The Pennsylvania State University

The Graduate School

MODEL OF VISUAL WORKING MEMORY BASED ON VISUAL KNOWLEDGE

A Dissertation in

Psychology
by

Shekoofeh Hedayati Zafarghandi

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2022

The dissertation of Shekoofeh Hedayati Zafarghandi was reviewed and approved by the following:

Brad Wyble
Associate Professor of Psychology
Dissertation Advisor
Chair of Committee

Richard Carlson
Professor of Psychology

Abhronil Sengupta
Assistant Professor of Electrical Engineering

Sharon Huang
Associate Professor of Information Sciences and Technology

Kristin Buss
Professor of Psychology and Human Development and Family Studies
Department Head

ABSTRACT

In the study of human cognition and cognitive neuroscience the concept of working memory has been a keystone capacity, which allows us to temporarily hold information, manipulate it, reproduce and store it into long term memory.

From its inception as a theoretical capacity dating back to 1970's, it has been assumed that working memory draws on existing long-term memories, as is obvious from our ability to easily remember complex spatial configurations that we are already familiar with (e.g., characters, words, images).

Despite widespread understanding that long term memory plays a key role in working memory, there have not yet been any computational models of how this might occur. By leveraging recent innovations in deep learning and combining them with decades of theory and behavioral data in both working memory and computational neuroscience, we provide the first model that meets this challenge. Our model describes how a neurocomputational memory system for latent representations (MLR) can store multiple items with just a single exposure to each one. The features of each item bound together, allowing each item to be individually retrieved either by its features or its sequential order.

We further explored the proposed model in other theories related to working memory, such as dual coding, visual creative imagery, and visual search. The model's simulations showed that verbal code is unnecessary in a task that requires to store visual detail information. Moreover, the simulations shed light on how combinatorial imagery might occur, and what neural codes might be used in a visual search paradigm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

During my PhD program, I received a great deal of support and assistance. First, I would like to thank my husband, my best friend and my collaborator, Pooyan Doozandeh. I am also grateful to my parents Maryam Sheikhaliyan and Ali Hedayati, and my brother Amir Hedayati for their love and sacrifices.

I am indebted to my advisor Brad Wyble who believed in me, and gave me constructive advice throughout these years. He has been a great mentor who taught me generosity, patience and kindness beside science.

I would like to express my gratitude to my other mentors and the committee members, Richard Carlson, Abhronil Sengupta and Sharon Huang for their helpful thoughts and comments on my dissertation.

Finally, special thanks to my lab mates and friends, Chloe Callahan-Flintoft, Ryan O'Donnell, Joyce Tam, Mandana Saebi, Elham Rahimi, Younes Shekarian, Ellie Nasr Azadani, Younes Karimi, Maryam Seifollahi and Akbar Mahdavi Shakib who have always been helpful and inspiring.

Part of this dissertation was supported by NSF grant no. 1734220 and Binational Science Foundation grant no. 2015299. The findings and conclusions do not necessarily reflect the view of the finding agencies.

Chapter 1

Introduction

Memory __i.e., the ability to encode and retrieve information__ is one of the core cognitive faculties studied by psychologists over decades. Based on how long a given information can be retained in memory, there are different kinds of memory such as sensory, short-term (i.e., working) and long-term memory (A. Baddeley, 1992; A. D. Baddeley, 1966; A. D. Baddeley & Warrington, 1970; G. A. Miller, 1956; Sperling, 1960). Sensory memory, that captures the information from our senses, lasts for less than a second (Sperling,1960). Short-term or working memory maintains the information for several seconds to minutes, while it also allows for manipulating the information (Baddeley,1966). Working memory has a limited capacity, but it is critical in doing complex tasks, as it enables us to keep track of information. For instance, having someone's address in mind, while receiving a direction from another person is feasible via working memory. Long-term memory, on the other hand, is a type of memory that holds information for more than several minutes, possibly for a life span. Long-term memory has a much higher capacity than working memory, for example, our semantic knowledge of the world (e.g., when World War II began) is under the category of long-term memory.

In the memory literature, there has been this notion of dependency of working memory on long-term memory since Atkinson & Shiffrin (1968). They proposed a model of memory called *modal model* that integrates sensory, short- and long-term memory at various stages of processing (Figure 1). In this view, the incoming sensory information derived from our senses (e.g., vision, auditory, etc.) enters the sensory memory, and is temporarily retained there for less than a second. Afterwards, this information is transferred into short-term memory, where it can

last for several more seconds. Based on this view, if the information is left unattended, it will decay in after several seconds. However, rehearsing the information using the control processes can keep the information for a longer period of time in short-term storage. Part of the short-term memory content can be copied over to long-term memory to be kept more permanently.

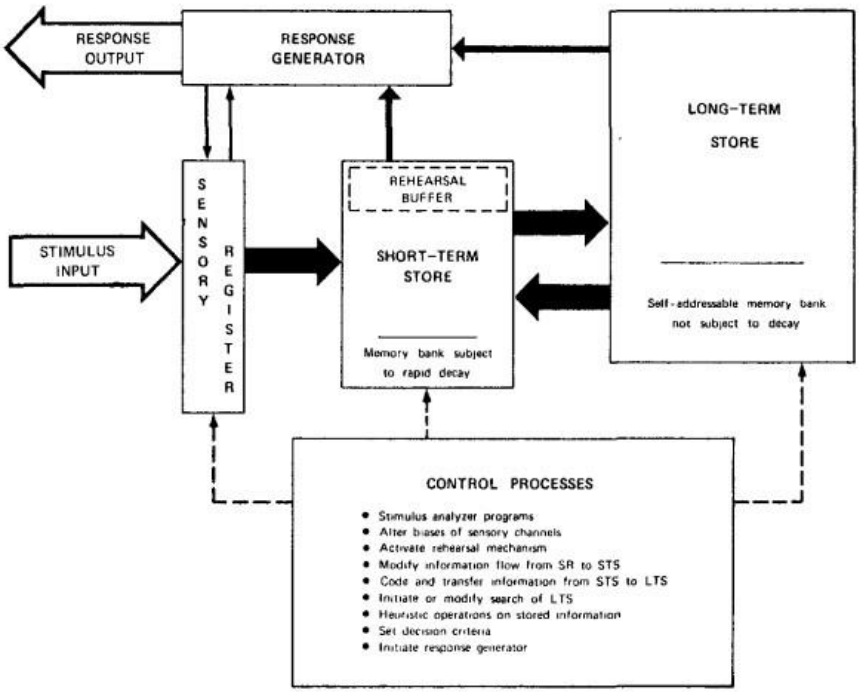

Figure 1. Schematic of modal model of memory consisting of sensory, short-term and long-term memory with control processes (Source, Atkinson & Shiffrin, 1968). Solid lines demonstrate the path for transferring information, however, this transfer does not mean that the content is removed from one component and is transferred to another. Rather, the information is copied from one storage to another one. Dashed lines indicate control signals (e.g., rehearsal, transfer, etc.) that are applied to different part of the system.

Consistently, Baddeley and Hitch (1974) proposed a multi-component model, in which visual short-term memory (i.e., visual sketchpad) is bidirectionally connected to visual semantics and episodic long-term memory (Figure 2). In this view, working memory is further divided into three subsystems of central executive and its slave components of visuospatial sketchpad and phonological loop. Central executive is the primary controlling system that regulates the flow of information into each of the subsystems. For instance, it decides what information should be stored in the visuospatial sketchpad based on the task demand. Visuospatial sketchpad and

phonological loop are the short-term storages to store visuospatial and verbal/acoustic information respectively. The reason for this distinction was derived from behavioral experiments in which people showed little to no interference when storing visual and verbal information concurrently compared to when they store only one type of information (visual or verbal). However, interference increased when information that was encoded were both visual or verbal (Baddeley & Hitch, 1974). Consistently, according to the articulatory suppression effect, verbal information is more impaired when people articulate an irrelevant word while they are doing a task causing the verbal information held in the phonological loop to decay (Richardson & Baddeley, 1975). Information in the short-term storage can be transferred to long-term memory in the form of language or visual semantics for verbal and visual information respectively, or in the form of episodic long-term memory irrespective of its type (visual or verbal). Note that this transfer is bidirectional, meaning that long-term storage can modulate the content of short-term memory.

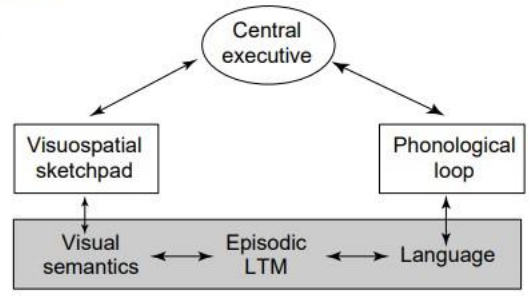

Figure 2. Multi-component model of working memory integrated with long-term memory (Source: (A. Baddeley, 2000). Working memory is composed of central executive, visuospatial sketchpad and phonological loop. Arrows indicate the flow of information for different parts of the system. Short-term storage of visual and verbal information can be transferred into long-term memory (LTM). Also, information in long-term memory can be transferred to short-term memory.

Another dominant account that linked working memory to long-term memory was proposed by (Ericsson & Kintsch, 1995). They proposed a unitary mechanism in which working memory was considered to be the part of long-term storage that was activated via focus of attention (Figure 3). In this embedded view that was further refined by (Cowan, 1988, 1999,

2001), working and long-term memory are not separate systems, rather working memory consists of activated parts of long-term memory in addition to attentional focus that determines what parts are to be activated.

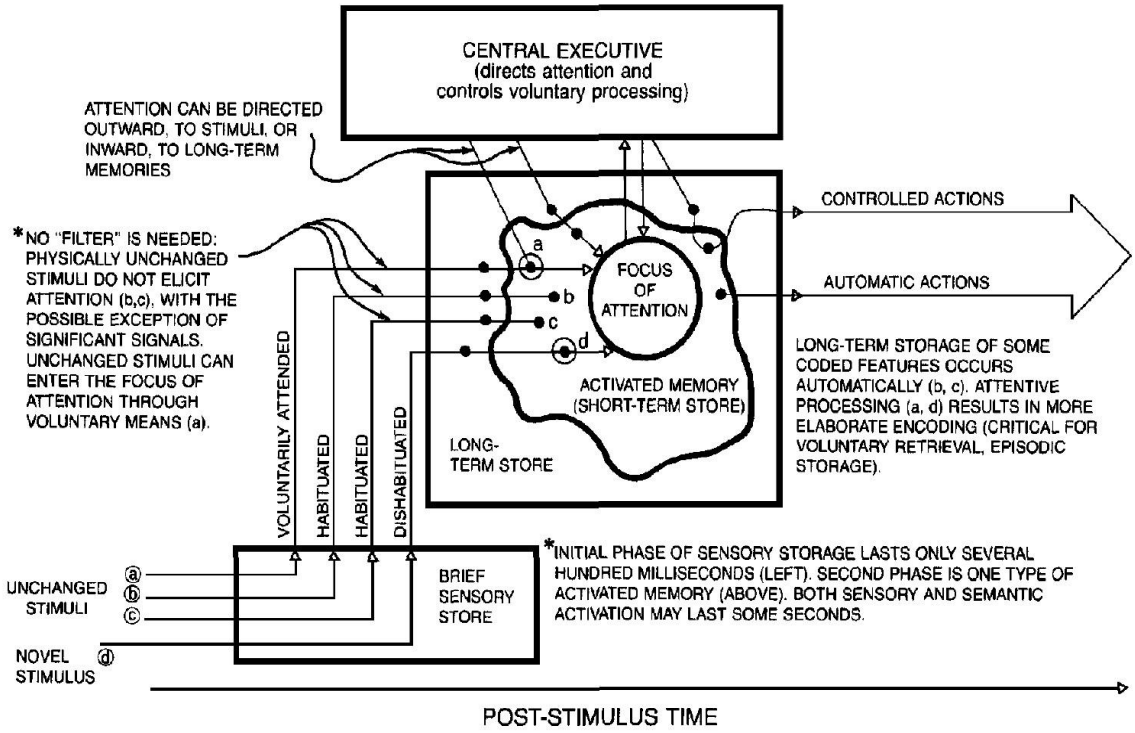

Figure 3. Schematic of activated long-term memory model (Source: Cowan, 1988). In this framework, short-term storage (i.e., working memory) is embedded as a subsystem within the long-term memory and is activated via the focus of attention. X-axis is the timing at which a given stimulus is perceived until it is stored in long-term memory. Arrows indicate the transfer of information from one form to another form. Central executive controls the attentional deployment and other voluntary processes.

Although the above models provide different explanations on stages of information processing with regard memory, they are all consistent with abundant behavioral data suggesting that working memory is inseparable from long-term knowledge (Alvarez & Cavanagh, 2004; Chen & Cowan, 2005, 2009; Hulme et al., 1991; Ngiam et al., 2019; Yu et al., 1985; Zhang & Simon, 1985; Zimmer & Fischer, 2020).

In this regard, Zimmer and Fischer (2020) conducted a series of experiments on two groups of Chinese literates and people unfamiliar with Chinese characters. They found that when showing Chinese characters, the former group can remember more items than the latter. The

results indicated that memory capacity is higher for familiar items that already existed in long-term memory. Consistently, (Brady et al., 2016) demonstrated that people have more working memory capacity for natural and meaningful images compared to arbitrary simple shapes such as colored squares, for they have had past experiences with natural objects. They concluded that our visual knowledge of long-term memory modulates working memory capacity.

Taken all these together, the importance of long-term knowledge in studying working memory has been acknowledged by numerous studies and were considered in major descriptive memory models, such as the modal model (Atkinson & Shiffrin, 1974), multi-component (Baddeley & Hitch, 1974) and activated long-term memory models (Cowan, 1988). These accounts have been useful to consider the implications of working and long-term memory integration in studying memory, but they barely provide explanations of how these memories are linked mechanistically and why we see certain effects in memory tasks. For example, we know that a first-grade student can temporarily remember a novel English character that she has not seen before. Similarly, Hedayati et al., (2022) conducted an one-trial experiment in which they showed people a hand-written Bengali character, and subjects could recognize the character among three other Bengali characters after 2.5 seconds with 95% accuracy without even knowing that they were participating in a memory task. If short-term memory is integrated with long-term memory, then how do we form rapid memories of configurations that we have not seen before? The above models cannot explain how novel memory formation occurs.

In another point, familiarity with a given object would make it more efficient to store in working memory (Zimmer & Fischer, 2020). For example, if we are given a series of digits, it is much easier to remember them if the sequence is drawn from a familiar phone number than if they are some random numbers. Nevertheless, none of the above accounts can explain why working memory is more efficient for familiar information.

The above questions are fundamental to understand the nature of working memory and its relation to long-term memory. We contend that without a working model it is really hard, if not impossible, to address the above questions and build an intuition of how working memory representations are formed.

In this regard, using recent advances in deep learning, we proposed a computational model that can provide mechanistic answers to the aforementioned questions and endow a new conceptualization of working memory via implementation.

The proposed model is named Memory for Latent Representations (i.e., MLR) and is comprised of two core elements: working memory and visual knowledge. Note that in the context of our work, visual knowledge emerges from long-term familiarity with particular shapes, or statistically common featural combinations that enables us to recognize and remember complex objects (i.e., the prototypical shape of a car, or the strokes that comprise a digit). The visual knowledge is achieved by training a neural network on sets of images to adjust the synaptic weights. As a result, visual knowledge constitutes a part of long-term memory that includes familiar items. These familiar items were drawn from a distribution that the model was trained on. For example, if the model was only trained on exemplars of cars, then cars would be familiar items, whereas bicycles that were not in the training set, would be considered as novel items.

Additionally, MLR is constrained by behavioral and neural data, meaning that it generates some core behaviors similar to human subjects' behavior in working memory tasks. Also, the model's architecture is consistent with the human visual system, for layers of the model correspond to levels of the visual ventral system from retina to inferotemporal cortex (IT cortex; Figure 5b). For example, when we see a car, an image of a car appears on the retina. This retinal image is transferred into the brain to be further processed by different levels of visual ventral stream such as V1, V4 and IT cortex. The neuronal representation becomes progressively compressed (i.e., represented by a smaller number of neurons) as information is transferred to the

higher levels of the visual system (Figure 5b). Furthermore, these representations become more abstract as they move from early levels (e.g., V1) to later levels (IT cortex). In this regard, the layers of the visual system in the MLR model constitute the visual knowledge hierarchy formed by compression of data, as visual ventral system represents familiar visual patterns with fewer neurons at successively later levels of the pathway (i.e., V1, V2, V4, IT) of the visual system (Bates & Jacobs, 2020).

In the context of neural networks, representations in latent spaces (i.e., latent representations) are basically the compressed form of input, such that similar items presumably reside closer to each other within a manifold. In this sense, we argued that latent spaces in a neural network can be mapped into the visual ventral system to form the visual knowledge hierarchy.

The MLR model simulates how latent representations of items embedded in the visual knowledge hierarchy are encoded into WM depending on their level of familiarity. Subsequently, the encoded items in WM can be retrieved by reactivating those same latent representations in the visual knowledge system.

As described earlier, we considered sets of constraints (behavioral and neural) in designing the MLR model. These constraints are necessary to make inferences about the underlying cognitive mechanisms, and this is how cognitive models are in contrast with commonly used neural network models that are used in other technical fields. The purpose of cognitive computational models is to understand a mechanism underlying cognitive constructs. The assumption is that when a given model generates a specific behavior, we can infer how that behavior was generated by looking into the internal machinery of the model. As a result, the goal of these models is not to achieve a better performance compared to the existing state-of-the-art deep learning models (e.g., better image classification accuracy), rather they aim to shed light on *how* a given cognitive function works, while it conserves some levels of biological plausibility.

That is, to legitimately expand theoretical conclusions about cognitive mechanisms there also needs to be biological constraints that limit the model's architecture. Cognitive models as such are especially useful to build a functional intuition of cognitive mechanisms. Moreover, since these models are actual implementations, they contain less ambiguity over descriptive models. Furthermore, their flexibility to simulate various conditions using different stimulus sets often result in insightful ideas that can be further tested via behavioral/neural experiments.

Brief Informal Description of MLR

Memory formation in MLR can *rapidly* and *selectively* encode specific attributes (e.g., shape, color, label) from one or more visual items within a distributed neural representation using a tokenized binding pool (Swan & Wyble, 2014). In MLR, information is encoded with varying levels of efficiency depending on the degree to which it matches representations embedded within the pre-trained visual knowledge hierarchy. The visual knowledge is built using gradient descent to train a modified variational autoencoder (VAE) on a set of stimuli combining handwritten digits (MNIST) and articles of clothing (fashion-MNIST; (LeCun, 1998; Xiao et al., 2017).We chose to build our model based on a fully connected VAE rather than a more complex convolutional network, because it is simple in terms of layers, and generates smooth latent spaces. More complex models would provide more detailed reconstructions, but our goal is to develop a clear and explainable theory rather than an optimized memory system.

In a trained VAE, familiar stimuli are encoded efficiently into a small-dimensional latent space (analogous to IT cortex) and classified into categorical labels, while novel stimuli can only be represented into higher-dimensional latents closer to the beginning of the visual pathway (analogous to V1 cortex).

As illustrated in Figure 5a, a visual stimulus is processed by the feedforward portion of the visual knowledge system to produce compressed shape and color representations of the object as well as a categorical label of each attribute. A binding pool stores a representation of selected features and/or labels according to a set of tunable parameters. These parameters control the proportion of different kinds of information that flow from the knowledge hierarchy into the memory trace. Each memory trace binds visual forms, colors, and categorical labels together into a single tokenized representation that can be stored alongside other tokenized representations of objects.

A clarifying assumption of MLR is that the memories are stored in a particular group of neurons that are allocated specifically to the role of memory storage and sit apart from the sensory areas themselves (Figure 4a). This account is in accordance with classic theories of prefrontal cortical involvement in working memory (Goldman-Rakic, 1995; E. K. Miller et al., 1996). This can be contrasted with models that imply distinct representations for visual and non-visual forms of memory (Baddeley & Hitch, 1974; Figure 4b), and embedded process models that distribute the storage of information through a variety of memory and sensory systems (Cowan, 1988, 1999; Morey, 2018; Pasternak & Greenlee, 2005; Teng & Kravitz, 2019; Figure 4c). Much thought and experimental evidence has been allocated to adjudicating between these hypothetical architectures (Cowan, 1999, 2001; Lee et al., 2013; Logie et al., 2020; Morey, 2018). Our goal is not to refute competing accounts at this point, but rather to provide a possible functional implementation of how knowledge *could* be combined with WM. We are advancing one particular instantiation of such an account, and will discuss a comparison between approaches in the discussion. However, it should be emphasized that the mechanism of storing latent representations in a binding pool as described here has some generality and could be adapted to other architectures (i.e., it would be easy to use two binding pools, one for visual and one for non-

visual information). A primary goal of this paper is to clarify potential implementations to develop computational formalisms for comparison of different architectures.

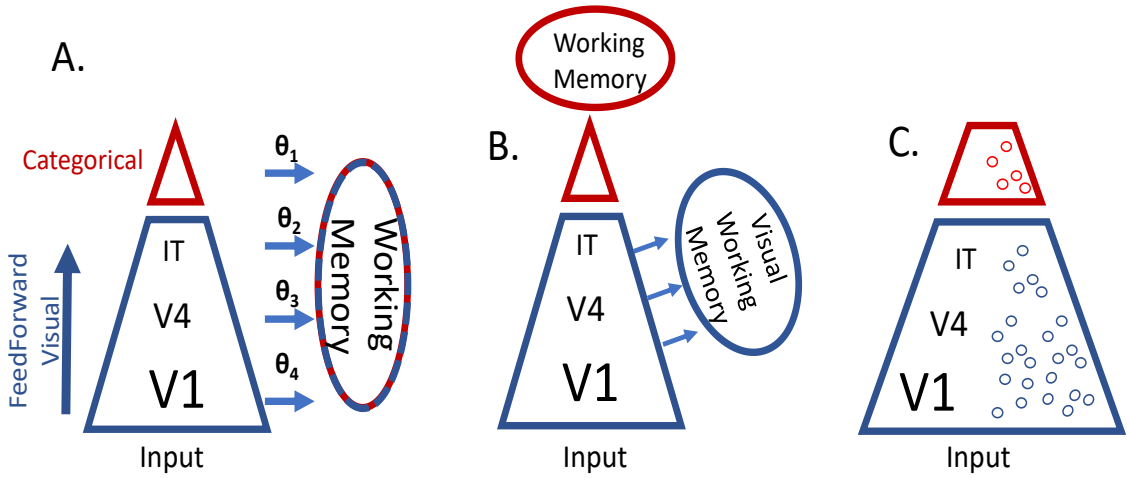

Figure 4. Three architectures for working memory as it relates to the visual system. ***A.*** *The MLR model as proposed here has a single memory representation that encompasses visual and categorical information in varying proportions according to task-dependent tunable parameters.* ***B.*** *A working memory model that suggests there are distinct systems for maintaining visual and non-visual forms of information* ***C.*** *A working memory model that has its representations embedded in the sensory system*

With this respect, the MLR model provides a functional intuition about how working memory builds representations of visual stimuli using the visual knowledge system. The behavioral constraints (i.e., functional requirements) as well as the neural constraints are listed here.

Functional Requirements

To motivate our account, we start with a list of empirical constraints that define the relevant functional capabilities of memory and several key neural plausibility constraints. MLR is constrained by empirical data obtained from human behavior, termed *functional constraints*. These constraints will be met in a qualitative fashion rather than by matching of specific empirical data points. Indeed, the requirements for a working memory model are more

comprehensive and it is very hard to imagine that a given model can meet all of them, however we selected a set of requirements that our proposed model could address in the next chapters.

Reconstructive

Although reconstruction of visual stimuli is not required in typical visual memory tasks, it is a form of retrieval, and people are able to draw or otherwise reconstruct the specific shape of remembered objects, particularly if subtle visual details need to be remembered (Bainbridge et al., 2019; Carmichael et al., 1932; Kosslyn, 1996). For instance, Carmichael et al., (1932) showed people sets of drawings and asked them later to draw what they could remember to assess how much detailed information was stored in their memory. The MLR accounts for this form of retrieval through pixel-wise reconstruction of images (described in Chapter 2).

Multiple codes

A memory of a familiar stimulus can be represented by a variety of different codes, from low level visual details to abstract categorical information (Potter, 2018; Potter et al., 1977; Potter & Faulconer, 1975). For instance, one can remember a specific visual detail of a car including its size, shape, color, etc. or a categorical information of the car in a more abstract sense. The MLR can store a combination of different kinds of information from a single stimulus, including its visual details or its categorical labels (described in Chapter 3)

Flexibility of feature representations

Within the visual memory of a single item, specific attributes (e.g., color, shape, etc.) are stored with varying degrees of precision. That is, depending on the task, some visual features are encoded more accurately than others. For example, (Swan et al., 2016) presented people with an oriented bar and asked them to report its color for several trails. In a surprise trail, subjects were asked to report the orientation of the bar on a continuous scale. The results indicated that memory for the color, which was task relevant, was more precise than the memory for the orientation, which was task irrelevant. Parameters in MLR control the ratio of distinct attributes of an object that are stored in memory. In the simulations here, color and shape are treated as distinct attributes but in a larger model, the set of tunable attributes could include any stimulus dimension for which there are distinct representations (described in Chapter 2).

Representing Novel stimuli

WM performance is more efficient for previously learned items (Yu et al., 1985; Zimmer & Fischer, 2020), however, humans can still encode novel configurations that they have not seen before. For example, Lake et al., (2011) presented subjects with unfamiliar hand-written characters such as Bengali, Sanskrit, etc. and asked them to draw the characters from memory. Subjects could successfully draw a given character by only one exposure. Similarly, MLR stores and retrieves novel shapes that it has not seen before such as Bengali characters, although those memory reconstructions are less precise compared to shape categories that the model was trained on such as hand-written MNIST digits (described in Chapter 3).

More Efficient representations of familiar items

Frequently-experienced objects drawn from long-term knowledge have compressed representations with high visual detail (Brady et al., 2008; Konkle et al., 2010). This allows more familiar objects to be stored in memory compared to novel stimuli (Hue & Erickson, 1988; Zimmer & Fischer, 2020). In the context of MLR, we define familiar as stimuli that the model has not been previously trained on, but are from the same distribution(s) that the model has been trained on. Novel stimuli are drawn from a distribution that is very different from the training distribution(s). MLR achieves more efficient representations of familiar items by encoding compressed representation of familiar items generated by a smaller number of neurons, whereas it resorts to encoding features represented in a larger number of neurons if the object is unfamiliar (described in Chapter 3).

Individuated Memory for Multiple items

Memory for one visual display or trial is able to store several different items and retrieve them individually using a variety of cues. For example, one could store a series of colored shapes and then retrieve a specific item based on one particular attribute, such as color, shape, location or serial order. In addition, people can store repetitions of the same item as well as the temporal order of different items (Bowman & Wyble, 2007; Kanwisher, 1991; Mozer, 1989; Swan & Wyble, 2014). MLR uses tokens as pointers to individuate different items, including repetitions (described in Chapter 2).

Storing multiple items and mutual interference

Storing multiple items and/or additional features of an item (i.e., shape, color, size, etc.) in WM causes interference that degrades the memory precision based on the number of items stored (Wilken & Ma, 2004) and the number of attributes within one stimulus(Swan et al., 2016) .The shared neural resources in MLR cause overlapping representations to interfere with one another, both for attributes within a stimulus and between stimuli (described in Chapter 2).

Content Addressability and Binding

Memory representations include a form of binding in which multiple attributes can be attached to one another, which is often thought of as object bindings when those attributes belong to distinct visual objects. Also, *content addressability* means that such bindings can be used to retrieve any attribute associated with the object from any other attribute. For example, a colored, oriented line can be accessed either through its color or orientation (Gorgoraptis et al., 2011). MLR can use any attribute or code associated with a token to retrieve other information from that token (described in Chapter 2).

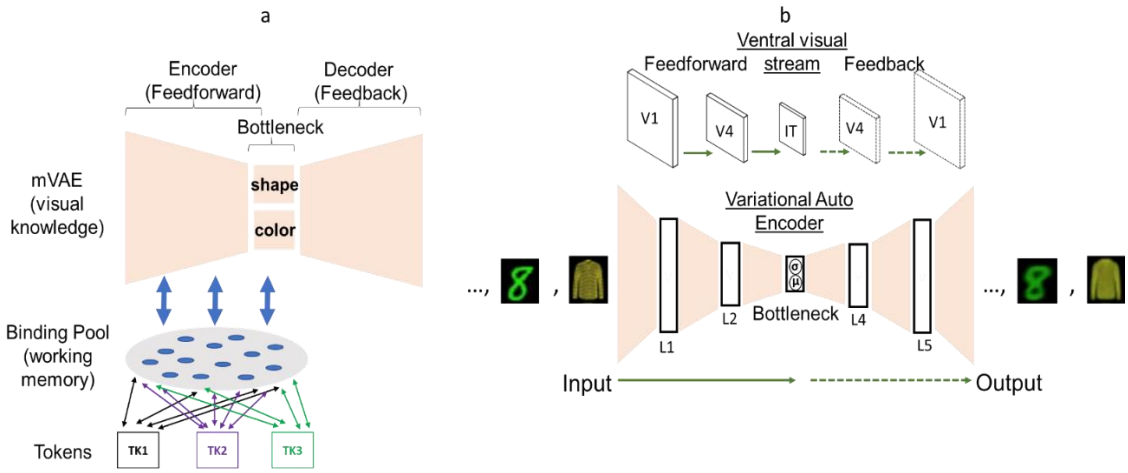

Figure 5. The simplified architecture of MLR with its two major elements: visual knowledge as represented by the modified VAE and the working memory as represented by the binding pool. We modified the bottleneck to represent shape and color in separate maps. b. Illustration of the architecture of a VAE (Kingma & Welling 2013) and its coarse neuroanatomical correspondence. In the neuroanatomical projection, solid arrows correspond to feedforward connections from V1 to IT cortex (or L1 to bottleneck in the VAE) and dashed arrows refer to feedback projections in the reverse direction (or from bottleneck to L5 in the VAE). The inputs were either colorized version of MNIST and f-MNIST. Note that one image at a time is fed into the VAE.

Neural Constraints

As mentioned earlier, a working memory model cannot acquire any architecture. Rather, the architecture should have some level of biological validity, because our goal is to eventually make inferences about how brain activations result in cognitive functions. Here, we describe a list of neural constraints that the MLR model is consistent with.

Rapid encoding and forgetting

Working memory representations are thought to be the result of persistent neural activity (Compte et al., 2000) or transient latent synaptic representations that can be rapidly changed (Rose et al., 2016; Szatmáry & Izhikevich, 2010). These representations allow for rapid encoding

and removal of visual information. The shared binding pool of MLR stores visual information by creating temporary activity states in a matrix that is intended as a generalization of either a population of self-sustaining neural attractors or other forms of rapidly modifiable representations (e.g., silent synapses; Rose et al. 2016). These activations are received via fixed randomly assigned weights from different layers of the visual knowledge hierarchy.

Hierarchical structure of ventral stream

The ventral visual pathway contains at least part of the visual knowledge that is gradually formed through extensive experience with the world. In primates, this pathway has cells that vary along a spectrum from receptive fields that are tuned to orientation and color in the earliest layers such as LGN and V1 cortex, up to cells that have large receptive fields and that are tuned to more complex shapes such as faces and complex configurations (Kanwisher et al., 1997). In MLR, visual knowledge is based on a VAE architecture as illustrated in Figure 5. The VAE resembles the hierarchical structure of the visual ventral stream with more generic representations at the early level and more compressed representations at higher levels (i.e., the bottleneck) with the number of neurons decreasing progressively. In a VAE, the layers from the bottleneck to the output translate between latent representations and a pixelwise representation. These are similar to the extensive feedback projections that extend backwards down the ventral stream from higher to lower order areas (Bullier, 2001; Lamme et al., 1998).

Training through synaptic weight adjustments

Our biological brain is an ever-changing system that alters its connectivity over time to represent the statistical regularities of the environment. Likewise, MLR learns statistical

regularities underlying visual categories through experiencing abundant exemplars that are used during the training phase. As a result of this training, the network's connectivity is tuned to better reconstruct the visual stimuli from compressed representations in the bottleneck. Moreover, training in the VAE occurs without explicit labels or supervision, akin to how a child can learn to see through exposure to patterned information.

Given the functional and neural constraints implemented in MLR, we conducted a series of simulations in the following chapters that enabled us to test the working memory mechanism against various tasks, and to learn about how a given behavior is generated.

Chapter 2

Building working memory representations from visual knowledge

This chapter is devoted to describing the architecture and mechanism of the MLR in more detail. Moreover, simulations of some of the functional requirements such as encoding flexibility, storing multiple items and content addressability are explained in this chapter.

MLR Architecture

The model is composed of two main components: a modified variational autoencoder (mVAE) operating as visual knowledge and a binding pool (BP), the memory storage that holds one or more objects (see Figure 6 for the detailed architecture).

mVAE

The VAE (Kingma & Welling, 2013) is an hourglass-shaped fully connected neural network consisting of three main elements: feedforward, bottleneck and feedback. Each layer of the mVAE consists of a number of neurons that all are connected to the neurons of the adjacent layers via initial random weights. These weights are adjusted throughout training to represent each stimulus.

The mVAE was trained by using a colorized variant of MNIST (LeCun, 1998) and fashion-MNIST (Xiao et al., 2017) stimulus sets prior to any memory storage simulations. The MNIST stimulus set consists of 70,000 images of 10 categories of digits (0-9) and fashion-MNIST set (i.e., f-MNIST) also has the same number of images as MNIST but for 10 categories

of clothing (T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneakers, Bag and Ankle boot). To add additional attributes to the dataset, we colorized all images using 10 distinct colors – red, blue, green, purple, yellow, cyan, orange, brown, pink, teal – with minor variations. Color values were [[0.9, 0.1, 0.1], [0.1, 0.9, 0.1], [0.2, 0.2, 0.9], [0.8, 0.2, 0.8], [0.9, 0.9, 0.2], [0.1, 0.9, 0.9], [0.9, 0.5, 0.2], [0.6, 0.4, 0.2], [0.9, 0.7, 0.7], [0.1, 0.5, 0.5]]. The color of each image was chosen by first selecting a prototype color and then adding random variation to each of the RGB channels from the range [-.1, .1]. While the mVAE major pathway was trained on the MNIST and f-MNIST, the skip connection was trained on the same images that were transformed by random rotations of +/- 90 degrees and random crop of size 28 with padding to be 8 (Figure 7).

We modified the VAE by dividing the bottleneck into two separate maps – a color map and a shape map – to represent each feature distinctively. This was important to demonstrate working memory flexibility by controlling which features (shape or color) are to be stored in working memory. The mVAE in MLR consists of 7 layers. Input layer ($L_{i;}$ dim= 28 x 28 x 3), Layer 1 (L_1; dim= 256), Layer 2 (L_2; dim= 128), bottleneck (color map, dim= 8; shape map, dim = 8), Layer 4 (L_4; dim= 256), Layer 5 (L_5; dim= 256) and the output layer (L_o; dim=28 x 28 x 3). A skip connection was added from L_1 to L_5. The BP layer is connected to the feedforward layers of mVAE bidirectionally.

Feedforward pathway

Translates information from a pixel representation into compressed latent spaces as a series of transitions through lower dimensional representations. This pathway receives the input images and constitutes input, L_1, L_2 and bottleneck. This is typically called the *encoder* in autoencoder models.

Shape and Color maps (bottleneck)

Typically, the bottleneck layer of a VAE that has the smallest number of neurons consists of one map. To generate distinct feature maps, we divided the bottleneck into two separate maps: one for representing shape and the other one for representing color. Each of the two maps is fully connected to the last layer of the feedforward pathway and the first layer of the feedback pathway.

Feedback pathway

Translates information from the compressed shape and color maps into pixel representations as a series of transitions through progressively higher dimensional representations. This pathway constitutes a bottleneck, L_4, L_5 and output. This is typically called the *decoder* in autoencoder models.

Skip Connection

Neural network representations are typically limited to their training sets, meaning that they can categorize or reconstruct objects that are drawn from the training set distribution. For example, a neural network that is trained to classify different types of vehicles, cannot perform well on images of houses. In the MLR model, we were able to reconstruct novel stimuli without involving the shape and color maps, by using a skip connection that linked the first layer to the last layer (i.e., L_1 to L_5) of mVAE. Anatomically, this would be the equivalent of a projection between layers with V1 cortex (Thomson, 2010).

Categorical labels

The mVAE is trained based on reconstruction and does not represent labels at any step. However, storing categorical labels is one of the ways an item can be encoded into working memory. In order to apply categorical labels to a given stimulus, we used a standard support vector machine classifier (SVM; Cortes & Vapnik, 1995). The SVM maps visual representations in the bottleneck onto discrete labels for different stimulus attributes such as shape or color. For instance, for a red 2, it generates a label "2" for its shape, and a label "red" for its color.

Binding Pool (BP)

The BP uses a modified formulation of the model described in Swan & Wyble (2014) and is similar to a Holographic Reduced Representation which has holistic , distributed representations across a fixed, large number of neurons (Cavanagh, 1973; Plate, 1995). Similar to classical ideas of working memory, the BP uses an activation-based storage mechanism by storing via sustained firing (Goldman-Rakic, 1995). Moreover, unlike an embedded working memory system (Cowan, 1999), the binding pool is assumed to be separate from the sensory system.

The BP is a one-dimensional matrix of neurons that is bidirectionally connected to each layer of the feedforward pathway (L_1, L_2, shape and color maps) as well as the outputs of the SVM classifiers which provide one-hot categorical labels of shape and color and allows for storing categorical labels. The BP stores a combined representation of the information from each of these sources for one or more stimuli in individuated representations indexed by tokens. The bidirectional connections allow information to be encoded into the BP, stored as a pattern of neural activity, and then projected back to the specific layers of the mVAE to produce selective

reconstruction of the encoded items. The connection between the BP and the latents is accomplished through normally generated, fixed weights. These are not trained through gradient descent but are assigned at the beginning of the simulation for a given model.

The binding pool is instantiated here in a mathematically idealized form of a single pool of neurons that is linked with an undifferentiated connection to a wide range of cortical areas. A key advantage of clustering neural activity associated with memory into a BP of general-purpose storage neurons is that higher-order processes have a straightforward path to control those representations, allowing them to be sustained, deleted, or instantiated into constituent cortical areas with a relatively small amount of circuitry. In the context of MLR, it is harder to imagine how a centralized executive control system could exert control over an embedded memory system, since those control circuits would have to synchronously infiltrate a large number of cortical areas in order to reactivate, manipulate or extinguish those memory representations. Binding information between different features within distinct objects is also simpler to implement in a binding pool architecture as demonstrated here because the information is physically clustered in a well-defined population of neurons. Thus, the arguments in favor of an account like MLR is that it places less demands on the interplay between functional networks.

Tokens

An additional supplementary element of the MLR is the tokens. Tokens are conceptually similar to the object files, which is a conjoined representation of features being presented closely in time or location (Kahneman et al., 1992; Marr, 1976). Tokens are instance specific, such that each item is associated with one token. In this sense, encoding is a process at which a given representation of an item in a latent space is linked to a token via the binding pool. Tokens are also akin to slots in working memory that each are allocated to an item to determine the capacity

limit of working memory in terms of the number of objects it can hold. However, unlike slots, tokens can interfere with each other via the shared resource of neurons within the BP.

Thus, tokens allow us to retrieve a specific stored item when needed. Without tokenizing and indexing each item, we cannot retrieve a specific object or attribute that is encoded into memory. In our simulations, each token is connected to a random subset of BP neurons comprising the 40% of total neurons.

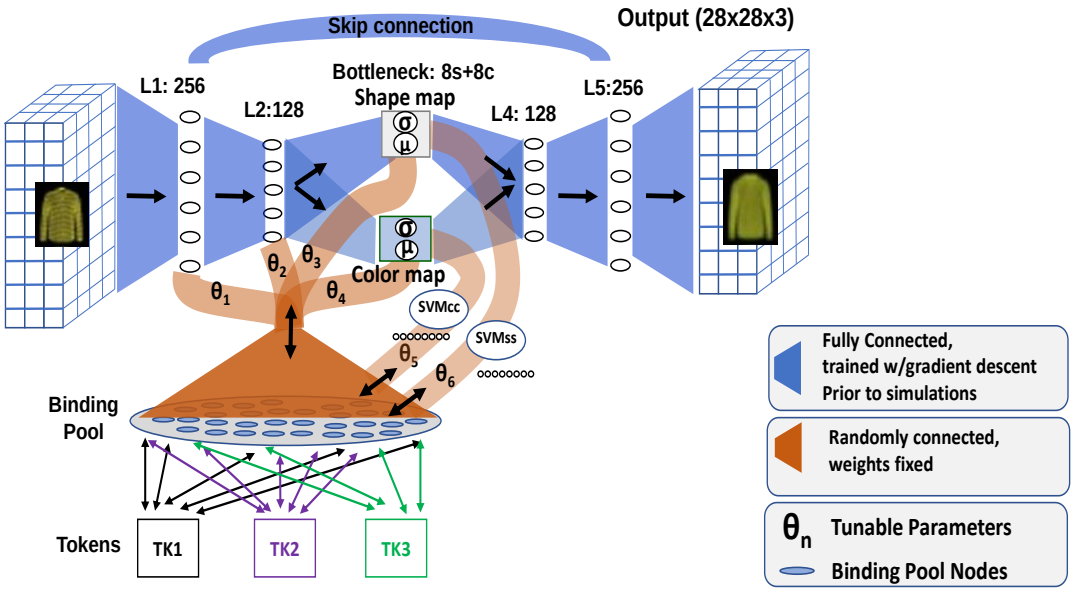

Figure 6. The complete MLR architecture that consists of the mVAE, binding pool, tokens, and classifiers for extracting labels (SVMss and SVMcc). Information flows in only one direction through the mVAE but can flow bidirectionally between the latent representations and the binding pool. Tokens are used to differentiate individual items. Note that three tokens are shown here but there is no limit to the number of tokens that can be allocated.

Figure 7. Colorization of MNIST and f-MNIST inputs using 10 prototypical colors with independent random variations on the RGB channels. Left: images used to train the mVAE. Right: Transformed images of the same dataset that were used to train the skip connection.

MLR Training

Typically, training a neural network constitutes finding the appropriate weight connections for predicting a set of test stimulus. VAEs are trained in an unsupervised manner, meaning that it does not use labels to map them to a given image for the purpose of training. This is akin to how a child can learn to see through exposure to patterned information.

In a VAE network, training is achieved by reconstructing the image. In other words, the model learns to encode and decode each image that is being trained on. The learning is obtained via optimizing the gradient descent by minimizing the reconstruction error at every epoch of training. The objective function proposed by (Kingma & Welling, 2013) that calculates the reconstruction error in a VAE is shown in equation 1. In this equation, $\emptyset$ and θ are the variational parameter and the generative parameter respectively. $q_{\emptyset}(z|x)$ represents the probabilistic *encoder* (posterior probability) by generating a distribution on the latent factor, z given the observed value of x. β is the regulation coefficient ($\beta = 1$ corresponds to the original VAE in Kingma & Welling (2013)). $P_{\theta}\ (x|z)$ represents the probabilistic *decoder* (likelihood probability) by estimating the distribution over x, given the latent factor, z. Finally, the first term ($E_{q_{\emptyset}(z|x)}[\log P_{\theta}\ (x|z)]$) is the reconstruction loss (i.e., expected log likelihood of the probability

distribution over the data points) and the second term ($D_{KL}(q_{\emptyset}(z|x)||P_{\theta}\ (z))$ is the Kullback-Leibler divergence between the encoder's distribution and the prior probability of $P(z)$ to measure how close these two distributions are.

$$L(\theta, \emptyset; x, z\ , \beta) = -E_{q_{\emptyset}(z|x)}[\log P_{\theta}\ (x|z)] + \ \beta * D_{KL}(q_{\emptyset}(z|x)||P(z)) \qquad [1]$$

Since we modified the VAE to represent shape and color distinctively, we used three objective functions to train the shape map, color map and the skip connection. As a result, the mVAE was trained on a total of 12,000 images with 200 epochs and a batch size of 100. Each batch was selected to train based on one of these three objective functions, and this was repeated for the entire training set for each epoch. All of the three objective functions were derived from equation 1.

Skip objective function: This function minimizes the reconstruction error for the input x represented by equation 2, where $l1$ is the activation of the first layer. This objective adjusted only the weights connecting the input to L_1, the skip connection to L_5 and connection from L_5 to the output.

$$L(\theta, \emptyset; x, l1) = -E_{q_{\emptyset}(l1|x)}[\log P_{\theta}\ (x|l1)] \qquad [2]$$

Shape objective function: This function converts the output images into grayscale images by averaging across the three RGB channels. Then the following objective was minimized with $\beta = 1$. This objective adjusted the weights connected to L_1, L_2, shape map, L_4 and L_5, while the color map and the skip connection were detached.

$$L(\theta, \emptyset; x, z_s, \beta) = -E_{q_{\emptyset}(z_s|x)}[\log P_{\theta}\ (x|z_s)] + \ \beta * \ D_{KL}(q_{\emptyset}(z_s|x)||P(z_s)) \qquad [3]$$

Color objective function: This function computes the maximum color value of RGB channels for each output image and converts the entire image to that color uniformly. That results in replacing each image with a uniform color patch containing no shape information. Then, it

minimized Equation 4 with $\beta = 1$. This objective adjusted the weights connected to L_1, L_2, color map, L_4 and L_5, while the shape map and the skip connection were detached.

$$L(\theta, \emptyset; x, z_c, \beta) = -E_{q_\emptyset(z_c|x)}[\log P_\theta\ (x|z_c)] + \ \beta * \ D_{KL}(q_\emptyset(z_c|x)||P(z_c)) \qquad [4]$$

The activation functions were ReLU (rectified linear unit) for the encoder and decoder, and sigmoid function for the last layer of the decoder.

Simulation Results

The mVAE disentanglement prior to memory encoding

Figure 8 shows reconstructions from shape, color and both maps respectively using SVMs. Note that when reconstructing only from the shape map, the color map activation was set to zero and vice versa.

SVMs were imported from the scikit-learn library as radial basis functions (kernel= 'rbf') with the decision function parameters to be C=10 and gamma='scale' respectively. Using classifiers, we were able to examine the amount of information within each map.

The results of classification accuracies in Table 1 show that color and shape representations were successfully disentangled in their corresponding maps. In other words, the shape map contained mostly shape information and the color map represented mostly color. This is a coarse approximation of the general finding that the ventral visual stream has specialization of cortical maps for different types of information (Cohen et al., 2014; Konkle & Caramazza, 2013). The benefit of such anatomical disentanglement in the context of a memory model like MLR is that it permits top-down modulation to select particular kinds of information for promotion to WM because the control signals only need to operate on the scale of selection regions of cortex, rather than individual neurons. That said, the *complete disentanglement* of

color and shape as we achieve here is not likely to be a real phenomenon but is very helpful for demonstrating the principles of encoding attributes selectively.

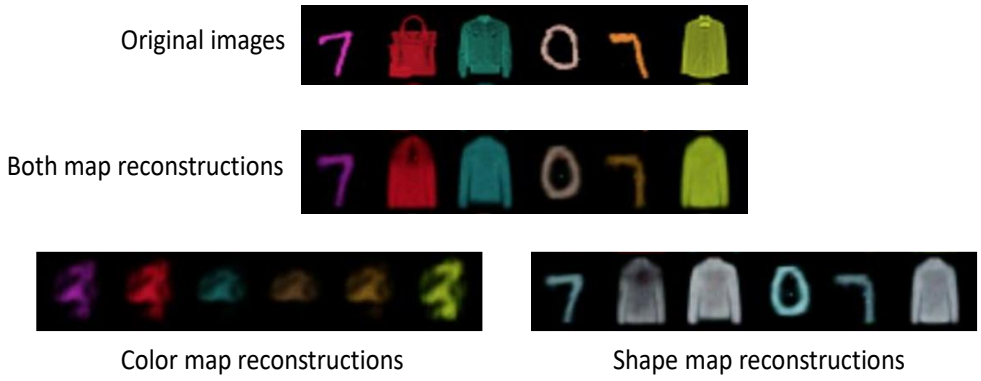

Figure 8. Reconstructions from the mVAE. Information from just one map is shown by setting the activations of the other map to 0. Both maps together produce a combined representation of shape and color, showing that the model is able to merge the two forms of information that are disentangled across the two maps. The model only processes one item at a time in these simulations, and these are combined into single figures for ease of visualization.

Table 1. Classifiers accuracies (%) of mVAE for information represented in shape and color maps

	Classifier type			
Shape/color maps	SVM_{SS}	SVM_{SC}	SVM_{CC}	SVM_{CS}
	84 (.20)	22 (.30)	87 (.40)	15 (.1)

Note. The table shows mean classification accuracies (%) for the mVAE for 10 trained models. The values in parentheses represent standard errors. SVM_{SS} represents an SVM trained on shape labels using data from the shape map. On the other hand, SVM_{SC} was trained to decode color labels from the shape map. SVM_{CC} Represents an SVM trained on color labels using data from the color map, whereas SVM_{CS} was trained to decode shape labels from the color map. Chance performance is 10% for classifiers trained on color labels (SVM_{CC} and SVM_{SC}) and 5% for classifiers trained on shape labels (SVM_{SS} and SVM_{CS}). As shown, SVM_{SC} and SVM_{CS} accuracies are above chance, but much smaller than SVM_{SS} and SVM_{CC} respectively.

BP encoding and retrieval of visual features

Once the mVAE was trained, memories could be constructed by shifting information from the latent spaces (L_1, L_2, or bottleneck) into the BP with 2500 neurons in total. The effective number of neurons representing each item was 1000 since 40% of the BP was allocated to each token. Such memories are constructed with a matrix multiplication of the activation values of a given latent space (i.e., L_1, L_2, shape and color map), by a randomly generated and fixed (i.e.,

untrained by gradient descent), normally distributed set of weights. This multiplication produces a vector of activation levels for each neuron in the BP. Multiple attributes can be combined into one representation in the BP by summing the activation values from multiple encoding features and then normalizing them. Equation 5 demonstrates the encoding of activations in the BP, where B_β represents each node in the BP, $N_{t,\beta}$ represents the connection matrix between the BP nodes and the token, X_f represents the activations in a given latent space, n is the number of neurons in the latent space that is being stored in the BP, and $L_{f,\beta}$ is the connection matrix between the latent space and the BP.

$$B_\beta = B_\beta + N_{t,\beta} \sum_{f=1}^{n} X_f L_{f,\beta} \quad [5]$$

Memory reconstructions to any given latent were accomplished by retrieving the associated token and multiplying the entire BP vector by the transpose of the same fixed weight matrix that was used during the encoding of that representation. As represented by Equation 6, the result is a noisy reconstruction of the original latent activity state, which can be processed by the rest of the mVAE in the same manner as visual inputs.

$$X_f = Z_t \sum_{\beta=1}^{n} B_\beta L_{f,\beta} N_{t,\beta} \quad [6]$$

Note that to determine which token was linked to a given visual form (e.g., a shape map representation), information can be passed from a given latent through the BP to determine which token has the strongest representation of that particular latent. Equation 7 illustrates the retrieval of a given token Z_t.Other parameters are similar to that of Equation 1.

$$Z_t = \sum_{\beta=1}^{n} B_\beta N_{t,\beta} \sum_{f=1}^{n} X_f L_{f,\beta} \quad [7]$$

Two methods were used to evaluate the quality of memory reconstructions of MLR. 1) Representations in the shape and color maps were classified by radial basis SVMs, which were trained to decode shape (one of 20) or color (one of 10) using the remaining 10,000 MNIST and 10,000 fashion MNIST as test set stimuli. The classification allowed us to assess the amount of

shape and color information in the shape and color maps before and after memory reconstruction. Note that we also used the same pre-trained classifiers to create the labels and to assess memory performance 2.) An alternative measure of the accuracy of reconstructing the original image was to correlate the reconstructed pixels with the original stimulus. We used this approach to quantify reconstructions of novel stimuli which have no pre-learned categories.

Projecting information from the latent representations into the BP and then back to the mVAE allows us to reconstruct the original activity pattern of that layer. Figure 9 illustrates examples of single items encoded individually and then reconstructed using the mVAE. Table 2 indicates the classifiers' accuracies averaged across 10 models for determining the shape and color of items according to which layer of the mVAE was encoded and then retrieved. According to the simulation results, it is evident that memory retrieval from shape and color maps is more precise than reconstructions from L_2 and L_1. Hence, compression resulted in more accurate memory retrieval, presumably due to the noisier reconstruction of the larger L_1 and L_2 latents. It is important to note that retrieval process of the familiar shapes requires images to pass through all the subsequent layers including the shape and color maps (e.g., L_2 representations are stored in the BP, then projected back to L_2 and then reconstructed by passing through the maps, L_4, L_5 and the output layer. Likewise, classifying the accuracy of the memory formed from the L_2 layer involves reconstructing the L_2 latent from the BP, then passing it forward to the shape and color maps and classifying those map activations with the SVMs).

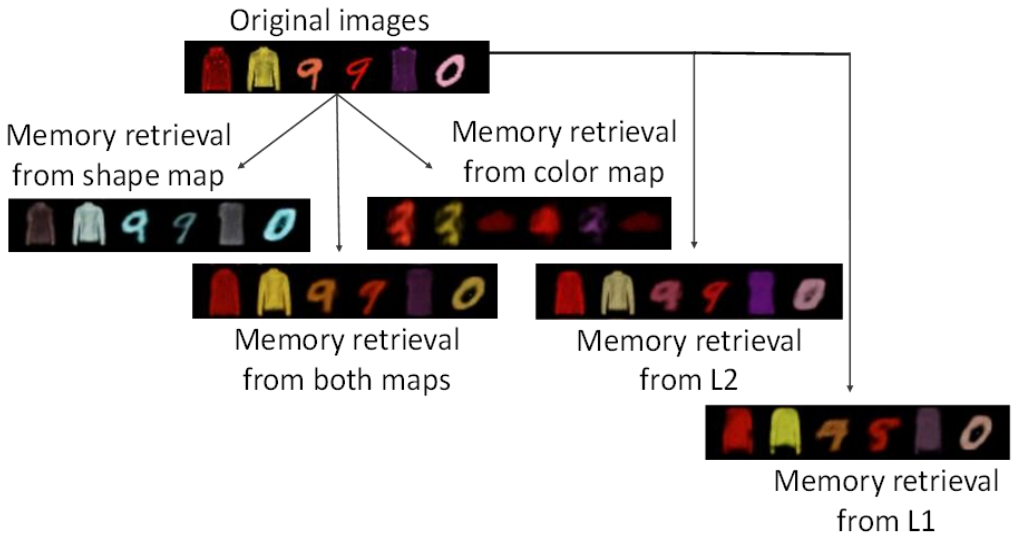

Figure 9. Demonstration of different latents that can be stored from one of the trained models (both shape and color were encoded in all conditions). Note that the reconstructions are visually less precise for memories formed from L_1 and L_2 latent spaces compared to the shape and color maps. Each item is stored individually in a separate BP, but the examples are combined into single images for ease of visualization.

Table 2. Mean classification accuracy (%) of shape and color information based on encoding conditions

	Classifier type			
Encoding conditions	SVM_{SS}	SVM_{SC}	SVM_{CC}	SVM_{CS}
Shape map and Color	83 (.43)	20 (.36)	80 (1.45)	15 (.29)
Shape map only	83 (.36)	20 (.45)	9 (.18)	4 (.22)
Color map Only	4 (.24)	8 (.69)	83 (.83)	15 (.33)
L2	75 (.55)	18 (.45)	69 (2.87)	14 (.34)
L1	44 (1.82)	13 (.85)	52 (2.96)	7 (.34)

Note. The table indicate means of classifier accuracies (%) after memory retrievals of a single stimulus from each layer for 10 independently trained models. The values in parentheses indicate standard errors. Rows correspond to different encoding conditions, showing which latent(s) were stored in the binding pool. Shape map only and color map only indicates that only shape or color of a stimulus was encoded and retrieved. L_1 and L_2 representations were passed forward to the shape and color maps after being stored in the BP to be classified.

Storing multiple items and mutual interference

As stated earlier, tokens allow individuation of different items in memory (Kanwisher, 1991; Mozer, 1989), which in this case occurs by allocating each item to a subset of the BP as introduced in earlier work (Bowman & Wyble, 2007).

Each token contacts a random, fixed proportion of the binding pool, effectively enabling those units for memory encoding while that token is active. Each token is connected to a random set of 40% (i.e., 1000) of total nodes (i.e., 2500). This means that when a given token is active, the subset of BP nodes it is connected to can be used to store and retrieve information, the remaining BP nodes will still hold their activation state, but can neither be encoded to, nor retrieved from. The subset of BP nodes associated with each token overlap with one another so that for any given token, 40% of its nodes overlap with any other token. As a result, with an increasing number of tokens stored in memory, the likelihood of interference between objects increases due to the overlap between token connectivity to the BP. There is no limit on the number of tokens, but the binding pool is assumed to be fixed in size. Given the fixed size of the binding pool, the interference between two items can be manipulated by increasing the subset of neurons allocated to each token. For instance, if we increased the token connectivity from 40% to 70%, the memory interference between two items would have been expected to increase accordingly. This mechanism enables multiple distinct sets of attributes to be stored in each token, effectively binding those attributes into one object. The tokens can be retrieved individually and in any order. Once stored in this way, a token can reactivate its portion of the BP to reconstruct the attributes associated with it.

Tokens thus provide object-based clustering of attributes in memory, and enable us to retrieve a specific item in memory in a case where multiple objects are encoded. Moreover, the shared resources in the BP that link tokens to the latent spaces causes interference between items,

as their representations partially overlap. Thus, representational quality degrades as additional items are added to memory in agreement with human behavior (Wilken & Ma, 2004).

Note that in this simulation we are reconstructing the actual shape and specific colors of the items, not just their categorical designations, and that explains the poor retrieval accuracy when four items are stored. As the number of items stored in memory gets larger, the quality of those representations visibly decreases as demonstrated in Figure 10. Note also that the color of the retrieved items becomes more similar as set size increases, reflecting the overlap in representation between the different items. This is emblematic of the interference observed in storing multiple visual stimuli in Huang & Sekuler (2010), such that the items representation shift towards one another, and is also consistent with previous studies that showed misbinding of colors between the stimuli as a form of interference with increased set size (Bays et al., 2009).

We also utilized classifiers to quantify the memory retrievals as a function of set size. Consistently, classifiers' accuracy showed poorer performance as the number of stored visual items increased (Table 3).

Figure 10. Illustration of the storage and retrieval of 1, 2, 3 and 4 items in memory. The interference increases as more items are stored in the BP. This results in inaccurate reconstructions of both shape and color, as well as the emergence of ensemble encoding.

Table 3. Accuracy of retrieving visual and categorical information (%) as a function of set size

conditions		
	shape	**color**
Set		
1	83 (.56)	81 (.59)

2	63 (.56)	56 (.79)
3	50 (.60)	45 (.55)
4	40 (.65)	38 (.56)

Note. Mean classifier accuracy (%) of retrieved items as a function of set size. The values in parentheses indicate standard errors computed over 10 independently trained models. The accuracy declines as more items are stored in memory consistent with human memory.

BP binding and content addressability

Human working memory is content addressable, meaning that it can retrieve an object using its particular feature (Gorgoraptis et al., 2011; Swan & Wyble, 2014). That is, when we memorize a red and a blue car, our memory is capable of retrieving either the red or the blue car given the model of the cars. In this regard, tokens enable content addressable recall in that a given attribute (e.g., the shape or color of a digit) can be used as a retrieval cue to determine which of several tokens was associated with that specific attribute. Then, that token can be activated to retrieve the other attributes associated with it. In other words, through token individuation, the BP is able to store the attributes of a given stimulus in a combined representation that allows it to link a particular shape to its color. In this sense, if two colored digits are stored in memory using the shape or color representations, memory can be probed by showing just the shape of one of the items and retrieving the token associated with that item. That token can then be used to retrieve the complete representation of the stimulus, including its color. Figure 11 illustrates an example of such binding and subsequent retrieval by a shape cue, and vice versa.

To test accuracy of binding retrieval, 500 digit-pairs were stored in the BP using the color and shape maps and two tokens. Afterwards, a grayscale MNIST was used as a retrieval cue to determine how often the model successfully retrieved the correct token based on this cue (Figure 13). When the two digits were from two different digit categories (e.g., a “2” and a “3”) the mean accuracy of retrieving the correct token across the 10 trained models was 88% (*SD*=1.73) against

a 50% chance. Tokens were used to retrieve the color map activation, which was then classified into a label, which resulted in an accuracy of 53% (SD = 2.1) with chance being 10% across correct and incorrect token retrievals. For the same MNIST digits (e.g., two 2's with a slightly different shape), the mean accuracy of retrieving the correct token across the 10 trained models is 73% (*SD*= 3.03), notably worse than when the digits were different but still far better than chance. The accuracy of retrieving the correct color from these tokens as estimated by the classifiers was 49% (SD = 2.42). This is a demonstration of retrieving a memory based on subtle variations in shape between categorically identical stimuli. This capacity is one of the predictions of the model, which is that human working memory is able to bind features to subtle variations in the shape of a highly familiar stimulus type for multiple stimuli.

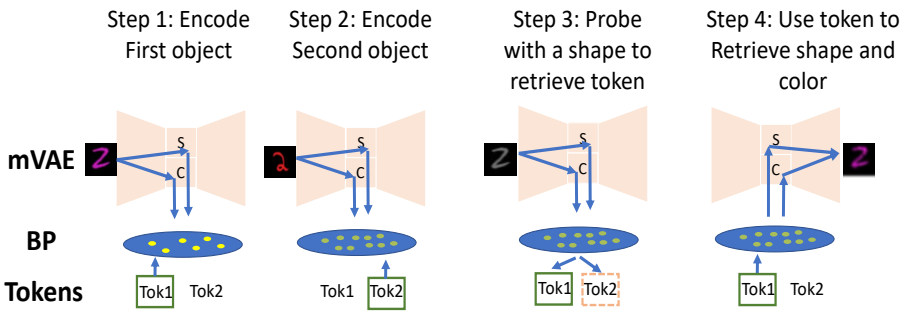

Figure 11. A diagram showing the flow of information during binding. The MLR stores two colored MNIST digits sequentially (step 1and step 2) in the BP. A grayscale shape cue is used to probe and retrieve the corresponding token (step 3). The resulting token is used to retrieve the shape and color of the cued input (step 4). The MNIST digits shown in this figure are not the result of direct simulation, but are just examples to show how binding process occurs.

Conclusions

In this Chapter we explained the mechanism underlying the MLR by describing its detailed architecture, the datasets it was trained on, and the training procedure.

We demonstrated simulations that were aligned with working memory functionalities. For instance, we indicated how working memory flexibility allows for storing and retrieving

specific attributes of items such as its shape or its color. We further showed that multiple items can be stored via tokens by indexing each individual item. Consistent with human behavior, we observed higher interference as more objects were encoded into working memory. Finally, we demonstrated the feature of content addressability of working memory. In that simulation, we indicated that using tokens, we could store two colored digits, and retrieve the digit by cuing it with its shape.

We further describe other functionalities of MLR in Chapter 3, such as how visual knowledge is used to store and retrieve novel information that is unfamiliar, and how familiarity resulting in a pre-defined categorical label can benefit the working memory encoding.

Chapter 3

Working memory based on familiarity/novelty

In Chapter 2, we described in detail how information in latents of the visual knowledge hierarchy is encoded into working memory. Those simulations were specifically for familiar items that had a compressed representation in the visual knowledge system throughout training and numerous exposures. However, human memory is capable of remembering a novel configuration as well. For example, Lake et al., (2011) presented participants with novel hand-written alphabets (e.g., Sanskrit, Bengali, etc.), and they were able to draw the alphabets from their memory. Thus, the question that arises here is how is visual knowledge used to form memories of novel information?

Furthermore, familiarity of items is also related to the code of memory representation. That is, a familiar item is followed by a categorical label. For instance, a car, in the context of our discussion, is considered to be a familiar object. Thus, an image of a car can have either a visual or a categorical representation. On the other hand, a configuration like "*X*\ "which is novel to many people does not belong to a specific category, and thus lack a categorical label. In this regard, this novel form can be remembered only via a visual code.

As a result, novel configurations lack two properties. One is lacking the categorical label, and the other is not having a compressed representation within the visual knowledge system. This would lead us to formalize familiarity in the context of compression and categorical labels.

As mentioned earlier, knowledge is the emergent feature of a trained sensory system. For instance, the connectivity of the visual system is adjusted via experiencing visual objects in a child, which results in a visual knowledge system. In this regard, MLR captures the two fundamental characteristics of visual knowledge (i.e., compression and categorical labels) that enables us to explain its interaction with working memory.

Compression: The amount of visual sensory input that we receive at every moment is enormous. Therefore, efficient data compression is essential given the limited-resources available to the perceptual system. It is likely that the visual ventral system represents familiar visual patterns with fewer neurons at successively later levels of the pathway (i.e., V1, V2, V4, IT) of the visual system (Bates & Jacobs, 2020). Hence, hierarchical visual knowledge can be formed from the compression of visual data (Ngiam et al., 2019; Norris & Kalm, 2021) by *learning*, via synaptic plasticity (Lamprecht & LeDoux, 2004), to encode and decode that data with high visual specificities. In this framework, later levels of the ventral stream (i.e., IT cortex) can represent specific shape patterns with minimal loss of visual details relative to earlier layers (i.e., V1) despite utilizing a smaller number of neurons to form the representation. This is due to connections between neurons encoding feature conjunctions in a hardwired fashion (VanRullen, 2009). Importantly, this compression is only effective for representations that are deeply familiar to the visual system (i.e., in which there have been thousands of exposures, sufficient to develop perceptual expertise, see (Pelli et al., 2006) and not for novel stimuli. Consistently, empirical data has shown long-term memory to have highly detailed representations of visual objects (Brady et al., 2008; Konkle et al., 2010).

Categorical representation: A familiar object also can have a conceptual/categorical representation. It has been demonstrated that whenever we perceive patterns that correspond to familiar concepts, that information is rapidly interpreted by the visual system to be translated into categorical codes that exist at an abstract level (L. Huang & Awh, 2018; Potter, 2018; Potter et al., 1977; Potter & Faulconer, 1975). For instance, for an experienced reader of the Roman alphabet, take note of the character: ‘A’. This character can be represented as a series of strokes with fine grained visual details that include the rightward slant, or the conceptual representation of A in a form of purely categorical information. The memory of seeing the familiar letter ‘A’ could be composed of one or both of these codes depending on task requirements.

The framework described here as visual knowledge, entailing the compression and categorical representation for visual data endows working memory with a doubly efficient representation of familiar objects. In other words, familiar objects benefit from compressed representations of visual information and also abstract categorical codes as they are encoded into WM, whereas novel objects lack such efficient representations. Figure 12 illustrates the diagram of hypothetical compressed and categorical representations of a handwritten digit '5' as it is being processed by the visual system. The key point here is that with increasing depth into the ventral stream the visual character is represented by progressively fewer neurons but the loss of detail is minimal as the stimulus category is familiar to the visual system. Moreover, this visual representation elicits a separate categorical representation which is even more compact than the visual representation, though it lacks the visual data.

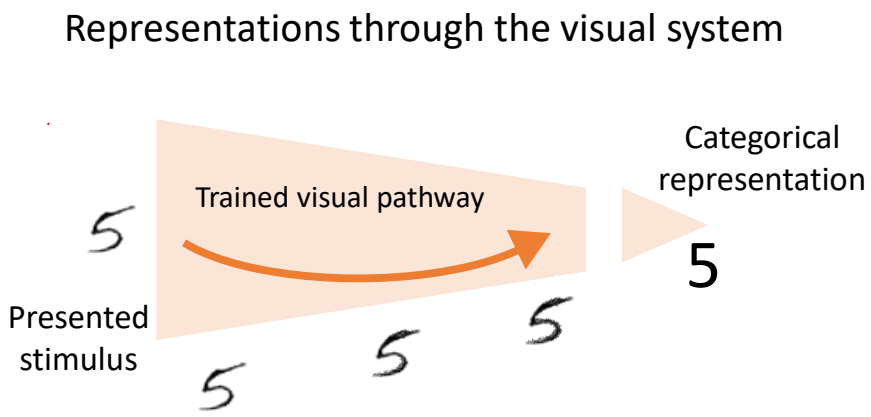

Figure 12. The compression and categorical representation of a single stimulus. The trained visual pathway represents the stimulus with specific visual details in all layers with little loss of visual specificities. The width of the cone reflects the number of neurons involved in the representation at different stages of processing. The final representation at the highest level would elicit a categorical representation that lacks the visual information.

Given the above framework of familiarity, using the MLR model, we will describe how novel items can be represented in working memory, and if an item is familiar, how we can encode it into working memory via its categorical label. We can also assess the memory capacity with

regard to an item's familiarity and its representational code (categorical vs. visual) in case it is familiar.

MLR Considerations to Represent Novelty

The detailed architecture of the MLR model and the description of each component was described in Chapter 2. In this section, we would like to remind the readers of what implementations were considered for representing novel information. An architectural consideration for this purpose was the skip connection. The skip connection linking the L_1 to L_5 was to circumvent the intermediate levels and prevent extensive compression. Biologically, this is similar to skip connections within the layers of V1 (Thomson, 2010). Another consideration was to train the skip connection on cropped and rotated images of MNIST and f-MNIST (random rotations of +/- 90 degrees and random crop of size 28 with padding to be 8, see Figure 7). As a result, the skip connection was trained to represent some patterns of lines and shapes, which were different from the actual dataset of MNIST and f-MNIST.

Storing Novel Stimuli

Novel shapes were 6 examples of colorized Bengali characters downloaded from www.omniglot.com. The colorization of Bengali characters was similar to that of MNIST and f-MNIST. The process of encoding novel shapes into BP is similar to familiar stimuli, except that the Bengali characters were stored from L_1 to the BP, then they are retrieved back to L_1, and were reconstructed via the skip connection. The skip connection is critical to reconstruct novel forms, since the nature of the compressed representations in the bottleneck force the reconstructions to

resemble familiar shapes. Figure 13 illustrates memory retrievals of the novel shapes using color/shape maps and L_1 representations.

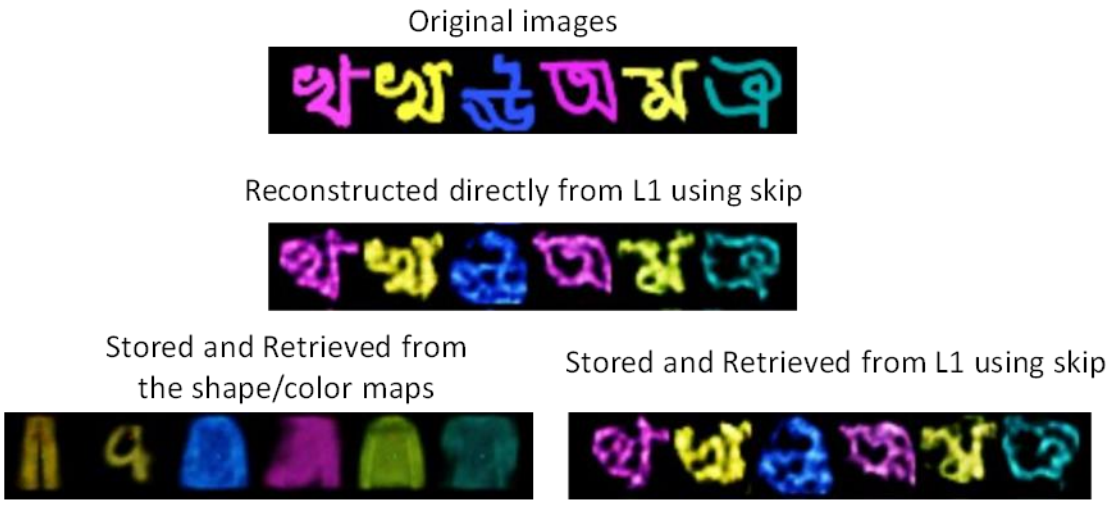

Figure 13. Illustration of storing a single novel Bengali character six times. The original images were reconstructed as familiar shapes when the BP stored the shape and color maps (bottom left). However, the successful reconstructions can be seen when the BP stores the L_1 activations and retrieve them via the skip connection. These representations are connected in one image for simplicity but each Bengali character was stored and retrieved from an empty memory store.

The results indicate that the novel shapes are better stored when they are encoded from L_1 and are reconstructed through the skip connection. On the other hand, reconstructions resulting from color/shape map encoding resemble the trained datasets, because the compressed representation in the bottleneck interprets the activations as what it has previously seen. These results suggest that unlike familiar items, novel configurations are better represented by the early levels of the visual system such as V1.

Different Codes of an Item

Shapes with known categories (i.e., familiar shapes) can take advantage of categorical memory. For instance, a given digit "*2*" can be remembered by its visual details, or with its categorical label. However, a Bengali character for someone who is not familiar with it does not have such categorical code. There are conflicting arguments about whether categorical and visual information are stored in a unitary memory system (Morrey, 2018), or they are encoded in separate storages (Baddeley & Hitch, 1974). The MLR model architecture can be adjusted based on any of these arrangements by having one BP or two separate BPs to store visual vs. nonvisual information. However, we examined the more parsimonious account of the unitary memory system (Morrey, 2018).

We used the trained classifiers on shape and color maps to estimate the categorical labels of each image from the shape/color maps. Then we converted the estimated categories into a localist or one-hot representation such that the estimated category is set to one and other possible categories are zero. Then, we stored the visual representation of shape and color maps along with the one-hot categorical labels. We examined the memory retrievals when shape and color and their corresponding labels were stored within one memory trace in the following conditions:

Condition 1 (encode visual, retrieve visual): shape and color map activations are stored together in the BP for each item; Then, either shape is retrieved (1s) or color is retrieved (1c). The retrieval accuracy was estimated by the same classifiers trained on the shape and color map representations (SVM_{SS} and SVM_{CC}).

Condition 2 (encode visual + categorical, retrieve visual): shape and color map activations are stored together in the BP along with shape and color labels for each item; either shape is retrieved (2s) or color is retrieved (2c). The retrieval accuracy was estimated as in condition 1.

Condition 3 (encode visual + categorical, retrieve categorical): shape and color map activations are stored in the BP alongside shape and color labels; either shape label is retrieved (3s) or color label is retrieved (3c). The retrieval accuracy of labels was computed by comparing the pre-encoding one-hot representations estimated by the classifiers for each item when it was first classified with the labels reconstructed from the BP.

condition 4 (encode 50% visual + categorical, retrieve categorical): This simulates prioritizing categorical information over visual details. This was similar to condition 3 except that the encoding parameters for the visual attributes was set at 0.5, meaning that activations of these maps were multiplied by .5 prior to encoding to prioritize categorical information over visual. When both shape and color maps are stored as visual information in the BP along with the one-hot coded labels, the visual information was not greatly perturbed (see condition 1 vs. 2 in Table 4 and Figure 14). This is because the one-hot labels are akin to a digital form of information that causes little interference with the stored visual details. It is also evident that retrieving labels results in higher accuracy compared to the visual information when all visual and categorical information are stored in one memory trace specifically for larger set sizes (condition 2 vs. condition 3 in Table 4 and Figure 14). By reducing the amount of visual information stored in memory down to 50%, the effect of set size on retrieving labels becomes smaller (condition 4 in Table 4, and Figure 14). On the other hand, accuracy of storing and retrieving only visual information is sensitive to the number of items (condition 1 in Table 4 and Figure 14). These simulation results match the common finding that people are able to remember several distinct familiar objects that have well-learned categorical labels (i.e., digits or familiar colors) with high accuracy up through approximately 3-5 items, while working memory for specific shape details is more limited (Alvarez & Cavanagh, 2004).

Condition 5 (encode categorical, retrieve categorical): This condition simulates a case in which no visual details are stored at all. This might not be a realistic condition, as it is hard to

imagine that there is absolutely no trace of visual information when people are shown a series of objects (i.e., this would preclude any memory of relative size, position, orientation, etc.). As shown in Figure 3, the capacity for encoding pure categorical information is high compared to the previous conditions when more items are stored. Note that while there is only a miniscule falloff in accuracy with set size here, interference does continue to increase beyond set size 4.

Table 4. Accuracy of retrieving visual and categorical information (%) as a function of set size

Set size	1s	1c	2s	2c	3s	3c	4s	4c	5s	5c
1	82.7 (.28)	79.9 (.65)	82.6 (.23)	79.9 (.65)	74.8 (.33)	81.4 (.34)	83 (.20)	86.6 (.34)	84.1 (.21)	87.2 (.36)
2	62.8 (.53)	56.3 (.58)	62.8 (.57)	56.1 (.55)	63.7 (.44)	72.3 (.42)	80.8 (.21)	84.7 (.39)	84.2 (.17)	87.1 (.42)
3	49.3 (.55)	44.4 (.52)	49.3 (.55)	44.5 (.51)	52.5 (.62)	61.8 (.44)	76.0 (.36)	80.0 (.40)	84.2 (.21)	86.9 (.38)
4	39.9 (.48)	37.5 (.54)	39.9 (.49)	37.5 (.56)	43.3 (.56)	53.2 (.51)	69.5 (.43)	73.7 (.37)	83.6 (.21)	85.8 (.37)

Note. Mean classifier accuracy (%) of retrieved items as a function of set size in conditions 1 and 2, and the mean accuracy of one-hot labels before and after storage in memory as a function of set size in conditions 3 and 4. The values in parentheses indicate standard errors computed over 10 independently trained models. In all cases the accuracy declines as more items are stored in memory, however, labels are more resistant to interference as shown in condition 4, especially when the amount of visual information stored in memory decreases.

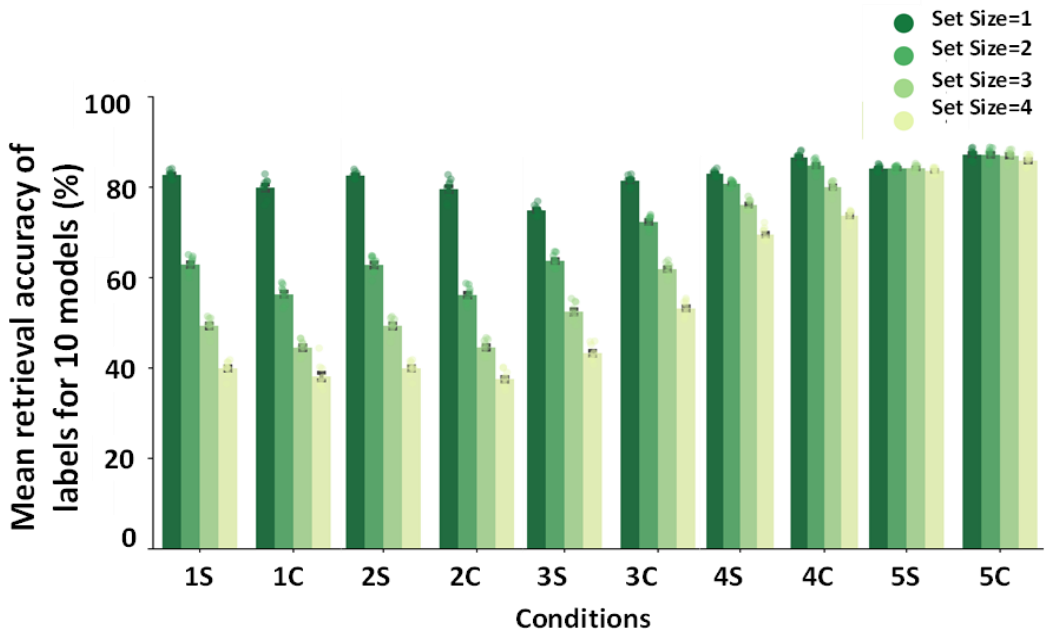

Figure 14. Visualization of Table 3 for mean accuracies of retrieved shapes and color maps (drawn from the classifiers) and labels. The results indicate that it is possible to store visual and categorical information in one memory trace in the BP. Moreover, categorical information did not interfere with visual detail information (compare condition 1 against condition 2). Furthermore, visual details were more susceptible to interference as the set size increased compared to categorical information.

More Efficient Storage of Familiar Items

It has been shown that people have higher memory capacity for familiar items drawn from long-term knowledge than novel stimuli (Chen & Cowan, 2005; Ngiam et al., 2019; Zimmer & Fischer, 2020). Based on studies on familiarity, we assume that natural images and their variations are familiar, because they can be mapped onto compact latent representations that are easier to remember. This means that a new picture of a familiar kind of object can be represented more efficiently than a new picture of an unfamiliar kind of object. Similarly, in the MLR model, familiar items for all simulations were drawn from the testing set of MNIST and f-MNIST images, such that the model was not trained on those specific images. Thus, those are new pictures of digits or fashion items but come from a familiar distribution

The MLR model can show how familiar items are stored more efficiently than unfamiliar ones, and therefore have less degradation of representations in working memory as the set size increases. As shown earlier, the BP better encodes the compressed shape and color representations for familiar items (see Figure 9), whereas novel shapes must be encoded from L_1 and then pass through the skip connection for precise retrieval (see Figure 13). To quantify the memory performance, instead of classifiers we compared the pixelwise cross-correlation of input and retrieved images as the function of set size for familiar and novel[1] stimuli, such that familiar shapes are encoded from the shape/color maps and novel shapes are encoded from L_1 and retrieved from the skip connection. The result of the cross-correlations for 500 repetitions are illustrated in Figure 15.

As it is shown, the correlation value declines as the set size increases, but more steeply for novel than familiar stimuli. Using cross-correlation, we also measured the memory performance for when familiar items are encoded from L_1 and retrieved via the skip connection, versus when novel items are encoded from the shape/color maps. The values have been summarized in Table 5. The shape/color map memory retrievals of the novel shapes are ~15% for all the set sizes, indicating that novel configurations cannot be represented by the highly compressed maps at the center of the mVAE. The results also revealed that the L_1 encoding of familiar shapes and retrieving it via the skip connection yielded a lower performance across all the set sizes compared to encoding of shape and color map representations. Hence, the compressed shape and color representations achieved by training allows for more precise memory representation for familiar shapes, whereas this efficient representation does not exist for novel

[1] Due to the limited number of images for Bengali characters as novel shapes, we augmented the data by doing slight rotation (10° rotation) and random crop with padding =8 on the 6 characters. This enabled us to do the permutations test for measuring cross-correlation.

configurations. Therefore, the model relies on the early-level representations of L_1 to store novel shapes, which in turn comes at the cost of less precise memory retrievals.

It should be noted that the efficient representation of the shape and color maps is limited to when the model *encodes* the information into memory. In other words, if the images were to be reconstructed from the mVAE without being stored in memory, the L_1 representations contain slightly more visual detail than the shape and color maps. This is illustrated in Figure 16 for familiar items, in which we computed the correlation values of a single input and its reconstruction when images were reconstructed from L_1 via the skip connection versus when they were reconstructed from the shape and color maps in the no memory storage condition (left panel). This was compared to memory retrievals from L_1 and shape/color maps, for which the retrievals from the latter were shown to be more precise (right panel).

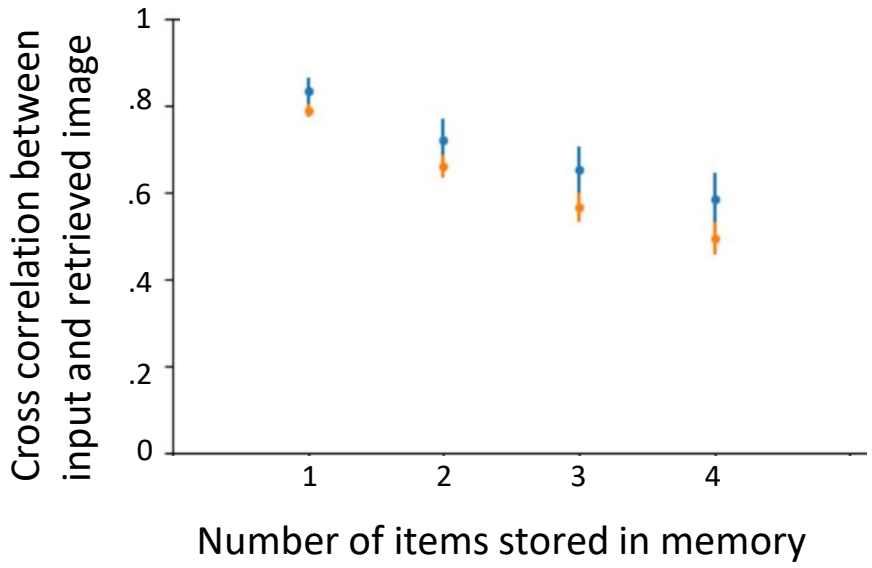

Figure 15. Mean cross-correlation of pixel values for 500 repetitions between input and retrieved images of 10 trained models for familiar (blue) and novel (Bengali, orange dots) shapes across different set sizes. Blue dots indicate the correlation for familiar stimuli when the shape/color maps were stored in the binding pool and retrieved via the mVAE feedback pathway. Orange dots indicate the correlation of a novel stimulus when the L_1 latent was stored and then reconstructed with the skip connection. Note that the reconstruction quality is lower for novel shapes and also those novel reconstructions deteriorate more rapidly as set size increases. The bars stand for standard errors.

Table 5. The correlation values between input and retrieval stimuli as a function of set size

	Stimuli type			
	Familiar		Novel	
Retrieval				
	S/C maps	L1-skip	S/C maps	L1-skip
Set size				
1	.84 (.03)	.79 (.03)	.15 (.06)	.79 (.01)
2	.72 (.05)	.69 (.04)	.15 (.05)	.66 (.03)
3	.65 (.05)	.61 (.04)	.14 (.06)	.56 (.03)

4	.59 (.06)	.51 (.05)	.14 (.06)	.50 (.04)

Note. The mean cross-correlation between stimuli and their retrievals for different set sizes across 10 trained models. S/C maps stands for shape and color maps. The values in parentheses are standard errors. The correlation values were measured in cases where the BP encoded the shape and color activations of the novel/familiar stimuli and then the stimuli were retrieved via the feedback pathway (retrieval condition: S/C maps). The correlation values were also measured in cases where the BP encoded the L_1 activations of the novel/familiar stimuli, and then the stimuli were retrieved via the skip connection (retrieval condition: L_1-skip).

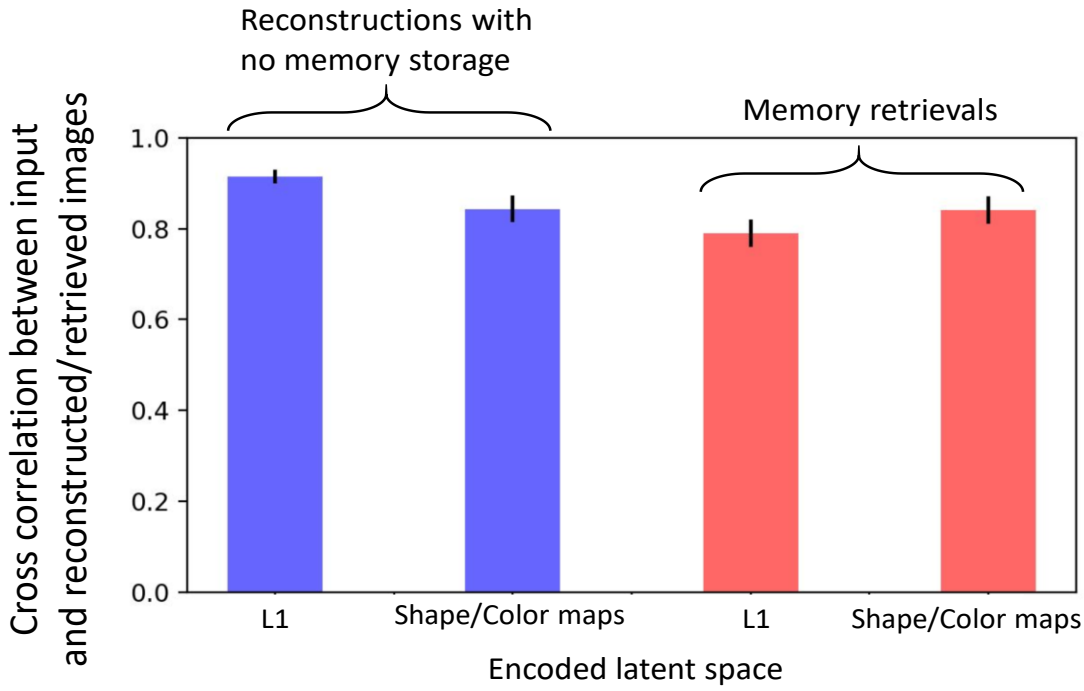

Figure 16 . The mean cross correlation values between a familiar input and a reconstructed image for 10 trained model. The blue bars are the mVAE reconstructions of L_1 via skip connection (left-blue) and the shape/color maps without being stored in memory (right-blue). The red bars are memory retrieval of one item when the L1 representation is stored in the BP and retrieved via the skip connection(left-red) vs. when the shape/color maps are stored in the BP and then retrieved via the decoder (right-red). The error bars represent standard error.

Empirical Validation of the MLR Model

Once a cognitive model is created, it needs to be validated by empirical data that were stemmed from the model's predictions. In partial validation of the model, we provide a series of predictions with empirical tests about the capabilities of working memory in storing visual

information. These capabilities were derived from the general properties of the MLR model and key assumptions that we have made in its construction

Predictions and Experiments

Prediction 1

Working memory experiments are typically performed with repetitive experience using the same stimuli that enable participants to develop fine-tuned expectations about the task demand. We posit, however, that working memory is typically used without such expectations in daily life and can store mental representations that are useful despite having no expectation of the memoranda or response. The specific prediction is that people can remember the fine-grained shape details of stimuli that they are not very familiar with even in the absence of specific expectations or experience in the task. MLR achieves this result by encoding the L_1 representations of the Bengali characters, and retrieving them via the skip connection. Cross-correlation of pixel values between the input and the retrieved image was .79 (*SE* = .01) for one item across 10 independently trained models.

Experiment 1

20 Penn State University undergraduates (Mean age = 19.55, 90% female, 20% left-handed) participated for course credit and signed an informed consent form prior to participation. All the experimental designs were approved by IRB at the Pennsylvania State University, and the scripts are available at this link (https://osf.io/tpzqk/). Participants were shown one Bengali character and were then asked to click on the exact character they remembered seeing from a

search array of four Bengali characters (Figure 17). Critically participants were only instructed to pay attention, and were uninformed as to the nature of the ensuing memory question until after viewing the image[2]. Five Bengali character categories were taken from the stimulus set downloaded from www.omniglot.com, which includes multiple different exemplar drawings of a Bengali character in grayscale. The experiment was developed in Psychopy (v2020.2.2, Peirce et al., 2019) before being translated to JavaScript using the PsychoJS package (v 2020.2) and run online via Pavlovia (Peirce et al., 2019). Each character was presented in the center of a grey screen at size (0.15x0.15 Psychopy height units, a normalized unit designed to fill a certain portion of the screen based on a predefined window size) for 1000ms, followed by a 1500ms delay. After viewing the image, participants were then instructed to click on the image they just saw. Four images including 3 distractors and the target were presented to the participants. On the first trial, non-target answer options were selected from different Bengali characters, and on trial 2 non-target answer options were different exemplars of the same character. Participants were not aware a 2nd trial would occur until after they completed the first. Accuracy scores were considered significantly above chance if a 95% bootstrapped CI (95% bCI) did not include the chance baseline (25%).

[2] The exact instructions for the experiment were provided separately on each trial, to minimize the potential of participant's predicting what they may need to remember during this experiment. Thus, the instructions occurring before trial 1 were as follows: "Thank you for participating in this experiment. You will be completing two separate experiments! This 1st experiment will be a very short, ONE TRIAL experiment where we show you some visual information. Because there is only one trial we need your full attention, as you only get ONE SHOT. So, keep your eyes on the fixation cross before the stimulus appears. Press the SPACEBAR when ready to begin." The instructions following trial 2 were as follows: ""That concludes our first experiment! We will now begin the 2nd, equally fast ONE TRIAL experiment. We will show you some new visual information. Again, we need your full attention, as you only get one trial. Press the SPACEBAR when ready to begin."

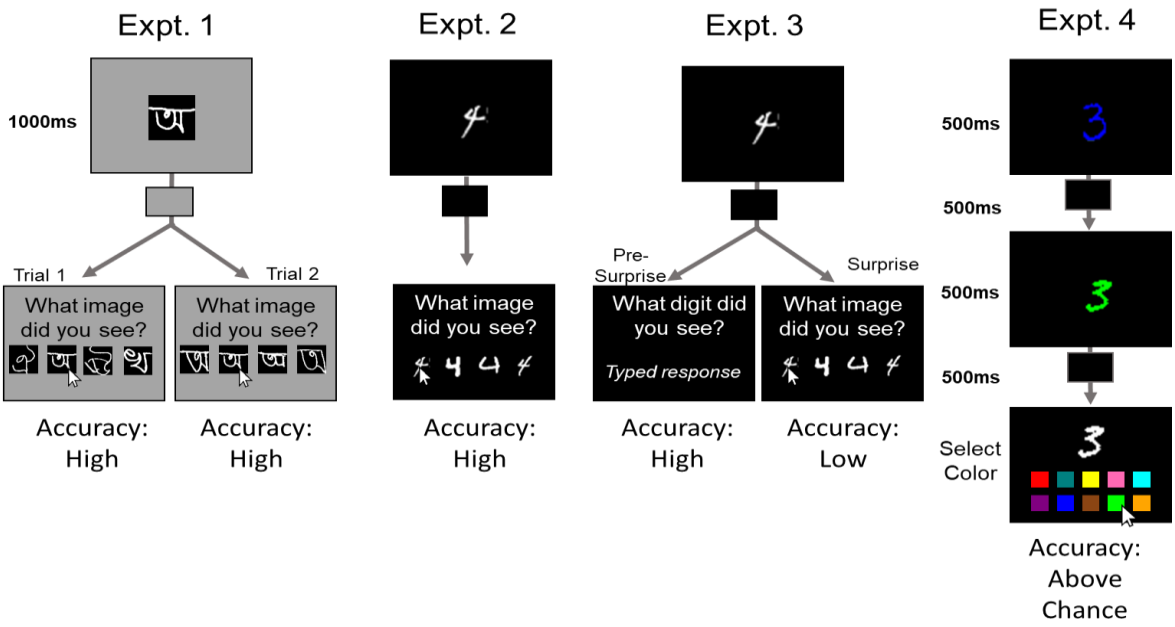

Figure 17. Trial layout for all experiments conducted on human participants. In Experiment 1, participants saw a grayscale Bengali stimulus before being asked to click which image they remembered seeing. The foils presented in the 4-afc varied between trial 1 and trial 2. They were not informed ahead of time that there would be a memory task. Experiment 2 was identical to Experiment 1, except the stimuli used were MNIST digits. In Experiment 3, participants viewed grayscale MNIST and were instructed to type in the category of the image (e.g., type '4' in displayed trial) for 31 consecutive trials before being surprised with a question asking them to click on the exact MNIST exemplar they remembered seeing. In Experiment 4, participants were instructed to remember the color-exemplar pairing of MNIST digits, before being cued with the specific exemplar and asked to click on the color that exemplar was. The key behavioral result is summarized below each condition, see text for details.

Experiment 1 results

Though participants were shown novel targets and given no instruction to remember the Bengali character presented to them, the mean accuracy on the very first trial was 95%, 95% bCI [85%,100%]. Participants were also highly accurate on the second trial which asked a more difficult question by requiring them to find the target image from the same category, M = 90%, 95% bCI [75%,100%].

Prediction 2

When people create a memory for a familiar stimulus, they can also have a memory for the specific shape of that stimulus, as well as being able to categorize it. This is particularly true when there is no specific expectation as to what response will be necessary. MLR predicts this by storing multiple codes of a given visual stimulus in one memory trace.

Experiment 2

A new sample of 20 Pennsylvania State University undergraduates (Mean Age = 18.6, 90% female, 5% left-handed) participated in this online experiment for course credit. Participants viewed one grayscale MNIST digit image (3, 4, 6, 7, and 9) on a black background before being asked to click on the exact image they remembered seeing (Figure 17). Again, participants were not informed there would be a memory task[3]. Thus, the first trial served as an unexpected memory test. Non-target options were exemplars from the same digit category (e.g., they saw four different instances of the digit 3, one of which was an exact match to what they had just seen; see Figure 17). Participants completed 5 trials in total, with a new digit category shown on each trial (i.e., digit categories were never repeated within an individual). All other components of Experiment 2 were identical to Experiment 1.

MLR simulates the lack of expectation via setting the encoding parameters of the shape and color maps to 1.0. and encoding both features as well as the categorical labels into memory.

[3] The specific instruction was: “This experiment will be a very short experiment where we show you some visual information. Because it is short and each of the 5 trials are unique, we need your full attention right from the start. Keep your eyes on the fixation cross before the stimulus appears. Press the SPACEBAR when ready to begin."

We replicated simulation 6 (Table 4, condition 2s) except with grayscale stimuli. The mean decoding accuracy of the retrieved shape across 10 trained models was 83% (*SE* = .43).

Experiment 2 results

Overall, the mean accuracy of reporting the MNIST exemplar without being specifically instructed was 85% (95% bCI [60%, 95%]) on the very first trial. Accuracy on subsequent trials was 85% (95% bCI [70%,100%]), 90% (95% bCI [75%, 100%]), 100%, and 100%. Thus, people could remember the shape of a familiar stimulus even when there was no expectation to report on the its shape and improved to perfect accuracy with a small amount of experience. This supports our assumption that even highly familiar stimuli are encoded at the specific shape level in the absence of expectation of what specific question will be asked.

Prediction 3

Our next prediction is that expectation can tune representations for highly familiar objects to represent categorical information at the expense of visual information. This will be tested by asking subjects to repeatedly report the identity of a digit, ignoring its specific shape, and then giving them an unexpected question about its shape after 30 trials. Thus, the same question about visual shape that could be easily answered in Experiment 2 should be hard to answer after memory encoding settings have been tuned to exclude visual details.

This prediction stems from the fact that MLR has modifiable parameters controlling the relative contribution of shape vs one-hot categorical representations as memories are constructed. We modified the parameters similar to simulation 6 (Table 4, condition 4) except that only 20% of shape information is stored in memory alongside the categorical label. The classification

accuracy of the decoded shape information demonstrated 24% (SE=1.6) shape accuracy, whereas the accuracy of retrieving the label was 84% (SE =.6). It should be noted that the ceiling accuracy of labels is constrained by the classifier accuracy, which is about 85% and thus lower than human performance in identifying an MNIST digit, which is near 100%.

Experiment 3

A new sample of 20 Pennsylvania State University undergraduates (Mean Age 18.8, 95% female, 5% left-handed) participated in this online experiment for course credit. The paradigm (Figure 17) resembles that used in standard Attribute Amnesia studies (Chen & Wyble, 2015). Participants viewed a grayscale MNIST digit (from any digit category 0 through 9), and were instructed to report the category of the image by typing the respective digit on the keyboard[4]. This task was repeated for 50 trials before participants were asked a surprise question on Trial 51: instead of identifying the image category, they had to select the specific category exemplar they remembered seeing (i.e., which specific "2" among a 4-AFC array of MNIST "2s"). On the surprise trial, participants reported the specific shape of the digit they just saw by clicking on the image that matches the target[5]. The display response consisted of the target and 3 distractors from the target category but with different shapes (see Figure 17).

Participants then completed 9 more exemplar identification trials (termed *control* trials). Significance for accuracy changes on the surprise trial was assessed by comparing surprise trial

[4] The exact instructions were as follows: "In this task, you will be presented with an image of a digit. Your task is to identify which digit it is. When asked to do so, using the number keys on the top of your keyboard (NOT the number pad), press the key of the number you remember seeing."

[5] The exact instructions presented on the surprise trial were as follows: "This is a surprise memory test! Try to remember the digit you last saw? What was its specific shape? Click the image that matches the image you just saw"

accuracy to accuracy on the 1st control trial via a permutation test (10,000 iterations). All other parameters of this study were identical to Experiment 2.

Experiment 3 results

The mean accuracy of identifying the target was 97% during the 50 pre-surprise trials. However, on the surprise trial, the accuracy of identifying the exact shape of the presented stimulus was 15%, 95%bCI [0,30%]. On the very next trial, when participants had an expectation to report such information, the accuracy of reporting the shape of the digit elevated to 100% on the very next trial. The difference between performance on the surprise and first control trials was significant as determined via a permutation test, $p < .0001$. This demonstrates that memory representations can be tuned to represent largely categorical information, with minimal specific shape information.

Prediction 4

Perhaps the most striking capability of WM that is unique to MLR among competing models is the ability to bind two specific colors to two different shapes, even when the two shapes are from the same category (binding accuracy: M=69.92%, SE= .50). This follows from the fact that tokens link shape map information to color map information. Therefore, when shown two different instances of an MNIST digit of a given category with different colors, participants should be able to report which color was bound to which specific shape (See Figure 13)

Experiment 4

A sample of 20 participants (Mean Age 21.9, 45% female, 15% left-handed) were recruited from the online website Prolific. On each trial (Figure 17), 2 MNIST exemplars from the same digit category were presented sequentially to the participants[6]. Each exemplar was randomly colored from a list of 10 options (Red, Green, Blue, Pink, Yellow, Orange, Purple, Teal, Cyan, and Brown), and colors did not repeat within a trial. Each digit was visible on screen for 500 ms, with a 500 ms ISI between exemplars and a 500 ms delay between the second exemplar and the response screen. One of the exemplars (counterbalanced across trials) was then presented to the participant in grayscale, and participants were instructed to click on the color that was paired with this exemplar (10-afc; chance = 10%). Unlike in previous experiments where no instruction was given, participants *were* explicitly instructed to remember the color-shape pairing.

Experiment 4 results

Participants were able to remember which color was linked to which specific MNIST exemplar. Overall, Participants correctly reported the target color 81.5% of the time[7], 95% bCI [75.5%, 87.25%], with swap errors (reporting the color of the other MNIST digit) occurring on average 9% of the time, 95% bCI [4.75%, 14%]. Importantly, participants were capable of completing this task on the first trial, as 17 of 20 participants (85%) reported the correct color on trial 1, 95% bCI [70%, 100%]. This shows that the ability to utilize one attribute of a bound

[6] The exact instructions presented on the surprise trial were as follows: "This is a surprise memory test! Try to remember the digit you last saw? What was its specific shape? Click the image that matches the image you just saw"

[7] Target presentation order had no influence on accuracy. Participants were as accurate at reporting the target's color when the first exemplar served as the probe (*M* = 84%) compared to the second (*M* = 79%), permutation p-value = 0.49.

representation to retrieve another attribute is a general capability of working memory, rather than a specific capacity that emerges through training.

Conclusions

In this Chapter we showed how familiar items are stored differently than novel information. We also demonstrated the memory capacity for categorical vs. visual information. We learned that based on the MLR model, memories of novel configurations can be formed from early layers of the visual system, where information has not been extensively compressed. On the other hand, familiar items that people have had several exposures to were better encoded from the higher levels of the visual system such as IT cortex which is equivalent to compressed representations of the MLR bottleneck. We also found that familiar objects are stored more efficiently if the set size increases. As a matter of fact, as set size increases, the accuracy of retrieving visual detail's information measured by cross correlation was higher for familiar vs. novel objects.

Finally, to partially validate our model with empirical data, we ran four experiments based on the model's predictions. In Experiments 1 and 2 we provided empirical evidence of the incredible flexibility of building memory representations from appropriate latent representations by showing that naïve subjects can retrieve the specific shape of both novel and familiar stimuli at the very first trial without being aware of the nature of the task, as no specific instructions or examples were provided prior to the brief exposure. This is important in validating the MLR model, as it demonstrates that the existing pathways for building memories of novel or highly familiar shapes, do not need to be recruited over multiple experiences or with forewarning. On

the other hand, Experiment 3 results showed that building expectations of remembering categorical information can diminish the memory of visual details and can then be rapidly readjusted to store the visual details on the trial immediately after the surprise test. In the model, this is achieved by tuning the model's weights for visual and categorical pathways. Finally, in Experiment 4 we showed that working memory stores shape-color bindings, allowing subtle shape differences to be used as a cue for retrieving a specific color, even for members of the same category of highly overtrained stimulus types like digits.

Chapter 4

Dual Coding and Its Implications for Memory

The process of designing the MLR model based on functional and neural constraints was described in previous chapters. Nonetheless, after the process of creating and validating the model is complete, it can be used for further predictions, as it facilitates conceptualizing new experiments. In the case of the MLR, we proposed several other working memory-related experiments in the following chapters that examine theories in the domain of dual coding, visual search and imagery. Not only does this help to integrate theories of related matter, but also it sheds light on limitations of the MLR model. Therefore, the MLR can continue to be modified and improved.

In this chapter we explore the formats of mental representations within working memory in various conditions, as this has been one of the important questions at the center of cognitive psychology since its early development (Paivio, 1978). In previous chapters, we described that people could store visual aspects of an object such as specific shape, color and size of an object, or the category of the object as a one-hot code representation. However, we did not dive into how verbal codes could be generated and stored into memory. For instance, how would memory differ if people see a picture of a car, versus if they receive the word “car” as an auditory input or as a word next to the picture of a car. The former represents the visual code of the car, whereas the latter would be considered as the verbal code input.

Reviewing the literature on representational formats reveals that researchers have different perspectives on this matter, and each perspective is supported by behavioral and physiological evidence. Propositional theorists contend that information is represented in a conceptual propositional format (Anderson & Bower, 1974; H. H. Clark & Chase, 1972; Reed,

1974). This format is in an abstract form and is similar to the way verbal codes are encoded (Anderson, 1978). As an example of what constitutes the proposition, Reed (1974) conducted a series of experiments in which subjects were presented with two sequential patterns and they were asked to respond whether the second pattern is part of (or similar to) the first pattern. They observed that some parts were easier for subjects to recognize, whereas some other parts were more difficult (i.e., subjects showed lower accuracy). They argued that participants encoded some structural proposition (e.g., two overlapping triangles) and used that for their recognition judgment. If they stored an image-like representation of the first item, there should not have been a difference in recognizing parts of the picture. In this regard, when people see a picture of a dog in a park, they may store it as a proposition: "There was a dog in a park" instead of a purely perceptual, image-like representation that contains most of the visual aspects of the scene.

On the other hand, proponents of mental imagery argued that people encode parts of images in a pictorial format (Kosslyn et al., 1978; Paivio, 1969). As an example, Carmichael et al., (1932) showed people a list of drawings with two kinds of verbal names that appropriately matched the drawing (e.g., a crescent moon had a label of moon and letter 'C'). Two groups of participants received the drawings with the one set of labels (e.g., one group received the label "moon", whereas the other group received the label "C" for the crescent moon shape). After subjects practiced the drawings with their labels, they were given the labels to draw the images from their memory. The results indicated that even though the drawings were distorted towards the corresponding label, people were still able to use mental imagery to draw the items using their memory, and therefore, there were some pictorial representations in their memory. In another study, Shepard & Metzler (1971) conducted an experiment in which they presented participants with 2-dimensional (2D) shapes of 3-dimensional (3D) geometrical objects, some of which were rotated. Subjects were asked to report if the 2D drawings matched the 3D objects. They hypothesized that if participants decomposed each shape into propositions and then put it back

together as an image, the time it would take to respond whether a given shape is the same or not had to be independent of how much each shape was rotated. However, they found that higher degrees of rotations resulted in a slower response time, and therefore objects were represented as images with their topology being preserved.

Consistently, Kosslyn et al., (1995) measured brain activations using PET (positron emission tomography) during imagery tasks. They presented participants with pictures of different sizes, and then asked them to close their eyes and imagine the objects in their actual size. They found increased activations in the visual cortex, primarily in V1, during the imagery task, and that the locus of maximal activation changed based on the size of the stimulus that was imagined. These findings suggested that objects were stored in a pictorial format, for imagery generated activations in the primary visual cortex as if the object was actually seen by the observer. As other neuroimaging techniques such as functional magnetic resonance imaging (fMRI) became more prevalent, the theory of mental imagery acquired even more attention. For instance, Albers et al., (2013) conducted an fMRI study, in which they showed participants a grating and asked them to either store it into memory or rotate it and keep it in their mind's eye (i.e., imagery task). In the imagery task, after a few seconds, subjects had to judge whether the presented probe was rotated clockwise or counter-clockwise. Authors were able to successfully decode the newly rotated images in the imagery task and the actual stored images held in the working memory task from early visual layers (V1, V2 and V3). These results, altogether, suggest that imagining objects involves increased activations in early visual processing similar to actually seeing the objects, and therefore supports the mental imagery theory.

Along with propositional and mental imagery accounts, Paivio, (1971) postulated that information is represented in a dual format of visual and verbal (Figure 18). According to this dual code theory, the verbal system processes the verbal code of visual, auditory and other modality-specific representations such as teacher, opinion, happy, school, etc. that are either

concrete or abstract. It is important to note that these verbal codes are arbitrary symbols denoting objects, feelings and ideas. On the other hand, the non-verbal system processes modality-specific images of shapes, sounds (e.g., horn of a bike), emotions (e.g., racing heart) or other non-linguistic events or objects. These representations are similar to perceptual experience, rather than arbitrary symbols like verbal codes (J. M. Clark & Paivio, 1991).

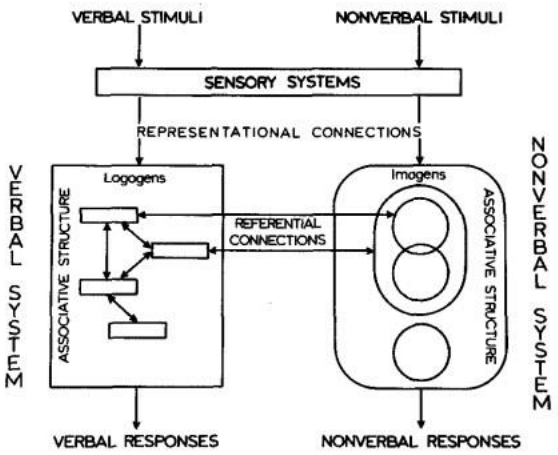

Figure 18. Verbal and non-verbal systems of dual coding theory (Source: Clark & Paivio (1991)). The schematic indicates two independent pathways to process verbal and visual information received from the sensory system.

There are also behavioral experiments supporting the dual code theory. For example, Paivio and Csapo (1969) presented subjects with a list of concrete words, abstract words and pictures in slow (2 items per second) and fast (5.3 items per second) rates. They assumed that showing items at a fast rate would prevent people from creating a verbal code of images. Subjects were asked to recall as many items as possible in each condition (i.e., free recall). The results indicated that people recalled more pictures accurately than concrete words, and more concrete words than abstract words in a slow rate presentation, whereas in the fast rate presentation the accuracy of recalling pictures and concrete words were similar. The authors hypothesized that when items are presented slowly, pictures are represented in a dual format of visual and verbal format, whereas in the fast rate condition, pictures are presented only in a visual format.

Therefore, people are more accurate in the slow condition compared to the fast condition in recalling the pictures. Paivio and Csapo (1973) also examined the representational code of picture superiority effect. The picture superiority effect indicates that if people are shown a list of pictures versus a list of words, performance in recalling pictures is superior to recalling words (Kirkpatrick, 1894). To test whether the picture superiority effect stemmed from dual or visual coding, Paivio and Csapo (1973) presented subjects with a sequential list of words and pictures. They found that if the picture accompanies its verbal code (e.g., showing a picture of a house and then the word "house") memory is better than if the same image or word is repeated (e.g., showing a picture of a house or the word "house" two times in the sequence). They concluded that the picture superiority effect may be the result of encoding both visual and verbal codes, such that an instant verbal code is generated when people see pictures (also see Hedayati & Wyble, 2020) . As a result, each of these codes enhance memory performance.

Reviewing the dual code theory, the account is characterized by independence of modality-specific codes such that they have additive features (Paivio, 1990, 1991). Independence also indicates that there exist separate pathways to process each code as demonstrated in Figure 18 (see also Bahrick & Bahrick, 1971). To show this independence empirically, (Thompson & Paivio, 1994) conducted an experiment in which they showed subjects either a list of pictures, a list of corresponding environmental sounds, picture- sound pairs (i.e., picture was presented simultaneously with its corresponding sound such as showing the drawing of a telephone and playing its corresponding sound) or repeated pictures. They hypothesized that if auditory and visual components were functionally independent in memory, they should contribute additively to recall each item. The results demonstrated higher accuracy for picture-sound pairs compared to all other conditions. This was consistent with dual code theory, as memory contains the modality-specific codes of items and therefore each code could be retrieved independently. Given all of the above studies, the answer to the question of *what code do we represent in memory?* seems to

depend on several factors. One of these factors is task demand. In other words, how the task is set up could highly influence what code is encoded into memory. In a study conducted by (Weldon & Roediger, 1987), they presented participants with a mixed list of words and pictures. Similar to previous studies, that in a free recall task, people could recall more pictures than words ((Kirkpatrick, 1894; Paivio & Csapo, 1973), however, when they primed the participants with word fragments (e.g., __ou__s__ for the word "house") or incomplete drawings of the presented pictures, the accuracy of recalling words were higher than pictures. This reversed picture superiority effect would be harder to explain via the dual coding theory, as according to this theory, pictures are better encoded because they are encoded with visual and verbal codes. In another study, Hedayati, et al. (2022), conducted an experiment in which they showed participants a handwritten MNIST digit for 50 trials and asked them to type in the digit category at the end of each trial. In trial 51, subjects received a surprise trial in which they were given four exemplars of the digit they just had seen (e.g., four different shapes of digit "2") and asked what specific digit they saw. The accuracy dropped from 97% to 15% at the surprise trial, indicating that people did not automatically store the details of the visual code due to the task demand in pre-surprise trials. However, in the subsequent post-surprise trials and once participants realized that the visual details of digits could be task relevant, the accuracy elevated to 100% (see Experiment 3 in Chapter 3). As a result, task demand could have a significant impact on what codes are represented in memory.

In this study, using the MLR model, we intended to examine the dual coding theory when visual details are task relevant. Particularly, we investigated the role of visual and verbal codes in retrieving the visual details information to see how each of these codes contribute to retrieve the specific shape of an object.

Dual Coding Storage with MLR

Dual code MLR architecture

We used a pre-trained VAE that was trained on grayscale MNIST dataset. This model was the simpler version of MLR that we used in Chapter 1 without having the color map and skip connection. The model's bottleneck consisted of a 4-dimensional map to represent only the shape of digits. This pre-trained VAE was considered as the visual pathway of the model that processed only the visual information. To test the dual code theory, we needed to include a separate pathway to process verbal information as demonstrated by Paivio (1971), thus visual and verbal codes could be represented in different areas of the brain. In this regard, we created a label network as a multiple layer perceptron (MLP) that processed the verbal one-hot code labels. This label network consisted of three layers with dimensions of 10, 7 and 4 respectively (Figure 19). Note that the input of the label network had the dimension of 10 to receive one-hot code labels for each digit.

Training and testing the MLR

The goal was to train the label network such that the model could generate images based on the verbal codes. In doing so, the MLP received one-hot codes as the inputs and learned to map them into the shape map activations within the VAE's bottleneck, while the VAE's encoder simultaneously processed the visual information of the corresponding stimulus. In other words, an image was passed through the VAE, which created activations in the shape map. At the same time, a one-hot coded label of the same image was fed to the MLP. The objective function used to train the MLP minimized the mean squared error (MSE) between the shape map activations and

the activations generated at the last layer of the MLP (i.e., 4-dimensional layer). For instance, the VAE first received an image of a “2” and that was subsequently represented in the 4-dimensional shape map. Then, the one-hot code of digit “2” was simultaneously fed into the label network as demonstrated in Figure 19. The MSE objective function minimized the difference between the label network and the encoder to map one-hot labels to the visual information. Note that during training of the label network, weights of the encoder were held constant by turning off the gradient. This allowed us to keep the shape map representation intact, while training the label network.

Once the label network was fully trained, we could reconstruct a given image only by providing a one-hot verbal label (Figure 19). To prevent the model from generating shapes of a category that are visually very similar (e.g., a one hot code of “2” would always generate an image of a 2 with the same specific shape), we added a normal Gaussian noise with a standard deviation of .2. This resulted in generating 2’s with different shapes.

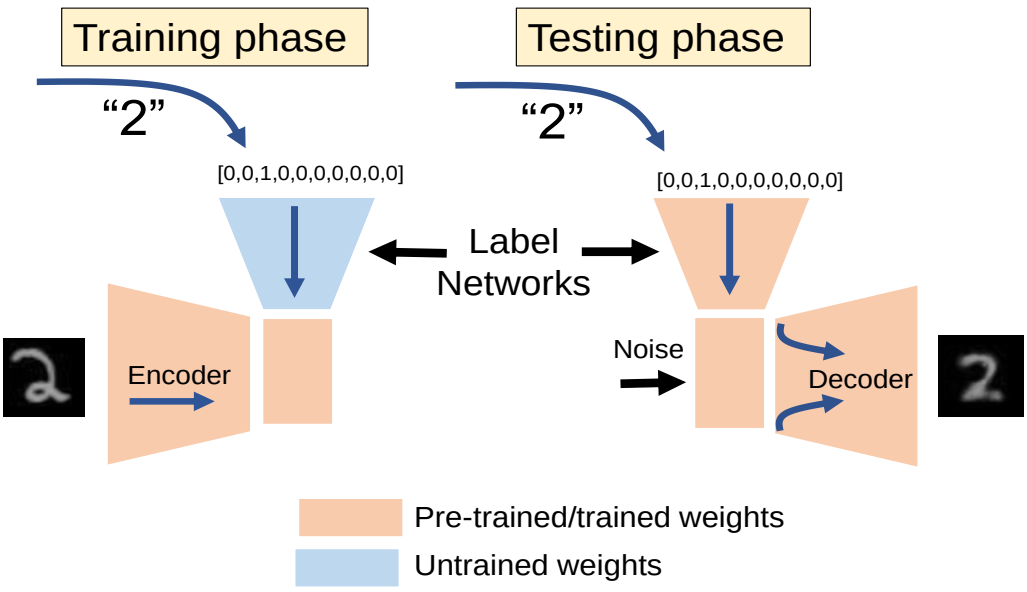

Figure 19. Illustration of the label network (green network consisting of two multiple layer perceptron) and the mVAE encoder during the training of the label network. The model receives an image and its categorical labels of shape and color simultaneously to map the labels into shape/color map activations

Simulation and results

We tested the memory performance by storing multiple items in the BP along with their one-hot code verbal labels. Figure 20, demonstrates the steps of storing and retrieving three items, one of which is cued and is considered as the target. Each of the items could be represented as verbal, visual or dual codes within the binding pool. Note that the stored items were always randomly selected from different categories of digits (e.g., 1,4,0). The BP activations were retrieved such that the visual information was projected back to the shape map and the verbal code was projected back to the first layer of the shape label network. Retrieval formulas were similar to formulas described in chapter 2.

For each stored item, there were three ways that it could be reconstructed. The visual code retrieval was the result of reconstructing the shape map via the decoder. The verbal code retrieval, on the other hand, was achieved by reconstructing the 4-dimensional space from the label network, and then decoded it via the decoder. Finally, to retrieve a given item via dual codes, we averaged the label network representation with the shape map and reconstructed the resulting activation via the decoder.

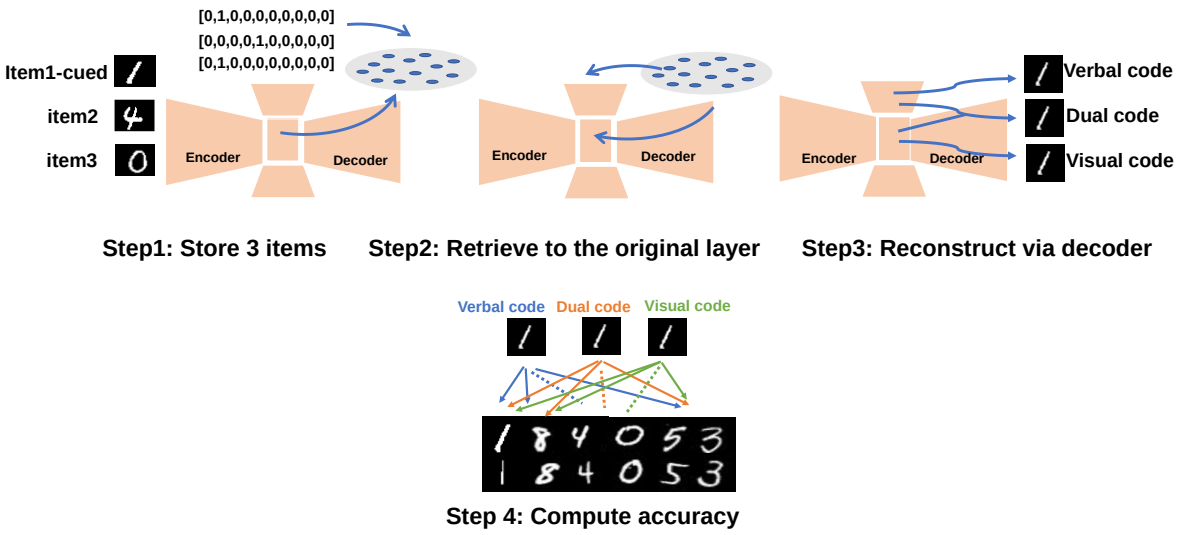

Figure 20. Simulation process of the recognition task. At step 1, visual codes of the shape map as well as the one-hot verbal codes are encoded into the BP. At step 2, both verbal and visual codes are retrieved via inverse matrix multiplication. At step 3, either visual, verbal or both codes are retrieved via the decoder. At step 4, the cross-correlation of each retrieved code of the cued item (aka target) is compared with a display consisting of items from the same or different categories of the target. The percentage of the times that the model detects the target (the one with the highest cross correlation value) is reported as the accuracy.

To compute the accuracy, we compared the cross correlation of visual, verbal and dual code retrievals of the cued target with images of the same and different categories. For instance, if there were three items stored in memory as 1, 4 and 0, the response options were two exemplars of each item in addition to two exemplars of three other categories that were not in memory (e.g., 8, 5 and 3). That led to the total of 12 comparisons for each retrieved code.

That is, we computed the cross correlation between each retrieved code of the cued target (i.e., visual, verbal and dual) and the response options. The item with the highest cross correlation with the cued item was selected as the target by the model. Accuracy was estimated by measuring the number of times that the model accurately detects the cued target for 100 repetitions and 5 independently trained models.

The mean accuracy of recognizing the target as a function of set size is shown in Figure 22 for each code type. The results indicated that discriminability between display items is best

obtained via visual (or dual) code, and that verbal code did not enhance the memory trace of visual details in the recognition task. This is because visual and dual code accuracies are relatively similar, particularly in set sizes greater than 1.

However, the baseline cross-correlation might differ in each condition. For example, it can be the case that more shape information exists in the dual code condition as measured by cross correlation, but the amount of visual detail for discriminating between options to compute accuracy does not reflect this. It is critical to clarify the terms *shape information* and *discrimination information* in the context of this study. Shape information refers to the number of bits of information shared between two images. On the other hand, discrimination information is the specific details that is critical to discriminate between images of the same category (Figure 21).

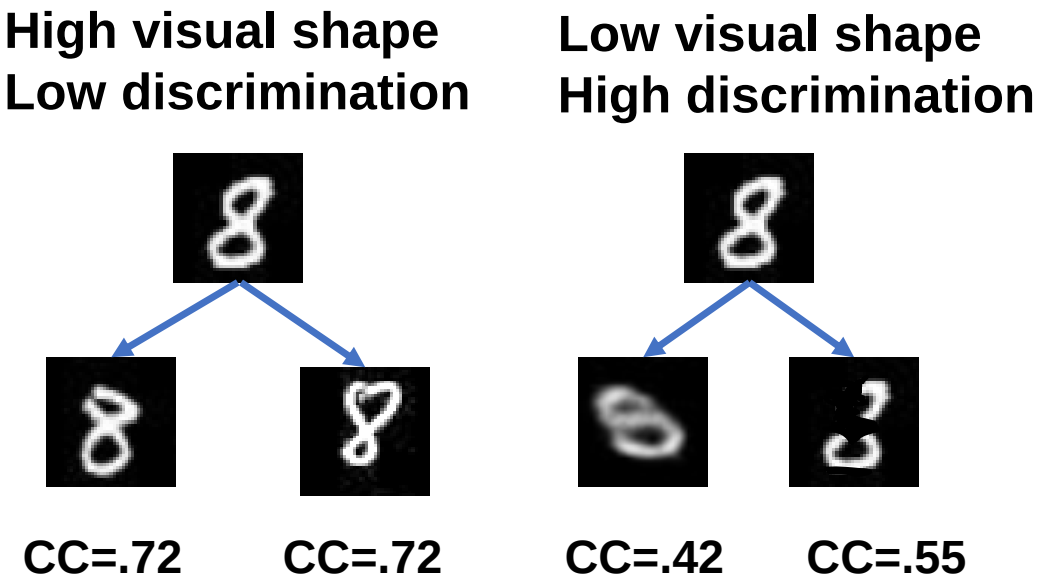

Figure 21. Demonstration of visual shape and discrimination information. CC represents the hypothetical measured cross correlation between the upper 8 and each of the images. A given 8 can have a high baseline cross correlation between two images of the same category (left), however the there's no difference between the CC's. On the other hand, the same 8 can have a low baseline cross correlation with two other images, but the discrimination information exists more in the one with higher CC.

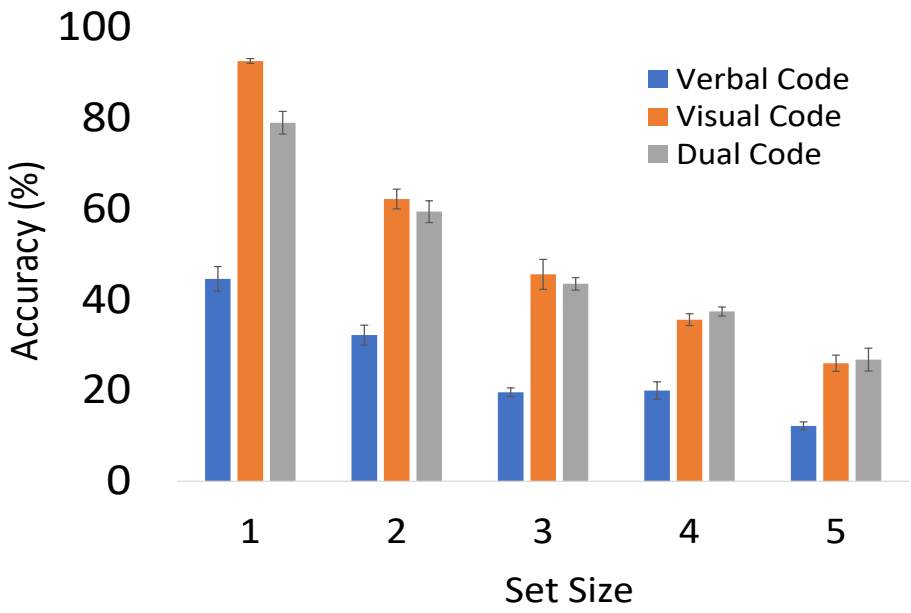

Figure 22. Mean accuracy of detecting the target estimated by computing the cross correlation between each retrieved code type (i.e., verbal, visual and dual) and the display items, and picking the highest cross correlation as the target. The accuracy was computed over 100 repetitions on 5 independently trained models for 1-5 set sizes. Error bars indicate standard error.

To examine the visual information that is shared between the target and its memory retrieval, we computed the cross correlations between the retrieved target and the ground truth target obtained from 100 repetitions for 5 independently trained models (Figure 23).

According to the cross-correlation results, the dual code does not necessarily contain more visual information, as the amount of cross correlation values is comparable to visual code. Furthermore, similar to the accuracy results, verbal codes demonstrated the lowest baseline values compared to the visual and dual conditions.

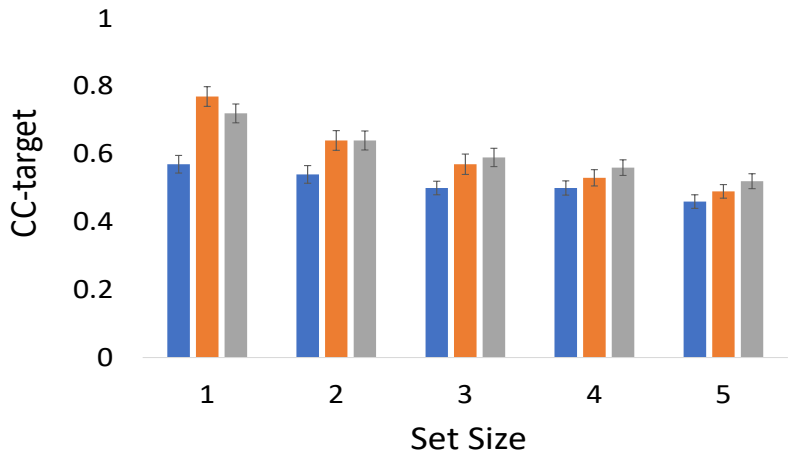

Figure 23. Mean cross correlation between the target and retrieved visual, verbal and dual codes for 100 repetitions on 5 independently trained models. Error bars indicate standard errors.

We have also examined the accuracy of retrieving the target category as a function of set size, by measuring the number of times that the model can accurately recognize the category of the target (Figure 24). The figure shows the percentage of times that the model recognized the correct category. As it is evident, the overall accuracy for each of the code types has been increased and the difference between the accuracy of verbal code and visual/dual codes is less compared to previous conditions demonstrated in Figure 22.

However, using visual codes are still preferred to accurately estimate the target category. Note that as mentioned before, the method of cross correlation is to compute how much shared visual information exists between two items. In this respect, to recognize a correct category of the target using visual details, visual codes are still a more reliable source of information than the verbal codes as they contain more of the necessary information. Consistently, no noticeable difference between visual and dual code was observed. That is, verbal code did not enhance the memory of the visual code when the item's category was being retrieved.

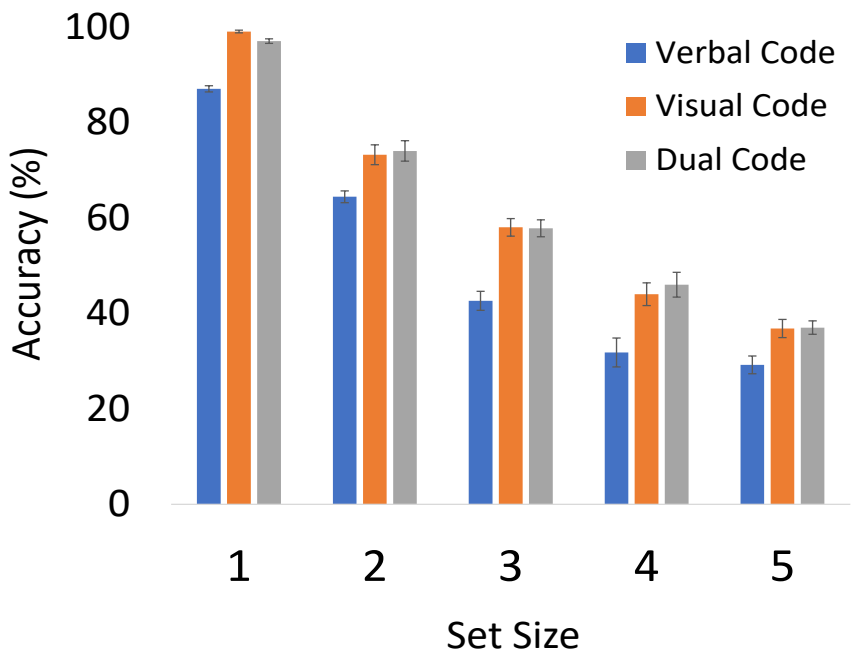

Figure 24. Mean accuracy of detecting the correct category via retrieved visual, verbal and dual codes for 100 repetitions over 5 independent trained models. Compared to Figure 20, the difference between visual and verbal code accuracy have been decreased meaning that using verbal codes we have more visual information to detect the target category, however, visual codes indicate better memory performance for recognizing the target category. Error bars indicate standard errors.

Dual Codes Decay in Memory

According to the model's simulations so far, our hypothesis is that verbal codes do not contribute to memory in a recognition task when visual details matter, but they might still be accessible. In other words, when we have the visual details in memory, we can generate a verbal code anytime from the visual information. Furthermore, in some visual memory tasks, to prevent subjects from generating and storing verbal memory they are asked to repeat a word while they are doing the task. This is known as articulatory suppression and is to encourage participants to rely on visual memory rather than verbal memory (Richardson & Baddeley, 1975). The question that arises here is whether there is a benefit in generating and storing verbal labels.

We hypothesize that one of the ways people benefit from encoding verbal code is that they are more robust over time. Experiments have shown that verbal coding of a visual code is more sustainable in memory than the visual code, such that in recalling an event, visual details associated with that event would fade away or get distorted at a faster rate than the verbal codes. For instance, Dunkin et al. (2015) suggested that verbal coding affects the perceptual memory of a colored square in terms of its distortion over time, and that verbal code can last for a few seconds in memory.

We tested our hypothesis with regard to verbal code robustness using the MLR model. It should be noted that the MLR model does not have a component to account for time in memory, and that is one of its limitations. However, we assumed that as information is retained longer in memory, further noise is accumulated and added to the memory. In this respect, increasing the added noise into memory is equivalent to maintaining the information for a longer period of time. Hence, we examined the accuracy of memory retrievals for verbal and visual codes when some amount of noise is added into memory in an ascending manner.

Visual and Verbal codes decaying over time in MLR

To test the robustness of visual and verbal codes in memory, we randomly selected two MNIST digits and stored them into the BP via shape map activations. Using SVM classifiers that were trained on the shape map, we estimated the categorical information of the stored images prior to their encoding in the BP (Figure 25). This was similar to encoding visual and categorical information in one memory trace described in Chapter 3.

Once the items were stored along with their categorical labels, ten random values were sampled from a normal gaussian distribution with mean 0 and standard deviation of 1.0. Sampled values were sorted ascendingly based on their absolute values, and were added to the BP

activations of the visual and verbal codes. Figure 23 shows the steps of the simulation from encoding to retrieving the items.

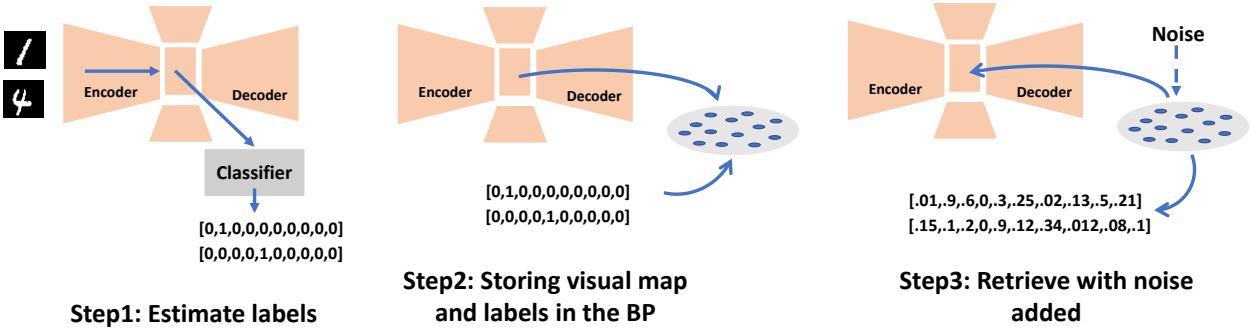

Figure 25. Items are encoded via the shape bottleneck, with a trained classifier to estimate the categorical code of the items (Step 1). One-hot verbal codes and the shape map visual code were simultaneously encoded into the BP (Step 2). Each code (visual and verbal) was retrieved, while an ascending vector of noise was added to the BP (Step 3).

Items were retrieved similar to the retrieval process explained in Chapter 2. To estimate the accuracy of retrieved visual codes, we applied the classifier to the retrieved visual codes within the shape map. On the other hand, the accuracy of verbal code retrieval was estimated by converting the retrieved verbal code into a one-hot code and comparing it with the one-hot codes estimated by the classifier prior to encoding.

This simulation is equivalent to an experiment in which people are shown two MNIST digits to remember and are asked to report their category in various time intervals. The MLR predicts that people can easily use the memory of verbal codes to report the category of digits, whereas memory of the visual codes are more susceptible to noise over time, and hence cannot feasibly be used to report the target category as the time interval increases.

We assessed the accuracy of visual and verbal codes with increased noise by averaging across 5 independently trained models with 100 repetitions for different set sizes. Figure 26 indicates the mean accuracy of retrieving visual and verbal codes for set size 1 and 2 with different levels of noise.

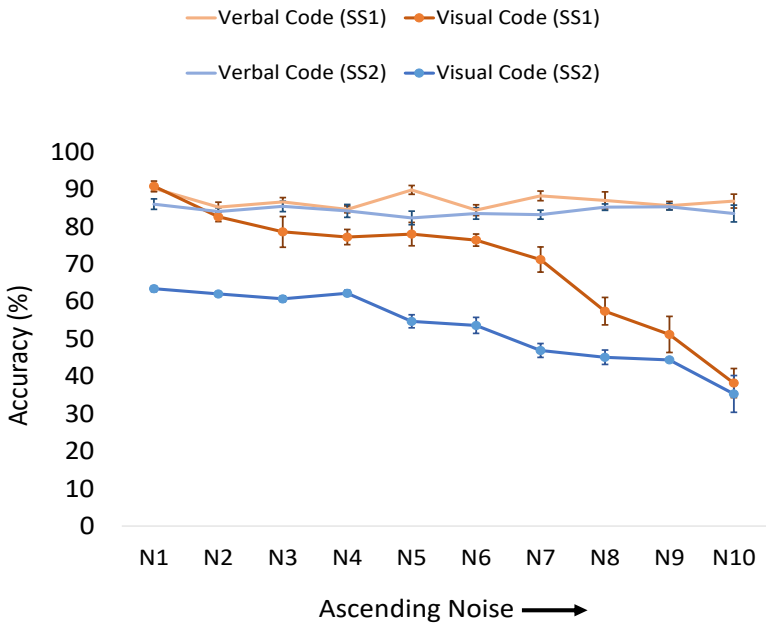

Figure 26. Mean accuracy of retrieving the target category using verbal vs visual code for set size 1 (SS1) and set size 2 (SS2). The orange lines indicate accuracies for SS1, whereas the blue lines show accuracies for SS2. As the amount of noise increases, the extent at which the memory of visual codes are impaired increases. On the other hand, verbal codes are more robust with increased injected noise.

Simulation results

As shown in Figure 26, with increased noise, the accuracy of retrieving the category of a given MNIST digit decreases via the visual code. This confirms our hypothesis that visual codes are susceptible to distortion over time with accumulated noise. On the other hand, verbal code is almost unaffected by the amount of noise added to the BP, meaning that over time, the categorical code is a more reliable source to retrieve a category of an item. It is important to note that unlike dual coding simulations, this simulation is not concerned with visual details information and the performance is evaluated by retrieving the category of the stored items.

Another notable difference between visual and verbal codes is the effect of set size. Increasing the set size, which can also be considered as a form of noise in memory (i.e., The interference in the BP increases as a function of set size due to shared neural resources between

items) impacts the visual more than the verbal code, suggesting that category of items can be retrieved better via verbal codes than visual codes when more items are stored in memory. This is consistent with simulations related to encoding visual and categorical information described in Chapter 3.

Conclusions

In this chapter, we examined the dual code theory in a recognition task using the MLR model. The MLR model predicted that visual codes are sufficient to recognize a target among distractors. Despite the fact that the label network was trained to map verbal codes into visual codes of the shape bottleneck, verbal codes encoded into memory did not contain enough details of the visual shape to be useful in recognizing the target. More importantly, comparing dual code with the visual code we found that verbal codes did not enhance memory of visual details, and that in a recognition task when it is critical to remember the visual details, visual code theory provides a simpler explanation for people's performance.

In another set of simulations, we showed that verbal code obtained from categorical information of visual items is more robust against noise in memory. That is, over time when noise is accumulated, visual codes get distorted. However, the extent to which verbal codes change in memory is much milder. As a result, verbal codes are beneficial to remember when items' categories are retrieved later in time, or when more items are to be encoded.

The simulation results highlighted the fact that addressing the question of "what code do we store in memory?" highly depends on factors such as task demand (e.g., free recall or recognition), stimulus type, number of items to encode, and the retrieval time. Therefore, either visual, verbal or both codes can be encoded into memory depending on those factors.

Chapter 5

Imagery and Creativity

According to mental representation theories in a previous chapter, we learned that mental imagery enables us to represent information in a pictorial format. Imagery, characterized by imagining objects in mind's eyes involves a process in which we reconstruct a previously seen object or a novel shape by combining familiar elements (e.g., imagining a cat's head on a horse body) in our minds, particularly in the absence of a direct stimulus (Kosslyn, 1996). But how do we imagine objects and combine them to create new forms?

For millennia philosophers have been interested in our ability to conjure memories and objects as a functional component of our visual system. Visual imagery allows us to re-experience remembered content and may enable us to mentally manipulate that information in a format that is similar to the original input from the eyes (Kosslyn et al., 2006). In this regard, Finks et al., (1989) conducted an experiment in which they asked people to imagine two predefined familiar shapes, and superimpose them while they have their eyes closed. Then, they asked subjects to report any emergent forms that they saw as the result of this superimposition. As an instance, they asked subjects to superimpose the letters 'X' and 'H'. The resulting emergent feature could be a butterfly, a bow-tie, four triangles, a letter M, or any other recognized forms. After participants reported all the emergent forms that they could imagine, they were asked to open their eyes and draw the final patterns that they had imagined. Following that, they were asked to examine their drawings and recognize other patterns that they had not seen in their mental image prior to drawing. The results demonstrated that a great deal of emergent patterns were generated prior to the drawing, and that people had a great ability to combine familiar patterns using mental imagery.

In a related experiment conducted by Finke and Slayton (1988), subjects were shown geometrical shapes and alphanumeric characters, and were asked to synthesize these items in their minds to create new patterns without specific instructions of how the images should be combined. For example, two squares and a triangle could be formed into the shape of a house. The results showed that people generated new forms in about 40% of the trials, and these patterns were rarely predicted by a naïve experimenter. Authors concluded that people are capable of generating synthetic shapes from simple parts using mental imagery without explicit instructions or training.

The above experiments highlighted the role of mental imagery in generating creative visual forms (Finke, 1996). From the theoretical vintage point, Finke et al., (1992) proposed a model of creative thinking named geneplore (Figure 27). The model consists of two cognitive phases of generative and exploratory processes. According to this framework, pre-inventive structures such as visualized forms and verbal/conceptual combinations are generated using the generative process. These generated patterns or concepts are further interpreted and evaluated by the exploratory processes, for example, in the case of superimposing two letters, subjects evaluate the emergent configuration after each letter was generated. Then the final form can be modified by additional generative processes (Finke, 1996).

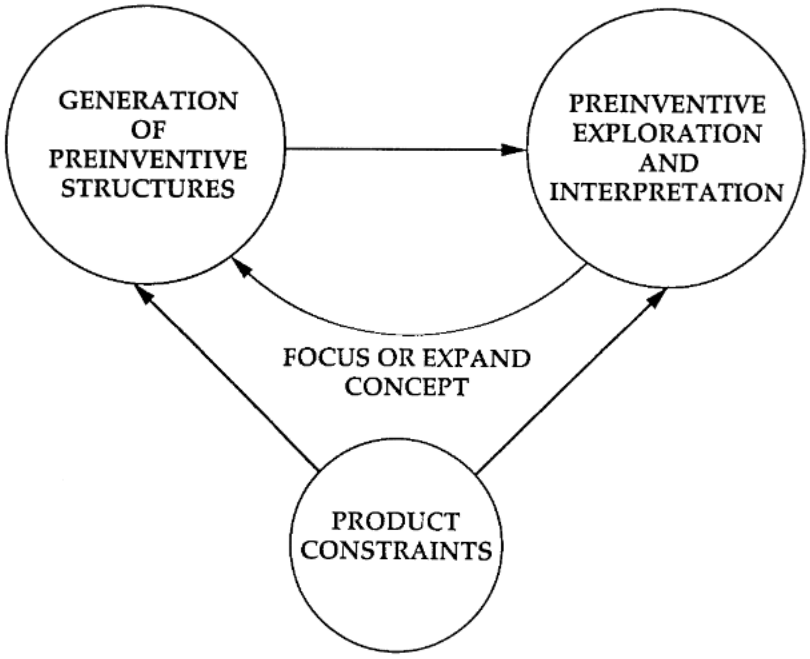

Figure 27. Geneplore model of creative thinking and imagery (source: Finke et al., 1992). Two core elements of generative and exploratory processes bring about creative forms or solutions. There is a bidirectional connection between these two cognitive mechanisms. At any point in the cycle of creative thinking, there could be specific constraints for what the final product would be.

As described above, literature on creative imagery has many informative experimental and theoretical investigations about how mental imagery is used to create new forms. However, none of these works provide a functional, and mechanistic explanation of how simple features and shapes are generated, and are combined in new ways. In this regard, the mechanism underlying creative imagery is yet to be explored.

In this study, we aimed to use the MLR model to form an intuition about how creativity in the context of mental imagery works. As mentioned earlier, computational models can provide a functional intuition about how information is transformed from sensory input, memories and visual knowledge, into an experienced visual form.

In recent years, in the field of artificial intelligence (AI), there have been models that alleged to demonstrate “creative” behavior (e.g., Ramesh et al., 2021). However, these models

are too complex to be understood as a model for human imagery and creativity. For instance, GPT-3 (Generative pre-trained Transformer; Brown et al., 2020) is a language model that can generate essays, translate between languages, and answer questions. The model consists of ~170 billion parameters and it was trained 570GB images from CcommonCrawl dataset. Another model proposed by Radford et al., (2021) named CLIP (Contrastive Language- Image Pre-training) can predict texts (or descriptions) for images in a zero-shot manner, meaning that it likely predicts accurate captions for image classes that were not in the training set. The model had billions of parameters and was trained on 400 million image-text pairs. Consistently, DALL-E (Ramesh et al., 2021) can generate new images corresponding to a description (e.g., a purse on an avocado chair). DALL-E consists of 12 billion parameters, and was trained on 250-million text-image pairs gathered from the internet. As a result, the datasets that were used to train these models were so large that it is impossible to understand the full scope of the examples they have been exposed to, and so it is difficult to determine the boundaries of creativity in such models.

While the generative output of these models is impressive, it is difficult to determine if such output is an interpolated sample that provides intermediate variation from the training set. Moreover, models with large training sets generate harmful and inappropriate contents which creates further barriers for using them as inspiration for studying creativity (Birhane et al., 2021). As demonstrated by the above models, the emphasis in AI research is often to build ever larger models with the idea that models that are large and trained on sufficiently massive data sets will approach human levels. However, the larger the systems get, the harder it will be to understand how they work, to know whether they are truly creative, and to use them as a lens to study human creativity. With this regard, truly creative means being able to utilize learned knowledge while combining elements in a way that generates a novel form.

Our goal is to move in the opposite direction, with a cognitively and biologically plausible model that uses simpler machinery. We aim to explore imagery/creativity using the

modified MLR model. Due to its generative feature and its biological relevance to the visual system we modified this model to explore visual imagery and creativity.

Bearing in mind the importance of avoiding inappropriate inference from artificial networks to neural systems (Guest & Martin, 2021), we are not arguing that the proposed MLR-based model is similar to human imagery and creativity. Rather, we are using this model to develop more accurate intuitions about how high dimensional systems can operate, which in turn allows us to think more clearly when developing theories.

Finally, previous research has shown the critical role of working memory in creative tasks (Benedek et al., 2014). Therefore, another advantage of using MLR is that it has an embedded working memory component that can, if necessary, be used to combine information and generate new forms. It is important to bear in mind that the common notion of creativity is generating something that is novel and useful. For the purpose of this study, we are focusing on novelty, because usefulness is outside of the scope of our model. We conducted a series of simulations to explore the ability of MLR in imagery and creativity tasks.

Imagery Model Architecture

The architecture of the model was similar to the dual-coding model except that there were two label networks (i.e., MLPs) for shape and color (Figure 28), and the training set to pre-train the mVAE was colored MNIST dataset as it was used in chapter 1. Each of the MLPs had 10 input neurons, 7 neurons in the middle layer and terminated on the 4-neuron shape or color map respectively. Similar to the dual-coding model, the label network started with a categorical (one-hot) coded representation and learned to map those codes onto representations in the shape and color maps that were generated by the mVAE's encoder based on the visual stimuli. Gradient descent minimized the mean squared error between the activation generated by the labels and the

activation generated by the corresponding image. After the training phase was completed, our model could create combinations of handwritten digits with different colors using the one-hot codes which we take as an analog of visual imagery.

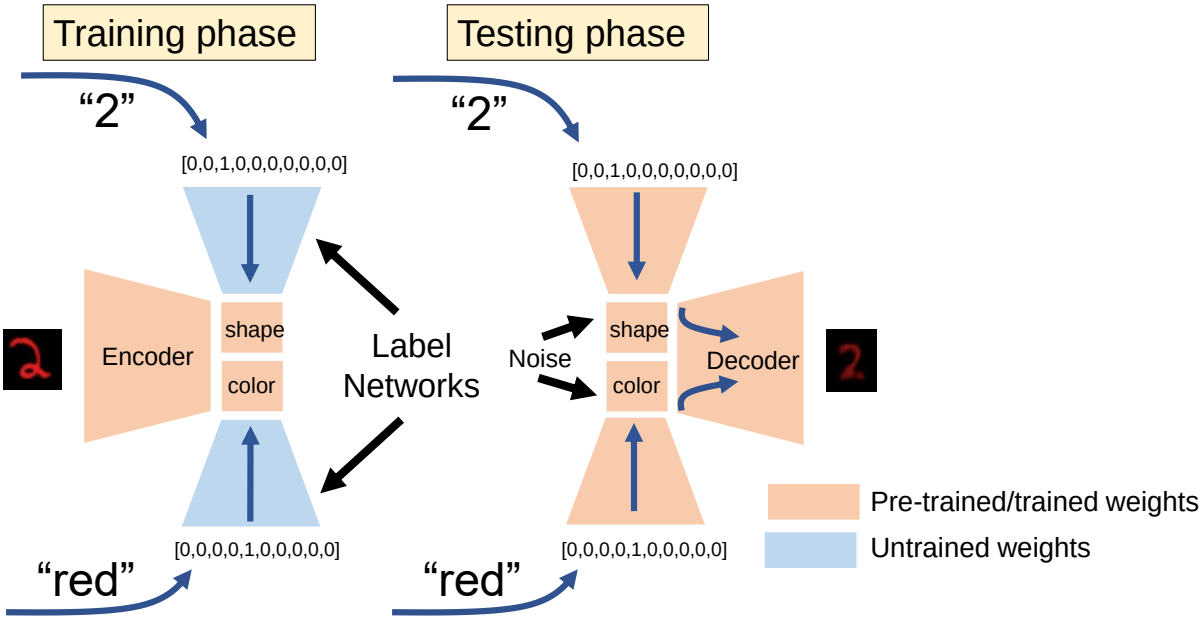

Figure 28. The proposed model’s architecture with label networks attached to a pre-trained mVAE. Initial training of the label networks is shown on the left (stage 1), and after training the label networks are able to generate arbitrary shape-color combinations in the absence of visual inputs (stage 2). The binding pool, which stores and retrieve from working memory is not shown

Simulation 1: Familiar Combination

Familiar combinations indicate color-shape combinations that the model has seen during the training. Figure 29 (Panel A) shows how imagining a “purple two” changes as we add increasing Gaussian noise to the shape/color maps. The label networks first activate a representation in the shape and color maps that collectively produce a purple 2 at the output. Increasing amounts of noise are then added to these latent representations which are passed through the decoder to generate the images. This simulation

demonstrates that random perturbation of a representation does not generate new shapes or colors.

Simulation 2: Novel Combinations

In simulation 1, the model was trained on all 100 combinations of the 10 shapes and 10 colors. To test the model's ability to generate novel shape-color combinations, in Simulation 2 we trained the model on red 0-4 and green 5-9 MNIST digits and tested its ability to generate red 5-9 or green 0-4 With 150 attempts. This model was unable to generate a combination outside of its training set. Figure 29B shows 50 examples of noisy imaginations of red 2, green 5 and green 2. The model generates only combinations of the red and green digits that it was trained on. In the lower third panel the label network tries to reconstruct a specific combination that it was not trained on (green 2) and the model is unable to do so.

Simulation 3: Novel Combinations with Overlapping Training

In the previous simulation, the two sets of training stimuli were completely disjoint (red 0-4, green 5-9). It is possible that partial overlap of color-digits in the training set would allow the mVAE decoder to generate more flexible representations that enable the reconstruction of novel digit-color pairings. To evaluate this possibility, a new model was trained with the following sets: red 0-3 and green 2-5, such that digits 2 and 3 were presented in both red and green. As shown in Figure 29C, even with this new training set, red 5 could not be generated but green 5 could be generated.

Also see Figure30 (left) for a t-SNE map illustration of the shape map and its comparison with a t-SNE map for a model that was trained on red and green digits with complete overlap

(Figure 30, right). The t-SNE maps allow for visualizing high dimensional data by mapping them into a lower dimensional space (Maaten & Hinton, 2008). The clusters within the maps show that red and green representations are more entangled based on their colors, and that clusters are more spread out when combinations have complete overlap.

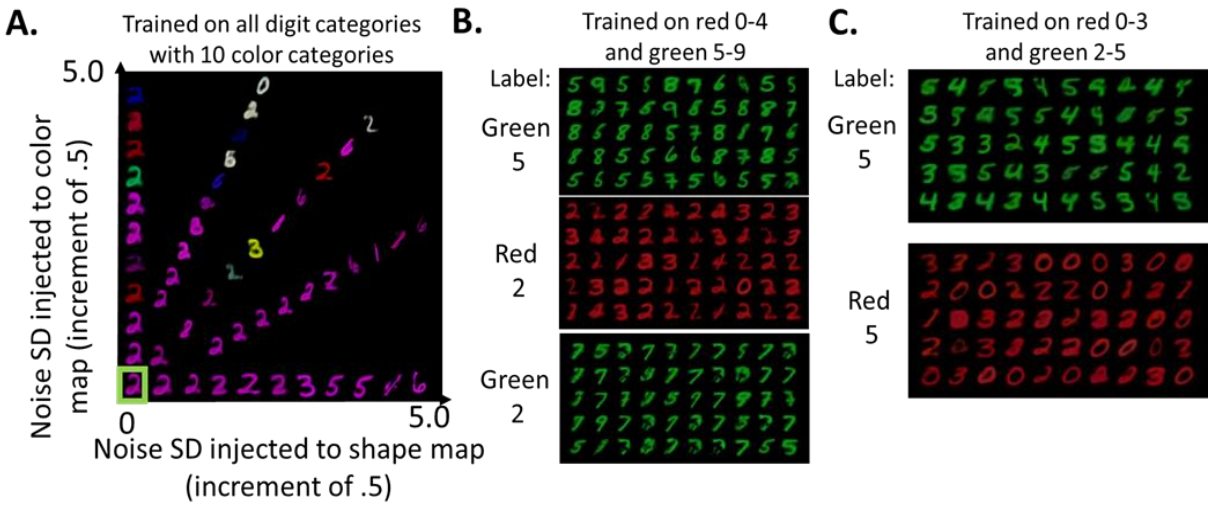

Figure 29. The image within the green rectangle is the reconstructed "purple 2" in the absence of noise on the model that was trained on all combinations of 10 digits and colors. Panel B, imagined images of the model that was trained on red 0-4 and green 5-9. The 50 images shown for each digit-color combination are the result of combining the label network's output with Gaussian noise (SD=1) added to the shape map. On the bottom, the label network tries and fails to generate a green 2. In Panel C, for the third simulation, even a model with overlapping shape and color combinations is unable to generate representations that are outside of its set of trained combinations.

To explore the latent space of the model that failed to generate a "green 2" combination, we trained classifiers on shape and color maps respectively. The average accuracy of decoding color from the shape map for 5 models was 75.8% (*SE*= 1.24, chance=50%) when the model was trained on non-overlapping digits (red 0-4 and green 5-9). On the other hand, the average accuracy of decoding color from the shape map was 60.6% (*SE*=1.46) when the model was trained on all combinations of red and green digits. These accuracies that are only moderately above chance level suggest that the model's inability to generate novel combinations is not due to the complete separation of colors as two clusters in the shape map. In other words, if colors were

represented separately in the shape map, we expected to see a very high classification accuracy (e.g., 95%), and that would have suggested that the way colors were represented in the shape map did not allow for creating novel combinations.

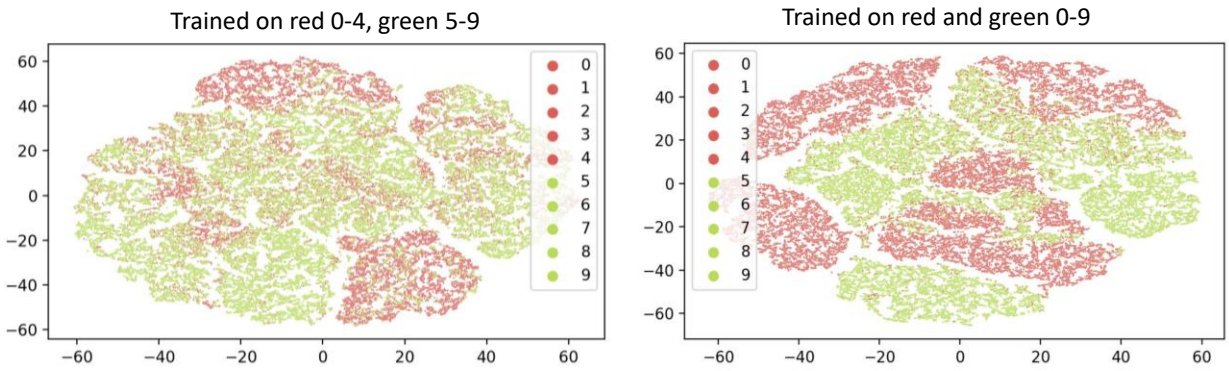

Figure 30. t-SNE illustration for the shape map when the model was trained on red 0-4 and green 5-9 (left) vs. when it was trained on red and green digits from 0 to 9 (right). The figure shows t-SNE reconstruction of the latent representation of the digits red 0-4 and green 5-9 for both models.

Simulation 4: Novel Combinations with Diverse and Overlapping Training

In this manipulation, we aimed to increase the latent space generalization by training the mVAE on MNIST and f-MNIST. By training the latent spaces on another dataset, we intended to increase the probability of representing variants of shapes within the shape bottleneck. We chose to use f-MNIST, because the dataset contains the same number of categories with articles of clothing that are quite different MNIST digits.

All of the training examples were represented in green and red, except that “2” was available only in red and “6” was available only in green. The results indicated that the mVAE could not combine the shape and color to generate new forms such as a “green 2” or “red 6”. Figure 31 indicates reconstruction examples directly from mVAE.

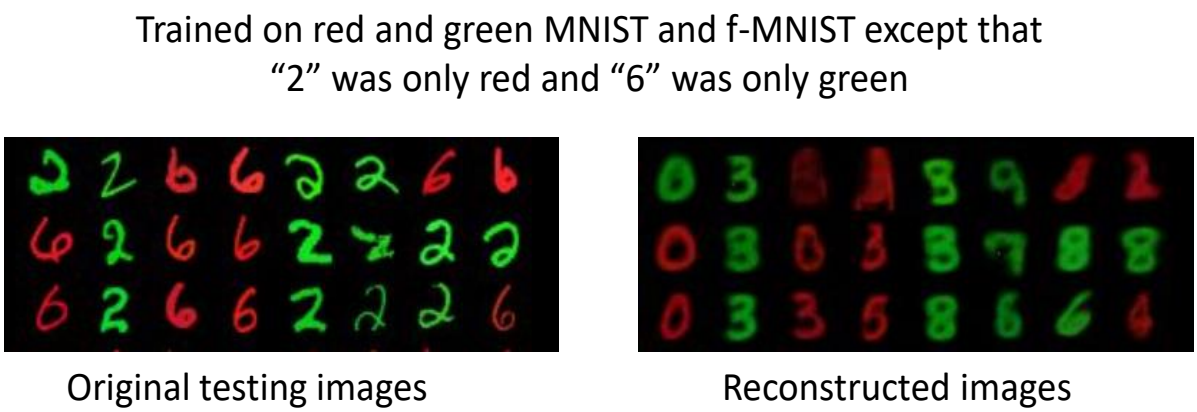

Figure 31. Examples of the direct reconstructions from mVAE for red 2s and green 6s, when the model was trained on red and green MNIST and f-MNIST, but "2" was presented in only red and "6" was presented only in green. The model is unable to generate green "2" and red "6".

Discussion of Results

MLR provides a simple, biologically plausible model of visual imagery that combines a hierarchical visual representation with a highly flexible working memory system. The training set uses clearly delineated combinations that we can use to study the building blocks of creativity in neural systems, such as imagery.

The model exhibits imagery in the sense that it can generate specific examples of targeted combinations of digits and colors based on labels provided by other putative cognitive systems (e.g., executive control or perhaps mechanisms associated with mind-wandering). For example, if asked to create a "“red 2" and provided with noise, it is able to generate a variety of images that are a red-2 in the absence of actual visual input. When noise is increased the imagined digits can change from one category to the next, but exhibit very little in the way of intermediate forms. Moreover, the model exhibits no ability to generate novel combinations of shape and color, such that it cannot generate a green 2 if it was trained on red 2’s and green 5’s. Even when the green and 2 latent representations have been activated by the label networks, the decoder has not

learned how to combine these specific representations at the output, so the failure to combine the features is in the decoder. Note that all of the green pixels required for a 2 can be found in the green 7 and 5, which means that the model has learned to generate green pixels along the bottom of the image, and yet, the closest approximation of a green 2 is a green 7. This suggests that there is no pathway connecting the latent representation of 2 in the shape map to green pixels in the output.

The simulations indicate that at least the comparatively simple VAE-based architecture is unable to generate truly novel combinations. While it remains an open question as to whether more complex models such as GPT-3, CLIP and DALL-E are genuinely creative, the example provided here suggests that we should develop more explicit means to evaluate the extent to which they can generate novel content.

Since MLR is inspired by cognitive and neurally plausible mechanisms of visual representation and working memory, it gives us a framework to explore possible mechanisms to induce creativity by using memory.

Creativity in humans is often a process of memory retrieval combined with recombination (Dietrich, 2004; Feldhusen, 2002) and the MLR model provided a conceptual platform to explore these ideas. Furthermore, because autoencoders like a VAE generate output that matches the format of the input, any output can be used to generate new training examples for the input space. This allows the model's expertise to grow to include a superset of combinations of real-world stimuli it has been trained on through self-generated replay. This requires additional mechanisms that can be explored via models that have cognitive components such as working or episodic memory. For instance, any given image can be generated and briefly stored in memory at the decoder's output to be combined with other pieces of output. The combination could then be used as new data for training. As a simple example of this, if the model is trained on a horizontal line and a vertical line but not a '+' shape, it would be unable to

simulate the imagination of a '+', since there is no representation of this more complex shape in any of its latent spaces. However, it would be possible for the model to retrieve a memory of the two lines and superimpose them on the output space, and then use this newly generated '+' shape as training data, allowing it to learn representations of the '+' in all of its latent spaces, as if that symbol had been present in its original training set.

More complex compositional learning and generation algorithms (e.g., Lake et al., 2011) could be used to achieve more complex representational combinations of shape as well. This form of providing additional training to the model based on recombination of its existing representations might be akin to the act of dreaming, in which hippocampal areas of the brain are often highly active and thought to be retrieving information. The advantage of using imagined visual forms as training data for the model is that these new forms become part of the permanent representational architecture of the system and thereby allow it to respond rapidly and efficiently to those forms in the future, despite having never experienced them before.

Other forms of representational manipulation and combination are also possible in a cognitively inspired system. For example, color information is thought to be represented in a way that allows it to be projected across surfaces even when it is not physically present (Figure 32).

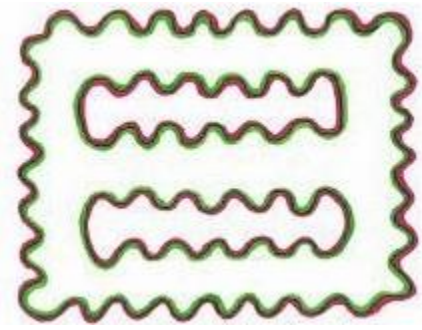

Figure 32. Watercolor illusion, in which the visual system expands the colors of red and green to create an illusory percept of colored surfaces (Pinna et al., 2001).

Similar mechanisms of color projection could conceivably be used to generate new color-shape combinations at the output layer. For example, if a given color is first projected from the decoder to the output space, and this is followed by a shape, the visual system might map the color onto the shape to create a new combination of shape and color. While it is premature to stipulate exactly what neural circuits would be responsible for such a transform, it is within the scope of functions that the intricate circuitry of the visual cortex could accommodate. As suggested above, these novel shape/color combinations could then be projected back through the model as a form of training, allowing them to be ensconced into the encoder, decoder and shape/color maps.

Such imagination of novel combinations thus seems within the scope of a model that has both memory and generative output. A better question than how does creativity occur is perhaps how a creative model would guide its creative imagination so that such retraining did not include all possible combinations of stimuli which might overload the model's representational space with conjunctions that never occur in the real world and thus are not helpful.

Another important point to consider is that there could be adaptive value in having an imaginative output at the earliest levels of vision (i.e., the output layer of MLR, that we consider to correspond to visual cortex which receives extensive feedback connections from higher level brain areas) to be constrained during waking behavior to imagine only stimuli that correspond to actual objects that have been experienced. A visual system without such controls might be overly prone to hallucinations that interfere with the ability to perceive. In such a case, the aspect of creative imagery that generates entirely novel forms in a visual format could be the province of generative visual workspaces that are not associated with perception. This would be unlike the framing of the MLR model, which assumes that the pixel level representations of both input and output correspond to early visual areas. It is notable that visual imagery of remembered items tends to activate early visual areas (Kosslyn et al., 1995, 2001) and this would support the MLR

interpretation, but it is possible that other brain areas are involved in the imagination of novel forms and this is supported by studies that fail to find visual cortex activation when using imagery to solve a problem (Knauff et al., 2000).

Conclusions

In conclusion, we have explored the ability of a modified VAE to create novel combinations of highly familiar visual features based on mental imagery. We found that mental imagery can occur for familiar combinations, whereas our attempt to create novel combinations with different manipulations was not successful. However, we suggested that when the model is coupled with memory systems and other aspects of representational interplay within visual areas, it is easy to theorize about how an autoencoder-based model could generate new forms and then use those as training data to expand the model's representational repertoire in a way that would be more consistent with some of the simplest aspects of creativity.

Chapter 6

Visual Search for Detailed Information

In the real world, from recognizing our friend in a crowd to finding a creamy peanut butter in the breakfast section of a grocery store, we are constantly scanning and searching the environment. Visual search is the process in which an observer actively searches the environment for a specific object among other items or features.

According to feature integration theory, visual search can be feature or conjunctive based (Treisman & Gelade, 1980). For instance, finding a red square amongst green squares is an example of a feature search that occurs pre-attentively and quickly, because one only needs to look for the color red feature. In this type of search, the target differs from distractors by a single unique feature, and that leads the target to pop out. On the other hand, detecting a red square among green squares and red circles is a conjunction search that requires the focus of attention to be directed to the target. In a conjunctive search, there is no single distinctive feature that discriminates the target, rather people have to look for a conjunction of features. In this case, the visual search is more difficult and slower because the focus of attention needs to be deployed to each item serially. Furthermore, unlike feature search, as the number of distractors increases, the search time increases as well (Treisman & Gelade, 1980).

In another proposed theory by Wolfe et al., (1989) named guided search model, pre-attentive processing of items guides the focus of attention to be directed to the target. In this regard, the attention map integrates the activation generated by the stimulus at the pre-attentive stage with the activation of top-down processes driven by the observer's goals to direct the focus of attention. In this framework, items are ranked based on their attentional priority. For example, if observers were asked to find red squares amongst green and red circles, the pre-attentive

processing determines that attention should be directed to red items on the screen first, and among those items they need to attend to the square shape to finally find the red square.

The above theories have been significantly influential in the visual search literature, however, in real life, people do not typically search for arbitrary shapes such as red circles and squares. Rather they search for more complex objects that are familiar to them. This search is relatively efficient and given the complexity of objects in our surroundings, it is hard to explain how attention is deployed via feature or conjunction search. In this regard, VanRullen (2009) argued that visual search within natural scenes, in which objects are familiar, occurs effortlessly through hardwired brain connectivity that results from daily experience. In other words, there are two types of binding mechanisms in the brain. One is on-demand binding similar to arbitrary conjunction search described in feature integration theory framework. And the other type of binding that people frequently use in real settings is hardwired binding, which allows processing familiar objects effortlessly. The latter resonates with the idea of the visual knowledge system containing the information of familiar objects embedded in the MLR model.

From the behavioral perspective, familiarity has been shown to affect the efficacy of visual search. For instance, Wang et al., (1994) presented participants with four different types of displays. The first display consisted of an unfamiliar target (e.g., rotated 5) and distractors (e.g., rotated 5's that were also flipped vertically). In the second condition, subjects saw a display consisting of an unfamiliar target (e.g., 5) amongst unfamiliar distractors (e.g., rotated 5's). The third condition consisted of showing people a familiar target (e.g., letter N) and distractors (e.g., Z). In all the three conditions, visual search time increased as more distractors appeared on the screen. However, in the condition in which the target was unfamiliar (e.g., flipped N) but distractors were familiar (e.g., N's) the search slope remained constant as a function of distractors' set size. Similarly, Shen and Reingold (2001) showed people a display consisting of a familiar target (e.g., letter N) and unfamiliar distractors (e.g., horizontally flipped N) and vice

versa. They found that searching for an unfamiliar target among familiar distractors was more efficient than searching for a familiar target among unfamiliar distractors.

The question that arises here is how familiarity affects the visual search efficacy? In other words, what mechanisms are involved in determining the search time for objects that are familiar to us compared to novel items that we have not seen before.

The concept of familiarity formalized by MLR suggests that familiar items are effectively compressed during visual processing. The MLR model learns neural codes of familiar items at different levels of compression. For instance, an image with 28x28x3 (2,352 dimensions) is compressed into 256 neurons at the first layer (i.e., L_1), and 4 neurons in the bottleneck. MLR simulations showed that learned items (i.e., familiar objects) are successfully compressed at all the levels, whereas novel items cannot be effectively compressed in small layers such as the bottleneck. Consistently, we showed that the only way that we are able to store novel items into memory and retrieve them later is by using the neural codes at early layers. In that sense, L_1 representations were stored in the BP and reconstructed via the skip connection to avoid further compression. On the other hand, familiar shapes were better encoded via bottleneck representations.

Using the MLR model, we were interested to know what neural codes are used in a visual search display depending on the familiarity of items. As described in previous research (Shen & Reingold, 2001; Wang et al., 1994), rotating or flipping familiar items would make it more difficult for subjects to use their visual knowledge and their sense of familiarity when processing the objects. Another related example is inverted faces that have shown to be harder to recognize than upright faces (Yin, 1969).

To explore the effect of familiarity in visual search, we presented participants with either a familiar MNIST digit (upright digit) or a novel MNIST digit (rotated 90° clockwise) among a

set of distractors. Using this method, we intended to keep visual complexity consistent across conditions, nonetheless eliminating the effect of familiarity.

However, familiarity is not the only factor determining the visual search efficacy. Discriminability, or in other words, how similar a target is to distractors is also an important element that impacts the speed at which a target is detected. For instance, as stated earlier, a pop-out search with a target that has a distinctive feature is normally completed faster than a conjunction search in which visual features are shared between the target and distractors (Treisman & Gelade, 1980). Similarly, Pashler (1987) conducted an experiment to investigate the effect of target-distractor similarity in a visual search display. He presented subjects with a display consisting of a letter 'C' as the target and letter 'G's as similar distractors to the target, and letter 'X's and 'L's as dissimilar distractors, such that the number of distractors were manipulated on each trial. Participants were asked to click on a specific key if the target (i.e., letter 'C') was present. The results showed that when the number of distractors that were perceptually similar to the target (i.e., letter 'G' in this case) increased, people were significantly slower in detecting the target. As a matter effect, less discriminability between the target and distractors made the visual search harder.

Discriminability can also be at the categorical level. For example, two different shapes of "2" s are less discriminable that a digit "2" vs. "3". Both because two shapes of the same item are more likely to be visually similar than two different objects, and also because "2" and "3" belong to different categories.

To manipulate discriminability in our experiment we conducted between versus within search conditions. That is, in the within search display target and distracts were from the same category (e.g., different exemplars of digit 2), whereas in the between search condition target and distractors were drawn from different digit categories (e.g., target is a "2" but distractors are "6", "9" and "3"). As a result, a target was categorically less discriminable in the within search than

in the between search, for in the within search subjects could not discriminate between items based on their category. In other words, in the within search condition people could not use categorical knowledge to accomplish the task because the target and distractors were categorically similar, however, between category search allowed for discriminating items based on their category without requiring to leverage on the perceptual visual details. Using this method, we aimed to compare the accuracy of finding the target template in between vs. within search with the model's simulation when information is stored from BN vs. L_1. More specifically, we were interested to exploring what neural codes (early level vs. late levels of the visual system) could better represent categorically similar/dissimilar items to accomplish a visual search task according to the behavioral data.

Visual Search Behavioral Experiment

Methods

Participants

twenty-five psychology undergraduate students at the Pennsylvania State University were recruited through a subject pool in exchange for course credits. Participants were instructed in English and they all reported normal or corrected-to-normal visual acuity.

Apparatus

Participants were positioned on a chinrest 60 cm from a 17-inch, 1024 x 768 CRT monitor. Experiment was performed in MATLAB (2012b) Psychtoolbox 3 (Kleiner et al., 2007) with a windows XP operating system. Responses were made using mouse clicks.

Stimuli and Procedure

Stimuli were drawn from MNIST dataset, excluding the category "zero" because its rotated version is the same as non-rotated. Each participant completed two blocks of experiment for within and between category search. Each block consisted of 192 trials such that half of the random trials were upright digit search (familiar) and the remaining trials were rotated digit search (novel).

At the beginning of each trial, a white fixation cross appeared at the center of a black screen (0,0,0 RGB) for 500 ms. Then, an upright or a rotated digit as the target template appeared at the location of a fixation cross for 300ms. Depending on the block of the experiment, participants observed four items located on vertices of a hypothetical square (diameter = 4.54° visual angle) with the three distractors being either from the same category as the target template (in within search block) or from different digit categories (between search block). Participants were instructed to make speeded and accurate responses by mouse clicking on the specific target template they had just seen. The visual search display items remained on the screen until a response was made. At the end of each trial, participants were given feedback on their response. Subjects' accuracy and reaction time were measured at the end of each trial. Reaction time (RT) data was analyzed by averaging the RTs across trials within each condition. No outlier was observed to be excluded from the dataset. Experimental blocks were counterbalanced across

participants, such that subjects with even subject number completed the within search block first and vice versa (Figure 33).

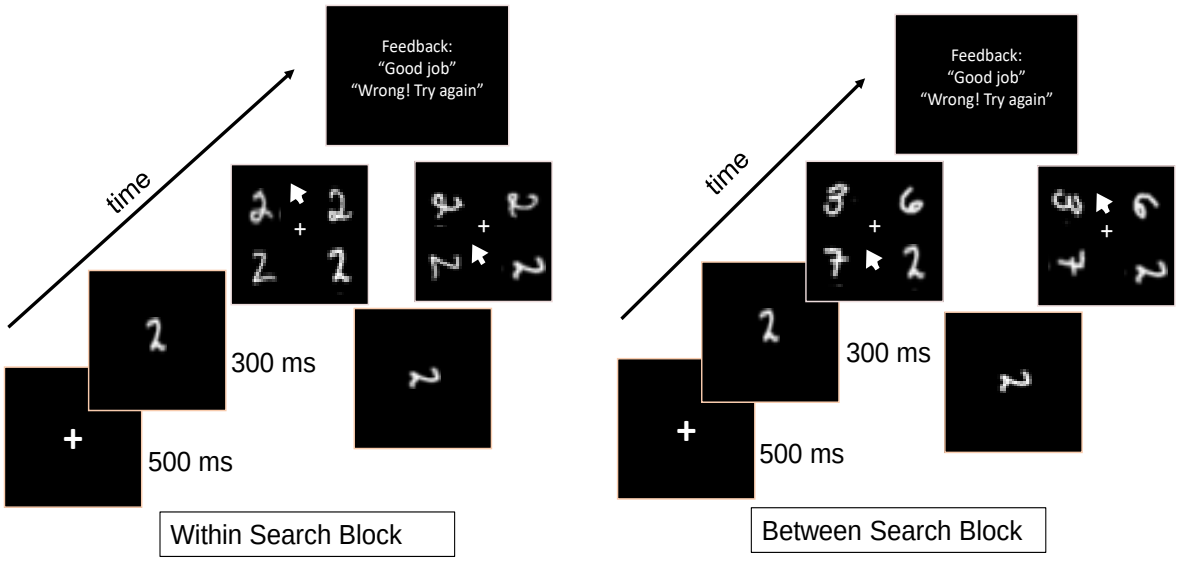

Figure 33. An example of a trial for within (left) and between search blocks. Participants were shown a fixation cross for 500 ms. A target template (an upright or rotated MNIST digit) appeared for 300 ms. Then participants were asked to find the target template among four options and click on it with the mouse. Subjects were given feedback based on the accuracy of their responses. The task instruction encouraged participants to be as accurate and as fast as possible.

Results and analysis

Accuracy and reaction time were measured across different conditions (Figure 34). A two-way analysis of variance (ANOVA) revealed that subjects in the between condition (M = .98, SD=.24) performed faster than the within condition block (M= 1.13, SD=.15), $F(1,24) = 48.96$, $p < .001$. This indicates that increased categorical discriminability improved the visual search speed. On the other hand, the main effect of familiarity was shown to be non-significant between upright (M= 1.04, SD=.20) and rotated (M=1.08, SD =.22), F (1,24) =.70, p =.40 conditions. The analysis also showed no interaction between block type (i.e., between vs. within) and stimulus type (upright vs. rotated), $F(1,24)$=.32, p =.57. Furthermore, subjects were more accurate in the

between condition (*M*=98.8, *SD*=1.9) than in the within condition (*M*=91.5, *SD*=5.3), *F*(1,24) = 82.7, $p < .01$. However, there was no main effect of stimulus type, such that people fairly performed the same in the upright condition (*M* = 95.4, *SD* =5.4) and rotated condition (*M*=95, *SD*= 5.4), *F* (1,24) = .29, *p* = .59.

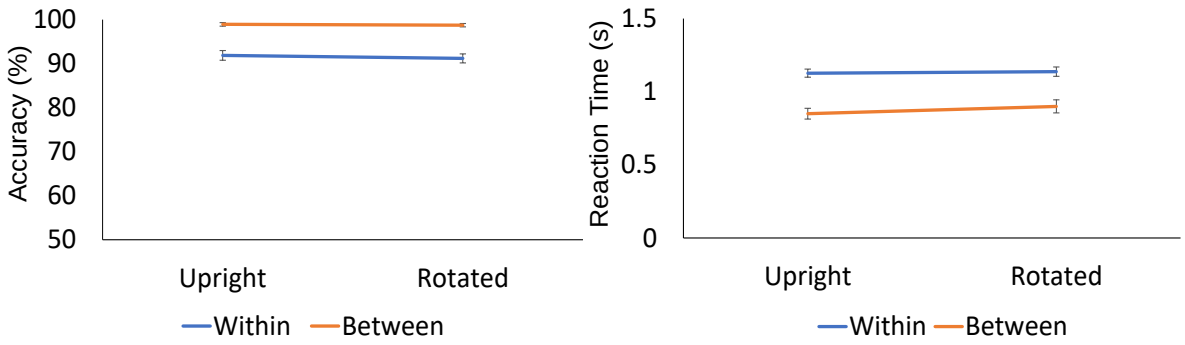

Figure 34. Mean accuracy (left panel) and reaction time (right panel) across participants for between (orange) and within (blue) conditions of upright (familiar) and rotated (novel) digits. Error bars indicate standard errors.

In general, subjects were more accurate and faster in the between condition than in the within condition. Surprisingly, familiarity did not have an effect on the visual search speed or accuracy, since people performed fairly the same in the upright vs. rotated condition.

In the case of the within condition, people were unable to distinguish the target based on its categorical code, and therefore they had to rely on visual details to find the specific shape of the target template. Similarly, familiarity did not improve the performance in terms of neither accuracy nor reaction time. Therefore, when required to remember the visual details, familiarity does not impact the visual search.

In the between search condition, the results demonstrate the effect of discriminability on visual search, but not familiarity. People were faster and more accurate in the between search condition where the target was categorically different from distractors. This could be the result of discriminability based on category, or perceptual/visual factors. In other words, when target is digit “2” subjects might just rely on the category of the target that was always different from

distractors' category, or they may have used the overall shape of the digit to distinguish between the items on the search display, since digits in different categories are also visually distinct. Interestingly, in rotated trials in which it is harder to extract the categorical information, people's performance is no different than the upright condition. That is, categorical knowledge as the result of familiarity did not improve the visual search performance. One hypothesis is that in the between search condition people mostly utilized visual discriminability than categorical discriminability.

To further explore our hypothesis and see which neural codes could possibly be used to explain the behavioral results, we carried out several simulations with the MLR model.

MLR Simulations for Visual Search

Using the MLR model that was trained on grayscale MNIST digits, we simulated the 4 conditions of between, within, upright and rotated stimuli with the two neural codes of bottleneck and L_1 (Figure 35). The model was pre-trained on upright MNIST digits. An MNIST digit as the target template (either upright or rotated) was shown to the model. Either the bottleneck (i.e., the most compressed representation) or the L_1 neural code (i.e., the least compressed representation) was stored into the BP with 1000 neurons. The bottleneck representation of the stored target template was reconstructed via the decoder, whereas the L1 representation was reconstructed via the skip connection. We already know that rotated digits could only be successfully reconstructed via L_1 neural code. However, the question is what neural code is used for the upright digits depending on the task. Once the stimulus is reconstructed, we computed the cross correlation between the retrieved target and four other images __including the target template__ that were either from the same category of the target (within search) or from different categories than target (between search). The item that had the highest cross correlation with the retrieved target

template was selected as the target by the model. Table 6 shows the mean accuracy and cross correlation with the selected target across 5 independently trained models with 500 repetitions for each condition.

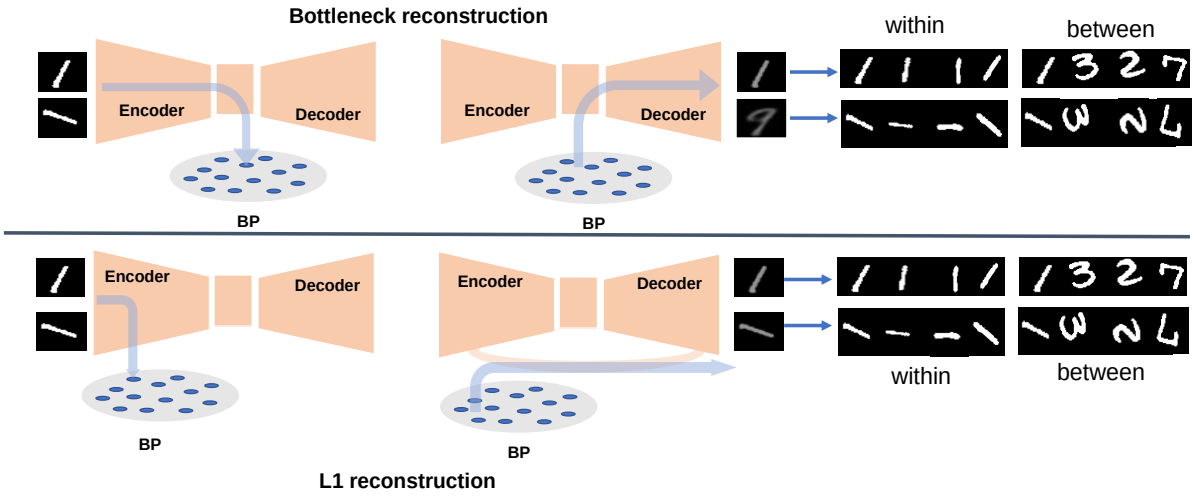

Figure 35. Visual search simulation using bottleneck (top panel) and L1 neural codes (bottom panel). Target template stored into the binding pool could be upright or rotated. The within search display consisted of items from the same category as the target (it included the target), whereas the between search display consisted digits from different categories than the target. The item that had the highest cross correlation with the retrieved target was selected as the target by the model.

Table 6. Mean accuracy and cross correlation (CC) of selected target.

	BN				L1			
	Within		Between		Within		Between	
	Upright	Rotated	Upright	Rotated	Upright	Rotated	Upright	Rotated
Acc (%)	74.6 (1.43)	50 (2.07)	95.6 (.67)	54.2 (2.47)	99.8 (.2)	99.8 (.2)	100 (0)	100 (0)
CC	.794 (.007)	.48 (.008)	.78 (.004)	.46 (.014)	.799 (.006)	.76 (.005)	.80 (.007)	.75 (.006)

Note. Mean accuracy and cross correlation (CC) of 5 independently trained models with 100 repetitions for between vs. within conditions, Bottleneck (BN) vs. L1 neural codes and familiar (upright) vs. novel stimuli (Rotated). Parenthesis indicate standard errors.

Discussion of the Results

As expected, rotated items were retrieved better using encoding L_1 activations both within and between search conditions. For the upright stimuli, accuracy was shown to be higher when items were encoded from L_1 than the bottleneck. Also, given the behavioral data, in the within search condition, upright and rotated stimuli had an accuracy of ~91%. Therefore, it is reasonable to think that both upright and rotated stimuli in the within condition were better encoded using the L_1 neural code, because they both have a similar accuracy of ~99%. Similarly, in the between search condition, according to the behavioral data the upright and rotated condition resulted in the same accuracy (i.e., ~98%). According to the model, similar to previous conditions, stimuli were better encoded and retrieved via L_1 activations. In conclusion, in all of the conditions (within, between, rotated and upright), visual search accuracy was higher when information was encoded from L_1. In other words, the model predicted that people use early level neural codes to accomplish a visual search task that consists of familiar or novel items, as L_1 neural codes provided a better discriminability. Moreover, in the case of within and between search conditions, visual details information in L_1 was shown to be critical to accurately find the target in the search display. For the within search, it is intuitive to think that early level representations are needed to discriminate subtle shape information, for all the items were selected from the same category.

With regard to the novelty effect that was not observed in the behavioral data, we hypothesized that since rotated and upright trials were randomly presented within a block, people relied on visual detail information that was better represented by L_1 than BN. Notably, even in the between search condition, where digits were categorically discriminable, it is possible that people did not utilize categorical information to accomplish the visual search task as opposed to our expectation. In other words, if people were to use categorical knowledge in the between search

task, we would have expected them to perform faster or more accurately in the upright vs. familiar trials, which was not the case. In this sense, it is possible that people did not form a compact representation in searching for familiar items, because having rotated stimuli in some trials might have forced participants to rely on early level representation that contains more visual discriminable information.

Conclusions

In this Chapter we explored how familiarity and categorical discriminability affected the visual search time. We conducted an experiment in which people were asked to find an upright or rotated MNIST digit among other digits that were categorically same or different. The behavioral results indicated that people were faster and more accurate to find a familiar or novel target from a set of categorically different stimuli (i.e., between search) than from categorically similar stimuli (i.e., within search). However, we did not find the main effect of familiarity, meaning that response time and accuracy was not different if the target was familiar or novel.

We further examined the neural substrates that may explain the results using the MLR model. The simulations indicated that regardless of familiarity or categorical discriminability, working memory stores L_1 representations that are equivalent to early levels of the visual ventral stream to accomplish the visual search task. In other words, the MLR model predicted that regardless of stimulus type (i.e., familiar or novel) L_1 representations contain more discriminable information to facilitate the visual search task.

General Conclusion

Working memory has been at the center of cognitive psychology and neuroscience research for decades. It is evident that this mental construct is inseparable from our previously learned knowledge, as it has been stated in many of the existing theories (Baddeley, 1992; Cowan, 1999). However, without an actual implementation of a working memory model that is coupled with visual knowledge, contrasting and evaluating the existing theories seems to be very challenging, if not impossible. The MLR model, constrained by computational, biological and behavioral findings, is the first attempt to shed light on the computational mechanisms of working memory in relation to visual knowledge, while highlighting the aspect of familiarity in storing and retrieving visual information.

As described in Chapters 2 and 3, the MLR meets the functional requirements of building flexible representations of familiar items, such that certain features are stored and retrieved via the BP memory. It also describes how storing more information causes interference among items, and in that sense is consistent with a limited resource theory of working memory model. Moreover, the model mechanistically demonstrates the content addressability of working memory, in which memory is retrieved based on its contents (e.g., the model is cued with a shape and can retrieve the whole item). The MLR model also describes how novel items could be stored in memory using early levels of the visual system. Storing novel items and how they are limited by memory capacity led us to formalize familiarity based on compression and categorization. We further showed that one-hot categorical labels could be stored along with the visual information causing little or no interference.

In Chapter 4, we investigated the dual code theory, and how it could improve the memory for visual details. We appended a label network to the MLR to map labels into visual details. This way, we were able to encode both visual and verbal codes into working memory, and see if that

improves the visual detail retrievals. The results suggested that when visual details are critical to remember, we may store only the visual code to accomplish a recognition task, as storing verbal information did not improve the memory.

We further extended our simulations to examine the robustness of each information type (verbal or visual) over time by noise perturbation. The results showed that visual information was more susceptible to noise than verbal code.

Chapter 5 was devoted to exploring imagery and creativity as one of the cognitive faculties that utilizes working memory representations. Using the label network that we implemented in Chapter 4, we could generate images of various shapes via a trained label network in the absence of a direct stimulus. We also explored combinatorial imagery by attempting to generate shape-color combinations that were not seen by the model. Even though the model was not able to generate new shape-color combinations by the proposed manipulations, the results of this chapter had theoretical implications on storing newly learned items in long-term memory using a VAE-based model.

In Chapter 6, we focused on visual search efficacy by manipulating categorical discriminability and familiarity in a behavioral study. We further, compared the results with MLR's simulations to see what neural codes are possibly stored to accomplish the visual search task. Interestingly, both the behavioral data and the model's simulation did not show the effect of novelty, meaning that familiar vs. novel stimuli could be stored using the same neural codes represented by the early levels of the visual system (i.e., L_1 representation in the MLR model).

We believe implementing models as such are valuable resources for generating intuitions about the formation of representations in working memory. The presented model, though limited in certain aspects (e.g., lack of representing time and space), encompassed a wide range of functional capabilities of working memory that enabled us to think more deeply about the existing theories, and how they might fit the MLR framework.

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427–1431.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249–277. https://doi.org/10.1037/0033-295X.85.4.249

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition*, *2*(3), 406–412.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Elsevier.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D. (1966). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology*, *18*(4), 302–309. https://doi.org/10.1080/14640746608400047

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Elsevier.

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(2), 176–189. https://doi.org/10.1016/S0022-5371(70)80048-2

Bahrick, H. P., & Bahrick, P. (1971). Independence of verbal and visual codes of the same stimuli. *Journal of Experimental Psychology*, *91*(2), 344.

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*(1), 1–13.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*(5), 891.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7.

Benedek, M., Jauk, E., Sommer, M., Arendasy, M., & Neubauer, A. C. (2014). Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. *Intelligence*, *46*, 73–83.

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *ArXiv Preprint ArXiv:2110.01963*.

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, *114*(1), 38.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2–3), 96–107.

Carmichael, L., Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology*, *15*(1), 73.

Cavanagh, J. P. (1973). *Holographic processes realizable in the neural realm: Prediction of short term memory performance.* [PhD Thesis]. ProQuest Information & Learning.

Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: A reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1235.

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *Quarterly Journal of Experimental Psychology*, *62*(7), 1420–1429.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K., & Alvarez, G. A. (2014). Processing multiple visual objects is limited by overlap in neural channels. *Proceedings of the National Academy of Sciences*, *111*(24), 8955–8960.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163.

Cowan, N. (1999). *An embedded-processes model of working memory.*

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Dietrich, A. (2004). The cognitive neuroscience of creativity. *Psychonomic Bulletin & Review*, *11*(6), 1011–1026.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211.

Feldhusen, J. F. (2002). Creativity: The knowledge base and children. *High Ability Studies*, *13*(2), 179–183.

Finke, R. A. (1996). Imagery, creativity, and emergent structure. *Consciousness and Cognition*, *5*(3), 381–393.

Finke, R. A., & Slayton, K. (1988). Explorations of creative visual synthesis in mental imagery. *Memory & Cognition*, *16*(3), 252–257. https://doi.org/10.3758/BF03197758

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*.

Finks, R. A., Pinker, S., & Farah, M. J. (1989). Reinterpreting visual patterns in mental imagery. *Cognitive Science*, *13*(1), 51–78.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.

Gorgoraptis, N., Catalao, R. F., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience, 31*(23), 8502–8511.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science, 16*(4), 789–802.

Hedayati, S., O'Donnell, R. E., & Wyble, B. (2022). A model of working memory for latent representations. *Nature Human Behaviour*, 1–11.

Hedayati, S., & Wyble, B. (2020). Memories of Visual Events Can Be Formed Without Specific Spatial Coordinates. *Journal of Cognition, 3*(1).

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision, 10*(2), 24–24.

Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports, 8*(1), 1–10.

Hue, C.-W., & Erickson, J. R. (1988). Short-term memory for Chinese characters and radicals. *Memory & Cognition, 16*(3), 196–205.

Hulme, C., Maughan, S., & Brown, G. D. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language, 30*(6), 685–701.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175–219.

Kanwisher, N. (1991). Repetition blindness and illusory conjunctions: Errors in binding visual types with visual tokens. *Journal of Experimental Psychology: Human Perception and Performance, 17*(2), 404.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *ArXiv Preprint ArXiv:1312.6114*.

Kirkpatrick, E. A. (1894). An experimental study of memory. *Psychological Review*, *1*(6), 602.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

Knauff, M., Kassubek, J., Mulack, T., & Greenlee, M. W. (2000). Cortical activation evoked by visual mental imagery as measured by fMRI. *Neuroreport*, *11*(18), 3957–3962.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, *33*(25), 10235–10242.

Kosslyn, S. M. (1996). *Image and brain: The resolution of the imagery debate*. MIT press.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, *2*(9), 635–642.

Kosslyn, S. M., Reiser, B. J., & Ball, T. M. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception & Performance*, 47–60.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.

Kosslyn, S. M., Thompson, W. L., Klm, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, *378*(6556), 496–498.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, *8*(4), 529–535.

Lamprecht, R., & LeDoux, J. (2004). Structural plasticity and memory. *Nature Reviews Neuroscience*, *5*(1), 45–54.

LeCun, Y. (1998). The MNIST database of handwritten digits. *Http://Yann. Lecun. Com/Exdb/Mnist/*.

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), 997–999.

Logie, R., Camos, V., & Cowan, N. (2020). *Working memory: The state of the science*.

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *275*(942), 483–519.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*(16), 5154–5167.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Morey, C. C. (2018). The case against specialized visual-spatial short-term memory. *Psychological Bulletin*, *144*(8), 849.

Mozer, M. C. (1989). Types and tokens in visual letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 287.

Ngiam, W. X., Brissenden, J. A., & Awh, E. (2019). "Memory compression" effects in visual working memory are contingent on explicit long-term memory. *Journal of Experimental Psychology: General*, *148*(8), 1373.

Norris, D., & Kalm, K. (2021). Chunking and data compression in verbal short-term memory. *Cognition*, *208*, 104534.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241–263. https://doi.org/10.1037/h0027272

Paivio, A. (1971). Imagery and language. In *Imagery* (pp. 7–32). Elsevier.

Paivio, A. (1978). A dual coding approach to perception and cognition. *Modes of Perceiving and Processing Information*, 39–51.

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255.

Paivio, A., & Csapo, K. (1969). Concrete image and verbal memory codes. *Journal of Experimental Psychology*, *80*(2p1), 279.

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, *5*(2), 176–206. https://doi.org/10.1016/0010-0285(73)90032-7

Pashler, H. (1987). Target-distractor discriminability in visual search. *Perception & Psychophysics*, *41*(4), 285–292. https://doi.org/10.3758/BF03208228

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research*, *46*(28), 4646–4674.

Pinna, B., Brelstaff, G., & Spillmann, L. (2001). Surface color from boundaries: A new 'watercolor' illusion. *Vision Research*, *41*(20), 2669–2676. https://doi.org/10.1016/S0042-6989(01)00105-5

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641.

Potter, M. C. (2018). The immediacy of conceptual processing. *On Concepts, Modules, and Language: Cognitive Science at Its Core*, *239*, 248.

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature*, *253*(5491), 437–438.

Potter, M. C., Valian, V. V., & Faulconer, B. A. (1977). Representation of a sentence and its pragmatic implications: Verbal, imagistic, or abstract? *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 1–12.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *ArXiv:2102.12092 [Cs]*. http://arxiv.org/abs/2102.12092

Reed, S. K. (1974). Structural descriptions and the limitations of visual images*. *Memory & Cognition*, *2*(2), 329–336. https://doi.org/10.3758/BF03209004

Richardson, J. T. E., & Baddeley, A. D. (1975). The effect of articulatory suppression in free recall. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 623–629. https://doi.org/10.1016/S0022-5371(75)80049-1

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759

Swan, G., Collins, J., & Wyble, B. (2016). Memory for a single object has differently variable precisions for relevant and irrelevant features. *Journal of Vision*, *16*(3), 32–32.

Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2136–2157.

Szatmáry, B., & Izhikevich, E. M. (2010). Spike-timing theory of working memory. *PLoS Computational Biology*, *6*(8), e1000879.

Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. *Nature Human Behaviour*, *3*(8), 827–836.

Thompson, V. A., & Paivio, A. (1994). Memory for Pictures and Sounds: Independence of Auditory and Visual Codes. *Canadian Journal of Experimental Psychology*, *48*(3), 380–396.

Thomson, A. M. (2010). Neocortical layer 6, a review. *Frontiers in Neuroanatomy*, *4*, 13.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

VanRullen, R. (2009). Binding hardwired versus on-demand feature conjunctions. *Visual Cognition*, *17*(1–2), 103–119.

Weldon, M. S., & Roediger, H. L. (1987). Altering retrieval demands reverses the picture superiority effect. *Memory & Cognition*, *15*(4), 269–280.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11–11.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *ArXiv Preprint ArXiv:1708.07747*.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

Yu, B., Zhang, W., Jing, Q., Peng, R., Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese and English language materials. *Memory & Cognition*, *13*(3), 202–207.

Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese words and idioms: Chunking and acoustical loop hypotheses. *Memory & Cognition*, *13*(3), 193–201.

Zimmer, H. D., & Fischer, B. (2020). Visual working memory of Chinese characters and expertise: The expert's memory advantage is based on long-term knowledge of visual word forms. *Frontiers in Psychology*, *11*, 516.

Vita

Shekoofeh Hedayati Zafarghandi

EDUCATION

2019 – 2022 Ph.D., Cognitive and Experimental Psychology
The Pennsylvania State University, University Park, PA

2017 – 2019 M.Sc. Cognitive and Experimental Psychology
The Pennsylvania State University, University Park, PA

2011 – 2016 B.Sc. Electrical Engineering, Systems and Control Engineering
The Iran University of Science and Technology, Tehran, Iran

SELECTED PUBLICATIONS

Doozandeh, P., & Hedayati, S. (2022). The effect of fidelity on transfer: A meta-analysis in the domain of troubleshooting. *IISE Transactions on Occupational Ergonomics and Human Factors*. In press.

Hedayati, S.**,** O'Donnell, R. E, Wyble, B. (2022)*:* A Model of Working Memory for Latent Representations. *Nature Human Behavior: 1-11*. Access link.

Hedayati, S., Beaty, R., Wyble, B. (2021): Finding the Building Blocks of visual imagery and creativity in a cognitively inspired model of working memory. *NeurIPS, SVRHM workshop.*

Hedayati, S., Wyble, B. (2021): How to interpret attentional blink findings? A practical MCMC tool to assess the attentional blink with the eSTST model. *Manuscript under revision at APP*. Pre-print is available at PsyArxiv.

Hedayati, S., Wyble, B. (2020) *Memories of Visual Events Can Be Formed without Specific Spatial Coordinates*. *Journal of Cognition, 3(1), 13.* Access link.

HONORS AND AWARDS

Recipient of the outstanding publication award, department of psychology, Penn State University, 2022.

Recipient of an Elsevier/Vision Research Virtual Travel Award for the Vision Sciences Society, 2021.