Convolutional neural network of deep learning in computer vision and image classification problems

K E Tokarev 1,2, Zotov V M1, Khavronina V N1 and Rodionova O V1

1 Volgograd State Agricultural University, 26, Universitetskiy Avenue, Volgograd, 400002, Russia
2 Volgograd State Technical University, 28, Lenina Avenue, Volgograd, 400005, Russia

E-mail: tke.vgsha@mail.ru

Abstract. The article considers the possibilities of using the deep learning convolutional neural network ResNet in computer vision and image classification problems. The interpretation of the ResNet network and the datasets used for its training are presented, as well as a method for training a deep convolutional neural network with stochastic depth, which allows significantly reducing errors in the test sample.

1. Introduction

In the last decade, convolutional neural networks have shown excellent results in recognizing objects in images. In 2015, the deep convolutional neural network ResNet (Residual Network) was introduced, which surpassed the accuracy of the human level of image classification on the ImageNet dataset. Deep convolutional neural networks have surpassed the human level of image classification. Deep networks extract low-, medium-, and high-level features in an end-to-end, multi-layered way, and increasing the number of stacked layers can enrich the «layers» of features.

The depth of the neural network is crucial, but deeper neural networks are harder to train, primarily due to vanishing/exploding gradients issues that hinder the convergence of neural networks. These problems were mostly solved by initialization normalization and batch normalization, but the researchers also found a degradation problem, when the error on the training and test samples increases with increasing network depth (Figure 1). Researchers attribute this degradation problem to the complexity of optimizing deep neural networks [1-5, 10].

Figure 1. Example of error analysis on a training sample and a test sample of neural networks.

2. Materials and methods

ResNet is a convolutional neural network consisting of residual blocks and skip-connections. All convolutional network architectures prior to ResNet consisted of a sequence of convolutional layers, with each convolutional filter trained to create the most representative feature maps. A distinctive feature of the ResNet architecture is that the convolutional blocks are trained to slightly correct the preceding
feature maps. In other words, each convolutional block can be interpreted as an increment that needs to be added to the feature maps to improve accuracy [3].

Figure 2. UML-diagram of interaction between the user and the components of the software system.

The ResNet architecture consists of 4 groups of residual blocks, where the groups differ in the number of blocks and the number of filters in the convolutional layers of the blocks. The output of each convolutional layer is the so-called feature maps, which show the presence of certain features in the original image. After each group of blocks, the size of the feature maps of the images decreases and the number of filters in the convolutional layers of the next group of blocks increases by 2 times. Also, the ResNet architecture varies depending on its depth. Figure 3 shows the ResNet architecture with 34 layers, and Figure 4 shows the ResNet architecture with different depths for the ImageNet dataset.

Figure 3. Construction and validation of deep-learning model (A - Network architecture of ResNet34. B - Schematic illustration of the work flow for training and validation of the model).

To train a convolutional neural network, the ImageNet dataset is used, which is a data set of millions of high-resolution images marked up using crowdsourcing, related to about 25 thousand classes, and PASCAL VOC, which contains standardized image data sets for recognizing object classes, standard tools for accessing data sets and annotations, allowing you to evaluate and compare methods, as well as analyze performance in recognizing object classes.

The new interpretation of the ResNet network can be considered as a set of multiple paths of different lengths. Moreover, using only short paths during the training phase makes it possible to train very deep networks. To confirm this observation, the paper proposes to represent the sequence of residual blocks as a set of possible paths. Figure 4 shows a sequence of three residual blocks on the left, and all possible paths on the right [6].
3. Results and discussion

In the course of the review and research of works devoted to image analysis using convolutional neural networks, the most optimal method of training deep neural networks with stochastic depth was identified [7], which allowed us to successfully train a 1202-layer ResNet. Very deep convolutional networks consisting of hundreds of layers resulted in a significant reduction in error on the test sample. However, there are several problems that make it difficult to train very deep neural networks: the disappearance of the gradient and the increase in training time. To solve these problems, the researchers proposed the stochastic depth algorithm, which uses a short neural network at the training stage and a deep one at the testing stage. To do this, a deep ResNet architecture is created with more than a thousand layers, but at the training stage it is reduced by randomly deleting a significant proportion of layers independently for each piece of data (Figure 4). This approach results in a network with a small expected depth at the training stage, but a large depth at the testing stage. It is experimentally revealed that training with stochastic depth significantly reduces the training time and error on test data. The reduction in training time can be explained by a shorter neural network architecture, which reduces the number of calculations in the learning process. The reduction in error on test data is due to two factors:

- reducing the depth during training reduces the chain of steps of forward propagation and gradient calculation, which helps to solve the problem of gradient disappearance;
- networks trained by this method can be interpreted as an ensemble of networks of different depths [8-10].
4. Conclusion

In the course of a review of studies on image analysis using convolutional neural networks, it was experimentally revealed that the use of the stochastic depth algorithm allows you to reduce the network operation time at the testing stage without losing accuracy. The main idea is to train a neural network with a relatively small number of parameters based on a pre-trained model, which will predict for each block, depending on the input image: use it or exclude it. This architecture is trained in such a way as to maximize the reward, which encourages the use of as few blocks as possible while maintaining the accuracy of the prediction. To do this, the target function takes into account the percentage of blocks left and the accuracy of the original network, provided that the predicted blocks are removed.

5. Acknowledgments

The research was conducted as a part of the grant of the President of the Russian Federation MK 592.2020.11.

References

[1] Hunt A J and Black A W 1996 Unit selection in a concatenative speech synthesis system using a large speech database IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings 1 373-376
[2] Tokuda K 1995 An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features Fourth European Conference on Speech Communication and Technology 757-760
[3] He K, Zhang X, Ren S and Sun J 2015 Deep residual learning for image recognition arXiv:1512.03385v1
[4] Van Den Oord 2016 A WaveNet: A generative model for raw audio SSW 125
[5] Wang Y 2017 Tacotron: Towards end-to-end speech synthesis arXiv:1703.10135
[6] Wilber M, Veit A and Belongie S 2016 Residual Networks Behave Like Ensembles of Relatively Shallow Networks arXiv:1605.06431
[7] Huang G, Sun Y, Liu Z, Sedra D and Weinberger K Q 2016 Deep networks with stochastic depth arXiv:1603.09382
[8] Wu Z 2018 Dynamic inference paths in residual networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 8817-8826
[9] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions arXiv:1409.4842v1
[10] K E Tokarev 2021 J. Phys.: Conf. Ser. 1801 012030
[11] K E Tokarev 2021 J. Phys.: Conf. Ser. 1801 012031