MODELLED NEURAL NETWORKS FOR MULTIPLE OBJECT TRACKING

Ivana Walter*

Singidunum University,
Belgrade, Serbia

Abstract:
Recent development in artificial intelligence brought deep learning and neural networks that are applied in various areas, e.g. robotics, surveillance, autonomous driving, automation and medicine. Multiple Object Tracking very commonly utilises those architectures and there are many different approaches for this task. Those solutions are based on different kinds of neural network structures and this paper provides comparison of the corresponding algorithms that could improve further research. The paper investigates the performance of the Faster R-CNN, the VITAL and the RetinaNet methods with practical results and examines their different architectures used for object detection. The requirements for models are the detection of objects' position and their classification. For tracking the instances, we use algorithms that are based on object detection systems. For registering the location of items Neural Networks use the IOU (Intersection of Union) in order to determine which bounding boxes should be examined and according to the IOU we distinguish positive and negative proposed bounding boxes. The negative predictions impact the performance and negatively contribute to the wanted signal. The results of the Faster R-CNN method present those challenges. The object classification could become difficult in the event of occlusion. The RetinaNet method provides distinguished detection and classification results that could be applied for the Faster R-CNN and the VITAL computations. There are many evolving implementations for object tracking. The VITAL detector that uses the GAN for the motion prediction was evaluated on the custom set of image sequences, that are used for deep neural network adaptive parameter regulation..

Keywords:
Neural Networks, Object Detection, CNN.

INTRODUCTION

The aim of object tracking is to keep detection of items in video sequences, and to determine their number, location, activity and their characteristics. In the future those features are going to be implemented in desktop computers, smartphones, tablets and digital billboards in order to enable interaction with humans [1].

Correspondence:
Ivana Walter

e-mail:
ivana.walter.20@singimail.rs

DOI: 10.15308/Sinteza-2021-165-169

Deep neural networks are applied for object detection challenges in object and background classification and various available datasets could be used for the neural network training. Object tracking tasks typically rely on object detection systems that could be divided into two categories: one-stage detectors and two-stage detectors.

Multi-target tracking algorithms that utilise the CNN features are two-stage detectors and are based on Region Proposal. In this approach the CNN (Convolutional Neural Network) calculates the dependences of positional features.

The idea is based on a pyramidal structure in order to first identify the concept of the lowest level before it is being transferred to the next layer for processing on higher levels. This model is inspired by the human vision that determines relevant elements of the whole scene.

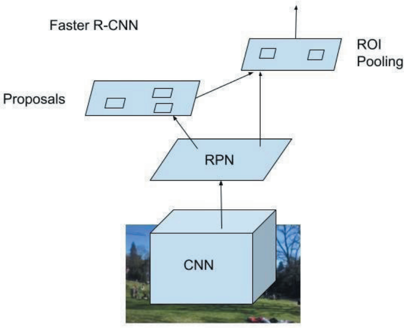

Figure 1 - Faster R-CNN [2]

The Faster R-CNN approach introduces the Region Proposal Network (RPN), that provides better performance than the Fast R-CNN and the R-CNN. Regions with Convolutional Neural Network (R-CNN) method is based on selective searching that proposes regions being classified one at a time with an output label and the bounding box. The Fast R-CNN provides better performance in generating Regions of Interest than the R-CNN.

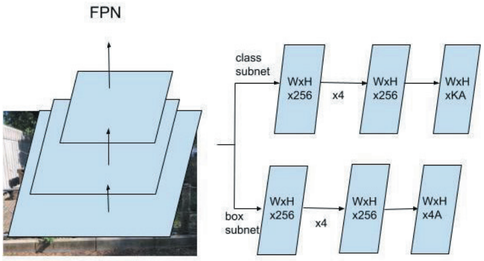

Figure 2 - RetinaNet [3]

Common one stage methods are YOLO and RetinaNet. You Only Look Once (YOLO) divides input images into cells and propagates them once through a neural network to calculate bounding boxes and class probabilities. The RetinaNet architecture is based on the Feature Pyramidal Network (FPN) that extracts features and takes original images on the input and calculates proportionally sized feature maps multiple times.

Generative Adversarial Networks (GANs) are a recent innovation in machine learning that use two different Neural Networks called Generator and Discriminator. The Generator provides data sets that are initially fake and are used for the Discriminator's training that can in this way learn the real data distribution.

The adaptive control is applied by collecting the input and output data in order to adapt the time wise dynamic implementation [4]. As the Discriminator might fail in classifying, the loss is computed based on the Discriminator's output in order to adapt the weights of the Generator's Neural Network. In this evaluation the given parameters of the RetinaNet algorithm were employed as the input values for tracking of motion using the VITAL method.

2. METHODOLOGY

For the purpose of this research frame sequencing of video captures was performed in python in order to provide a set of images for models' evaluation. The RetinaNet model was utilised for object detection that gave accurate results for the bounding box prediction and object classification. The set of sequential images was then used for training the GAN-based VITAL model. The RetinaNet could also provide exceptional results for video-based object tracking.

The VITAL involves Deep Learning used to simultaneously detect features in order to be aware of motion changes. It utilises the Sample Generator to collect data, extract regions and to provide proposed bounding boxes at the output. Reinforcement learning is approach that approximates values for input signals in unpredictable circumstances [5]

In the Faster R-CNN algorithm one image from the custom image set was transferred through the CNN. The object location is presented with a pair of values: (x, y) that correspond to coordinates of pixels in the image [ymin, xmin, ymax, xmax]. The image is then being resized in order to be adapted for features extraction. Feature detection is essential for object detection and it is used to determine correspondences in order to generate models. The key point features that are examined are specific positions of objects in images. Those features could be matched based on their orientation and local appearance and could indicate object boundaries and occlusion events [6]. The VGG16 pretrained CNN Network processes the adapted image and extracts features with dimensions of 50 x 50 pixels. Around 2000 anchor points are generated on the image. Point features are used to detect the corresponding object locations in different images and that is applied for the category recognition. It is important to determine the key points in order to perform the successful matching in the event of occlusion and motion changes. For each anchor, several anchor boxes are generated. The anchor boxes represent predicted bounding boxes of a certain height and width. They are printed across the image and they are used for multiple object detection. The Intersection of Union (IOU) of bounding boxes is then calculated and if it is higher than 0.7, the object detection should be performed.

$$I_{o}U = \frac{|P \cap G|}{|P \cup G|} = \frac{|I|}{|U|}$$

Equation 1 – Intersection over Union [7]

In equation 1 P represents the predicted bounding box and G represents the ground truth box.

The Non Max Suppression (NMS) algorithm reduces the number of predicted bounding boxes for the particular object. It computes Regions of Interest (ROI) with positive labels, where the IOU is higher than 0.7. The region proposal algorithm processes input images and predicts where potential objects could be, without knowing if there are objects in that location. Regions of Interest (ROI) with IOU lower than 0.3 are classified as negative labels.

3. RESULTS

Figure 3 represents the accurate detection results using the RetinaNet. However, in object-groups' detections the overlapping of bounding boxes could be significant that makes overview of detected items unclear. The algorithm provides us with the bounding boxes coordinates, the classification and the class probability value.

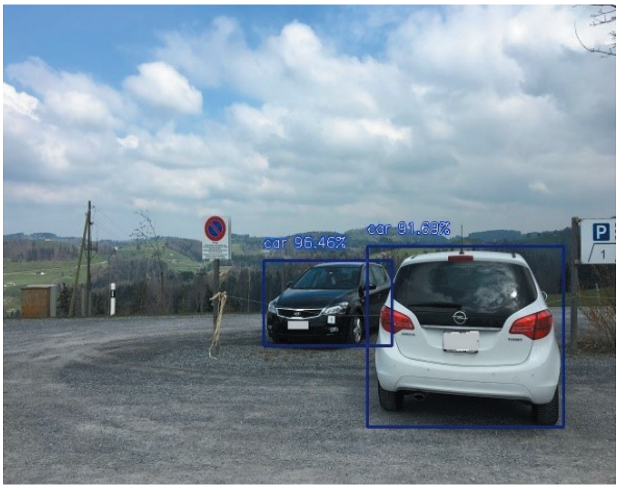

Figure 3 - RetinaNet detection

The VITAL algorithm applies adversarial learning that tracks motion and generates the corresponding bounding boxes that are used for the adaptation of neural network input features. Figures 4, 5, 6 and 7 represent the results of the VITAL method evaluated on the custom set of images. It computes the bounding box corresponding to the positional changes over time (Figure 6). Figures 4 and 5 display the algorithm values of the applied cost sensitive loss. The locations of frames on image sequences are being calculated with the consideration of the distance from the ground truth locations [8]. As the input data for the VITAL algorithm the resulting bounding boxes of specific objects from the RetinaNet computation were used together with the custom set of sequence images.

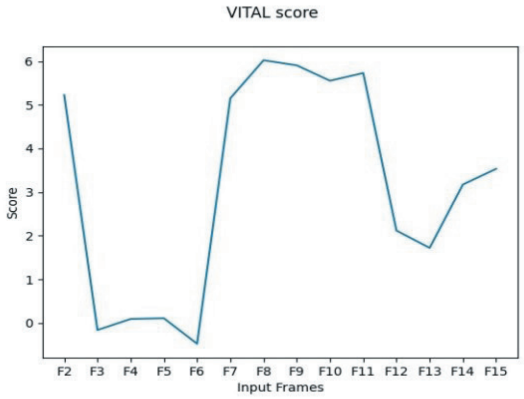

Figure 4 - VITAL score

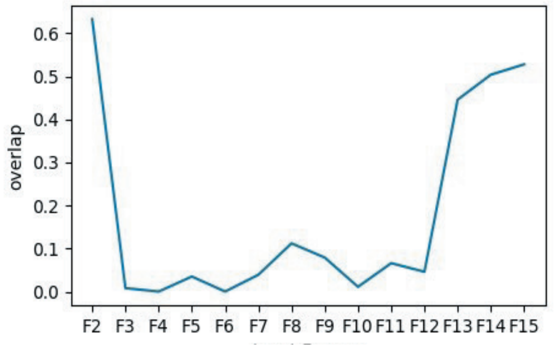

Figure 5 - VITAL overlap

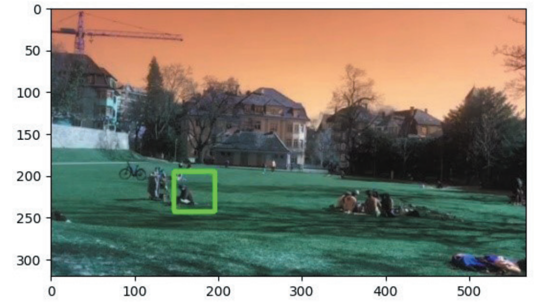

Figure 6 - VITAL estimated bounding box

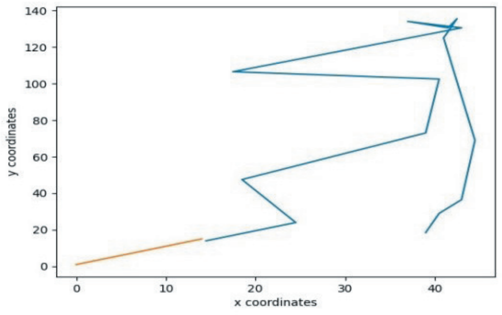

Figure 7 - VITAL tracking trajectory using middle values

In the Faster R-CNN algorithm localized items of interest are specified using bounding boxes. For the Faster R-CNN examination the coordinates extracted from the RetinaNet evaluation were used. One of the difficulties in the Faster R-CNN is that it generates many positive samples that overlap and affect slower performance during the process of recognition. The loss function [2] is calculated by the following equation, where j indicates the anchor, p_j represents the detected class probability, p_j* represents the actual label probability, t_j represents the predicted object position and t_j* is the real position, N_{CLS}, N_{REG} and λ are used for the normalization and balancing:

$$L\left(p_j,t_j\right)=\left(\frac{1}{N_{CLS}}\right)*\sum_j L_{CLS}\left(p_j,p_j^*\right)+$$
$$\lambda*\left(\frac{1}{N_{REG}}\right)*\sum_j p_j^* * L_{REG}\left(t_j,t_j^*\right)$$

Equation 2 – Image Loss function [2]

The resulting loss function was L = 0.6940.

In vehicle tracking the data augmentation is necessary for the classifier and the detector training and the Faster R-CNN performs detailed scanning of images in order to predict the bounding boxes for object detection [9].

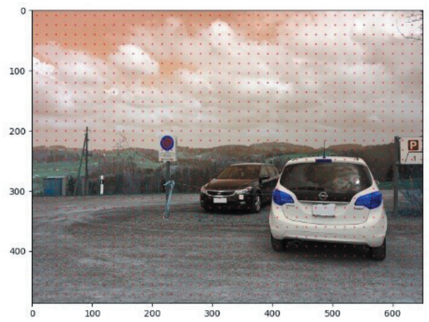

Figure 8 – About 2000 generated anchor points

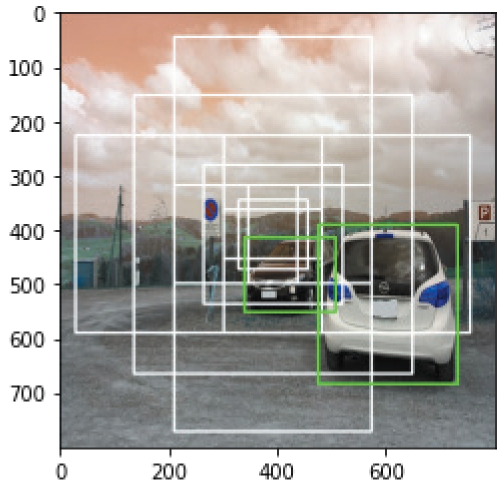

Figure 9 – Generated anchor boxes for one anchor point

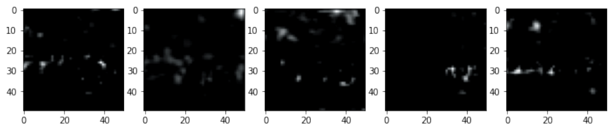

Figure 10 – Extracted features

4. CONCLUSION

Different approaches in object tracking and neural network-based architectures are examined in this paper and the various ideas behind object detection are associated. The reliable outcome of the RetinaNet could be applied for the Faster R-CNN and the VITAL approaches as they require the input data values for the items tracking. For multiple objects tracking the results of these algorithms could be applied for computing data correlation, data association, re-identification, and optimization parameters, that provides a wide field for following improvements.

REFERENCES

[1] M. R. Solomon, Consumer Behavior: Bying, Having and Being, Pearson, 2017.

[2] K. H. R. G. J. S. Shaoqing Ren, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in [*Online*] *Available at https://arxiv.org/pdf/1506.01497.pdf*, 2016.

[3] P. G. R. G. K. H. P. D. Tsung-Yi Lin, "Focal Loss for Dense Object Detection," in *[Online] Available at https://arxiv.org/abs/1708.02002,* 2017.

[4] L. F. R. S. P. G. a. M. M. Marko Tanaskovic, "Adaptive model predictive control for constrained linear systems," in *European Control Conference (ECC)*, Zürich, Switzerland, 2013.

[5] M. B. S. S. S. Milos S. Stankovic, "Distributed Value Function Approximation for Collaborative Multi-Agent Reinforcement Learning," in *https://arxiv.org/abs/2006.10443*, 2021.

[6] R. Szeliski, Computer Vision: Algorithms and Applications, Springer, 2010.

[7] H. a. T. N. a. G. J. a. S. A. a. R. I. a. S. S. Rezatofighi, "Generalized Intersection over Union," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] C. M. X. W. L. G. L. B. W. Z. C. S. R. L. M.-H. Y. Yibing Song, "VITAL: VIsual Tracking via Adversarial Learning," in *https://arxiv.org/abs/1804.04273*, 2018.

[9] J. W. X. Y. Yongjie Zhang, "Real-time vehicle detection and tracking in video based on," Beijing 100084, China, 2017.