Understanding Image Memorability

Nicole C. Rust1,* and Vahid Mehrpour1

Why are some images easier to remember than others? Here, we review recent developments in our understanding of ‘image memorability’, including its behavioral characteristics, its neural correlates, and the optimization principles from which it originates. We highlight work that has used large behavioral data sets to leverage memorability scores computed for individual images. These studies demonstrate that the mapping of image content to image memorability is not only predictable, but also non-intuitive and multifaceted. This work has also led to insights into the neural correlates of image memorability, by way of the discovery of a type of population response magnitude variation that emerges in high-level visual cortex as well as higher stages of deep neural networks trained to categorize objects.

The Innovation: Quantifying Memorability for Individual Images

In a now classic paper, 'Learning 10,000 pictures', Lionel Standing [1] documented the remarkable ability that humans have to remember whether they’ve seen an image before, even after seeing thousands of images, each only once and only for a few seconds. Standing also determined that there is no indication that this form of ‘image recognition memory’ saturates as a function of the number of images viewed, at least up to 10,000 images. Nearly 50 years later, Standing’s results remain robust [2] but yet we still understand little about how our brains manage to remember images so well. However, over the past handful of years, rapid advances have been made toward understanding an issue that brings us one step closer to understanding of image recognition memory: ‘image memorability’, or the systematic variation with which some images are better remembered than others. Progress in understanding image memorability has dovetailed with progresses in object and scene identification [3,4], and similar to that work, has leveraged advances in the training of deep neural networks [5,6], the acquisition of large human behavioral data sets [5,7,8], and population-based approaches for recording neural activity at single-unit resolution [6].

Many of the recent advances in our understanding of what drives image memorability variation can be traced back to the development that memorability can be reliably quantified for individual images [7–9] (Box 1). This development facilitated a better understanding of memorability in multiple ways. First, compared with foundational studies that focused on how one or a few factors impacted image memorability (e.g., image coloration [10] or object distinctiveness [11–13]), image memorability scores for a set of images provided a measure of total memorability variation that could be understood as the weighted combination of several factors (described in detail later) [5,7,8]. This in turn led to an appreciation that the factors that determine image memorability are, naïvely, not intuitive [7], and that a considerable amount of memorability variation remained (and still remains) unexplained [5,8]. Second, despite an incomplete understanding of how image content determines image memorability, the existence of image memorability scores allowed for the pursuit of the representational correlates of image memorability in the brain [6,14–16], as well as neural networks trained to categorize objects [5–7] (Box 2). These investigations led to the discovery of a population response magnitude coding scheme that emerges at higher stages of the visual hierarchy and is predictive of how well images will be remembered [6].

1Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

*Correspondence: nrust@psych.upenn.edu (N.C. Rust).

https://doi.org/10.1016/j.tics.2020.04.001
The Behavioral Characteristics of Image Memorability

The Factors that Drive Image Memorability Are Predictable but not Intuitive

Image memorability is typically investigated in the context of a recognition memory task (“Have you seen this image before?”), quantified as described in Box 1. In developing the techniques...
to measure memorability, researchers were careful to consider the sources of memorability variation. If what makes an image more or less memorable is dominated by individual differences, measures of subject-averaged memorability performance will fail to capture it. However, if a large fraction of memorability variation is shared across different individuals, memorability can be meaningfully captured at the resolution of individual images through subject-averaged measures of memory behavior. Consistent with the latter notion, memorability scores (Box 1) are highly correlated across random splits of large subject pools (average Spearman rank correlations: $\rho = 0.68$ for faces [8,17] and $\rho = 0.68$–0.78 for images that contain faces as well as other content [5,7,18]), thereby establishing that image memorability scores do in fact capture reliable properties of individual images. Remarkably, memorability variation is not only consistent across different human subjects, but is also correlated between humans and rhesus monkeys [6], suggesting that memorability does not depend on extensive experience with the objects contained in images (such as cars and fire hydrants) and it does not require the capacity for language.

One striking finding is that, naïvely, subjects are bad at predicting how memorable images are: when untrained subjects were asked to predict memorability, their predictions and actual memorability scores were at best weakly correlated [7,8,17]. Rather, subjects’ memorability predictions were strongly correlated with judgments of image esthetics and interestingness, which were each only weakly correlated with memorability (but were strongly correlated with one another) [7]. Examples of this are shown in Figure 1A, including one image that was predicted to be forgettable but had a high memorability score (Figure 1A, top) and another image that was predicted to be memorable but had a low memorability score (Figure 1A, bottom). This suggests that we have misguided notions about what makes images memorable. However, that is not to say that image memorability is not predictable: a convolutional neural network trained to predict image memorability (‘MemNet’ [5]), predicts memorability scores for untrained images near the noise ceiling present in the human behavioral data (Spearman’s rank correlations between MemNet predictions and human behavior $\rho = 0.64$; Spearman’s rank correlations between random splits of the human subject pool for the same image set $\rho = 0.68$ [5]). This confirms that, while image memorability is counterintuitive for humans, it is predictable from image pixel patterns (Box 2).

![Figure 1. Examples of Memorable and Forgettable Images.](image)

Image memorability scores are labeled to the left of each image. (A) Examples illustrating that naïve untrained subjects have misguided notions about image memorability. Top: example of an image that was predicted to be forgettable by naïve subjects but that had a high memorability score. Bottom: example of an image that was predicted to be memorable by naïve subjects but that had a low memorability score. (B) Images containing people tend to be highly memorable, whereas nature scenes tend to have low memorability. (C) Atypical depictions of objects tend to be more memorable than typical depictions. (D) Two example images produced by GANalyze from the same seed image, with enhanced and reduced memorability. Images from [7] (A) and [28] (D).
The Mapping of Image Content to Image Memorability Is Multifaceted

Several foundational studies published before the acquisition of memorability scores for individual images documented a number of different types of image content that impact image memorability. These include lower-level image properties, such as the presentation of images in color as opposed to grayscale [10] and viewing images in 3D as opposed to 2D [19]. These also include higher-level image properties, such as distinctiveness and atypicality, particularly for faces [11–13]. In addition, these earlier studies established that conceptual distinctiveness (such as category membership) impacts memorability in a manner that cannot be accounted for by perceptual distinctiveness alone [20,21]. At the same time, this earlier work emphasized the rich details with which visual memories are stored, including the fact that we remember considerable detail about the configurations and contexts that we view objects in [2] and that we rely on specific detailed information to remember whether we have seen an image previously [22]. However, we are typically bad at remembering random patterns unless they take on object-like qualities [23], suggesting that visual memory is not driven entirely by visual details. These results have been summarized as ‘meaningfulness’ contributing to memorability, because images that are meaningful insofar as they contain recognizable content are better remembered than those that do not [24].

The quantification of memorability scores for individual images built on these foundational results to determine the relative importance of different known factors, discover new factors, and determine the fraction of total image memorability variation that all factors together could explain, once combined. Examples of the largest factors that drive image memorability variation include the fact that images containing people are on average highly memorable (average memorability score = 0.82), in contrast to images of nature scenes, which have lower average memorability (average memorability score = 0.61) [7] (Figure 1B). Images of atypical content, such as a chair shaped like a hand [25], are also typically highly memorable (average memorability score = 0.83) [5] (Figure 1C). The emotional valence of an image also impacts its memorability, where images that evoke disgust, amusement, and fear are on average more memorable, whereas images that evoke awe and contentment are, on average, less memorable [5]. As described earlier, memorability is only weakly correlated with subjective judgments of image esthetics and interestingness [5]. Similarly, while low-level image properties such as color and simple image features do contribute to memorability variation, they only account weakly for it [7,26].

When considered together, how much memorability variation do all known factors account for? To address this question, one study scored and regressed 127 semantic attributes (e.g., open/closed; static/dynamic; frightening/funny, person/not, etc.) against image memorability scores and found that together they accounted for ~75% of explainable variance (where explainable variance refers to the variance preserved across random splits of human subjects [7]). This suggests that, at the same time that we can describe many of the principles that dictate image memorability, a considerable fraction of this variation remains uncharacterized.

When restricted to images of faces, similar principles to those established for images at large generally hold. In line with earlier literature [11–13], studies targeted at measuring the memorability of individual face images found that atypical faces tended to be more memorable than typical ones [8]. At the same time, the perceived distinctiveness of faces, measured as the combined influence of several terms (e.g., atypical, uncommon), could not fully explain face memorability [8,27]. Regressions of 20 attributes (e.g., interesting/boring, calm/aggressive, etc.) against image memorability scores revealed several additional attributes that enhance the memorability of faces, including faces that are judged as intelligent, responsible, trustworthy, attractive, and kind [8,17]. When considered together, scorings for all 20 attributes only accounted for ~75%
of explainable variance in image memorability scores [8], indicating that, similar to images at large, a considerable amount of image memorability variation for faces remains unexplained by these attributes. In addition, these studies determined that across different views of a face, memorability tends to be preserved [17].

Further insight into the relationship between image content and image memorability has been gained from a technological advance that combined the power of Generative Adversarial Networks (GANs) with a deep neural network trained to predict image memorability (MemNet [5], described earlier). This network, GANalyze, receives images as input and produces new images with minimal modifications to image content but parametrically manipulated memorability [28] (Figure 1D). Human behavioral experiments confirmed that the network successfully manipulates image memorability, as intended [28]. What image properties does the network manipulate to enhance image memorability? As might be expected, the network tends to make images brighter and more colorful. In addition, the network adjusts a handful of previously unappreciated factors, including: increasing object size, centering the objects within images, and uncluttering the backgrounds of the objects. Intriguingly, the network also tends to make objects more square or circular.

**Image Memorability for Recognition Memory versus Recollection**

Image memorability has been investigated most extensively by probing ‘recognition memory’: asking subjects to report whether images are novel or familiar (Box 1). A distinct but complementary memory task, investigated extensively for lists of words [29], is one that requires subjects to view images and then recall what they have seen, absent any cue. Are the same images that are most difficult to remember in a recognition memory task also the most difficult to recall? One recent study took on the challenge of quantifying memorability variation for image recollection by asking subjects to view images and then later draw them [30]. They found no relationship between image memorability variation quantified for recollection versus recognition memory [30]. These results are consistent with notions that image memorability variation for these two memory tasks may be distinct.

**Memorability and other Cognitive Phenomena**

Image memorability is both correlated with, and distinguishable from, several other cognitive phenomena. One of these is visual salience, which refers to the fact that, when we look at images, certain regions tend to pop out and grab our attention. The visual salience of an image can be quantified based on the regularity with which patterns of fixations during free viewing are consistent across subjects [31], and several studies have determined that measures of visual salience are correlated with image memorability [5,26,32–34]. Similar to memorability, images tend to be more salient when they contain one or a few objects [26,31], including images that are presented more close-up and in an uncluttered context [5]. However, when images contain multiple objects and multiple points of fixation, the correlation between memorability and salience drops considerably [26], suggesting that memorability and salience are distinguishable. Similarly, differences in memorability exist across face images that are identical in saliency in terms of their shapes, parts, image features, and fixation patterns [8,17]. Together, these results suggest that, while the factors that determine visual salience and memorability are correlated, they are not one and the same. By a similar logic, a recent study investigated the relationship between memorability and several cognitive factors, including manipulations of bottom-up attention (through spatial cuing and visual search), manipulations of top-down attention (through cognitive control and depth of encoding), and priming [35]. None of these factors were able to account for modulations of memorability, suggesting that memorability is distinct from these other phenomena.
Image Memorability Depends on Image Set Context
Image memorability scores are highly replicable when images are viewed in the context of a sequence of other, randomly selected images. When the same images are viewed in a sequence of images selected from the same category (e.g., when a picture of a lighthouse is embedded in a sequence of other lighthouse images), image memorability scores remain equally reliable across subjects, but they take on new values [33]. Contextual changes in the magnitude and sign of image memorability scores relative to the random benchmark can be predicted by how distinct an image is from other images within the new image set (quantified based on the image activation patterns of a deep neural network trained for object and scene categorization) [33]. In the categorical context, images that produce the most similar activation patterns to the others in the set are the ones that undergo the largest decrements in memorability, whereas memorability can increase modestly for the images that are the most distinct. These results establish that a full account of image memorability requires a description not only of individual image identities, but also the context within which those images are embedded.

The Neural Correlates of Image Memorability
Image Memorability Is Reflected in the Magnitude of the Response to Novel Images
What neural mechanisms shape image memorability? In principle, image memorability variation could be a consequence of variation in how images are represented when they are viewed as novel, which is then carried over to variation in how well those images are remembered. Alternatively, image memorability variation could emerge for the first time when images are viewed as familiar, implying that memorability arises via the mechanisms involved in memory storage and/or memory signaling. As described later, existing evidence supports the former account (while not ruling out added contributions from the latter).

The first reports of the neural correlates of image memorability variation utilized human fMRI to pinpoint its locus [14,15]. These studies reported that image memorability within a category (e.g., faces or scenes) could be classified by increases in blood oxygen level-dependent (BOLD) activation in high-level visual cortex, as well by decoding the patterns of BOLD activation across voxels, as subjects viewed novel images. Similar results were determined with human electroencephalogram (EEG) responses to ambiguous pictures of faces, where images that evoked larger N170 activity during the novel viewing period were more likely to be remembered [24]. Similar results were also reported in a study whereby image identity was more robustly decoded from human magnetoencephalography (MEG) responses for more as opposed to less memorable novel images, even under conditions in which images were not remembered at all [i.e., in a rapid serial visual presentation (RSVP) sequence at short durations and with masking [16]]. Together, these results suggest that the neural correlates of image memorability are reflected in the visual representations of images when they are viewed for the first time. These results are complementary but distinct from earlier work on ‘subsequent memory effects’ where items are sorted for each subject into ‘remembered’ versus ‘not remembered’ and remembered items tend to evoke indicators of higher neural activity during memory encoding (reviewed by [36]). Subsequent memory effects have been recapitulated in fMRI image memorability investigations, both within high level visual cortex as well as other structures (e.g., the medial temporal lobe, and prefrontal and parietal cortex [14,15]). In comparison, the memorability activation patterns described earlier remain whether subjects remember images or not [15,16,24,27], and are limited to high-level visual cortex as well as the medial temporal lobe [14,15]. These results are consistent with the interpretation that image memorability effects follow from properties that are associated with images as opposed to other factors (such as stimulus-independent fluctuations in the attentional state of an observer).
Conceptualizing Image Memorability and Object Identity Representations

What exactly differs in the visual representations of novel images that are more as compared with less memorable? To address this question, one study recorded population activity at single-unit resolution from monkey inferotemporal cortex (ITC) as the monkeys performed a visual memory task similar to that described in Box 1 [6]. This study reported a strong correlation between image memorability scores and the overall magnitude of the ITC population response to novel images (Pearson correlation $\rho = 0.62$), where the most memorable images evoked ~20% larger magnitude responses versus images that were the least memorable. Notably, the existence of population magnitude variation in response to natural images had not previously been appreciated before investigations of image memorability in ITC, despite extensive investigation of visual representations in this structure (reviewed in [3,4,37]). Population response magnitude variation or equivalently ‘magnitude coding’ in ITC may have been overlooked due to its relatively subtle (albeit measurable) impact on perceptual behavior [38,39], coupled with assumptions that this type of variation is largely eliminated by neural mechanisms that work to maintain constant global firing rates across a cortical population (such as homeostatic plasticity [40] and divisive normalization [41]). By contrast, investigations of memorability demonstrate that population response magnitude variation can be considerable (up to 20%) and that this variation strongly covaries with at least one type of behavioral change: how well images will be remembered.

How can current accounts of object representations in high-level visual cortex be extended to incorporate image memorability variation? That is, how might brain areas such as ITC reflect representations in which different images of the same object are identified as the same and, simultaneously, some of those images are more memorable than others? The multiplexing of object identity and memorability in ITC has been proposed to happen through two complementary coding schemes: a spike pattern coding scheme for object identity, and spike magnitude coding scheme for image memorability [6]. In the case of object identity, spike pattern coding is a consequence of individual ITC neurons that are selectively responsive or ‘tuned’ for the high-level image properties that define objects. This translates into representations of different images that are reflected by different population spike patterns, and in the high-dimensional neural representational space that is typically used to conceptualize object representations (Figure 2), different angular directions for different images [37,42]. Given that ITC neurons tend to maintain their rank-order object selectivity across different transformations of an object (e.g., changes in position or background [37,42]), representations of different images containing the same object tend to cluster in this space (Figure 2) [37]. In this format, object identity can be easily decoded, for example, by determining the object cluster that a particular ITC population response pattern is most similar to (or by other variants of linear population decoding [37]). In comparison, memorability is reflected by the magnitude of the ITC population response, thus allowing for some images within an object cluster to be more memorable than others (Figure 2).

While Figure 2 provides a simple and intuitive account of ITC that is supported by considerable evidence, there are also results that it cannot account for [27]. Namely, more memorable images have more similar fMRI voxel response patterns compared with less memorable images, as assessed by a representational similarity analysis [14,15]. This result, replicated across two studies, suggests that, in addition to the reflection of memorability via a magnitude coding scheme in ITC (Figure 2), some aspect of memorability may be reflected via spike pattern coding. This result conflicts with proposals in which memorability follows from a multidimensional representational space where more typical (and less memorable) objects are represented more centrally, while more distinctive (and more memorable) items are represented more distantly [43,44]. One question going forward will be to understand how this finding integrates with the depiction presented in Figure 2.
Image Memorability Is Reflected in Deep Artificial Neural Networks

Where does the population response magnitude variation associated with image memorability come from? Intriguingly, image memorability variation and its brain-analogous correlates emerge at higher stages of deep artificial neural networks trained to categorize objects (Box 2) [6], and this result holds across different deep neural network architectures and optimization schemes (e.g., AlexNet [45], the Hybrid CNN [46], and VGG-16 [47]) [6]. Specifically, at earlier stages of these networks (e.g., the V1-analogous layers), population response magnitude only weakly correlates with image memorability, consistent with observations that low-level image properties only contribute weakly to image memorability variation [7] (Figure 3). Correlations between population response magnitude and image memorability grow in strength across the hierarchy of these networks up to the layers that are analogous to high-level visual cortex [6] (Figure 3). At the very highest layers of the network, correlations often begin to drop (e.g., the fully connected layers of VGG-16, Figure 3), consistent with representations that become categorically invariant (i.e., these layers increasingly respond in an identical manner to all images containing the same object). Stated differently, layers analogous to high-level visual cortex in deep neural networks trained to categorize objects respond more vigorously to some images than others, and the vigor of these responses is predictive of the images that we find most memorable. These findings complement illustrations that deep neural networks trained to categorize objects have a functional organization that bears considerable resemblance to the brain areas that comprise the form processing pathway in humans and nonhuman primates [3,4]. However, what makes the memorability result so compelling is its emergence: compared with optimizing a deep neural
network for object categorization and finding that brain-like object representations emerge, memorability representations emerge in deep neural networks that are trained for object categorization but not explicitly to predict image memorability (nor do they, once trained, have a memory trace of anything that they have ‘seen’).

These results suggest that image memorability variation is shaped by the optimizations required for object-based (as opposed to memory-based) processing. However, intuitively, how do object-based optimizations shape network activity? For example, atypical images tend to be more memorable than typical ones; how is image atypicality reflected, as well as shaped, in a deep neural network trained for object categorization? Many questions remain.

Figure 3. Representation of Image Memorability across Different Layers of a Deep Artificial Neural Network Trained for Object Categorization, VGG-16. (A) Pearson correlation between population response magnitude and image memorability scores computed for different layers of one neural network, VGG-16 [47]. The correlations for both a randomly connected version of the network and a fully trained network are shown, replotted from [6]. ‘Conv’, convolutional layer; ‘FC’, fully connected layer; ‘ReLU’, rectified linear unit. (B) Depiction of the VGG-16 network architecture. This architecture includes 13 convolutional layers (Conv), three fully connected layers (FC), and five Max pooling layers. The last layer of VGG-16 (FC8) has been omitted in both panels, because it reflects the output of a 1000-way classification. Adopted from [69] (B).
The fact that different human individuals [5,7,8,17,33], and even humans and rhesus monkeys [6],
distinguished from memory: insofar as image memorability variation arises entirely from variation
an image before. However, at the level of underlying mechanism, image memorability could be
relationship between neural representations of image memorability and visual salience.
response magnitude, both in high-level visual cortex (Figure 2) and in deep neural networks
straightforward, where image memorability is strongly predicted by variation in population
response magnitude, both in high-level visual cortex (Figure 2) and in deep neural networks
trained to categorize objects (Figure 3) [6]. The apparent misalignment between the seeming
complexity of image memorability behavior and the seeming simplicity of its neural correlates likely
reflects the fact that many different factors combine to determine population response magnitude
in high-level visual cortex. At the same time, whether the neural correlates of image memorability
prove to be much more elaborate than what has been revealed thus far remains to be seen.

Going forward, image memorability can provide an important complement to object identification
behavior for probing and constraining descriptions of how representations of different behaviorally
relevant variables are transformed across the primate visual form processing pathway [3,4,48].
Analogous to the alignment of human and monkey object identification behavior [49,50], humans
and monkeys tend to find the same images memorable and forgettable [5], thereby allowing for
these investigations to be conducted with neural data collected in the brains of animals at high
spatial and temporal resolution, as well as causal tests of existing hypotheses via perturbation
approaches [51,52]. Moreover, the strong relationship between image memorability behavior
and population response magnitude in high-level visual cortex (Figure 2) prompts a host of qualita-
tively new questions about how different types of memorability variation are reflected, as well as the
relationship between neural representations of image memorability and visual salience.

At the behavioral level, image memorability cannot be isolated from memory, because image
memorability is the systematic variation in the ability of subjects to report whether they have seen
an image before. However, at the level of underlying mechanism, image memorability could be
distinguished from memory: insofar as image memorability variation arises entirely from variation
in the robustness of visual representations (which, in turn, has consequences for memory storage),
one can regard the source of image memorability variation as ‘visual processing’. That said, it is
crucial to appreciate that current accounts of image memorability are fundamentally incomplete,
because they lack descriptions of how visual representations are transformed into visual memories.

Box 3. Visual Memory Storage and Visual Familiarity Signaling

Visual memory storage and the signaling of whether an image is novel or familiar have been linked to processing within ITC
and its primary projection area, perirhinal cortex. Within those structures, familiarity is reflected by repetition suppression:
adaptation-like reductions in the population response to repeated images [57–61]. Potentially consistent with image memorability behavior, repetition suppression acts multiplicatively [62–64], and more repetition suppression is expected to follow when more memorable images are repeated (as a consequence of more vigorous responses). As an illustrative example, the same proportional reduction (e.g., 10%) applied to a larger compared with a smaller quantity (e.g., 100 versus 10) will result in a larger reduction (e.g., 10 versus one, respectively). However, whether repetition suppression in ITC and perirhinal cortex can fully account for image memorability behavior is still unclear. Other evidence suggests a role for the hippocampus in supporting visual recognition memory behavior [65,66], particularly in scenarios where subjects are asked to judge the familiarity of an image that is visually similar to one that has been viewed previously [67]. In those cases, the hippocampus is proposed to contribute via a process known as pattern separation [68]. Outside the medial temporal lobe, frontoparietal regions have been implicated in supporting visual memory behavior by way of differential responses elicited by these structures to images that are remembered versus those that are not [14,15].
While several different brain areas have been implicated in visual memory storage (Box 3), we do not understand how processing in these different regions combines to support memory. What is clear is that image memorability variation can serve as an important way to continuously manipulate visual memory in future investigations of its neural correlates and thereby facilitate answers to a question that has largely eluded the field for 50 years: how do our brains manage to remember images so well [1]?

References

1. Standing, L. (1973) Learning 10,000 pictures. Q. J. Exp. Psychol. 25, 207–222