Holographic Reduced Representations for Working Memory Concept Encoding

by
Grayson M. Dubois

A thesis presented to the Honors College of Middle Tennessee State University in partial fulfillment of the requirements for graduation from the University Honors College

Spring 2016

Introduction

The field of artificial intelligence (AI) is synergistic with a wide range of disciplines but artificial neural networks (ANNs) is perhaps the most prolific subfield. Not only are biological principles of neural computation and neuroanatomy adapted to solve engineering problems, but ANNs also serve as formal, testable hypotheses of brain function and learning in the cognitive sciences. Still, since ANN models often employ distributed encoding (DE), most have limited application in other areas of AI where symbolic encoding (SE) is the norm (e.g. planning, reasoning, robotics).

For example, there is extensive evidence that the brain contains a working memory (WM) system that actively maintains a small amount of task-essential information that focuses attention on the most task-relevant features, supports learning that transfers across tasks, limits the search space for perceptual systems, provides a means to avoid the out-of-sight/out-of-mind problem and more robust behavior in the face of irrelevant events [1,2]. The prefrontal cortex and mesolimbic dopamine system have been implicated as the functional components of WM in humans and animals, and biologically-based ANNs for WM have been developed based on electrophysiological, neuroimaging, and neuropsychological studies [3]. A software library, the working memory toolkit (WMtk), was developed to aid the integration of ANN-based WM into robotic systems by mitigating the details of ANN design and providing a simple DE interface [4].

An example of the capabilities of the WMtk can be seen in a robotic simulation written using the toolkit based on the *delayed saccade task (*DST) [4]. In the DST, the robot is required to focus attention on a crosshair in the center of the screen. After a variable time delay, a target object will appear in the periphery of the screen, but the robot must continue to focus on the crosshair in the face of this distraction. After some time, the target object disappears and the robot must continue to focus on the crosshair. Finally, the crosshair disappears and the robot must then look at (or *saccade* to) the location where the target object appeared during the task. Rather than programming the robot to solve the DST, the WMtk allows the robot to *learn* how to solve the DST by repeatedly attempting the task as a series of *episodes*. The robot's WM learns to both override automatic behaviors (such as immediate saccades) and store task-relevant information (such as target locations) in order to guide future actions. Importantly, the robot is given feedback (positive reward) only at the very end of correctly performed episodes. Even under these conditions, the WMtk learned to correctly manage items in WM and attain proficiency on the DST within just hundreds of episodes.

Despite the fact that the WMtk can solve common tests of working memory performance such as the DST, the DE/SE distinction is problematic for the WMtk since DE/SE conversion needs to be programmed directly by the user and tuned specifically to each learning task. A technique called holographic reduced representation (HRR) [5] may provide the technical assistance needed to overcome this limitation. HRRs provide a framework for creating and combining symbolic concepts using a distributed formalism that is compatible with ANNs. The name HRR summarizes how many different concepts, each represented by separate, unique vectors, can be combined and reduced to a single vector that represents the combined knowledge of the concepts while still retaining information about each constituent concept which is closely related to the concept of holographic storage. By replacing the DE interface of the WMtk with an HRR interface, DE/SE conversion would be automated, concepts learned from one task would naturally carry over to new tasks, and additional cognitive phenomena (e.g. chunking) may be investigated. *Therefore, our specific aim is to develop and test a holographic reduced representation engine, and integrate it with the Working Memory Toolkit.*

Methodology

Work on this project will be performed in 2 phases, each separated into three parts. The first phase will take place during the spring semester, and during that time I will create an engine to generate and manipulate HRRs. This engine will provide me with a means of working with simple and complex concepts such as those used by the WMtk. It may be useful to other researchers in the cognitive sciences who wish to follow an HRR approach to concept encoding, but the primary purpose of this engine is to augment the WMtk to allow more powerful manipulation of concepts as well as to simplify the interface for users programming with the toolkit. This augmentation of the WMtk is the second phase of my thesis project. Over the summer, I will add the HRR engine to the toolkit, and make the necessary updates to the WMtk source code to fully utilize the capabilities of the HRR engine.

Phase 1: Creating an HRR Engine for concept encoding. My 3-part plan for creating the HRR engine consists of (1a) learning and gathering information, (1b) developing a conjunctive encoding engine, and (1c) creating a conjunctive decoding engine. These phases will be implemented as follows:

1a) Gathering information / background for HRRs and HRR generation. I will need to spend the first month of this project researching methods for generating HRRs, and manipulating them to encode and decode conjunctive concepts using circular convolution. I will identify the mathematical formulas necessary and organize the steps required to achieve encoding and decoding of HRRs.

1b) Developing a conjunctive encoding engine. Eventually, I will have enough information to write the algorithm for conjunctively encoding HRRs, and will create a library in C++ that can be included in any future projects. This encoding will consist of taking concepts in the form of common words, creating a representation for each one, constructing a table of each concept and all possible combinations of concepts. As an example, the encoding engine could take two concepts, “blue”, and “horseshoe”, and create HRRs for each, as well as a new HRR for the complex concept “blue horseshoe.” The encoding engine will be complete by the end of March 2016.

1c) Developing a conjunctive decoding engine. As decoding is much more complicated than encoding, I will create the decoding capabilities of the engine during the third phase of the process. Using the information gathered in Phase 1a, I will write and optimize the algorithm for decoding

HRRs, and add it to the HRR generation engine created in Phase 1b. The decoding capabilities of the engine will include taking a complex concept (such as **ab**) and another concept (**a**), and performing involution operations on them to extract the other constituent concept (**b**). In this way, we could perform an involution operation on the unique concepts "blue horseshoe" and "blue" to retrieve the concept "horseshoe."

Phase 2: Integrating the HRR engine into the Working Memory toolkit. Once I have created the HRR engine, I will spend the summer working on incorporating it into the Wmtk. The next three steps of the project are: (2a) learning the mechanics of the WMtk, (2b) altering the toolkit to utilize the HRR engine for concept encoding and manipulation, and (2c) testing the capabilities of the augmented toolkit against the original.

2a). Learning the mechanics of the Working Memory toolkit. In order to make the necessary changes to the WMtk to incorporate the HRR engine, I will need to understand its design and functionality. I will spend the first month of summer break learning temporal difference learning algorithms utilized in the toolkit, as well as the chunking of concepts in working memory and how the adaptive critic utilizes this information to determine whether an action is favorable or not.

2b). Adding the HRR engine to the Working Memory toolkit. Once I have a firm grasp of how the toolkit functions, I can begin replacing the existing concept encoding interface with the HRR engine. I will set up the engine to automate the process of generating representations for concepts, and establish the capabilities of quickly and easily manipulating complex concepts in working memory. This will make it easier for future users of the toolkit, as they will no longer need to manually construct the representations for the concepts they use in their simulation. Instead, they will be able to create a list of concepts, and the HRR engine will automatically handle the encoding, decoding, and manipulation of these concepts.

2c) Testing the capabilities of the augmented toolkit against the original. Towards the beginning of the fall semester, the HRR engine will be fully integrated into the WMtk, and I can begin qualitatively and quantitatively testing its capabilities. The primary test I will perform will be to create a new version of the DST, similar to that which was created using the original toolkit [4]. I will check for qualitative improvements in the ease of use in terms creating a new simulation using the toolkit, and will quantitatively compare the learning rates using HRRs as opposed to the manually-encoded concept representations used in the original toolkit.

Results and Future Impact

Adding a holographic reduced representation engine to the Working Memory toolkit will open up many possibilities for future research and development in modeling working memory. Not only will it antiquate the tedious process of manually programming representations for concepts, but it will also give the toolkit the power and versatility to allow future researchers to tackle more complex issues in cognitive science, such as transferability of learned information between similar tasks, chunking similar concepts together in memory, and storing concepts across a continuum, rather than in discrete space. For example, colors could be represented across a scale, rather than storing a few base colors as concepts, and using complex concepts formed from combinations of these base colors (like red and blue) to represent other colors (such as purple). With the augmentation of the WMtk, I will be opening up the playing field to cognitive scientists to more effectively solve issues in ANN-based computing, and discover more about how the human brain functions.

Appendix

AI – Artificial Intelligence

ANN – Artificial Neural Networks

DE – Distributed Encoding

SE – Symbolic Encoding

WM – Working Memory

WMtk – Working Memory toolkit

HRR – Holographic Reduced Representation

References

[1] A. Baddeley. *Working Memory*, volume 11 of *Oxford Psychology Series*. Clarendon Press, Oxford, 1986.

[2] N. C. Waugh and D. A. Norman. Primary memory. *Psychological Review*, 72:89–104, 1965.

[3] R. C. O’Reilly, D. C. Noelle, T. S. Braver, and J. D. Cohen. Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex*, 12:246–257, 2002.

[4] J. L. Phillips and D. C. Noelle. Working Memory for Robots: Inspirations from Computational Neuroscience. in *Proceedings of the 5th International Conference on Development and Learning*, 2006.

[5] T. A. Plate. Holographic reduced representations. *IEEE Trans. Neural Networks*, vol. 6, no. 3, pp. 623–641, May 1995.