ISSN: 2664-9640

An improved YOLOV4 applied to video detection of face masks

Xin Zhang[1], Shunyong Zhou[2]

[1] School of Automation & Information Engineering, Sichuan University of Science & Engineering, Yibin 644000, China;

[2] Artificial Intelligence Key laboratory of Sichuan Province, Yibin 644000, China.

Abstract

Deep learning has great potential in many practical applications in different fields, one of which is target detection. Due to the demestic novel coronavirus, it is required to standardize the wearing of masks in public places, especially in important places such as high-speed rail stations with heavy traffic. It is necessary to pay more attention to preventing the novel coronavirus. Therefore, real-time mask wearing detection for pedestrians has become critical. The improved target method YOLOV4 is used to realize the detection of wearing masks. By replacing the backbone network CSPdarknet53 in YOLOV4 with a lightweight convolutional structure MobileNetV3 network, the parameters of the neural network are reduced. The experiment results shows that under the same configuration, more pictures can be detected in one second, and real-time video detection can be faster while ensuring accurary. The FPS of video detection increased from 21 to 32 in YOLOV4. When the threshold was 0.5, the average accurary rate mAP increased from 92.935% of YOLOV4 to 93.39%. Compared with other detection methods, it still maintains a good detection speed, which is better than YOLOV3, SSD and other methods.

Keywords

Target detection, YOLOV4, CSPdarknet53, MobileNetV3.

1. Introduction

With the gradual expansion of the impact of the Novel Coronavirus, it is necessary to wear masks and conduct body temperature monitoring in crwoded places such as railway stations, passenger stations, and airports to prevent the spread of the Novel Coronavirus. At present, the main task is to manually check and ask questions and supervise the wearing of masks by the staff. This method involves large number of manual detection of body temperature and monitoring of the wearing of masks by passengers. This method has problems such as a lot of waste of human resources and low efficiency when the flow of people is large. The rapid development of computer vision technology allows us to use the intergration of camera and computer to detect whether the person wearing a mask is supplemented by infrared thermal imaging system to detect body temperature, which can achieve the purpose of non-contact automatic detection[1].

In the past few years, after deep learning has achieved widespread success in image classification, a large number of convolutional neural networks have been used for detection tasks, and these methods can be divided into two categories[2]: region-based methods and regression-based methods. Region-based methods mainly include R-CNN, SPP-Net[3], Fast R-CNN, Faster R-CNN[4] and R-FCN[5]. In order to slove the problem of the balance between detection speed and accuracy, a regression-based detection method is proposed. This type of methods can directly obtain the coordinate position and classification score of the detected object, mainly including YOLO, SSD, YOLOV2 and YOLOV3[6]. Among them, YOLO has serious

ISSN: 2664-9640

positioning errors, so the detection accuracy is not high. SSD[7] is based on the VGG network fusion of feature representastion ability of the system, while ensuring real-time, greatly improving the detection accuracy, but SSD does not consider the convolutional layer when fusing multiple convolutional features[8]. YOLOV2 uses a series of methods to optimize the model structure of YOLO[9], which significantly improves the detection speed, while the detection accuracy is the same as that of the SSD[10], but the basic network of YOLOV2 is relatively simple and dose not improve the detection accuracy[11]. YOLOV3 used deep residual network to extract image features, and realizes multi-sacle prediction, and obtains the best balance of detection accuracy and speed[12]. However, the minimum feature map size used by YOLOV3 to extract features is 13*13,which is relative to that in SSD. 1*1 is still too large[13], causing YOLOV3 to have a poor detection effect on some medium or large-sized objects, and will cause problems of false detection, missed detection or repeated detection[14].

Based on the YOLOV4 algorithm, this paper proposes an improved YOLOV4 mask wearing detection algorithm to achieve fast and accurate detection of mask wearing. In order to improve the detection rate, MobileNet is used to replace the backbone features network CSPdarknet53 of the YOLOV4 algorithm, and h-swish is used in the structure to replace the swish activation function, which reduces the amount of calculation and improves performance. In addition, a deep separable convolution is used instead of the ordinary convolution used in YOLOV4, and the mask data collected through imagenet is used as a dataset. Experiments show that compared with YOLOV4 algorithm, the improved algorithm in this paper improves the detection accuracy and detection speed, which is conductive to real-time video detection.

2. Classic YOLOV4 neural network

2.1. Basci features of YOLOV4

The YOLOV4 network is an upgraded version of YOLOV3. The backbone feature extraction network has changed from DarkNet53 to CSPDarkNet53, using the feature pyramid SPP and PAN. Tips for training include Mosaci data enhancement, Label Smoothing, CIOU, learning rate cosine annealing attenuation. The activation function is changed from Relu to using Mish activation function, etc. The Msih activation function is represented, see **Formula 1**.

$$\mathbf{Mish} = \mathbf{x} * \mathbf{tan\,h}(\mathbf{l\,n}(\mathbf{1} + \mathbf{e^x})) \tag{1}$$

IOU is the concept of ratio and is not sensitive to the scale of the target object[15]. However, the commonly used Bbox regression loss optimization and IOU optimization are not completely equivalent, and the ordinary IOU cannot directly optimize the parts that do not overlap. Therefore, the excellent idea of CIOU is proposed. CIOU tasks into account the distance between the target and the anchor frame, the overlap rate, the sacle and the penalty items, so that the return of the target frame becomes more stable, and there will be no training like IOU and GIOU. Problems such as divergence in the process, and the penalty factor tasks into account the aspect ratio of the predicted frame to fit the target frame. The calculation of CIOU, is determined as Formula 2.

$$\mathbf{COIU} = \mathbf{IOU} - \frac{\boldsymbol{\rho}^2(\mathbf{b},\mathbf{b}^{\mathbf{gt}})}{\mathbf{c}^2} - \boldsymbol{\alpha}\mathbf{v} \tag{2}$$

Among them: $\boldsymbol{\rho}^2(\mathbf{b},\mathbf{b}^{\mathbf{gt}})$respectively represents the Euclidean distance between the center points of the prediction box and the real frame[16], and **c** represents the diagonal distance of the smallest closure area that can contain both the prediction box and the real frame[17]. And the expression of $\boldsymbol{\alpha}$, is represented as **Formula 3.**

$$\boldsymbol{\alpha} = \frac{\mathbf{v}}{\mathbf{1-IOU+v}} \tag{3}$$

Also the expression of **v,** is shown as **Formula 4.**

$$\mathbf{v} = \frac{\mathbf{4}}{\boldsymbol{\pi}^2}(\mathbf{arctan}\frac{\mathbf{w^{gt}}}{\mathbf{h^{gt}}} - \mathbf{arctan}\frac{\mathbf{w}}{\mathbf{h}})^2 \tag{4}$$

ISSN: 2664-9640

The classic YOLOV4 neural network has many of the advantages, but it is difficult to maintain a high level of video detection speed.

2.2. YOLOV4's network structure

In this section, we shall elaborate the details of YOLOv4.

YOLOV4 consists of:

Backbone: CSPDarknet53

Neck: SPP, PAN

Head: YOLOV3

YOLOV4's network structure, see **Figure 1**.

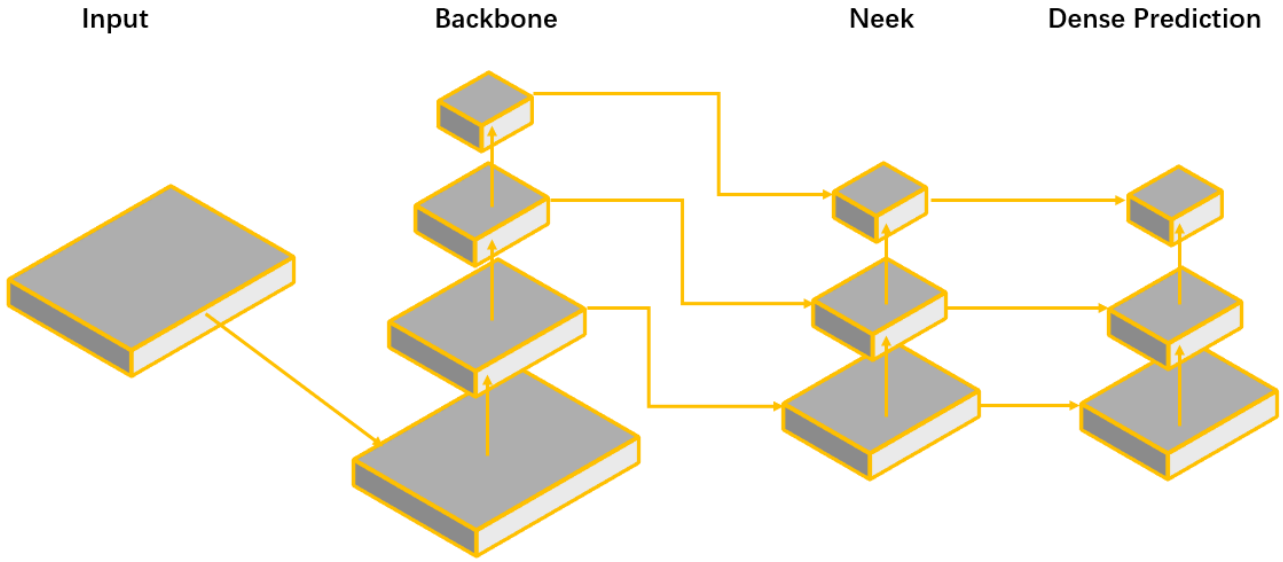

Figure 1: YOLOV4's network structure

3. Improved YOLOV4 mask detection model

3.1. The deep learning network of MobileNet-YOLOV4

The MobileNetV1 is a lightweight deep neural network proposed by Google for embedded devices such as mobile phones. The core idea used is depthwise separable convolution[18].

The MobileNetV2 is an upgraded version of MobileNet. It has a very important feature that uses Inverted resblock[19]. The entire MobileNetV2 is composed of Inverted resblock.

The MobileNetV3 uses a special bneck structure. It combines the following four characteristics. The first is MobileNetV2's inverse residual structure with a linear bottleneck. The second is depth separable convolution of MobileNetV1. The third is a lightweight attention model. Finally, using h-swish instead of swish function[20].

The network structure of the entire MobileNetV3, see Table 1.

Table 1: Network structure of the entire MobileNetV3

Input	Operator	exp size	out	SE	NL	s
$224^2 \times 3$	conv2d	-	16	-	HS	2
$112^2 \times 16$	bneck,3 × 3	16	16	-	RE	1
$112^2 \times 16$	Bneck,3 × 3	64	24	-	RE	2
$56^2 \times 24$	bneck,3 × 3	72	24	-	RE	1
$56^2 \times 24$	bneck,5× 5	72	40	✓	RE	2

ISSN: 2664-9640

$28^2 \times 40$	bneck,5× 5	120	40	✓	RE	1
$28^2 \times 40$	bneck,5× 5	120	40	✓	RE	1
$28^2 \times 40$	bneck,3 × 3	240	80	-	HS	2
$14^2 \times 80$	bneck,3 × 3	200	80	-	HS	1
$14^2 \times 80$	bneck,3 × 3	184	80	-	HS	1
$14^2 \times 80$	bneck,3 × 3	184	80	-	HS	1
$14^2 \times 80$	bneck,3 × 3	480	112	✓	HS	1
$14^2 \times 112$	bneck,3 × 3	672	112	✓	HS	1
$14^2 \times 112$	bneck,5× 5	672	160	✓	HS	2
$7^2 \times 160$	bneck,5× 5	960	160	✓	HS	1
$7^2 \times 160$	bneck,5× 5	960	160	✓	HS	1
$7^2 \times 160$	conv2d,1× 1	-	960	-	HS	1
$7^2 \times 160$	pool,7× 7	-	-	-	-	1
$1^2 \times 960$	conv2d,1× 1,NBN	-	1280	-	HS	1
$1^2 \times 1280$	conv2d,1× 1,NBN	-	k	-	-	1

In this table, the first column Input represents the shape change of each feature layer of MobileNetV3. The second column of Operator represents the block structure that the feature layer will experience each time. We can see that in MobileNetV3, feature extraction has gone through many block structures. The third and fourth columns respectively represent the number of channels in the block inner inverse residual structure after rising, and the number of channels in the characteristic layer when input to block.

Figure 2 shows the special block structure of MobileNetV3.

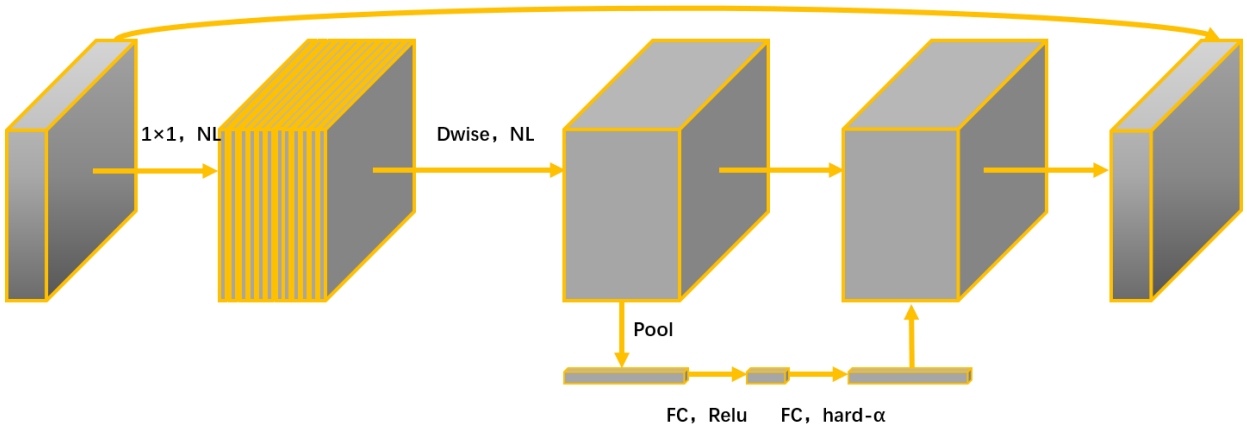

Figure 2: The special block structure of MobileNetV3

It can be seen that MobileNetV3 combines the ideas of the following three models: MobileNetV1's depthwise separable convolutions, MobileNetV2's inverted residual with linear bottleneck and MnasNet based Lightweight attention model of squeeze and excitation structure. The advantages of the above three structures are combined to design an efficient MobileNetV3 module.

The original h-swish activation function, see Formula 5.

$$\mathbf{swish}\ \mathbf{x} = \mathbf{x} \cdot \boldsymbol{\sigma}(\mathbf{x}) \tag{5}$$

However, swish is too computationally expensive, so a new calulation method is adopted, see Formula 6.

ISSN: 2664-9640

$$h-swish[x]=x\frac{ReLu6(x+3)}{6} \quad (6)$$

The nonlinearity brings many advantages while maintaining accuracy. First, Relu can be implemented in many software and hardware frameworks, and secondly, it avoids the loss of numerical accuracy during quantization and runs fast. This nonlinear change increases the delay of the model by 15%. But the network effect it brings has a positive boost to accuracy and delay, and the remaining overhead can be eliminated by fusing nonlinearity with the previous layer.

For YOLOV4, we need to use the three effective features obtained by the backbone feature extraction network to bulid a strengthened feature pyramid. Using MobileNetV3, we can obtain three effective feature layers corresponding to each network. These three effective feature layers can be used to replace the effective feature layers of the original yolov4 backbone network CSPDarknet53. This completes the construction of our MobileNet-YOLOV4 model.

3.2. Train and test on the mask dataset

The ImageNet dataset has downloaded 2707 photos about recognizing masks. These photos have been marked by labelimg software. Colored rectangles can be seen on the marked photos. The face of the person is marked by colored rectangles. The mask is framed and there is no extra space left. The corresponding parameters can be obtained by using the marked photos, such as the center coordinates of the mask frame, and the width and height of the frame.

Train and test under the three prediction models of SSD, YOLOV4 and MobileNetV3-YOLOV4. Comparison of their respective performance includes detection speed and detection accuracy, and analysis of the detection effect of real-time video.

First, we divide the 2707 marked photos into the training set and the test set by 9:1. Then change the categories we want to divide into two categories, wearing a mask and not wearing a mask. Using the VOC2007 pre-training weight as our initial weight, and then perform one hundred rounds of training, and finally get the weight after training. Its loss is very small, only 0.93. We can use the trained weight as a prediction weight.

The threshold value needs to be set in the prediction stage, here is set to 0.5, if the threshold is exceeded, it is considered as a positive sample mask, otherwise it is a negative sample without a mask.

Parameter settings required in the experiment, see Table 2.

Parameter	Value
Learning rate	0.001
epoch	100
Batch size	4
momentum	0.9
Weight_dacay	0.005
Learning rate stsp	1000
Learning rate factor	0.1
NMS	0.3

The experiment environment is AMD Ryzen 7 5800H with Radeon Graphics @ 3.20Ghz, 16G running memory, Nvidia GeForce RTX 3060 Laptop GPU, windows 10, 64-bit operating system, pytorch deep learning framework.

Perform training on the dataset according to the parameter settings in the above table, and after obtaining the corresponding optimal weights. The new weights are used to evaluate the test set. In order to test the effectiveness and real-time performance of the improved YOLOV4 algorithm

ISSN: 2664-9640

for face mask wearing detection. Average accuracy, Recall and mAP are used to test the effectiveness of the improved algorithm.

Predict the photos of the test set, and summarize the Average accuracy, Recall, and mAP under the same equipment and the same prediction conditions. The evaluation of the positive sample prediction is averaged after multiple experiments, see Table 3.

Detection model	Average accuracy	Recall
SSD	97.5%	97.56%
YOLOV3	97.11%	96.86%
YOLOV4	97.15%	96.86%
MobileNetV3-YOLOV4	97.12%	95.82%

The evaluation of the negative sample prediction is averaged after multiple experiments, see Table 4.

Detection model	Average accuracy	Recall
SSD	90.76%	85.00%
YOLOV3	90.16%	91.00%
YOLOV4	88.72%	89.00%
MobileNetV3-YOLOV4	89.66%	83.00%

The mAP prediction, see Table 5.

Detection model	mAP
SSD	94.13%
YOLOV3	93.63%
YOLOV4	92.935%
MobileNetV3-YOLOV4	93.39%

Through the analysis of the prediction results, it can be seen that the improved alogrithm MobileNet-YOLOV4 on the basis of YOLOV4 compared to YOLOV4 algorithm has improved the AP by 0.94% for negative samples. In addition, in the detection of positive samples, the two alogrithm are basically the same. And for the overall sample, the improved algorithm is 0.455% higher than the original algorithm's mAP, which is better than the original algorithm.

3.3. Real-time video detection and analysis

First, a pedestrian corssing video which was downloaded was taken by a street camera. The import of this video can be regarded as real-time mask detection and recognition, and the trained model is used to predict the video.The blue box represents the detection of a face, the red box represents the detection of a face with a mask, and the upper right corner of the box has a probability value.This probability value represents the probability of detecting a face and a face wearing a mask.

FPS and mAP are two important evaluation indicatiors of target detection algorithms. FPS is used to evaluate the speed of target detection, that is, the number of pictures that can be processed per second or the time required to process a picture to evaluate the detection speed. The shorter the time, The faster the speed. FPS is how many frames the target network can process per second. FPS is simply understood as the refresh rate of the image, that is, how many frames per second. Assuming that the target detection network processes 1 frame, it takes 0.02s. At this time, FPS is 50.

ISSN: 2664-9640

In real-time video detection of targets, the more targets appear in a frame, the greater the amount of calculation required by the detection algorithm. For example, when there are only two or three goals at a certain moment, the fps can reach 60, but at another moment, there are a lot of goals, and the fps is only less than half. Therefore, for the accuracy and general practicability of the experiment, the prediction was selected during the crowded time period, including the detction of two kinds of objects with and without masks.

For YOLOV4 and MobileNetV3-YOLOV4, the detection results of real-time video are as follows.

YOLOV4 real-time video detection effect, see Figure 3.

Figure 3: YOLOV4 real-time video detection effect

MobileNetV3-YOLOV4 real-time video detection effect, see **Figure 4**.

Figure 4: MobileNetV3-YOLOV4 real-time video detection effect

Through comparison, we can know that in place with dense human figures, that is, when there are many targets in a frame of image, the real-time detection effect of MoboleNetV3-YOLOV4 is better than YOLOV4, YOLOV4 is only 21, and the improved YOLOV4 reaches 32, directly

ISSN: 2664-9640

improved performance by 52%, which means that more photos can be detected in the same time and more information can be obtained.

4. Conclusion

This paper proposes an improved algorithm MobileNetV3-YOLOV4 for mask detection based on the YOLOV4 algorithm. The improved algorithm mainly uses MobileNetV3's lightweight convolutional neural network to replace YOLOV4's backbone feature network CSPDarknet53 to reduce model parameters. In the experiment, the mask dataset on ImageNet was used to train and test the model. Under the same conditions, the SSD, YOLOV4, MobileNetV3-YOLOV4 were trained and tested. The improverd YOLOV4 average detection accuracy is basically the same as YOLOV4. The speed is better than the YOLOV4 algorithm. In real-time video detection, the FPS has been increased from 21 to 32, achieving a very satisfactory detection effect.

References

[1] Guan Junlin, Zhixin: A method of mask wearing detection based on YOLOV4 convolutional neuarl network, Modern Information Technology, Vol. 4(2020) No.11, p.683-685.

[2] Zhang Fukai, Yang Feng, Li Ce: Fast vehicle detection method based on improved YOLOV3, Computer Engineering and Applications, Vol. 55(2019) No.02, p.12-20.

[3] K.He, X.Zhang, S.Ren, J.Sun: Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans Pattern Anal Mach.Intell, Vol. 37(2015) No.09, p.1904-1916.

[4] S.Ren, K.He, R.Girshick: Faster R-CNN:towards real-time object detection with region proposal networks, IEEE Trans Anal Mach Intell, Vol. 39(2017) No.06, p.1137-1149.

[5] Dai J, Li Y, He K, Sun J: R-Fun:Object detection via region-based fully convolutional networks, Proceedings of the Advances in Neural information processing systems, Vol. 1(2016) No.01, p.379-387.

[6] Liu Y, Cao S, Lasang P, Modular lightweight network for road object detection using a feature fusinon approach, IEEE Trans Syst Man Cybern, Vol. 51(2021) No.08, p.4716-4728.

[7] Wei Wei, Pu Wei, Liu Yi: Application of improved YOLOV3 in Aerial Target Detection, Computer Engineering and Applications, Vol. 56(2020) No.07, p.17-23.

[8] S.Ren, K.He, R.B.Girshick, J.Sun: Faster R-cnn: Towards real-time object detection with region proposal networks, IEEE Trans Pattern Anal Mach Intell, Vol. 39(2015) No.11, p.1137-1149.

[9] K.He, X.Zhang, S.Ren, J.Sun: Deep residual learning for image recognition, IEEE conference on Computer Vision and Pattern Recognition, (CVPR) (2016) , p.770-778.

[10] C.Pan, W.Q.Yan: Object detection based on saturation of visual perception Multimed, Tool.Appl, Vol. 79(2020) No.27, p.485-533.

[11] Felzenszwalb PF, Girshick RB, Mcallester D: Object Detection with Discriminatively Trained Part-Based Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32(2010) No.09, p.1627-1645.

[12] Asman A.J, Landman B.A: Non-loacl statistical label fusion for multi-atlas segmentation, Med Image Anal, Vol. 17(2013) No.02, p.194-208.

[13] Dalal N, Triggs B: Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 01(2020) No.27, p.886-893.

[14] Farbman Z, Fattal R, Lischinski D: Convolution pyramids, ACM Trans Graph, Vol. 30(2011) No.06, p.175.

[15] Gao L, Chen P, Yu S: Demonstration of convolution kernel operation on resistive cross-point array, IEEE Electron Device Lett, Vol. 37(2016) No.07, p.870-873.

[16] Hahm N, Hong B.I: An approximation by neural networks with a fixed weight, Comput Math Appl, Vol. 47(2004) No.47, p.1897-1903.

ISSN: 2664-9640

[17] Mohammad A, Masouros C, Andreopoulos Y: Complexity-scalable neural-network-based MIMO detection with learnable weight scaling, IEEE Trans Commun, Vol. 68(2020) No.10, p.6101-6113.

[18] Wang Y, Wang C, Zhang H, Dong Y, Wei S: Automatic ship detection based on retinanet using multi-resolution gaofen-3 imagery, Remote Sens, Vol. 11(2019) No.05, p.531.

[19] Ahmadi M, Sharifi A, Khalili S: presentation of a developed sub-epidemic model for estimation of the COVID-19 pandemic and assessment of travel-related risks in Iran, Environ. Sci. Pollut. Res. Int, Vol. 28(2021) No.12, p.14521-14529.

[20] Ahmadi M, Jafarzadeh-Ghoushchi S, Taghizadeh R, Sharifi A: Presentation of a new hybrid approach for forecasting economic growth using artificial intelligence approaches, Neural Comput. Appl, Vol. 31(2019) No.12, p.8661-8680.