Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Population Statistics Algorithm Based on MobileNet

To cite this article: Xiaoqin Feng *et al* 2019 *J. Phys.: Conf. Ser.* **1237** 022045

Population Statistics Algorithm Based on MobileNet

Xiaoqin Feng, Rong Xie, Junyang Sheng, Shuo Zhang

Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, Beijing 100124, China

Email: fengxqin@emails.bjut.edu.cn

Abstract. In today's society, intelligent video surveillance plays an important role in social security, traffic scheduling, national security and other fields. One of the research hotspots is people statistics based on image processing, which has strategic significance in practical applications. Aimed at the problem that the low accuracy in the actual application scenario, the limited hardware resources, and the low operation efficiency, this paper proposes a multi-feature target detection model based on the lightweight deep learning network MobileNet [1], which can be used in intelligent terminals. The basic feature-extraction network MobileNet as a lightweight network can provide a flexible alternative configuration in terms of efficiency and accuracy. The underlying detection network selects a single deep nerual network, named SSD [2]. The algorithm can achieve multi-scale target detection, and uses the target position and category to perform one-time regression. In this paper, the activation function of SSD is changed into SeLU (scaled exponential linear units) [3], which improves the robustness of the algorithm. At the same time, the work of sample diversity and data enhancement has been made, and the characteristics of the human body above the shoulders have been fully utilized. Experiments have shown that the improved network structure based on MobileNet has higher detection accuracy, lower delay, excellent robustness, while the number of model parameters is effectively reduced.

1. Introduction

With the development of economy and society, the population of China has an exponential trend, and the statistical research on the number of people in various fields of life has become a realistic demand. At the same time, with the continuous development of computer vision technology, image processing technology has broad application prospects and development space in this field. However, in practical application scenarios, due to target occlusion, contour non-rigidity, attitude variability, image resolution, ambient illumination and complex environment, etc., it poses a huge challenge to accurate population statistics. In addition, due to the limited resources of embedded intelligent terminals, the computational complexity of image algorithms based on deep learning makes the application of algorithms a major research problem. Designing a simple, effective, robust, real-time, and accurate population statistics algorithm has become a challenge.

At present, the population statistics algorithm combines multiple techniques such as background analysis, feature extraction, object detection, target segmentation, and target tracking. According to different means, it is mainly divided into two categories: 1) Indirect method, also known as feature-based method, refers to establishing a functional relationship between target characteristics and number of people to perform metric calculation. Hashemzadeh M [5]et al. proposed population density and population occlusion characteristics based on key points and foreground segmentation to estimate the number of people; Chang QL [6] et al proposed to calculate the population occlusion factor by normalizing the foreground and corner information to pass the two features through backward propagation(BP). The network completes the number of people, etc. The above method is mainly applied to scenes where pedestrians are highly dense, by extracting features in the image and establishing an estimated function statistic. 2) Direct method: The core idea is to extract the target features, learn a classification model to achieve target detection, and the classic pedestrian detection model of HOG+SVM proposed by Navneet Dala [7]. The classification model of SVM is established by extracting the HOG (gradient histogram feature) feature of the image; similarly, the Haar-like+Adaboost cascade classification model proposed by Paul Viola [8]. In addition, the convolutional feature based on deep learning combined with traditional machine learning classification has become the current mainstream method, with high precision and strong generalization ability [4].

At present, people believe that the shortcomings in this aspect are in three parts [9]: 1) Shallow learning. The accuracy is not up to the requirement. Although a better method is proposed, such as the DPM algorithm based on the human body component, it has a better detection effect in a complicated environment. However, because of the existence of multiple classification models, the efficiency is difficult to improve [10] 2) Deep learning. Based on various CNN algorithms, it is mainly divided into one-stage and two-stage. Typical examples include RCNN, Faster-RCNN, SSD, etc. Because of their strong feature expression ability, these algorithms can achieve a better detection results and strong generalization ability in combination with traditional machine learning classifiers. However, the depth of its model and the huge amount of computation make it difficult to use in real life. 3) Data labeling problem, currently there is no standard data set for the population statistics application scenario, such as the PASCAL-VOC data set, the standard is the whole body. The coverage area is relatively wide and there are other types of coverage. This paper mainly constructs a comprehensive human head detection data set, based on the current popular lightweight network MobileNet to achieve image feature extraction. The lower layer adopts the SSD single detection model and adds the activation function SeLU. Experiments show that the improved network structure based on MobileNet has higher detection accuracy, lower latency, better robustness, and the number of model parameters is effectively reduced.

2. Model

This paper selects MobileNet as the feature extraction network and SSD as the underlying target detection framework. There are two versions of MobileNet, and the activation function of the overall model structure change into SeLU function, which improves the robustness of the algorithm

2.1. MobileNet:Feature extraction network

MobileNet, mainly a lightweight deep neural network proposed by Google to solve the problem that mobile embedded terminals cannot be applied. Its research direction lies in the aspect of model compression, and its core idea is the ingenious decomposition of convolution kernel. It can effectively reduce network parameters while taking into account optimization delay.

2.1.1. MobileNetV1

The network design is based on a streamlined architecture that uses a deeply separable convolution to build a lightweight deep neural network. Introducing two global hyperparameters: the width parameter and the resolution parameter, effectively balancing the efficiency and accuracy. These two parameters allow us to choose the right model for our application based on actual problems.

Depthwise convolution applies convolution kernels to each channel, the 1*1 pointwise convolution is used to combine the output of channel convolutions. This kind of idea can achieve the same effect as traditional convolution, but it can effectively reduce the size of the model. Fig 1, 2 show a comparison of standard convolution and depthwise separable convolution.

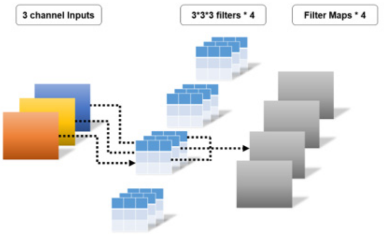


Fig.1 Standard Convolution: The three-channel RGB image input as M*M passes through the convolution process of 3*3*3*4 convolution kernel (eg: input channel: 3, output channel: 4), and finally outputs 4 feature maps. The specific size also refers to the padding and strides parameters

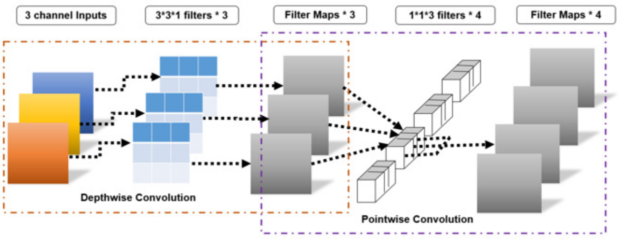


Fig.2 Depthwise Separable Convolution: Decompose a standard convolution operation into two steps, namely Depthwise Convolution and Pointwise Convolution.

It can be found that, unlike traditional convolution operations, Depthwise Convolution performs a separate convolution for each channel, that is, a channel is convolved by only one convolution kernel. The number of Feature maps after depth convolution is the same as the channel of input layer, and the feature maps cannot be extended. And this operation performs convolution operation independently for each channel of the input layer, and cannot effectively utilize the feature information of different channels in the same spatial position. Therefore, Pointwise Convolution is required to combine these feature maps to generate a new feature map. In this separation operation, the amount of calculation is greatly reduced, and the calculation ratio (CCR=DW Conv Cost / Std Conv Cost) of the two is calculated by the following formula (width hyper-parameter and resolution hyper-parameter are omitted)

$$\mathrm{CCR} = \frac{G_k^2 * M * G_f^2 + M * N * G_f^2}{G_k^2 * M * N * G_f^2}$$

$$\mathrm{CCR} = 1/N + 1/G_k^2 \tag{1}$$

Where N is the channel of feature maps, usually larger, usually greater than 10. G_k^2 is the size of the convolution kernel, typically 3*3, and the ratio is a number less than one.Therefore, the DW convolution reduces the amount of computation compared to the Standard convolution. In addition, the formula for calculating the network after adding two hyper-parameters is as follows:

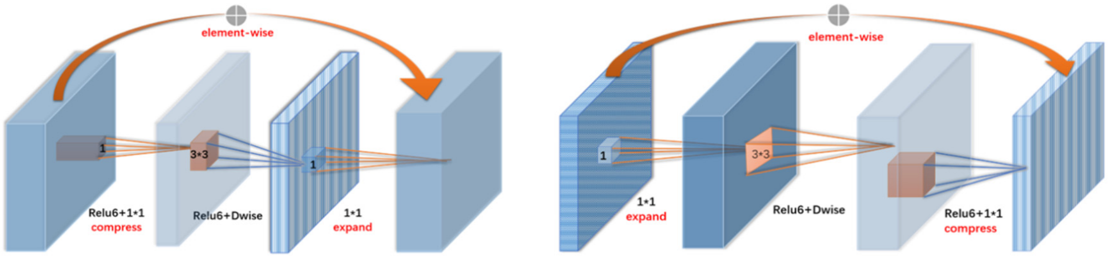
$$G_k^2 * \alpha M * \rho G_f^2 + \alpha M * \alpha N * \rho G_f^2 \tag{2}$$

Where α is the width parameter, the role is to change the input and output channels, reduce the number of feature maps, the value is 0~1; ρ is the resolution parameter, the role is to change the resolution of the input layer. The combination of the two can further reduce the amount of calculation.

2.1.2. MobileNetV2

On the basis of MobileNetV1, an upgraded version is proposed, which ensures the accuracy of the model and reduces the amount of calculation [11].

Two basic structures are proposed: 1) Linear Bottlenecks, which replaces the ReLU layer with a linear transformation layer to reduce information loss. 2) Inverted Residual Structure, which uses a shortcut to link the block's input and output (element-wise). Compared to the original residual structure, this reverses the order of transformation of the internal data dimensions, which saves memory and enriches feature information. Fig 3 shows the comparison between the original residual structure and the Invert residual structure:



a) Origin residual block

b) Invert residual block

Fig.3 Comparing the original RBlock, MobileNetV2: 1) 'expand' the channel of the feature map after a 1*1 Conv layer; 2) extract the feature information via the 3*3 Conv layer; 3) after a 1*1 Conv Layer, 'compresses' the feature map channel back to its original size. The original is to first "compress" and then "expand"

2.2. *Improved SSD Structure*

SSD is a single-shot detection model. The core idea is to obtain the location and category of the target in a regression manner on the multi-scale feature map. However, there is a problem of inaccurate detection of small targets. Through analysis, the factors affecting the detection of small targets are mainly the resolution of the feature map and the global information and feature extraction capabilities. The residual block structure of MobileNetV2 can improve the high-resolution low-level feature expression of the feature map used for SSD detection. In addition, the original SSD framework uses the ReLU activation function, which is modified into a SeLU activation function. The network obtained by the activation function has self-normalization characteristics. Among them, ShaoHua. et al. compared SeLU with other activation functions, which proved its effectiveness and robustness, even surpassing batch normalization.

Comparison of ReLU & SeLU: People know that the addition of a neuron to an activation function has the ability of non-linear representation, which is the biggest difference between neural networks and linear classifiers. Compared to the original activation function Sigmoid, ReLU has three differences: 1) unilateral suppression 2) wider excitation boundary 3) sparse activation. ReLU can transfer gradients very well: after repeated back propagation, the gradient will not be greatly reduced, suitable for training deep neural networks. However, experiments have shown that the ReLU activation function is prone to training interruptions, and ReLU-enforced sparse processing reduces the effective parameter capacity of the model (when x < 0, the negative gradient is zeroed and will no longer be activated by any data. It is called neuron "necrosis"). One of the similarities between ReLU and Sigmod is that the results are all positive, reducing the ability to express features

In 2017, the literature [24] introduced a new activation function SeLU (scaled exponential linear units), introducing the properties of self-normalization. The following are the mathematical formulas for the SeLU and ReLU:

$$ReLU_{(z)} = \begin{cases} z & z > 0 \\ 0 & z \leq 0 \end{cases} \tag{3}$$

$$SeLU_{(z)} = \gamma \begin{cases} z & z > 0 \\ \alpha(exp(z) - 1) & z \leq 0 \end{cases} \tag{4}$$

$$where\ \gamma \approx 1.0507, \ \alpha \approx 1.673$$

SeLU mainly uses a function F to establish the mapping relationship of the neural network layer. At the same time, the parameters are transformed to a fixed mean and variance to achieve the normalized effect. Compared with ReLU, it has the following advantages: 1) Strong convergence property, even if there is noise and interference in the data, it will converge faster after multi-layer forward and backward propagation; 2) Regularization effect, enhance Algorithm robustness; 3) In addition, for the excitation values that do not approximate the unit variance, the variance has upper bound and lower bound, so gradient disappearance and gradient explosion are almost impossible.

Model Structure:This article selects the SSD detection framework based on MobileNet. Adding 8 convolution layers behind the conv13 of MobileNet. A total of 6 layers are extracted for detection, and the activation function is SeLU. Fig4 is the model structure diagram of this paper.

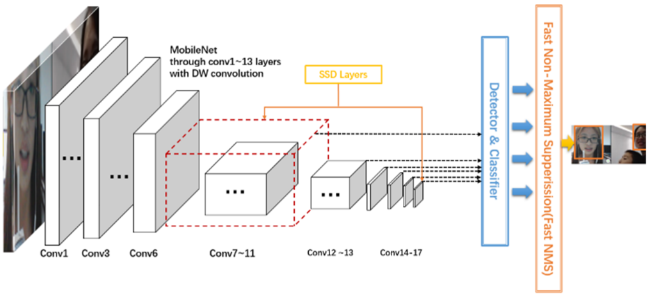


Fig.4 Model structure, the left part is the feature extraction network MobileNet, which contains the output of the conv1-conv13 layer. After the conv13, 8 layers of convolution are added, and 4 layers and conv11 and conv13 of MobileNet are selected as the scale feature map of the SSD, and finally attributed to a prediction unit. The activation function in the middle uses the SeLU function.

3. Experiments

In this section, introducing the experimental operating environment, data set, evaluation criteria, and training process in turn. Finally, the effects of different models on the same data set are compared.

3.1. Experimental Preparation

TABLE I. DATASETS

DataSet	***Labels***	***Datasize***	***Pixelsize***
Pascal VOC 2007-06	21	9963	500*375
Pascal VOC 2012-11	21	17125	500*375
INRIA	2	2416	96*160
MyData	2	5000	300*375

The experimental hardware environment is Ubuntu16.04+cuda8.0+GTX1080. Using Keras as a deep learning framework to define different models, and the back end is a Tensorflow implementation that supports GPU computing. The data set is shown in Table I. In order to construct the omnidirectional data set of the human head, the INRIA pedestrian detection data set and the PASCAL-VOC 2007+2012 database are used. However, the data that meets the requirements is still relatively small. Therefore, a large number of image data containing the target (person) are captured online and relabeled using the LabelImg tool and the PASCAL-VOC labeling format was used. In order to get all-round and feature-rich data, not only focus on the facial skin when marking, and the features above the shoulder will be added to the target frame. The PASCAL-VOC data has also been recalibrated to obtain richer and more accurate feature information. In addition, over-fitting of data information is avoided, and the generalization ability of the model is increased. The code and dataset will be published later on github.

In order to enhance the complexity of the data, series of data enhancement work was putted into the model, such as random clipping, brightness transformation, etc.

3.2. Evaluation Criterion

In this paper, the target detection model is selected, and people counting in different scenarios is applied. RMSE is selected as the evaluation criterion. The calculation formula is:

$$RMSE(X,f) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x^{i}) - y^{(i)})^{2}} \tag{5}$$

Where $\boldsymbol{y}^{(i)}$ is the number of original people in each picture, $\boldsymbol{f}(\boldsymbol{x}^{i})$ is the number of people detected, and m is the number of video frames detected. In addition, this article uses the frames per second (FPS) to measure the target detection speed and uses 25 fps as the real-time consideration threshold.

3.3. Experimental Results and Analysis

In order to compare the network parameters, detection effect and real-time performance, the popular VGG convolution network is choosed, which is more popular now. Like AlexNet, the network believes that the increase of the depth of convolution neural network and the use of small convolution kernels have a great effect on the final classification and recognition effect of the network [12]. However, it also loses the real-time performance of detection and is difficult to be used in practice.

TABLE II. RESULTS

BaseModel	***params***	***RMSE(m=500)***	***FPS***

VGG16(ReLU)
MobileNetV1(ReLU)
MobileNetV2(ReLU)
	26M	0.89	46

	7M	1.25	67
	6M	0.93	85

VGG16(SeLU)
MobileNetV1(SeLU)
MobileNetV2(SeLU)
	26M	0.82	47

	7M	1.19	67
	6M	0.87	85

The VGG16 + SSD and MobileNet + SSD detection frameworks are constructed, in which the input size of the picture is 300 * 300.The parameters of the model include the number of bbox, scales, aspect_ratios , position offset, and so on. The weight of the basic feature network is selected by ImageNet pre-training weights, and the number of training rounds is 10000+. Finally, a video is selected for comparison. Finally, a video is selected for comparison, and the RMSE is calculated based on the actual number of people and the number of people in each frame. A series of statistics were made on the operating parameters and efficiency of the model. Table 2 compares the results and compares them in terms of parameter size, model accuracy, and model efficiency.

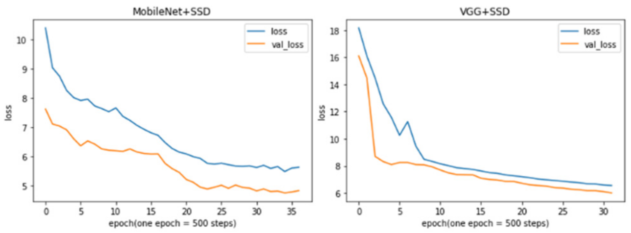


Fig.5 In Loss, the contrast between MobileNet and VGG

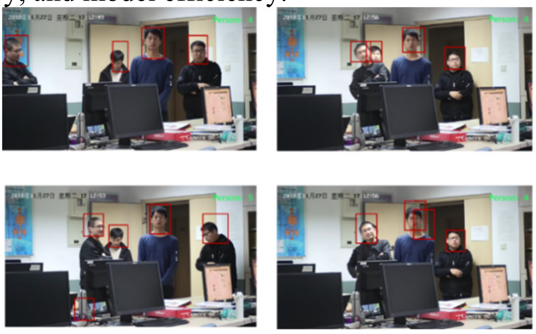


Fig.6 Partial visualization results for MobileNetV1+SSD

Analysis of the results: 1) Through the visualization of the loss (Fig 7), it can be seen that both MobileNet and VGG can converge to a range more quickly, showing only 0-40epoch results (one period = 500 steps); 2) all models are Verify on the video (Fig 8). The experiment saves each frame of the video as an image and extracts some of the results from this article to the network MobileNet + SSD. There are several types of phenomena: a) better recognition persons with bow and partial occlusion; b) cases with recognition errors; c) cases with repeated face detection 3) Comparing the final evaluation criteria RMSE (Table II), MobileNet is slightly worse than VGG. In real-time testing, the time to infer a frame is much better than VGG. In practical applications, it usually involves modules such as video capture, encoding, preprocessing and decoding. This is a work that needs continuous improvement, and it is hoped that it will be real-time in the future.

4. Conclusion & Future

In this paper, the existing statistical algorithms in various fields are analyzed. Starting from the two directions of efficiency and accuracy, this paper proposes a population statistics algorithm based on lightweight network MobileNet, and makes some optimization to improve the robustness of the algorithm. The labeling of the data set is modified for the scenario of population statistics. Three kinds of data sets are used for experiments, and data enhancement is added to the later training. Compared with other complex networks, it loses a bit of accuracy, but in real time, it has been greatly improved.

It is a good direction to do some work on the number of people based on image processing: one is that in the selection of the data set, the characteristics of the crowd in different scenes have a certain tendency, such as the crowd on the bus, the body has occlusion, but The head information is relatively complete; the crowd at the mouth of the passage, the different passages, the density of the crowd and the way of moving are different. The characteristics of the single passage are very comprehensive, which is very suitable for counting statistics and identifying relevant. The second is to do more research work on model compression, such as model structure, model parameter precision transformation, distributed parallel computing and so on. At the same time, for the statistical application of people in different scenarios, different model structures are proposed, or a certain prior knowledge is extracted and model training is added. The algorithm model can automatically adapt to the statistics of people in different scenarios, which is a research direction in the future.

References

[1] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017

[2] Liu W , Anguelov D , Erhan D , et al. SSD: Single Shot MultiBox Detector[J]. 2015.

[3] Klambauer G, Unterthiner T, Mayr A, et al. Self-Normalizing Neural Networks[J]. 2017.

[4] Zhang Junjun, Shi Zhiguang, Li Jicheng. Research Status and Trends of Population Statistics and Population Density Estimation Techniques [J]. Computer Engineering and Science, 2018.

[5] Hashemzadeh M , Farajzadeh N . Combining Keypoint-Based and Segment-Based Features for Counting People in Crowded Scenes[M]. Elsevier Science Inc. 2016.

[6] Qinglong C, Hongshan X, Li Ning. A statistical method for population size based on normalized foreground and corner information[J]. Journal of Electronics & Information Technology, 2014(2).

[7] Dalal N , Triggs B . Histograms of Oriented Gradients for Human Detection[C]// null. IEEE Computer Society, 2005.

[8] Viola P , Jones M . Rapid Object Detection using a Boosted Cascade of Simple Features[C]// null. IEEE Computer Society, 2001.

[9] Gao Fei, Minqiang F, Minqian W, et al. Research on the Method of Population Statistics Based on the Definition of Hot Spots[J]. Computer Science, 2017(S1): 183-188+211.

[10] Guoshu Z, Qiuzhen Z, Wang Hui. Statistics of video indoors based on deep learning SSD model[J]. Industrial Control Computer, 2017(11): 51-53.

[11] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[J]. 2018.

[12] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.