Hierarchical AI - from neurons to psychology

Serb, Alexantrou
Zepler Institute
Univ. of Southampton
Southampton, UK
A.Serb@soton.ac.uk

Themistoklis Prodromakis
Zepler Institute
Univ. of Southampton
Southampton, UK

***Abstract*—The undeniable successes of deep learning (and more generally statistical learning) in bringing pattern matching to the market is still just the tip of the iceberg of AI. In this talk we will look at a very high level overview of AI as a whole and see how it can be interpreted at very different levels of abstraction. Each level of abstraction boasts its own vocabulary and is suited to understanding different aspects of the general problem of artificial intelligence. We will walk through four "levels of abstraction of AI" ranging from physical implementation all the way to semantic processing, and will investigate how memory technologies can play a vital role in their successful implementation. The aim is to show how innovation in the domain of memory tech can unlock the potential of AI to attack problems much more general than simple pattern matching and thus pave the way to the next wave of AI on the market.**

Keywords — AI, perspective, education

I. Overview of Talk

Currently AI is understood primarily as statistical learning, very frequently implemented in a connectionist style (using artificial neural networks – ANNs, e.g. deep learning [1], convolutional neural networks [2] etc.). The zoo of ANN topologies expands rapidly every year, but all share an important aspect: a voracious and ever increasing appetite for memory resources with modern networks such as VGG19 [3] using close to 100M weights. Simultaneously, hardware engineers strive to build microchips and larger systems that are tailor-made for the type of computational task required by statistical learning-based AI. As a result systems such as TrueNorth [4], BrainScaleS [5], SpiNNaker [6] and many more have emerged and are constantly being developed and upgraded. These architectures need to offer an answer to the problem of memory storage and access, which they all do in their own manner but all within the constraints and limits of traditional Complementary Metal-Oxide Silicon (CMOS) electronics.

Against this backdrop of the state-of-art we make the argument that developing truly intelligent AI in an efficient hardware substrate will require moving well beyond machine learning specialists and traditional CMOS electronics engineers. In order to identify areas of interest we begin by reviewing the hierarchical structure of AI proposed in [7]. According to that work, a natural way to compartmentalise the area follows a 5-level structure:

- Level 1: The physical layer. A layer of transistors, electrons, voltage and currents. This is where important concepts such as signals and memory are grounded in the physical world, as well as where the memristor community [8] is focussing efforts for building artificial, ultra-compact and low-power synapses made of all kinds of materials and featuring a swath of different properties [9]–[13].

- Level 2: The functional layer. A layer of logic gates, artificial neurons, activation function circuits, multipliers and (arithmetic) accumulators. This is where fundamental mathematical functions are constructed – the Lego blocks used for the next layer. Memory devices here can be used to instantiate full neurons [14] as well as the 'active/regenerative cabling' required to emulate axonal signal transmission [15].

- Level 3: The computational layer. The layer of modular neural networks, ANN microcircuits (e.g. convolutional kernels) and in general, standard ANN topologies (Boltzmann machines, fully connected networks, reservoirs, etc.). This is where much of the machine learning community lives and breathes. Memory devices enter the scene here typically in the form of arrays, for example crossbars [16] (memristive or otherwise) that typically emulate the synaptic connectivity matrix of general purpose neural networks [17], [18]. Memory technologists have to solve a host of problems not visible at the individual device level in order to allow crossbars to work adequately well. These include the sneak path problem [19] and developing appropriate read-out techniques [20].

- Level 4: The semantic layer. A layer of symbols, concepts, and symbolic manipulation that provides insights in data that simple interpolation cannot match without increasing the data sample by orders of magnitude. Includes work on the mathematics of high-dimensional vectors and their implementation in hardware (e.g. [21]–[23] and many more). The memory technology aspect here is directly intertwined with the concepts of in-memory computation and the 'arithmetic-logic memory' [24], [25]. Specifically here we are interested in building a memory fabric capable of performing some of the fundamental operations of hyperdimensional computing directly in-memory [26].

- Level 5: The agency layer. A layer of stimuli, motivations, reactions, models (internal and of the outside world). This layer closer resembles psychology than

machine learning. It is yet unclear exactly how memory technology will exert influence at this level.

Naturally, the boundaries of each level are blurry and any specialist in any layer will need a good working understanding of at least the layers adjoining it. However, the above breakdown illustrates how memory technology from fundamental physics to integrated memory systems design is a key component of physically embodied AI system, whose design exerts a profound impact on the performance of the overall system.

On striking aspect of this overview that doesn't become immediately clear is advance of volatile memory technologies into the domains previously considered the exclusive territory of 'signal electronics'. For example, volatile synapses (level 1) and the memristors implementing the active cabling of the 'neurisor' (level 2) encroach on territory that a majority of engineers would still consider the domain of monostable circuits and RCs. This equivalence and interchangeability between electronic conduction-based and ionic motion-based paradigms illustrates that the fundamental hardware principles underlying both memory and computation need not be treated as separate, but rather a more flexible outlook would be opportune.

In conclusion, memory technology and mem-tech hardware developers and companies have a significant role to play in the ongoing AI revolution. Improvements across all levels of mem-tech, from physics to system-level design, can be expected to dramatically impact the future evolution of AI. They will also play a substantial role in the next wave of AI, which is expected to be the rise of semantic-level computing. It is important to remember that whilst this seems like an algorithm-level innovation the hardware community has still got a lot to contribute on the matter.

ACKNOWLEDGEMENTS

Many thanks to a vast number of people for stimulating conversations (Chris Eliasmith and Terry Stewart, Univ. of Waterloo; Ivan Kobyzev, Borealis, Evgeny Osipov, Univ. of Lulea, Michael Hopkins, Univ. of Manchester and many, many more).

REFERENCES

[1] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014.

[4] F. Akopyan *et al.*, "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[5] K. Meier, "A mixed-signal universal neuromorphic computing system," *Tech. Dig. - Int. Electron Devices Meet. IEDM*, vol. 2016-Febru, no. 1, pp. 4.6.1-4.6.4, 2015.

[6] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, May 2014.

[7] A. Serb and T. Prodromakis, "A system of different layers of abstraction for artificial intelligence," Jul. 2019.

[8] R. Waser and M. Aono, "Nanoionics-based resistive switching memories.," *Nat. Mater.*, vol. 6, no. 11, pp. 833–40, Nov. 2007.

[9] C. Du, W. Ma, T. Chang, P. Sheridan, and W. D. Lu, "Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics," *Adv. Funct. Mater.*, vol. 25, no. 27, pp. 4290–4299, Jul. 2015.

[10] S. L. Wei, E. Vasilaki, A. Khiat, I. Salaoru, R. Berdan, and T. Prodromakis, "Emulating long-term synaptic dynamics with memristive devices," Sep. 2015.

[11] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, "HfO2-based memristors for neuromorphic applications," in *Proceedings - IEEE International Symposium on Circuits and Systems*, 2016, vol. 2016-July, pp. 393–396.

[12] N. Du *et al.*, "Single pairing spike-timing dependent plasticity in BiFeO3 memristors with a time window of 25 ms to 125 μs.," *Front. Neurosci.*, vol. 9, p. 227, Jan. 2015.

[13] A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses," *Nat. Commun.*, vol. 7, p. 12611, Sep. 2016.

[14] Z. Wang *et al.*, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nat. Electron.*, vol. 1, no. 2, 2018.

[15] M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, "A scalable neuristor built with Mott memristors.," *Nat. Mater.*, vol. 12, no. 2, pp. 114–7, Feb. 2013.

[16] J. E. Green *et al.*, "A 160-kilobit molecular electronic memory patterned at 10(11) bits per square centimetre.," *Nature*, vol. 445, no. 7126, pp. 414–7, Jan. 2007.

[17] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, May 2015.

[18] A. Shafiee *et al.*, "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," in *Proceedings - 2016 43rd International Symposium on Computer Architecture, ISCA 2016*, 2016.

[19] E. Linn, R. Rosezin, C. Kügeler, and R. Waser, "Complementary resistive switches for passive nanocrossbar memories.," *Nat. Mater.*, vol. 9, no. 5, pp. 403–406, May 2010.

[20] A. Serb, W. Redman-White, C. Papavassiliou, and T. Prodromakis, "Practical Determination of Individual Element Resistive States in Selectorless RRAM Arrays," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 63, no. 6, pp. 827–835, Jun. 2016.

[21] P. Kanerva, "Fully Distributed Representation," *Proc. 1997 Real World Comput. Symp.*, no. c, pp. 358–365, 1997.

[22] T. A. Plate, “Holographic Reduced Representations,” *IEEE Trans. Neural Networks*, vol. 6, no. 3, pp. 623–641, May 1995.

[23] A. Serb, I. Kobyzev, J. Wang, and T. Prodromakis, “A semi-holographic hyperdimensional representation system for hardware-friendly cognitive computing,” Jul. 2019.

[24] Y. Levy *et al.*, “Logic operations in memory using a memristive Akers array,” *Microelectronics J.*, vol. 45, no. 11, pp. 1429–1437, Nov. 2014.

[25] E. Lehtonen and M. Laiho, “Stateful implication logic with memristors,” in *2009 IEEE/ACM International Symposium on Nanoscale Architectures*, 2009, pp. 33–36.

[26] G. Karunaratne, M. Le Gallo, G. Cherubini, L. Benini, A. Rahimi, and A. Sebastian, “In-memory hyperdimensional computing,” Jun. 2019.