Person Detection from a Top-View Perspective

Baldascino Giovanni, Loisi Squarcella, Marco Mameli, Marina Paolanti, Giulia Pazzaglia, Emanuele Frontoni

***Abstract*—This article presents a method for identifying people from top-view perspective using the most recent approaches in order to be able to identify how many and what people are within an image. These technologies have been proposed to increase the efficiency and accuracy of detection of people compared to the previous ones and compared to the classic human approach. The dataset was collected in a real scenario, consisting of 17,000 images, each containing at least one person, and an .XML file containing the annotation. The network used was the SSD, Single-shot Detector, concentrating the work on three architectures: SSD7, SSD300 and SSD512. The initial approach was to train the three networks from scratch and, at the same time, fine-tune the weights coming from pre-trained networks on the PascalVOC 07 + 12 dataset; a further study was carried out by modifying the feature extractor present in the networks, replacing the VGG16 with the ResNet50. The metrics used to evaluate the models were the same for all networks: the AP (Average Precision), the Recall, F1 Score and IoU (Intersection over Union). An attempt was made to modify the loss function using the evaluation metrics. The results obtained were satisfactory for the networks trained from scratch, both with VGG16 and with ResNet50, while fine tuning proved to be the worst.**

I. Introduction

Convolutional neural networks are the state of the art for the solution of problems of classification, localization and identification of objects within images or videos. We speak of classification and localization if the input images contain at most one element of a single class and we want to identify both the element and the coordinates of the bounding box. If there are multiple elements of the input image any class and you are interested in the classification and localization of each of them, then we speak of object detection. One branch of object detection is person detection which, as can be seen from the name, is responsible for identifying how many and which people are within an image. If we think of today's object detection as a technical aesthetics under the power of deep learning, then turning back the clock 20 years we would witness "the wisdom of cold weapon era". Most of the early object detection algorithms were built based on handcrafted features. Due to the lack of effective image representation at that time, people have no choice but to design sophisticated feature representations, and a variety of speed up skills to exhaust the usage of limited computing resources. As the performance of hand-crafted features became saturated, object detection has reached a plateau after 2010. In 2012, the world saw the rebirth of convolutional neural networks. As a deep convolutional network is able to learn robust and high-level feature representations of an image, a natural question is whether we can bring it to object detection? R. Girshick et al. [5][6] took the lead to break the deadlocks in 2014 by proposing the Regions with CNN features (RCNN) for object detection. Since then, object detection started to evolve at an unprecedented speed. In recent years, technological development has allowed an increasing use of neural networks based on the recognition of objects and/or people. An example is found overseas, in the U.S.A., where Amazon has opened supermarkets based mainly on the recognition of people, through the use of cameras and sensors, thus increasing safety and ensuring high reliability. Developments using technologies of this type are almost infinite. This article is structured as follows: all related works are presented in section II. Section III discusses the methods used for the dataset and the networks with the related experimental protocols. Section IV presents all the results obtained and the related discussions. The conclusions are presented in Section V. Finally section VI with references.

Dipartimento di Ingegneria dell'Informazione (DII), Università Politecnica delle Marche, Via Brecce Bianche 12, 60131, Ancona (Italy), m.mameli@univpm.it, m.paolanti@univpm.it, g.pazzaglia@univpm.it, e.frontoni@univpm.it, s1096539@studenti.univpm.it, s1097405@studenti.univpm.it

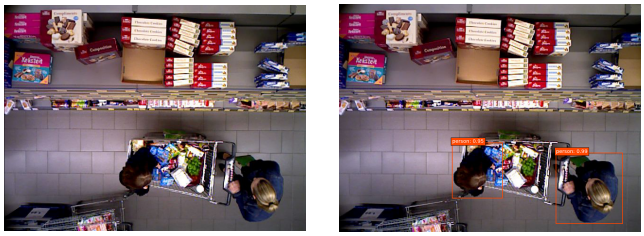
II. Related Works

19 years ago, P. Viola and M. Jones achieved real-time detection of human faces for the first time; the detection algorithm, which was later referred to the "Viola-Jones(VJ) detector", was herein given by the authors' names in memory of their significant contributions. In the past two decades, it is widely accepted that the progress of object detection has generally gone through two historical periods: "traditional object detection period (before 2014)" and "deep learning based detection period (after 2014)". In deep learning era, object detection can be grouped into two genres: "two-stage detection" and "one-stage detection", where the former frames the detection as a "coarseto-fine" process while the later frames it as to "complete in one step" [5][6]. Two-stage detectors have high localization and object recognition accuracy, whereas the one-stage detectors achieve high inference speed.

Misbah Ahmad et al. [1] talk about the importance that person detection is having in recent years and in addition offers a clear differentiation of the architectures that are based on Blobs and those that are based on Features.

Wei Liu et al. [2] propose a detection approach using a single neural network: the SSD; this approach ranks among the best among those based on the extraction of features, as opposed to blob-based architectures.

Zhengxia Zou et al. [3] and L. Jiao et al. [4] list and describe in detail all the different existing networks, including the SSD, for person detection and in addition to outlining the advantages and disadvantages they offer an explanation of the standard metrics.



(a) Image without prediction (b) Image with prediction

Fig. 1: An example of recognizing multiple people

III. Materials and Methods

This section analyzes the methodologies used for the study of the problem. Section A describes the Dataset and the annotations. Section B presents the architectures used for the detection. Training settings are exposed in Section C. Finally Section D lists the metrics used for evaluation.

A. Dataset

The dataset, annotated using LabelImg, an image annotation tool, was acquired by three cameras over several days. The scenario is a retail environment, where people can be present also grouped or with shopping carts or carrying shopping baskets. Starting from a dataset of about 50,000 images, an initial skimming was done by labeling only the images with one or more people. Made up of 17,000 images, the dataset was divided into 30% Test Set and the remaining 70% divided into 70% Training Set and 30% Validation Set.

B. Detection Network and Features Extraction

Person Detection from a top view perspective (Fig. 1) has been implemented with the SSD using three different architectures: SSD300, SSD512 and SSD7. Starting from the original Caffe implementation provided by Wei Liu et al. [2], the networks have been adapted to the needs of the dataset used, to be able to train them from scratch. At the same time, always from Wei Liu et al. [2], weights of pretrained nets were obtained on the PascalVOC 07 + 12 dataset; fine tuning was necessary to move from the 20 classes, present in PascalVOC, to the single "person" class. The feature extractor of the original Caffe implementation is implemented with VGG16. A change was made to the backbone of the networks, replacing it with the ResNet50. VGG16 was composed of five blocks, each consisting of some convolutional layers and a final MaxPooling. The installed ResNet50 consists of four blocks, the first of which consists of a ZeroPadding, a convolutional layer, a BatchNormalization, another ZeroPadding and a final MaxPooling; Skip Connections have been implemented through two types of sub-blocks: convolutional blocks and identity blocks. The remaining three blocks of the ResNet50 are composed of a convolutional block and some identity blocks. In the Table I and Table II we see the architectures of the VGG16 and ResNet50 used respectively. The activation function used for all networks was the ReLu. Adam was used as an optimizer.

Block	Type	Filter
1	Convolution	64
	Convolution	64
	MaxPooling	
2	Convolution	128
	Convolution	128
	MaxPooling	
3	Convolution	256
	Convolution	256
	Convolution	256
	MaxPooling	
4	Convolution	512
	Convolution	512
	Convolution	512
	MaxPooling	
5	Convolution	512
	Convolution	512
	Convolution	512
	MaxPooling	

TABLE I: VGG16 backbone architecture

Block	Type	Filter
1	ZeroPadding	
	Convolution	64
	BatchNormalization	
	ReLu	
	ZeroPadding	
	MaxPooling	
2	ConvBlock	[64, 64, 256]
	IdenBlock	[64, 64, 256]
	IdenBlock	[64, 64, 256]
3	ConvBlock	[128, 128, 512]
	IdenBlock	[128, 128, 512]
	IdenBlock	[128, 128, 512]
	IdenBlock	[128, 128, 512]
4	ConvBlock	[256, 256, 1024]
	IdenBlock	[256, 256, 1024]
	IdenBlock	[256, 256, 1024]
	IdenBlock	[256, 256, 1024]
	IdenBlock	[256, 256, 1024]
	IdenBlock	[256, 256, 1024]

TABLE II: ResNet50 backbone architecture

C. Training Settings

For each training carried out, the learning rate was set at 0.001, the number of epochs of 40 for SSD300 (Fig. 2 and Fig. 3) and SSD512 (Fig. 4 and Fig. 5), both with VGG16 and with ResNet50, while 70 for SSD7 (Fig. 6). The batch-size for SSD300, with both extractors used, has been set to 32 while for SSD512 with VGG16 at 16 and with ResNet50 at 12 because the ResNet50 has a higher memory occupation and it was not possible to keep the batch size of the same value. The steps for epoch have been calculated by applying the formula (1), thus having 260 steps for both architectures of the SSD300, 520 steps for the SSD512 with VGG16 and 690 steps for the SSD512 with ResNet50.

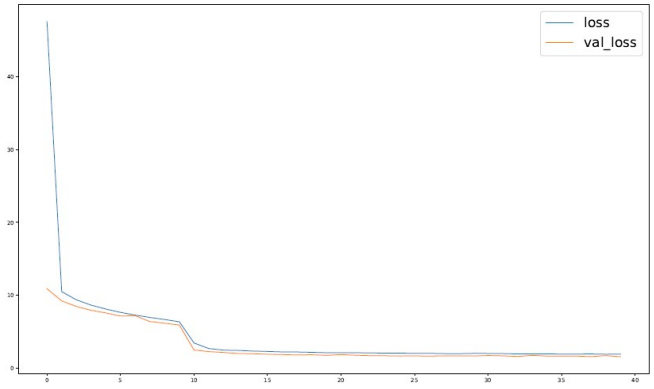


Fig. 2: SSD300-VGG16 Training and Validation Loss

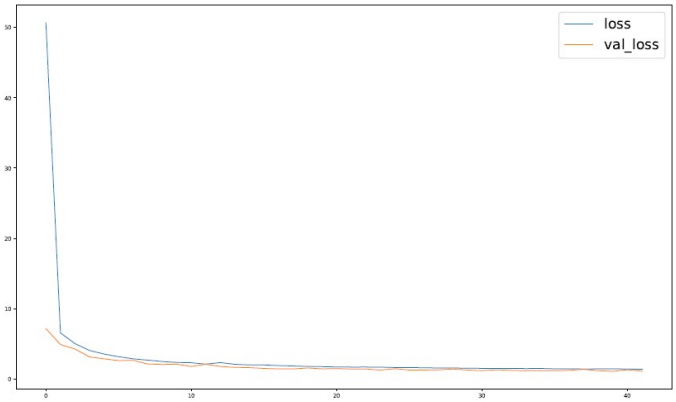


Fig. 3: SSD300-ResNet50 Training and Validation Loss

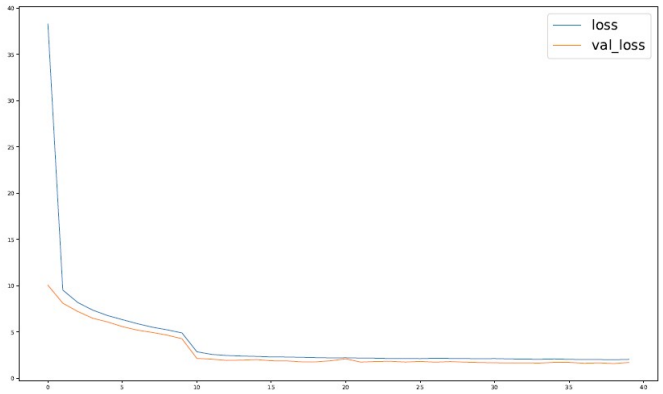


Fig. 4: SSD512-VGG16 Training and Validation Loss

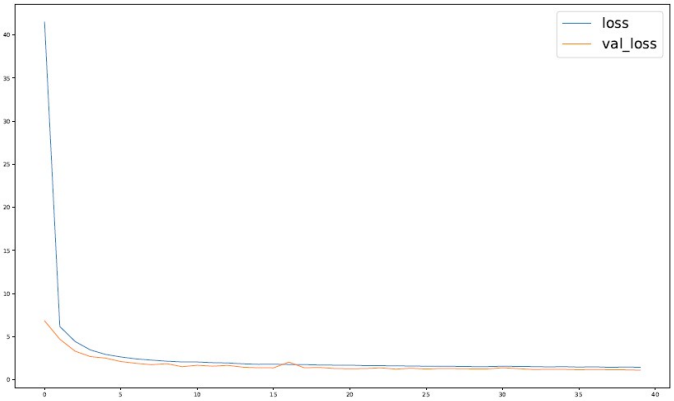


Fig. 5: SSD512-ResNet50 Training and Validation Loss

$$Steps = \frac{TrainingSet}{BatchSize} \tag{1}$$

The Loss function consists of two terms: Lconf and Lloc where N is the matched default boxes.

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \tag{2}$$

Lloc is the Localization Loss which is the smooth L1 loss between the predicted box (l) and the ground-truth box (g) parameters. Lconf is the Confidence Loss which is the softmax loss over confidences (c).

D. Perforance Metrics

The metrics used to evaluate the performance of the networks were manifold: the AP and the Recall calculated using the official PascalVoc script; precision and recall are defined as follows:

$$Prec = \frac{TP}{TP + FP} \quad \text{and} \quad Rec = \frac{TP}{TP + FN} \tag{3}$$

Where TP is True Positives, FP is False Positives and FN is False Negative. For further analysis, two additional metrics were used: F1 Score which measures the accuracy of a test

$$F_1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{4}$$

and the IoU, or Jaccard index, calculated from the relationship between the intersection and the union of the ground-through with the box predicted; a necessary clarification is that the IOU was used as a metric only during the testing phase to assess that the Bounding Boxes were actually correct.

IV. Results and Discussion

The results are shown in the Table III.

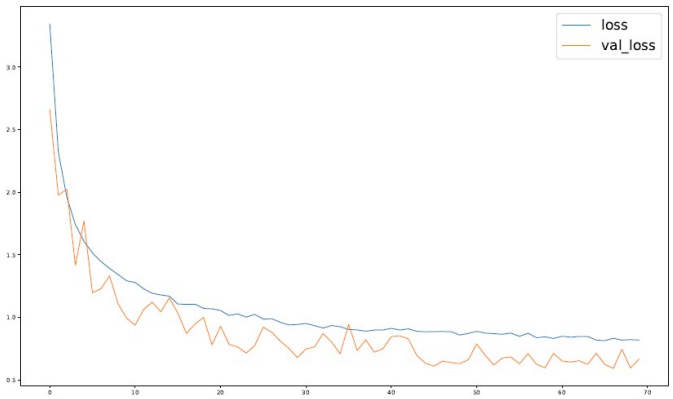


Fig. 6: SSD7 Training and Validation Loss

Fig. 7: Prediction of the SSD 512 with VGG16, in green the ground truth and in orange the prediction

Network	AP	Rec	F1	IoU
SSD300 - VGG16	0.908	0.818	0.861	0.879
SSD300 - ResNet50	0.909	0.618	0.736	0.842
SSD300 - Fine Tuning	0.643	0.887	0.746	0.712
SSD512 - VGG16	0.908	0.912	0.910	0.801
SSD512 - ResNet50	0.909	0.872	0.890	0.846
SSD512 - Fine Tuning	0.744	0.934	0.828	0.885

TABLE III: Table of results

In the results table there is no SSD7 since during the training phase it showed loss values that are too fluctuating to evaluate. Focusing, first of all, on the networks trained from scratch, the results show how the average precision obtained is excellent for all the networks, with very similar values but never lower than 90%; while, as regards recall, although all networks offer equally good results, those proposed by VGG-16 have proved to be better than those of resnet-50. There is a further evaluation with the F1-score values which confirms that the best architecture, for this problem, is the SSD512 with VGG-16 as seen in Fig. 7, where the network also predicts a person without ground truth. Secondly, the results relating to fine tuning offer better recall values than the networks trained from scratch but all at the expense of the average precision which is clearly lower; the F1-score confirming the fact that fine tuning proved to be the worst approach. Finally we find exposed the IoU values, which as already specified were used only in the testing phase, which proved to be acceptable, showing good predictions.

V. Conclusion

The approach offered in this paper has proved to be very good with the aim of solving the problem of person detection from top-view.

Future works could be based on the modification of the feature extractor with other architectures, for example DenseNet, to confirm or not that VGG16 remains the best for this type of detection. Another development could focus on the use of IoU metrics as Localization Loss in the calculation of the Loss Function, replacing the Confidence Loss with another metric, using machines with more memory for computing.

References

[1] Misbah Ahmad, Imran Ahmed, Kaleem Ullah, Iqbal khan, Ayesha Khattak, Awais Adnan, ”Person Detection from Overhead View: A Survey”.

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, ”SSD: Single Shot MultiBox Detector”.

[3] Zhengxia Zou, Zhenwei Shi,Member, IEEE, Yuhong Guo, and Jieping Ye,Senior Member, ”Object Detection in 20 Years: A Survey,” IEEE.

[4] L. Jiao et al., ”A Survey of Deep Learning-Based Object Detection,” in IEEE Access, 2019, vol. 7, pp. 128837-128868.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ”Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp.580-587.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ”Region based convolution with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097-1105.