Investigation of optimal configurations of a convolutional neural network for the identification of objects in real-time

M A Isayev[1], D A Savelyev[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

e-mail : michailisaev.home@gmail.com

Abstract. The comparison of different convolutional neural networks which are the core of the most actual solutions in the computer vision area is considers in hhe paper. The study includes benchmarks of this state-of-the-art solutions by some criteria, such as mAP (mean average precision), FPS (frames per seconds), for the possibility of real-time usability. It is concluded on the best convolutional neural network model and deep learning methods that were used at particular solution.

1. Introduction

At present, the field of computer vision is actively developing, especially at the time of emergence of convolutional neural networks (CNN) [1, 2] and unmanned devices [3]. Another integral part of computer vision field is the detection of objects and classifying based on special object features [4]. Object detection is successfully used in vehicle tracking, positioning, surveillance [5-7]. The difference between the classification and detection algorithms is that the detection algorithms contain the boundaries of the region of interest (object) and are defined in the image. It is also worth noting the big difference between the concepts of classification and clustering - the main difference between classification and clustering is that when solving the classification problem, groups of objects are already known, while clusters are already determined at the moment of solving the clustering problem. Obviously, for object detection tasks, you should not use a regular neural network with a fully connected layer at the end. This is due to the fact that, as a rule, the length of the output layer is dynamic, which is associated with a non-fixed number of appearing objects.

One approach to solve this problem is to obtain different areas from an image (Region of Interests) and use convolutional neural networks to determine the presence of an object within this area. This solution does not take into account the possibility of a different location of the object and different proportions of the side. Consequently, it will be necessary to process a huge number of such areas, which is so expensive in terms of computational power. Another solution is special algorithms that were developed for the problem of detecting objects in real time. [8].

Solutions in the field of image recognition in real time are divided into two main families: Region Proposes (the frame regions are alternately proposed and classified) and Single Shot (all objects are immediately detected on the resulting image). The first family includes neural networks such as R-

CNN, Fast R-CNN, Faster R-CNN [9-11]. The second family includes YOLO, SSD [5, 12]. Neural networks that use recognition by region have a rather slow recognition time in the qualitative determination of objects.

In this paper, we study the determination of the optimal solution for the problem of detecting objects in real time based on testing solutions of R-CNN, R-FCN, SSD (VGG-16), YOLOv3 (Darknet-53), based on metrics such as mAP and FPS.

2. A convolutional neural network used to identify objects

The following parameters were used for the study: dataset - PASCAL VOC 2012 [8]. Faster R-CNN, R-FCN, SSD (VGG-16), YOLOv3 (Darknet-53) solutions were tested. Faster R-CNN uses RPN (Region Proposal Network) instead of the slow selective search algorithm. RPN is a complete replacement for the selective algorithm, and works as follows: at the last level of the original CNN, a 3x3 sliding window bypasses the feature map and reduces its dimension and for each sliding window position, RPN generates many possible areas based on the k boundaries of a possible object. R-FCN, or Region-based Fully Convolutional Net is a fully connected network and raises one of the main problems in the design of neural networks. On the one hand, when performing a classification of an object, it is necessary to train the model on the property of the invariance of the object's location: despite where the object appears in the image, the object must be uniquely determined. On the other hand, it is necessary that the trained model selects the boundaries of the object in the place of the image where it appears (local variation). The compromise between variance and location invariance is to use positional scorecards. The input image is processed by CNN, adding a fully connected layer to create a storage of position-sensitive rating maps that RoI generates. Next, for each region, the assessment storage is checked for the fact whether this region is the corresponding position of some object.

SSD, in contrast to Faster R-CNN, which uses algorithms for the regional classification and generation of prediction domains, simultaneously determines the frame of the object, as well as its class at the time of image processing. The SSD sends the image for processing through a series of convolutional layers, receiving several sets of feature maps, for each position in each of these feature maps, a 3x3 convolutional filter is used to obtain a set of reference coordinates of the image boundaries, where for each set of coordinates, the offset and the probability of being within the boundaries of these object coordinates.

YOLOv3, like SSD, belongs to the Single Shot family, and also uses softmax with independent logistic classifiers to calculate the similarity of the input data with a particular class. Instead of using MSE (mean squared error) to compute a classification error, YOLOv3 uses binary cross-entropy for each class label. To determine the coordinates of the boundaries of the object, YOLOv3 uses the k-means clustering algorithm.

3. An anthropometric model-based method for extracting facial specified features to improve classification

Detection of objects in images and videos using neural networks, including the identification of [13] real-time faces [14] is an important task today. In particular, V.S. Gorbatsevich and al. propose original iterative proposal clustering (IPC) algorithm for aggregation of output face proposals formed by CNN and the 2-level “weak pyramid” providing better detection quality on the testing sets containing both small and huge images [15].

To extract important features from person`s face at the first stage, it is necessary to mark the key points on the person’s face, which will determine the relative position of the main elements of the person’s face and take the necessary measurements, which in the next stage are fed to the classifier. figure 1 shows the 68 face landmarks on the left and the main anthropometric parameters on the right.
The key points of a person’s face are used to mark the image of protruding regions of a person’s face, such as: eyes, eyebrows, nose, mouth, jaw line. At the moment, they are actively used in applications for aligning faces in an image, presenting a model of a pose of a human head, detecting flicker, etc.

The detection of such points is a subtask of determining the shape and shape of the input object. Obtaining an image as an input, the predictor tries to localize the key points, taking into account the shape and shape of the object [16].

At the stage when the necessary key points are known and their coordinates relative to the whole image, the necessary proportions are calculated based on the anthropometric model [17]. Facial measurements are presented at the figure 2.

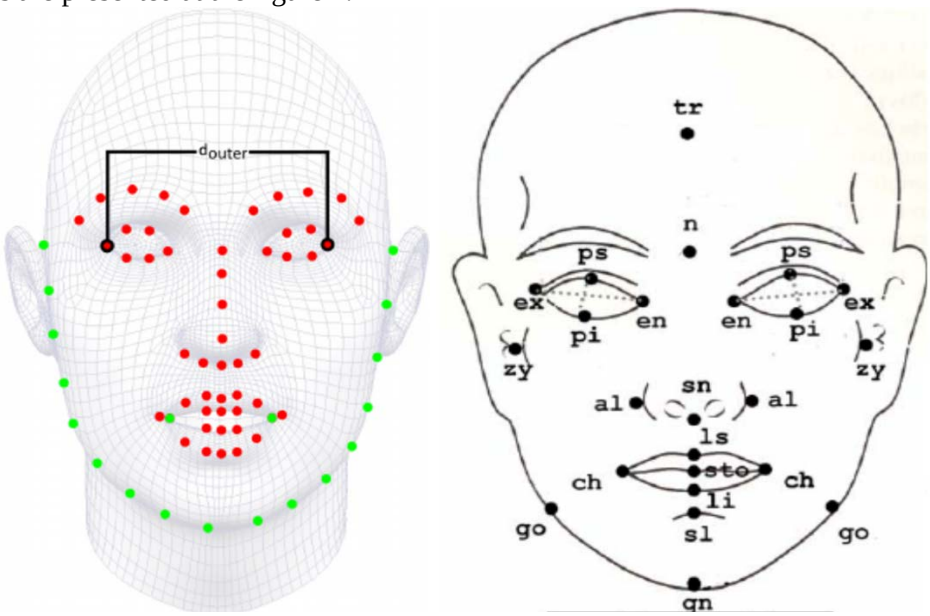

Figure 1. The important facial measurements as a main feature for signs extraction.

$$FI = \frac{n-gn}{zy-zy};\ MI = \frac{sto-gn}{go-go};\ II = \frac{en-en}{ex-ex};\ OWI = \frac{ex-en}{en-en};$$
$$EFI = \frac{ps-pi}{ex-en};\ NI = \frac{al-al}{n-sn};\ VHI = \frac{ls-sto}{sto-li};\ MFWI = \frac{ch-ch}{zy-zy}.$$

Figure 2. Facial indexes for anthropometric model.

4. The method based on hybrid Hesse filter for extracting facial specified features to improve classification

The Hesse filter is a tool for detecting wrinkles in the input image. After all, it is no secret that the appearance of wrinkles on a person's face is the most expected changes in a person's face and skin as his age increases.

The algorithm is based on the use of the Hesse matrix and directional gradient [17]. The Hessian hybrid filter detects wrinkles by computing the Hessian matrix for each pixel of the input image. The maximum eigenvalues of the Hessian matrix indicate that a particular pixel of the image belongs to the outline of the wrinkle, regardless of its position. The eigenvalues in this context are independent vector measurements for the components of the second derivatives at each point. A small value of the eigenvalue indicates a weak rate of change of the surface of the face in the corresponding direction of the vector of eigenvalues, and vice versa.

First of all, a two-dimensional image is converted into an image with only one shade. We denote the image as I (x, y). The gradient (Gx, Gy) is calculated from the single-channel image as: ΔI (x, y) = (∂I / ∂x, ∂I / ∂y), where ∂I / ∂x, ∂I / ∂y are directional gradients, that is, ∂I / ∂x = Gx, and ∂I / ∂y = Gy. The directional gradient significantly smoothes the image, but retains the data that are of interest in the original task. Gy is used as the input image for the Hessian hybrid filter, that is, as input to calculate the Hessian matrix H, which is used to extract horizontal lines. To modify the algorithm for detecting vertical wrinkles, Gx is used as input to the modified Hessian hybrid filter.

In this case, Hessian matrix determines as follows:

$$H(x,y,\sigma)=\begin{bmatrix}\frac{\partial^2 I(x,y)}{\partial I(y)\partial I(y)} & \frac{\partial^2 I(x,y)}{\partial I(x)\partial I(y)}\\ \frac{\partial^2 I(x,y)}{\partial I(x)\partial I(y)} & \frac{\partial^2 I(x,y)}{\partial I(x)\partial I(x)}\end{bmatrix}=\begin{bmatrix}H_a & H_b\\ H_b & H_c\end{bmatrix}. \quad (1)$$

Figure 3 shows the result of the work of the Hesse hybrid filter. This method is a reliable tool for extracting age characteristics from the input image, but with all the reliability of this method, it should be understood that wrinkles may appear at all at different times, depending on race, lifestyle and genes [17].

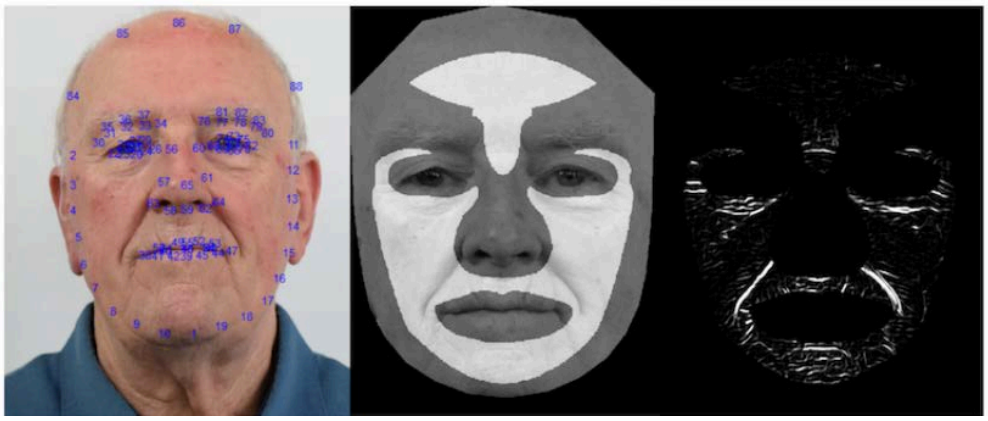

Figure 3. Visualization of the Hesse hybrid filter (HHF) work.

At the stage when arrays of features are received, these age signs are fed to the input of the classifier, for its subsequent training. In the framework of this work, classifiers were used based on the random forest algorithm.

Random forest is an ensemble decision tree algorithm because the final prediction, in the case of a regression problem, is an average of the predictions of each individual decision tree, in classification; it's the average of the most frequent prediction [18]. So, the algorithm takes the average of many decision trees to arrive at a final prediction, as shown on figure 4.

As part of this work, the following architectures of convolutional neural networks were used, which were trained to solve the problem to the identification of objects: InceptionV3, ResNet50.

It is quite obvious that convolutional neural network models, their configuration and training is very resource-intensive, even on modern computers. In view of this, many researchers have proposed and developed modifications of convolutional neural networks (such as residual networks and inception blocks), the task of which was to simplify the initial problem with resources [19].

Comparison of the results of age estimation task by the considered methods showed: accuracy of RF + HHF method is 81.1%, accuracy of Inception v3 is 80.8%, and accuracy of ResNet50 is 85.7%.

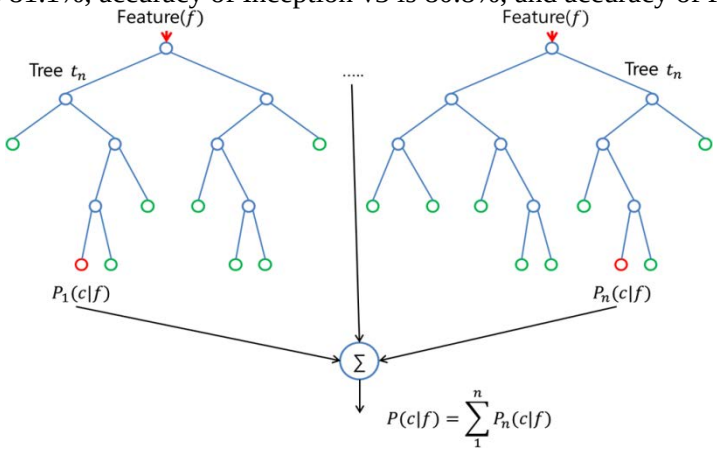

Figure 4. Visualization of the Random forest (RF) method.

The generalized result of the objects identification research is given in table 1.

Table 1. Comparison of the results of objects identification by the considered neural networks in p. 2.

Solution	Dataset	mAP, %
R-FCN	COCO + VOC 12	59.8
Faster R-CNN	COCO + VOC 12	60.15
SSD	PASCAL VOC 12	64.00
YOLOv3	COCO	63.35

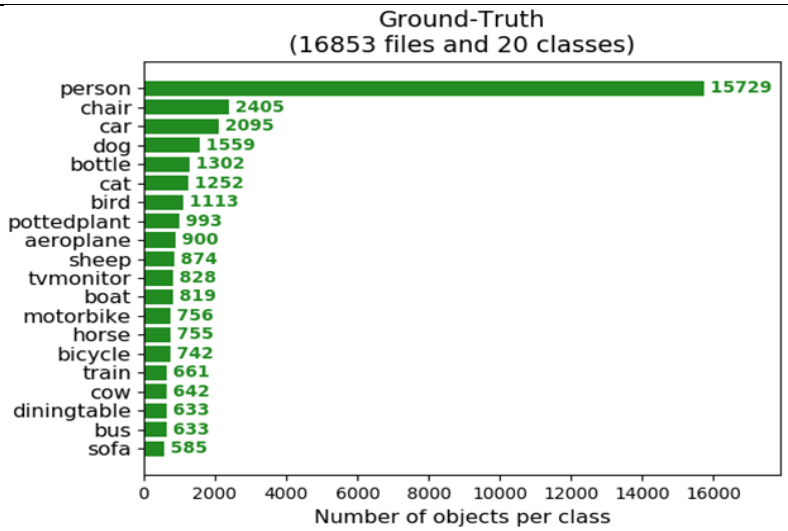

Figure 5. The number of ground-truth labels per class.

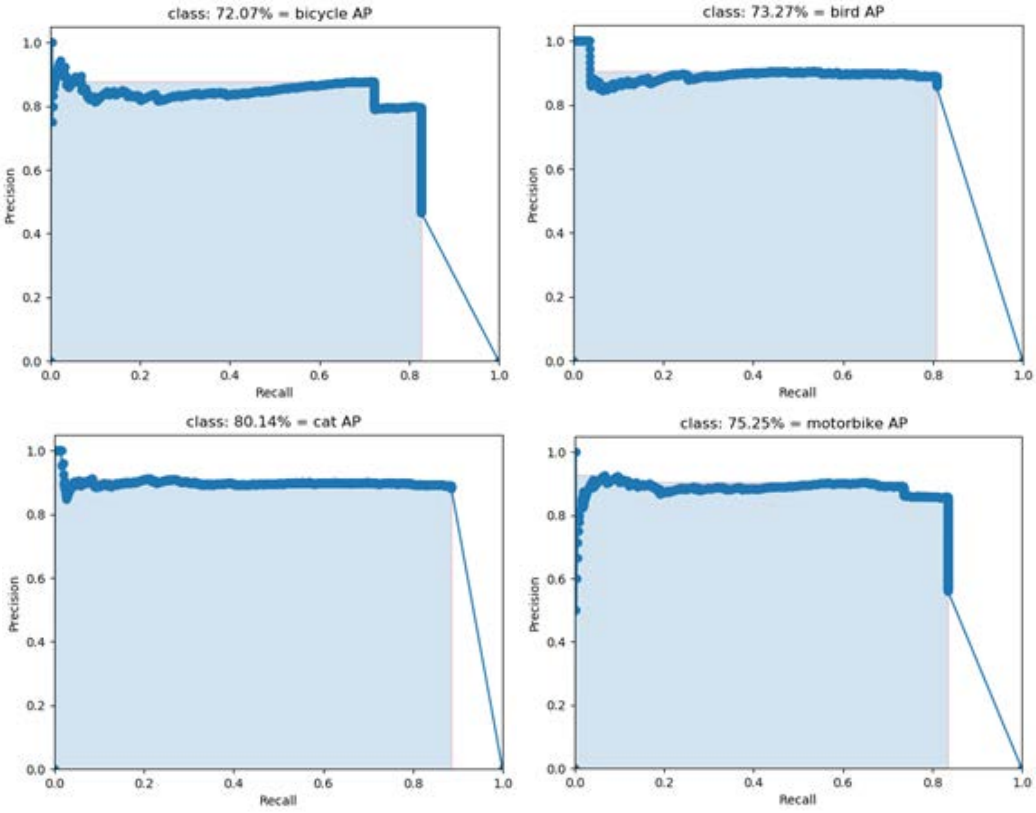

Figure 6. The results of AP estimations for some class of objects.

Figure 5 shows the number of labeled data by class, and figure 6 shows the results of the average detection accuracy for some classes of objects when using the YOLOv3 solution.

5. Conclusion

We compared various convolutional neural networks in this paper. The study includes comparing data on computer vision solutions for such criteria as mAP, FPS, i.e. the possibility of using them in real time. Based on the study, it was shown that the most suitable solution for real-time object identification task is YOLOv3. Despite not the highest mAP rating, YOLOv3 has a high processing speed of the video stream. Therefore, YOLOv3 has great prospects as a tool for tracking and detecting objects in a video stream.

As a result of the study, it is shown that convolutional neural networks successfully cope with the task of automatically determining a person's biological age on his face: : accuracy of RF + HHF method is 81.1%, accuracy of Inception v3 is 80.8%, and accuracy of ResNet50 is 85.7%.

6. References

[1] Shi, W, Caballero J, Huszar F, Totz J, Aitken A P, Bishop R, Rueckert D and Wang Z 2016 Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1874-1883

[2] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision *Proceedings of the IEEE conference on computer vision and pattern recognition* 2818-2826

[3] Pestana J, Sanchez-Lopez J L, Saripalli S and Campoy P 2014 Computer vision based general object following for gps-denied multirotor unmanned vehicles *American Control Conference (ACC)* 1886-1891

[4] Magdeev, R, Tashlinskii A G 2019 Efficiency of object identification for binary images *Computer Optics* **43(2)** 277-281 DOI: 10.18287/2412-6179-2019-43-2-277-281

[5] Redmon J, Divvala S, Girshick R and Farhadi A 2016 You only look once: Unified, real-time object detection *Proceedings of the IEEE conference on computer vision and pattern recognition* 779-788

[6] Protsenko V I, Kazanskiy N L and Serafimovich P G 2015 Real-time analysis of parameters of multiple object detection systems *Computer Optics* **39(4)** 582-591 DOI: 10.18287/0134-2452-2015-39-4-582-591

[7] Kazanskiy N L, Protsenko V I and Serafimovich P G 2017 Perfomance analisys of real-time face detection system baced on stream data mining frameworks *Procedia Engineering* **201** 806-816 DOI: 10.1016/j.proeng.2017.09.602

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

[9] Dai J, Li Y, He K, Sun J 2016 R-fcn: Object detection via region-based fully convolutional networks *Advances in neural information processing systems* 379-387

[10] Girshick R 2015 Fast r-cnn *Proceedings of the IEEE international conference on computer vision* 1440-1448

[11] Ren S, He K, Girshick R and Sun J 2015 Faster r-cnn: Towards real-time object detection with region proposal networks *Advances in neural information processing systems* 91-99

[12] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C 2016 Ssd: Single shot multibox detector *European conference on computer vision* (Springer, Cham) 21-37

[13] Nemirovskiy V B, Stoyanov A K 2017 Clustering face images *Computer Optics* **41(1)** 59-66 DOI: 10.18287/2412-6179-2017-41-1-59-66

[14] Vizilter Yu V, Gorbatsevich V S, Vorotnikov A V and Kostromov N A 2017 Real-time face identification via CNN and boosted hashing forest *Computer Optics* **41(2)** 254-265 DOI: 10.18287/2412-6179-2017-41-2-254-265

[15] Gorbatsevich V S, Moiseenko A S and Vizilter Y V 2019 FaceDetectNet: Face detection via

fully-convolutional network *Computer Optics* **43(1)** 63-71 DOI: 10.18287/2412-6179-2019-43-1-63-71

[16] Krizhevsky A, Hinton G andSutskever I 2012 ImageNet Classification with Deep Convolutional Neural Networks *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems* **1(1)** 1097-1105

[17] Karthikeyan D, Balakrishnan G 2018 A comprehensive age estimation on face images using hybrid filter-based feature extraction *Biomedical Research Medical Diagnosis and Study of Biomedical Imaging Systems and Applications* 472-480

[18] Bosch A, Zisserman A and Munoz X 2007 Image Classification using Random Forests and Ferns *11th International Conference on Computer Vision* 1-8

[19] Szegedy C, Ioffe S, Vanhoucke V and Alemi A 2017 Inception - v4, Inception-ResNet and the Impact of Residual Connections on Learning *Thirty-First AAAI Conference on Artificial Intelligence* 4278-4284

Acknowledgments

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation.