ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse113932020.pdf
https://doi.org/10.30534/ijatcse/2020/113932020

Neural Network System for Processing Large-Volume Diagnostic Data

Zavalko N.A.[1], Krasyukova N.L.[2], Plotitsyna L.A.[3], Gladyshev A.G.[4], Boyko A.N.[5]
[1]Financial University under the Government of the Russian Federation, 125993, Russia
zavalko@mail.ru
[2]Financial University under the Government of the Russian Federation, 125993, Russia
krasyukova@mail.ru
[3]Financial University under the Government of the Russian Federation, 125993, Russia
plotitsyna@mail.ru
[4]Financial University under the Government of the Russian Federation, 125993, Russia
gladyshev@mail.ru
[5]Financial University under the Government of the Russian Federation, 125993, Russia
boykoan@mail.ru

ABSTRACT

Today, big data and its Analytics are effectively used in many areas. In information technologies, the results of analysis can be used for security purposes. Based on customer identification models, you can improve the quality of service. Online transactions help detect fraud, risk management in business and trade predicts risk factors, astronomy has the ability to learn more about the universe, the health care system provides better services, and so on.

Over the past decades, the volume of information in the health system has increased significantly, which has led to it being considered big data. This industry has historically generated a huge amount of data related to record keeping and patient care, enabling a wide range of medical functions, including clinical decision support, sensor-based health and food safety monitoring, disease surveillance and population health management, etc.

For example, a cancer diagnosis requires petabytes of data from various sources to determine the condition of the disease and the patient's survival potential. In addition, today the use of information technology for major medical services reduces the cost of healthcare, improving its quality through preventive and personalized care and continuous monitoring. In this context, savings in the US alone reach $ 300 billion annually.

Key words: Neural network, diagnostic data, processing, data analyze.

1. INTRODUCTION

The U.S. healthcare network includes more than 2,700 organizations, hospitals, and Wellness systems, 90,000 outpatient clinics, and 400,000 doctors. A large database of clinical and financial information has been accumulated with data on both patients and deliveries. Based on the data collected, reports on clinical trials, resource utilization, and so on were obtained that allowed better decisions to be made and improved treatment processes in approximately 330 hospitals, while saving approximately 29,000 lives and reducing healthcare costs by almost $ 7 billion.

Telemetry has become particularly relevant – remote study of various processes, objects or phenomena by measuring their parameters and transmitting this information over a distance.

In medicine, telemetry is used to register and transmit information about a patient's condition over a distance:

- a patient who is outside of a treatment or consultation center;
- an athlete in the process of training;
- monitoring the health status of astronauts working in orbit;
- control of the physiological functions of the patient's body in the pressure chamber, cardiology or intensive care unit, etc.

If in clinical diagnostics the main actions with medical equipment are performed by the doctor and medical technical personnel together, then in telemetry procedures, which also form the basis of diagnostics, the actions of the doctor and the engineer can be separated.

Thus, it is relevant to create an intelligent analysis of medical data, thanks to which it is possible to aggregate and analyze the information obtained from heterogeneous sources of different types of information, including numerical, graphical and textual data for the purpose of decision-making.

2. MATERIALS AND METHODS

A medical image is a structural and functional image of human organs intended for the diagnosis of diseases and the study of anatomy-physiological characteristics of the body. Sometimes it is also called a diagnostic image.

The typical composition of medical data accumulated by a medical institution, text, graphic, and numeric data of the following types [1-3]:

- laboratory data;
- instrumental and visualization methods of diagnostics;
- indicators of comprehensive clinical examination.

Based on this data, we analyze:

- skin;
- respiratory system;
- cardiovascular system;
- system of digestive organs;
- urinary system;
- musculoskeletal system;
- endocrine system;
- nervous system and sense organs.

Based on the composition of the output data, the tasks of medical data mining are defined:

- search for generated query data stored in relational form;
- classification of patients ' symptoms to solve the problem of primary diagnosis;
- parallel and distributed processing of large data sets.

The following requirements are put forward for the system under development: low cost; geographical distribution of resources; large amount of heterogeneous information; scalability. The main purpose of the developed system is to monitor the state of human health on the basis of accumulated images and sensor indicators [59].

The composition and relationship of components of a neural network expert system is expressed as the conversion of input values into output values:

$$S \subset X \times Y \quad (1)$$

Where, $X(i) = \left[x_1(i), x_2(i), ..., x_n(i)\right]$ - the survey vector, $Y = \left\{y_1, y_2, ..., y_Q\right\}$ - is a set of classes.

Thus, the state of a person based on the survey vector is classified by the system for a given number of reference images, which include physiological functions.

For the initial values, a set of problems is constructed, the solution of which belongs to the set $D_X = D_{\text{interview}} \cup D_{\text{photo}} \cup D_{NN}$, where $D_{\text{interview}}$ - the task of creating a survey vector, D_{photo} - the task of processing images of a human, D_{NN} - the task of making a diagnosis using a neural network.

Reflection $T : X \to Y$ allows you to find for each $X(i)$ find $y_j \in Y$ ($j = \overline{1, Q}$ number of classes), which is the solution to the problem D_X .

The topic of unbalanced classification is becoming increasingly popular. Unbalance occurs when the classes presented in the task show a skewed distribution, that is, there are smaller (minority) and larger (majority) classes [4-7]. This means that the data set elements are unevenly distributed among classes. This class imbalance can manifest itself in different areas of activity, such as microanalysis, medical diagnosis, oil reservoir recognition, or intrusion detection systems. Most algorithms simply calculate accuracy based on the percentage of correctly classified observations. However, if distorted distributions are studied, the results may be unreliable, because minority classes have minimal impact on overall accuracy [8-10]. As a result, performance in the majority class can overwhelm poor performance in the minority class. In addition, the emergence of Big Data has created new problems related to processing speed, reliability and diversity of data. One approach to solve this problem is to use various sampling strategies (a set of techniques for splitting the initial sample into working and control areas with certain rules), which can be divided into two groups: random and special.

In [6], three possible schemes are considered: sampling majority class examples, resampling minority examples, and using hybrid methods. Random resampling removes instances from the majority class until the class distribution is fully balanced.

More complex methods are proposed based on the generation of artificial samples. The idea of the method is to reprogram the minority (SMOTE) using the sampling method [9] to generate new examples, by interpolating between several minority data that are located next to each other.

On the other hand, this can lead to the problem of over-generalization. Various adaptive sampling methods have been proposed to solve the problem of over-generalization and increase the representation of the minority. These methods are mainly intended for balancing data before the training stage, however, the presence of additional characteristics may impair performance [11].

Based on the results of this work, we selected popular methods for pre-processing and classifying data. The MapReduce work environment is the most common framework for providing reliable and scalable Big Data solutions. To use this scheme, any algorithm must be divided into two main stages: "Map" and "Reduce". The first one is dedicated to splitting data for processing, while the second one collects and aggregates results.

Within the framework of the theory of building cloud systems, a special place is occupied by the issue of organizing traffic routing in virtual networks. In most cases, routing functions in the PaaS environment are performed by applications that are also virtualized according to the concept of SaaS solutions. One of the tasks in implementing cloud computing structures [12], in particular meta-computing and SuperServer virtualization systems, is the rapid transfer of large data arrays. The task is more complicated if the physical systems that the cloud architecture is built on are separated, and switching is implemented using two or more intermediate nodes. Therefore, it is important to minimize the interaction time of computing nodes in overlay networks when transmitting large data arrays.

The method is based on entering and using the intermediate data storage function as SaaS routers. In order not to reduce the reliability of communication networks, these routers must be placed on physical routers.

Due to this, high-bandwidth communication channels will be used. As a result, the stream type of traffic will still be transmitted, while the elastic type will be transmitted via intermediate storage systems, if necessary. In fact, the locations of data storage systems are justified by the limits of low-speed sections of the transmitted data routes, especially the limits of the last-mile channels. At the same time, there may not be a need for additional intermediate data storage systems for transmitting large amounts of data on the backbone sections of communication networks. The developed method provides end-to-end streaming data transmission and more efficient transmission of large data arrays by implementing storage systems at the above points in overlay networks.

Further research should be devoted to the development of methods for optimizing the parameters of the developed method and methods for adapting these parameters to real data transmission conditions, taking into account the dynamics of changes in the global Internet network with a high level of separation of the overlay network.

Within the framework of this task, we consider several variants of models for the functioning of multilets within the framework of the system described above.

1. a Multilet is a transport of a data storage device. A special feature is the physical contact of the multilet with the data storage device. In terms of implementation, it is somewhat similar to the "Hummingbird" principle, where accuracy plays an important role. On the one hand, this option is labor – intensive, and on the other-quite energy-efficient in the case of creating an optimal robotic solution. The algorithm of functioning of the multicopter consists of these steps.

Step 1. Multinet starts moving on a predetermined route after the assessment of environmental factors.

Step 2. Multilet moves to the point defined in its geo-positioning system with the necessary maneuvers to be able to capture and / or install the data carrier of the sensor network storage system.

Step 3. Multilet is forwarded to another collection point/installation media of the storage system.

Step 4. If the route point is the last, the multicopter returns to the starting position. If this version of the model is used, the following problems arise that need to be solved:

- in the case of taking the last data carrier of the system for storing information in the sensor network, where should the new data be stored;
- how can we implement a mechanism for retrieving data storage devices in practice, taking into account the presence of a MEGASYSTEM with a large number of sensor network systems;
- how to implement a mechanism for exchanging media at the starting point within the automation of the proposed model.

2. Multilet is a data carrier by collecting data via a wireless or wired communication channel from a sensor network data acquisition device. There is a big difference in the use of a leading and wireless method of collecting information. Thus, the leading method has all the disadvantages in terms of the complexity of the management and implementation model. In contrast, a wireless data transmission method with a holographic directional pattern of transmitting systems can provide data transfer speeds of up to 7 GB / s (IEEE 802.11 ad standard). In the case of a pie chart, the data transfer speed may be significantly reduced. The algorithm of functioning of the multicopter and TS and in in the case of switched data collection consists of these steps.

Step 1. Multinet starts moving on a predetermined route after the assessment of environmental factors.

Step 2. Multilet moves to the point defined in its geo-positioning system and performs the necessary maneuvers to be able to connect to the sensor network storage system.

Step 3. the process of transferring data from the sensor network data storage system to the storage system itself, the multilet.

Step 4. Multilet is routed to another data collection point in the storage system of another sensor network segment.

Step 5. If the route point is the last, the multicopter returns to the starting position.

The algorithm of functioning of the multicopter and TS-in in the case of wireless data collection consists of these steps.

Step 1. Multinet starts moving on a predetermined route after the assessment of environmental factors.

Step 2. the Multilet moves to the point defined in its geo-positioning system and performs the necessary maneuvers to adjust the wireless data transmission system, which is based on the use of antenna systems with a needle pattern of the signal.

Step 3. the process of transferring data From the sensor network data storage system to the storage system itself, the multilet.

Step 4. Multilet is routed to another data collection point in the storage system of another sensor network segment.

Step 5. If the route point is the last, the multicopter returns to the starting position. If you use this version of the functioning model, the following problems arise that need to be solved: what standards and protocols should be used in terms of energy efficiency and speed for collecting information by a multicopter; if you use wire switching, how to implement this model in practice.

3. Multilet is a data carrier by collecting data via a wireless communication channel as a result of moving the data acquisition device in the sensor network. In this case, it is more appropriate to use IEEE 802.11 AC data transfer technologies. When moving a multilet near a sensor network storage device all data must be transferred to the multilet storage systems.

Neural network training is a complex part of deep learning, because it requires a large set of data and a large amount of computing power. Previously, it was expensive, difficult and long - needed powerful GPUs, video cards and memory. The boom in deep learning is due to the widespread availability of GPUs that speed up and reduce the cost of computing, virtually unlimited data storage capabilities, and the development of "big data"technology.

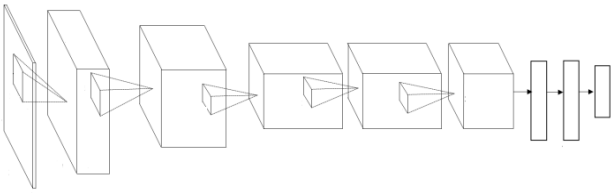

Figure 1: DNN architecture

Multi-layer DNNs are difficult to teach the standard gradient descent method. However, modern high-performance computers, in particular parallel GPUs, allow for faster DNN learning by one order of magnitude compared to serial code for standard computers (figure 1). A neural network has a two-dimensional layer of winning neurons whose weights are shared. A simple maximum pool technique [6] identifies the winning neurons by dividing the layers into quadratic regions of local inhibition, selecting the most active neuron from each region. The winners represent the first layer with reduced sampling and lower resolution, which transmits data to the next layer.

The following model of a parallel convolutional neural network is implemented (Fig. 2). First level includes eight consecutive layers of nonlinear neurons.

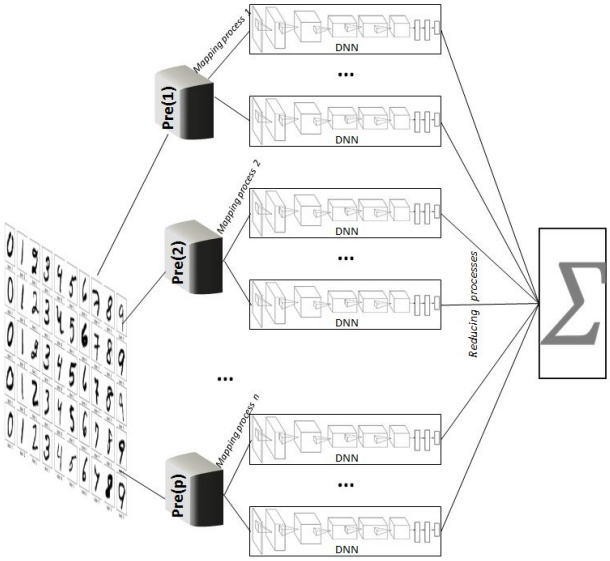

Figure 2 - model of a parallel convolutional neural network

Random DNN weights are iteratively trained to minimize classification error on a set of labeled training images. The generalization performance is then tested on a separate set of test images. The proposed model combines several approaches.

DNN having two-dimensional layers of neurons consists of a sequence of convolutional and maxpooling layers. Each layer, which has a lot of cards, only receives connections from the previous layer. This hierarchical feature allows you to separate the pixel intensity of the raw input image into a vector view, which is classified by two or three fully connected layers. All parameters are adjusted by minimizing classification errors with the training set. The maximum pooling layer determines the gain of neurons by dividing the layers into quadratic regions of local inhibition, selecting the most active neuron of each region. The winners of some layer represent a smaller, lower-resolution sample layer, feeding the next layer in the hierarchy.

At some point, lowering the selection automatically results in a 1-dimensional layer. From now on, only trivial one-dimensional winner regions are possible, "winner gets everything", i.e. the middle part of the structure becomes a standard multi-layer perceptron (MLP). In these places, the DNN becomes minimal, for example, only 2x2 or 3x3 neurons.

3. CONCLUSION

Many tasks of image processing, analysis, and recognition involve automating medical diagnostic processes, which include classifying images to prepare a preliminary diagnosis, making it easier for the doctor to make a final decision. The main part of research in the field of deep learning is aimed at achieving the highest possible accuracy of image classification and recognition by deep convolutional neural networks (DCNN). However, not enough attention is paid to the problem of DCNN learning time.

Two sets of data were used for computational experiments.

1.to solve the problem of recognizing the skin disease of melanoma, the neural network Inception-v3 was used, which uses the decomposition of the nxn filter by two consecutive filters 1xN and Nxe1. Implemented a distributed selection of functions in large medical data using the Hadoop MapReduce platform.

The neural network was trained on the recognition of 2000 photos of skin formations to distinguish benign from malignant tumors using the portal of International cooperation in the field of skin imaging (ISIC 2017: analysis of skin lesions on the way to detect melanoma). The use of the proposed approach for the diagnosis of skin diseases provides significant support in the diagnosis of both dermatologists and General practitioners.

2.the MNIST (Modified National Institute of Standards and Technology) database is used to solve the problem of digit recognition. IT is a large database of samples of handwritten numbers. The database is a standard proposed by the US National Institute of standards and technology for the purpose of calibrating and comparing image recognition methods using machine learning primarily based on neural networks.

For experimental evaluation of the selected parallel programming environments, we used IntelXeonE5-26502 v4 and GeForceGTX 1070 GPU-24 streaming multiprocessors (sm), 1920 CUDA cores with a frequency of 1.5 GHz.

REFERENCES

1. L. Huang and L. Wang, "**Accelerated Monte Carlo simulations with restricted Boltzmann machines**," Phys. Rev. B 95, 035105 (2017).

2. N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer. (2015) **SMOTE: synthetic minority over-sampling technique**. J ArtifIntell Res 16:321–357. https://doi.org/10.1613/jair.953
3. H. He, Y. Bai, E. García, S. Li. (2008) **ADASYN: adaptive synthetic sampling approach for imbalanced learning**. In: Proceedings of the 2008 IEEE international joint conference neural networks (IJCNN'08), pp 1322–1328.
4. C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap. (2012) **DBSMOTE: density-based synthetic minority over-sampling technique**. ApplIntell 36(3):664–684. https://doi.org/10.1007/s10489-011-0287-y
5. P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "**Machine-learning-assisted materials discovery using failed experiments**," Nature 533, 73–76 (2016). https://doi.org/10.1038/nature17439
6. A. Paszke et al. **Enet: A deep neural network architecture for real-time semantic segmentation** //arXiv preprint arXiv:1606.02147. – 2016.
7. G. Levi, T. Hassner. **Age and gender classification using convolutional neural networks** //Proceedings of the iEEE conference on computer vision and pattern recognition workshops. – 2015. – C. 34-42. https://doi.org/10.1109/CVPRW.2015.7301352
8. D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. **Flexible, high performance convolutional neural networks for image classification**. In International Joint Conference on Artificial Intelligence, pages 1237–1242, 2011.
9. G. Carleo, C. Ignacio, C. Kyle, D. Laurent, S. Maria, T. Naftali, V.-M. Leslie, and Z. Lenka, "**Machine learning and the physical sciences**," arXiv:1903.10563 (2019). https://doi.org/10.1103/RevModPhys.91.045002
10. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, "**All optical neural network with nonlinear activation functions**," arXiv: 1904.10819 (2019). https://doi.org/10.1364/OPTICA.6.001132
11. Z. Habybellah, M. Bassiri, S. Belaaouad, M. Radid&S. Benmokhtar. (2019). **Training and development of professional skills: An analysis of activity in professional skills.** International Journal of Advanced Trends in Computer Science and Engineering, 8(5), 2029–2033. https://doi.org/10.30534/ijatcse/2019/28852019
12. S. Hachmoud, A. Hachmoud, A. Meddaoui& H. Allali. (2019). **Analysis of students online learning behavior in a pedagogical model combining blended learning and competency based approach**. International Journal of Advanced Trends in Computer Science and Engineering, 8(6), 3389–3395. https://doi.org/10.30534/ijatcse/2019/113862019