ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 166

7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)

Multi-Label Human Activity Recognition on Image Using Deep Learning

Pavel Nikolaev
Department 316
Moscow Aviation Institute (National Research University)
Moscow, Russia
npavel89@gmail.com

***Abstract*—This paper describes the model of convolutional neural network which is designed for multi-label human activity recognition. The possibilities of using activity recognition systems in the daily life of a person are considered. As part of this work, the study is conducted for the method of recognizing human activity on an image that can be obtained from a surveillance camera. To obtain more accurate recognition results, the network model used technology of transfer learning. Several pre-trained convolutional networks are considered using two types of transfer learning in order to find the best solution. The deep learning networks for solving the problem are implemented in Python using deep learning libraries. Considered models are trained to recognize binary multi-label human activity. Training and testing are performed on images collected by the author. The article also provides the obtained training and testing results of different models of convolutional neural networks. The data obtained are tabulated and also presented in graphical form.**

Keywords—image recognition; computer vision; human activity recognition; machine learning; artificial neural networks; deep learning; convolutional neural networks; transfer learning

I. INTRODUCTION

Systems for determining human activity can be applied in many areas. There are many different options for the use of such systems: security systems, medical and social services, educational and work processes, smart home [1].

For example, in security systems activity recognition can be used to detect suspicious, dangerous or unusual behaviors. In the medical field they can be used to monitor patients and disabled people. In the educational process such systems can be used to monitor the situation in classrooms, in the work process - when tracking the activities of employees.

And finally, ample opportunities for human activity recognition systems open up when they are integrated with home and building automation systems (the so-called smart homes or intelligent buildings). Through the recognition of current human activity it is possible to realize automatic control and monitoring of various devices and systems. For example, a smart system records that a person is in the room and determines that he is reading a book now. In the control system can be set such an algorithm that automatically increases the brightness of the light when reading a book. Many such scenarios can be implemented.

II. METHODS OF HUMAN ACTIVITY RECOGNITION

There are several methods for recognizing human activity. Some works consider using data from sensors of smartphones and other devices [2, 3], as well as from wearable sensors [4, 5, 6]. There are also methods for human activity recognition by image or video [7, 8]. One of these methods is to recognize the image in which there is a person performing a certain activity.

The author of this work is working to create the system for human activity recognition in the image obtained from a surveillance camera. In [1] we proposed the method consisting in the segmentation of a person and related subjects and the further classification of his activity.

However, most research on the subject is considered only an unambiguous classification of human activity. But often the activity performed by a person can be attributed to different classes: for example, at the same time a person is working at the computer and talking on the phone. In such a situation recognition system must define multiple classes of activity instead of one.

Multi-label image classification is one of the most challenging problems in computer vision [9]. Single-label image classification deals with images that are associated with a single label from a finite set of disjoint labels [10]. Multi-label classification implies that there may be several labels for one image. For example, there are 4 classes of images. Then with a single-label classification, one category can be represented as [1,0,0,0], and with a multi-label (two-label) classification - [1,1,0,0].

The first problem is considered to be much easier to solve than the problem of multi-label classification of images.

The purpose of this work is to consider the multi-label classification of human activity on image. In this case, the image recognition will be carried out using deep convolutional networks. Accordingly, we are tasked to train the convolutional neural network to determine several classes of human activity in the image at once.

III. CONVOLUTION NETWORK MODEL

A. Convolution neural networks in image recognition tasks

The most effective method of deep learning in solving problems of image recognition (image classification, detection of objects in an image and image segmentation) are convolutional neural networks (CNN). CNN are effectively used in computer vision for image recognition after the victory of the AlexNet network [11] in ImageNet LSVRC-2012 competition on visual recognition [12].

The structure of a convolutional network for image classification usually consists of convolutional and fully

connected parts. The convolutional part is designed to highlight the characteristics features of the classified image. The fully connected part is intended to determine the category to which the image belongs.

The convolutional part of the neural network consists of a set of alternating layers of convolution and pooling. During the convolution operation, templates are extracted from the input image (input feature map), and using the same transformations to all templates, the output feature map is produced [13]. This process is carried out by the sliding window method – a small window (kernel), usually 3 × 3, slides along the input image (feature map). At the same time, the operation of element-wise multiplication of the kernel by the part of the input image through which the kernel passes is performed.

During the pooling operation, the resolution of the feature map is reduced. Usually, the max-pooling operation is used as the pooling operation [14].

After all the operations of convolution and pooling, feature maps are formed, which are then combined and transmitted to the input of the fully connected part. The fully connected part consists of one or several fully connected layers. The output of this part of the network and the entire convolutional neural network is the values of the neurons of the last layer, which means the probability that the input image belongs to some class.

B. Transfer learning

We can train the convolutional neural network from scratch. If the network consists of a small number of layers, then it is not possible to achieve good results in recognition. You can add more layers, but training a deep neural network which contains tens or hundreds of millions parameters is a non-trivial task that takes hours, days or even weeks [15].

In such cases, it is advisable not to train the convolutional network from scratch, but to use the pre-trained neural network. This technology is called transfer learning. Using the technology of transfer learning can significantly improve the accuracy of image recognition compared with learning from scratch. Also, this method is best used in the case when the data set for training is not large enough.

The transfer learning strategy is to use a network which has already trained on a large dataset. The main idea of this technology consists of training a machine learning algorithm on a new task while exploiting knowledge that the algorithm has already learned on a previously related task [16].

According to [13], there are two ways to use technology of transfer learning: feature extraction and fine-tuning.

In the first case (feature extraction), the following operation algorithm is used:

- the convolutional base of the pre-trained network is frozen;
- replacing fully connected layers for classification;
- training of fully connected layers is carried out.

In the second case (fine-tuning), a slightly different work algorithm is used:

- only part of the convolutional base of the pre-trained network is frozen, and the last block of the convolutional part usually remains unfrozen;
- replacing fully connected layers for classification is also performed;
- training of unfrozen convolutional and fully connected layers is carried out.

C. Network model for multi-label activity classification

Our convolutional neural network model for the multilabel classification of human activity is based on the pre-trained neural network.

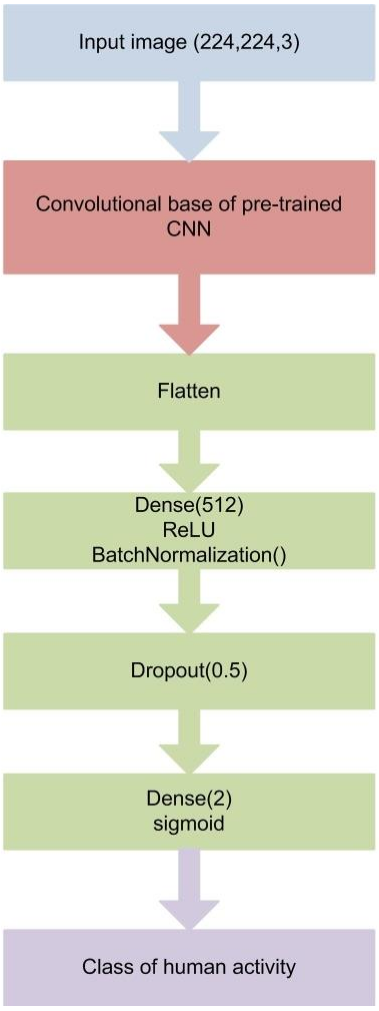

Fig. 1. Structure of used CNN model

In this work we used the following type of transfer learning: freezing of the convolutional base of the pre-trained network and the replacement of fully connected layers of by others. The network structure is shown in Figure 1.

IV. DATASET

The convolutional neural networks with different pre-trained networks have been trained to recognize two types of human activity: reading and drinking. The author has collected the set of 2700 images:

- 900 images with reading person;

- 900 images with drinking person;
- 900 images wit person doing both activities.

We have divided this data set into 3 parts.

- training set (60% of all images – 1620 image);
- validation set (10% of all images – 270 images);
- testing set (30% of all images – 810 images).

In the process of training, augmentation of data was used to expand the training data set. The following transformations were applied to the images from the training set:

- mirror image horizontally;
- zooming from 0.8 to 1.2;
- image rotation by 10 degrees and fill the blank areas with black.

In this paper, we consider several pre-trained neural networks. In all cases the same fully connected unit is used. This fully connected block consists of the fully connected layer with 512 neurons with the ReLu activation function. Batch normalization layer and dropout with coefficient of 0.5 are used. Using these methods, you can accelerate the training of the neural network and prevent its retraining, respectively.

The output fully connected layer contains only two neurons, because we managed to collect a suitable data set for only two types of human activity. The activation function used on the last layer is sigmoid.

V. EXPERIMENTS

We have trained several convolutional networks for image recognition using transfer learning. Such pre-trained networks were used as VGG16, VGG19 [17] and InceptionV3 [18]. ResNet50 [19] was also considered, but with its application it was not possible to achieve significant results in solving the set task. Therefore, we did not include the data on the application of this network in our work.

In total, we have trained 6 models of convolutional neural networks based on the VGG16, VGG19 and InceptionV3. Three networks have been built using the first method of transferring learning – feature extraction. In all cases, their convolutional bases have been frozen and fully connected blocks have been replaced.

Three other networks have been built with fine-tuning. In models based on VGG16 and VGG19, the last convolutional blocks have been subjected to additional training ("block_5"). In the model based on InceptionV3 layers have been frozen up to 248 layers, the remaining layers have been retrained. Fully connected blocks have been also replaced here.

The convolutional networks have been implemented in Python 3.6. Also the following deep learning libraries have been used:

- Keras 2.2.4;
- TensorFlow 1.12.0, version under GPU;
- cuDNN 7.1.4.

We have trained of our convolutional networks on the GPU Nvidia GeForce GTX 1060 6GB.

We used the following parameters for the neural networks training:

- size of input image – 224x224x3
- optimizer – stochastic gradient descent (SGD);
- the learning rate – 0.01;
- batch-size – 64 samples;
- loss function – binary_crossentropy;
- metric for assessing the correctness of the class definition – accuracy;
- the number of epochs for training – 50.

Table 1 presents the results of training for various models of convolutional networks, as well as the results of testing the best weights on validation and test data sets. The best weights were those weights at which the losses on the validation data set were minimal.

Figures 2-7 show the accuracy and losses of the considered models of CNN during 50 epochs of training.

VI. CONCLUSION

In this paper, several models of convolutional neural networks based on pre-trained networks have been considered. Pre-trained networks have been applied using two methods of transfer learning technology: feature extraction and fine-tuning. The use of these models has been researched as part of solving the problem of multi-label classification of human activity on the example of recognizing two classes: a person reading a book and a person drinking something.

As can be seen from the data in Table 1, the smallest recognition losses on the test set have been obtained using the VGG16 and VGG19 networks. It is worth noting that some of the best results have been achieved with the second method of transfer learning – fine-tuning.

TABLE I. THE BEST RESULTS OF TRAINING AND TESTING CNN

CNN model	Training set		Validation set		Testing set	
	Accuracy	*Loss*	*Accuracy*	*Loss*	*Accuracy*	*Loss*
VGG16 (feature extraction)	85.96%	0.3313	85.68%	0.3507	86.65%	0.3281
VGG16 (fine-tuning)	89.18%	0.2721	84.47%	0.3502	86.85%	0.3201
VGG19 (feature extraction)	79.29%	0.4303	85.44%	0.3510	82.16%	0.3864
VGG19 (fine-tuning)	90.90%	0.2239	87.62%	0.3249	84.12%	0.3799
InceptionV3 (feature extraction)	88.22%	0.2876	83.50%	0.4680	85.16%	0.4536
InceptionV3 (fine-tuning)	92.43%	0.1975	86.89%	0.4609	85.81%	0.4387

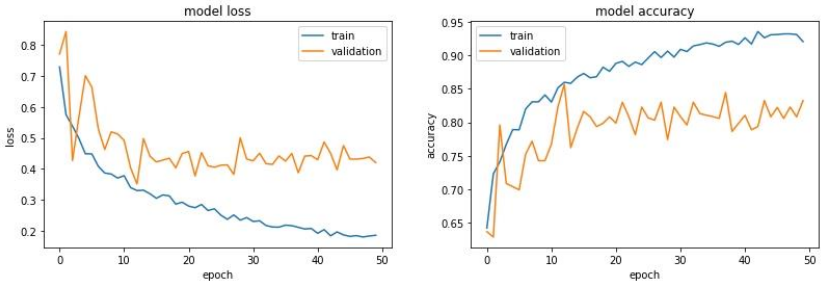

Fig. 2. Accuracy and losses of the CNN based on VGG16 with feature extraction

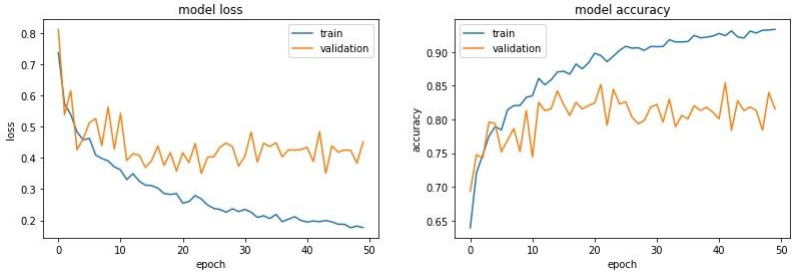

Fig. 3. Accuracy and losses of the CNN based on VGG16 with fine-tuning

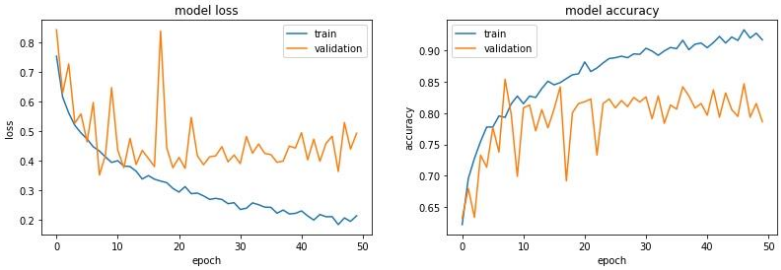

Fig. 4. Accuracy and losses of the CNN based on VGG19 with feature extraction

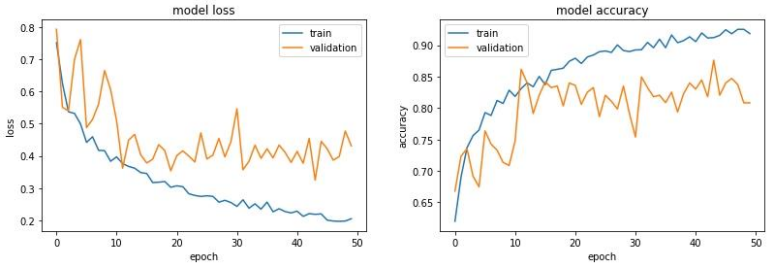

Fig. 5. Accuracy and losses of the CNN based on VGG19 with fine-tuning

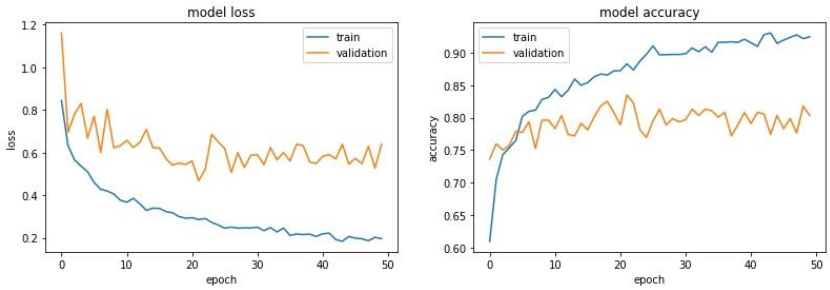

Fig. 6. Accuracy and losses of the CNN based on InceptionV3with feature extraction

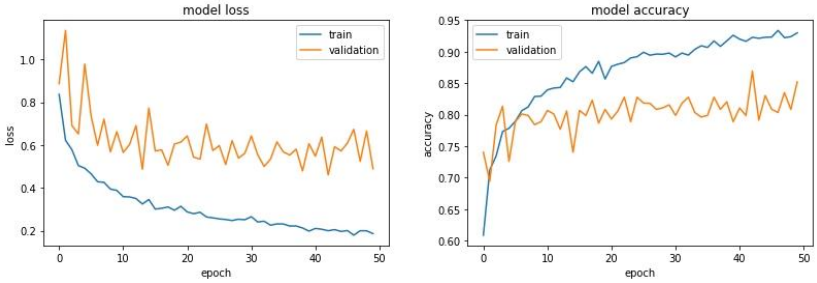

Fig. 7. Accuracy and losses of the CNN based on InceptionV3 with fine-tuning

REFERENCES

[1] P.L. Nikolaev, “Analysis of human activity by deep learning,” in Sistemnyj administrator, vol. 12 (193), pp. 80-83, 2018. (In Russian)

[2] J.A. Morales, D. Akopian, S. Agaian, “Human activity recognition by smartphones regardless of device orientation,” in Proc. of SPIE Conference on Mobile Devices and Multimedia: Enabling Technologies, Altorithms, and Applications, 2014, doi: 10.1117/12.2043180.

[3] B. Sefen, S. Baumbach, A. Dengel, S. Abdennadher, “Human activity recognition: using sensor data of smartphones and smartwatches,” in Proc. of the 8th International Conference on Agents and Artificial Intelligence, vol. 2, 2016, pp. 488–493, doi: 10.5220/0005816004880493.

[4] O. Lara, M. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1192-1209, 2013, doi: 10.1109/surv.2012.110112.00192.

[5] B. Chikhaoui, F. Gouineau, "Towards Automatic Feature Extraction for Activity Recognition from Wearable Sensors: A Deep Learning Approach," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, doi: 10.1109/icdmw.2017.97.

[6] W. Jiang, Z. Yin, "Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks," in Proc. of the 23rd ACM international conference on Multimedia - MM '15, 2015, doi: 10.1145/2733373.2806333.

[7] A. Kumar, A. Kumar, S. Kumar Singh, R. Kala, "Human Activity Recognition in Real-Times Environments using Skeleton Joints," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 7, pp. 61-69, 2016, doi: 10.9781/ijimai.2016.379.

[8] T. Subetha, S. Chitrakala, "A survey on human activity recognition from videos," in Proc. of the 2016 International Conference on Information Communication and Embedded Systems (ICICES), 2016, doi: 10.1109/icices.2016.7518920.

[9] L. Li, S. Wang, S. Jiang and Q. Huang, "Attentive Recurrent Neural Network for Weak-supervised Multi-label Image Classification," in Proc. of 2018 ACM Multimedia Conference on Multimedia Conference - MM '18, 2018, pp. 1092–1100, doi: 10.1145/3240508.3240649.

[10] S. He, C. Xu, T. Guo, C. Xu, D. Tao, “Reinforced Multi-Label Image Classification by Exploring Curriculum,” The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018, pp. 3183-3190.

[11] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, pp. 1097-1105, 2012.

[12] P.L. Nikolaev, “Adaptive system of intelligent environment based on neural networks,” in Proc. of IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine" (ITSMSSM 2017), Advances in Computer Science Research (ACSR), vol. 72, pp. 152-155, doi: 10.2991/itsmssm-17.2017.32

[13] F. Chollet, Glubokoe obuchenie na Python in St. Peterburg, Russia: Piter (In Russian), 2018.

[14] S. Nikolenko, A. Kadurin, E. Arhangelskaya, Glubokoe obuchenie in St. Peterburg, Russia: Piter (In Russian), 2018.

[15] N. Warakagoda, Ø. Midtgaard, “Transfer-learning with deep neural networks for mine recognition in sonar images,” in Proc. of the Institute of Acoustics, vol. 40., pt.2, 2018.

[16] M. Sabatelli, M. Kestemont, W. Daelemans, P. Geurts, “Deep Transfer Learning for Art Classification Problems,” in Computer Vision – ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science, vol 11130, pp. 631-646, 2019, doi: 10.1007/978-3-030-11012-3_48

[17] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations (ICLR), 2015.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016, doi: 10.1109/cvpr.2016.308.

[19] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016, doi: 10.1109/cvpr.2016.90.