Image-based Technology for Creating a Catalog of E-commerce Goods Using Neural Networks and Attention Model

V. Sorokina, S. Ablameyko

***Abstract*—**We propose a technology for automatic creation of a catalog of e-commerce products based on neural networks. An e-commerce product catalog is a set of processed product photos. The technology includes parts for determining the product itself (object detection task) and its main color, segmentation and automatic cropping (human body parts detection task). The novelty of the presented work lies in the use of convolutional neural network in conjunction with an attention model that is able to highlight significant and filter out insignificant information about objects in an input image. The obtained results demonstrate that the use of the attention model has a positive effect on both the quality of the trained network and the rate of convergence, which is reflected in improved metrics for object recognition and segmentation. The technology is integrated into the system developed by us which aim is to cut the images and prepare e-commerce catalogue.

***Keywords*—**Attention model, Convolutional neural network (CNN), Segmentation, YOLACT.

I. INTRODUCTION

In e-commerce, one of the key components of success is the product catalog. A product catalog is a set of images representing a product. Creating a catalog is the process of preparing photographs from the moment of shooting to the presentation of the processed set. Online store sales are highly dependent on product images. It's important to use the right ecommerce images to drive traffic, answer questions visually, and turn site visitors into buyers. The above mentioned is achieved by conforming the image to certain e-commerce standards such as image size (minimum 500x500 pixels), background (predominantly white), object position in the photo, etc.

With the rapid growth of e-commerce and the advent of artificial intelligence algorithms, traditional content management systems are giving way to automated, scalable systems.

For instance, when preparing images of clothing a full-length photograph of a person representing several items of clothing at the same time is usually used. The task is to determine the position of the object, to make a segmentation, to crop the image and create a color swatch - a sample of the color of the product. Currently, the process of preparing images for catalogs is done manually.

To automate this process, we have developed a technology that allows by solving the problem of object detection, segmentation, selection of human body parts, as well as determining the dominant color of the product, to automatically prepare an

V. Sorokina, Belarusian State University, Minsk, Belarus (e-mail: Viktoria.sorokina.96@gmail.com).
S. Ablameyko, Belarusian State University, Minsk, Belarus; United Institute of Informatics Problems of the NASB, Minsk, Belarus (e-mail: ablameyko@bsu.by).

e-commerce product catalog.

II. Literature Review

It is well known that images play an important role in e-commerce. Currently, neural networks for e-commerce products are mainly used for solving the classification, product recognition and segmentation tasks.

In this paper, e-commerce product recognition is considered as a specific research question related to the task of object recognition. At present, computer vision methods have already become widespread in the problem of object recognition, but their application for e-commerce product recognition is not yet so perfect. The task of e-commerce product recognition is more complex than standard object detection because some specific situations (pre-trained models couldn't be used due to the difference of the trained images; need to collect specific dataset with e-commerce products; there are crossed categories that impact on the quality of trained model, etc.) must be taken into account. However, the same methods are used to solve it.

The research [1] is devoted to building a smart system for selecting optimal product images in e-commerce. There were used 2 types of algorithms: based on machine learning methods, and shallow networks in combination with ResNet neural network as a base. The disadvantage of this approach is the high sensitivity of the model to the training set.

In paper [2], an algorithm using k-means clustering method with calculating the distance between image classes was proposed. The constructed model showed a high classification accuracy, however, preliminary image processing is required. This processing involves noise removal, color alignment and segmentation of the objects themselves.

Lucas Bossard et al. [3] proposed clothing classification and developed a dataset of over 80,000 images for clothing classification using random forest, transductive support vector machine (TSVM) and transfer forest algorithms.

In paper [4], an improved recognition model was developed using regional convolutional neural networks (R-CNN). The model was trained on top of the AlexNet convolutional network and used the weights of the pretrained ImageNet network. The disadvantages of the model are the complexity of implementation, the inability to achieve real-time speed, and the use of non-standard layers.

One of the most used networks in last object recognition and segmentation tasks is YOLACT [5]. The key point of YOLACT is speed - it is the fastest method of object recognition, as well as real-time segmentation of instances (when it was introduced). However, due to the fact that YOLACT is a fully convolutional neural network, training is slow and difficult, since each individual stage must be trained separately. To solve this problem, we propose the use of an attention model.

To solve the problem of human body parts detection we used well-known OpenPose model that we modified with attention model as well.

III. Research Method

A. E-commerce image recognition technology

In this paper a technology for automatically preparing images of e-commerce products and creating a catalog is proposed. The technology scheme is shown in Figure 1.

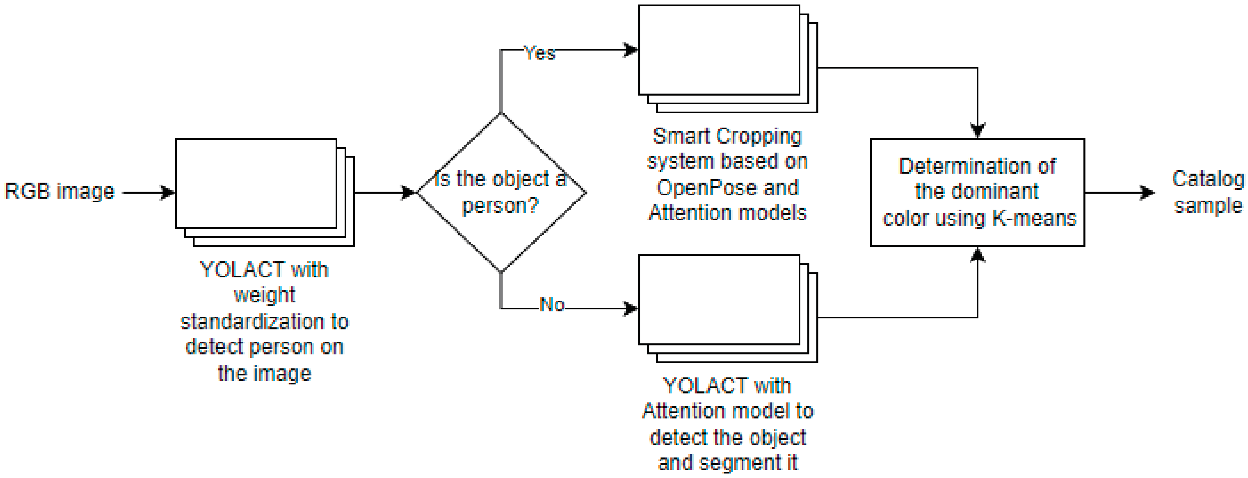
RGB image
YOLACT with weight standardization to detect person on the image
Is the object a person?
Yes
No
Smart Cropping system based on OpenPose and Attention models
YOLACT with Attention model to detect the object and segment it
Determination of the dominant color using K-means
Catalog sample

Fig.1 Scheme of the technology

The system developed on the basis of this technology is a program with a web interface implemented using the Flask framework. The input for the program is an RGB image of a product/group of products. Next, the image sequentially passes through the algorithms described below.

At the beginning, the type of object in the image is determined. For this purpose, we implemented the algorithm based on the YOLACT neural network [5] and weights standardization, a detailed description of which is presented in [6].

The aim of the previous step is to obtain a class label "human" / "non-human". If this is a person, then the Smart Cropping algorithm [7] is triggered, where the key points of the human body are defined, the positional relationship between them is calculated, after which it is used to crop the original image and create a set of images representing the goods. This algorithm is capable of preparing images of different clothing, hats and shoes. In this article, the OpenPose architecture [8] modified by the attention model [9] is used to determine the parts of the human body. The constructed model allows not only to structure parts of the human body due to the similarity fields of parts, but also to define parts of the human body in more detail due to stimuli to enhance significant and suppress non-significant objects in the image, which is achieved by constructing a two-dimensional evaluation matrix for each heat map.

If the type of object is not a person, then the classification and detection of the position of this object take place. These tasks are implemented using the YOLACT convolutional neural network, as well as the attention model. The next step is the segmentation that is needed to perform background removal. The attention model captures cross-channel feature correlations while maintaining an independent representation in the metastructure. The network module performs a set of transformations on low-dimensional embeddings and combines their out-put. Each transformation involves applying a channel-by-channel attention model to capture feature map interdependencies. Each transformation has the same topology. This approach allows us to speed up training using the same implementation as the unified CNN operators.

The last step in the technology of creating an e-commerce product catalog is to determine the dominant color of the product and generate a sample - a small image (usually 50x50 px) filled with the dominant color. For this, the k-means method [10] is used. By choosing the right value of k, the center of gravity of the largest cluster will be a pretty good representation of the dominant color in the image.

B. Neural network selection

When choosing an architecture for object detection and segmentation, we were guided by 3 factors. The first factor is the speed of prediction because the inference should be in real time. The second factor relates to the ability to support the recognition of intra-class variation - objects that have only minor differences. The third factors implies on the usage of only one model that could solve both object detection and segmentation task.

At the time of research, YOLACT was the fastest model for object recognition and instance segmentation [5]. This fact is the reason why its architecture was chosen for this work.

YOLACT is a convolutional neural network. The idea behind the convolution is that the elements of kernel are presented by weights. Input channels of the image are processed by each of that cores.

During creation of YOLACT architecture the goal was to solve the segmentation task by appending a mask branch to the existing one-stage model. The same idea is used in Mask R-CNN for Faster R-CNN, but without an explicit function localization step (for example, re-combining functions). The YOLACT architecture is shown in Fig. 2 [5].

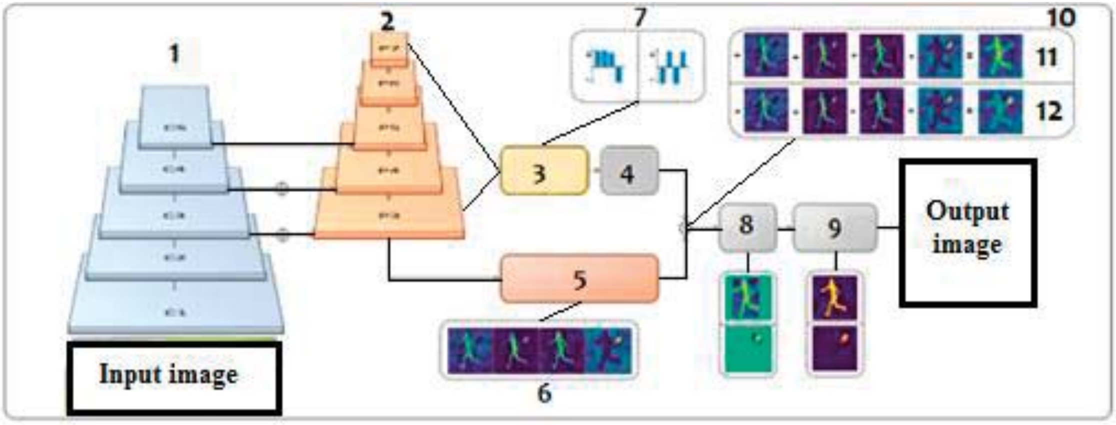

Fig.2 YOLACT architecture [5]: 1 – feature map, 2 – feature pyramid, 3 – prediction layers, 4 – NMS (non-maximum suppression technique), 5 – protonet, 6 – prototypes, 7 – mask coefficients, 8 –pruning, 9 – threshold, 10 – ensemble, 11, 12 – detection

The following loss functions were used during training process:

- classification Loss ($Lcls$);
- bounding box regression Loss ($Lbox$);
- mask Loss ($Lmask$) with weights 1, 1.5 и 6.125 accordingly.

Functions $Lcls$ and $Lbox$ are defined similar to [11]. Then, to calculate $Lmask$, pixel binary cross entropy (BCE) between obtained masks (M) and ground truth masks (Mgt) is used: $Lmask = BCE\ (M;\ Mgt)$.

ResNet-101 was used as a backbone. The size of the image is 800×800 pixels.

The ResNet-101 architecture is shown in Fig. 3.

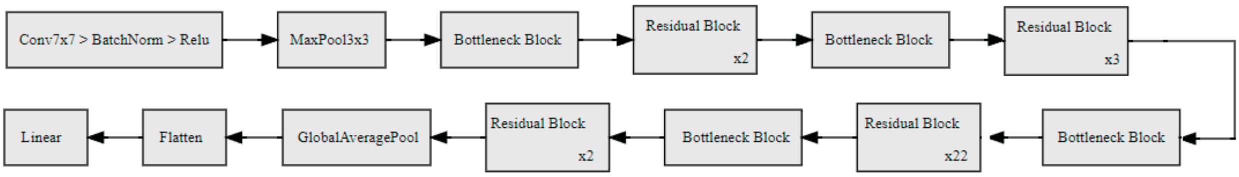

Fig.3 Architecture of ResNet-101

For human body parts detection we selected neural network OpenPose.

OpenPose [8] is a computer vision library that is specifically designed for the detection and tracking of human body parts. It uses a neural network-based approach to analyze images or videos and generate a set of key points corresponding to different parts of the body. These key points can be used to track movements, gestures, and poses of human subjects in real-time.

In the context of e-commerce, OpenPose is used as a component of an automated catalog creation system proposed in the study discussed in the article. The system can use OpenPose to identify human body parts in product images, such as models wearing clothing or accessories. This information can then be used to automatically crop the image to focus on the product itself, rather than the entire image. The architecture of OpenPose is shown in Fig. 4.

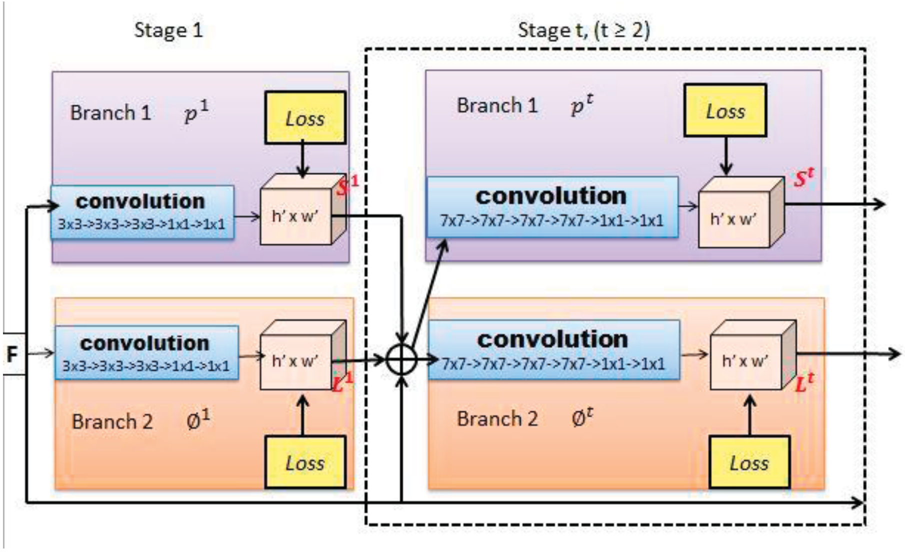

Fig.4 OpenPose architecture

C. Attention model and its application for e-commerce tasks

Attention models are a type of neural network architecture that have gained increasing popularity in recent years, particularly in natural language processing and computer vision tasks. The attention mechanism allows the model to focus on certain parts of the input data that are deemed most relevant, while ignoring irrelevant or redundant information. This mechanism can significantly improve the performance of the model, especially when working with large or complex data sets.

In the context of e-commerce, attention models have been used for various tasks such as product recommendation, image classification, and object detection. In this study, the authors propose a novel approach that combines convolutional neural networks (CNNs) with attention models to create an automated catalog of e-commerce products. The attention mechanism allows the network to focus on significant features of the product image, such as its shape, color, and texture, while ignoring irrelevant background or noise. This can lead to more accurate and efficient product recognition, segmentation, and cropping.

We propose modification of the ResNet-101 [12] backbone of the YOLACT neural network using an attention model. The idea is to focus only on relevant image features when solving the e-commerce product recognition and segmentation problem.

The attention model captures cross-channel feature correlations while maintaining an independent representation in the metastructure. The network module performs a set of transformations on low-dimensional embeddings and combines their output. Each transformation involves application of a channel-by-channel attention model to capture feature map interdependencies. Each transformation has the same topology. This approach allows us to speed up learning using the same implementation as the unified CNN operators. The resulting computational block is called the attention division block. Combination of several blocks of attention forms the necessary architecture.

The attention division block consists of a group of feature maps and attention division operators. Features are divided into groups, which are controlled by the cardinality hyperparameter of this group K. A new base hyperparameter R is also added, which reflects the number of divisions within the group K in such a way that the total number of feature groups $G = KR$. The attention model is inserted into the ResNet-101 network as follows: after the pooling layer two consecutive fully connected layers where the number of groups equal to the cardinality of the group are put in to predict the attention weights of each block. This approach makes it possible to combine the first 1×1 convolutional layers into one layer. 3x3 convolutional layers can be represented as a single group convolution with $R * K$ groups. Consequently, the attention model block has a modular structure using standard convolutional neural network operators. The internal structure of the ResNet-101 network is replaced by the attention model block. The architecture of the attention model block is shown in Fig. 5.

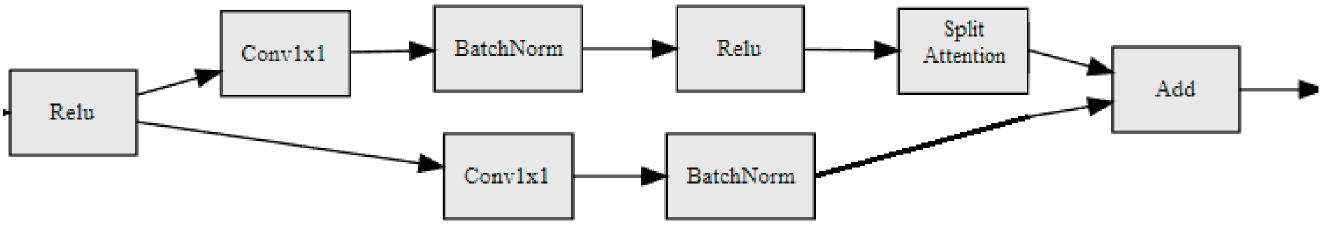

Fig.5 ResNet-101 block with attention model architecture

In the Smart Cropping system attention model has been integrated into VGG-19 [13] that is a part of the OpenPose architecture.

We included 2 key changes in VGG-19 architecture and the algorithm is the following (Fig. 6):

- after layers 7, 10, and 13 (highlighted in blue in Fig. 6), attention estimators are inserted, on the basis of which a binary mask is calculated, where 0 is irrelevant information for the desired object, and 1 is important. The mask, represented by the matrix, is then multiplied by the original result of the layer for which it was calculated, for example, 7, thus overestimating attention;
- the last fully connected layer was replaced with a fully connected layer, the input of which is the results of 3 attention estimators.

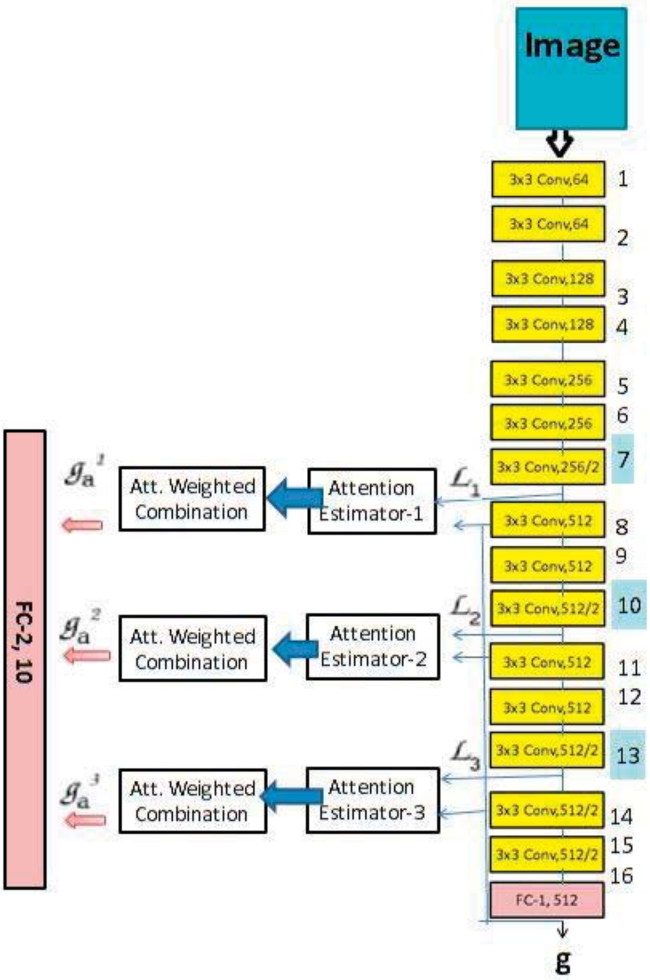

Fig.6 VGG-19 with attention model

IV. Experiments and Results

In the course of the research, the following algorithms were developed:

- image cropping task – an algorithm based on the OpenPose architecture using a modified attention model for VGG-19. The algorithm showed an 8% improvement in accuracy and is able to recognize 23 key points on the human body.
- the task of detecting an object in an image - an algorithm using the YOLACT convolutional neural network using the attention model. Model training was aimed at recognizing 26 classes of e-commerce objects: model (full-length person), shoes (four classes), clothing (five classes), food (five classes), cosmetics (five classes), kitchen appliances, accessories, and background class. The attention model was used in the basis of the ResNet-101 neural network to highlight the most significant features of an object. It allowed to improve object recognition by an average of 3%.
- segmentation task – an algorithm based on the YOLACT convolutional neural network using weights standardization. Model training was aimed at recognizing 21 classes of e-commerce objects. Weight standardization was used in convolutional layers in the forward pass of neural network training. This made it possible to improve object classification by an average of 3%, and object detection by 4%.
- the task of determining the main color of the product in the im-age - an algorithm based on k-means that solves the problem of clustering colors found on the image that was cropped using the previous algorithm.

In all cases, an NVIDIA T4 GPU and the COCO data set [14] were used for training, the size of the original image was at least 800x800 px. The algorithms were implemented on the PyTorch framework.

The results are shown in Fig. 7.

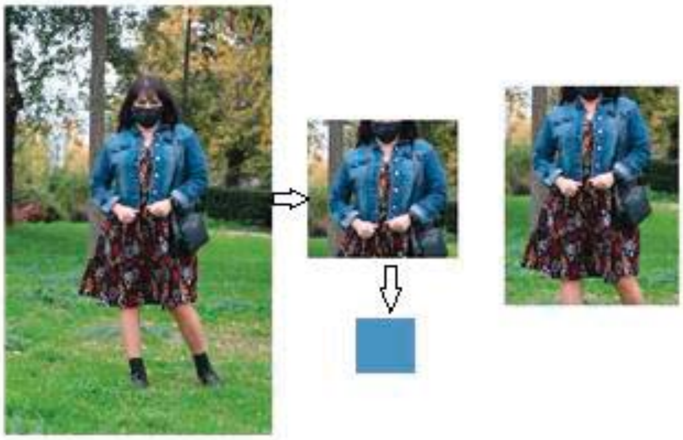

Fig.7 Results of applying the technology to images from real online store

The results of classical YOLACT, YOLACT with standardization of weights and with the use of the attention model, are shown in Table 1.

Table 1. Average precision of the trained neural networks.

Method	*FPS*	*AP*	*AP_{50}*	*AP_{75}*	*AP_S*	*AP_M*	*AP_L*
YOLACT	28,31	33,7	53,5	35,9	17,2	35,6	45,7
YOLACT with weights standardization	28,31	36,8	59,2	38,2	22,4	37,2	47,2
YOLACT with attention model	28,31	37,4	59,9	39,3	25,2	37,7	48,4

In Table 2 comparison of the Smart Cropping results with other models such as CMU-Pose [15], Mask-RCNN [16], OpenPose.is shown.

Table 2. Comparison of the Smart Cropping system with other models

Method	**AP**	**AP^{50}**	**AP^{75}**	**AP^{M}**	**AP^{L}**
CMU-Pose	61.8	84.9	67.5	57.1	68.2
Mask-RCNN	63.1	87.3	68.7	57.8	71.4
OpenPose	65.3	85.2	71.3	62.2	70.7
Smart Cropping	67.4	85.3	72.9	63.5	71.1

V. Discussion

In this paper, usage of the attention model for the problem of e-commerce product image recognition and segmentation to create an automatic catalog of e-commerce goods was proposed. As a result, the model makes it possible to detect feature correlations on different network layers in order to highlight significant and filter out insignificant information about the object in the image. The obtained results are of higher quality than Mask R-CNN [16] and FCIS [17] (for the same class of products).

It has been established that the attention model allows the network to focus more accurately on the features of the object, which affects the quality of the trained network and the rate of convergence.

However, the problem of data quality can be identified. If the image does not meet the standards for uniformity, dullness, blur, etc., then standard auto-correction methods do not give a tangible result.

A possible way to solve these problems is to use automatic data preprocessing to improve image quality.

We can also extract the following limitations of the study:

- synthetic origin of the training dataset due to the lack of various background of the images;
- small value of batch size conditioned by available computation power;
- no automatic data preprocessing to improve image quality.

VI. Conclusion

Solving such problems as detecting an object in an image, segmenting it, cropping it, and determining the dominant color of an object made it possible to create a technology for automatically preparing images for an electronic product catalog. This technology was tested on real images of goods from one online store.

The results obtained in the developed program can be used to prepare images of other categories of goods.

As more data becomes available and new technologies appear, we constantly improve the built system. For example, we’re planning to evaluate the relevance of an image using the title and description of a product.

References

[1] Chaudhuri, A. A Smart System for Selection of Optimal Product Images in E-Commerce / A. Chaudhuri, A. [et al.] // IEEE International Conference on Big Data (Big Data) – IEEE, 2018. – pp. 1728–1736.

[2] Zhang, X. Content-Based E-Commerce Image Classification Research / X. Zhang [et al.] // IEEE Access – 2020. – vol. 8, pp. 160213-160220.

[3] Bossard, L. Apparel classification with style / L. Bossard [et al.] // Asian conference on computer vision – Berlin, 2012 – P. 321–335.

[4] Lao, B. Convolutional neural networks for fashion classification and object detection / B. Lao and K. Jagadeesh // CCCV 2015 Computer Vision. – 2015 – PP. 120–129.

[5] Bolya, D. YOLACT: Real-time Instance Segmentation / D. Bolya [et al.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9157-9166

[6] Neural network training acceleration by weight standardization in segmentation of electronic commerce images / V. Sorokina, S. Ablameyko // Studies in Computational Intelligence – 2020. – Vol.976, PP.237-244.

[7] Viktoria Sorokina and Sergey Ablameyko. Extraction of Human Body Parts from the Image Using Convolutional Neural Network and Attention Mode. Proceedings of 15th International conference “Pattern Recognition and Information Processing”, Minsk, UIIP NASB, 2021, pp.84-88.

[8]] Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / Zhe Cao [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 21-26 July 2017. – Honolulu, 2017. – P. 1302-1310.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference on Learning Representations.

[10] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). California: University of California Press.

[11] Instance-sensitive fully convolutional networks / J. Dai [et al.] // 14th European Conf. on Computer Vision, Amsterdam, 11-14 October 2016. – Amsterdam, 2016. – Vol. 9910. – P. 534–549.

[12] He, K. Deep Residual Learning for Image Recognition / K. He [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition – Las Vegas, 2016. – P. 770-778

[13] Very Deep Convolutional Networks for Large-Scale Image Recognition / Karen Simonyan and Andrew Zisserman // International Conference on Learning Representations, San Diego, May 7-9, 2015. – San Diego, 2015. – P. 1137–1149

[14] COCO dataset // COCO 2018 Keypoint Detection Task [Electronic resource] – Mode of access: http://cocodataset.org/#overview – Date of access: 05.04.2019.

[15] The CMU Pose, Illumination, and Expression (PIE) database / T. Sim, S. Baker and M. Bsat, // Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition – 2002. – P. 53-58

[16] He, K. Mask R-CNN / K. He // IEEE Intern. Conf. on Computer Vision (ICCV), Venice, 22–29 October 2017. – Venice, 2017. – P. 2 980–2 988.

[17] Fully convolutional instance-aware semantic segmentation /Yi Li [et al.] // IEEE Intern. Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, 21–26 July 2017. – Honolulu, 2017. – P. 4 438–4 446.