Azerbaijan Journal of High Performance Computing, Vol 1, Issue 2, 2018, pp. 179-184
https://doi.org/10.32010/26166127.2018.1.2.179.184

Convergence of HPC and AI: Two Directions of Connection

Nigar Ismayilova, Elviz Ismayilov

Department of General and Applied Mathematics, HPC Advance Research Center, Azerbaijan State Oil and Industry University, nigar.ismailova@asoiu.edu.az, elviz.ismailov@asoiu.edu.az

*Correspondence: Nigar Ismayilova, Department of General and Applied Mathematics, HPC Advance Research Center, Azerbaijan State Oil and Industry University, nigar.ismailova@asoiu.edu.az

Abstract

This paper examines the role of HPC systems in the solution of the most AI problems, on the other hand, assesses the impact of the application of AI methods on the resolution of different tasks in distributed systems. The findings from the literature review illustrate how these two main fields of science and information technologies can be converged together for the achievement of the goals in designing and developing of the intelligent agents with the high level of intelligence, also for the solution of optimization problems in distributed systems with the increasing complexity.

Keywords: HPC, AI, fuzzy load balancing, exascale load balancing, predicting HPC

1. Introduction

The existence of enormous amount of the data, importance of learning intelligent agents by large-scale training data for implementing their high-quality application, requirement of individual approach for each branch of artificial intelligence (AI), time extending for learning deep neural networks in most cases form days to weeks, etc. makes unavoidable application of different high performance computing (HPC) systems (cloud, grid, P2P, supercomputer) [1] – [2].

Appropriation of machine learning is endorsed to be an essential tool for the solution of problems in such fields as medical diagnosis, designing of intelligent systems, predicting, etc. Nowadays, developing systems based on machine learning requires analysis of huge amount of data regardless of supervised or unsupervised training. Under such circumstances, usage of HPC systems avoids delay emerging in training procedure and reduces several times required period [3] – [4].

There also alternative methods of AI are available for solution of different problems in design and development of the HPC systems, such as predicting of running times, resource utilization, optimization of load balancing, job scheduling, resource discovery and process migration [5] – [6].

This work aims to analyze the convergence of these two main directions of computer science and propose new challenges for further research and application. Investigation of AI and HPC integration was carried out in two directions: the

possibility of usage of parallelization techniques for reducing time in machine learning via HPC systems, also HPC models for data processing in Big data analytics and handling problems appeared on the development of the distributed computing systems (Figure 1).

Figure 1. Convergence of AI and HPC by different points of view

2. Usage of HPC in learning systems

Recent methods concentrate on the solution of training problems in developing intelligent agents by proposing different approaches for convergence of HPC and neural networks [2]-[4]. However, the main disadvantage of present systems is establishing only for neural networks of small size and existence of limitations for the number of parameters. Researchers introduced more specific research for the development of supercomputers only for recognition, mining or intellectual games, some companies and researchers have also suggested establishing multichip architecture for CNN and DNN.

Reviewing of large-scale data, mining of significant and unknown knowledge and facts from the enormous amount of information are the challenging problems of machine learning and big data. Superposition of the algorithm and parallelizing of the iterations makes it possible to decrease the time required for data analysis by a solution of the problem in HPC systems [5].

The main requirements for the application of deep learning are related to hardware attributes. Indeed, it is necessary to establish a hardware platform for storage of large-scale training data and appropriate for a large number of neural

network parameters. From this perspective utilization of HPC systems leads to achieve superior results. On the other hand, there are numerous efforts for reducing hardware requirements by application of different optimization problem algorithms which ensure satisfactory results [6].

One of the main problems of AI is single appointment characteristic of intelligent agents. Thus text recognition system developed for any language would not be able to show the results with the identical quality. Application of HPC contributes utilization of machine learning system for AI tools with different engagements, speech recognition of two different languages as English and Chinese mandarin proposed and investigated by Amodei and others would be a good model for the importance of HPC in machine learning [1].

3. Applications of machine learning algorithms in big data

Big data analytics, extraction of useful knowledge from the massive scale of information, causal inference through the texts, text summarization, text reviewing are the main problems of current research not only in applied sciences, but it also becomes to the main subject of social sciences as political science, sociology, psychology [10].

Problems of decision making and predicting through classification, analysis, summarization and mental processing of existing information are in the central interest of both HPC and AI companies and scientific laboratories. Successful solution of this problem is illustrated by Elsebakhi and et al. for analyzing big data through the use of the system modeled on machine learning, data mining, and statistics [8]. Suthaharan (2014) has used the representation-learning technique to predict attacks in computer networks and cloud environment [9]. An important question integrated with big data classification developed via machine learning techniques is the difficulty of using big data analytics system trained in the particular dataset for another kind of large-scale information. However, the proposed approaches for application of machine learning techniques in big data analytics have many problems related to data itself, so it is necessary to develop tools to handle mislabeled data, missing values, high dimensionality, and imbalance of training data and for implementation of noise [11].

Although many authors have conducted many machine learning algorithms for analyzing big data, there is still a problem in usability front. One of the main challenges for researchers in this field is developing useful tools for managing, visualization and understanding of results [12].

4. AI tools for load balancing in HPC systems

Another direction of convergence between AI and HPC is involvement of machine learning techniques during several steps of HPC systems as load balancing, resource discovery, process migration, job scheduling. In this regard application of learning algorithms for one of the main problems of HPC as a job running time predicting and achieving satisfying results can be mentioned as an example [7].

As load balancing in HPC systems remains the best assignment between tasks and machines, application of combinatorial optimization problem solutions can

become a useful tool. In a recent study by several researchers were analyzed, compared and critiqued relevance of using genetic algorithms for task scheduling and load balancing in different HPC systems [13-15]. Necessity of predicting resource utilization for incoming tasks and as a result handling the task scheduling problem enables to use neural networks in distributed systems [16], the application of neural networks for forecasting loads in cloud data centers under the condition of minimizing energy costs was successfully tested in cloud computing platforms by Prevost and others [17].

Nowadays increasing the complexity of supercomputers complicates the definition of the best assignment between tasks and machines in HPC systems and parallelism at the level of hundreds of million processors [22]. Therefore, handling of load balancing, moreover resource discovery, process migration in Exascale computing environment requires new approaches and different methods of AI would give beneficial results in developing of Exascale computing systems. In this case, traditional discrete load balancing mapping finally becomes useless and appears necessity for hybrid load balancing mapping which can be characterized as continuous function [23]. The primary practical solution for this problem might be the application of the AI methods and fuzzy logic. Representing load balancing assignment by fuzzy graphs or definition of fuzzy relations between processes or resources at different time moments might be one of the best solutions for job scheduling in distributed systems.

5. Conclusion

On this basis, we conclude that developing of the intelligent systems to improve the living conditions of the people would be impossible without application of the appropriate distributed systems — this statement due to the requirement for large-scale training data in the shortest period.

At the same time, another aspect of the research argued that, a solution of the control optimization problems, also job scheduling, process migration, time predicting, resource discovery in the modern complex supercomputers would be achieved by using of different soft computing methods as neural networks, fuzzy logic, constraint satisfaction problems, etc.

In future work, investigating the methods for handling load balancing function in Exascale environment, where complexity is more than traditional distributed systems, and quite possibly the emergence of dynamic and interactive event, application of AI methods could continue developing of the powerful supercomputers and the productive distributed systems in the world.

Reference

[1] Sammut, C., & Webb, G. I. (Eds.). (2011). *Encyclopedia of machine learning.* Springer Science & Business Media.

[2] Ganapathi, A., Datta, K., Fox, A., & Patterson, D. (2009, March). A case for machine learning to optimize multicore performance. In *Proceedings of the First USENIX conference on Hot topics in parallelism* (pp. 1-1). Berkeley, CA: USENIX Association.

[3] Chien, S. W. D., Sishtla, C. P., Markidis, S., Zhang, J., Peng, I. B., & Laure, E.

(2018). An Evaluation of the TensorFlow Programming Model for Solving Traditional HPC Problems. In *International Conference on Exascale Applications and Software* (p. 34). The University of Edinburgh.

[4] Pittino, F., Diversi, R., Benini, L., & Bartolini, A. (2018). Robust online identification of thermal models for in-production HPC clusters with machine learning-based data selection. *arXiv preprint arXiv:1810.01865.*

[5] Hamada, S., Akiyama, S., & Namiki, M. (2018). Reactive NaN Repair for Applying Approximate Memory to Numerical Applications. *arXiv preprint arXiv:1804.00705.*

[6] Berral, J. L., Goiri, Í., Nou, R., Julià, F., Guitart, J., Gavaldà, R., & Torres, J. (2010, April). Towards energy-aware scheduling in data centers using machine learning. In *Proceedings of the 1st International Conference on energy-Efficient Computing and Networking* (pp. 215-224). ACM.

[7] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, J. (2016, June). Deep speech 2: End-to-end speech recognition in English and mandarin. In *International Conference on Machine Learning* (pp. 173-182)

[8] Cireşan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745.*

[7] Esmaeilzadeh, H., Sampson, A., Ceze, L., & Burger, D. (2012, December). Neural acceleration for general-purpose approximate programs. In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 449-460). IEEE Computer Society.

[8] Temam, O. (2012). A defect-tolerant accelerator for emerging high-performance applications. *ACM SIGARCH Computer Architecture News, 40*(3), 356-367.

[9] Boehm, M., Tatikonda, S., Reinwald, B., Sen, P., Tian, Y., Burdick, D. R., & Vaithyanathan, S. (2014). Hybrid parallelization strategies for large-scale machine learning in SystemML. *Proceedings of the VLDB Endowment, 7*(7), 553-564.

[10] Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., & Andrew, N. (2013, February). Deep learning with COTS HPC systems. In *International Conference on Machine Learning* (pp. 1337-1345).

[11] Gaussier, E., Glesser, D., Reis, V., & Trystram, D. (2015, November). Improving backfilling by using machine learning to predict running times. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (p. 64). ACM.

[12] Elsebakhi, E., Lee, F., Schendel, E., Haque, A., Kathireason, N., Pathare, T., ... & Al-Ali, R. (2015). Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. *Journal of Computational Science, 11,* 69-81.

[13] Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review, 41*(4), 70-73.

[14] Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics, 48*(1), 80-83.

[15] Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosys-

tem. *Journal of Big Data, 2*(1), 24.

[16] Madden, S. (2012). From databases to big data. *IEEE Internet Computing,* (3), 4-6.

[17] Sim, K. M., & Sun, W. H. (2003). Ant colony optimization for routing and load-balancing: survey and new directions. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 33*(5), 560-572.

[18] Dasgupta, K., Mandal, B., Dutta, P., Mandal, J. K., & Dam, S. (2013). A genetic algorithm (ga) based load balancing strategy for cloud computing. *Procedia Technology, 10,* 340-347.

[19] Dam, S., Mandal, G., Dasgupta, K., & Dutta, P. (2015, February). Genetic algorithm and gravitational emulation based hybrid load balancing strategy in cloud computing. In *2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)* (pp. 1-7). IEEE.

[20] Sigal, L., & Glauberman, A. (2012). *U.S. Patent No. 8,185,909.* Washington, DC: U.S. Patent and Trademark Office.

[21] Prevost, J. J., Nagothu, K., Kelley, B., & Jamshidi, M. (2011, June). Prediction of cloud data center networks loads using stochastic and neural models. In *System of Systems Engineering (SoSE), 2011 6th International Conference on* (pp. 276-281). IEEE.

[22] Geist, A., & Lucas, R. (2009). Major computer science challenges at exascale. *The International Journal of High Performance Computing Applications, 23*(4), 427-436.

[23] Ramyani Saleh, S., Mousavi Khaneghah, E., Shadnoush, N., & Aliev, A. R. (2018). A mathematical framework for managing interactive communication distortions in exascale organizations. *Cogent Business & Management, 5: 1545356, 1-23.*

Submitted 29.06.2018
Accepted 20.10.2018