The Brain's Computational Efficiency derives from using Sparse Distributed Representations

Rod Rinkus (rod@neurithmicsystems.com)
Neurithmic Systems, 468 Waltham St.
Newton, MA 02465 USA

Abstract:

Machine learning (ML) representation formats have been dominated by: a) *localism*, wherein individual items are represented by *single* units, e.g., Bayes Nets, HMMs; and b) *fully distributed* representations (FDR), wherein items are represented by unique activation patterns over *all* the units, e.g., Deep Learning (DL). DL has had great success vis-a-vis classification accuracy and learning complex mappings (e.g., AlphaGo). But, without massive machine parallelism (MP), e.g., GPUs, and thus high power, DL learning is intractably slow. The brain is also massively parallel, but uses only 20 watts and moreover, the forms of MP used in DL, model / data parallelism and shared parameters, are patently non-biological, suggesting DL's core principles do not emulate biological intelligence. We claim that a basic disconnect between DL/ML and biology and the key to biological intelligence is that instead of FDR or localism, the brain uses *sparse distributed* representations (SDR), i.e., "cell assemblies", wherein items are represented by small sets of binary units, which may overlap, and where the pattern of overlaps embeds the similarity/statistical structure (generative model) of the domain. We've previously described an SDR-based, extremely efficient, one-shot learning algorithm. Here, we discuss fundamental differences between the mainstream localist/FDR-based and our SDR-based approaches.

Keywords: Sparse distributed representations (SDR); cell assemblies; one-shot learning; generative model; episodic memory; semantic memory; locality-sensitive hashing; nearest neighbor methods

Current State of Machine Learning

Deep learning methods, i.e., Deep Belief Nets, ConvNets, LSTM, have achieved great successes in recent years, significantly lowering error rates on numerous ML benchmarks, and learning complex mappings, i.e., action policies, e.g., AlphaGo. Several conceptual / algorithmic advances have been crucial, including: a) using contrastive divergence (CD) for unsupervised training of restricted Boltzmann machines (RBMs) (Hinton, Osindero, & Teh, 2006); b) level-by-level unsupervised *pretraining* of a stack of RBMs, which greatly helps a final, supervised (Backpropagation) learning phase; c) *dropout*, i.e., random exclusion of a large fraction (e.g., half) of a level's weights on any one training input; d) tied/shared weights; e) the addition of reinforcement learning (RL) to the learning protocol (Silver et al., 2016); and f) *adversarial* learning.

However, the two items usually mentioned first as fueling the rise and success of DL are: 1) availability of cheap, massive *machine parallelism* (MP), i.e., GPUs; and 2) availability of massive amounts of training data. There are two major causes for concern regarding item 1. First, it is readily acknowledged that without massive MP, DL learning times are unacceptably long, i.e., do not scale to "big data", e.g., for the watershed ImageNet result (Krizhevsky, Sutskever, & Hinton, 2012), learning took a week on two GPUs: how long would it have taken on a single CPU? Thus, the current situation for DL is that learning either requires lots of power or takes too long. Secondly, though the brain is massively parallel, the forms of parallelism used in DL, model and data parallelism, and tied/shared weights, are clearly non-biological in nature, and in fact, actually accentuate the processor-memory distinction despite the hardware community's understanding that most (~80-90%) of the energy used in computation is expended in moving data between memory and processor. This raises questions as to whether DL learning algorithms and representations might fundamentally differ from those of the brain.

The second item above, availability of massive training data, is also problematic because it engenders a view of learning that may be quite at odds with human learning. It is increasingly acknowledged that much of learning, particularly of declarative knowledge, appears to be single or few-trial. Certainly, by definition, *episodic* memories of specific events are formed on the basis of single trials (we discuss hippocampus below). However, owing to their origins in gradient-following and energy-based concepts, e.g., Backpropagation, Boltzmann machines, DL methods, by and large, involve gradual, small weight changes, based on numerous samples and epochs. While consistent with traditional statistical theory, e.g., more samples yields more reliable estimates, that framework does not fully leverage the structural regularity (in space and time) of natural (physical) domains. That structural regularity means that individual inputs contain far more information, i.e., higher-order statistics, than is used in DL learning methods. The DL community has recently begun to address this, cf. Hinton's "dark knowledge" (Hinton, Vinyals, & Dean, 2015), though with respect to supervised learning and still, with FDR. This structural regularity of natural domains is also what's exploited in Compressive Sensing [see (Ganguli & Sompolinsky, 2012) for a relevant assessment]. Our SDR-based approach described below also exploits this regularity, but under a very different learning paradigm involving singular, maximal weight changes (from 0 to 1), rather than numerous small weight changes.

A Radically Different Approach

In prior work, we described an SDR-based, cortex-inspired model of spatial / spatiotemporal learning and probabilistic inference (Rinkus, 1996, 2010, 2017), for which: a) the time to learn a new input or retrieve the *closest matching* stored input remains *fixed* as the number of stored inputs grows; and b) the number of items storable (to a criterion retrieval accuracy) grows faster-than-linearly in the number of units. The model, now called Sparsey, can be viewed as a form of adaptive locality-sensitive hashing, though we emphasize that to our knowledge, no other published algorithm has fixed time learning and best-match retrieval; see (Wang, Liu, Kumar, & Chang, 2016) for a relevant review.

Sparsey's coding field (proposed as an analog of the cortical macrocolumn) consists of *Q* WTA competitive modules (CMs), each containing *K* binary units. Thus, its sparsity is structurally fixed: all codes consist of exactly *Q* units, one per CM. The key to Sparsey is that its learning algorithm assigns more similar inputs to more highly intersecting SDR codes without requiring explicit comparison of a new input to the previously stored inputs. Briefly, this is accomplished by computing a measure of the input's familiarity, *G*, and using it to bias the softmax choice of winner in each CM. The more familiar the input, the greater the bias for the units in the CM that are contributing most to *G*. But these will be precisely the units that will have won and had their afferent weights increased for prior similar (correlated) inputs. Thus, as *G* goes to 1, the dynamics moves toward pattern completion, i.e., retrieval. As *G* drops toward zero (complete novelty), the activation function approaches the constant function, resulting in choosing a winner uniformly randomly in each CM, which maximizes the average Hamming distance of the newly chosen code with the set of already stored codes, i.e., pattern separation, as appropriate for learning a novel input.

Stepping back, we can ask: Is the brain's fundamental purpose to remember inputs / events that an organism actually experiences, i.e., episodic memory, or to learn the class structure of the world, i.e., semantic memory? Clearly, the vast majority of ML research has focused on classification and learning statistical (generative) models of domains, with relatively little concern for being able to recall, in full detail, individual experiences that may have occurred remotely in a model's operational life. In fact, *catastrophic forgetting*, in which new learning erases old learning has been a perennial problem for ML/DL, right up to the present, as discussed in (Kirkpatrick et al., 2017), which presents a novel solution.

Of course, real brains possess both episodic and semantic memory. The prevailing view in the computational neuroscience and DL-related communities is that the hippocampus forms memory traces based on single trials and replays the traces (e.g., during sleep), which gradually embeds (consolidates) permanent traces in cortex. The recent "memory networks" or "Neural Turing Machine" models (Graves, Wayne, & Danihelka, 2014) are consistent with this view. However, it nevertheless reifies the processor-memory distinction, which again, incurs higher power needs for moving data. In contrast, in Sparsey, since the process of assigning SDRs preserves similarity, the similarity/statistical structure of the inputs emerges, as a *computationally free side effect* of storing episodic memory traces, in the pattern of overlaps amongst those traces, i.e., the episodic and semantic memories are physically co-located. Sparsey may thus capture a more primitive cortical circuit that performed both functions, memory and classification.

Acknowledgements

Portions of this work were supported by ONR N00014-12-C-0539, DARPA FA8650-13-C-7432, and NIH 5 T32 NS07292.

References

Ganguli, S., & Sompolinsky, H. (2012). Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis. *Annual Review of Neuroscience, 35*, 485-508.

Graves, A., Wayne, G., & Danihelka, I. (2014). *Neural Turing Machines*. (arXiv:1410.5401).

Hinton, G., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation, 18*(7), 1527-1554.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. (arXiv:1503.02531).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS, 114*(13), 3521-3526.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). *Imagenet classification with deep convolutional neural networks*. Paper presented at the NIPS.

Rinkus, G. (1996). *A Combinatorial Neural Network Exhibiting Episodic and Semantic Memory Properties for Spatio-Temporal Patterns*. (PhD), Boston U., Boston.

Rinkus, G. (2010). A cortical sparse distributed coding model linking mini- and macrocolumn-scale functionality. *Frontiers in Neuroanatomy, 4*.

Rinkus, G. (2017). *AA Radically New Theory of how the Brain Represents and Computes with Probabilities*. (arXiv:1701.07879).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484-489.

Wang, J., Liu, W., Kumar, S., & Chang, S. F. (2016). Learning to Hash for Indexing Big Data - A Survey. *Proceedings of the Ieee, 104*(1), 34-57.