A neural correlate of image memorability

Andrew Jaegle*[1], Vahid Mehrpour*[1], Yalda Mohsenzadeh*[2], Travis Meyer[1], Aude Oliva[2], Nicole Rust[1]
*=Authors contributed equally. Listed in alphabetical order.
[1]Department of Psychology, University of Pennsylvania
[2]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Some images are easy to remember while others are easily forgotten. While variation in image memorability is consistent across individuals, we lack a full account of its neural correlates. By analyzing data collected from inferotemporal cortex (IT) as monkeys performed a visual memory task, we demonstrate that a simple property of the visual encoding of an image, its population response magnitude, is strongly correlated with its memorability. These results establish a novel behavioral role for the magnitude of the IT response, which lies largely orthogonal to the coding scheme that IT uses to represent object identity. To investigate the origin of IT memorability modulation, we also probed convolutional neural network models trained to categorize objects. We found brain-analogous correlates of memorability that grew in strength across the hierarchy of these networks, suggesting that this memorability correlate is likely to arise from the optimizations required for visual as opposed to mnemonic processing.

We have a remarkable ability to remember the images that we have seen, even after a single viewing [1, 2]. Although this capacity appears general and may serve a wide variety of functions, we remember some images better than others [3]. Image memorability is consistent across individuals [3, 4], however, a full account of the sources of image memorability has remained elusive. For example, while some types of natural image content are known to impact memorability – such as images with people, which tend to be more memorable than scenes [3], and abnormal objects, such as chair shaped like a hand, which tend to be more memorable than typical objects [4] – we lack a complete account of how image content determines image memorability.

What neural processes determine memorability? The sources of memorability could range from variation in the perceptual organization of images in visual cortex to the processes that support memory formation and/or memory recall. The neural correlates of memorability are likely to reside at higher stages of the visual form processing pathway, where image memorability can be decoded from human fMRI activity patterns [5, 6], and more memorable images evoke larger fMRI BOLD responses [5]. However, we lack a deeper understanding of how the representations of memorable and non-memorable images differ. Similarly, some insight into the neural correlates of memorability can be gained from convolutional neural network (CNN) models trained for object classification, which have been demonstrated to mimic other (i.e. object identity) representations in the form-processing pathway (reviewed by [7]). Image memorability can be reasonably decoded from the higher layers of at least one of these networks [8], but we do not understand how memorability is reflected in this CNN nor whether this CNN reflects memorability like the brain.

The fact that image memorability is linearly decodable in higher visual brain areas such as inferotemporal cortex (IT) [5, 6] could imply that information about image memorability is

represented in the same fashion as information about object identity in these areas. Within IT, representations of image and object identity are generally thought to be encoded as different patterns of spikes across the IT population, consistent with neurons that are individually “tuned” for distinct image and object properties. In a population representational space, these distinct spike patterns translate into population response vectors that point in different directions, and information about object identity is formatted such that it can be accessed from IT neural responses via a weighted linear decoder (Fig. 1a; reviewed by [9]). Similarly, image memorability could be represented by population vector direction in IT. However, under this proposal, it is not clear how our experience of image identity and image memorability would be represented as by the same population of neurons, i.e. the fact that one image of a person can be more memorable than another image of that same person, and at the same time, identity information, such as the class of an object, explains only a limited amount of how memorable an image will be [3].

Here we present an alternative proposal, hinted at by the fact that more memorable images evoke larger fMRI responses [5]: we propose that memorability variation is determined principally by the magnitude of the IT population response, or similarly, the total number of spikes across the IT population (Fig. 1). This scenario is consistent with general accounts in which visual perceptual processing precedes memory storage and images that evoke larger numbers of spikes, and consequently have more robust visual representations, are remembered best. This scenario incorporates a representational scheme for memorability that is orthogonal to the scheme IT uses to support object identity, and it is thus attractive from the perspective that it would provide a straightforward account of how IT multiplexes visual information about image content (as the population vector direction) as well as memorability (as population vector magnitude). The plausibility of this scenario rests on whether there is sufficient variation in population response magnitude across the class of natural images to account for memorability, given the host of homeostatic and normalization mechanisms that act to maintain constant grand mean firing rates across a cortical population [10].

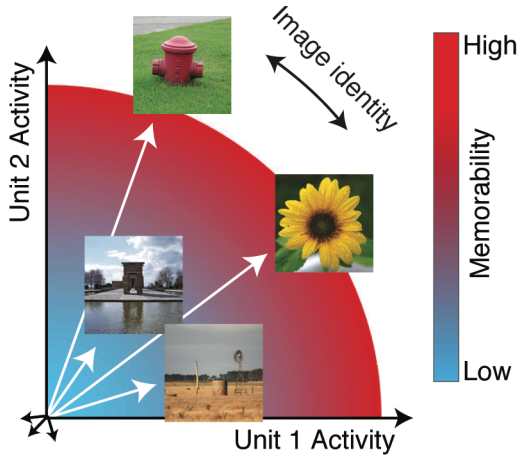


Figure 1. *The hypothesis: the magnitude of the IT population response encodes image memorability.* In geometric depictions of how IT represents image identity, the population response to an image is depicted as a vector in an N-dimensional space, where N indicates the number of neurons in the population, and identity is encoded by the direction of the population vector. Here we test the hypothesis

that image memorability is encoded by the magnitude (or equivalently length) of the IT population vector, where images that produce larger population responses are more memorable.

To test the hypothesis presented in Fig. 1, we obtained image memorability scores by passing images through a model designed to predict image memorability for humans ([4]; Supp. Fig. 1). The neural data, also reported in [11], were recorded from IT as two rhesus monkeys performed a single-exposure visual memory task in which they reported whether images were novel (never before seen) or were familiar (seen once previously; Fig. 2a). In each experimental session, neural populations with an average size of 26 units were recorded, across 27 sessions in total. After screening for responsive units, data were concatenated across sessions into a larger pseudopopulation in a manner that aligned images with similar memorability scores (see Methods and Supp. Fig. 1). The resulting pseudopopulation contained the responses of 707 IT units to 107 images, averaged across novel and familiar presentations.

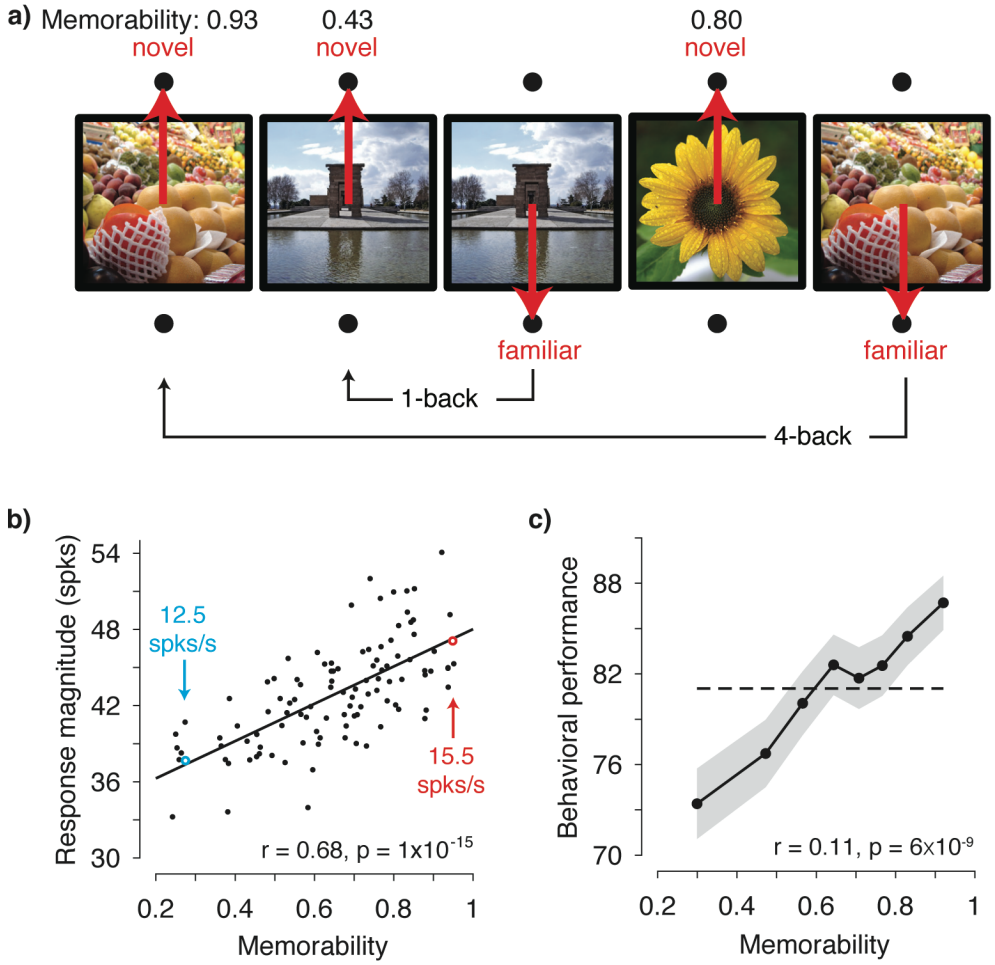


Figure 2. *IT population response magnitude strongly correlates with image memorability.* **a)** The monkeys' task involved viewing each image for 400 ms and then reporting whether the image was novel or familiar with an eye movement to one of two response targets. The probability of a novel versus familiar image was fixed at 50% and images were repeated with delays ranging from 0 to 63 intervening

trials (4.5 s to 4.8 min). Shown are 5 example trials with image memorability scores labeled. The memorability of each image was scored from 0-1, where the score reflects the predicted chance-corrected hit rate for detecting a familiar image (i.e., 0 maps to chance and 1 maps to ceiling, [4]). **b)** The relationship between image memorability scores and IT population response magnitudes. Each point corresponds to a different image (N=107 images). Population response magnitudes were computed as the L2 norm $\left(\sqrt{\sum_{i=1}^{N} r_i^2}\right)$, where r_i is the spike count response of the ith unit, across a pseudopopulation of 707 units. Spikes were counted in an 80 ms window positioned 180 to 260 ms following stimulus onset (see Supp. Fig 2c for different window positions). The Pearson correlation and its p-value are labeled. The solid line depicts the linear regression fit to the data. For reference, the mean firing rates for two example images are also labeled (see also Supp. Fig 3b). **c)** Mean and standard error (across experimental sessions) of monkey behavioral performance on the memory task as a function of human-based image memorability scores. Performance was binned across images with neighboring memorability scores and pooled across monkeys (see Supp. Fig 4 for plots by individual). The dashed line corresponds to the grand average performance, and if there were no correlation, all points should fall near this line. The point-biserial correlation and its p-value, computed for the raw data (i.e. continuous memorability scores and binary performance values for each image in each session) are labeled.

Fig. 2b shows the correlation between image memorability and IT population response magnitudes, which was strong and highly significant (Pearson correlation: $r = 0.68$; $p = 1\times10^{-15}$). This correlation remained strong when parsed by the data collected from each monkey individually (Supp. Fig. 2a-b) and, after accounting for the time required for signals to reach IT, across the entire 400 ms viewing period (Supp. Fig. 3a). The correlation also remained strong when computed for a quantity closely related to response magnitude, grand mean firing rate (Supp. Fig. 3b), as well as when the highest firing units were excluded from the analysis (Supp. Fig. 3c).

The strength of the correlation between memorability and IT response magnitude is notable given the species difference, as the memorability scores were derived from a model designed to predict what humans find memorable whereas the neural data were collected from rhesus monkeys. Likewise, we found that estimates of human memorability scores were predictive of the images that the monkeys found most memorable during the single-exposure visual memory task (Fig. 2c).

As described above, image memorability can be reasonably decoded from at least one CNN trained to categorize objects and scenes, but not explicitly to score memorability [8]. This hints at the fact that the neural correlate of memorability variation may be a consequence of the optimizations required for visual (as opposed to mnemonic) processing, however, before making this conclusion, one would want to establish that this CNN reflects memorability in a manner analogous to the brain. We found that this was the case: the correlation between image memorability scores and their corresponding population response magnitudes was significantly higher in the trained as compared to a randomly initialized version of the network in all layers, and the strength of this correlation generally increased across the hierarchy (Fig. 3). These results were also replicated in other CNNs trained for object classification, where correlation strength also systematically increased across the hierarchy throughout much of the network (Supp. Fig. 4), suggesting that this signature is not unique to this particular architecture or training procedure. These results suggest that variation in population response magnitude across images is likely to be a natural consequence of visual systems that classify objects, and that this variation is directly related to variation in image memorability.

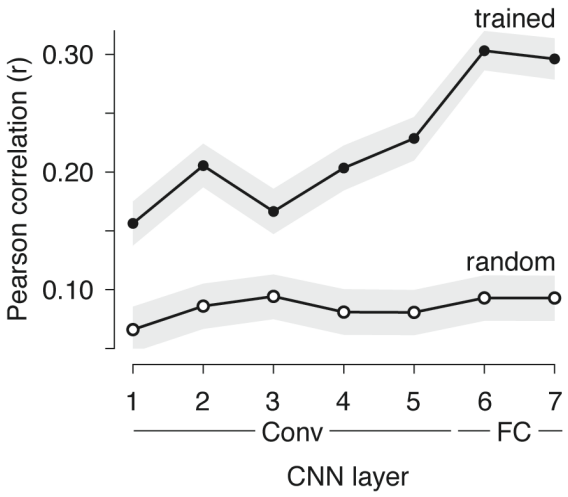


Figure 3. *Correlations between memorability and population response increase in strength across layers of a CNN trained to classify objects and scenes.* Mean and 95% CIs of the Pearson correlations between image memorability and population response magnitude for each hierarchical layer of the CNN described in [8], up to the last hidden layer. "Conv": convolutional layer; "FC": fully connected layer. p-values for a one-sided comparison that correlation strength was larger for the trained than the randomly connected network: $p < 0.0001$ for all layers.

Discussion

Here we have demonstrated that variation in the ability of humans and monkeys to remember images is strongly correlated with the magnitude of the population response in IT cortex. These results indicate that memorability is reflected in IT via a representational scheme that lies largely orthogonal to the one IT uses for encoding object identity (Fig. 1). For example, investigations of how monkey IT and its human analogs represent objects using 'representational similarity analysis' typically begin by normalizing population response vector magnitude to be the same for all images such that all that is left is the direction of the population response pattern, under the assumption that population vector magnitude is irrelevant for encoding object or image identity [12]. Before our study, data from human fMRI had pinpointed the locus of memorability to the human analog of IT, but we did not understand "how" the representations of memorable and non-memorable images differed. Our results point to a simple and coherent account of how IT multiplexes representations of visual and memorability information using two complementary representational schemes (Fig. 1).

How might variation in IT population response magnitudes lead to variation in how visual memories are stored? These results are consistent with general accounts of memory in which visual processing precedes memory storage and images with more robust visual representations are those that are best remembered. Our results demonstrate that despite the host of homeostatic mechanisms that contribute to maintaining constant global firing rates across a cortical population [10], changes in image content can result in IT population response magnitudes that differ by up to 19% (Fig. 2b; Supp. Fig. 3b). Of course one naturally expects that classes of images that are known to be more robustly represented in IT should be better remembered – for example, natural images should be better remembered than their scrambled counterparts [e.g. 13]. The significance of our result follows from the unexpected finding that there is variation in the robustness of visual representations within the class of natural images

that correlates with our understanding of the content that makes images more or less memorable. For example, unusual objects, such as a chair shaped like a hand, are known to be more memorable than typical objects, but the fact that unusual objects have more robust visual representations has not been previously established. As such, our results give insight not only into visual memorability, but also vision itself.

Our neural data were recorded from the brains of monkeys that could both see and remember what they had seen. To tease apart whether the origin of memorability could be attributed to optimizations for visual as opposed to mnemonic processing, we investigated CNNs optimized to categorize objects but not explicitly trained to predict the memorability of images. Prior to our study, memorability was demonstrated to be linearly decodable from higher layers of one of these CNNs, but it was unclear how memorability was reflected in this CNN and how that compared to the brain. Additionally, while this class of models has been demonstrated to mimic many aspects of how IT represents visual object identity (reviewed by [7]), image memorability has a distinct representational scheme from identity (Fig. 1), and in the context of the many illustrations that CNNs solve the same problems as brains using different strategies (e.g. [14]), it need not have been the case that CNNs reflected memorability in the same way as the brain. The fact that CNNs trained for object recognition mimic the neural representation of a distinct behavior – visual memorability – is compelling evidence that this strategy of multiplexing visual identity and memorability results from the computational requirements of optimizing for robust object representations. These modeling results also offer insight into the nature of the mechanism underlying memorability. The brain perceives and remembers using both feedforward and feedback processing, and this processing is modulated by top-down and bottom-up attention. Because of this, it is difficult to pinpoint the locus of an effect like the one we describe to any single mechanism using neural data alone. The fact that variations in response magnitudes that correlate with memorability emerge from static, feed-forward, and fixed networks suggests that memorability variation is unlikely to follow primarily from the types of attentional mechanisms that require top-down processing, recurrent processing, or plasticity beyond that required for wiring up a system to identify objects.

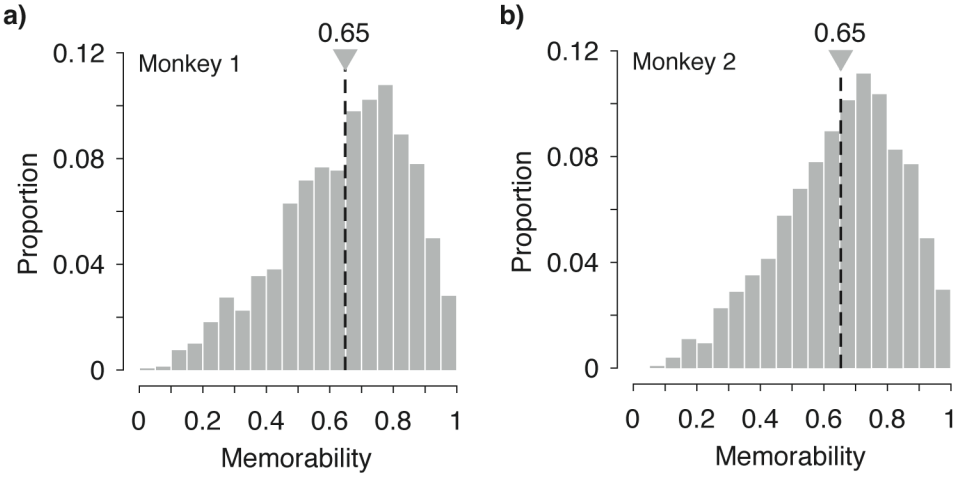
FUNDING

This work was supported by the National Eye Institute of the US National Institutes of Health (grant number R01EY020851); the Simons Foundation (through an award from the Simons Collaboration on the Global Brain); and the McKnight Endowment for Neuroscience.

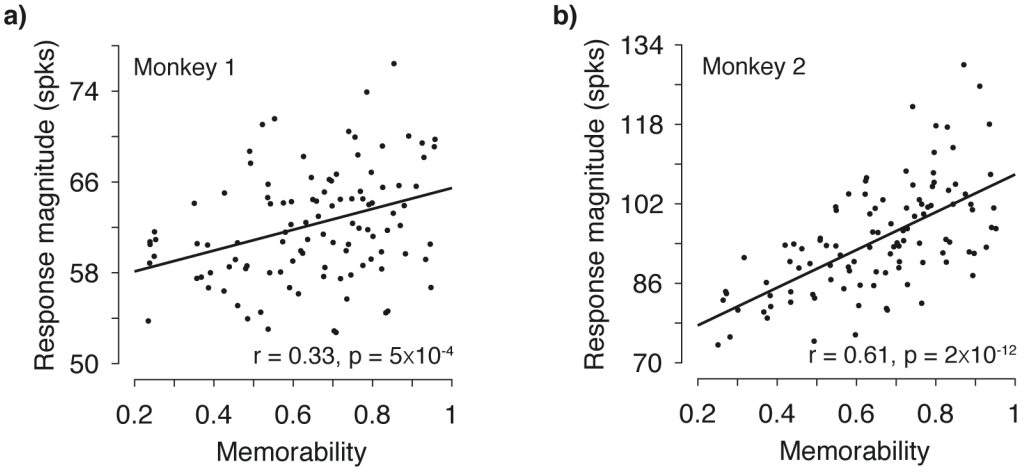
CONFLICTS OF INTEREST

None.

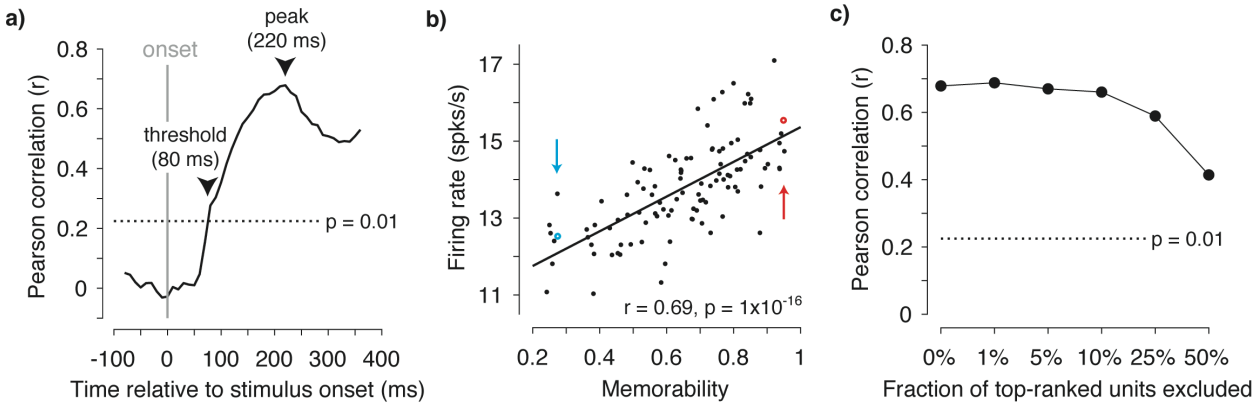
SUPPLEMENTAL FIGURES



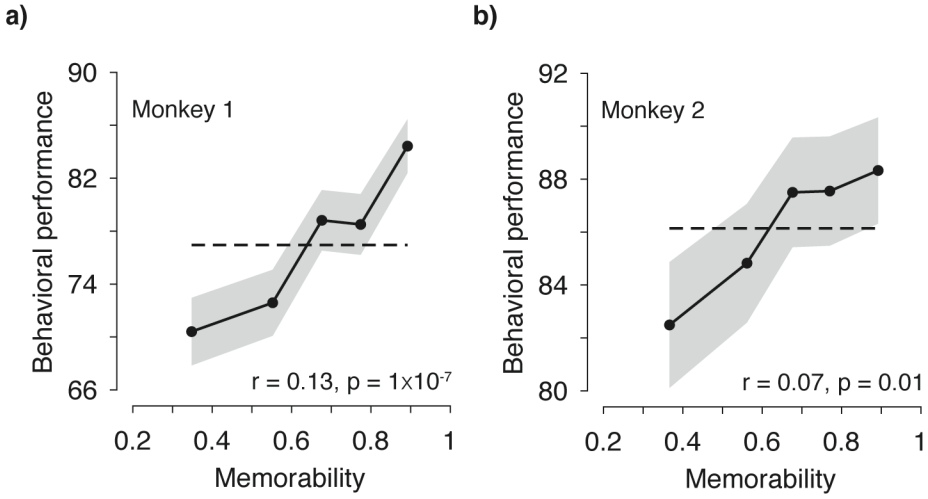
Supplemental Figure 1. *Distributions of memorability scores for the images used in these experiments.* Memorability scores range from 0-1, where the score reflects the predicted chance-corrected hit rate for detecting a familiar image and 0 maps to chance (see Methods and [4]).



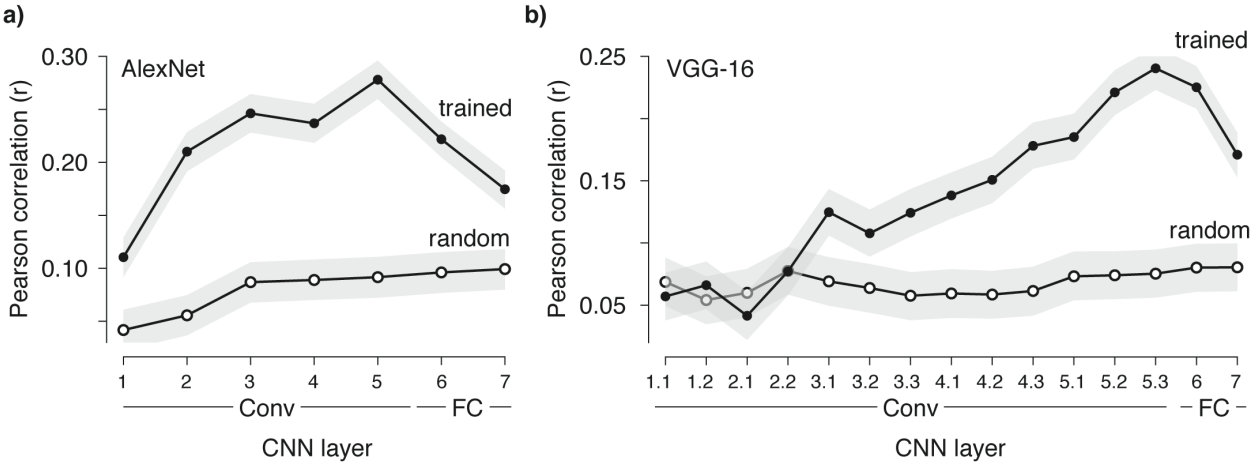
Supplemental Figure 2. *The correlation of memorability and population response magnitude, for each monkey individually.* **a-b)** Fig 2b replotted for each monkey individually (monkey 1: 353 units; monkey 2 354 units). To compensate for parsing the data, the spike count window was increased to 250 ms in these plots (positioned 150 ms – 400 ms) relative to the 80 ms window depicted in Fig. 2. The Pearson correlation and its p-value are labeled. The following two points were included in computing the correlations but fall outside the boundaries of the plot or are obscured by text: Monkey 1 (panel a): memorability = 0.86, response magnitude = 83.6; Monkey 2 (panel b): memorability = 0.57, response magnitude = 71.5. Solid lines depict the linear regression fits to the data.



Supplemental Figure 3. *The correlation of memorability and the IT population response, applied to different time windows, assessed with firing rate, and determined with top-ranked firing units removed.* **a)** The same analysis described for Fig. 2b, but applied to 80 ms windows shifted at different positions relative to stimulus onset, where the correlations are plotted against the center of each time bin. Fig. 2b is shown at the peak of this plot (220 ms). Also shown (dotted line) is the critical correlation threshold for $p < 0.01$, which the population reached at 80 ms following stimulus onset. **b)** Correlations between memorability and grand mean firing rate across the 707 units (in contrast to the plots of response magnitude in Fig. 2b). The two example images from 2b are indicated. Solid line depicts the linear regression fit to the data. **c)** The analysis in Fig. 2b with N% top-ranked firing rate units excluded from the pseudopopulation for different N. The dotted line indicates the critical correlation for the significance level $p = 0.01$.



Supplemental Figure 4. *Human-based memorability scores predict what monkeys find memorable.* The analysis presented in Fig. 2c, applied to each monkey individually. To compensate for parsing the data, the data is parsed into 5 bins as opposed to the 7 bins in Fig. 2c. The dashed lines correspond to the grand average performance, and if there were no correlation, all points should fall near this line. The point-biserial correlation and its p-value, computed for the raw data, are labeled.



Supplemental Figure 5. *Correlations between memorability and population response magnitude are also reflected in two other CNNs.* Mean and 95% CIs of the Pearson correlations between image memorability and population response magnitude for each hierarchical layer for two CNNs, including **a)** AlexNet [15], **b)** VGG-16 [16], up to the last hidden layer. “Conv”: convolutional layer; “FC”: fully connected layer. p-values for a one-sided comparison that correlation strength was larger for the trained than the randomly connected network, AlexNet: $p < 0.0001$ for all layers; VGG-16: $p = 0.8$, 0.2, 0.9, and 0.5 for Conv 1.1, 1.2, 2.1, and 2.2, respectively, $p = 0.0008$ for Conv 3.2, and $p < 0.0001$ for all other layers.

References

1. Standing, L., *Learning 10,000 pictures.* Q J Exp Psychol, 1973. **25**(2): p. 207-22.

2. Brady, T.F., et al., *Visual long-term memory has a massive storage capacity for object details.* Proc Natl Acad Sci U S A, 2008. **105**(38): p. 14325-9.

3. Isola, P., et al. *What makes an image memorable?* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

4. Khosla, A., et al. *Understanding and predicting image memorability at a large scale*. in *International Conference on Computer Vision (ICCV)*. 2015.

5. Bainbridge, W.A., D.D. Dilks, and A. Oliva, *Memorability: A stimulus-driven perceptual neural signature distinctive from memory.* Neuroimage, 2017. **149**: p. 141-152.

6. Bainbridge, W.A. and J. Rissman, *Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval.* Sci Rep, 2018. **8**(1): p. 8679.

7. Yamins, D.L. and J.J. DiCarlo, *Using goal-driven deep learning models to understand sensory cortex.* Nat Neurosci, 2016. **19**(3): p. 356-65.

8. Zhou, B., et al. *Learning deep features for scene recognition using places database.* in *Neural Inofrmation Processing Systems (NIPS)*. 2014.

9. DiCarlo, J.J., D. Zoccolan, and N.C. Rust, *How does the brain solve visual object recognition?* Neuron, 2012. **73**(3): p. 415-34.

10. Turrigiano, G., *Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function.* Cold Spring Harb Perspect Biol, 2012. **4**(1): p. a005736.

11. Meyer, T. and N.C. Rust, *Single-exposure visual memory judgments are reflected in inferotemporal cortex.* eLife, 2018. **7:e32259**.

12. Kriegeskorte, N., M. Mur, and P. Bandettini, *Representational similarity analysis - connecting the branches of systems neuroscience.* Front Syst Neurosci, 2008. **2**: p. 4.

13. Rust, N.C. and J.J. Dicarlo, *Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT.* J Neurosci, 2010. **30**(39): p. 12978-95.

14. Berardino, A., et al. *Eigen-distortions of hierarchical representations*. in *Adv. Neural Information Processing Systems (NeurIPS*17)*. 2017.

15. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *ImageNet classification with deep convolutional neural networks.* . in *International Conference on Neural Information Processing Systems (NIPS)*. 2012.

16. Simonyan, K. and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. in *International Conference on Machine Learning (ICLR)*. 2015.

METHODS:

As an overview, three types of data are included in this paper: 1) Behavioral and neural data collected from two rhesus monkeys that were performing a single-exposure visual memory task; 2) Human-based memorability scores for the images used in the monkey experiments, and 3**)** The responses of units at different layers of three convolutional neural network models trained to classify objects and scenes. The Methods associated with each type of data are described below.

Behavioral and neural data collected from two rhesus monkeys that were performing a single-exposure visual memory task

Experiments were performed on two adult male rhesus macaque monkeys (*Macaca mulatta)* with implanted head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee. Monkey behavioral and neural data were also included in an earlier report that examined the relationship between behavioral reports of familiarity as a function of the time between novel and familiar presentations (e.g., “rates of forgetting”) and neural responses in IT cortex [1]. The results presented here cannot be inferred from that report.

The single-exposure visual memory task:

All behavioral training and testing were performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli were presented on an LCD monitor with an 85 Hz refresh rate using customized software (http://mworks-project.org).

Each trial of the monkeys’ task involved viewing one image for at least 400 ms and indicating whether it was novel (had never been seen before) or familiar (had been seen exactly once) with an eye movement to one of two response targets. Images were never presented more than twice (once as novel and then as familiar) during the entire training and testing period of the experiment. Trials were initiated by the monkey fixating on a red square (0.25°) on the center of a gray screen, within an invisible square window of ±1.5°, followed by a 200 ms delay before a 4° stimulus appeared. The monkeys had to maintain fixation of the stimulus for 400 ms, at which time the red square turned green (go cue) and the monkey made a saccade to the target indicating that the stimulus was novel or familiar. In monkey 1, response targets appeared at stimulus onset; in monkey 2, response targets appeared at the time of the go cue. In both cases, targets were positioned 8° above or below the stimulus. The association between the target (up vs. down) and the report (novel vs. familiar) was swapped between the two animals. The image remained on the screen until a fixation break was detected. The first image presented in each session was always a novel image. The probability of a trial containing a novel vs. familiar image quickly converged to 50% for each class. Delays between novel and familiar presentations were pseudorandomly selected from a uniform distribution, in powers of two (n-back = 1, 2, 4, 8, 16, 32 and 64 trials corresponding to mean delays of 4.5s, 9s, 18s, 36s, 1.2 min, 2.4 min, and 4.8 min, respectively).

The images used in these experiments were collected via an automated procedure that downloaded images from the Internet. Images smaller than 96*96 pixels were not considered and eligible images were cropped to be square and resized to 256*256 pixels. An algorithm

removed duplicate images. Within the training and testing history for each monkey, images were not repeated.

The activity of neurons in IT was recorded via a single recording chamber in each monkey. Chamber placement was guided by anatomical magnetic resonance images in both monkeys. The region of IT recorded was located on the ventral surface of the brain, over an area that spanned 5 mm lateral to the anterior middle temporal sulcus and 14-17 mm anterior to the ear canals. Recording sessions began after the monkeys were fully trained on the task and after the depth and extent of IT was mapped within the recording chamber. Combined recording and behavioral training sessions happened 4-5 times per week across a span of 5 weeks (monkey 1) and 4 weeks (monkey 2). Neural activity was recorded with 24-channel U-probes (Plexon, Inc) with linearly arranged recording sites spaced with 100 μm intervals. Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored using the Grapevine Data Acquisition System (Ripple, Inc.). Spike sorting was done manually offline (Plexon Offline Sorter). At least one candidate unit was identified on each recording channel, and 2-3 units were occasionally identified on the same channel. Spike sorting was performed blind to any experimental conditions to avoid bias. For quality control, recording sessions were screened based on their neural recording stability across the session, their numbers of visually responsive units, and the numbers of behavioral trials completed. A multi-channel recording session was included in the analysis if: (1) the recording session was stable, quantified as the grand mean firing rate across channels changing less than 2-fold across the session; (2) over 50% of neurons were visually responsive (a loose criterion based on our previous experience in IT), assessed by a visual inspection of rasters; and (3) the number of successfully completed novel/familiar pairs of trials exceeded 100. In monkey 1, 21 sessions were recorded and 6 were removed (2 from each of the 3 criterion). In monkey 2, 16 sessions were recorded and 4 were removed (1, 2 and 1 due to criterion 1, 2 and 3, respectively). The resulting data set included 15 sessions for monkey 1 (n = 403 candidate units), and 12 sessions for monkey 2 (n = 396 candidate units). Both monkeys performed many hundreds of trials during each session (~600-1000, corresponding to ~300-500 images each repeated twice). The data reported here correspond to the subset of images for which the monkeys' behavioral reports were recorded for both novel and familiar presentations (e.g. trials in which the monkeys did not prematurely break fixation during either the novel or the familiar presentation of an image). Finally, units were screened for stimulus-evoked activity via a comparison of their responses in a 200 ms period before stimulus onset (-200 ms – 0 ms) versus after stimulus onset (80 – 280 ms) with a two-sided t-test, $p < 0.01$. This yielded 353 (of 403) units for monkey 1 and 354 (out of 396) units for monkey 2.

To perform our analyses, we concatenated units across sessions to create a larger pseudopopulation. In the case of the pooled data, this included 27 sessions in total (15 sessions from monkey 1 and 12 from monkey 2). When creating this pseudopopulation, we aligned data across sessions in a manner that preserved whether the trials were presented as novel or familiar, their n-back separation, and image memorability scores (obtained using methods described below). More specifically, the responses for each unit always contained sets of novel/familiar pairings of the same images, and pseudopopulation responses across units were always aligned for novel/familiar pairs that contained the same n-back separation and images with similar memorability scores. When the number of images in a session exceeded the number required to construct the pseudopopulation, a subset of images were selected separately for each n-back by ranking images within that n-back by their memorability scores,

preserving the lowest-ranked and highest-ranked images within that session, and selecting the number of additional images required as those with memorability scores that were evenly spaced between the two extreme memorability scores for that session. The resulting pseudopopulation consisted of the responses to 107 images presented as both novel and familiar (i.e. 15, 15, 16, 17, 17, 15 and 12 trials at 1, 2, 4, 8, 16, 32 and 64-back, respectively). To perform the neural analyses (Fig 2b, Supp Figs 2, 3), a memorability score for each of the 107 pseudopopulation images was computed as the mean of the memorability scores across all the actual images that were aligned to produce that pseudopopulation response. The average standard deviation across the set of memorability scores used to produce each pseudopopulation response was 0.05, where memorability ranges 0-1. To perform behavioral analyses (Fig 2c, Supp Fig 4), the memorability score as well as binary performance values (correct/wrong at reporting that a familiar image was familiar) were retained for each of the 107 images, across each of the 27 sessions.

Human-based memorability scores for the images used in the monkey experiments

We obtained memorability scores for the images used in the monkey experiments using MemNet [2] estimates. MemNet is a convolutional neural network (CNN) trained to estimate image memorability on a large-scale dataset of natural images (LaMem [2], publicly available at memorability.csail.mit.edu). LaMem consists of 60K images drawn from a diverse range of sources (See [2] for more detail). Each image in this dataset is associated with a memorability score based on human performances in an online memory game on Amazon's Mechanical Turk. Behavioral performances were corrected for the delay interval between first and second presentation to produce a single memorability score for each image. After training, MemNet estimates visual memorability of natural images near the upper bound imposed by human performance: MemNet estimates reach 0.64 rank correlation with mean human-estimated memorability, while the upper bound of consistency between human scores has a rank correlation of 0.68. Here we treat MemNet memorability estimates as a proxy for human memorability scores.

The memorability scores were obtained using the network weights reported in [2] and publicly available at http://memorability.csail.mit.edu/download.html. This network was originally trained using the Caffe framework [3], and we ported the trained network to Pytorch [4] using the caffe-to-torch-to-pytorch package at https://github.com/fanq15/caffe_to_torch_to_pytorch. Before passing images into MemNet, we preprocessed them as described in [5]: we resized images to 256 × 256 pixels (with bilinear interpolation), subtracted the mean RGB image intensity (computed over the dataset used for pretraining, as described in [5]), and then produced 10 crops of size 227 × 227 pixels. The 10 crops were obtained by cropping the full image at the center and at each of the four corners and by flipping each of these 5 cropped images about the vertical axis. All 10 crops were passed through MemNet. The average of these 10 scores was used as the mean prediction of the model for the input image. This mean prediction was then linearly transformed to obtain the estimated memorability score:

$$\text{Memorability_score} = \min(\max((\text{output} - \text{mean_pred})*2 + \text{additive_mean}, 0), 1)$$

where following [2], we set mean_pred = 0.7626 and additive_mean = 0.65.

The responses of units at different layers of CNN models trained to classify objects and scenes.

We evaluated the correlation between response magnitude and image memorability on images from the LaMem dataset [2] using three commonly used convolutional neural networks (CNNs). All reported models were evaluated on the full test set of split 1 of LaMem, which contains 10,000 images. We chose to use LaMem images, as each image in this dataset is labeled with a memorability score computed directly from human behavioral performance (i.e. not estimated with a model; see above and [2] for details of data collection and memorability score computation). All networks were run in TensorFlow 1.10 ([6], software available from tensorflow.org), using custom Python evaluation code.

The results presented in Fig 3 were obtained by running images from this dataset through HybridCNN [5]. HybridCNN is a network with an identical architecture to AlexNet [7]. HybridCNN was first trained to classify natural images of objects and scenes using data from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, a 1000-way object classification dataset [8], as well as the Places 183-way scene classification dataset [5], for a combined 1183-way classification task. For details of training, see [5]. Results were obtained using the network weights reported in [5] and publicly available at http://places.csail.mit.edu/downloadCNN.html. This network was originally trained using the Caffe framework [3], and we ported the trained network to TensorFlow using the caffe-tensorflow package https://github.com/ethereon/caffe-tensorflow. Random initialization baselines were obtained using the same architecture, but randomly sampling the weights using the initialization algorithm described in [9].

Before passing images into each network, we preprocessed them as described in [5] and above: we resized images to 256 × 256 pixels (with bilinear interpolation), subtracted the mean RGB image intensity (computed over the training dataset), and then cropped the central 227 x 227 and passed it into the network. The response magnitude (L2 norm) of each layer was computed over the full output vector of each hidden layer. In all cases, we show the magnitude of hidden layer output after applying the nonlinear operation. Results for the two networks presented in the supplement (Supp Fig 5) were obtained in an identical manner, except for the image preprocessing step. For each network, images were preprocessed as described in the original papers (AlexNet: [7], VGG-16: [10]).

For all three networks (HybridCNN, AlexNet, and VGG-16), we computed correlations for all convolutional and fully-connected hidden layers. The Pearson correlation coefficient was used to measure correlation. All correlations were computed over the full set of 10,000 images described above. 95% confidence intervals for the correlation coefficient of each layer were obtained by bootstrapping over the set of 10,000 per-image layer magnitudes and memorability scores. 95% confidence intervals were estimated empirically as the upper and lower 97.5%-centiles of the bootstrapped correlation coefficients for each layer and condition. Bootstrapped resampling was performed independently for each layer and each condition (trained or randomly connected). In all cases, bootstrap estimates were performed using 10,000 samples (with replacement) of the full dataset of 10,000 images. The bootstrapping procedure was also used to conduct one-tailed tests to determine whether the correlations between memorability and response magnitude were stronger in the trained as compared to the randomly initialized network at each layer separately. p-values were estimated by taking pairs of correlation coefficients computed on the bootstrapped data for each condition and measuring the rate at which the correlation for the random layer exceeded the correlation for the trained layer.

REFERENCES:

1. Meyer, T. and N.C. Rust, *Single-exposure visual memory judgments are reflected in inferotemporal cortex.* eLife, 2018. **7:e32259**.

2. Khosla, A., et al. *Understanding and predicting image memorability at a large scale*. in *International Conference on Computer Vision (ICCV)*. 2015.

3. Jia, Y., et al. *Caffe: convolutional architecture for fast feature embedding.* in *22nd ACM International conference on Multimedia (MM'14)*. 2014. New York.

4. Paszke, A., et al. *Automatic differentiation in PyTorch*. in *Neural Information Processing Systems (NIPS)*. 2017.

5. Zhou, B., et al. *Learning deep features for scene recognition using places database.* in *Neural Inofrmation Processing Systems (NIPS)*. 2014.

6. Abadi, M., et al., *TensorFlow: Large-Scale machine learning on heterogeneous distriuted systems* 2016: arXiv.

7. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *ImageNet classification with deep convolutional neural networks.* . in *International Conference on Neural Information Processing Systems (NIPS)*. 2012.

8. Deng, J., et al. *ImageNet: A large-scale hiearchical image database*. in *IEEE Computer Vision and Pattern Recognition (CVPR)*. 2009.

9. Giorot, X. and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks*. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010.

10. Simonyan, K. and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. in *International Conference on Machine Learning (ICLR)*. 2015.