Image memorability is predicted by activity across different stages of convolutional neural networks and the human ventral stream

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: memory, convolutional neural networks, scenes, recognition, memorability

Conflict of interest statement:
The authors declare no competing financial interests.

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 study set up, as well as Uma Balaji, Sharran Chakravorty, Amber Shergill, Abigail Swire, Swathi Tata, Carlo Vignali, and Aarya Wadke for their assistance in data collection. The authors also thank Scott Fraundorf for providing a consultation on statistical analyses.

Abstract

What makes some images memorable while others are forgettable? The features of an image can be represented at multiple levels – from low-level visual properties to high-level meaning. Across two behavioral studies and a neuroimaging study, we addressed the question of how image memorability is influenced by different levels of the visual hierarchy. In a first behavioral study, we combined a convolutional neural network (CNN) with behavioral prospective assignment, by using one of 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 that were identified as discriminable using low-level CNN layers, or identified as similar in high-level layers. A second study replicated the first experiment's results using images from a single semantic category (houses), but found that similarity predicted memorability at a slightly less high-level that holds representations of objects, suggesting this level is more important for remembering images from the same category. Finally, we analyzed neural activity collected through functional magnetic resonance imaging (fMRI) scans as independent participants viewed the same scene images. Pattern similarity analyses revealed an analogous relationship in the ventral stream between image discriminability/similarity and level of the visual hierarchy. Discriminability in early visual areas, and similarity later in the ventral stream, each predicted greater image memorability. Together, this research shows that discriminability at different visual levels can be used to predict image memorability through both CNN models and neural activity in the human ventral stream.

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. What makes certain places or pictures more likely to be remembered than others?

The images that are more likely to be remembered than others (Isola, Xiao, Parikh, Torralba, & Oliva, 2014) 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, Dilks, & Oliva, 2017; Dubey, Peterson, Khosla, Yang, & Ghanem, 2015; Isola et al., 2014), as do participants' subjective predictions (Isola et al., 2014). Images that are distinctive have been shown to be particularly memorable (Bartlett, Hurry, & Thorley, 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, Yousif, Scholl, & Chun, 2018) though the nature of predictive low-level and high-level content remains unclear. Determining why, when, and how different factors influence image memorability is a necessary step for modelling the relationship between visual perception and memory, and to enable applications such as selecting memorable health-related or educational images.

A recent tool that is increasingly used to characterize images is a convolutional neural network (CNN) trained for object recognition (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009). The hierarchical organization of these trained multi-layer models has similarities with the human visual system (Kriegeskorte, 2015; Lindsay & Miller, 2018), in which visual information is represented at progressively higher stages (Coutanche, Solomon, & Thompson-Schill, 2016). In analyzing an image, early CNN layers extract basic visual properties, which become increasingly high-level, until ultimately classifying the image (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015). Although primarily employed in the vision sciences, the memory field also has much to gain from using these models.

Here, we leverage the multiple layers of a CNN to determine why certain images are more memorable than others, at multiple visual stages. Uniquely, we use a CNN model in a prospective assignment, in which levels of a CNN are used to present images to participants. In Study 1, we presented 100 participants with images of natural scenes (featuring a variety of different objects) that were selected based on the images' similarity and discriminability from one of four stages of a trained CNN (i.e., one of four groups). A surprise recognition memory test (with matched foils) showed that images that were distinguishable in the earliest layer of the CNN (i.e., edges and blobs) were more likely to be remembered than those that were most similar. In contrast, images that had the highest similarity in the last layer (i.e., semantic categorization) were more likely to be remembered. Study 2 repeated the above CNN layer analysis in an independent group of participants that viewed exemplars from the same semantic category (houses). Once again, images that were the most discriminable at the earliest stage of the CNN were more likely to be remembered, replicating the first study. For this set of images within the same category, images that were more similar in a layer preceding the final layer were more likely to be remembered. Finally, study 3 analyzes how neural activity patterns recorded as separate participants viewed the same images as used in Study 1 relate to their memorability. We observed a gradient that paralleled our above behavioral results – with higher pattern discriminability in the early visual system, and greater pattern similarity in its later stages. These findings reveal the multiple sources of memorability that are present for any given image, which have corresponding signatures in the patterns of trained CNNs and the human brain.

Materials and Methods

Participants

In Study 1, participants were recruited until 100 contributed usable data (25 in each condition), in line with prior research investigating recognition memory for scenes (Konkle, Brady, Alvarez, & Oliva, 2010b, 2010a). 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). The remaining 100 participants' data were included in all analyses and results.

For Study 2, thirty-two participants (15 females, 17 males, *M* age = 19.7 years, *SD* = 1.2 years) were included in the analyses. 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 studies 1 and 2. Participants in studies 1 and 2 were compensated through course credit for their participation.

The images in Study 1 were taken from the BOLD5000 dataset because this resource also provides functional magnetic resonance imaging (fMRI) data collected as participants viewed the images (Chang et al., 2019). The dataset includes data collected from four participants as they viewed the large number of images, allowing for item-wise analyses within each individual. Two of the available subjects were not analyzed because one only had 9 functional sessions collected instead of the full 15, and another showed very low temporal signal-to-noise (tSNR) values in anterior regions of the ventral stream (Binder et al., 2011). The two analyzed subjects are designated here as Subject A (female; age = 26) and Subject B (female; age = 24).

Stimuli and Materials

Stimuli for Study 1 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 Study 2, 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 generally centered within the image.

Procedure

CNN metrics

Across all studies, 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 had been trained using more 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 by conducting pairwise correlations between their sets of feature weights, giving a 1,000 x 1,000 matrix of image similarity for each layer in Study 1, and a 120 x 120 matrix of image similarity for each layer in Study 2. 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 study, according to each of the CNN stages.

Participant assignment

For Study 1, participants were randomly assigned to one of four conditions. Participants in each condition were presented with images selected based on the corresponding CNN layer (layer 1, 3, 5, and 8). Condition assignment thus dictated which set of images would be presented to a participant. The above CNN metrics of image similarity/discriminability (in each layer) was 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. 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.

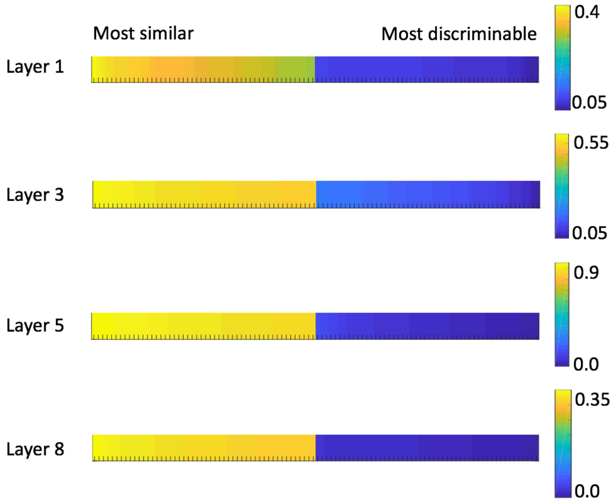

Figure 1. Range of Fisher-*Z* transformed *r*-values for the 100 images (depicted by tick marks along color spectrum) in each of four conditions. Color bars depict similarity values for each layer.

Because the Study 2 dataset was collected prior to calculating CNN metrics, the similarity/discriminability metrics of this study were calculated retrospectively, and related to performance in a continuous manner.

Paradigm

Study 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 judged whether the image was indoor or outdoor. Images remained onscreen for 4 seconds (s) regardless of participants' 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 final recognition memory test, participants judged whether an image was old (seen previously in the study) or new (not seen in the study). 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 study.

Study 2 consisted of two similar key phases as Study 1: 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 study) 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 otherwise shorten the session), participants were also instructed to estimate the appraisal value for new houses. All images were presented in a random order. Upon completion of the final recognition memory test, participants were debriefed about the purposes of the study.

In study 3, the BOLD5000 fMRI dataset was collected as participants viewed images from 250 categories corresponding to SUN dataset categories (Chang et al., 2019). Each participant was scanned across 15 separate functional imaging sessions. Eight sessions included 9 imaging runs, and 7 sessions included 10 imaging runs, with 37 images viewed in a randomized order during each run. Participants were scanned using a 3T Siemens Verio MR scanner that used a 32-channel phased array head coil. Each participant underwent a T1-weighted anatomical scan (TR = 2300 ms, TE = 1.97 ms, 1.00 mm isovoxel resolution) and T2-weighted functional scans (TR = 2000 ms, TE = 30 ms, voxel size = 2.00 mm isovoxel resolution). A localizer was used to functionally define early visual regions for each subject (for full details, see Chang et al., 2019).

Analyses

Study 1

For Study 1, we first calculated participants' accuracy during the initial encoding phase task (indoor vs. outdoor). Four participants' data 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, Davidson, & Bates, 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 ratios and 95% confidence intervals (on the odds) as a measure of effect size.

Study 2

Prior to the key analyses of Study 2, a sense-check was performed to ensure that participants were engaging with the task, which included giving a possible value for each house. To check for task engagement, 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 on these expensive and inexpensive houses. Participants who estimated that the cheaper houses were more expensive than the more expensive houses were removed, as it suggested they did not understand or follow the task directions (or were merely guessing random values). This excluded seven participants' data from further analysis.

The key analyses of Study 2 followed a similar analysis plan as Study 1, again performing logistic mixed effects regression models. Due to only having one group of participants (instead of the four in Study 1), regression models were used to compare recognition memory for images based on similarity values take from each of the four examined CNN layers. 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 previously) and a continuous predictor of the image's similarity values in the relevant layer (because these images were not prospectively selected into condition, as in Study 1). Participant was included as a random effect within the models. We report standardized coefficient estimates (β), as well as odds ratios and 95% confidence intervals (on the odds) as a measure of effect size.

Study 3

The fMRI data downloaded from the BOLD5000 online repository is preprocessed to the specifications outlined in the original study (Chang et al., 2019). Each subject's cortical parcellated Destrieux Atlas was defined using automated Freesurfer segmentation (Fischl et al., 2002; Fischl et al., 2004). We examined activity patterns along each individual's ventral stream by anatomically defining regions progressing from the occipital pole to anterior temporal lobe – occipitotemporal cortex (including the collateral sulcus and fusiform gyrus), ventral temporal cortex, parahippocampal gyrus, and the calcarine sulcus, as well as the functionally defined early visual regions that were acquired via localizer (Bressler et al. 2013).

In order to analyze how activity pattern discriminability changed along the posterior – anterior dimension of the ventral stream, we used a 3 voxel-radius searchlight (Kriegeskorte, Goebel, & Bandettini, 2006), giving 5,655 and 5,843 searchlights for the two subjects. The center of each searchlight was indexed by its coordinates (i.e., coronal slice) for analysis of its anterior – posterior location. For each searchlight, the activity patterns (vectors of beta coefficients) underlying each of the 1,000 scenes were subjected to the same correlation procedure that was used on the CNN features in Study 1, resulting in a single mean correlation coefficient that reflected the similarity (or discriminability) of the activity patterns for each image within that searchlight region. Next, as in Study 1, we applied binary logistic mixed effects regression models to predict subjects' judgments as to whether or not they had previously seen an image during the encoding phase, using the activity pattern similarity measure as a continuous predictor. The resulting standardized coefficient estimates (β) from the regression reflected the relationship between activity pattern similarity and image memorability in the 100 behavioral subjects from Study 1 (who observed the same images) for each searchlight.

In addition to the above analyses, we examined temporal signal-to-noise ratio (tSNR) to ensure that regions examined in the planned analyses had sufficient tSNR. Following prior work (Binder et al., 2011), we set a minimum tSNR threshold of being greater than 20 in at least 85% of a region's voxels. First, concerned with the susceptibility of the anterior portion of the temporal lobe to signal loss, we calculated tSNR in the anterior half of the ventral stream. Two of the three available subjects passed the 85% threshold in this area (Subject A: 87.1%; Subject B: 89.9%), but a third's tSNR was significantly below this threshold (with only 63.8% of voxels with a tSNR greater than 20). We did not analyze this subject further, because such significant signal loss precluded an informative analysis of the posterior - anterior dimension of the ventral stream. Within the two analyzed subjects, the vast majority of searchlights (98.41% and 98.37%) met the required tSNR criteria (85% of searchlight voxels showing a tSNR > 20). Searchlights not meeting this criterion (predominantly in the most anterior temporal lobe) were censored from analyses. After applying the tSNR threshold, 98.41% and 98.37% of searchlights met the required tSNR criteria. Searchlights not meeting the threshold were focused predominantly in the anterior temporal lobe.

Results

We investigated how discriminability of images at different stages of the visual hierarchy (through a CNN in studies 1 and 2, and in the human ventral stream in study 3) influence memorability.

Study 1

We first assessed participants' ability to remember the presented images compared to novel foils. On average, previously presented images were 18.26 times more likely to be judged as having been seen, than were the matched foils (B = 2.90, *p* < .001, 95% confidence interval in odds [16.94, 19.69]). Our question of interest was how memory for previously presented images would differ based on the CNN layer used to select them. A test for polynomial effects across all 100 participants to predict memory performance across the four conditions revealed a significant linear interaction between similarity of the images and layer condition (B = 0.34, odds = 1.41, *p* < .001, [1.17, 1.69]). This interaction between similarity and layer was key to detecting the relationship, as rhere was no main effect of images categorized as similar versus discriminable when the layers were collapsed (B = 0.01, odds = 1.01, *p* = .795, [0.92, 1.11]). Neither a quadratic nor cubic function fit the data better than linear (*p*s > .193).

Each of the four conditions were then separately examined using individual regression models. Within the earliest layer (layer 1), images that were more similar in the respective layer were *less* likely to be correctly recognized as seen before than were images that were more discriminable (B = -0.28, odds = 0.76, *p* = .003, [0.63, 0.91]; Figure 2). For ease of interpretability, this means that images categorized as discriminable were 1.32 times more likely to be correctly recognized as having been seen before, than images that were categorized as similar. Within the last layer (layer 8), images that were categorized as similar were 1.27 times *more* likely to be correctly recognized as having been seen before than images that were categorized as discriminable (B = 0.24, *p* = .010, [1.06, 1.52]). No differences were observed between similar and discriminable images in the middle two layers (3 and 5) (*p*s > .609).

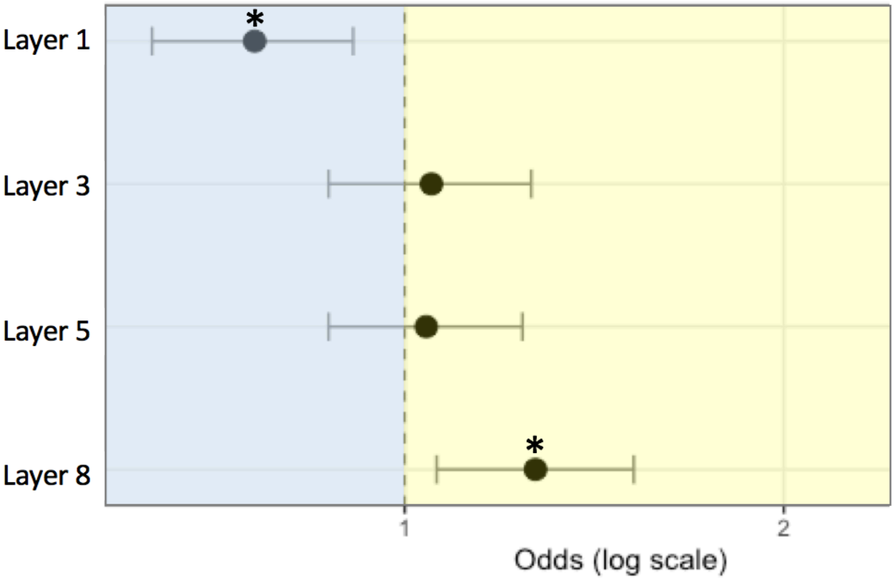

Figure 2. Odds of participants correctly judging an image as having been previously presented, based on similarity or discriminability across four conditions. Odds in blue reflect better memory performance for discriminable images (compared to similar images) and odds in yellow reflect better memory performance for similar images (compared to discriminable images). Error bars reflect 95% confidence interval on the odds. * represents statistical significance ($p < .05$).

We also implemented regression models with continuous predictors of image similarity (instead of categorical predictors). For every one standard deviation decrease in similarity (Fisher-*Z* *r*-values) based on layer 1, an image was *more* likely to be correctly recognized as seen before (β = -0.14, odds = 0.87, $p = .004$, [0.79, 0.96]). In other words, for each one standard deviation decrease in similarity, the odds of judging an image as seen before were 15% greater. In the model using features from layer 8, for every one standard deviation increase in similarity, an image was 1.12 times more likely to be correctly recognized as seen before (β = 0.11, $p = .016$, [1.02, 1.22]). No differences were observed for the middle two layers (3 and 5) (*p*s > .669).

Study 2

Study 2 examined recognition performance of a large set of images within the same semantic category (houses) in an independent set of participants. Replicating the results of Study 1, greater discriminability in the earliest layer (1) predicted superior subsequent memory performance (β = -0.17, odds = 0.84, $p < .001$, [0.76, 0.93]). For every one standard deviation decrease in similarity, the odds of correctly judging an image as seen before were 19% greater. On the other hand, greater similarity within layer 5 predicted stronger subsequent memory performance for an image: for every one standard deviation increase in similarity, an image was

1.16 times more likely to be correctly recognized as old (β = 0.15, *p* = .003, [1.05, 1.28]). No differences were observed for either layer 3 or layer 8 (*p*s > .163).

Study 3

We next tested for a relationship between memorability and activity pattern discriminability (similarity) along the human ventral stream in a model that matched the CNN analyses above (but with voxels taking the place of CNN features). Regression models predicted image memorability (calculated using behavioral judgments from participants in Study 1) based on pattern similarity in searchlights of two independent subjects. Searchlights were indexed by their posterior-to-anterior location. In an item-analysis within searchlights of both examined subjects, as searchlights progressed in an anterior direction, the odds ratio of a participant in Study 1 judging an image as having been seen before was predicted by greater neural pattern similarity (Subject A: B = .00039, *p* < .001; Subject B: B = .00315, *p* < .001; Figures 3 and 4). The increase in the odds ratios reflect a shift from greater discriminability predicting memorability, to greater similarity predicting memorability, along the ventral stream.

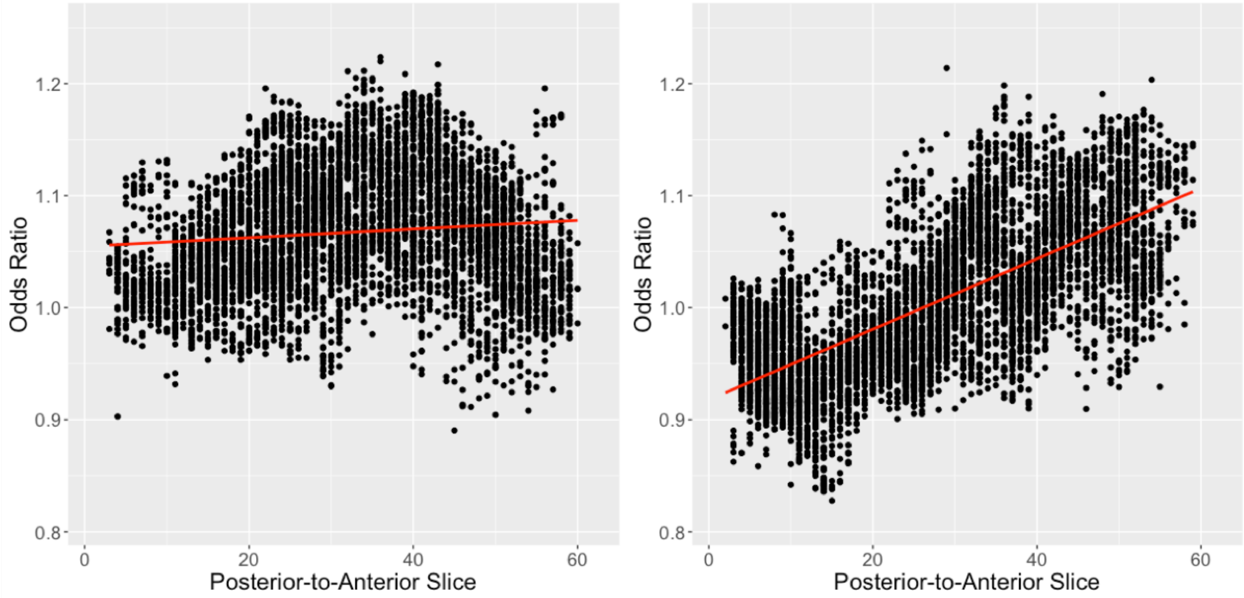

Figure 3. Odds ratios of Subject A (left) and Subject B (right) reflecting image memorability progressing from posterior to anterior regions across the visual stream. Odds ratios above 1 reflect a better memory performance for images with more similar patterns of neural activity and odds ratios below 1 reflect better memory performance for images with more discriminable patterns of neural activity. The red line in each figure represents the fitted linear regression line.

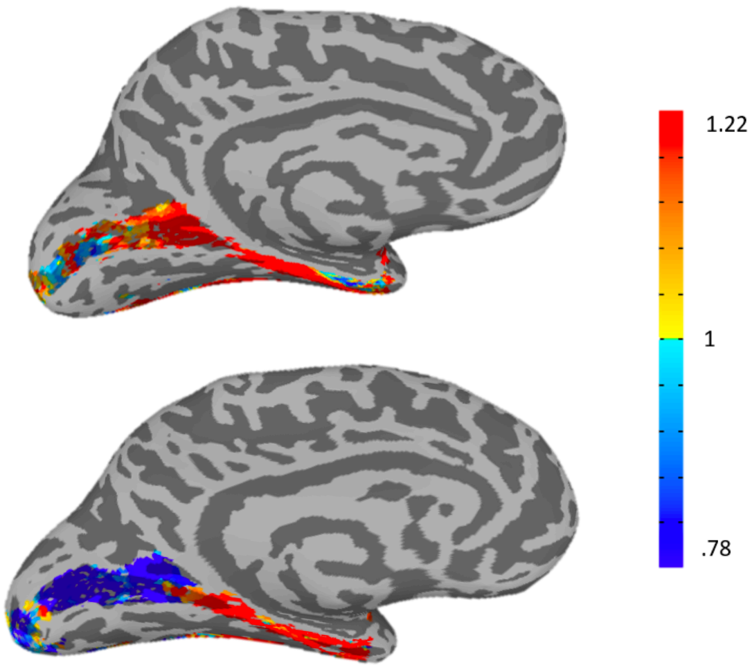

Figure 4. Odds ratios for searchlight pattern similarity in two scanned subjects predicting image memorability based on an independent behavioral group (from Study 1). Odds ratios are displayed at the center of each searchlight for Subject A (top) and Subject B (bottom). Odds ratios above 1 reflect better memory performance for images with more similar patterns of neural activity, whereas odds ratios below 1 reflect better memory performance for images with more discriminable patterns of neural activity.

Discussion

We have investigated how image discriminability across the visual hierarchy differentially predicts the likelihood that an image will be remembered by an observer. In Study 1, we found –in a prospective assignment paradigm– that participants remembered more scene images if they were selected based on high discriminability in low-level visual properties (earliest CNN layer), or high similarity in higher-order properties (final CNN layer). In Study 2, we examined how CNN layers predict memorability for a set of images from the same semantic category (houses). These results replicated the importance of discriminability in the earliest CNN layer, with memorability associated with greater similarity at a mid-high level stage (layer 5). Finally, Study 3 conducted an item-wise analysis of pattern similarity in searchlights along the ventral stream for two independent participants who viewed the same images of Study 1. The imaging findings paralleled the shift from greater discriminability to similarity in the CNN behavioral findings of Studies 1 and 2. A positive trajectory was observed along the ventral stream, with an increasingly positive relationship between activity-pattern similarity and memorability, even when memorability was determined from independent subjects. These neural results support the existence of a gradient in the human visual system for the link between neural discriminability and memorability.

Our findings that similarity and discriminability can support memorability at different levels of the visual hierarchy –in CNN models and data from the human ventral stream– help to reconcile several seemingly conflicting findings within the field. For instance, in some prior research, image memorability has been associated with the presence of discriminable features (Bartlett et al., 1984; Bruce, Burton, & Dench, 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 two behavioral studies is the different CNN layers that had a similarity relationship with memorability (layer 8 in Study 1, and layer 5 in Study 2). Notably, a likely reason for this difference is the inclusion of scene images from a variety of semantic categories in Study 1, but just one category (houses) for Study 2. The final CNN layer (8), which classifies images, 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 influence of image similarity can occur at different levels of the visual hierarchy– highlights the critical factor of the kind (and level) of features that distinguish a remembered image from others in a particular set.

An aspect of the design that is worth highlighting is the prospective assignment performed in Study 1. Typically, studies of memorability test how memory for a large set of intermixed images varies with various 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. Interference between presented items can detrimentally affect retrieval performance (Ciranni & Shimamura, 1999), and perceptual and conceptual image properties are known to affect responses to the targets of memory tests (Huebner & Gegenfurtner, 2012). Prospectively assigning participants to a condition that only includes images selected from one of the four layers ensured that other layers could not be influencing the observed group differences. Future memory 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.

One limitation of Study 3 is that item analyses were conducted in two participants, based on the design of the employed open dataset in which a small number of participants were shown a very large number of scenes many times, across multiple sessions. This design is optimal for within-subject item analyses, and is consistent with a number of fMRI studies of vision that have also used high-powered studies of several subjects (e.g., Kamitani & Tong, 2005, 2006; Naselaris et al., 2015). Our own analysis compared activity of these subjects' visual systems with memorability in 100 behavioral participants, so that the small number of subjects impacts the visual system results, rather than memorability per se, but it is nonetheless a limitation that could be addressed in further studies.

To summarize, we find that high image discriminability at early visual levels, and high similarity at later visual levels, predict image memorability in CNNs and the human visual system. A prospective assignment approach demonstrated the ability to select images for greater memorability based on CNN metrics. Differences in the critical visual levels for images of varied scenes versus images from the same semantic category, revealed that the variability across images plays a key role in which visual stages contain relevant metrics for memorability.

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

Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., … Weaver, K. E. (2011). Mapping Anterior Temporal Lobe Language Areas with FMRI: A Multi-Center Normative Study. *NeuroImage*, *54*(2), 1465–1475. https://doi.org/10.1016/j.neuroimage.2010.09.048

Bressler DW, Fortenbaugh FC, Robertson LC, Silver MA (2013) Visual spatial attention enhances the amplitude of positive and negative fMRI responses to visual stimulation in an eccentricity-dependent manner. Vision Res 85:104-112, doi:10.1016/j.visres.2013.03.009, pmid:23562388.

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 *Scientific Psychology Series*. *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 147–191). Mahwah, NJ, US: 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. In *CVPR*.

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

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. Neuron, 33(3), 341–355.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., … Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, 14(1), 11–22.

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

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kamitani, Y., & Tong, F. (2006). Decoding Seen and Attended Motion Directions from Activity in the Human Visual Cortex. *Current Biology*, *16*(11), 1096–1102. https://doi.org/10.1016/j.cub.2006.04.003

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

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103\

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).

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

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38). doi: 10.1073/pnas.1719616115

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

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, *105*, 215–228. https://doi.org/10.1016/j.neuroimage.2014.10.018

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