proceedings

MDPI

Proceedings

Cognitive Computing Architectures for Machine (Deep) Learning at Scale †

Samir Mittal

CEO and Founder, SCUTI AI, Palo Alto, CA 94306, USA; samir@scutiai.com; Tel.: +1-952-807-3598

† Presented at the IS4SI 2017 Summit DIGITALISATION FOR A SUSTAINABLE SOCIETY, Gothenburg, Sweden, 12–16 June 2017.

Published: 9 June 2017

Abstract: The paper reviews existing models for organizing information for machine learning systems in heterogeneous computing environments. In this context, we focus on structured knowledge representations as they have played a key role in enabling machine learning at scale. The paper highlights recent case studies where knowledge structures when combined with the knowledge of the distributed computation graph have accelerated machine-learning applications by 10 times or more. We extend these concepts to the design of Cognitive Distributed Learning Systems to resolve critical bottlenecks in real-time machine learning applications such as Predictive Analytics and Recommender Systems.

Keywords: machine learning; cognitive computing; distributed computing; knowledge structures; heterogeneous computing

1. Introduction

As depicted in Figure 1, the information technology landscape has evolved from providing business intelligence through analytics to enabling operational intelligence with real-time insights. Businesses are able to generate real-time insights to harness perishable in-the-moment opportunities. These insights are synthesized through analysis of live data within the contextual lens of multi-domain historical data. Inference rules that were traditionally generated with classical ETL processes are being quickly replaced by Machine Learning (ML) and Deep Learning (DL) techniques to generate knowledge based inference engines.

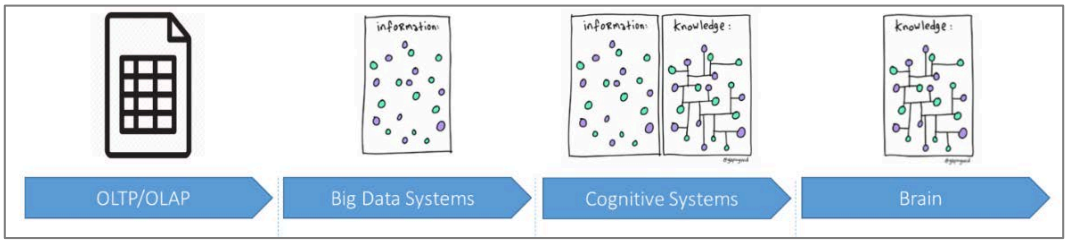

Figure 1. Evolution of Information Technology landscape (from [1]).

Advancements in knowledge based learning systems are accelerating. A rich eco-system of algorithms and open source frameworks, implemented on commodity infrastructure, holds the promise of universally accessible capability. Society scale impact is expected in autonomous driving, contextualized recommendations, and personalized medicine and in other verticals. As shown in Figure 2, machine translation capability is approaching human quality, and automated large-scale image classification is beating human capability (see Figure 3). In medicine, ML/DL systems have

started to outperform doctors in detecting breast cancer in radiology images [2] It is the dawn of the golden age of data, and we have only begun to unlock key capabilities through learning structures.

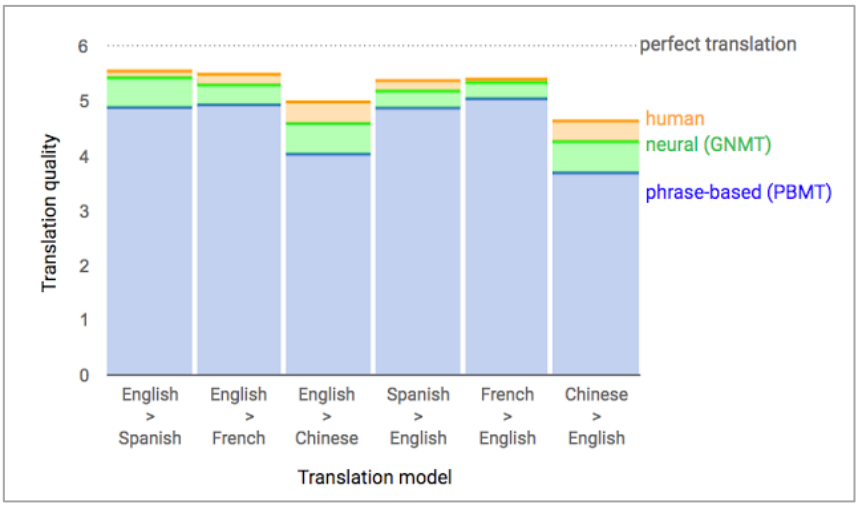

Figure 2. DL/ML language translation capability [3].

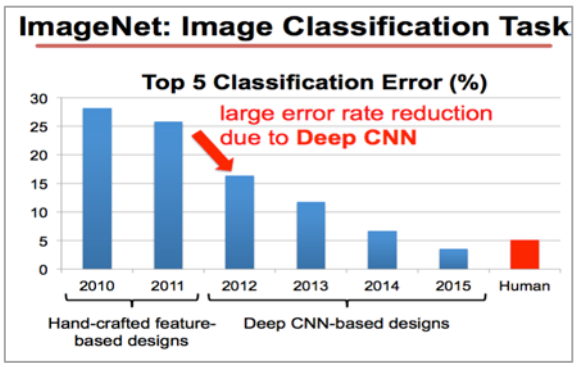

Figure 3. Automated image classification [4].

2. Challenges with Machine Learning

Building machine learning systems is hard. The most visible and the best success stories require 100's if not 1000's of engineers. The prediction accuracy of learning systems improves with more data and larger models. Computation requirements grow non-linearly with the complexity of the task at hand (Figure 4). This creates acute challenges relating to data dimensionality, complex model development, slow experiments, and scalability of production deployments.

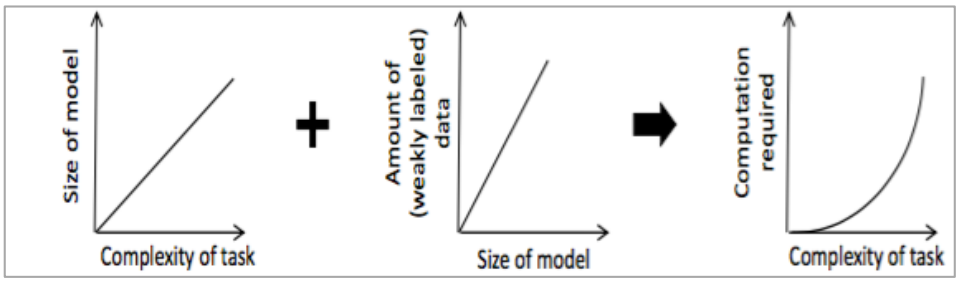

Figure 4. Computational requirements for scale-out learning (from [5]).

The data and the computation pipelines in DL/ML systems are complex. As explained in a recent paper [6], DL/ML code comprises only 10% of real world ML systems. The bulk of the effort is consumed in infrastructure and data management (Figure 5). Automating much of this pipeline has

become a focus of recent research activity, so as to make ML/DL systems universally accessible, and refocus the activities of the domain expert in building production quality systems [7].

Training ML/DL systems are compute intensive tasks, where models can take exaflops to compute while processing and generating petabytes of input and intermediate data. The compute complexity is high; medium sized experiments and popular benchmarks can take days to run [8], severely compromising the productivity of the data scientist. Distributed scaling stalls only after a dozen nodes due to locking, messaging, synchronization and data locality issues. The rate of data avalanche is beating the growth rates from Moore's law, resulting in diminished economic returns at scale (Figures 6 and 7).

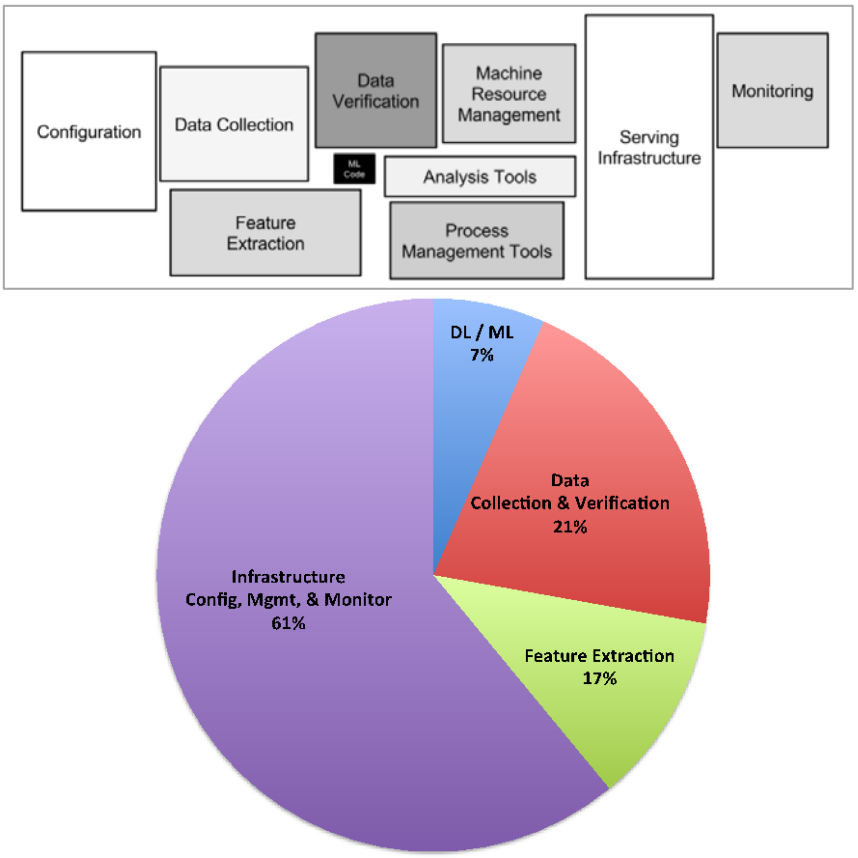

Figure 5. The hidden technical debt in machine learning systems (adapted from [6]).

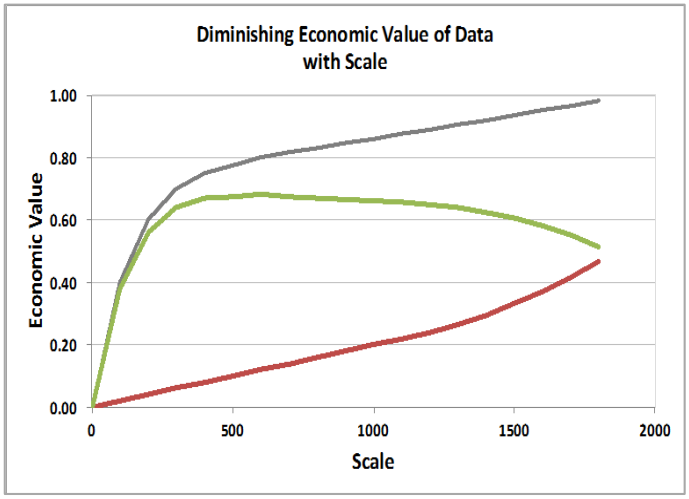

Figure 6. Diminishing economic value of data with scale.

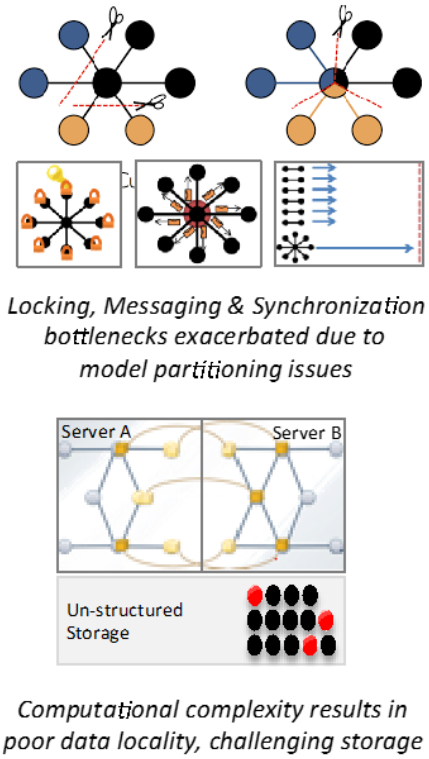

Figure 7. Key issues in infra management.

3. Cognitive Computing Stack

To address the issues highlighted above, recent research has turned to self-managing, self-aware models of computing [9]. The idea is to enable autonomic management of the compute infrastructure to the application intent with regards to application performance and data security. Policies defining resource, workload and computation process rules regulate application performance (Figure 8). This strategy promises exciting and new results in the following areas [10]:

- **Separation of Concerns and Information hiding**: Infrastructure and process optimization concerns are separated from application logic to hide complexity that allows domain experts to re-focus on ML/DL breakthroughs.
- **Autonomic scaling:** Provides features that economize scaling, and allow applications to achieve high performance without human engineering. It extracts the best performance of many-core systems, while optimizing for reliability and data movement—the primary impediments in designing scale-out DL/ML systems.
- **Ability to handle complexity with scale:** The system can accumulate knowledge and act on it to adaptively tune its behavior to robustly achieve desired goals within the performance, power, and resilience envelopes specified in the policy framework.
- **Computation resiliency and trusted results:** Improve the resiliency of data, applications, software, and hardware systems, as well as the trustworthiness of the results produced through in-situ fault detection, fault prediction and trust verification mechanisms.

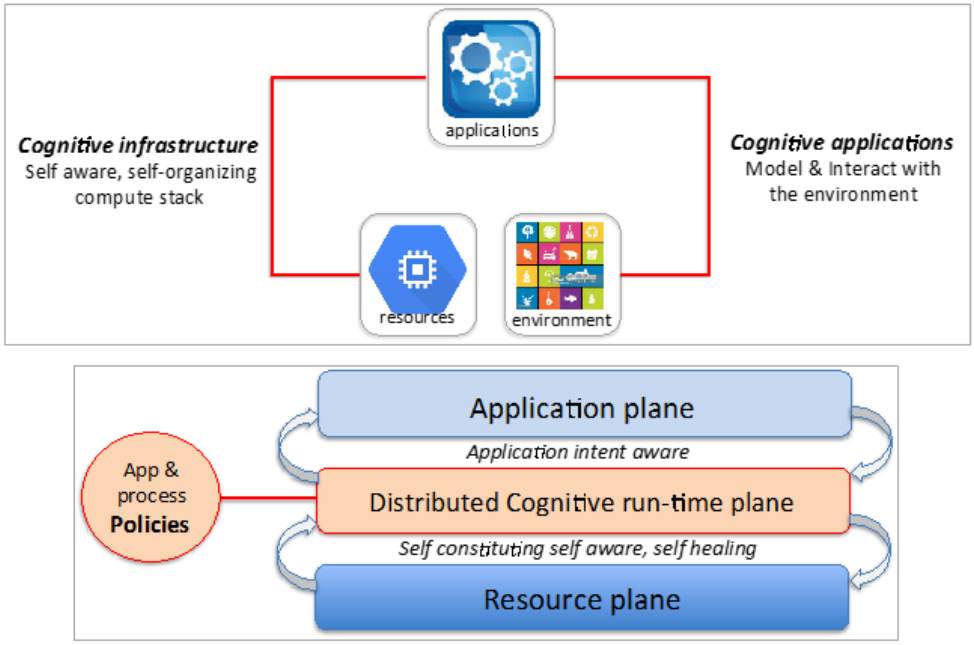

Figure 8. Cognitive distributed computing model by Dr. Rao Mikkilineni [11].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Storage Implications of Cognitive Computing. Available online: http://www.snia.org/sites/default/files/DSI/2016/presentations/gen_sessions/BalintFleischer-JianLI_Storage_Implications_Cognitive_Computing_1-04.pdf (accessed on 6 July 2017).
2. Liu, Y.; Gadepalli, K.; Norouzi, M.; Dahl, G.E.; Kohlberger, T.; Boyko, A.; Venugopalan, S.; Timofeev, A.; Nelson, P.Q.; Corrado, G.S.; et al. Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv* **2017**, arXiv:1703.02442.
3. A Neural Network for Machine Translation, at Production Scale. Available online: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html (accessed on 6 July 2017).
4. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252.
5. Chilimbi, T.M.; Suzue, Y.; Apacible, J.; Kalyanaraman, K. Project Adam: Building an Efficient and Scalable Deep Learning Training System. In Proceedings of the 11th Usenix Conference on Operating Systems Design and Implementatio (OSDI), Broomfield, CO, USA, 6–8 October 2014; Volume 14.
6. Sculley, D.; Hold, G.; Golovin, D.; Davydov, E.; Phillips, H.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.F.; Dennison, D. Hidden technical debt in machine learning systems. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2503–2511.
7. Palkar, S.; Thomas, J.J.; Shanbhag, A.; Narayanan, D.; Pirk, H.; Schwarzkopf, M.; Amarasinghe, S.; Zaharia, M. Weld: A common runtime for high performance data analytics. In Proceedings of the Conference on Innovative Data Systems Research (CIDR), Chaminade, CA, USA, 8–11 January 2017.
8. Keuper, J.; Preundt, F.-J. Distributed training of deep neural networks: Theoretical and practical limits of parallel scalability. In Proceedings of the Workshop on Machine Learning in High Performance Computing Environments, Salt Lake City, UT, USA, 13–18 November 2016.
9. Mikkilineni, R.; Comparini, A.; Morana, G. The Turing O-Machine and the DIME Network Architecture: Injecting the Architectural Resiliency into Distributed Computing. In Proceedings of the Turing-100—The Alan Turing Centenary, Manchester, UK, 22–25 June 2012.

10. Rockville, M.D. *Machine Learning and Understanding for Intelligent Extreme Scale Scientific Computing and Discovery*; DOE Workshop Report; Workshop Organizing Committee: Rockville, MD, USA, 2015.
11. Dr. Rao Mikkilineni, Available online: https://www.linkedin.com/in/raomikkilineni/ (accessed on 6 July 2017).