A Model of Working Memory for Latent Representations

Shekoofeh Hedayati $*^{1}$, Ryan E. O′Donnell[1], Brad Wyble[1]

[1]Department of Psychology, The Pennsylvania State University, State College, Pennsylvania, USA

Corresponding author:

*Email: Shokoufeh.hed@gmail.com

Abstract

We propose a mechanistic explanation of how working memories (WM) are built and reconstructed from the latent representations of visual knowledge. The proposed model features a variational autoencoder with an architecture that corresponds broadly to the human visual system and an activation-based binding pool of neurons that links latent space activities to tokenized representations. The simulation results revealed that new pictures of familiar types of items can be encoded and retrieved efficiently from higher levels of the visual hierarchy whereas, truly novel patterns are better stored using only early layers. Moreover, a given stimulus in WM can have multiple codes, representing visual detail, in addition to categorical information. Finally, we validated our model's assumptions by testing a series of predictions against behavioral results obtained from WM tasks. The model provides a demonstration of how visual knowledge yields compact visual representation for efficient memory encoding.

Introduction

In the study of cognition, working memory (WM) is thought to be responsible for temporarily holding and manipulating information to enable complex cognitive operations. Characterizing WM is an integral part of the birth of cognitive psychology, as decades of research have centered on the question of discovering the capacity and nature of this short-term memory system[1].

One of the central issues in many discussions over the structure of WM is how it is affected by previously learned knowledge[2–7]. Knowledge that emerges from long-term familiarity with particular shapes, or statistically common featural combinations enables us to recognize and remember complex objects (i.e., the prototypical shape of a car, or the strokes that comprise a digit). It is widely acknowledged that such information is crucial for building WM representations [7–9], but there has been little attempt to mechanistically implement the role of visual knowledge in WM models in spite of abundant behavioral research in this domain [10–18]. For instance, performance on immediate recall of a list of words is limited by the number of pre-learned chunks represented in long-term knowledge[11,12,19] , and readers trained to read Chinese are better able to remember Chinese characters than other readers [17].

Even prior to these findings, there has been extensive theoretical discussion of the necessity to link WM to long-term memory representations. The modal model of memory [20] proposed that representations in long-term memory could be transferred to a short-term storage. Later, the multicomponent model of WM suggested that the short-term storage of visual information (i.e., visuospatial sketchpad) is dependent on visual semantics and episodic long-term memory [21,22]. This idea is also carried by theories of activated long-term memory account [3,5,7,8,23]. In such

accounts WM representations are built by activating pre-existing representations within long-term memory.

The above accounts imply that WM is integrated with long-term knowledge, but their lack of computational specificity has made it challenging to understand this integration. To fill this gap, we implemented a computational WM model in conjunction with a visual knowledge system. This model is named Memory for Latent Representations (MLR) and it provides a new conceptualization of WM that achieves a range of functional benchmarks and forces us to formally specify our intuitions about how visual information is represented in the mind [24]. Our approach is abductive, in which a likely explanation is proposed for a set of data.

We consider the problem of WM models to exist in the M-open class (as opposed to M-closed and M-complete classes [25]), in which a true model is unattainable due to its extreme complexity but it is possible to build and test approximations that are constrained by behavior and biology to formalize our account of memory structure and function.

The proposed MLR model simulates how latent representations of items embedded in the visual knowledge hierarchy are encoded into WM depending on their familiarity. For the purposes of this work, 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).

After memoranda have been encoded in WM, they can be retrieved by reactivating those same latent representations in the visual knowledge system. Functional constraints for the model are inspired in part by previous works [7,26] and include the following capabilities.

Information about the shape of remembered objects can be regenerated [27,28].A familiar stimulus can be represented by different codes, varying from visual details up to abstract categorical information [29].Specific attributes of a given stimulus can be stored depending on their relevance for a task [30].WM performance is more efficient for familiar types of stimuli[17], but it is possible to remember novel shapes[27]. WM can store multiple items (even repeated items), each consisting of a bound combination of stimulus attributes and these can be individually retrieved according to the content of those attributes-[31].There is storage interference between stored memoranda which degrades the memory of constituent attributes according to the number of items stored in memory. [32].

Results

To build a model that exhibits these capabilities, we have created the MLR model which consists of a variational autoencoder[33] (i.e., VAE) to represent the visual knowledge hierarchy, and a binding pool[26] to store token-bound representations of the VAE's latent spaces. We modified the VAE (mVAE) to represent the color and shape distinctively in the network's compact latents and trained it on the MNIST[34] and fashion-MNIST[35] datasets using a modified version of the original VAE objective function. Figure 1 illustrates the MLR's model simplified architecture, and the correspondence of the mVAE to the visual ventral stream.

Simulation results

The mVAE disentanglement prior to memory encoding: Classification accuracies of trained support vector machines[36] (SVM) of shape and color have been summarized in Table 1 in the no-encoding condition. The results of the mean classification accuracies for 10 trained models and 10 repetitions for each model show that color and shape representations were successfully disentangled in their corresponding maps (Extended Data Figure 1). This is a coarse approximation of the general finding that the ventral visual stream has specialization of cortical maps for different types of information[37,38]. The benefit of such anatomical disentanglement in the context of a memory model like MLR is that it permits top-down modulation to easily select particular kinds of information for promotion to WM, because the control signals only need to operate on the scale of selecting regions of cortex, rather than individual neurons. The nearly complete disentanglement of color and shape as we achieve here is an exaggeration of the visual system but is helpful for demonstrating the principles of encoding attributes selectively.

The Binding Pool can encode and retrieve information: Projecting information from any given latent representation into the BP and then back to the mVAE allows us to store and reconstruct the original activity pattern of any layer in the encoder or shape/color maps. Figure 2a illustrates examples of single items encoded individually and then reconstructed using the mVAE. Table 1 indicates the classifiers' accuracies of 10 randomly generated BPs for each model across 10 separately trained 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, memory retrieval from shape and color maps is more precise than reconstructions from L_1 and L_2. Hence, compression by deeper layers allowed more accurate memory retrieval due to the relative ease of reconstructing the precise activity pattern on the smaller latent spaces of the shape and color maps. In other words, the BP encoding is lossy, particularly for layers that have more neurons, such as L_1 and L_2.

Storing multiple attributes and codes of one stimulus*:* The MLR can flexibly store specific attributes of a given stimulus such that BP representations are more efficiently allocated for a particular task[30,39]. According to Table 1, the classification accuracy of retrieving color was improved when the shape information was not stored even for a set size of one. The reverse relationship was shown also. The randomized weights between the latent spaces and the BP result in overlapping activation patterns for different attributes and therefore interference, however the impact of interference on accuracy depends on the number of BP nodes as well as the number of attributes that are being encoded. For instance, decreasing the size of the BP from 2,500 to 1000 resulted in increased interference (82% vs 76% accuracy of classifying a retrieved stimulus).

Encoding of Novel stimuli: In the preceding simulations, MLR was tested on specific MNIST images that it had not been trained on, but were from the same distribution as the training set. In this sense, they were new pictures of familiar kinds of stimuli. MLR also can store and retrieve truly novel shapes from a distribution that does not overlap with its training set (i.e., Bengali characters[40]). This is done by encoding the L_1 latent into the BP and retrieving it via the skip connection. The skip connection is critical to reconstruct novel forms, since the nature of the compressed representations in the shape and color maps force any representations that pass through those maps to resemble familiar shapes (Figure 2b).

One might ask how does the MLR model know whether to use a skip connection or the shape/color latents to store and retrieve objects. MLR can estimate the novelty of a given stimulus according to the reconstruction error, with large errors indicating novelty (see the section "Detectability of novel vs familiar shapes" in methods*)*

The accuracy of detecting a familiar item was 99.5% (*SE* = .23), whereas detecting a novel shape was 96.42% (*SE* = .58). Such novelty detection could be used to implement control mechanisms that determine which latent representations are used for memory storage although such control signals are not implemented in this version of MLR. Note that all the simulation results we reported here do not include novelty detection for the sake of simplicity.

Encoding multiple visual items*:* Tokens allow individuation of different items in memory[41,42] , by linking each token to a random subset of the binding pool as introduced in an earlier work[43]. Accordingly, tokens have overlapping memory representations, such that multiple items stored in memory interfere with one another causing a progressive degradation of memory quality as memory load increases[32] (Figure 2c).

.

Note that we are reconstructing the actual shape and specific colors of the items, not just their categorical designations and the color/shape of the retrieved items are biased toward the other items 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[44], and is also consistent with previous studies that showed misbinding of colors between stored stimuli as a form of interference with increased set size[45]. Classification accuracies of retrieved items are shown in Figure 3 condition 1.

Multiple codes for multiple objects: When appropriate for a given task, MLR can store categorical labels of information alongside the visual information in a combined memory trace[29,46–48]. By converting the output of a classifier into a localist (i.e., one-hot representation in which the estimated category has the value of 1, and other categories are set to zero) representation, a neural code of the categorical label can be stored into the BP, summing with the representations of the shape and color maps. Thus, a localist representation of category can be stored in the BP alongside the memory for the visual details of a given item within a single token just by adding the BP activation values together. This brings additional interference; however, the categorical codes are fundamentally dissimilar in character to the codes within the shape and color maps and thus do not systematically bias the memory for visual details.

To assess the accuracy of memory retrievals for visual and categorical information as a function of set size we consider five encoding conditions replicated for set sizes 1-4. All the conditions' results are summarized in Figure 3 (See Supplementary Table 1).

In 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}).

In 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. When both shape and color maps are stored as visual information in the BP along with the localist labels, the visual information was not greatly perturbed (see condition 1 vs. 2 in Figure 3) suggesting that there is little cost to remembering labels along with visual details.

In 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 localist representations estimated by the classifiers for each item when it was first classified with the labels reconstructed from the BP. Note that accuracy for remembered labels in condition 3 for larger set sizes is higher compared to condition 2. This is because the labels are akin to a digital form of encoding that can more easily be reconstructed in the presence of noise.

In condition 4 (encode 50% visual + categorical, retrieve categorical) the encoding is similar to condition 3 except that the encoding parameters for the visual maps was set at 0.5, meaning that activations of these maps were multiplied by .5 prior to encoding. This simulates prioritizing categorical information over visual details. This simulation reveals that we can parametrically adjust the relative proportion of visual details stored, producing a progressive improvement in the accuracy of retrieved labels (compared condition 4 to condition 3) at higher set sizes.

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[10].

In condition 5 (encode categorical, retrieve categorical) we simulate 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 (see Extended Data Figure 2).

BP binding and content addressability*:* Token individuation allows content addressability[31], such that 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 (Extended Data Figure 3).

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 WM is able to bind features to subtle variations in the shape of a highly familiar stimulus type for multiple stimuli (see Experiment 4 below for the human data).

More efficient storage of familiar information*:* Human memory has higher memory capacity for familiar items drawn from long-term knowledge than novel stimuli[11,13,17]. 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 shows how familiar items are stored more efficiently than unfamiliar ones, and therefore have less degradation of representations in WM as the set size increases. As shown earlier, the BP better encodes the compressed shape and color representations for familiar items (Figure 2a) because it can use the smaller shape and color maps, whereas novel types of shapes must be encoded from the larger L_1 latent and the reconstruction, then passes through the skip connection (Figure 2B). To quantify the memory performance for familiar and novel stimuli, we compared the pixelwise cross-correlation of input

and retrieved images as the function of set size, with familiar shapes being 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 Table 2 (Data visualization in Extended Data Figure 4).

The correlation value always 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 2. Note that the baseline cross-correlation between an input and the reconstructed pattern for a familiar stimulus passing through the shape and color maps is 0.85 (SD = .027) when there is no binding pool involved, therefore, the reason that the cross correlation is not closer to 1.0 for a set size of one is primarily due to the compression inherent in the mVAE, rather than memory encoding/retrieval.

The shape/color map memory retrievals of the novel shapes have correlations of .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 configurations. Therefore, the model relies on the early-level representations of L_1 to store novel

shapes, though the quality of these memories is lower than for familiar shapes using the shape and color maps.

Behavioral experiment results

In partial validation of the model, we provide 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. Figure 4 shows the summary of the experiments and the results while methods are provided at the end of the paper. All group averages are reported alongside a 95% bootstrapped confidence interval (bCI).

Experiment 1 results: 20 participants from Pennsylvania State University were shown Bengali characters and given no warning about the nature of the stimuli or that there was going to be a memory test. The stimulus was always different in trial 1 and 2. The mean accuracy on the first trial, where the target was foiled by 3 Bengali characters from different categories at 95%, bCI = [85%,100%], significantly greater than chance (25%). Participants were also highly accurate on the second trial which required them to find the target image from 3 foils of the same category, M = 90%, bCI = [75%,100%]. This supports the assumption that the pathways used to build memories of novel stimuli are always available and can be recruited on the fly with no advanced preparation.

Experiment 2 results: 20 participants were again given minimal instructions as in Experiment 1, but were now shown a single MNIST digit instead of a Bengali character. Accuracy on the first trial when the memory test was unexpected was 85%, 95% bCI = [60%, 95%], significantly

greater than chance (25%). Accuracy on subsequent trials was 85% (bCI = [70%,100%]), 90% (bCI = [75%, 100%]), 100%, and 100%. This supports our assumption that even highly familiar stimuli are encoded with memory of visual details in the absence of expectation of what specific question will be asked.

Experiment 3 results: 20 participants were led to expect that only category memory was required for report by showing them 50 trials in which they reported the categorical identity of an MNIST digit. The mean accuracy of identifying the target was 97% during these pre-surprise trials, bCI = [96%, 98.3%]. On trial 51, participants were unexpectedly asked to choose the specific visual form of the digit they saw, and accuracy dropped to 15%, bCI = [0%, 30%]. On the next trial, when participants now expected to report visual details, the accuracy of reporting the shape of the digit elevated to 100%. This difference was statistically significant according to a one-tailed permutation test, *difference* = 85%, $p < .0001$, bCI = [70%, 100%]. This demonstrates that memory encoding parameters are flexible and can be tuned to minimize visual detail information when only category is expected to be relevant. These parameters can also be rapidly modified to re-enable visual detail memory, within the span of just one trial.

Experiment 4 results: 20 Participants recruited from Prolific completed 20 trials in which they were shown two colored MNIST digits from the same category and were then asked to report the color of one digit, cued by its specific shape. Overall, participants correctly reported the target color 81.5% of the time, bCI = [75.5%, 87.25%], with swap errors (reporting the color of the other MNIST digit) occurring on average 9% of the time, bCI = [4.75%, 14%]. Importantly, 17 of 20 participants (85%) reported the correct color on trial one, bCI = [70%, 100%], which was significantly above chance (10%). This finding shows that visual details can be used for binding

and retrieval of item representations even for overtrained stimuli that belong to the same category.

Discussion

The MLR model provides a plausible account of rapid memory formation that utilizes a limited neural resource to represent visual and categorical information in an active state. The model mechanistically illustrates how WM representations could build on long-term knowledge traces to store familiar items more efficiently, while also preserving the ability to encode novel visual patterns. Using a generative model such as a VAE, we were able to build a knowledge system based on synaptic plasticity trained with gradient descent using back propagation. The VAE shares similarities with the hierarchical structure of the visual ventral stream (Figure 1) with more generic representations at the early level and more compressed representations at higher levels that can only represent familiar stimuli. In a VAE the decoder corresponds roughly to the extensive feedback projections that extend backwards down the ventral stream from higher to lower order areas[49,50].

When paired with a binding pool model of working memory[26], the MLR model was able to build generative memories of small visual images for both familiar and novel stimuli and it also exhibited the ability to tradeoff memory for visual details against memory for categorical information. Furthermore, the MLR is a cognitive model of human WM, in that it fulfills numerous requirements as proposed by Oberauer (2009)[7].

For instance, MLR can build new structural representations, which refers to the ability to quickly link or dissolve representations that bind existing representations together into novel configurations. MLR can store novel spatial arrangements of line segments (i.e., Bengali characters).

MLR can manipulate structural representations which refers to the ability to access information that is currently stored in memory and to implement cognitive operations on it. As a pure memory model MLR does not represent complex cognitive operations, but it has tunable parameters that control the flow of information to determine what specific attribute(s) or labels are encoded into WM and also allows for regeneration of the original input stimulus based on select attributes which is essential for some kinds of manipulation.

MLR has flexible reconfiguration which refers to findings that WM is a general-purpose mechanism that can be reconfigured to perform a variety of tasks. This flexibility is at the heart of MLR's mechanism for weighting which latent spaces are projected into the binding pool and can be accomplished quickly by nonspecific modulation of connection strengths along a pre-existing pathway.

MLR representations are partially decoupled from long-term memory, meaning that WM must be able to store and retrieve information in a way that is distinct from information stored in long-term memory. The binding pool exhibits this property by creating active representations that are separate from the latent spaces embedded in the visual knowledge.

MLR draws on long-term memory by building efficient memories using existing long-term knowledge representations when they are available. The BP in MLR can use the most compact latent space that is available to encode a familiar stimulus.

MLR allows for the transfer of useful information into long-term memory. It must be possible to convert or "train" WM representations into long-term memory representations. This capability is enabled by the generative aspect of MLR. Memory consolidation could occur by regenerating remembered representations and then using those to drive perceptual learning (or gradient descent in an artificial neural network). To maintain the previously learned knowledge, techniques such as interleaving previous samples or using generative replay can be used [51]. Alternatively, copies of the binding pool could serve as compressed representations to be encoded into episodic memory.

In addition to these functional requirements, we also consider the architectural benefits of the MLR which is that clustering neural activity associated with memory into a binding pool of general-purpose storage neurons provides a straightforward path for higher order processes to control memory function, allowing them to be sustained, deleted, or instantiated into constituent cortical areas. Binding information between different attributes within distinct objects is also simpler to implement in a binding pool architecture because the memories are physically clustered in a well-defined population of neurons instead of being distributed across a large expanse of sensory cortex.

MLR is not intended as a complete model of working memory as there are many functional, empirical, and computational aspects that have not yet been considered. These limitations include accounting for spatial locations, temporal effects, attention, and executive control. Their

omission is not intended to signal that they are unimportant, but rather is an admission that a formal implementation of a cognitive function so flexible as WM is beyond the scope of any single paper (see [7,23,52–55] for extensive discussion on other aspects of WM). Rather, the MLR model is intended as a nucleus of a storage mechanism to store memories in a way that is linked to visual knowledge and that is extensible to a broader range of empirical phenomena and capacities.

MLR gives us a working implementation of how memories can exploit long-term knowledge using either or both of compression and categorization. When images are drawn from MNIST or f-MNIST datasets as familiar stimuli, the visual knowledge provides a compressed representation in the shape map and also learned categorical labels derived from the shape map. In contrast, entirely novel shapes could leverage only the less compressed, generic representations at early layers to encode them into WM. Subsequently, we demonstrated that the advantage of storing compressed format of known shapes is having less interference between items compared to when early level representations of novel shapes are stored in memory. Moreover, we showed attribute binding for individual items by encoding two instances of the same digit with different colors in WM and cuing one of the shapes to retrieve the whole item. Finally, we demonstrated that the MLR could leverage the existing knowledge to detect the novelty or familiarity of a presented stimulus.

Figure 5 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 form is represented by progressively fewer neurons, but the loss of detail is minimal as the stimulus is drawn from a distribution that the

system has been trained on, or has experience with. Moreover, this visual representation can elicit a separate categorical representation that is even more compact than the visual representation, though it lacks all visual details.

We also provided empirical evidence in Experiments 1 and 2 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 that only categorical information is important for a task can diminish the memory of visual details, but this expectation can 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 WM 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.

Methods

All the experimental designs involving human participants were approved by IRB at the Pennsylvania State University. All subjects participated for course credit and acknowledged consent electronically prior to participation. In the behavioral experiments, no statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications using similar methodology[30,56]. Furthermore, in Experiments 1-2 subjects were blind to the nature of the task, and in Experiment 3 they were not aware of the surprise trial. Since each experiment consisted of only one group of subjects, no randomized assignment was performed.

The Architecture of MLR

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

mVAE: The VAE[33] is an hourglass shape fully connected neural network consisting of three main elements – the encoder, bottleneck and the decoder– which are trained by using a colorized variant of MNIST[34] and fashion-MNIST[35] stimulus sets prior to any memory storage simulations. The code for the original VAE was retrieved from a GitHub repository at: https://github.com/lyeoni/pytorch-mnist-VAE/blob/master/pytorch-mnist-VAE.ipynb.

We modified the original VAE by dividing the bottleneck into two separate maps – a color map and a shape map – to represent each feature distinctively.

Encoder: Translates information from a pixel representation into compressed latent spaces as series of transitions through lower dimensional representations.

Shape and Color maps*:* 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 encoder and the first layer of the decoder.

Decoder: Translates information from the compressed shape and color maps into pixel images through progressively higher dimensional representations. We consider the decoder to be analogous to the feedback pathways in the visual system that descend back down to primary visual cortex from deeper areas like inferotemporal cortex. Generation of a remembered stimulus at the output is not considered analogous to a motoric reconstruction but rather a reconstruction of details in an imagined visual representation.

Skip Connection*:* To allow memory reconstruction of novel stimuli without involving the shape and color maps, a skip connection was added to the mVAE that linked the first layer to the last layer. Anatomically, this would be the equivalent of a projection between layers within V1 cortex[57].

Categorical labels*:* In order to apply categorical labels to a given stimulus, we used a standard support vector machine classifier[36]. The SVM maps representations in the latent spaces onto discrete labels for different stimulus attributes such as shape or color.

Binding Pool (BP): The BP uses a modified formulation of the model described in the original binding pool paper[26] and is similar to a Holographic Reduced Representation[58]. It is a one-

dimensional matrix that is bidirectionally connected to each layer of the encoding pathway (L_1, L_2, shape and color maps) as well as the outputs of the SVM classifiers which provide one-hot or localist (i.e., the estimated category has the highest value of 1, while other categories are set to zero) representation of categorical labels of shape and color. 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 randomized, normally distributed, fixed weights. These are not trained through gradient descent but are assigned at the beginning of the simulation for a given model.

Tokens: The tokens function as object files[59,60] for each specific stimulus (e.g., token 1 stores stimulus 1). Having tokens allows multiple items to be stored within a single pool of neurons. The tokens only indicate which neurons of the BP are associated with an object representation, and do not actually store item-specific information.

The MLR implementation

Architecture: The mVAE 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= 128), 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 size of the shape and color maps were chosen to be equal for simplicity, but one can adopt optimization methods to determine the dimension of each map based on the complexity of representations. The BP layer is connected to the encoder layers

of mVAE bidirectionally (Extended Data Figure 5). Multiple tokens were connected to the binding pool nodes to individuate the items stored in memory, and there is no limit to the number of tokens one can add, although storing information in more tokens will cause increasing interference. Two layers of 20 and 10 neurons were allocated to represent the categorical information of shape and color labels estimated by the SVMss and SVMcc respectively and these were also connected to the binding pool.

Dataset: Training was done using the MNIST[34] stimulus set consisting of 70,000 images of 10 categories of digits (0-9) and fashion-MNIST[35] set, which has the same structure but for 10 categories of clothing (T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneakers, Bag and Ankle boot). To add an additional attribute to the dataset, we colorized all images that were originally grey using 10 distinct colors applied uniformly to the images – 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 uniform random variation to each of the RGB channels from the range [-.1, .1]. One triple of red-green-blue color values were generated for a given image and then multiplied by the greyscale value of that image such that all non-black pixels had the same ratio of red, green and blue color values. 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 (Extended Data Figure 6).

Training and testing the mVAE*:* The mVAE was trained on 120,000 images from MNIST and f-MNIST with 200 epochs and a batch size of 100. Three objective functions were used to train the shape and color maps and the skip connection. 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. It should be noted that with autoencoders training occurs without explicit labels or supervision, akin to how a child can learn to see through exposure to patterned information.

All three objectives to train the mVAE were derived from 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[33]). $P_\theta\ (x|z)$ 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.

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

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.

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

Shape objective function: This function converted the output images into grey scale 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.

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

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 the 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.

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

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

BP memory encoding of latents: Once the mVAE was trained, memories could be constructed by projecting information from the latent spaces into the BP which had 2500 neurons in total. The effective number of neurons representing each item was 1000 since 40% of the BP was allocated to each token. The size of the BP was determined such that it could accommodate the storage of multiple latents of the mVAE, and store multiple items, including novel stimuli. However, future works can explore optimizing the BP size, such as by encouraging sparsity.

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) or one-hot categorical labels, by a randomly generated and fixed (i.e., untrained by gradient descent), normally distributed set of weights with the mean = 0 and standard deviation of 1.0. The weights are randomly re-generated for each simulated trial. However, they remain fixed each time that the binding pool function is called.

This multiplication produces a level of activation for each neuron in the BP. Multiple attributes can be combined into one representation in the BP by summing the activation values from multiple encodings 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, which consists of ones and zeros such that a randomly selected 40% of the weights between a given token and the binding pool are set to 1, and the remainder are set to zeros.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 as modified by the task dependent encoding parameter. Summing over the binding pool nodes, we could compute the binding pool activation for all the neurons.

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

BP memory encoding of categorical labels*:* The color and shape category labels estimated by an SVM classifier, as an analog of categorical representations, could also be encoded into the BP. The shape labels were extracted from SVM_{SS} (i.e., an SVM trained to decode shape labels from the shape map) and the color labels were extracted from SVM_{CC}. (i.e., an SVM trained to decode color labels from color map). Shape was a localist (i.e., one-hot) code in a vector of

length 20 (10 digits and 10 fashion items), while color used a vector of length 10. Either or both vectors could be added to a BP representation through matrix multiplication described above. Reconstructions from the BP were converted into a one-hot vector with a max function.

One-shot encoding of novel shapes in BP: Novel shapes were 6 examples of colorized Bengali characters[40]. The colorization of Bengali characters was similar to that of MNIST and f-MNIST. The colored novel images were used as inputs to the model, and activations from L_1 and shape and color maps were encoded and retrieved from the BP to compare the efficiency of encoding from these layers. Due to the limited number of images for Bengali characters as novel shapes, we augmented the data by doing slight rotation (random from –10° to +10° rotation) and random crop with padding =8 on the 6 characters. This enabled us to do the permutations test for measuring cross-correlation.

Storing multiple items: 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. Moreover, 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 (see[26] for more details).

Memory Maintenance in BP*:* The binding pool is a simple implementation of a persistent-trace model that holds the vector of activation produced by the encoding operation(s). This is consistent with self-excitatory neural attractors, or silent synaptic storage[61]. The silent synaptic storage could be implemented by arranging small ensembles in the BP with interconnecting synapses that can store information through intracellular currents, and then reconstructing the attractor states via a trigger. The specific mechanism of trace-maintenance was not a crucial question in this implementation as there was no time course or delay of activity over time and the biophysical details of the neurons were not implemented.

Token Retrieval*:* To determine which token was linked to a cued 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 6 illustrates the retrieval activation of a given token Z_t.Other parameters are similar to that of Equation 5.

609 $Z_t = \sum_{\beta=1}^{n} B_\beta N_{t,\beta} \sum_{f=1}^{n} X_f L_{f,\beta}$ (6)

610 To test the binding accuracy, 500 digit-pairs were stored in the BP one at a time using the color
611 and shape maps and two tokens. Afterwards, a grayscale MNIST was used as a retrieval cue to
612 determine how often the model successfully retrieved the correct token based on this cue
613 (Extended Data Figure 2).

614 **Memory Reconstruction and model's evaluation***:* Memory reconstructions to any given latent
615 or one-hot (i.e., localist) vector were accomplished by retrieving the associated token and
616 multiplying the BP nodes that are linked to the corresponding token by the transpose of the same
617 fixed weight matrix that was used during the encoding of that representation. As represented by
618 Equation 7, the result is a noisy reconstruction of the original latent activity state, which can be
619 processed by feedforward activation through the rest of the mVAE . K is the normalization
620 factor that represents the sum of the active BP neurons for each item. To improve the L_1
621 reconstructions for the novel shapes, we implemented an extra transformation by increasing the
622 difference between active and inactive nodes, such that we added 2.0 to the active neurons and
623 subtracted 3.0 from nodes that had a zero activation prior to encoding in the BP . Finally, when
624 the latent L_1 received back the activations from the BP, we set the negative neurons to zero.

625 $X_f = 1/K(Z_t \sum_{\beta=1}^{n} B_\beta L_{f,\beta} N_{t,\beta})$ (7)

626 Two methods were used to evaluate the quality of memory reconstructions of MLR. 1)
627 Representations in the shape and color maps were classified by radial basis support vector
628 machines[36] (SVM), which were trained to decode shape (one of 20) or color (one of 10) using the
629 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.

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. For instance, 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. 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. Cross-correlations were normally computed over 500 repetitions.

Detectability of novel vs familiar shapes: In all the simulations presented above, the model does not decide whether the presented stimulus is familiar or novel. However, we built this mechanism into the model as a novelty detectability feature. To do this, every stimulus is reconstructed straight from the mVAE by passing through the latent space. We computed the cross-correlation between an item and its reconstruction. The model categorizes the stimulus to be familiar if the cross-correlation is above the .5 threshold. Accordingly, a given stimulus is detected to be novel if the cross-correlation is less than .5. This was repeated for 100 repetitions across the 10 trained models.

Behavioral Experiment Methods

For Experiments 1 through 3, participants were Penn State University undergraduates who participated in exchange for course credit. For Experiment 4, participants were recruited online via Prolific and compensated $1 USD for their participation in this 5-minute study. All participants provided informed consent before completing a study.

Experiment 1: 20 Penn State University undergraduates (Mean age = 19.55, 90% female, 20% left-handed) participated in this experiment. On each trial, participants were shown one randomly selected Bengali character and then asked to click on the exact character they remembered seeing from a search array of four Bengali characters. Critically participants were only instructed to pay attention and were otherwise uninformed about what would happen until after viewing the image. 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." Participants were then shown a second trial beginning with the instructions: "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." Participants were not aware a 2nd trial would occur until after they completed the first, and the presented target on trial 1 was always different from the trial 2 target.

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[62]. 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. The response screen, which consisted of the target image and 3 non-target Bengali characters was then presented to the participants. The response screen varied between trial 1 and trial 2. On the first trial, non-target answer options were selected from different Bengali character categories, and on trial 2 non-target answer options were different exemplars of the same character category. Accuracy scores were considered significantly above chance if a 95% bootstrapped confidence interval (95% bCI) did not include the chance baseline (25%).

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. Again, participants were not informed there would be a memory task. The exact instructions were as follows: “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.". Thus, the first trial served as an unexpected memory test format. 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). 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.

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 resembles that used in 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. 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 (e.g., which specific "2" among an array of four 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. The display response matched the design of Experiment 2: the target was presented alongside 3 non-target distractors selected from the target's category but with different shapes.

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 accuracy to accuracy on the 1st control trial via a one-tailed permutation test with 10,000 iterations[63]. All other parameters of this study were identical to Experiment 2.

Experiment 4: A sample of 20 participants (Mean Age 21.9, 45% female, 15% left-handed) were recruited from the online website Prolific. Participants were tasked with reporting the color

of an MNIST digit using its shape as the retrieval cue. On each trial, 2 MNIST exemplars from the same digit category were presented sequentially to the participant. 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 blank 500 ms interval 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 alternatives; chance = 10%). Unlike in previous experiments where no instruction was given, participants were explicitly instructed to remember the color-shape pairing. Participants completed 20 trials in total. Accuracy scores were considered significantly above chance if a 95% bCI did not include the chance baseline.

Data availability

The datasets for the behavioral experiments that were analyzed in this study are publicly available on the open science framework (OSF) [https://osf.io/tpzqk/]. Also, datasets that were analyzed and generated the simulations for the model can be found through the GitHub link [https://github.com/Shekoo93/MLR]

Code availability

The codes for the behavioral experiments, running the paradigm and analyzing the data are publicly available on OSF [https://osf.io/tpzqk/]. All the code for the MLR model including the simulations that generated the figures and the analysis presented in the tables are provided at the GitHub link [https://github.com/Shekoo93/MLR].

Acknowledgements

We would like to thank John Collins, Dwight Kravitz, Dimitris Pinotsis, Joyce Tam, Chloe Callahan-Flintoft and Pooyan Doozandeh for their helpful comments during the preparation of this manuscript. This work was supported by NSF grant 1734220 to B.W and Binational Science Foundation grant 2015299 to B.W. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors Contributions Statements

S.H conceptualized and wrote the paper, coded the model, and performed the simulations. B.W helped with the writing and conceptualizing, as well as writing the code. R.O coded and

performed the behavioral experiments, analyzed the behavioral data, and wrote the result section and methods of the behavioral experiments.

Competing Interests Statement

The authors declare no competing interests.

Tables

Table 1. Mean classification accuracy (%) of shape and color information of a single item stored and retrieved from memory for encodings from different latents

	Classifier type			
Encoding conditions	SVM_{SS}	SVM_{SC}	SVM_{CC}	SVM_{CS}
No encoding	84.2 (.02)	21.9 (.03)	87.2 (.04)	14.7 (.02)
Shape map and Color map	82.7 (.14)	20.9 (.14)	79.6 (.44)	14.4 (.14)
Shape map only	83.5 (.14)	21.5 (.15)	9.6 (.06)	4.5 (.08)
Color map Only	5.1 (.08)	11.4 (.33)	83.2 (.35)	14.5 (.13)
L2	74.3 (.32)	18.4 (.18)	71.3 (.65)	13.2 (.13)
L1	45.6 (.8)	13.3 (.22)	55.8 (1.1)	8.3 (.17)

The table indicate means of classifier accuracies (%) after memory retrievals of a single stimulus from each layer for 10 BPs (10 random connections matrix from BP to mVAE) for each model across 10 independently trained models. SVM_{SS} represents an SVM trained on shape labels using data from the shape map while 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}). The values in parentheses indicate standard error. Rows correspond to different encoding conditions,

showing which latent(s) were stored in the binding pool. In the top row, "No encoding", corresponds to a classification of the shape and color latents without storage into memory and this represents the theoretical maximum that the memory encoding/retrieval could obtain. 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.

Table 2. The correlation values between input and retrieval stimuli as a function of set size

	Stimuli type			
	Familiar		Novel	
Retrieval				
Set size	S/C maps	L1-skip	S/C maps	L1-skip
1	.84 (.03)	.75 (.03)	.15 (.06)	.78 (.01)
2	.72 (.05)	.66 (.04)	.14 (.06)	.65 (.03)
3	.65 (.05)	.58 (.04)	.14 (.06)	.55 (.03)
4	.58 (.06)	.53 (.04)	.14 (.06)	.48 (.04)

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 decoder 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).

Figures legends

Fig1. The simplified architecture of MLR. The model has two major elements including visual knowledge represented by mVAE and working memory shown as Binding Pool. We modified the bottleneck of a VAE to

represent shape and color in separate maps. The figure also shows the architecture of the mVAE 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 back down to V1. The inputs were either colorized version of MNIST or f-MNIST. Note that model was shown one image at a time.

Fig2. Memory retrievals from the MLR. A. Memory reconstructions from different latents in a trained model for familiar images. Selective shape or color map retrievals were achieved by setting the other map activations to zero. Note that the familiar items' reconstructions are visually less precise for memories formed from L_1 and L_2 latent spaces compared to the shape and color maps. B. Reconstruction of novel items using the L1/Skip connections and the shape/color maps. Novel shapes are reconstructed more accurately from the L_1 latent and the skip connection. Each item is stored individually in a separate BP, but the examples in A and B are combined into single images for ease of visualization. C. 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.

Fig 3. Retrieval accuracies. 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, 4 and 5. Error bars represent 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 and 5, especially when the amount of visual information stored in memory decreases. Each dot represents the accuracy of a given model over 10,000 repetitions.

Fig4. 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 stimulus and 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 50 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.

Fig5. 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.

References

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

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

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

4. Cowan, N. Short-term memory based on activated long-term memory: A review in response to Norris (2017). (2019).

5. Ericsson, K. A. & Kintsch, W. Long-term working memory. *Psychol. Rev.* **102**, 211 (1995).

6. Norris, D. Short-term memory and long-term memory are still different. *Psychol. Bull.* **143**, 992 (2017).

7. Oberauer, K. Design for a working memory. *Psychol. Learn. Motiv.* **51**, 45–100 (2009).

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

9. Brady, T. F., Konkle, T. & Alvarez, G. A. Compression in visual working memory: using statistical regularities to form more efficient memory representations. *J. Exp. Psychol. Gen.* **138**, 487 (2009).

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

11. Chen, Z. & Cowan, N. Chunk limits and length limits in immediate recall: a reconciliation. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 1235 (2005).

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

13. Ngiam, W. X., Brissenden, J. A. & Awh, E. "Memory compression" effects in visual working memory are contingent on explicit long-term memory. *J. Exp. Psychol. Gen.* **148**, 1373 (2019).

14. Ngiam, W. X., Khaw, K. L., Holcombe, A. O. & Goodbourn, P. T. Visual working memory for letters varies with familiarity but not complexity. *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 1761 (2019).

15. Yu, B. *et al.* STM capacity for Chinese and English language materials. *Mem. Cognit.* **13**, 202–207 (1985).

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

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

18. Brady, T. F., Störmer, V. S. & Alvarez, G. A. Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proc. Natl. Acad. Sci.* **113**, 7459–7464 (2016).

19. Hulme, C., Stuart, G., Brown, G. D. & Morin, C. High-and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *J. Mem. Lang.* **49**, 500–518 (2003).

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

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

22. Baddeley, A. The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* **4**, 417–423 (2000).

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

24. How Computational Modeling Can Force Theory Building in Psychological Science - Olivia Guest, Andrea E. Martin, 2021. https://journals.sagepub.com/doi/abs/10.1177/1745691620970585.

25. Clarke, J. L., Clarke, B. & Yu, C.-W. Prediction in M-complete problems with limited sample size. *Bayesian Anal.* **8**, 647–690 (2013).

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

27. Lake, B., Salakhutdinov, R., Gross, J. & Tenenbaum, J. One shot learning of simple visual concepts. in *Proceedings of the annual meeting of the cognitive science society* vol. 33 (2011).

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

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

30. Chen, H. & Wyble, B. Amnesia for object attributes: Failure to report attended information that had just reached conscious awareness. *Psychol. Sci.* **26**, 203–210 (2015).

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

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

33. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *ArXiv Prepr. ArXiv13126114* (2013).

34. LeCun, Y. The MNIST database of handwritten digits. *Httpyann Lecun Comexdbmnist* (1998).

35. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv Prepr. ArXiv170807747* (2017).

36. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

37. Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K. & Alvarez, G. A. Processing multiple visual objects is limited by overlap in neural channels. *Proc. Natl. Acad. Sci.* **111**, 8955–8960 (2014).

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

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

40. Omniglot - the online encyclopedia of writing systems and languages. https://omniglot.com/.

41. Kanwisher, N. Repetition blindness and illusory conjunctions: errors in binding visual types with visual tokens. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 404 (1991).

42. Mozer, M. C. Types and tokens in visual letter perception. *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 287 (1989).

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

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

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

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

47. Potter, M. C. The immediacy of conceptual processing. *Concepts Modul. Lang. Cogn. Sci. Its Core* **239**, 248 (2018).

48. Bae, G.-Y., Olkkonen, M., Allred, S. R. & Flombaum, J. I. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *J. Exp. Psychol. Gen.* **144**, 744–763 (2015).

49. Bullier, J. Integrated model of visual processing. *Brain Res. Rev.* **36**, 96–107 (2001).

50. Lamme, V. A., Super, H. & Spekreijse, H. Feedforward, horizontal, and feedback processing in the visual cortex. *Curr. Opin. Neurobiol.* **8**, 529–535 (1998).

51. van de Ven, G. M., Siegelmann, H. T. & Tolias, A. S. Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* **11**, 4069 (2020).

52. Barrouillet, P., Gavens, N., Vergauwe, E., Gaillard, V. & Camos, V. Working memory span development: a time-based resource-sharing model account. *Dev. Psychol.* **45**, 477 (2009).

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

54. Oberauer, K. *et al.* Benchmarks for models of short-term and working memory. *Psychol. Bull.* **144**, 885 (2018).

55. Schneegans, S. & Bays, P. M. New perspectives on binding in visual working memory. *Br. J. Psychol.* **110**, 207–244 (2019).

56. Chen, H. *et al.* Does attribute amnesia occur with the presentation of complex, meaningful stimuli? The answer is,“it depends”. *Mem. Cognit.* **47**, 1133–1144 (2019).

57. Thomson, A. M. Neocortical layer 6, a review. *Front. Neuroanat.* **4**, 13 (2010).

58. Plate, T. A. Holographic reduced representations. *IEEE Trans. Neural Netw.* **6**, 623–641 (1995).

59. Marr, D. Early processing of visual information. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **275**, 483–519 (1976).

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

61. Rose, N. S. *et al.* Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).

62. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).

63. Wilcox, R. R. *Introduction to robust estimation and hypothesis testing*. (Academic press, 2011).