Energy Efficient Computing with Hyperdimensional Vector Space Models

Sayeef Salahuddin
EECS, UC Berkeley, USA
sayeef@berkeley.edu

Computing is changing. Over the past six decades, the semiconductor industry has been immensely successful in providing exponentially increasing computational complexity at an ever-reducing cost and energy footprint. This incessant march has been supported by a set of well-defined abstraction layers, starting from robust switching devices that support a deterministic Boolean algebra to a scalable and stored program architecture, which is Turing complete, and hence capable of tackling a wide variety of computational challenge. Unfortunately this abstraction chain is being challenged as scaling continues to nanometer dimensions and also by exciting new applications that must support a myriad of new data types. Maintaining a deterministic model ultimately puts a lower bound on the amount of energy scaling that can be obtained, set in place by fundamental physics that governs the operation and also by the variability and reliability of the underlying nanoscale devices [1,2].

At the same time, it is clear that the nature of computing itself is evolving rapidly. For a vast number of applications, cognitive functions such as classification, recognition, synthesis, decision-making and learning are gaining rapid importance in a world that is infused with sensing modalities and in need of efficient information-extraction. This is in sharp contrast to the past when the central objective of computing was to perform calculations on numbers and produce results with extreme numerical accuracy. Indeed, the recent success of Deep Learning networks - based on the artificial neural nets (ANNs) of the past - is finding ever expanding applications from speech and image recognition to predicting the effects of mutations in non-coding DNA on gene expression and disease [3-10]. Recently the AlphaGo program developed by Google DeepMind has convincingly beaten a professional human player [11].

While these success stories allude to an intriguing future for Learning Machines, the reality is that each of these programs needs a large cluster of digital computers to run several days to train and be able to perform the required computation. In doing so, they also dissipate gigantic amounts of energy. Thus the combination of the digital computer and deep learning algorithms, while very important as proof of the concept, is neither realistic nor scalable for broad societal adoption. To realize the full potential of the learning machines significant advances are essential in every aspect of the computing hierarchy.

Conventional deep learning algorithms require brute force training of billions of weights in repetitive iterations at every layer of a several-layer stack. As a result learning is slow and every change of weight, that involves changing the value of a state variable, expends energy. Shuttling data through the networks is also power hungry. Currently there are no guarantees regarding the optimality or efficiency of operations in these networks. It is therefore essential that new algorithms be found, which are capable of 'online' or one-shot learning, and for which there exists computational theories that bound the resource usage and complexity for a given task. At the same time, significant technological advances are required to create new physical devices, that will form the building blocks of future learning machines, so that the operating voltage can be substantially lowered. These devices must be integrated and organized to create efficient architectures that are tuned to the intricacies of the corresponding learning algorithms to achieve optimum energy usage.

Recently, with the support of National Science Foundation (NSF) and Semiconductor Research Corporation (SRC), we, a group of PIs from UC Berkeley and Stanford University, have embarked on a project, ENIGMA, that is exploring a new computational model called the Hyper Dimensional (HD) Computing [12-14]. In this formalism, information is represented in ultra high dimensional vectors. Such vectors can then be mathematically manipulated to not only classify but also to bind, associate and perform other types of cognitive operations in a straightforward manner. In addition, these mathematical operations also ensure that the resulting hyper vector is unique and thus the learning is one shot. Thus HD computing can substantially reduce the number of operations needed by conventional deep learning algorithms, thereby providing tremendous energy savings. In this talk, I shall give a brief overview of the computational model and also how potential hardware implementation of such a model can be achieved.

References

[1] Semiconductor Industries Association (SIA) and Semiconductor Research Corporation (SRC), Rebooting the IT Revolution, a Call for Action, (2015), https://www.src.org/newsroom/rebooting-the-it-revolution.pdf

[2] Stanley Williams, Erik P. DeBenedictis, OSTP Nanotechnology Inspired Grand Challenge: Sensible Machines (extended version 2.5), October 20, 2015, http://rebootingcomputing.ieee.org/images/files/pdf/SensibleMachines_v2.5_N_IEEE.pdf

[3] Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012).

[4] Farabet, C., Couprie, C., Najman, L. & LeCun, Y. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929 (2013).

[5] Tompson, J., Jain, A., LeCun, Y. & Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In Proc. Advances in Neural Information Processing Systems 27 1799–1807 (2014).

[6] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9. 2015.

[7] Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocky, J. Strategies for training large scale neural network language models. In Proc. Automatic Speech Recognition and Understanding 196– 201 (2011).

[8] Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. and Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6), pp.82-97.

[9] Sainath, T., Mohamed, A.-R., Kingsbury, B. & Ramabhadran, B. Deep convolutional neural networks for LVCSR. In Proc. Acoustics, Speech and Signal Processing 8614–8618 (2013).

[10] Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R. and Morris, Q., 2015. The human splicing code reveals new insights into the genetic determinants of disease. Science, 347(6218), p.1254806.

[11] https://gogameguru.com/tag/deepmind-alphago-lee-sedol/

[12] P. Kanerva. Sparse Distributed Memory. MIT Press, 1988.

[13] P. Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. Cognitive Computation, 1(2):139–159, 2009.

[14] P. Kanerva. Computing with 10,000-bitwords. In Proc. or the 52nd Annual Allerton Conference on Communication, Control, and Computing, 2014.