Adaptation supports short-term memory in a visual change detection task

Brian Hu[1,†], Marina E. Garrett[1], Peter A. Groblewski[1], Douglas R. Ollerenshaw[1], Jiaqi Shang[1], Kate Roll[1], Sahar Manavi[1], Christof Koch[1], Shawn R. Olsen[1], Stefan Mihalas[1,†]

1 Allen Institute for Brain Science, Seattle, WA

† Corresponding authors: brianh@alleninstitute.org, stefanm@alleninstitute.org

Abstract

The maintenance of short-term memories is critical for survival in a dynamically changing world. Previous studies suggest that this memory can be stored in the form of persistent neural activity or using a synaptic mechanism, such as with short-term plasticity. Here, we compare the predictions of these two mechanisms to neural and behavioral measurements in a visual change detection task. Mice were trained to respond to changes in a repeated sequence of natural images while neural activity was recorded using two-photon calcium imaging. We trained two different models to perform the same visual change detection task. Compared to a recurrent neural network (RNN), we find that a feedforward neural network with short-term synaptic depression (STPNet) is more consistent with the pattern of behavioral responses to image changes in mice and the observed adaptation in neural responses across repeated image presentations. These results suggest that simple neural adaptation mechanisms may serve as an important bottom-up memory signal in this task, which can be used by downstream areas in the decision-making process.

Introduction

Short-term memory, or the ability to temporarily store and maintain task-relevant information, is a fundamental cognitive function. A large body of experimental and computational work suggests that this information can be maintained in persistent neural activity arising from local recurrent connections [28] or even cortical-subcortical loops [11] (for a recent review, we refer the reader to [31]). Both sustained and sequential forms of persistent activity have been observed across tasks and brain regions [26], including the prefrontal cortex (PFC) [7, 18, 24].

An alternative mechanism for the maintenance of short-term memories is via short-term synaptic facilitation, using presynaptic residual calcium as a memory buffer [2, 19]. However, recent experimental work suggests that the excitatory synapses in early sensory areas such as mouse primary visual cortex are predominantly depressing [25]. Synaptic depression can shape visual processing by providing temporal context as to whether a stimulus was recently experienced (resulting in lower postsynaptic activity) or is novel (resulting in higher postsynaptic activity). Adaptation at the synaptic level or intrinsic firing rate adaptation [10, 29] may then also play an important functional role in the brain.

Here, we study short-term memory using a visual change detection task involving a memory delay [8]. The task is to report whether a stimulus presented before the delay period is the same as the stimulus presented after the delay period. Task difficulty can be modulated by changing the duration of the delay period or the identity of the compared stimuli. In humans, the ability to detect changes in static, flashed scenes is surprisingly poor, giving rise to the phenomenon of "change blindness" [23]. In mice, previous studies using the change detection paradigm mainly presented parametric stimuli such as gratings [14]. In our study, the stimuli consist of flashed natural images (Figure 1). A hallmark of the recorded neural responses in these studies is suppression to familiar stimuli and enhancement to novel stimuli, which is consistent with findings related to mismatch negativity [12] and stimulus-specific adaptation [10]. In the

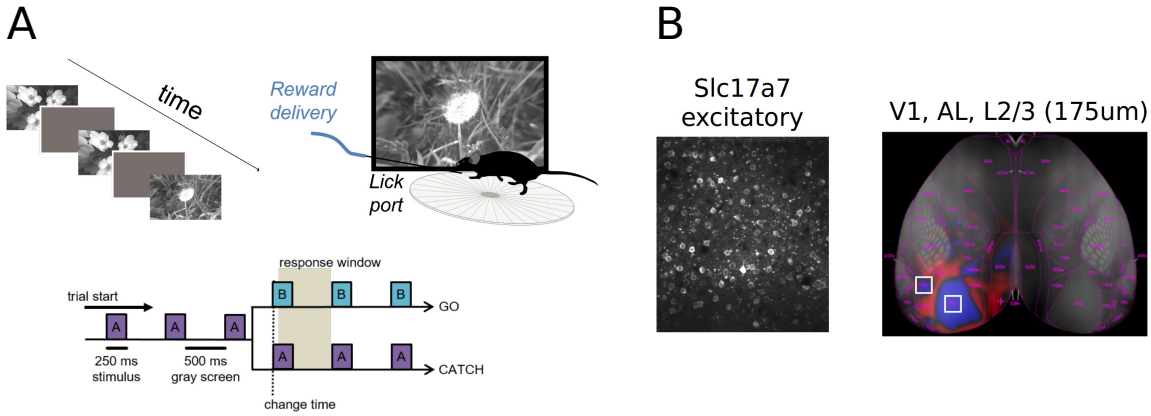

Figure 1. Overview of the visual change detection task and experiments [8]. (A) Schematic of the change detection task. Mice are trained to detect changes in a sequence of flashed natural images, giving rise to a task structure involving go and catch trials. Images were presented for 250 ms followed by a 500 ms gray screen. (B) All analyses were restricted to the pan-excitatory Cre line Slc17a7 and mouse visual areas VISp and VISal (layer 2/3).

auditory cortex, a recurrent network model with synaptic depression has been shown to explain many of the effects of stimulus-specific adaptation [30].

We use computational models to study the form of short-term memory used in the task (Figure 2). Specifically, we compare the output of a recurrent neural network in which the memory trace is stored in persistent neural activity, with a feedforward neural network with depressing synapses in which the memory trace is stored in short-term synaptic efficacies. We find that the purely feedforward model with depressing synapses can better capture the observed adaptation in neural responses across repeated stimuli, as well as the asymmetric pattern of behavioral responses to different image changes in mice. This adaptation model also makes specific predictions for neural and behavioral responses when stimuli are omitted from a repeated image sequence, which is consistent with the observed data. While there is undeniable evidence that in other tasks short-term memory can be stored in persistent neural activity, our study suggests that short-term synaptic depression can also provide neural circuits with a useful form of short-term memory on the order of seconds.

Results

Asymmetry in the detectability of natural images

We tested two types of models in the context of the change detection task, one based on persistent neural activity and one based on short-term synaptic depression (Figure 2). Input activity to the models was sparse, matching the observed responses in a large-scale survey of mouse visual cortex [4]. Both models reached similar performance as the mice in the experiments, as quantified by d-prime, hit rates, and false alarm rates (Figure 3A-C). We performed a more fine-grained quantification of behavior by computing response probability matrices, which measure the probability of responding to each image transition (Figure 3D). Off-diagonal elements of this matrix represent go trials, while diagonal elements represent catch trials. As expected, both the mice and the models respond with higher probability on go trials versus catch trials. We also observed that on average mice show a striking behavioral asymmetry in their responses, with certain image transitions being either easier or harder to detect. Here, the asymmetry is with respect to the diagonal of the response probability matrix, where the response probability for the transition from image i to image j is different than image j to image i (for example, compare the difference in response probabilities for the image transition at the top-right and bottom-left of the matrix).

We quantified this asymmetry by computing a matrix symmetry value Q for each response probability matrix (see Methods for more details). To facilitate comparison between the experiments and the models,

the response probability matrices are always ordered by the mean detectability of images across mice in the experiments. The experimental response probability matrices exhibited low matrix symmetry Q (mean=-0.07, std=0.05). We find that the STPNet model also shows an asymmetry in its responses, as characterized by a higher Pearson correlation of its response probability matrix with the experimental response probability matrix ($R = 0.79, p < 0.001$), as well as a lower matrix symmetry value Q (mean=-0.02, std=0.07). On the other hand, the RNN model showed a weaker correlation with the experimental response probability matrix ($R = 0.65, p < 0.001$) and a largely symmetric response probability matrix, quantified by a high matrix symmetry value Q (mean=0.63, std=0.05). Although both models are able to perform the the change detection task, we find that the STPNet model better matches the observed behavioral data (Figure 3E).

Adaptation as a hallmark of neural and model responses

One hallmark of the recorded neural responses is strong adaptation, with many neurons showing reduced responses to repeated image presentations (Figure 4A). We quantified this adaptation by computing a change modulation index for each neuron (see Methods for details). The change modulation index compares the mean response of the neuron on the first presentation of a change image (i.e. a go trial) with the mean response of the neuron on the corresponding pre-change image (Figure 4B-C). The change modulation index is bounded between -1 and 1, with positive values indicating higher responses to change images versus repeat images (adaptation) and negative values indicating higher responses to repeat images versus change images (facilitation). Neurons showed positive change modulation indices (mean=0.25, sem=0.01) that were significantly different than zero ($p < 0.001$, Wilcoxon signed-rank test), indicating larger responses to change images versus repeat images. We also computed change modulation indices for the input units of the two models, and found that only the STPNet model showed significant positive change modulation indices (mean=0.14, sem=0.004, $p < 0.001$, Wilcoxon signed-rank test). In contrast, the RNN model had change modulation indices that were not significantly different than zero (mean=0.0, sem=0.001, $p = 0.26$, Wilcoxon signed-rank test). The change modulation indices are slightly higher for the experiments than for the models (Figure 4C-D, note the difference in scales along the x-axis). Given that both models are able to solve the change detection task, these differences in the change modulation indices suggest different representations of short-term memory being used by the two models.

To further understand what task-relevant information is encoded during the change detection task, we performed a set of decoding analyses using the recorded populations of neurons from primary visual cortex (VISp) or one higher visual area (VISal). As a previous study did not observe major differences between these two areas [8], we also combined these areas for the analyses reported here (see Methods for details). Both image identity and image change were decodable from population activity above chance in the experiments (Figure 4E). Mean decoding accuracy was 86.7% and 79.4% for image identity and image change, respectively; chance level was 12.5% for decoding image identity and 50% for decoding image change. The number of recorded neurons per experiment ranged from 57 to 310 (mean=157, std=77). Mean decoding accuracy for image identity and image change both saturated after including approximately 100 neurons (Supplementary Figure S1). Interestingly, we observed a strong positive correlation ($R = 0.65, p < 0.001$) between the change decoder weight and the change modulation index for each neuron (Figure 4F). These results suggest that neurons which showed greater adaptation were more informative of image change, and their responses to the change image could be used to decode whether the presented image was a change image or a repeat image. We performed a similar analysis using model hidden units (which are directly connected to the output neuron), and found that both models also showed a strong positive correlation between the output weights and the change modulation indices (Supplementary Figure S2). Finally, we wanted to understand if the trained change decoders were predictive of the mice's actual behavior on a trial-to-trial basis. We quantified the agreement between each change decoder's predictions and the mouse's actual responses by computing a Jaccard score, which measures the degree of similarity between two binary vectors. Using a shuffle control, we found that the change decoder showed greater agreement with the mice's behavior than that simply expected by chance (69.1% true versus 57% shuffled, Figure 4G).

Low-dimensional analyses of neural and model activity

To better understand the structure and dynamics of population activity during the change detection task, we performed PCA dimensionality reduction on the population of recorded neurons from VISp or VISal and studied how this representation changed over the course of a trial. We found that the first two principal components captured on average 50.4% of the variance in VISp and 44.3% of the variance in VISal. We found that the representations of images cluster to different regions of the low-dimensional space (Figure 5A). We also find that repeated presentations of the same image pulls the representations closer to the origin of the PCA space (Figure 5B). We quantified this effect by calculating a distance from each image representation to the center of the PCA space (see Methods for details). We did this using either the Euclidean distance of the full population activity, or a low-dimensional view of this activity by using only a subset of the principal components. For the experiments, we observed a decrease in distance from the origin with increasing number of image repeats, which is consistent with the adaptation effect observed at the single-neuron level (Figure 5C). In many cases, the first principal component was sufficient to capture this effect, suggesting a low-dimensional correlate of image repetition. This effect was reproducible across multiple mice and visual areas (Figure 5D).

For the STPNet model, we also find that the different images map to different regions of the PCA space, and that this representation moves closer to the origin with increasing image repetitions (Figure 5E-F). In contrast, for the RNN model, we found that the representations largely cluster together based on image identity, and do not move significantly with increasing image repetitions (Figure 5J-K). The distance from the origin decreases rapidly for the STPNet model, which is similar as in the experiments, but remains relatively constant for the RNN model (compare Figure 5G,L). These findings held over the ensemble of models that were tested (Figure 5H,M). As the origin of PCA space represents the mean across all image presentations, this suggests that the representations of images becomes closer to the gray inter-stimulus interval with increasing number of stimulus repetitions. We hypothesize that the "pulling" of neural activities towards the center of the PCA space may be a useful mechanism to form a clear decision boundary which can be used to determine whether an image is novel or repeated. We note that similar results were obtained using other dimensionality reduction techniques (Supplementary Figures S3,S4).

Models make different predictions on image omissions

During the experiment, a small fraction of image presentations were omitted, effectively increasing the inter-stimulus duration between image presentations[1] (Figure 6A). The presence of stimulus omissions allows us to test the two different mechanisms of short-term memory, which make different predictions under this condition. Models based on persistent neural activity should show sustained activation during the omissions, which is what we observe in the RNN model. In contrast, both the neural data from the experiments and the STPNet model do not show significant activity during the omitted stimulus presentations (Figure 6B).

Furthermore, we find that the RNN model shows a high response probability for omitted stimulus presentations, and a lower response probability for post-omitted stimulus presentations. This suggests that the RNN model views the absence of an image as a change in its input, and responds to this change accordingly (note that all models were trained without omissions, and only tested with omissions in line with the experiments). In contrast, both mice and the STPNet model show lower response probabilities for the omitted stimulus presentations, but higher response probabilities following omission. This is consistent with a synaptic depression mechanism in which the increased inter-stimulus interval allows synapses to recover from adaptation, and effectively corresponds to a "decay" in the short-term memory of the previously seen image. The STPNet model makes the prediction then that increased inter-stimulus intervals will result in decreased model performance, specifically in the form of increased false alarm rates.

[1] In rare cases, two consecutive stimuli could be omitted, further extending the inter-stimulus duration, but we did not explicitly analyze this condition due to the low number of available trials.

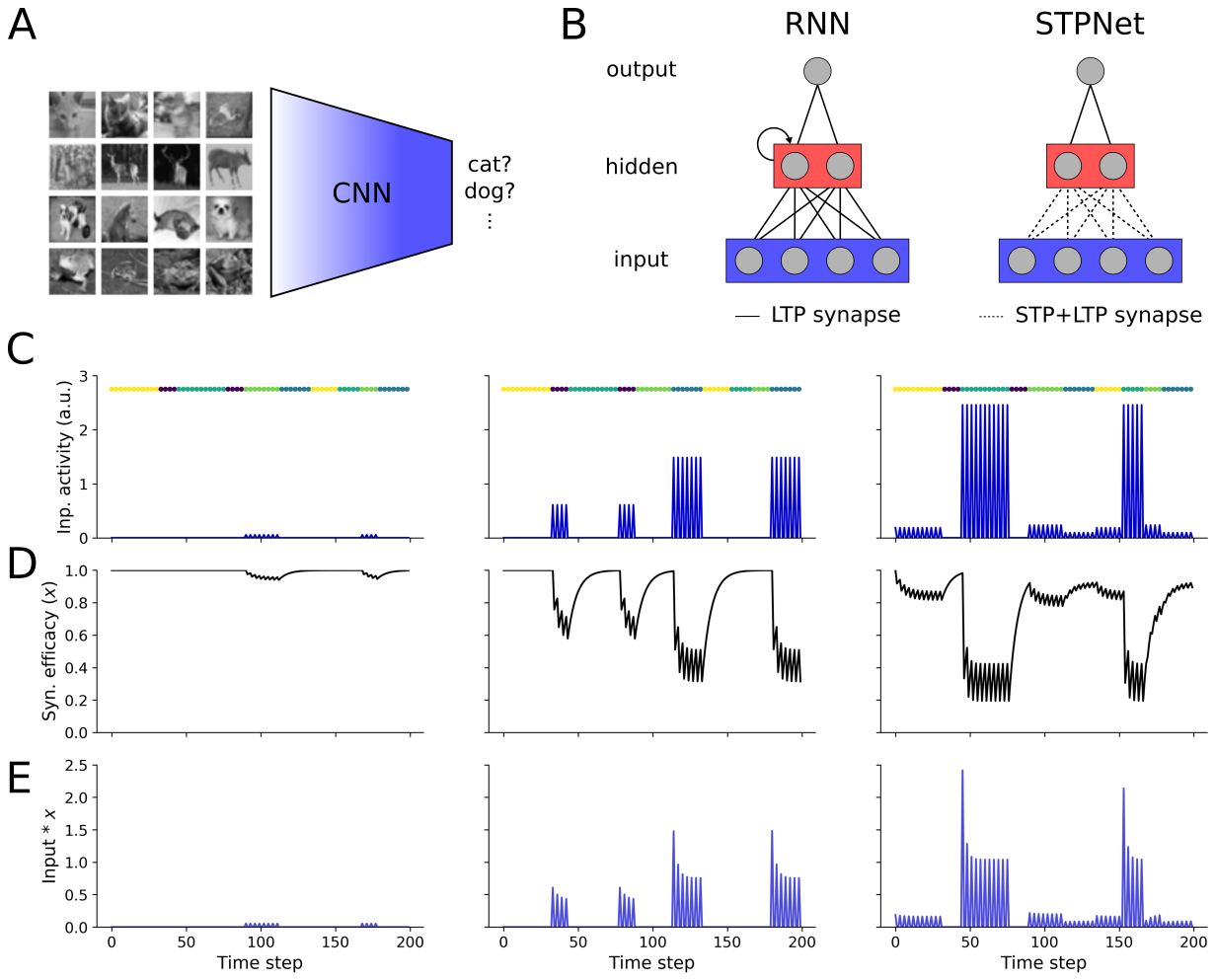

Figure 2. Overview of the models tested. (A) Image features are derived from the last fully-connected layer of a convolutional neural network (CNN) pre-trained on a grayscale version of the CIFAR-10 object recognition task. This encoder network maps an input image to a lower-dimensional feature space, which serves as input to the models described in panel B. (B) Two models of short-term memory were tested based on persistent neural activity (RNN, left) or short-term synaptic plasticity (STPNet, right). Models consisted of three layers, roughly corresponding to sensory, association, and motor areas. Model weights were trained with backpropagation (LTP synapses, solid lines), with the input synapses in STPNet also being modulated by short-term synaptic depression (STP synapses, dotted lines). (C) Input activity of three example units during the change detection task. Images were presented for 250 ms (one time step) followed by a 500 ms gray screen (two time steps). The left unit ends up responding to only one image and weakly. The center unit responds to two images, one stronger than the other. The right unit responds to four images, in a graded fashion. Image presentation times are color-coded and shown above each plot. (D) Input-dependent changes in synaptic efficacy (depression) for the units shown in panel C. (E) Input activity modulated by short-term depression for the units shown in panel C.

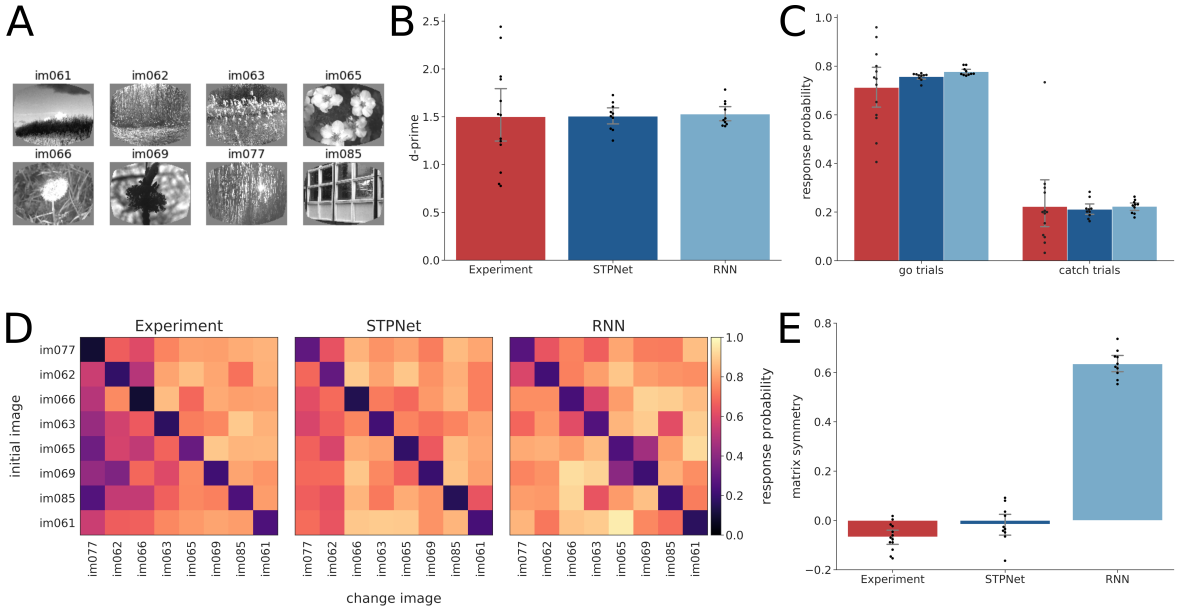

Figure 3. Quantification of behavioral responses in the experiments and the models. (A) Set of eight images that were used during the task. (B) D-prime values computed for the experiments (red) and the two models (blue). Error bars show 95% confidence intervals on the mean. The same color convention is used throughout all figures. (C) Hit rate (left) and false alarm rate (right) for the experiments and the models. Error bars show 95% confidence intervals on the mean. (D) Average response probability matrices which show the probability of responding to each of the possible 64 image transitions during the task (go trials off the diagonal, catch trials along the diagonal). The experimental response probability matrix shows an asymmetry in the detectability of images, which is partly captured by the STPNet model, but not the RNN model. (E) Quantification of response probability matrix asymmetry. Matrix symmetry is bounded between –1 and 1, with negative values indicating asymmetry and positive values indicating symmetry. The experiments and the STPNet model show a small degree of matrix asymmetry, while the RNN model shows high matrix symmetry. Error bars indicate 95% confidence intervals on the mean.

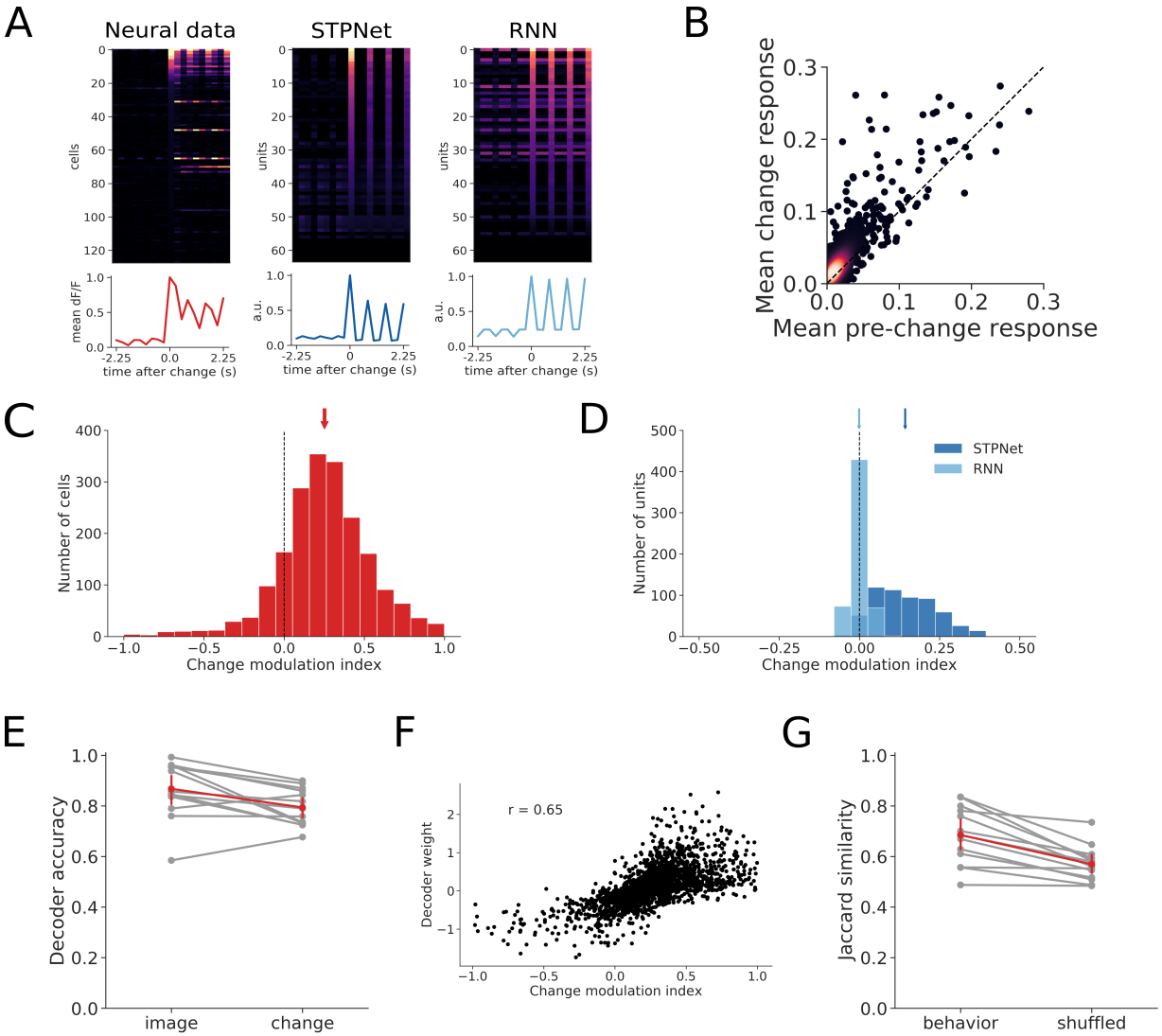

Figure 4. Quantification of change modulation in the experiments and the models. (A) Example responses centered around the change image time, for the experiments and the two models. The change image was chosen to be the preferred image for either the cell or the model unit under study. The line plots below show the population average time course. (B) A scatter plot of the mean dF/F response on change images versus the mean dF/F response on pre-change images for all neurons recorded in the experiments. Points above the unity line indicate larger mean response on change images versus pre-change images. (C) Quantification of the change modulation index (defined in Methods) in the experiment for the cells shown in panel (B). The mean change modulation index for all cells is positive (mean=0.25, sem=0.01) and significantly different than zero ($p < 0.001$, Wilcoxon signed-rank test). (D) Quantification of the change modulation index for units in the first layer of the two models. The STPNet model shows a positive mean change modulation index (mean=0.14, sem=0.004), while the RNN model shows a change modulation index close to zero (mean=0.0, sem=0.001). The change modulation index is significantly different than zero for the STPNet model ($p < 0.001$), but not the RNN model ($p = 0.26$). (E) Mean three-fold cross-validated decoding accuracy for image identity and image change across sessions ($N = 13$). Both image identity and image change were decodable above chance from population activity. Chance accuracy is 12.5% for image identity and 50% for image change. (F) Scatter plot of average decoder weights and change modulation indices for the neurons recorded in the experiments. A strong positive correlation ($r = 0.65$) is evident, suggesting that neurons which adapt are informative for change detection. (G) Jaccard similarity between decoder predictions and actual mouse behavior. The decoder predicts mouse behavior better than chance level, as determined by a shuffle control.

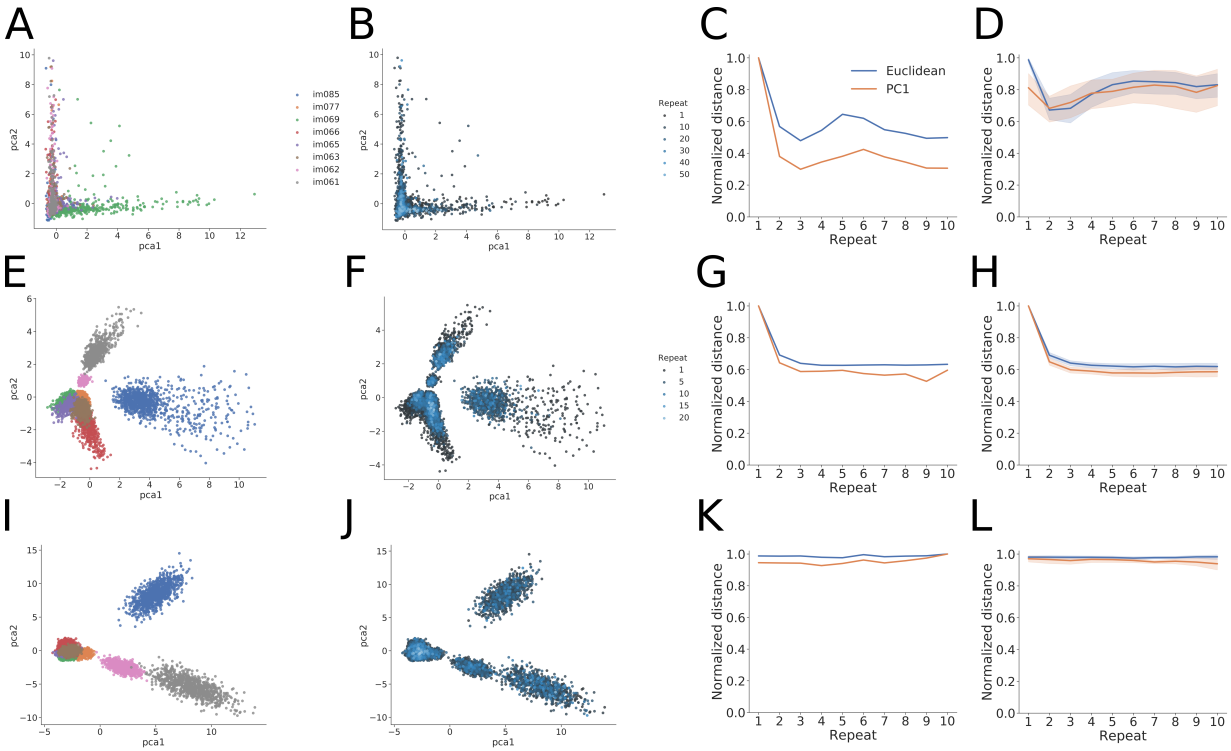

Figure 5. Low-dimensional analyses of population activity for an example experiment and trained examples of the two computational models. (A) PCA analysis showing projection of population activity onto the first two PCs. Colors label the identity of the shown image. The different images can be well separated in this low-dimensional space. (B) The same projection as in panel (A), with the color now showing the number of image repeats (more blue hue indicates higher repeats). The experiments show a "collapse" toward the origin of the low-dimensional space with increasing number of repeats. (C) Quantification of the population-level adaptation effect as a function of number of repeats. The distance from the origin of the low-dimensional space is plotted using either using the Euclidean distance (blue) or the absolute value of the first PC (orange). The experiment shows decreasing distances with repeats, and the fact that the first PC tracks the Euclidean distance indicates a low-dimensional effect. (D) Same convention as in panel (C), now showing the measured distances over all experimental sessions ($N = 13$). The mean and 95% confidence intervals are shown for each distance plot. (E-H). Same as (A-D), but for the STPNet model, which shows similar trends as for the experiments. (I-L) Same as (A-D), but for the RNN model. The RNN model shows differences with the experiments and STPNet model, mainly a constant distance as a function of image repeats.

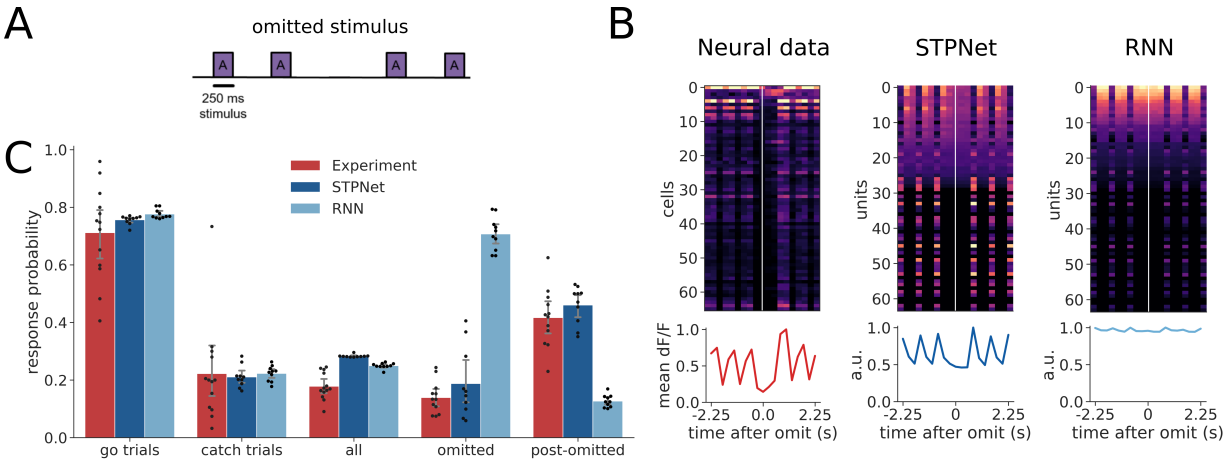

Figure 6. Quantification of response probabilities in the experiments and models. (A) Schematic of omitted stimulus presentations during the experiments. A small percentage (5%) of stimulus presentations were omitted during the experiment. (B) Example responses centered around the omitted stimulus presentation time, for the experiments and the two computational models. Cells in the experiments and STPNet did not respond during the omission, while many cells in the RNN model were active during the omission. The line plots below show the population average time course. (C) Average response probabilities for go trials, catch trials, all stimulus presentations, omitted stimulus presentations, and post-omitted stimulus presentations. Response probabilities on go trials, catch trials, and all stimulus presentations are similar across the experiments and models. Both the experiments and STPNet model show an elevated response probability following omission (post-omitted stimulus presentations), while the RNN model shows a high response probability on the omitted stimulus presentations. Error bars indicate 95% confidence intervals.

Discussion

Understanding the neural mechanisms of short-term memory

Our models suggest that short-term synaptic depression may also serve as a substrate for short-term memory in neural circuits of early sensory cortex, and it is the main memory source in the described change detection task. Multiple lines of evidence support our proposed model with depression (STPNet)– the asymmetric pattern of behavioral responses for mice trained on the change detection task, the adapting profile of neural responses across image repetitions, and the pattern of responses to omitted image presentations. Image repetition causes the input synapses in our model to adapt, reducing information about image identity. Upon presentation of a change image, a new set of input units are activated (release from inhibition). Images with more salient features (e.g. low-frequency, high-contrast edges) will strongly drive the input units in our model, while less salient images will only drive the input units weakly. The level of input unit activity influences the responses of hidden units, which is then used to directly decode image change. As a result, synaptic depression acts as a temporal filter which enables the comparison of repeated and novel stimuli, with stimuli which are different from those encountered in the recent past showing higher activity than for repeated stimuli.

In contrast, while a model which stores short-term memory in persistent neural activity was also able to solve the task, the representation used by the model was much less consistent with the constraints given by the observed data. Over all the models that we trained and tested, we empirically found that recurrent neural networks tend to show more symmetric responses. Understanding why these networks show symmetric responses is left for future work. Although we cannot rule out the contribution of persistent neural activity in other cortical areas, our results suggest that in early sensory cortex with the change detection task used here, a mechanism as simple as depressing synapses on the neurons representing the sensory input may be sufficient to capture the neural dynamics associated with behavior. Importantly, future work will require the use of causal optogenetic perturbations to test and confirm the results of our experimental and modeling analyses.

Comparison to other models

A large body of literature suggests that both persistent neural activity and short-term synaptic plasticity can be used in tasks that involve short-term memory. Attractor-based models suggest that memories can be maintained in persistent neural activity across a recurrently-connected population of neurons. More recently, many studies have shown that recurrent neural networks trained in a task-based manner can reproduce experimentally observed neural dynamics [3]. A related line of research shows that sequential neural activity, where neurons within a population are activated at different time points can also be used to store short-term memories [9, 13, 20]. The sustained or sequential nature of the persistent neural code used can vary with the task and other factors [21].

Related to our study, it has been shown that recurrent neural networks with synapses modulated by short-term synaptic plasticity can use both persistent neural activity and synaptic efficacies to store short-term memories [17]. Furthermore, these authors found that models were more likely to use persistent neural activity in cases where the short-term memory information has to be manipulated rather than simply maintained, e.g. during a delayed match-to-rotated sample task. Our work differs in that we show that synaptic depression at the input synapses alone can be also used as a bottom-up memory signal (see [30] for a similar computational modeling study in auditory cortex). Importantly, we note that the short-term synaptic depression model we used here is presynaptic and non-Hebbian, corresponding to changes which happen at a much faster timescale than typically observed during learning.

In the experiments, neurons showed varying levels of adaptation (some neurons did not adapt) while in the models with persistent activity, the majority of units showed sustained responses. Interestingly, in the context of predictive coding, a two-channel code has been proposed, with one channel showing transient responses that exhibit a similar adaptation profile in neurons across repeated stimuli, and the other channel showing sustained responses in a smaller subset of the population across all image presentations [14]. It remains to be seen whether these types of models can be reconciled with our current findings.

Model predictions

Our model makes several predictions which can be tested experimentally. First, the timescale of the short-term memory is intimately linked with the timescale of short-term depression (i.e. the recovery time constant parameter τ_x in our model). Short-term depression operates on the order of milliseconds to seconds, and we predict that in tasks with delay periods which exceed this timescale, the short-term memory will be lost as the synaptic efficacies are released from depression and return to baseline.

Changing the length of the delay period during the task will allow us to estimate the timescale of synaptic depression and determine whether this is a fixed intrinsic property or can be adapted to the task at hand. Our model also predicts that interrupting the presentation of one or more of the repeated stimuli prior to the change image will also have an impact on behavioral performance. The effect of this manipulation will be to allow the synapses to recover, resulting in an inability to distinguish the repeat image from the change image. As a result of this manipulation, we also predict a higher false alarm rate on "catch" trials, as the memory of the repeated image has been altered. However, when pushing the system to a condition in which adaptation is no longer a useful cue, we believe that the animals will employ a different strategy, potentially using persistent neural activity to solve the task.

Scope and limitations of the model

In our current model, we showed that short-term synaptic depression was sufficient to solve the change detection task. We did not consider the role of short-term facilitation, which may also play an important role in memory maintenance. On other memory tasks, the interaction of short-term facilitation and depression may be needed to account for more complex neural dynamics and behavior. Our current models also propose a dichotomy between persistent neural activity and short-term synaptic plasticity mechanisms, which is unlikely to be true in the brain. Future work will combine these two types of memory mechanisms and study their interaction for more complex memory tasks, such as category change detection tasks [5, 6]. We have also not accounted for other important dimensions of the data which may contribute to how the mice solves the change detection task. In our current study, our model only considers the responses of excitatory neurons, but understanding the contributions of different inhibitory cell types and the emergence of temporal expectation signals with experience remains an exciting area of future study [8].

Our work also has the potential to inform new model architectures in the field of artificial intelligence, where the neural networks currently used employ static synapses that have no history or temporal dependence. We believe that incorporating short-term synaptic plasticity may allow these models to be more robust and have the capability to store short-term memories.

Methods and Materials

Visual change detection task

A cohort of mice were trained to perform a visual change detection task involving natural images. Training was performed in a standardized manner, using an automated protocol in which mice progressed through a series of training stages involving static gratings, then flashed gratings, and finally natural images [8]. Stimuli were spherically warped to account for the variable distance from the eye to the periphery of the display monitor (all figures show the unwarped stimuli for ease of visualization). A set of eight natural images ("image set A") was presented, with all images being drawn from a larger set of images previously used as part of the Allen Brain Observatory [4]. Natural images were presented in grayscale, contrast normalized, and matched to have equal mean luminance. The mice were trained on image set A, and had to respond by licking whenever the image within a repeated sequence of images changed (Figure 1). Each image was presented for 250 ms, followed by a blank inter-stimulus interval of mean luminance gray for 500 ms. A small fraction (5%) of image presentations were chosen to be omitted, with omissions never preceding an image change (go trials) or sham image change (catch trials). Importantly, these omissions were only shown during imaging sessions, but never during training. The number of repeat presentations of an image was drawn from a truncated exponential distribution, set to a minimum of 4 repeats and a

maximum of 11 repeats. Correct responses were rewarded with water, while premature licks were penalized with a "time-out" period where the trial logic timer was reset. During the task, neural activity was recorded from transgenic mice expressing GCaMP6f in excitatory neurons (Slc17a7-IRES2-Cre; CamK2-tTA; Ai93) in layer 2/3 of primary visual cortex (VISp) and one higher visual area (VISal) using two-photon calcium imaging [8]. All analyses are performed on imaging sessions with simultaneously collected behavioral and physiological data. We combined the data from VISp and VISal for all analyses, as earlier work did not observe major differences between these two areas.

Neural network models

We compared two models of short-term memory– one based on persistent neural activity (RNN), and another based on short-term synaptic plasticity, specifically synaptic depression (STPNet). For both models, we used the same feedforward neural network to embed the set of natural images into a lower-dimensional feature space (Figure 2). To this end, we first trained a convolutional neural network with two convolutional layers and two fully-connected layers on a grayscale version of the CIFAR-10 dataset [16]. As we were not concerned with state-of-the-art performance on the image classification task, we did not perform an extensive hyperparameter search and only trained this encoder network for ten epochs using stochastic gradient descent with a learning rate of 0.01 and a momentum value of 0.9. After training, we froze the weights of this network, and used the output of the last fully-connected layer before the classifier as input to both the RNN and STPNet models. Although the learned weights in our models can be both positive and negative, our models are meant to only capture the responses of excitatory neurons in the experiments, and here we do not explicitly consider the role of inhibitory neurons. When applied to the change detection task, input images were first resized to 32x32 pixels. To ensure non-negative unit activations within the encoder network, we used the rectified linear nonlinearity at the output of each layer of the network. Furthermore, to make our model more biologically plausible, we also introduced independent multiplicative Gaussian noise (mean=0, std=0.5) to the output of the encoder network and to the hidden units to make the responses non-deterministic. All models were trained and simulated using the Pytorch neural network framework (v. 1.1.0) [22].

Short-term synaptic plasticity

For the model with short-term synaptic plasticity (STPNet), we included dynamics of short-term depression at the synapses from the output of the encoder network (Figure 2). We emphasize that this plasticity is presynaptic and non-Hebbian, but can interact with the weight updates from backpropagation during neural network training, similar to the idea of "fast weights" from machine learning [1]. We modeled short-term depression based on the model proposed by Tsodyks and Markram [27], but instead of using spikes, we used a firing rate approximation in terms of the activations of units within the network:

$$\frac{dx(t)}{dt} = \frac{1 - x(t)}{\tau_x} - Ux(t)r(t) \tag{1}$$

where $x(t)$ represents the fraction of neurotransmitter available, U represents the calcium concentration in the presynaptic terminal, $r(t)$ is the presynaptic activity at time t and τ_x is the recovery time constant of the synapse. The total amount of input at the postsynaptic unit at time t becomes:

$$I(t) = Wx(t)r(t) \tag{2}$$

where W is the synaptic efficacy of the synapse without short-term synaptic plasticity. For simplicity, all synapses from the same presynaptic unit were assumed to be modulated similarly. Unless otherwise noted, we fix U to be 0.5 with a τ_x of three seconds, corresponding to a regime where short-term depression dominates [25]. The exact choice of U and τ_x did not qualitatively change our results. We discretized time with a time step Δt of 250 ms, corresponding to the duration of each stimulus presentation during the experiments. We solved for Equation 1 using the implicit backward Euler method.

Neural network training

To model the mice's behavioral responses, we treat the change detection task as a binary classification problem where the network has to determine whether the presented image is a change image (represented by the label 1) or a repeat image (represented by the label 0). In the case of the RNN model, the network was a basic recurrent neural network with a hidden layer of sixteen neurons connected to a single output neuron representing the decision variable. In the case of the STPNet model, the network was a two-layer neural network with a single hidden layer of sixteen neurons connected to a single output neuron. To prevent the model from being over-confident in its predictions and to capture the inherent behavioral variability in mice, we sampled the model's responses from a Bernoulli distribution based on the sigmoid of its output, which was bounded between 0 and 1. The models were trained using sequences of images drawn from the same set of images and distribution of change times as those used in the experiments. The models were trained using stochastic gradient descent and the ADAM optimizer [15] with the default settings (learning rate of 0.001 and beta values of 0.9, 0.999), using a binary cross entropy loss with a positive weight of 5. We trained a total of 10 networks for each type of model with different random seeds. All reported results, unless otherwise noted, are averages over the trained networks. We masked the loss during inter-stimulus intervals (so that only model responses on image presentations contributed to the loss), and we also used an additional L2 penalty with a weight of 0.001 on the activations of hidden units to promote sparsity. The models were trained for a maximum of 5000 epochs, with a d-prime early stop criterion of 1.5 (based on the hit and false alarm rates observed in mice). We used a patience of 5, meaning the d-prime value had to be above the d-prime early stop criterion for at least 5 epochs before training stopped. The exact choice of model hyperparameters did not qualitatively change our results.

Analyses of experiment and model data

Quantifying response asymmetry

We quantified behavioral performance using the response probability matrix, which captures the pattern of responses to each pair of image transitions (Figure 3C). In this matrix, catch trials (trials where no image change occurred) are along the diagonal and go trials (trials where an image change occurred) are off the diagonal. For an image set of eight natural images, there are a total of 64 possible image transitions. The response probability matrix was computed for both the experiments and the models. We quantified similarity between the model and mice response probability matrices using the Pearson correlation coefficient. In addition, we quantified the degree of symmetry in the response probability matrices using the following metric Q:

$$Q = \frac{\| M_{sym} \| - \| M_{anti-sym} \|}{\| M_{sym} \| + \| M_{anti-sym} \|} \tag{3}$$

$$M_{sym} = 1/2\ (M + M^T) \tag{4}$$

$$M_{anti-sym} = 1/2\ (M - M^T) \tag{5}$$

where M is the mean-subtracted response probability matrix (ignoring the elements along the diagonal), M^T is the transpose of this matrix, M_{sym} and $M_{anti-sym}$ are the symmetric and anti-symmetric components of the matrix M, respectively, and $\| . \|$ represents the Frobenius norm. This metric is bounded between -1 and 1, with more symmetric matrices having values closer to 1 and more anti-symmetric matrices having values closer to -1.

Change modulation index

One possible signature of neural adaptation is stronger responses for change images compared to repeat images. To test this hypothesis, we computed the change modulation index for each neuron by comparing mean responses to change and pre-change images (i.e. the images presented on go trials and the images presented just before a change image, respectively). The change modulation index (CMI) is defined as:

$$CMI = \frac{r_{change} - r_{pre-change}}{r_{change} + r_{pre-change}} \tag{6}$$

where r_{change} is the mean response to change images and $r_{pre-change}$ is the mean response to pre-change images. The change modulation index is bounded between -1 and 1, with neurons that respond more strongly to change images having values closer to 1 (suggesting adaptation) and neurons that respond more strongly to pre-change images having values close to -1 (suggesting facilitation). In our analyses, we excluded neurons that had a negative mean dF/F response to either the change or pre-change images, which would result in change modulation indices outside the range [-1, 1]. Alternate modulation indices based on either comparing the mean response to go versus catch trials, or the mean response to change images versus repeat images a fixed number of repeats later, gave qualitatively similar results.

Decoding analyses

We also performed decoding analyses to determine what task-relevant information could be reliably decoded from the population of recorded neurons. Specifically, we trained linear support vector machines (SVMs) to decode either image identity on each image presentation (multi-class classification) or whether a given image was a change or pre-change image (binary classification). We used change and pre-change images to avoid potential problems with imbalanced data when using go and catch trials. We performed three-fold cross-validation, and report mean decoding accuracy averaged across the three folds. For each experimental session, we used all recorded neurons in our decoding analyses, only excluding neurons which had a negative mean dF/F response to either the change or pre-change images. We also tested the effect of number of neurons on mean decoding accuracy by randomly sampling a fixed number of neurons from each session ($N = 10$ cross-validated samples) and training the decoders using this smaller subset of neurons (Supplementary Figure S1). To understand whether neurons which showed greater change modulation were more important for change decoding, we computed the Pearson correlation coefficient between the weights from the change decoder and the corresponding change modulation indices for each neuron. We also performed a similar analysis on the hidden units from the two models, using the hidden unit to output unit weights as a measure of change decoding (Supplementary Figure S2). To assess whether the trained decoders were predictive of mice's behavioral responses, we used the Jaccard score to quantify the similarity between the decoder's predictions and the mice's actual choices. Importantly, while the decoders were trained on change and pre-change images, the decoders were tested on go (equivalent to change images) and catch trials. As a control, we shuffled the decoder's predictions ($N = 1000$ times) and computed the Jaccard score between the shuffled predictions and the mice's choices, which gives a measure of similarity based on random chance alone.

Dimensionality reduction

To visualize the structure of population activity during the change detection task, we also performed PCA dimensionality reduction on the activity of model units and neurons recorded in the experiments (Figure 5). We retained those principal components (PCs) which captured 95% of the variance of the neural activity, the number of which could be different for each experiment. We then computed a Euclidean distance metric using either the neural activities themselves (which is equivalent to using all PCs) or using a subset of the PCs approximating a low-dimension projection of the data (typically just the first PC). We computed this distance from the origin as a function of the number of image repeats. Similar results were obtained using linear discriminant analysis (LDA) or isomap as dimensionality reduction techniques (Supplementary Figures S3,S4).

Omitted image presentations

A small fraction (5%) of image presentations were omitted during the experiments. In addition to the response probability on go and catch trials, we quantified the response probability to these omitted images, as well as to the post-omitted images (defined as image presentations that came after the omission). For

the two models, we conditioned the responses to post-omitted images on the omitted image response, i.e. if the model responded to the omitted image, its response probability on the post-omitted image presentation was always defined to be zero. This was not required for the experiments due to existing trial logic which aborted trials on the omitted image if the mouse responded prematurely.

Acknowledgments

We wish to thank the Allen Institute for Brain Science founder, Paul G. Allen, for his vision, encouragement, and support. We would also like to thank Matthew Valley, Corbett Bennett, and Alex Piet for helpful discussions.

Competing interests

The authors declare no competing interests.

References

[1] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, pages 4331–4339, 2016.

[2] O. Barak and M. Tsodyks. Working models of working memory. *Current opinion in neurobiology*, 25:20–24, 2014.

[3] W. Chaisangmongkon, S. K. Swaminathan, D. J. Freedman, and X.-J. Wang. Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron*, 93(6):1504–1517, 2017.

[4] S. E. de Vries, J. A. Lecoq, M. A. Buice, P. A. Groblewski, G. K. Ocker, M. Oliver, D. Feng, N. Cain, P. Ledochowitsch, D. Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, 2020.

[5] D. J. Freedman and J. A. Assad. Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–88, 2006.

[6] D. J. Freedman and J. A. Assad. Neuronal mechanisms of visual categorization: an abstract view on decision making. *Annual review of neuroscience*, 39:129–147, 2016.

[7] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, 61(2):331–349, 1989.

[8] M. Garrett, S. Manavi, K. Roll, D. R. Ollerenshaw, P. A. Groblewski, N. D. Ponvert, J. T. Kiggins, L. Casal, K. Mace, A. Williford, et al. Experience shapes activity dynamics and stimulus coding of vip inhibitory cells. *eLife*, 9, 2020.

[9] M. S. Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, 2009.

[10] K. Grill-Spector, R. Henson, and A. Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006.

[11] Z. V. Guo, H. K. Inagaki, K. Daie, S. Druckmann, C. R. Gerfen, and K. Svoboda. Maintenance of persistent activity in a frontal thalamocortical loop. *Nature*, 545(7653):181, 2017.

[12] J. P. Hamm and R. Yuste. Somatostatin interneurons control a key component of mismatch negativity in mouse visual cortex. *Cell reports*, 16(3):597–604, 2016.

[13] C. D. Harvey, P. Coen, and D. W. Tank. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62, 2012.

[14] J. Homann, S. A. Koay, A. M. Glidden, D. W. Tank, and M. J. Berry. Predictive coding of novel versus familiar stimuli in the primary visual cortex. *bioRxiv*, page 197608, 2017.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[17] N. Y. Masse, G. R. Yang, H. F. Song, X.-J. Wang, and D. J. Freedman. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature neuroscience*, page 1, 2019.

[18] E. K. Miller, C. A. Erickson, and R. Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of neuroscience*, 16(16):5154–5167, 1996.

[19] G. Mongillo, O. Barak, and M. Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543–1546, 2008.

[20] J. D. Murray, A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo, and X.-J. Wang. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*, 114(2):394–399, 2017.

[21] A. E. Orhan and W. J. Ma. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nature neuroscience*, 22(2):275, 2019.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[23] R. A. Rensink, J. K. O'Regan, and J. J. Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological science*, 8(5):368–373, 1997.

[24] R. Romo, C. D. Brody, A. Hernández, and L. Lemus. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470, 1999.

[25] S. C. Seeman, L. Campagnola, P. A. Davoudian, A. Hoggarth, T. A. Hage, A. Bosma-Moody, C. A. Baker, J. H. Lee, S. Mihalas, C. Teeter, et al. Sparse recurrent excitatory connectivity in the microcircuit of the adult mouse and human cortex. *bioRxiv*, page 292706, 2018.

[26] K. K. Sreenivasan and M. D'Esposito. The what, where and how of delay activity. *Nature Reviews Neuroscience*, 20(8):466–481, 2019.

[27] M. Tsodyks, K. Pawelzik, and H. Markram. Neural networks with dynamic synapses. *Neural computation*, 10(4):821–835, 1998.

[28] X.-J. Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463, 2001.

[29] C. J. Whitmire and G. B. Stanley. Rapid sensory adaptation redux: a circuit perspective. *Neuron*, 92(2):298–315, 2016.

[30] T. S. Yarden and I. Nelken. Stimulus-specific adaptation in a recurrent network model of primary auditory cortex. *PLoS computational biology*, 13(3):e1005437, 2017.

[31] J. Zylberberg and B. W. Strowbridge. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annual review of neuroscience*, 40:603–627, 2017.

Supporting Information

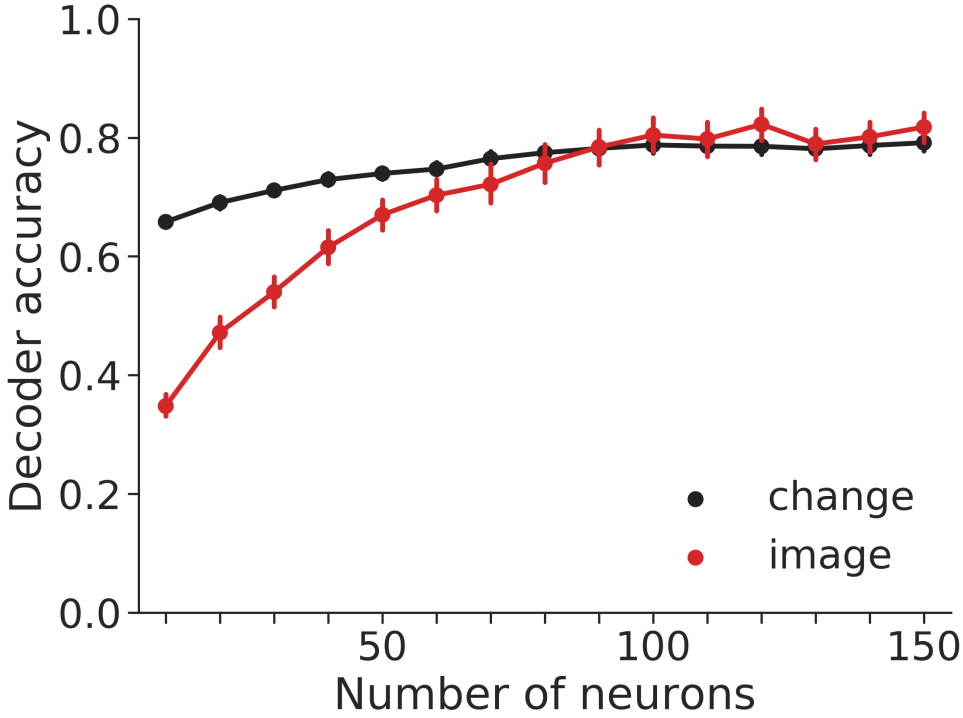

Figure S1. Decoder accuracy as a function of number of neurons. Decoder accuracy for both image change (black) and image identity (red) increases with the number of neurons included in the analysis, saturating around 100 neurons. Neurons were added in increments of 10, randomly sampled using 10 cross-validated samples per experimental session. For each number of neurons, only sessions which had enough neurons were used in the analysis. Decoder accuracy is averaged across sessions and samples. Error bars show 95% confidence intervals on the mean.

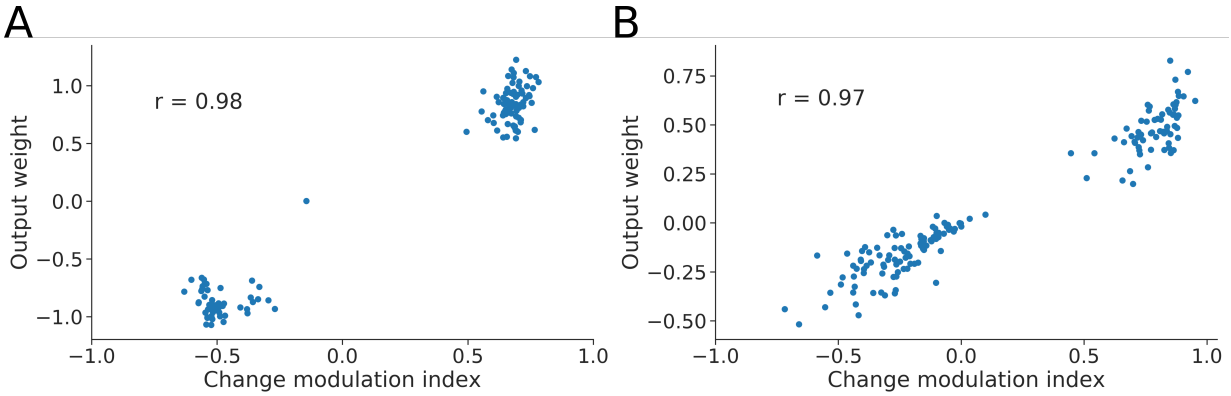

Figure S2. Comparison of model hidden unit output weights and change modulation indices. (A) Scatter plot of output weights and change modulation indices for the hidden units in the STPNet model ($N = 10$ models). Two clusters emerge, with one showing positive change modulation indices and output weights (corresponding to adaptation) and the other showing negative change modulation indices and output weights (corresponding to facilitation). A strong positive correlation ($r = 0.98$) is evident as for neural data from the experiments, suggesting that units which adapt are informative for change detection. (B) Same convention as in (A), for the RNN model. There is also a strong positive correlation ($r = 0.97$).

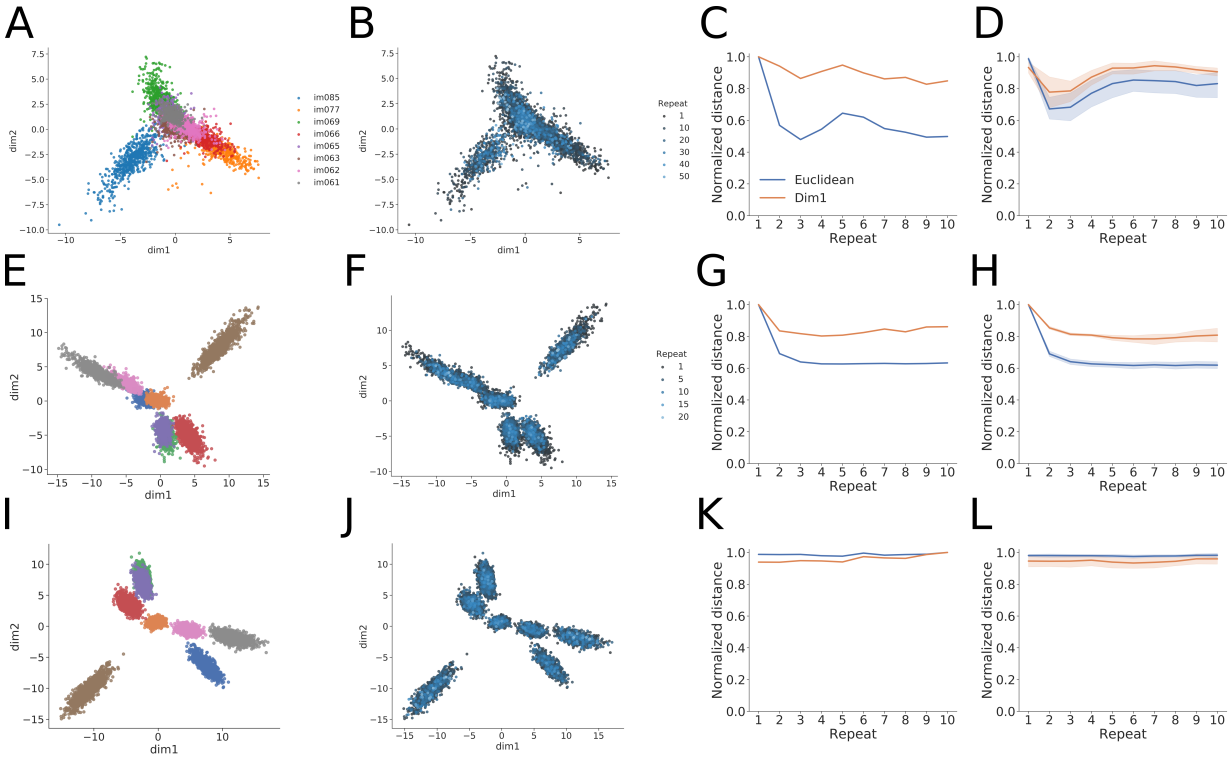

Figure S3. Low-dimensional analyses of population activity for an example experiment and trained examples of the two computational models. (A) LDA analysis showing projection of population activity onto the first two LDA dimensions. Colors label the identity of the shown image. The different images can be well separated in this low-dimensional space. (B) The same projection as in panel (A), with the color now showing the number of image repeats (more blue hue indicates higher repeats). The experiments show a "collapse" toward the origin of the low-dimensional space with increasing number of repeats. (C) Quantification of the population-level adaptation effect as a function of number of repeats. The distance from the origin of the low-dimensional space is plotted using either using the Euclidean distance (blue) or the absolute value of the first LDA dimension (orange). The experiment shows decreasing distances with repeats, and the fact that the first LDA dimension tracks the Euclidean distance indicates a low-dimensional effect. (D) Same convention as in panel (C), now showing the measured distances over all experimental sessions ($N = 13$). The mean and 95% confidence intervals are shown for each distance plot. (E-H). Same as (A-D), but for the STPNet model, which shows similar trends as for the experiments. (I-L) Same as (A-D), but for the RNN model. The RNN model shows differences with the experiments and STPNet model, mainly a constant distance as a function of image repeats.

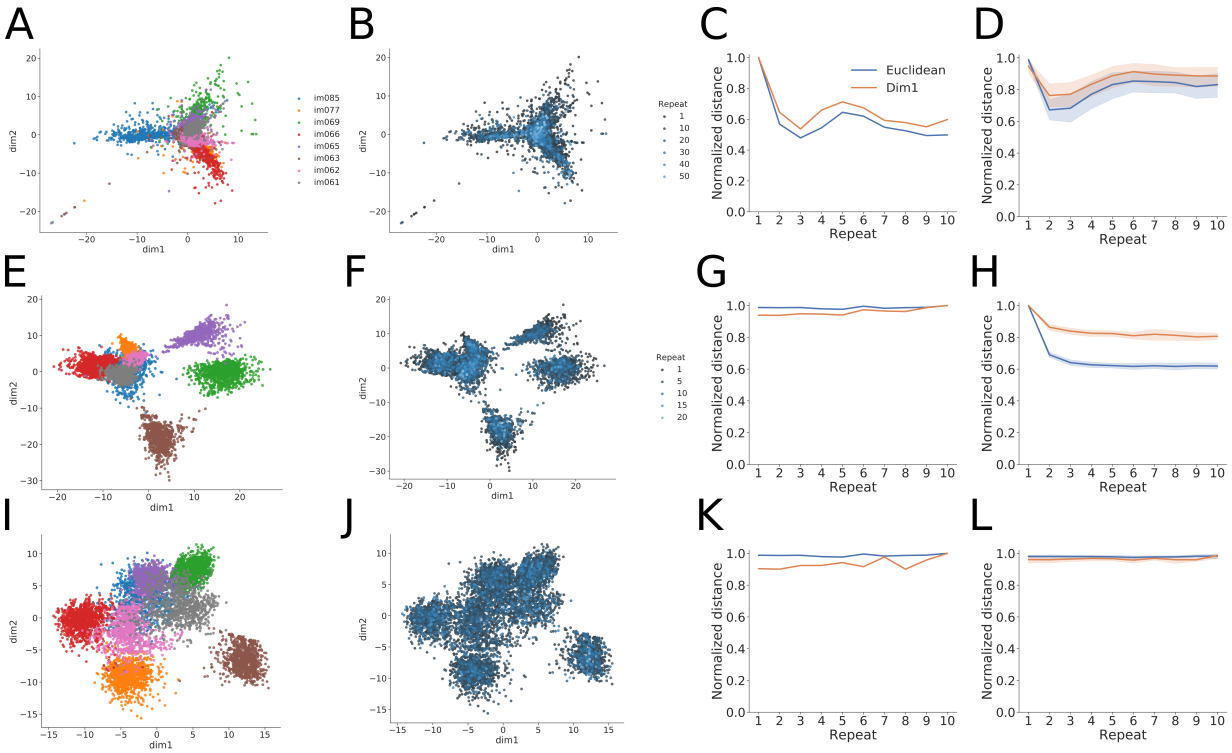

Figure S4. Low-dimensional analyses of population activity for an example experiment and trained examples of the two computational models. (A) Isomap analysis showing projection of population activity onto the first two isomap dimensions. Colors label the identity of the shown image. The different images can be well separated in this low-dimensional space. (B) The same projection as in panel (A), with the color now showing the number of image repeats (more blue hue indicates higher repeats). The experiments show a "collapse" toward the origin of the low-dimensional space with increasing number of repeats. (C) Quantification of the population-level adaptation effect as a function of number of repeats. The distance from the origin of the low-dimensional space is plotted using either using the Euclidean distance (blue) or the absolute value of the first isomap dimension (orange). The experiment shows decreasing distances with repeats, and the fact that the first isomap dimension tracks the Euclidean distance indicates a low-dimensional effect. (D) Same convention as in panel (C), now showing the measured distances over all experimental sessions ($N = 13$). The mean and 95% confidence intervals are shown for each distance plot. (E-H). Same as (A-D), but for the STPNet model, which shows similar trends as for the experiments. (I-L) Same as (A-D), but for the RNN model. The RNN model shows differences with the experiments and STPNet model, mainly a constant distance as a function of image repeats.