 doi:10.20944/preprints202005.0274.v1

Article

Big Data system for medical images analysis

Janusz Bobulski [1,‡] 0000-0003-3345-604X and Mariusz Kubanek [1,‡] 0000-0001-9651-9525

[1] Czestochowa University of Technology, Department of Computer Science, Czestochowa, 73 Dabrowskiego Str., Poland;
januszb@icis.pcz.pl, mariusz.kubanek@icis.pcz.pl
* Correspondence: januszb@icis.pcz.pl
‡ These authors contributed equally to this work.

Abstract: Big Data in medicine includes possibly fast processing of large data sets, both current and historical in purpose supporting the diagnosis and therapy of patients' diseases. Support systems for these activities may include pre-programmed rules based on data obtained from the interview medical and automatic analysis of test results diagnostic results will lead to classification of observations to a specific disease entity. The current revolution using Big Data significantly expands the role of computer science in achieving these goals, which is why we propose a Big Data computer data processing system using artificial intelligence to analyze and process medical images.

Keywords: Big data, deep learning, intelligent systems, medical imaging, multi-data processing.

1. Introduction

The demand for solutions offering effective analytical tools has been increasing in recent years. This trend can also be observed in the field of Big Data analysis. Almost all sectors are interested in these solutions, although undoubtedly business organizations excel in the use of this type of analysis. However, it can be seen that the healthcare sector is also more likely to use these solutions. Theory and practice show that Big Data analytics in this sector can contribute, among others to: improve patient care, designate and implement appropriate paths (methods) for patient treatment, support clinical treatment and diagnostic support. However, Big Data is also associated with a certain type challenges in the form of complexity, threats to security and privacy as well as the demand for new technologies and human skills [1,2].

In Big Data there is a big problem with the integration and homogeneity of data, because they come from different sources and there are different formats, for example, the format of dates in the US and in Europe, whether upper and lower case letters in the name of their own. This results in the need to control their condition [3]. The key factor here is the correctness and quality of the data, because in the case of their lack the results of the system will be unsatisfactory, according to the principle: garbage at the input-garbage at the output. Therefore, the system should be equipped with a data control and conditioning module, analogous to signal processing systems where the input signal is filtered to eliminate interference [4,5].

In modern IT, we have two main trends, big data and deep learning. The development of these fields will significantly affect technological progress. By adding Computer Vision to this you can talk about the systems of the future. That is why we decided to develop a system for analyzing large volumes of images using artificial intelligence for the needs of medicine [6].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2020 doi:10.20944/preprints202005.0274.v1

2. Big Data architectures

2.1. Lambda architecture

Lambda architecture is a popular architecture used in Big Data systems. It allows simultaneous access to large data sets and their parallel processing. The main feature of the Lambda architecture is the presence of two identical data streams (Fig 1), where one is processed in real time and the other in batch mode [7].
In real time mode, data is processed continuously, and thanks to the short time of data access it is possible to quickly search for information. In this case, access to historical data is not possible and not all operations are possible. The quality and reliability of data in this mode is lower. Batch mode is more reliable, however, due to its longer processing time, the real mode allows you to process data in real time.
In batch mode, calculations are made for the entire data set and take much longer, but the data received is of high quality and contains a full history. This dataset has an indivisible form that only needs to be expanded without removing data from it. This ensures data consistency and access to historical data. Data views are created based on real and batch data in access mode. Their aggregation causes them to bond in such a way that they are visible as a whole. Views give the opportunity to perform various ad hoc analyzes while providing quick access to data.
The idea of Lambda architecture provides a compromise between batch processing and real-time processing. The biggest disadvantage of this solution is the need to maintain two independent applications - one for feeding the batch layer and the other for the real-time layer. The tools used in each layer are different, hence the need to use different solutions for each mode, which makes this architecture more complicated and more expensive to maintain.

Kappa architecture is an alternative without the main disadvantage of Lambda cite 9 architecture. Its idea results from four main assumptions:

1. Everything is a stream that any data source can generate.
2. The data is immutable and can be reused at any time.
3. The KISS - Keep principle is short and simple.
4. You can restore the data state at any time.

The data in the stream must remain unchanged and original, otherwise you will not be able to get consistent calculation results. In fig. 1 we can see the structure of the Kappa architecture. We have here a real-time mode and an access mode that performs the same functions. There is no Batch layer mode here, which has become redundant, because the story can be restored at any time. Because of these advantages, we chose this architecture for our system.

2.2. Elements of the system

The general scheme of the system is shown in the figure 2. Main parts of the systems are:

1. System: A system for processing any type of data that is able to handle the system, having its own language and grammar, can learn, is intelligent and able to find solutions to new problems; the goal of the system is to gather knowledge, not data.
2. User: a person, device or system having the ability to communicate with the system,
3. Data: a set of information in the form of a data stream as well as a file.
4. External commands: queries from the user in natural language.
5. Methods library: a set of functions that will work on a set of external data provided to the system, based on algorithms of artificial intelligence, deep learning, application algorithms, data processing and analysis; this library will be used by the processing unit.
6. Pre-processing: recognition and interpretation of input data; recognition of data forms and their processing to the necessary internal form with the analysis of correctness and error signalling.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2020 doi:10.20944/preprints202005.0274.v1

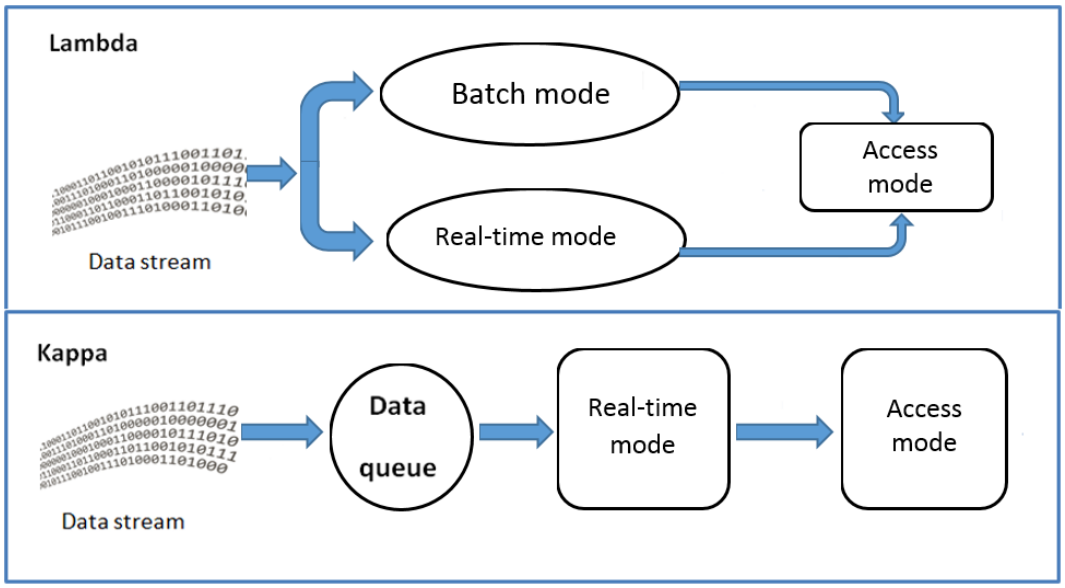

Figure 1. Comparison of Lambda and Kappa architectures.

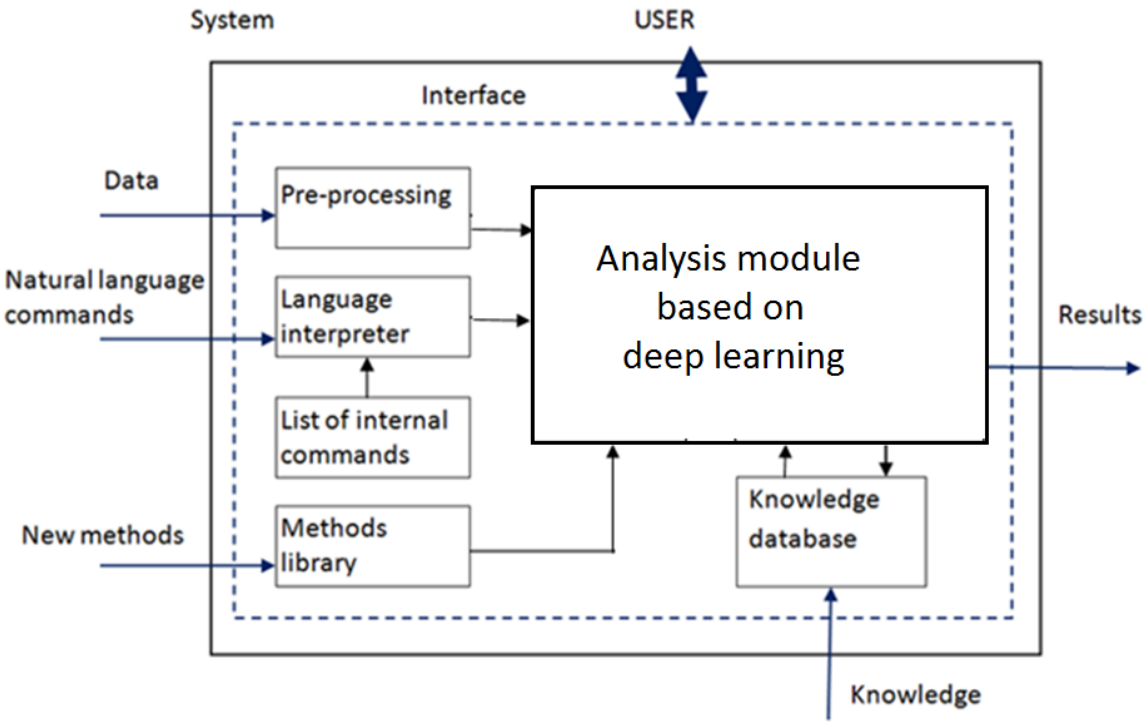

Figure 2. Architecture of propose system.

 doi:10.20944/preprints202005.0274.v1

7. Language interpreter: program that verifies the correctness of external commands, intelligent, interactive parser and generator of executable code or an internal language suitable for the system to perform in the processing.
8. List of internal commands: a set of internal system instructions that are not available to an external user.
9. Knowledge database: Knowledge of the system stored in accordance with specific rules; objects and relations between them introduced from outside or deduced by the system and stored in the system in graph database.
10. Analysis module: based on deep learning and artificial inteligence.
11. System output: results - a data stream obtained in the results of data processing by the system.

Preliminary analysis of data should lead to the formulation of only such statistical statements regarding data for which the value of "true" is the invariant of correct transformations, e.g. median, mean or range. Therefore, the pre-processing unit should control the quality of data, because according to the GIGO principle (garbage in - garbage out) - entering erroneous data results in erroneous results - conclusions. Poor quality data make it difficult to draw correct conclusions and, as a result, knowledge exploration and rational decision making. Loaded data and dependencies derived from them can have serious consequences when it comes to formulating laws and rules. Pre-processing should include data cleaning and transformation to prepare for exploration. It is estimated that the initial data processing is 70-80% of the knowledge discovery process. Data after verification of, for example, the range, will be converted to the internal format of meta-based tags. Thanks to this we will obtain a unified data structure. An additional benefit of this stage of data processing will be their standardization. A detailed description of the data format will be developed in further works and parallel processing of multiple data streams is also planned.

3. Data model for knowledge database

Database model is understood as a set of rules that describe the structure of data in a given database. Allowed operations or data structure are defined by specifying representations of entities and relationships allowed in a given model. In the context of Big Data processing, NoSQL (Not only Structured Query Language) databases that have no SQL database (Structured Query Language) restrictions apply.

The graph model of data representation is a model with labelled and directed multi-graphs, which contains attributes. Labelled, because it has labels for all edges. Directed, because it has edges with certain direction (from the source to the end node). The presence of attributes in the graph causes the attribute list variable to be assigned to each node and edge. The attribute is the value associated with the name.

A multigraph can have multiple edges between two nodes. Therefore, different edges can connect two nodes many times, without paying attention to whether these edges have the same source, end node and label.

In GBD, graph is a native form of information storage. It is often stated that GBD provides index-free neighbourhoods. In GBD, information about vertex neighbours is stored locally. For comparison, in RBD there would be a reference to an index that would show the vicinity of the vertices. Therefore, in the graph option, the neighbourhood search time depends on the number of vertices neighbours. In the relational option, however, on the number of all edges [11,12].

Semi-structured data is often modeled as large tables, with lots of columns, blank for most rows, which significantly reduces performance. The alternative in the form of modeling this data, with the participation of many table connections, is also not an efficient solution, considering the costs of joins during queries. In addition, normalization of graph structures in RBD degrades performance during queries. Such a decrease in performance is related to the recursive nature of e.g. file trees or social networks, as well as to the form of recording data as relationships. Each operation performed on the

 doi:10.20944/preprints202005.0274.v1

edge of the graph gives the effect of joining between tables in a relational database. This is a very slow operation, which in addition is not scalable [13]. GBD support a graph model that allows direct storage of specific objects and relationships between them in a database. GDB should allow access to query methods that cope not only with stored objects, but with the graph structure itself. The best known example is traverse, which in its simplest form can be used to obtain adjacent vertices in the graph. The use of the graph database in the propose System has the additional advantage of being able to search graphs in depth and breadth [14].

The use of GDB technology will result in the creation of a semantic data network. The semantic net is a graph in which individual elements have their meaning. A semantic graph is not a specific product, specification or standard. It's more of an idea or vision than technology. The goal of the semantic graph project is to make data available for processing: people and machines so that they can be used not only for display purposes, but also for automation, integration and reuse in many different applications, such as intelligent agents. They will use distributed databases in the form of semantic networks. This in turn will allow the creation of an automatic infrastructure that, if properly designed, will make a significant contribution to the evolution of human knowledge [15]. In addition, the semantic web enriches the current web with annotations written in a machine-process able language, which can also be associated with each other [16].

4. Analysis module on the base deep learning

In connection with the growing need to provide remote diagnostics systems, new methods are being sought to enable automatic performance of tasks that until now were only possible for people, e.g. speech understanding, object recognition or the entire image context. One of these techniques is deep learning. This rapidly growing field of machine learning, based on deep (having many hidden layers) neural networks, has become an indispensable tool that allows computers to solve problems of perception of the world around us. Deep learning is a technique known for many years. The earliest deep learning models, consisting of many layers of nonlinear features, date back to the 1960s. In 1965, Ivakhnenko and Lapa [17] published the architecture of the first deep one-way (feedforward) neural network based on polynomial activation functions. Unfortunately, in those days models based on deep learning were not effective. Only recently has deep learning become a key approach used, among others, in image recognition. One of the reasons for this progress is the ability to perform calculations on larger models, thanks to the availability of faster processors and graphics cards. In addition, better regularization methods are now known that allow the training of large networks that respond to complex problems. Another important aspect is the ability to provide models with the right amount of data (Big Data) necessary to successfully train a neural network. The Big Data trend has made machine learning much simpler because the significance of the main limitation of estimation techniques (incorrect reasoning based on a small data set) has been reduced [18].

Many aspects have contributed to the success of the deep neural networks observed over the past few years. First of all, thanks to the development of the Big Data trend and much greater availability of various databases, neural networks using deep learning algorithms can draw appropriate conclusions. It turned out that these algorithms become very effective when we provide models with the right amount of data necessary to make proper observations, min. 5000 tagged data in each class. Another reason for developing deep learning is the ability to perform calculations on larger models. Networks consisting of a small number of neurons are not able to model complex problems. However, the more neurons, the more intelligent our system becomes. The use of large models is now possible due to the availability of much faster processors and graphics cards, as well as better methods of preventing overfitting of the network, such as early termination, L2 regularization (adding an additional factor to the error function that imposes a penalty on high weights obtained by the network), or dropout (removing a selected portion of the activation function at random). The greater availability of data used to train models and much better hardware resources meant that deep learning could compete with other algorithms in the world-wide image recognition professions 'ImageNet Large Scale Visual

 doi:10.20944/preprints202005.0274.v1

Recognition'. In 2012, convolutional neural networks [19] achieved first place for the first time, reducing the error of detecting the correct category in the first five results (top-5 error) from 26.1% to 15.3%. Since then, deep neural networks have been winning this competition every year, and now the top-5 error has been reduced by up to 3.6%. Due to the rapid development of the deep learning technique, several models developed on large data sets have been created, which allow classification of objects with great efficiency - Inception model [20] trained using ImageNet containing 1000 categories of images is subject to a top-5 error of 5.6%). In addition, using transfer learning [21], it is possible to re-train existing proven and effective models for new tasks by using existing network weights for all layers except the last one, which is removed and re-trained using a new data set. Transfer learning opens up a number of options for using deep learning algorithms when the training data set is too small to learn full deep representation or when we have limited network training time. Recently, several papers have been devoted to the issue of transfer learning used for the purpose of medical diagnostics in order to improve and improve the quality of services provided, e.g. celiac disease diagnosis and detection of anomalies in duodenal endoscopic images [22]. In both works, the revolutionary neural networks trained using the ImageNet image database [23] were re-trained on a new smaller set of images. The applied transfer learning procedure allowed to significantly reduce the time of network training while maintaining high accuracy of classification. In addition, the same model has been adapted to two completely different tasks, giving very good results in both cases. On this basis, we can conclude that known effective models can be used for virtually any object classification task, as long as we provide the appropriate input data to the networks. In this regard, it becomes very important to create publicly available collections of medical images that could be used to overtrain deep neural networks. The trained models can then be used for remote medical diagnostics.

5. Conclusion

Thanks to the development of deep learning algorithms, image analysis can be performed in real time on generally available desktop computers while maintaining high accuracy of obtained results. However, it should be remembered that models based on deep learning become effective only when we provide them with the necessary amount of data to make the right observations. The development of the Big Data trend has meant that we have much more information at our disposal, thanks to which we can create universal deep neural networks and adapt them to our needs using the transfer learning technique. In order to be able to use them also in medicine and remote medical diagnostics, it is very important to create further databases of photographs of people and their features as well as medical photos, which will be used to create neural networks and appropriate inference when analyzing new images. Big Data is a one of the most important challenges of the modern digital world. Big Data as a complex of IT issues requires the introduction of new data analysis techniques and technological solutions that will allow to extract valuable and useful knowledge from them. New technologies of data collection and processing force interdisciplinary research and the need to combine existing solutions. Future large IT systems will be based on techniques that use Deep Learning, Computer Vision, Big Data and others [24,25], so new technologies should be developed that can process large amounts of data and extract useful knowledge for medical world.

Funding: This work was supported by the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019 - 2022 project number 020/RID/2018/19 the amount of financing 3,000,000 $. https://rid.pcz.pl

References

1. H. Buhl, M. Röglinger, F. Moser, J. Heidemann J., *Big Data, Business & Information Systems Engineering*, **2013**, *5 (2)*, 2013, pp. 65–69.
2. P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, G. Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill, USA, 2012.

3. C. Jinchuan, C. Yueguo, D. Xiaoyong, L. Cuiping, L. Jiaheng, Z. Suyun, Z. Xuan, *Big data challenge: a data management perspective, Frontiers of Computer Science*, **2013**, SP Higher Education Press, vol. 7, issue 2, pp. 157-164.
4. A. Katal, M. Wazid, R.H. Goudar, *Big Data: Issues, Challenges, Tools and Good Practices*, **2013** Sixth Interna-tional Conference on Contemporary Computing (IC3), IEEE, Noida, , pp. 404-409.
5. L. Doug, *Data Management: Controlling Data Volume, Velocity, and Variety*, Application Delivery Strategies, META Group, Gartner, 2011.
6. J. Bobulski, M. Kubanek, *Design of the BLINDS System for Processing and Analysis of Big Data - A Pre-processing Data Analysis Module*, **2019** Advances in Intelligent Systems and Computing, *889*, pp. 132-139.
7. N. Marz, J. Warren, *Big Data Principles and Best Practices*, Manning Publications Co.,2015.
8. L.A. Zadeh, *Computing with Words: Principal Concepts and Ideas*, Springer Publishing Company, Incorporated,2012.
9. R.C. Schank, *Conceptual Information Processing*, Yale University, New Haven, Connecticut, 1975.
10. J. P. Sadalage, M. Fowler, *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*, Pearson Education, 2013.
11. E. Barbierato, M. Gribaudo, M. Iacono, *Performance evaluation of NoSQL big-data applications using multi-formalism models*, **2014** Future Generation Computer Science, *37*, pp. 345-353.
12. V. Kolomičenko, *Analysis and Experimental Comparison of Graph Databases*, Masters's Thesis, Charles University in Prague, 2013.
13. I. Robinson, J. Webber, E. Eifrem, *Graph Databases*, O'Reilly Media, 2013.
14. D. Slotwinski, *Graph databases - technology review*, AGH, 2010.
15. T. Berners-Lee, J. Hendler, O. Lassila, *The semanticweb*, **2001**,Scientific american, *284(5)*, pp.34–43.
16. J. Euzenat, P. Shvaiko, *Ontology Matching*, Springer Science & Business Media, 2007.
17. A.G. Ivakhnenko, V.G. Lapa, *Cybernetic Predicting Devices*, CCM Information Corporation, 1965.
18. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2017).
19. A. Krizhevsky, I. Sutskever, G.E Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, **2012** NIPS'2012: Neural Information Processing Systems, Lake Tahoe, Nevada.
20. C.Szegedy, *Going deeper with convolutions*, **2015** IEEE Conference on Computer Vision and Pattern Recognition CVPR, Boston, MA, pp. 1-9.
21. G. Wimmer, A. Vécsei, A. Uhl, *CNN transfer learning for the automated diagnosis of celiac disease*, **2006** The 6th International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, pp. 1-6.
22. G. Ianiro, S. Bibbò, S. PecereS, A. Gasbarrini, G. Cammarota, *Current technologies for the endoscopic assessment of duodenal villous pattern in celiac disease*, **2015**, Comput Biol Med. *1(65)*, pp. 308–314.
23. ImageNet database, http://www.image-net.org, Last accsess 20 Jan 2019.
24. J. Bobulski, M. Kubanek, *CNN use for plastic garbage classification method*, **2019** 25TH ACM SIGKDD Conference On Knowledge Discovery And Data Mining, Workshop on Data Mining and AI for Conservation, 4-8 August 2019, Anchorage, Alaska, USA.
25. M. Kubanek, J. Bobulski, J. Kulawik, *A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network*, **2019**, Symmetry, *11*, 1185, 2019.