STRUCTURAL APPROACH TO THE DEEP LEARNING METHOD

L.A. Sevastianov[1,3], A.L. Sevastianov[1], E.A. Ayrjan[2], A.V. Korolkova[1], D.S. Kulyabov[1,2], I. Pokorny[4]

[1]*Department of Applied Probability and Informatics, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

[2]*Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russian Federation*

[3]*Bogoliubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russian Federation*

[4]*Technical University of Košice, Košice, Slovakia*

E-mail: kulyabov-ds@rudn.ru

According to the modest opinion of the authors, data science, machine learning, neural networks are extremely relevant areas of science. Around them there was an expectation that attracted lovers of quick results. This does not detract from the importance and benefit of these areas of science. In the paper, the authors took an attempt to briefly describe their experience in organizing work with data analysis using neural networks and machine learning methods.

Keywords: machine learning, neural networks, data analysis

Leonid Sevastianov, Anton Sevastianov, Edik Ayrjan, Anna Korolkova,
Dmitry Kulyabov, Imrikh Pokorny

1. Introduction

Historically, the term big data is directly related to machine learning. The increase of interest in big data appeared in the second half of the 2000s. It was primarily associated with the advent of the practical ability to collect, store and process really big data, because without big data the stable operation of the algorithms that existed at that time (standard machine learning algorithms: SVM, Random Forest, etc.) was impossible. But the most important reason of increase interest in big data was the appearance of an ability to monetize big data algorithms, which were used primarily to study customer preferences and behavior.

The period of special popularity began with the advent of neural networks. Namely, convolutional neural networks have won a number of image recognition contests [4], because convolution networks made it possible to solve the problems of classifying images and objects in an image, detecting objects, recognizing objects and people, improving images [3, 5, 6]. The next step was recurrent networks, since word processing algorithms, speech to text conversion, machine translation systems have become the area of recurrent networks application.

Machine learning technology issues

It seems to us that at the present moment, machine learning methods, and especially methods based on neural networks, are a source of unnecessary expectations. They became a kind of new silver bullet [1]. In fact, most of the problems, witch now are solved with the help of machine learning methods and neural networks have been solved for a long time. Undoubtedly, neural networks have indicated some progress in areas such as image recognition and machine translation. However, the limits of their applicability were reached very quickly. In addition, it was not possible to find any fundamentally new applications, especially allowing monetize the results. Actually, this, in our opinion, is the reason for the presence of a large number of qualitatively free tools: Tensorflow [https://www.tensorflow.org/], PyTorch [https://pytorch.org/], Keras [< https://keras.io/>], MXNet [https://mxnet.apache.org/], Microsoft Cognitive Toolkit [https://docs.microsoft.com/en-us/cognitive-toolkit/], Flux [https://fluxml.ai/].

The hope that it is possible to create a technology that can solve any scientific and technical problems has led to a sharp decrease in the level of culture of analytical thinking. The profession of a data specialist (Data Scientist, Data Engineer) has become sharply popular also because the magic flair is created around this specialty. A typical data specialist does not proceed from a problem, but from a method. They promise to solve all the problems. They do not ask anything. What for? After all, there is data. The data will tell you everything. The more data, the better. In this case, some arbitrary data is used, if only there would be more of them. As a result, a certain model is constructed, according to which calculations are carried out. The result is often uncheckable. All this is fueled by various online competitions, such as Kaggle [https://www.kaggle.com/].

The time of universal specialists has passed. The following competencies are required from a real data scientist: an analytical approach, a culture of working with data, the ability to put forward hypotheses.

2. Data analysis project structure

Starting data specialists think that working with data is a continuous creation. They believe that they will be involved in complex but interesting projects. At the same time, it is enough for them to learn the skills of working with the appropriate software. In fact, this is a hard routine work requiring high concentration and a large theoretical and practical background. The vast majority of the work of any data specialist is extremely far from creativity in its romantic sense. This can only be achieved by experience and nothing more.

Based on our experience, we will describe what a typical data analysis project represents. Undoubtedly, everyone who is seriously involved in this field will say the same. Unfortunately, this information is considered so obvious that experts are in no hurry to codify it. We tried to close this gap, however, by no means claiming originality.

2.1. Theoretical preparation

The first step is to formulate the requirements for the new project. Most likely, we initially do not know anything about what data we should have. To understand this, you need to follow a few simple steps:

- We must delve into the problem statement, clearly define the purpose of the study, understand what is the problem.
- We must understand what result is required to get from the project, choose the metrics by which the project result will be evaluated.
- We must decide by what method the problem can be solved. In fact, this is the most important step determining the next steps. The choice of approach depends on what type of result you need to get at the end. For example:
  - if we need a yes / no answer,then a Bayesian classifier is suitable;
  - if we need an answer in the form of a numerical sign, then regression models are suitable;
  - if we need to determine the probabilities of certain outcomes, then a predictive model is suitable;
  - if we need to identify relationships, then a descriptive approach is used;
  - and so on. In fact, knowledge of these approaches and the ability to navigate them is a necessary feature of a data analysis scientist.
- We must set the data requirements, the format of required data, select a suitable data source (if it is possible).

2.2. Data handling

After theoretical preparation, we should, in fact, begin to work with data. Data needs to be collected, analyzed and prepared for the study. Data collection, data analysis, data preparation – all this takes approximately 70%–90% of the entire project time. At the same time, this stage is very far from any kind of romance.

So, let's start collecting data from existing sources and form a sample with which we will continue to work. To do this, we must follow a series of simple steps.

- We must make sure that our data sources are accessible, reliable and of high quality.
- We need to understand if we were able to get the data we wanted.
- Based on this, we should perhaps revise our data requirements.
- Perhaps, we will have to decide if we need some additional data.

Then we need to check the quality of the data and the properties of the received data. To do this, one should study the statistical characteristics of the data, find the distributions that they satisfy, identify possible correlations (to eliminate redundancy), check the data for emissions:

- the central position is studied (middle, median, mode);
- emissions are searched for and variability is estimated (variance, standard deviation);
- histograms of the distribution of variables are built.

And finally, we must prepare the data for further work:

- we remove duplicates;
- we process missing or incorrect data;
- we check and correct formatting errors.

Actually, at this stage, the extraction and selection of features with which our model will work is performed. Errors at this stage can be critical. If we ask an excessive amount of features, the model can be retrained. If we ask an insufficient number of features, then the model may be untrained.

2.3. Project development

When the type of model is defined and there is a training sample, we develop the model and test it on a set of features. Due to the large amount of available software, the software implementation of the model is currently the least problem.

Now the cycle begins (we hope the final one) of working with the model:

- run training set;
- testing on a test sample. It is necessary to check whether the model works as intended;
- verification of the result. It is necessary to make sure that the data in the model are correctly used and interpreted and the result obtained does not go beyond the statistical error;
- and now we can start all over again.

2.4. Project implementation

The implementation phase is especially highlighted. In comparison to all the previous steps,this phase looks extremely boring. And so it is often omitted. As a result, the project result is not a complex of programs, but a set of disparate scripts. And the result is that research can be of use, remains unknown and worthless.

This stage involves the preparation of self-sufficient software modules, their placement in public repositories, the presence of intelligible documentation. And most importantly, it is necessary to maintain feedback with potential users of the project.

3. Conclusion

Having witnessed more than one wave of the increased popularity of technology [2], the authors think that the excitement around big data is on the decline. At the same time, there is a danger that the technology, which is acutely fashionable now, may lose its adherents in the future. But machine learning technologies, in spite of fashion, have very definite scientific applications. Therefore, the authors see the purpose of this article in enlightening novice researchers who begin to engage in science.

Acknowledgments

The publication has been prepared with the support of the ``RUDN University Program 5-100'' (Leonid A. Sevastianov, Anton L. Sevastianov, Anna V. Korolkova, Dmitry S. Kulyabov). The reported study was funded by Russian Foundation for Basic Research (RFBR), project number 19-01-00645 (Edik A. Ayrjan).

References

[1] Brooks, F.P.J. 1986. No silver bullet-essence and accidents of software engineering. Proceedings of the IFIP Tenth World Computing Conference. (1986), 1069–1076.

[2] Fenn, J. and Raskino, M. 2008. Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Harvard Business Press.

[3] He, K., Zhang, X., Ren, S. and Sun, J. 2016. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016-Decem, (2016), 770–778.

[4] Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2, (2012), 1097–1105.

[5] Simonyan, K. and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. (2015), 1–14.

[6] Zeiler, M.D. and Fergus, R. 2014. Visualizing and understanding convolutional networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 8689 LNCS, PART 1 (2014), 818–833.