection, Deep learning, Neural Networks, Single shot detector, CNN.

I.Introduction

Over the years, detecting people has attracted a lot of attention. Deep learning has become one of the most sought-after techniques in various fields of machine learning as well as to object detection [1]. Many advanced methods based on such as R-CNN, RCNN faster have been in the object detection area and have achieved great accuracy over numerous training and project works which have furthermore improved its accuracy. These approaches have reached a higher definition, but their network structure is more complex. The basic concept of SSD is largely based on a single pass neural network. It cuts off the binding space of the binding boxes in the default box set above the feature variations and scales with the feature map location. It then generates the availability scores for each item category in each default box and generates adjustments to better match the structure of the object [2]. The SSD model consists of mainly two structures: Base structure and Auxiliary structure.

The Base Network is the first part of the model based on the standard architecture used for high-resolution image classification [3]. The Auxiliary Network has features that focus mainly on objects with different levels or aspect ratios. SSD consists of two parts: remove feature maps, use convolution filters to locate objects. Our model-building project uses a single-box Single-shot detector using a Chokepoint database that contains video and image views and overcomes system issues with better accuracy and efficiency. To speed up the integration of the high-level model, the Adaptive and Momentum bind (AdaMod) optimizer will be used to convert the flexible training level of the high-level model into the training process. Use Chokepoint as our database will help train the best performance model for the crowd as the chokepoint database captures the status of the crowd.

II.Related Work

The author in [1] used "Visual Saliency as a regional proposal algorithm that proved useful in the research work. Human Detection has improved by introducing video window frames in the HOG + SVM separator". Using the Deep Multi-Layer Network to predict intelligence to grow interest regions makes modeling possible. A detailed study of the frameworks of deep learning methods is provided by the authors of a paper [2] dealing with a variety of issues, such as closure, derangement and convenience, and stages of conversion on R-CNN. The review begins with a standard detection pipeline that provides the basis for other related activities. Then summarize the other three common functions, namely the discovery of the object, the discovery of the face and the discovery of pedestrians.

The authors of [6] use VGG16 - With a faster extraction network, the model will work better. • Real-time acquisition is extremely slow and all it offers is good precision of local and phased performance. SSD - Accuracy detection is better with automatic boxes, which have a negative impact on speed •

*Corresponding author: Susmita Goswami
Student, Department of Computer Science and Engineering, SRMIST, Chennai

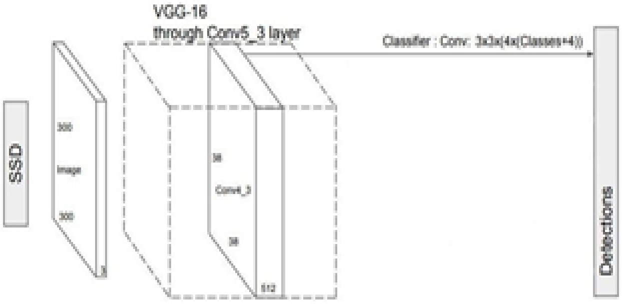

Fig 1 Architecture of VGG16

The research paper provides an important guide for those interested in continuing research in human detection. “In the action recognition study, the appropriate data for capturing the action should be selected. Also, a logical algorithm should be used to detect human action. With practical learning problems, deep learning methods have great effectiveness”[7]. “In addition to the separation of first-person and individual-action actions, communication recognition, and action acquisition have become prominent new research topics. The ongoing work of modern human detection on static videos is well documented” [5]. Each method has its advantages and disadvantages that are discussed.

“A brief study of the discovery of a different object and algorithms for the classification of an object found in textbooks and comparative studies of different methods, used for the detection of an object, and the classification of an object”[4]. “Object classification can be performed by using different methods like supervised learning, non-supervised learning and semi-supervised learning.It can be summarized that background subtraction is the simplest method that provides complete information about an object in comparison with optical flow and frame difference for detecting objects”[2].

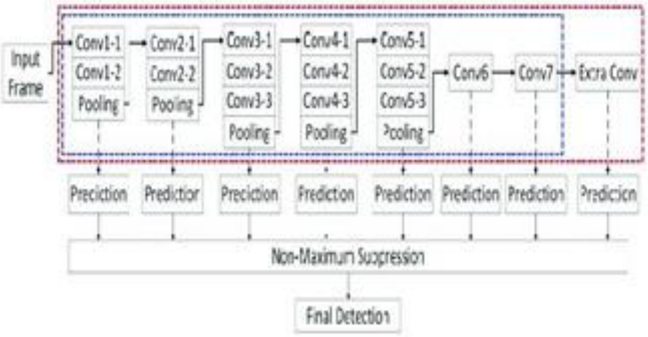

Fig 2 Framework of Object detection

“Neural Networks regional-convolutional systems and shot detector systems are two types of detection systems. Over the past few years, new facilities have been built to address the challenges of R-CNN and its successors which enable real-time discovery”[6]. “The most popular is YOLO (You Just Look Once) and SSD Multi-box (Single Shot Detector) “[6]. “SSD speeds up the process by eliminating the need for a region proposal network. To get the precision drop, the SSD uses several enhancements including limited features and default boxes. This upgrade allows the SSD to be similar to R-CNN's fast accuracy which is used with low-resolution images, which further speeds up the process”[6].” It achieves real-time processing speed and in some cases even beats the accuracy of Faster R-CNN “[4]. “It has no delegated region proposal network and predicts the boundary boxes and the classes directly from feature maps in one single pass by removing the proposed regional suggestion and using low-resolution images” [4]. The model has improved so as to it can run real-time and still beat the state-of-the-art accuracy of Faster R-CNN.

II.I Single Shot Detector

SSD is a deep learning method and is one of the fastest algorithms that use a single convolutional neural network to find an object and also separate the independent elements that will be placed in the image. An SSD (Single-Shot Detector) uses a complete solution when the network can detect all objects within the image in more than one way (hence - 'one shot' or 'look once') via convnet. The SSD has two components: the spine model and the SSD head. Object detection using SSD is done in two stages.

1. Feature mapping
2. The use of convolution filters to find objects in images.

The process of tying the SSD binding box was inspired by Szegedy's work on MultiBox, a way to make suggestions for connecting a fast connecting box. The 1x1 combination helps to reduce the size because the size will decrease but the width and height will remain the same. SSD has a basic VGG-16 network followed by layers of multi-box Conv. The VGG-16 SDD architecture is a standard CNN architecture for high-resolution image editing but without the terminal layers. VGG-16 is used for feature derivation.

“The construction of the SSD builds on the revered design of the VGG-16 but eliminates the completely connected layers”[3]. “The reason why VGG-16 is used as a basic network is because of its strong performance in high-quality image classification and its problem-solving where learning transfer helps improve outcomes. Instead of the first fully integrated VGG layers, a set of auxiliary layers has been installed (from conv6 onwards), enabling features to increase and continuously reduce the input size in subsequent layers”[3].

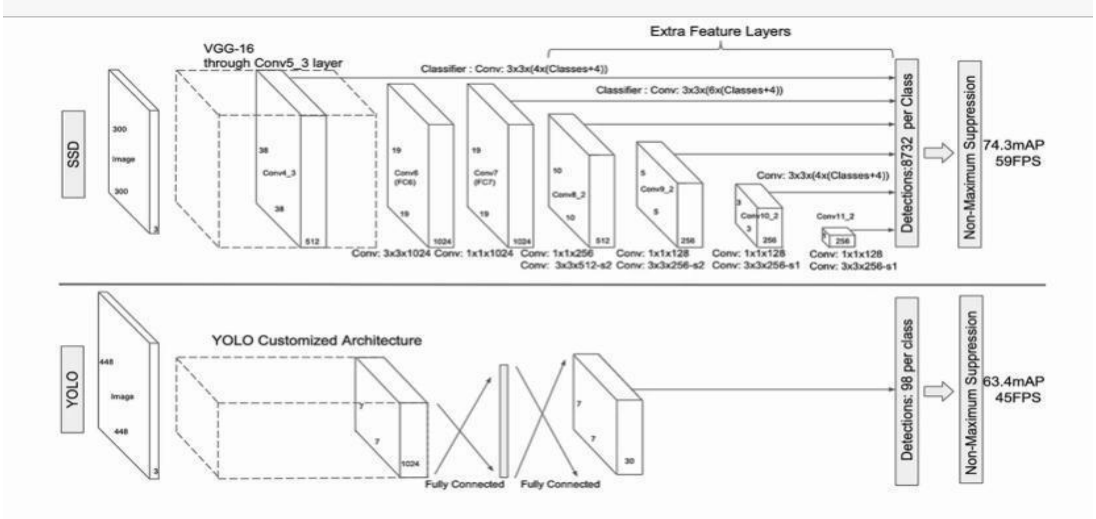

Fig 3 Diagrammatic comparison of SSD and YOLO

II.II Multibox

Multi-Box, a fast-paced agnostic box system that incorporates suggestions.Two components that are integrated into the multibox loss function includes: Loss of Confidence & Loss of Location. The function for it looks like:

loss = loss_of_confidence + alpha * loss_of_location

II.II.I MultiBox Priors And IoU

Multi-boxes, researchers created priors, which are pre-installed computers, measuring size boxes that are closely related to the distribution of true ground boxes. Many of the boxes started out as important as projections and attempts to retreat close to the true ground boxes. Finally, Multi-boxes only maintain high K predictions that reduce spatial loss (LOC) and confidence (CONF).

II.II.II Feature Maps

Feature maps which are the results of CNN blocks on each level are nothing but patterns/textures/patches of the image on a smaller level which ultimately are fed into the next layers wherein these maps become prominent in size i.e shapes/objects. These feature maps are key in the precision of detection.

II.II.III Non-Maximum Suppression (NMS)

The pruning of bounding boxes is done using a process known as non-maximum suppression wherein boxes with low confidence loss and IoU boxes are rejected to keep the peak predictions for the same object. This is to ensure only that high and top predictions are kept and maintained, while the irregular or boxes with fewer scores are repudiated.

III Inferences

The review concludes that the use of SSD is beneficial for real-time access. A good plan for detection is beneficial. SSD has the potential to be trained to the end of the best accuracy [4]. SSD makes a lot of speculation and has better input by location, rating, and feature rating. With the above enhancements, SSD is able to reduce the image stabilization to 300 × 300 with comparable accuracy performance [3]. The detection of a person is a difficult task from the point of view of the machine as it is influenced by a number of visual cues due to the changes in appearance, clothing, lighting, and background, but previous knowledge of these limitations can improve the performance of the detection.

IV Proposed System

Currently SSD's accuracy is performing well only at certain scenarios. We can increase its accuracy with better design boundary boxes and smaller default boxes and enhance association between prediction object result and positioning precision by IoU loss branches. Using Chokepoint as our dataset will help train the model for better performance for a crowd as chokepoint dataset includes sequences in crowd scenario.The pruning technique to be used is to be decided later. The Adaptive and Momental Bound (AdaMod) optimizer can be used to change the corresponding level of the advanced model that is too high in training to speed up the integration speed of the advanced model. There's also the issue of spatial redundancy in the models, which we're trying to eliminate for better compression.

V Conclusion and Future Work

Our future work would be focused on building this model completely from scratch and training this model to have the best accuracy possible with various datasets. Currently SSD's accuracy is performing well only at certain scenarios. We can boost its accuracy by using better design boundary boxes and smaller default boxes, as well as IoU loss branches to improve the connection between prediction object score and positioning accuracy. The (Adam) optimizer will be used to change the model's adaptive learning rate.

VI REFERENCES

”**Human Detection and Tracking for Video Surveillance: A Cognitive Science Approach.**” by Vandit Gajjar,Ayesha Gurnanim,Yash Khandhediya (L.D College of Engineering,Ahmedabad).

”**Object Detection with Deep Learning : A review**” by Zhong-Qiu Zhao, Member, IEEE, Peng Zheng, Shou-tao Xu, and Xindong Wu, Fellow, IEEE

”**Object detection: Comparison of VGG16 and SSD**” by Sabhatina Selvam

“**A Comprehensive Survey of Vision-Based Human Action Recognition Methods**” by Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du and Duan-Sheng Chen.

“**Survey on Human Detection Techniques in Real Time Video**” by Paresh M Tank, Hitesh A Patel, (Lecturer, Department of Computer Engineering, B & B Institute of Technology, V.V.Nagar, Gujarat, India).

“**A Survey on Object Detection and Classification Methods**” by N.Kiaee, E. Hashemizadeh, N.Zarrinpanjeh.

“**SSD Object Detection Model Based on Multi-Frequency Feature Theory**” by Jinling Li; Qingshan Hou; Jinsheng Xing; Jianguo Ju.

“**Deep Learning for Generic Object Detection: A Survey”** by Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu & Matti Pietikäinen .\

Object detection method of multi-view SSD based on deep learning” by Tang Cong,Ling Yongshun, Zheng Kedong,Yang Xing,Zheng Chao,Yang Hua, Jin Wei1.

A lightweight small object detection algorithm based on improved SSD” by Wu Tianshu,Zhang Zhijia,Liu Yunpeng,Pei Wenhui1,Chen Hongye.