2019 International Russian Automation Conference (RusAutoCon)

Transfer Learning Approach to Fine-Grained Image Classification

Valentin A. Golodov
School of Electrical Engineering and Computer Science
South Ural State University
Chelyabinsk, Russia
golodovva@susu.ru

Mariya S. Dubrovina
School of Electrical Engineering and Computer Science
South Ural State University
Chelyabinsk, Russia
lacus_veris@mail.ru

Anastasiya S. Paziy
School of Electrical Engineering and Computer Science
South Ural State University
Chelyabinsk, Russia
paziy_anastasiya@mail.ru

***Abstract*—This paper deals with fine-grained image classification in which instances from different classes share common parts but have wide variation in shape and appearance. Introduction gives a short review of existing works on this topic such as weakly supervised learning, discriminative localization network. Various approaches to the solution were considered and the method of the transfer of learning for the fine-grained image classification is considered as one the most promising methods. Convolutional neural networks are used. The approach demonstrates its effectiveness and allowed obtaining high accuracy of recognition on the dog breed recognition problem. Next step is to add more breeds and improve the recognition result.**

Keywords— recognition, convolutional neural networks, fine-grained image classification, transfer learning

I. Introduction

Classification of very similar images is one of the most important and difficult tasks in the field of computer vision today. This task involves the recognition of subcategories of a certain general category. This may be recognition of types of birds, plants, cars, planes, dogs, and more. The problem of fine-grained image classification remains open today, as there are difficulties in trying to classify image data.

First, in one subcategory there can be a huge variety of objects, and this difference can affect the further classification of the object. For example, we can consider the image, which shows one of the subspecies of bearded vulture (Fig. 1 (a–c)). In these three images, the bird is depicted from different angles so that even an ordinary person could classify them as different species of birds.

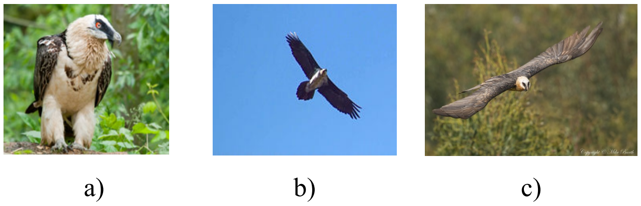

Fig. 1. Bearded vulture

Second, there may also be a huge variety of objects belonging to different subcategories, but having similar external features. Consider the images of three different species of birds. This figure shows the following bird species: rook, raven and crow Fig. 2 a), Fig. 2 b) and Fig. 2 c) respectively.

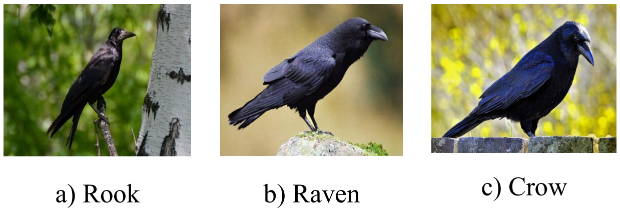

Fig. 2. Birds

Anyone without ornithology knowledges can hardly see difference between these species of birds, because the appearance of the birds look almost the same.

However, if you look at the images in more detail, you can identify distinctive areas of objects, such as the shape of the beak, tail and head, which can help in their further classification. Thus, the selection of these areas of the object is a priority in the field of fine-grained image classification.

II. Approaches to Fine-Grained Image Classification

There are many works on this topic [1]. Thus, with the use of convolutional neural networks, studies were conducted in the field of classification of bird species [1], [2], varieties of different colors [3] and also breeds of dogs [4]. This set of fields is determined by existing of large and free datasets but the approach can be considered in any field with very similar data in the dataset. The most promising approaches use neural networks of different architectures [5]–[9].

Lets consider two few examples of neural networks applications to the fine-grained image classification.

A. Dog breed classification

The most famous work on fine-grained image classification on the example of dog breed classification was carried out by researchers from Columbia and Maryland University in 2012 [10]. They proposed a new approach for this task. The selection of important features on the object (eyes, nose, ears),

The work was supported, act 211 Government of the Russian Federation, contract № 02.A03.21.0011.

978-1-7281-0265-8/19/$31.00 ©2019 IEEE

which will be used for classification, significantly increases the accuracy of recognition. On the dataset of 8351 images of 133 breeds, including not only classes, but also 66808 distinctive labels, 8 on each image, they were able to achieve 67 percent accuracy in the first prediction and 93 percent accuracy that the breed will be in the top ten predictions. Fig. 3 presents samples of labeled images by workers and MTurk (a crowdsourcing online marketplace that allows individuals or businesses to coordinate human actions to perform tasks that a computer cannot currently perform).

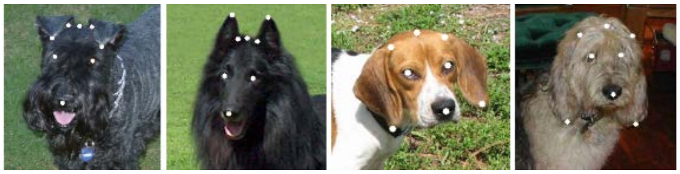

Fig. 3. Labeled images

However, it is not working if the dog's head on the test image is turned sideways or looks down, the algorithm is not able to determine the key points on the object and accordingly recognize it correctly. Also, the disadvantage of this approach is that the allocation of distinctive features on the object is difficult and expensive, since the placement of marks on 8351 images was done by people. This approach can be used only for specific task.

B. *Weakly supervised learning*

Another approach was proposed by a group of scientists from the Institute of Computer Science and Technology and Beijing University in China [11]. They proposed a method of training with weak data markup (weakly supervised learning) [12] for fast classification of fine-grained images.

In this method, an through n-path neural network was designed to simultaneously localize distinctive areas on the object and encode distinctive features. The main objective of this study was not only to improve the accuracy of recognition of difficult to distinguish images, but also to increase the speed of recognition and eliminate the dependence of recognition accuracy on the annotation of objects and its parts in the image. The schematic architecture is shown in Fig. 4. It consists of two subnets:

- multi-level attention extraction network, MAEN;
- discriminative localization network, DLN.

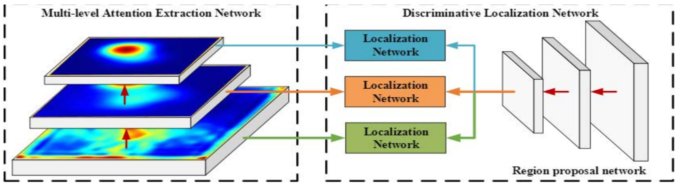

Fig. 4. Weakly supervised learning network

MEAN bounding box provides distinctive areas for the automatic exercise DLN, however, the allocation of data areas MAEN is not very neat. Also, notice that MAEN is used only at the training stage.

DLN optimizes the derived from MEAN initialized distinctive areas to find the most important key areas that will help distinguish one subcategory from the others.

The combination of these networks helps to reduce the disadvantages of these networks and to achieve the best accuracy in classification.

Consider each subnet in a bit more detail.

1) *Multi-level attention extraction network*

In convolutional neural networks, different feature maps reflect different image characteristics. In Fig. 5 various feature maps of bird images that were obtained from convolutional layers "Conv 4_3", "Conv5_3" and "Conv_cam" in the multi-level attention extraction network are presented. As you can see, different convolutional layers focus on different parts of the image and provide additional information that helps to improve classification accuracy.

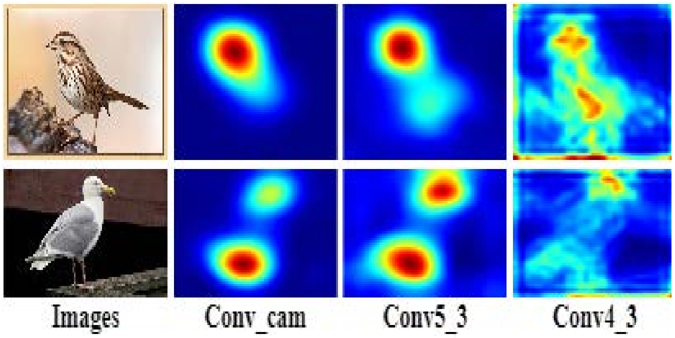

Fig. 5. Feature maps

The input of the subnet is a cropped image, and the output is n different feature maps from n convolutional layers to account for each pixel in the image when classifying it. Then, bounding boxes are generated for feature map data, instead of a fully-connected layer, the output uses global average pooling, followed by a softmax layer. By summing feature maps with appropriate weights, a feature map is generated for each image on a specific convolutional layer. At this step, n feature maps based on n different convolutional layers will be generated. In the final step, a binarization operation is applied on each card with an adaptive threshold, which is calculated using the OTSU algorithm [13].

Then we take the bounding box, which covers the largest part of the distinctive area of the image. Similarly, we obtain n bounding boxes of distinctive areas that are used in the DLN learning phase.

2) *Discriminative Localization Network*

Received n card from MAIN signs used at the training stage DLN. In order to obtain the best result, using the information obtained from MAEN, an end-to-end n-path neural network based on Faster R-CNN [14] was designed, which consists of several networks for localization of regions and one network for area assumption. Faster-CNN was used to speed up the object recognition process while maintaining good recognition accuracy. However, for the purpose of this task it was necessary to upgrade Faster-CNN.

At the training stage of Faster R-CNN it is necessary to provide true bounding boxes of distinctive areas on the image. But since this approach is difficult and expensive, and on some types of tasks it is almost impossible to perform, authors used bounding boxes obtained from MAEN.

In order to improve recognition accuracy, it was decided to use a multi-level network for feature extraction. However, the network architecture is limited by the architecture of the Faster R-CNN, so it was decided to design a n-path through network with multiple networks, localization regions, and one network for assuming a region in which all network localization use the characteristics of the convolutional layer received from the network according to the assumption region.

Since this approach uses multilevel feature extraction, it is necessary to use n networks to localize areas.

Each network is connected to the RPN using the region of interest (RoI) pooling layer, which is used to extract a fixed length vector from the feature map for each proposed area generated by the RPN. Each vector is fed to the input of the network for localization of a region and passed forth below was derived two conclusions: the predicted sub-category and bounding box of the distinctive region.

The approach gives the following results:

1) On the basis of the images of the CUB-200-2011 [15], which includes images of 200 11788 subcategories of different species of birds, has been achieved recognition accuracy of 85.71%.

2) on the basis of Cars-196 images [16], which includes 16185 images of 196 different subcategories of machines, the recognition accuracy was 92.30 %.

The disadvantage of this approach is the complexity of adaptability to particular problems. For the correct operation of the algorithm requires a fairly large set of well-marked data. In addition, it is necessary to have a lot of computing resources for the training.

III. Transfer Learning in Fine-Grained Classification

This work devoted to develop easy to use common approach to the fine grained image classification. Very similar dogs breeds recognition task is considered an example of usage.

A. *Simple Sequential Network*

Initially, for fine-grained image classification it was planned to develop its own neural network from scratch, but when training this network, the accuracy of recognition of dog breeds was quite low and amounted to only 22%. Then it was decided to replace the last fully-connected layer with global average pooling, experiments show that this approach can significantly improve the accuracy of recognition [17].

Using global average pooling approach, the recognition accuracy increased to 52%. Training was 90 epochs, each era was spent about 30 minutes. The change of the coefficient of learning rate, activation functions, adding another layer of convolution – is not able to increase recognition accuracy. The schematic architecture of this network is presented below.

Input-Conv-MaxPool-Conv-MaxPool-Conv-MaxPool-Flatten-GlobalAvgPool-SoftMax

Where:

- Conv-convolution layer. Multiplies the filter values by the original pixel values of the image (piecemeal multiplication), after which all these multiplications are summed.
- MaxPooling is down sampling strategy in Convolutional Neural Networks.
- Flatten lauer transforms multiple pooled feature maps into a long vector of input data that you then pass through the artificial neural network to have it processed further.
- Global average pooling layers to minimize overfitting by reducing the total number of parameters in the model. Replacing the FC layer with the GAP layer significantly increased the recognition accuracy.
- Soft maximum function commonly used for the last layer of deep neural networks for classification problems.

B. *Using Existing Network*

Next improving step was to use the transfer of training. The essence of this approach is that pre-trained neural networks that have been trained on a large set of data can be used to solve problems in other areas. This approach not only improves accuracy by initially training the network on a large amount of data, but also reduces the time and resource costs of training. Successful applications of the transfer of learning approach in different tasks are published [18]–[22].

As ready pre-trained neural network it was decided to use a neural network model InceptionV3 [23]. Also network VGG16 [24] have been used, but the obtained recognition accuracy was 67%.

InceptionV3 is a multi-layered deep neural network, trained on a large set of ImageNet images, capable of recognizing about 1000 different categories. Model IncepcionV3 Inception uses modules that, in fact, are mini versions of the large model. The architecture of this network is shown in Fig. 6.

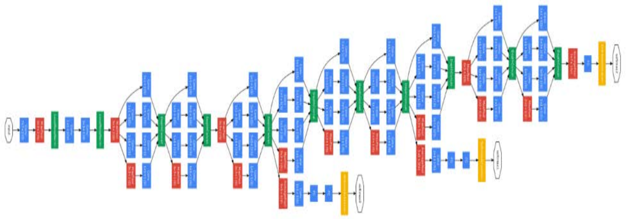

Fig. 6. Inception V3 architecture

This network model allows you to use multiple bundles in parallel on each layer, concatenating (combining) the resulting feature maps before moving to the next layer.

Assume that the next layer is the initial module. Then each of the feature maps obtained by convolution on the previous layer will pass through a mixture of convolutions of the current layer. Thus, it is not necessary to know in advance which convolution is better to use, for example, 3 × 3 and then 5 × 5. Instead, you can just use all the convolutions and let the model choose the best one. In addition, this architecture allows the model to store both local features through smaller convolutions and high-level features using large convolutions. Fig. 7 shows the architecture of one Inception module

Since this pre-training network will be used for a specific task and, accordingly, on another data set, it is necessary to remove the last classification layer from the model. Thus, will only be used for the convolutional part of the network see Fig. 8.

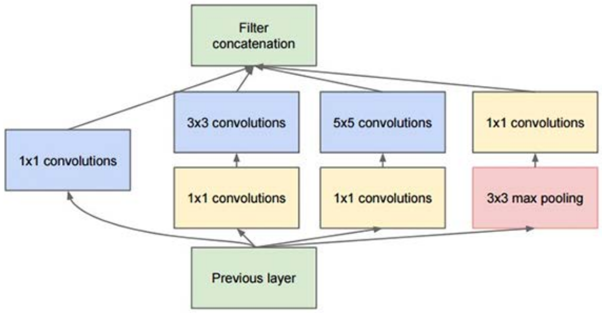

Fig. 7. Inception V3 module

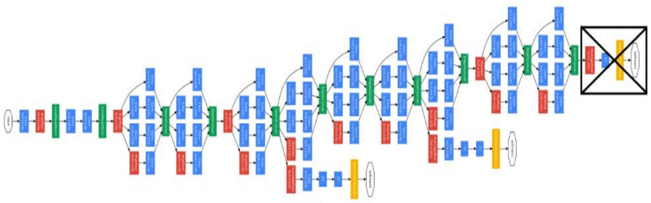

Fig. 8. Convolutional part of the inception network

So, you will need to feed cropped images to the InceptionV3 input and extract feature maps from the last layer of the network, which will then be used as input to another classification network. The architecture of this classification network is shown in Fig. 9.

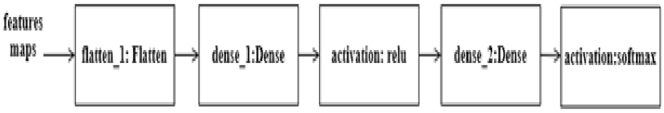

Fig. 9. Classifier network

Thus, InceptionV3 will act as a separate module for data preprocessing before classifier input.

IV. Implementation and Results

A. Dataset

```
base_model=applications.InceptionV3(include_top=False,
                                    weights='imagenet')
```

The paper uses data from Stanford Dog dataset [25], which was developed specifically for the classification of hard-to-distinguish images. This database includes 20580 images of 120 breeds of dogs. But since for some species the number of images wasn't enough, we also used images from other sources [26], [27]. It was decided to use 21 breed of dog for recognition within the framework of the task. Thus, the training set amounted to 4260 images, validation set of 850 images, a test set of 210 images.

B. Realization

Neural network was implemented using Python 3 programming language and keras-tensorflow deep learning library. Loading convolution layers may be easy performed by loading pretrained network with following command.

Fig. 10. Loading pretrained network

Using the include_top=False parameter, we do not load the classification part of the InceptionV3 network. Thus, only the convolutional part of the network will be loaded.

Classifier network Fig. 9, uses an image preprocessed by InceptionV3 as an input.

```
model = Sequential()
model.add(Flatten(input_shape=train_data.shape[1:]))
model.add(Dense(256, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(num_classes,activation='softmax'))
```

Fig. 11. Classifier network implementation

Since the number of images for each subcategory is evenly distributed, accuracy was used as a metric.

Most useful for the classification purposes categorical cross entropy loss function was used. Training uses Adam optimizer that gives recognition accuracy 90% and validation loss 0.92.

Then adamax optimizer with a learning rate coefficient of 0.0001 was used for the training that raise recognition accuracy to 93%, and the validation loss decrease to 0.62.

Recognition of the dog breed in the image takes about of 1 - 2 minutes. This is due to the fact that image preprocessing takes a lot of time, due to the large number of layers in the neural network InceptionV3.

A possible solution to this problem is to increase the computational resources for the algorithm or using quantizing of deep convolutional networks. Some another way is to use segmentation as preprocessing [28]. Using of the proposed algorithm on platform with sufficiently low computation resources, for the tasks where recognition time is important, will be impractical in current release. Full code is available in github repository by the link https://github.com/lacusver/dog-breed-classificationtf.

All results are given in the Table I.

TABLE I. RESULTS

Experiment	Accuracy	Validation loss
Simple Sequential Network	22%	-
Simple Sequential Network with global average pooling	52%	-
Transfer learing with Adam training of classifier	90%	0.92
Transfer learing with Adamax training of classifier	93%	0.62

V. CONCLUSIONS

Thus, fine-grained image classification is an actual and difficult task today. Recently, good results have been obtained in this area. The method of transfer of learning was proposed to classification of hard-to-distinguish images, neural network for the recognition of dog breeds was developed. This network uses a pre-trained neural network as a preprocessing. It allows to achieve a sufficiently high accuracy of recognition. However, the use of this method in realtime tasks is impractical because of its computational complexity. In the future, it is planned to find solutions to address this shortcoming, as well as to expand the training sample while maintaining the accuracy of recognition.

REFERENCES

[1] S. Branson, G. Van Horn, P. Perona, and S. Belongie, “Improved Bird Species Recognition Using Pose Normalized Deep Convolutional Nets,” Proc. of the British Machine Vision Conf., pp. 87.1–87.14, 2014.

[2] S. Haobin, Z. Renyu, and S. Gang, “Fine-Grained Bird Classification Based on Low-Dimensional Bilinear Model,” IEEE 3rd Int. Conf. on Image, Vision and Computing, pp. 424–428, 2018.

[3] M. E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” Proc. - 6th Indian Conf. on Computer Vision, Graphics and Image Processing, pp. 722–729, 2008.

[4] X. Wang, V. Ly, S. Sorensen, and C. Kambhamettu, “Dog breed classification via landmarks,” 2014 IEEE Int. Conf. on Image Processing, pp. 5237–5241, 2014.

[5] A. Sinkov, G. Asyaev, A. Mursalimov, and K. Nikolskaya, “Neural networks in data mining,” 2nd Int. Conf. on Industrial Engineering, Applications and Manufacturing, pp. 1–5, 2016.

[6] H. Ge, X. Tu, M. Xie, and Z. Ma, “The Smaller the Better: Fine-Grained Image Classification with Compressed Networks,” 11th Int. Symposium on Computational Intelligence and Design, pp. 37–40, 2018.

[7] W. Shi, Y. Gong, X. Tao, D. Cheng, and N. Zheng, “Fine-Grained Image Classification Using Modified DCNNs Trained by Cascaded Softmax and Generalized Large-Margin Losses,” IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 3, pp. 683–694, 2019.

[8] M. Srinivas, Y.-Y. Lin, and H.-Y. M. Liao, “Deep dictionary learning for fine-grained image classification,” IEEE Int. Conf. on Image Processing, pp. 835–839, 2017.

[9] Z. Lin, “A Unified Matrix-Based Convolutional Neural Network for Fine-Grained Image Classification of Wheat Leaf Diseases,” IEEE Access, vol. 7, pp. 11570–11590, 2019.

[10] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur, Dog Breed Classification Using Part Localization. Berlin: Springer, 2012.

[11] Z. H. Zhou, “A brief introduction to weakly supervised learning,” National Science Review, vol. 5, no. 1. pp. 44–53, 2018.

[12] X. He, Y. Peng, and J. Zhao, “Fast Fine-grained Image Classification via Weakly Supervised Discriminative Localization,” 2017.

[13] D. Liu and J. Yu, “Otsu Method and K-means,” Ninth Int. Conf. on Hybrid Intelligent Systems, pp. 344–349, 2009.

[14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.

[15] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset”.

[16] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” Proc. of the IEEE Int. Conf. on Computer Vision, pp. 554–561, 2013.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization”.

[18] E. Chalmers, E. B. Contreras, B. Robertson, A. Luczak, and A. Gruber, “Learning to Predict Consequences as a Method of Knowledge Transfer in Reinforcement Learning,” IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 6, pp. 2259–2270, 2018.

[19] B. Zoph, G. Brain, V. Vasudevan, J. Shlens, and Q. V Le Google Brain, “Learning Transferable Architectures for Scalable Image Recognition”.

[20] W. Sun and X. Qian, “An improved transfer learning algorithm for document categorization based on data sets reconstruction,” Proc. of the 10th World Congress on Intelligent Control and Automation, pp. 575–578, 2012.

[21] B. Alothman and P. Rattadilok, “Towards using transfer learning for Botnet Detection,” 12th Int. Conf. for Internet Technology and Secured Transactions, pp. 281–282, 2017.

[22] C. Qiu, “Transfer Learning for Small-Scale Fish Image Classification,” 2018 OCEANS - MTS/IEEE Kobe Techno-Oceans, pp. 1–5, 2018.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” 2015.

[24] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 2014.

[25] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs”.

[26] Cane Corso (Italian watchtower). Cane Corso Club - national Cane Corso breed club. [Online]. Available: http://www.corsoclub.ru/.

[27] Alaskan Malamutes is the largest database in Russia. [Online]. Available: http://malamuts.ru/.

[28] V. A. Golodov and A. N. Fedorov, “Application of the Discriminative Loss Function to the Biomedical Image Instance Segmentation,” Proc. of the Conf. IEEE DSDT, 2018.