Title: Image memorability is predicted at different stages of a convolutional neural network

Author names and affiliations:
Griffin E. Koch[1,2,3*], Essang Akpan[1,2], and Marc N. Coutanche[1,2,3,4]

[1] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA 15260
[2] Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, USA 15260
[3] Center for the Neural Basis of Cognition, Pittsburgh, PA, USA 15260
[4] Brain Institute, University of Pittsburgh, Pittsburgh, PA, USA 15260

Corresponding author:
Griffin E. Koch; griffinkoch@pitt.edu

Keywords:
Memorability, Convolutional Neural Networks, vision, recognition, memory

Abstract

The features of an image can be represented at multiple levels – from low-level visual properties to high-level meaning. What factors drive some images to be memorable while others are forgettable? Across three behavioral experiments, we addressed this question. In a first behavioral experiment, we combined a convolutional neural network (CNN) with behavioral prospective assignment, by using four CNN layers to select the scene images that each of one hundred participants experience. We found that participants remembered more images when they were assigned to view stimuli selected based on their discriminability in low-level CNN layers, or similarity in high-level layers. A second experiment replicated these results in an independent sample of participants. The third experiment investigated how memorability is influenced when images fall within a single semantic category (houses). We replicated results from the first two experiments at lower levels, but found that similarity predicted memorability at a mid-high level, rather than the highest level observed for scenes from multiple categories. This mid-high level contains representations for objects and object-parts, which are more important for discriminating images from the same category. Together, this research provides evidence that discriminability at different visual levels, modeled using a CNN, predicts image memorability.

When faced with the option to dine at a new restaurant, we might rely on the familiarity of a certain building, or the look of a specific logo, perhaps from a commercial we once saw. Advertising and marketing departments certainly hope this is the case, but what makes certain places and pictures more likely to be remembered than others? Determining why, when, and how images are memorable is a necessary step for modelling the relationship between visual perception and memory, and applying research, such as to select memorable health-related or educational images.

The particular images that are more likely to be remembered than others are remarkably consistent across individuals (Bainbridge, Isola, & Oliva, 2013; Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015; Isola et al., 2014). Yet, the exact reasons that some images are more memorable than others remain to be determined. Simple visual features, such as spatial frequency, hue, and saturation, struggle to predict an image's memorability (Bainbridge et al., 2017; Dubey et al., 2015; Isola et al., 2014), as do participants' own subjective predictions about whether an image will be easily remembered (Isola et al., 2014). Images that are visually distinctive have been shown to be particularly memorable (Bartlett et al., 1984; Busey, 2001; Huebner & Gegenfurtner, 2012; Lukavský & Děchtěrenko, 2017), and some form of high-level content plays a role, based on the negative consequences of rearranging visual features (Lin et al., 2018) though the nature of predictive low-level and high-level content remains unclear.

The intrinsic visual properties of images play a more important role in memorability than even individual differences in observers (Bylinskii et al., 2015). An image's properties range from low-level features such as color and edges, to high-level semantic components such as a scene's category (Epstein & Baker, 2019). Researchers have previously quantified how visual features relate to memory with algorithms that account for the color and spatial frequencies of

images (Huebner & Gegenfurtner, 2012) but such algorithms are not able to simultaneously account for higher-level properties. This is important because higher-level visual properties are thought to be central to memorability, to the degree that it has been stated that "memorability is not low-level vision" (Bainbridge et al., 2017, p. 149). Bainbridge and colleagues (2017) recently used functional magnetic resonance imaging (fMRI) and representational similarity analysis (RSA) to identify brain regions with pattern of activity that reflect an image's memorability-based similarity space, which included regions involved in higher-level visual processing, such as later ventral stream areas and more traditional memory-related areas within the medial temporal lobe. The authors noted, however, that controlling for low-level visual features in their stimulus set, weakened the ability to find relationships between memorability and lower-level visual processing areas, such as early visual cortex. Thus, while these results supported claims that higher-level representations play a role in memorability, they do not necessarily rule out a role for lower-level representations. The question therefore remains as to how lower-level *and* higher-level features (which are both present in any image) may play complementary (or even competing) roles in memorability.

Recently, rather than considering stimuli in terms of independent visual features, vision scientists have utilized convolutional neural networks (CNNs) with a hierarchical organization to characterize images (Deng et al., 2009; Kriegeskorte, 2015). These object recognition CNNs often have similarities with the human visual system (Lindsay & Miller, 2018), in which visual features are represented at increasingly high levels (Coutanche et al., 2016). Early CNN layers extract basic visual properties, which become increasingly high-level in later layers, until the final layer classifies the image (Krizhevsky et al., 2012; LeCun et al., 2015). Although primarily employed in the vision sciences, CNNs are not only applicable to perception and vision, but also

to the memory field (though they are currently rarely utilized). Because CNN models allow multiple feature representations to be extracted for a single image, they make an excellent tool for measuring the impact that varied visual features may have on behavioral outcomes, such as memory performance.

In this set of experiments, we leverage the multiple layers of object recognition CNNs to determine why certain images are more memorable than others, at multiple visual stages. Uniquely, we use a CNN model to select the images that participants will view, in a prospective assignment design. Typically, research within the memory field involves collecting measures of memory performance, which are then related to others measures of interest (e.g., visual properties; Dubey, Peterson, Khosla, Yang, & Ghanem, 2015). Yet, by *retrospectively* relating memory performance to the features of items, a causal relationship between features and memory performance is not established. In contrast, prospective assignment –common in more clinical settings– provides stronger evidence of causality.

Across three behavioral experiments, we test how the hierarchy of visual properties (as measured through the features at different layers within a CNN) determine an image's memorability. In Experiments 1 and 2, we utilized prospective assignment to present participants with scene images from multiple categories that were either similar or discriminable based on CNN levels. In Experiment 3, we investigated how CNN levels predict memorability for images from the same semantic category.

Results

We investigated how image memorability is influenced by visual features at different levels.

Experiment 1

In the first experiment, we contrasted participants' ability to remember both discriminable and similar images of scenes, compared to foils, that were selected based on a level of a trained CNN. On average, the odds of identifying an image as having been seen before were 18.59 times greater for the previously presented images than for matched foils (B = 2.92, $p < .001$, 95% confidence interval in odds [17.24, 20.04]). Our key question of interest was how the memory of previously presented images would differ based on which CNN layer had been used to select them. A test for polynomial effects across all 100 participants to predict memory performance for both discriminable and similar images across the four conditions (i.e., CNN layers: 1, 3, 5 or 8) revealed a significant linear interaction between image similarity (i.e., how similar an image's features from the corresponding CNN layer are to other image's features from the same layer) and layer condition (B = 0.33, odds = 1.39, $p < .001$, [1.16, 1.67]). This similarity x layer interaction was key to detecting the relationship, as there was no main effect of similarity when the layers were collapsed (B = 0.02, odds = 1.02, $p = .742$, [0.93, 1.11]). Neither a quadratic nor cubic function fit the data better than linear (ps > .197).

Each of the four conditions were then separately examined using individual regression models. Within images selected based on the earliest layer (layer 1), more similar images were less likely to be correctly recognized as old than were more discriminable images (B = -0.26, odds = 0.77, $p = .005$, [0.64, 0.93], Cohen's $d = 0.35$). For ease of interpretability, this means that the odds of correctly recognizing an image as having been seen before were 1.30 times greater for images categorized as discriminable, than for images categorized as similar. Within the images selected based on the last layer (layer 8), the odds of correctly recognizing an image as old were 1.26 times greater for images categorized as similar, than for images categorized as

discriminable (B = 0.23, p = .012, [1.05, 1.52], Cohen's d = 0.38). No differences were observed between similar and discriminable images in the middle two layers (3 and 5) (ps > .581). Mean hit rates (correctly recognizing an image as old) are shown in Figure 1.

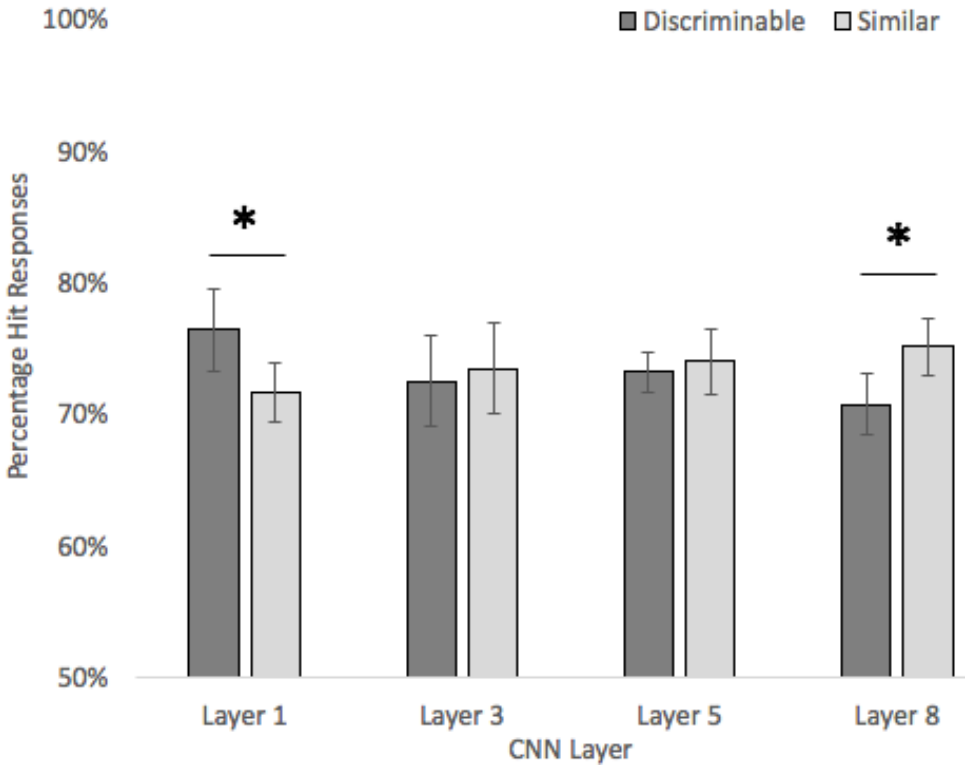

Figure 1. Mean hit rates for previously presented images categorized as discriminable or similar based on the four layers of the CNN. Error bars reflect standard error of the mean. * represents statistical significance (p < .05).

We also implemented regression models with continuous (rather than similar/discriminable categorical) predictors of image similarity. For every one standard deviation *decrease* in similarity (Fisher-Z r-values) in layer 1, the odds of correctly recognizing an image as old were 1.14 times greater (β = -0.13, odds = 0.88, p = .006, [0.80, 0.96]). For layer 8, for every one standard deviation *increase* in similarity, the odds of correctly recognizing an

image as old were 1.12 times greater ($\beta = 0.11$, $p = .018$, [1.02, 1.22]). No differences were observed for the middle two layers (3 and 5) (ps > .674).

Experiment 2

We next asked whether these relationships between discriminability and memorability would generalize to an independent sample of participants that did not have a verbal free recall task prior to testing recognition memory (as was the case for Experiment 1). We followed the same approach as in Experiment 1, but only assigned participants to either layer 1 or layer 8, as these were the two layers with significant effects in Experiment 1. Again, as expected, participants were able to correctly recognize images that had been seen before, compared to the novel foils, with the odds of identifying an image as old being 21.98 times greater for previously presented images than for matched foils (B = 3.09, $p < .001$, [19.67, 24.57]).

As in Experiment 1, each of the two conditions were separately examined using individual regression models. Within the images based on the earliest layer (layer 1), more similar images were less likely to be correctly recognized as old than were images that were more discriminable (B = -0.71, odds = 0.49, $p < .001$, [0.41, 0.59], Cohen's $d = 1.33$). The odds of correctly recognizing an image as old were 2.04 times greater for images categorized as discriminable, than categorized as similar. In contrast, for images selected based on the last layer (layer 8), the odds of correctly recognizing an image as old were 1.62 times greater for images categorized as similar, than as discriminable (B = 0.48, $p < .001$, [1.35, 1.94], Cohen's $d = 0.91$). Mean hit rates (correctly recognizing an image as old) Experiment 2 are shown in Figure 2.

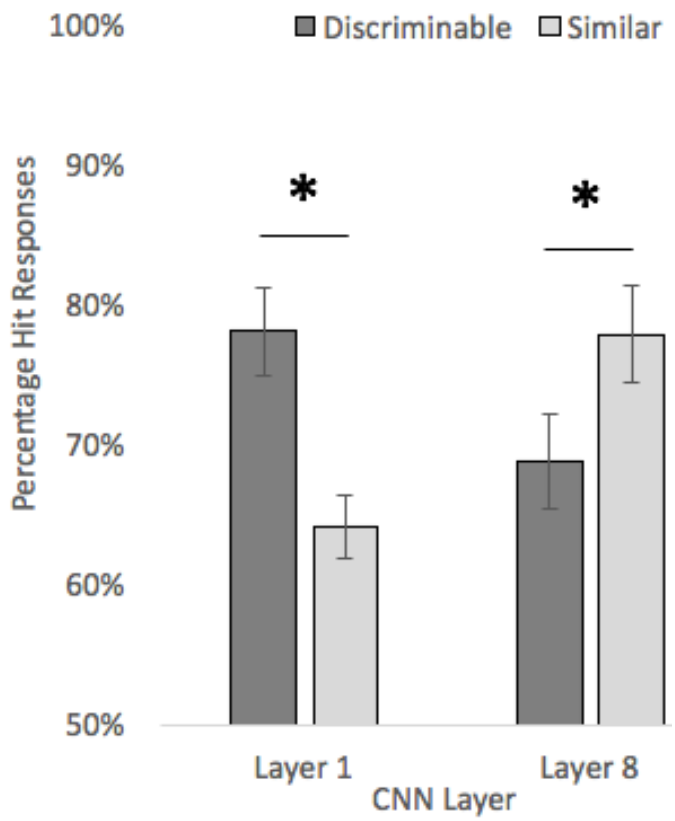

Figure 2. Mean hit rate (correctly recognizing an image as having been seen before) for images categorized as discriminable or similar based on the first and final layers of the CNN. Error bars reflect standard error of the mean. * represents statistical significance ($p < .05$).

As in Experiment 1, we also implemented regression models with continuous (instead of categorical) predictors of image similarity. For every one standard deviation decrease in similarity (Fisher-*Z* *r*-values) based on layer 1, the odds of correctly recognizing a previously presented image as old were 1.41 times greater ($\beta = -0.34$, odds = 0.71, $p < .001$, [0.65, 0.78]). In the model using features from layer 8, for every one standard deviation increase in similarity, the odds of correctly recognizing a previously presented image as old were 1.27 times greater ($\beta = 0.24$, $p < .001$, [1.16, 1.39]).

Experiment 3

Experiment 3 examined memory recognition for a large set of images within one semantic category (houses) in an independent set of participants. This experiment retrospectively tests for a relationship between visual discriminability and recognition memory performance, rather than selecting images based on particular CNN layers, so these results were analyzed in a continuous fashion. In this new dataset, we replicated the layer 1 results of Experiment 1 and 2, in which greater discriminability in the earliest layer predicted superior subsequent memory performance ($\beta = -0.17$, odds = 0.84, $p < .001$, [0.76, 0.93]). For every one standard deviation decrease in layer 1 similarity, the odds of correctly recognizing a previously presented image as old were 1.19 times greater. In contrast to Experiments 1 and 2, in which greater similarity in the final layer (8) predicted stronger memorability of the varied scenes, this third experiment identified that greater similarity within *layer 5* predicted stronger subsequent memory performance for these images coming from just one semantic category. For every one standard deviation increase in similarity, the odds of correctly recognizing a previously presented image as old were 1.16 times greater ($\beta = 0.15$, $p = .003$, [1.05, 1.28]). No differences were observed for either layer 3 or layer 8 (ps > .163).

Discussion

We have investigated how image discriminability at multiple visual levels predicts the likelihood that an image will be remembered. In Experiment 1, we found –in a prospective assignment paradigm– that participants remembered more scene images if the images were selected based on high discriminability in low-level visual properties (earliest CNN layer), or high similarity in higher-order properties (final CNN layer). Experiment 2 replicated these

results in an independent sample when the recognition test occurred immediately after image presentation. In Experiment 3, we examined how discriminability across CNN layers predicted memorability for images in one semantic category (houses), rather than the multiple categories in the first two experiments. In this third experiment, discriminability in the earliest CNN layer again positively predicted memorability, but for this same-category set of images, similarity at a mid-high level (layer 5) –representing object parts– predicted memorability.

Our findings that similarity and discriminability can support memorability at different levels help to reconcile several seemingly conflicting findings. For instance, in some prior research, image memorability has been associated with the presence of discriminable features (Bartlett et al., 1984; Bruce et al., 1994; Lukavský & Děchtěrenko, 2017), whereas other investigations have found an association with similarity (Bainbridge et al., 2017; Bainbridge & Rissman, 2018). Our evidence suggests that both are true – discriminability and similarity are each important predictors for whether an image will be remembered, though they operate at different stages of visual processing.

A notable finding across the three behavioral experiments is the different CNN layers that had a positive similarity relationship with memorability (layer 8 in Experiments 1 and 2; layer 5 in Experiment 3). The likely basis for this difference is the use of scene images from multiple semantic categories in the first two experiments, and only one (houses) in Experiment 3. The final CNN layer (layer 8), which classifies images into categories, is particularly important for a stimulus set that covers a multitude of classes (e.g., igloo, field, etc.). In contrast, stimuli from the same class are better differentiated by variability in objects and other mid-level features, which are extracted in layer 5. This finding –that the positive influence of image similarity can occur at different higher levels– highlights the critical factor of the kind (and level) of features

that distinguish a remembered image from others in a particular set. It is possible that our finding of increased memorability for images that are similar in high-level visual features could derive from links to idiosyncratic memory as suggested by Charest and colleagues (Charest et al., 2014) Such images could be the types of images that evoke the most personal episodic memories amongst people, thus resulting in increased memorability. Additionally, our finding is in line with recent neuroimaging evidence from Bainbridge and colleagues which showed that later regions within the visual ventral stream were sensitive to differences in image memorability (Bainbridge et al., 2017). These authors suggest that their findings could be in part due to feedback from memory regions within the medial temporal lobe.

The consistent results of discriminability at the lowest level predicting memorability across all three behavioral experiments highlights the robustness of these effects. Therefore, it is not likely that our findings are driven by differences in encoding tasks, nor additional forms of memory testing prior to the recognition task of interest.

An aspect of the design worth highlighting is the prospective assignment performed in Experiments 1 and 2. Typically, experiments of memorability test how memory for a large set of intermixed images varies with several metrics, such as visual properties. This approach can be valuable for identifying potential underlying predictors of memorability, but by *retrospectively* relating features to memory performance, the relationship is necessarily correlational. In contrast, prospective assignment –common in clinical trials– provides stronger evidence of causality because the hypothesized dimension of interest (here, CNN layer discriminability) is used to allocate participants to different conditions in advance, and allows for greater confidence in the reason for differing outcomes (memory performance) across the groups. In addition to giving us greater confidence in the cause of differences in image memorability, this approach also

minimized any potential interference from images selected using other layers (which are given to different groups). Future memorability research might consider prospectively partitioning presented items based on hypothesized features of interest, as opposed to the more common method of retrospectively relating memory performance to stimulus-level features in a large set of presented stimuli.

The use of CNNs in the field of cognitive psychology, and specifically relating to memorability, remain underutilized. For the scope of these experiments, we were interested in the underlying levels that are represented by the layers of a given CNN (in this case, AlexNet), although we note that other CNNs have shown greater similarity with human brain activity. Future research may attempt to more closely link the specific representations between CNNs and neural activity in relation to memorability, however, the specific modeling of visual processing within the human brain, while interesting, is beyond the scope of our analyses here. The current work builds on prior research which has shown the promise of adapting CNNs to identify features relevant for memorability (Khosla et al., 2015). Our work contrasts the representations at multiple individual layers, as a way to examine the contribution of each level to memorability.

Materials and Methods

Participants

In Experiment 1, participants were recruited until 100 contributed usable data (25 in each condition), in line with prior research investigating recognition memory for scenes (Konkle et al., 2010a, 2010b). Participants were native English speakers with normal or corrected-to-normal vision, without a learning or attention disorder, and from the University of Pittsburgh community (49 females, 51 males, mean (*M*) age = 19.6 years, standard deviation (*SD*) = 1.7 years). Four

participants' data were not analyzed after the initial encoding phase due to low task accuracy (described in more detail below), so were removed and excluded from the target of 100 participants.

For Experiment 2, participants were recruited until 50 contributed usable data (25 in each condition). Participants were recruited using the same criteria as Experiment 1, but were an independent sample (25 females, 25 males, *M* age = 19.7 years, *SD* = 1.8 years). Three participants' data were not analyzed after the initial encoding phase due to low task accuracy (described in more detail below), so were removed and excluded from the target of 50 participants.

For Experiment 3, data from thirty-two participants (15 females, 17 males, *M* age = 19.7 years, *SD* = 1.2 years) were analyzed. Participants who were run but did not contribute to data analysis included four who experienced technical malfunctions, and seven who failed sense-checks of task behavior during the encoding phase (described below). All participants were native English speakers with normal or corrected-to-normal vision, and without a learning or attention disorder. The institutional review board (IRB) at the University of Pittsburgh approved all measures prior to all experiments. Participants were compensated through course credit for their participation.

Stimuli and Materials

Stimuli for Experiments 1 and 2 consisted of 1,000 images from the *Scenes* collection within the BOLD5000 dataset (Chang et al., 2019). Scenes ranged across semantic categories (e.g., airport, restaurant, soccer field) and included at most four images from one semantic category (e.g., four images of different soccer fields). In Experiment 3, stimuli consisted of 120

images of houses. House images were collected from real estate websites displaying houses found within the Northeastern United States, with a majority being from Pennsylvania. These images depicted houses using the same viewpoint ("front-on") and were generally centered within the image.

Procedure

CNN metrics

Across all experiments, each image was submitted to the pre-trained AlexNet CNN model (Deng et al., 2009; Krizhevsky, Sutskever, & Hinton, 2012) through the MatConvNet MATLAB toolbox (http://www.vlfeat.org/matconvnet/; Vedaldi & Lenc, 2014). This CNN is one of the most well-known and accessible convolutional neural network models, and had been trained using more than 1.2 million images as part of the ImageNet object classification challenge. For each image, the CNN's feature weights were extracted from four different layers of the CNN: earliest (convolutional layer 1), early-middle (convolutional layer 3), late-middle (convolutional layer 5), and last (fully-connected layer 8).

We focused on image discriminability (the inverse of similarity) by measuring the similarity between the presented images at each CNN layer with pairwise correlations between their sets of feature weights, giving a 1,000 x 1,000 matrix of image similarity for each layer in Experiments 1 and 2, and a 120 x 120 matrix of image similarity for each layer in Experiment 3. We Fisher-*Z* transformed the resulting correlation coefficients (*r*-values), and averaged these for each image to give a value reflecting its average similarity (or discriminability) with other images in the experiment, according to each of the CNN stages.

Participant assignment

For Experiment 1, participants were randomly assigned to one of four conditions. Participants in each condition were presented with images selected based on a CNN layer (layer 1, 3, 5 or 8). This subset of layers from the CNN was selected so that a diverse group of images could be presented to separate samples of participants. By spacing the layers throughout the CNN we hoped to increase the diversity of images, as representations within sequential layers will be more correlated with each other than with alternate layers.

Condition assignment dictated which set of images would be presented to a participant. The above CNN metrics of image similarity/discriminability (in each layer) were used to determine the images that were presented to each group. Each group was presented with the 50 most similar (highest *r*-values) and 50 most discriminable (lowest *r*-values) images based on its corresponding layer. See Tables 1 and 2 for a breakdown of image overlap between layers as well as similar/discriminable categorization. To allow for foils from the same semantic category (defined in the stimulus dataset) to be used in the subsequent recognition task (described below), the 50 images included a maximum of two images from the same semantic category, so that a third image from the same semantic category could be used as a foil. For Experiment 2, participants were randomly assigned to one of two conditions (layer 1 or layer 8). Participants in Experiment 3 were presented with all house images.

Table 1. Percentage of overlapping images

	Similar				Discriminable			
	Layer 1	Layer 3	Layer 5	Layer 8	Layer 1	Layer 3	Layer 5	Layer 8
Layer 1	100%	0%	0%	0%	100%	0%	0%	0%
Layer 3	0%	100%	46%	16%	0%	100%	72%	40%
Layer 5	0%	46%	100%	14%	0%	72%	100%	44%
Layer 8	0%	16%	14%	100%	0%	40%	44%	100%

Note. Values represent the number of images overlapping between layers (as a percentage of the total number of possible images). The table shows the percentage of overlapping images within those images categorized as *similar*, and within those images categorized as *discriminable* between each pair of layers.

Table 2. Percentage of mixed overlapping images

		Discriminable			
		Layer 1	Layer 3	Layer 5	Layer 8
Similar	Layer 1	0%	32%	42%	34%
	Layer 3	26%	0%	0%	0%
	Layer 5	24%	0%	0%	0%
	Layer 8	12%	0%	0%	0%

Note. Values represent the number of images overlapping between layers (as a percentage of the total number of possible images) in different *similar/discriminable* categorizations. As an example, the value representing 26% indicates that 26% of the images categorized as *discriminable* within layer 3 were categorized as *similar* within layer 1. In contrast, the value representing 32% indicates that 32% of the images categorized as *similar* within layer 3 were categorized as *discriminable* within layer 1.

Paradigm

Experiment 1 consisted of three key phases: initial encoding, free recall, and final recognition test. During the encoding phase, participants were presented with each of the 50 similar and 50 discriminable images of scenes from the corresponding layer (intermixed in a random order). Participants were asked to indicate whether the image was indoor or outdoor. Images remained onscreen for 4 seconds (s) regardless of a participant's response, to allow equal encoding time across all images. A 2 s inter-trial interval followed each presented image. Upon completion of the encoding phase, participants played a game of Tetris for five minutes to

prevent visual rehearsal. After Tetris, the free recall phase consisted of participants describing as many scenes as they could remember by typing as much detail as possible (not analyzed in this paper). Lastly, during a surprise recognition memory test, participants judged whether an image was old (seen previously in the experiment) or new (not seen in the experiment). The stimuli included the 100 previously seen images and 100 novel foils randomly drawn from the same semantic category as the old images (e.g., one igloo scene foil if an igloo scene was initially presented). The recognition test images were shown in a random order, and remained onscreen until participants responded (maximum 4 s). A 2 s inter-trial interval followed each image. Upon completion of this final recognition memory test, participants were debriefed about the purposes of the experiment.

Experiment 2 followed the same procedures as described above, except that participants did not complete a free recall phase following the game of Tetris, and instead moved directly to the surprise recognition memory test, in order to ensure that the Experiment 1 recognition memory effects were not being affected by the prior free recall test.

Experiment 3 (originally run to test an unrelated question) consisted of two similar key phases as Experiments 1 and 2: initial encoding and final recognition memory test. During the encoding phase, participants were presented with 60 images of houses (in a random order) and were asked to appraise the price of the house. Images remained onscreen until participants submitted an appraisal, at which point they were provided with pseudo-random feedback about the “true” appraisal of the house (randomly selected from a distribution of values). This sequence continued until participants had viewed and appraised all 60 house images. Upon completion of the encoding phase, participants solved basic arithmetic problems for five minutes to clear their working memory. Participants were then given a surprise recognition memory test, in which they

judged whether 120 images were old (the 60 seen previously in the experiment) or new (60 not previously seen) and indicated the appraisal presented during encoding. To avoid an incentive to indicate that houses were not seen (which would shorten the session), participants were also instructed to estimate the appraisal value for houses indicated as 'new'. All images were presented in a random order. Upon completion of the final recognition memory test, participants were debriefed about the purposes of the experiment.

Analyses

For Experiments 1 and 2, we first calculated participants' accuracy during the initial encoding phase task (indoor vs. outdoor). Data from four participants from Experiment 1, and three from Experiment 2, were not analyzed further due to having accuracy scores that were more than two standard deviations below the mean of the full group. Behavioral results are reported based on signal detection theory implemented through logistic mixed effects regression models (Baayen et al., 2008). In each regression model, the dependent variable was the participant's judgment as to whether or not they had previously seen the image during the encoding phase. We included fixed effects terms for image type (i.e., whether or not the image was shown during the encoding phase, and if shown, whether it was in the top 50 most similar or top 50 most discriminable for that layer), as well as the participant's group (i.e., from which layer of the CNN the images were drawn). Additionally, a variable for participant was included as a random effect. Trials with no response were removed prior to conducting the regression models. We report unstandardized coefficient estimates (B) in logits for models with categorical predictors and standardized coefficient estimates (β) for models with continuous predictors, as well as odds and 95% confidence intervals (on the odds) as a measure of effect size.

In Experiment 3, a sense-check was first performed to ensure that participants were engaging with the task, which included giving a possible value for each house. An independent rater (blind to participant responses) selected the three most expensive houses and three cheapest houses based on appearance from those presented during the encoding phase. We compared the estimated appraisals offered by each subject to these expensive and inexpensive houses. Participants were removed if they estimated that the cheapest houses were more expensive than the most expensive houses, as this suggested they did not understand or follow the task directions (or were merely guessing randomly). This excluded seven participants' data from further analysis. The key analyses of Experiment 3 followed a similar analysis plan as Experiments 1 and 2, with logistic mixed effects regression models. Due to only having one group of participants (instead of the four in Experiment 1 and two in Experiment 2), regression models were used to compare recognition memory for images based on similarity values taken from each of the four examined CNN layers. In each regression model, the dependent variable was the participant's judgment of whether or not they had previously seen the image during the encoding phase. We included fixed effects terms for image type (i.e., whether or not the image was shown previously) and a continuous predictor of the image's similarity values in the relevant layer. Participant was included as a random effect within the models. We report standardized coefficient estimates (β), as well as odds and 95% confidence intervals (on the odds) as a measure of effect size.

Acknowledgements

G.E.K. was supported by the Behavioral Brain Research Training Program from the National Institutes of Health (grant number T32GM081760). The authors thank Brandon Carlos and Cory Derringer for their help with preliminary experiment set up, as well as Uma Balaji, Sharran Chakravorty, Amber Shergill, Abigail Swire, Swathi Tata, Carlo Vignali, Dan Volpone, Aarya Wadke, Grace Waldow, and Joanna Wang for their assistance in data collection. The authors also thank Scott Fraundorf for advice on statistical analyses.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage*, *149*, 141–152. https://doi.org/10.1016/j.neuroimage.2017.01.063

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-26467-5

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*(3), 219–228. https://doi.org/10.3758/BF03197669

Bruce, V., Burton, A. M., & Dench, N. (1994). What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *47*(1), 119–141.

Busey, T. A. (2001). Formal models of familiarity and memorability in face recognition. In *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 147–191). Lawrence Erlbaum Associates Publishers.

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(6), 1403–1414.

Coutanche, M. N., Solomon, S. H., & Thompson-Schill, S. L. (2016). A meta-analysis of fMRI decoding: Quantifying influences on human visual population codes. *Neuropsychologia*, *82*, 134–141. https://doi.org/10.1016/j.neuropsychologia.2016.01.018

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *ImageNet: A Large-Scale Hierarchical Image Database*. 8.

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What Makes an Object Memorable? *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 1089–1097. https://doi.org/10.1109/ICCV.2015.130

Epstein, R. A., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annual Review of Vision Science*, *5*(1), null. https://doi.org/10.1146/annurev-vision-091718-014809

Huebner, G. M., & Gegenfurtner, K. R. (2012). Conceptual and visual features contribute to visual memory for natural images. *PloS One*, *7*(6), e37575. https://doi.org/10.1371/journal.pone.0037575

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and Predicting Image Memorability at a Large Scale. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2390–2398. https://doi.org/10.1109/ICCV.2015.275

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Conceptual Distinctiveness Supports Detailed Visual Long-Term Memory for Real-World Objects. *Journal of Experimental Psychology. General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science*, *21*(11), 1551–1556. https://doi.org/10.1177/0956797610385359

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–

1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lin, Q., Yousif, S., Scholl, B. J., & Chun, M. M. (2018). *Visual memorability in the absence of semantic content*. https://doi.org/10.1167/18.10.1302

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, *7*, e38105. https://doi.org/10.7554/eLife.38105

Lukavský, J., & Děchtěrenko, F. (2017). Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception, & Psychophysics*, *79*(7), 2044–2054. https://doi.org/10.3758/s13414-017-1375-9

Vedaldi, A., & Lenc, K. (2014). MatConvNet—Convolutional Neural Networks for MATLAB. *ArXiv:1412.4564 [Cs]*. http://arxiv.org/abs/1412.4564