**Abstract**

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. 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. However, the DS/SE distinction is problematic to the toolkit since the DE/SE conversion has to be programmed directly by the user and tuned specifically to each learning task. A technique called holographic reduced representation (HRR) may provide 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. We wrote a software engine for managing HRRs and rebuilt the WMtk around the HRR Engine (HRRE). The HRRE automates the encoding process and manages a store of known concepts, while allowing the user to program the WM component using a simple string-passing interface for the state and environment data. By replacing the DE interface of the WMtk with the HRRE, DE/SE conversion is automated, concepts learned from one task will naturally carry over to new tasks, and additional cognitive phenomena (e.g. chunking) may be investigated.

**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. 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. 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.

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) 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.

**Background**

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.

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 was to develop and test a holographic reduced representation engine, and integrate it with the Working Memory Toolkit.

**Methods**

Work on this project was performed in 2 phases, each separated into three parts. During the first phase we created an engine to generate and manipulate HRRs. This HRR Engine (HRRE) provided us with a means of encoding, storing, and manipulating representations of concepts used in the WMtk. During the second phase, we rebuilt the WMtk around the HRRE, replacing the original DE interface with a simple string-passing interface, and automating the encoding and manipulation of concepts using HRRs.

**Phase 1: Creating an HRR Engine for concept encoding.** Our 3-part process for creating the HRRE consisted of (1a) researching holographic reduced representation, (1b) developing a conjunctive encoding engine, and (1c) developing a conjunctive decoding engine. These phases were implemented as follows:

*1a) Researching Holographic Reduced Representation*. HRR is a robust method of representing symbolic concepts in a distributed form that can be combined to make holographic representations for complex concepts containing the information for each of the constituent concepts. With HRRs, it is possible to use symbolic concepts with ANNs. It works by generating a vector of double values, calculated in various ways depending on what type of HRRs you want to create. We used unitary HRRs in our engine. This distributed vector can then be assigned to the data structure that holds your concept. We used a string, where the string value is just the name of the concept we wanted the HRR to represent. For example, to generate an HRR for the concept *red*, we generated a unitary HRR, and assigned it to the string value “red”.

The base data structure for our HRRE is a dictionary, where the string name for the concept is the key and the HRR representing that concept is the value.

*1b) Developing a conjunctive encoding engine.* After setting up the base dictionary for the engine’s concept memory and building the functionality for HRR generation, we needed to add a conjunctive encoding function to the engine for combining the HRRs to form complex concepts. We combine the representations for the concepts using the circular convolution operation. This operation is what makes holographic representations reduced, as it combines information from two HRRs into a single HRR of the same size. Tony Plate gives a method of circular convolution by constructing the outer product of the two HRRs and summing the elements along the trans-diagonals, as indicated in the figure taken from Tony Plate’s paper. [Figure 4 from Plate’s paper on Circular Convolution]. This order n­2 operation (where n is the size of the HRRs being convolved) results in an HRR of size n with information from both constituent HRRs. This operation can be performed in order n time, however, by implementing fast-fourier transforms (FFT). We use FFT in the HRRE for convolution operations.

*1c) Developing a conjunctive decoding engine.*