Towards A Distributed Federated Brain Architecture using Cognitive IoT Devices

Dinesh Verma
Distributed Cognitive Systems,
IBM T. J. Watson Research Center
Yorktown Heights, NY, U.S.A.
e-mail: dverma@us.ibm.com

Graham Bent
Emerging Technology Services
IBM UK
Hursley Park, Hants, UK
e-mail: gbent@uk.ibm.com

Ian Taylor
Cardiff University, Cardiff, Wales, UK
& University of Notre Dame, USA
email: TaylorIJ1@cardiff.ac.uk

***Abstract*—Cognitive Computer Systems (CCS) like IBM Watson implement a brain-like system in a centralized location. Limitations of current networks and organization structure necessitate the development of a distributed cognitive system, in effect a distributed federated brain. This distributed federated brain is composed of the different types of devices in the system, ranging from hand-held devices at the edge of the network to large systems in the cloud. It needs to demonstrate the properties of resilience, proactivity, agility and collaboration. In this vision, we discuss the factors that drive the need for the distributed brain, its technical requirements, and propose an architecture to attain the concept of a distributed brain for military coalition operations. We provide a roadmap that can attain this vision, moving intelligence from a centralized cloud location to a distributed collection of smart devices which are connected together using a cognitive Internet of Things technology.**

Keywords-distributed brain, IoT, distributed analytics, distributed learning, cognitive computing, symbolic vector representations

I. INTRODUCTION

Cognitive computing [1] refers to an approach for developing computer systems that augment human capabilities in a seamless manner. A Cognitive Computing System (CCS) consists of humans and software that work together, with the computer systems showing characteristics of a human brain to assist humans in an intuitive manner. CCS's improve their capabilities over time, learning from the environment and changing their characteristics without requiring manual programming or reprogramming. The IBM Watson Jeopardy [2] machine is a well known example of a CCS, but not the only cognitive system being worked upon. A CCS can be envisioned as the cyber-equivalent of a human brain implemented in software. A CCS can be viewed as a specialized instance of the broader notion of distributed cognition [3] which refers to a socio-technical system in which cognitive processing routines are distributed across the constituent social and technological elements

Most current industrial CCS's adopt a cloud-centric approach, with the brain component such as a deep learning algorithm being a centralized entity. Data, whether for training purposes, or for analysis, is uploaded to a central site, where the brain software processes it. While the centralized approach has proven successful in several domains, it does suffer from a number of limitations. When the data volume is large, the delays and costs associated with uploading the information to a central site, whether to a cloud site or a data center, may render cognitive computing solutions slow and expensive. As the processing power of distributed devices increases over time, decentralized cognitive solutions have the potential to become more responsive, scalable and inexpensive. Thus, there is a need to move from the paradigm of a centralized brain to a distributed brain. Furthermore, in many cases, the distributed brain may leverage assets across several administrative domains resulting in a federated distributed brain.

In this paper, we examine the challenge of creating a distributed federated brain, and propose an architecture and a roadmap for attaining this vision. The rest of the paper is organized as follows. Section II provides the motivating factors behind the need for a distributed federated brain. Section III provides a definition of the distributed federated brain, and discusses the technical challenges that need to be addressed to attain this vision. The following Section proposes a high level architecture which can be used to create a distributed brain. Section V discusses the concept of cognitive Internet of Things (IoT) and approach proposed for the physical realization of the distributed brain. Section VI discusses a possible evolution of the capabilities of such a system. Finally, Section VII lays out a roadmap for how the cognitive capabilities of the 'distributed federated brain' may increase over time.

II. MOTIVATIONS FOR DISTRIBUTED CCS

While the concept of a CCS has primarily focused on a centralized processing paradigm. When all entities within a network are connected together with a high speed reliable inexpensive network, centralized CCS has many advantages. However, there are many situations where such network connectivity is not present. Furthermore, there are several conditions under which a centralized CCS may underperform compared to a distributed CCS, even where the network connectivity is favorable.

Situations where network connectivity can be problematic include environments with mobile endpoints, including automobiles, ships, drones, trains and robotic mules, which need to move over a wide geographical area. Connectivity to a cloud site for such devices can only be provided by wireless communication networks such as cellular or satellite. These networks have high latencies and can be very expensive when a large amount of data needs to be transmitted.

 ISBN: 978-1-61208-531-9

In some areas, even cellular or satellite communications may be absent or be of poor quality. Mountainous terrains, hilly areas, or underground mines are likely to have poor network connectivity. There are many areas in the world where networking infrastructure is inadequate. In military coalition operations, which occur in areas with poor infrastructure, connectivity to a backend cloud system can be very sporadic and frequently absent. Any cognitive capabilities in these situations have to be provided in a manner that does not depend on continuous network connectivity to a cloud site.

Even when adequate network connectivity to cloud servers is available, there are many scenarios where a distributed CCS approach will be better. For instance, when devices are generating a significant amount of data, extracting insights from the data can be computed more efficiently near the location of data generation, as opposed to moving the data to a central location. As an example, consider a CCS which relies on video input to train itself. The code which extracts patterns from the video data, or finds interesting events in the video code, is likely to be much smaller, e.g. around a few Megabytes in size compared to streaming high resolution video, which at the rate of 4-8Mbps and can easily run into tens of Gigabytes. In these cases, it will be more efficient to move the code near the source of data, and extract patterns near the point where data is generated. The smaller of the two elements needed for cognition, code implementing intelligence or data which needs examination, needs to be moved for optimal performance.

Another scenario where distributed CCS is needed despite sufficient network connectivity occurs in relation to issues of regulatory compliance. Many types of data, e.g. healthcare data, are subject to regulations which may prevent it being sent to a central location for processing. Several countries restrict information on their citizens to be moved across borders. Extracting insights from data, subject to such restriction, requires a distributed cognitive infrastructure.

In some cases, security, privacy and licensing concerns may prevent the movement of data to a central location. In other cases, cost considerations may lead to a distributed cognitive infrastructure i.e. if a distributed cognitive system reduces the workload on the cloud site, it reduces the cost of cloud hosting. Furthermore, given the increasing processing capacity of end points like smart-phones, drones and robotic mules, this reduction in cost can be performed without impacting the cognitive capabilities of the system.

Because of the above motivating factors, we need to develop technologies that can enable distributed cognition that can leverage, but not be reliant on, a centralized infrastructure. The reasons for distributed cognition have a strong commonality with the driving forces behind approaches such as fog computing [4] or mobile edge computing [5].

III. DEFINITION AND TECHNICAL CHALLENGES

With the explosion in low cost phones, wearable devices and the IoT, future computing environments will have a diverse set of small elements capable of computation, storage and communication. Leveraging cognitive software on all of these devices leads to the concept of the distributed federated brain. The distributed federated brain is a socio-technical hybrid system capable of taking proactive actions based on the current and anticipated future situation on the ground. It is composed from the different types of devices present in the environment (sensors, hand-held devices, UAVs, robots, backend cloud computing sites, data center server farms etc.), along with the people who use those devices. This system provides a self-organizing self-healing predictive analytics capability, which is capable of functioning as a whole even when connectivity to the backend systems is missing. It will leverage all the services offered by a wired backend infrastructure (e.g. a backend cloud system, data center or available cellular network infrastructure) but it will not be critically dependent on a continuous form of connectivity to the backend.

The distributed federated brain operates seamlessly across networks and systems belonging to different organizations. In the context of military coalition operations, it uses assets belonging to coalition members or sub-groups within a single coalition member, while complying with any policies and guidelines required by individual coalition members. In the context of a civilian infrastructure, the brain uses assets across several enterprises and consumers, taking into account any restrictions imposed by the owners of the assets.

The distributed brain can analyze the situation on the ground in real-time, anticipate the situation likely to happen in the future, and determine whether the situation requires human involvement. If the situation does not require human involvement, the brain would undertake the most appropriate automatic action to the situation. When the situation needs human involvement, the brain will recommend alternative courses of actions, along with their pros and cons.

The brain is frequently charged with performing tasks that require creating dynamic groups on a short notice. Such dynamic groups may be transient and short-lived (days or hours), but could also last for a longer period (months). Differences in the pedigree of disparate systems belonging to different organizations necessitate the development of approaches that work with partial visibility, partial trust, and cultural differences, while simultaneously dealing with the challenges of a dynamically changing situation in which power, computation and connectivity may be severely constrained.

The ‘distributed brain’, therefore, needs to have several key properties. It must be self-healing and resilient, since it has to operate in an environment where elements may lose connectivity to backend systems, and any of the small component systems may disappear in an unpredictable manner. It has to react rapidly to changing situations on the ground, so it must be predictive and proactive in the decisions it makes. To deal with a dynamic environment, the system must be self-configuring, agile and adaptive. Since it is dynamically assembled from a large number of independent components, it needs to be a cooperative and collaborative collective of individual components. Humans and machines have different types of analytic and cognitive

 ISBN: 978-1-61208-531-9

capabilities. The 'distributed brain' must integrate human analytics capabilities into the machine analytics capability in a seamless manner.

A number of challenges confront the attempt to develop the distributed brain. Some insight into the nature of these challenges is provided by a consideration of the following four attributes:

Composability: How do we compose smaller elements into a larger aggregate that works like a seamless whole? What are the principles that link the attributes of a component to the larger whole, and how can we compose components belonging to different organizations with partial visibility and control in an environment with limited resources?

Interactivity: How do different computing elements and people interact with each other, both with other members of the groups and to external stimuli from the environment? How should we model and understand the interactions between different elements and information sources? How do different sub-brains work together as a larger aggregate brain?

Optimality: How can elements work together to obtain the optimal results in an environment with constrained resources? How can analytics be performed so that optimal performance is obtained automatically, instead of requiring complex manual optimization?

Autonomy: How can elements work together in a proactive manner understanding future situations sufficiently well to operate with a degree of autonomous behavior? How can a system determine that autonomous operation is inappropriate and human intervention is needed? How can different elements simplify the cognitive burden involved to best assist humans in the loop when intervention is needed?

The four attributes are not independent, and progress along any one attribute can positively address the attainment of the other attributes. If we want to decompose the problem further into relatively independent technical topics, we can identify six key topics which can collectively provide approaches to obtain the four attributes identified above. These six key topics are:

- *Software Defined Federations*: understand the principles by which different elements across a federated environment could be composed to form a virtualized larger element, and the properties of different types of architectures that enable such composition.
- *Generative Policy Models*: explore architectures and algorithms which enable devices in a federated environment to automatically determine their own operational policies, under the loose guidance of a higher level manager, but not be a slave to the higher level manager.
- *Agile Composition*: understand the architectures and principles which will allow different digital assets, such as code or data, to find each other in an optimal manner to generate insights.
- *Complex Adaptive Human Systems*: understand the properties of groups of humans working with machines, and understand how such groups would react to external stimuli and interact with other groups.
- *Instinctive Analytics*: create new techniques by which data and services can be automatically advertised, discovered and matched together to create analytics workflows that are autonomous and optimal.
- *Anticipatory Situational Understanding*: create new analytics approaches that can attain proactive situation understanding by autonomous systems and help create intelligent advisors for human-in-the-loop systems.

These six key topics are being investigated by an alliance of several universities, industrial and government research laboratories from the UK and the USA as part of the International Technology alliance in Distributed Analytics and Information Sciences (DAIS ITA) [6].

IV. ARCHITECTURE ENABLING DISTRIBUTED BRAIN

The ultimate goal of DAIS ITA is to investigate the basic science that would enable the creation of a distributed CCS that can perform analytics on demand across heterogeneous networks of interconnected devices. Some of the capabilities of such a system will be to (i) understand user requests for analysis, (ii) seamlessly compose the desired analytics functions from other functions and services available in the network, (iii) identify the right data set needed for the analytics, and (iv) bring together the data and analytics required to perform the function.

One approach that is being considered is a CCS architecture inspired by the cloud computing paradigm [7], comprising (i) a cognition layer, (ii) a platform layer, (iii) infrastructure layer and (iv) a management layer, with the six key topics mapping onto the layers as shown in Figure 1.

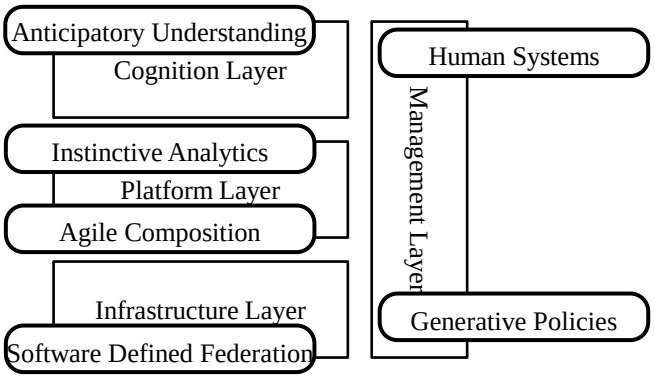

Figure 1. Key Topics in relation to layered architecture

An alternative approach maintains the CCS architecture, but considers the challenge from the perspective of an interacting network of cognitive micro-services. Our micro-services are cognitive in the sense that they comply with established principles of cognition such as those defined by the Core Cognitive Criteria (CCC) [8] which to a large extent incorporates the four key attributes of *composability*, *interactivity*, *optimality* and *autonomy* described above. In this approach micro-services are considered as semantic concepts and can be processed as such. It is in this sense our

 ISBN: 978-1-61208-531-9

distributed CCS can truly be described as a 'distributed federated brain.

Our challenge is to develop an approach where micro-services are self-describing, can self-discover other micro-services (including data services, network services, policy and security services) and where micro-services are self-allocating and self-provisioning in the sense that they can optimally position themselves or be invoked within a network to perform the tasks demanded by users. To achieve these goals, we require a common way to represent our cognitive services and their capabilities. Distributed cognitive processing is then the patterns of information flow and influence that occur across the network. The resulting cognitive phenomena are a property of the larger systemic organization, rather than a property of the individual micro-services.

V. EXAMPLE OF A COGNITIVE IOT

To illustrate how our 'distributed federated brain' concept might be realized, we consider how it can be applied to micro-service architectures in IoT context. Micro-services are an approach to developing a single application as a suite of small services, each deployed and running independently while communicating with each other via lightweight mechanisms. They typically require minimum centralized management and may be written in different programming languages and use different data storage technologies. They are widely adopted in the industry by companies like Netflix and Amazon, with a large number of developers, to streamline the software development lifecycle. They are also the basic building blocks of the IoT and can also be immensely useful in military and coalition scenarios being considered by the DAIS ITA, where each of the individual services may belong to different partners but a common goal needs to be achieved by composing them dynamically.

Applications that use micro-service architectures, may be composed of hundreds or even thousands of micro-services [9]. To be able to learn feasible composition of micro-services, dynamically compose new workflow graphs, and run learning algorithms on these workflows, we propose that micro-services self-describe in the form of vectors that capture not only the functionality that the service offers but also how it may be composed with other available services in the network, i.e., the feasible sequences of the service calls. These vector representations need to capture not just the semantic meaning of the service composition of which the micro-service is a part but also the order in which the micro-services are called. One possible representation is to use vector symbolic architectures such as the Holographic Reduced Representations (HRR) [10] which use convolution algebra for compositional distributed representations, a form of symbolic binding and unbinding. Other potential vector representations include binary spatter coding (BSC) or random permutation (RP) [11]. These types of representation have been demonstrated to be capable of supporting a wide range of cognitive tasks including reasoning [12], semantic composition [13], analogical mapping [14] and representing word meaning and order [15]. HRR's form the basis of the Semantic Vector Pointer Unified Network Architecture (SPAUN) [16] which claims to be a biologically plausible implementation of spiking neural network computation in a brain like manner, and can be used in military coalition contexts[17].

In our distributed brain model, we envisage micro-services being distributed across a heterogeneous network with micro-services being owned by different organizations. Rather than searching for micro-services and then centrally compiling a workflow, as in the standard service oriented architecture model, in our proposed model each micro-service learns its role (i.e. position in the workflow) in each of the service workflows in which it has been invoked and binds this into its own symbolic vector representation (this is an online learning task). Essentially, the resulting vector is the micro-service's memory of all of the workflow contexts in which it has been used. A user can request a high level task to be performed by declaratively specifying the precise service composition they require using a symbolic vector representation of the workflow. Alternatively, we are investigating how users can specify the service requirement (e.g. using natural language request) and the nearest matching service compositions are discovered automatically using semantic matching to automatically compute the corresponding symbolic vector representation. The resulting vector is broadcast to all nodes on the network that are capable of invoking micro-services and the micro-services respond by configuring themselves (self–provision and self-allocate) to match the request. Details of the online learning, matching and self-organization are described in [18].

VI. EVOLUTION OF COGNITIVE COMPUTING SYSTEMS

From an evolution perspective, we can envision how CCS's will progress over time. This is illustrated in Figure 2.

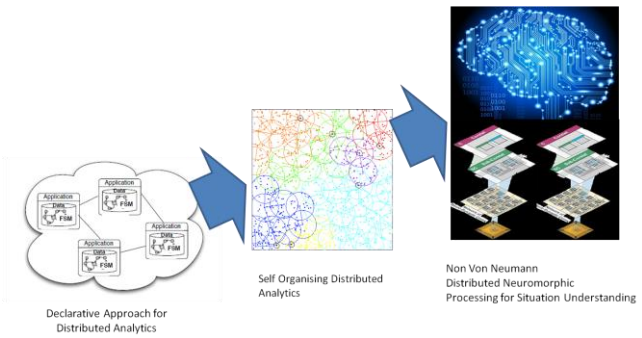

Figure 2. Evolution of a distributed federated brain

Our approach begins with the declarative approach described above, in which users request analytic services which are self-composed from multiple cognitive micro-services. Using extensions of the vector symbolic representations we are investigating the mechanisms by which these micro-services can best organize themselves not only in in response to user demand but also within the constraints of network availability, location of data, policy and security requirements. We envisage this self-organization to include the four principles of *composability*,

 ISBN: 978-1-61208-531-9

interactivity, *optimality* and *autonomy* outlined above, including the capability to for cognitive micro-services to learn from their environment and to exhibit proactive situation understanding and adapt accordingly.

The use of vector symbolic representation in the SPAUN architecture suggests that a future distributed federated brain may operate as a highly parallel non Von Neumann machine where the micro-services are themselves implemented as neuromorphic machines using spiking neural network processing. Such machines, which mimic the processing capabilities of the neo-cortex, have the distinct advantage of being extremely low power, operate at much lower frequencies than conventional microprocessors and potentially have less stringent bandwidth and latency requirements for inter service communication [14, 15]. Using a symbolic vector representation lends itself to both a conventional computing paradigm and to a neuromorphic computing model or a hybrid approach. For this reason, we believe that this is a fruitful area of research that will produce valuable insights as we move towards our goal of a distributed federated brain.

VII. A ROADMAP FOR LEARNING IN CCS

While the current state for cognitive computing systems is that of a centralized environment, the eventual state will be that of a fully distributed cognitive system with a peer to peer relationship among different nodes in the system. Learning is a core capability of such systems. From a learning perspective, the roadmap in Figure 3 shows one way in which cognitive systems may progress in their learning capabilities over time.

The working of any cognitive computing system can be defined into two distinct functions, the first being that of analyzing data to understand the patterns that lie within it, and the second one trying to assess the current situation on the ground, as to whether it matches one of the previously encountered patterns. We can refer to the first step as learning and the second as inference.

From a physical infrastructure perspective, we can divide the devices into two categories, cloud and edge. The cloud consists of a central location, while the edge consists of devices not in the cloud. Depending on the physical topology of the system, the edge may consist of mobile devices, sensors, gateways or other network devices.

In the roadmap shown in Figure 3, the current state is that of cloud based cognition in which the edge devices are just feeders of data. They send in information to the cloud based site, and the cloud performs both learning and inference for them.

The next stage of distributed cognition in CCS consists of the situation when learning happens in the cloud, while inference happens in the edge devices. In this stage, the cloud interprets all the data that it receives to create models of knowledge, e.g. a trained neural network, or a calculated decision tree, and sends that model to the edge devices that are not in the cloud. The edge devices use those models of knowledge to perform the task of inference.

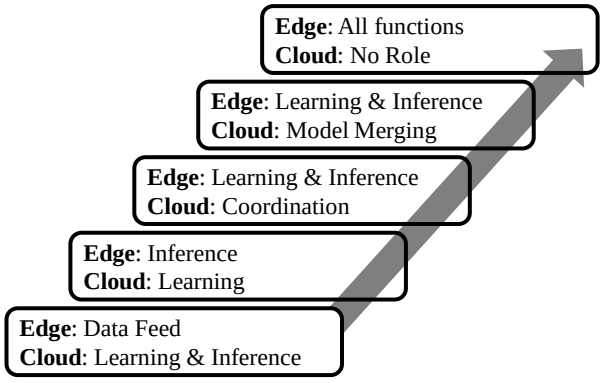

Figure 3. Roadmap for CCS Capabilities

Note that in this stage of CCS, none of the individual elements at the edge is cognitive on its own. However, when they are taken together, and the capabilities in edge devices combined with the capabilities in the cloud, cognitive computing capabilities are realized. A higher-level cognitive function is realized by the coordinated activity of distinct elements, each engaged in its own form of processing, some of which is cognitive (e.g. learning in the cloud) and some of which is not cognitive (functions at the edge-devices).

In the third stage, the edge devices perform both tasks of inference and learning, but rely on the cloud for coordinating their learning. The cloud can provide coordination such as directing different edge nodes to learn about different types of information, and then share the learnt models with each other. As an example, the cloud may instruct one edge node to learn models for identifying cars, another to learn models for identifying trucks, and yet another to learn models for identifying planes. The models can then be exchanged among the edge-devices, each of whom benefit from the models learnt by the other edge devices. In this stage, cognitive functions are enabled at the edge, while the task in the cloud, that of coordination, can be considered non-cognitive (standard computer processing). The net result is a distributed cognitive computing system.

In the next stage, the edge devices learn models that may potentially be for the same type of information. Since models learnt by one edge device may not always match with the models of the other edge device, a merging of the models needs to be performed. The cloud provides this capability for merging models. In this stage, both the edge devices as well as the cloud based system are performing cognitive processing. Distributed cognition is obtained by a combination of many different cognitive elements, some on the edge and some in the cloud.

In the final stage of distributed cognition, the role of the cloud can be dispensed with, and the merging of the models happens using peer to peer information exchanges among the edge devices.

 ISBN: 978-1-61208-531-9

ACKNOWLEDGMENT

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

REFERENCES

[1] J. Kelly III & S. Hamm, Smart Machines: IBM's Watson and the Era of Cognitive Computing. Columbia University Press, 2013.

[2] D. Ferrucci, Introduction to “this is watson”. IBM Journal of Research and Development. vol 56, no 3, May 2012.

[3] J. Hollan, E. Hutchins & D. Kirsh, Distributed cognition: Toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction (TOCHI), vol. 7(no 2), pp. 174-196, 2000.

[4] F. Bonomi, R. Milito, J. Zhu and S. Addepalli, Fog computing and its role in the internet of things, Proc. First ACM workshop on Mobile cloud computing, Aug 2012

[5] A. Ahmed and E. Ahmed, A survey on mobile edge computing. Proc. IEEE International Conference on Intelligent Systems and Control (ISCO), Jan 2016.

[6] T. Pham, G. Cirincione, A. Swami, G. Pearson, & C. Williams, Distributed analytics and information science. In IEEE International Conference on Information Fusion (Fusion), July 2015.

[7] W. Kim, Cloud computing architecture. International Journal of Web and Grid Services. Jan 2013.

[8] C. Eliasith, How to build a brain, Oxford University Press 2013

[9] Microservices. https://developer.ibm.com/cloudarchitecture/2016/10/20/meeting-microservices/

[10] T. A. Plate, Holographic reduced representations.IEEE Transactions on Neural Networks, 6 , 623– 641.(1995)

[11] P. Kanerva, Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors, Cogn Comput (2009) 1:139–159 DOI 10.1007/s12559-009-9009-8, January 2009

[12] D. Widdows and T. Cohen, Reasoning with Vectors: A Continuous Model for Fast Robust Inference, Log J IGPL; 23(2):141–173 October 2015.

[13] A. Neelakantan, B. Roth, A. McCallum, Compositional Vector Space Models for Knowledge Base Inference, Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches: Papers from the 2015 AAAI Spring Symposium

[14] R.W. Gayler and S.D. Levey, A distributed basis for analogical mapping,

[15] M.N. Jones and D.J.K. Mewhort, Representing word meaning and order information in a composite holographic lexicon, Psychological Review 2007, Vol. 114, No. 1, 1–37

[16] C. Eliasmith et. al,. A large-scale model of the functioning brain. Science. vol 30 no. 338, pp. 1202-1205, Nov. 2012.

[17] F. Bergamaschi et al, Smart coalition systems: a deep machine learning approach, Human-machine Interface and Machine Learning Approaches II, SPIE Conference 10190, April 2017 (in prep)

[18] P.A. Merolla et al., A million spiking-neuron integrated circuit with a scalable communication network and interface. Science. 345 (6197): 668. doi:10.1126/science.1254642. PMID 25104385.

 ISBN: 978-1-61208-531-9