THE UNIVERSITY OF CHICAGO

The Features that Drive the Memorability of Objects

By

Max Kramer

June 2022

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Wilma Bainbridge
Preceptor: Elizabeth Huppert

ABSTRACT

Despite decades of study of memory, it remains unclear what makes an image memorable. There is considerable debate surrounding the underlying determinants of memory, including the roles of semantic (e.g., animacy, utility) and visual features (e.g., brightness) as well as whether the most prototypical or most atypical items are best remembered. Prior studies have also relied on constrained stimulus sets, preventing a generalized view of the features that may contribute to memory. Here, we collected over one million memory ratings (N=13,946) for THINGS (Hebart et al., 2019), a naturalistic dataset of 26,107 object images designed to comprehensively sample concrete objects. We uncover a model of object features that is significantly able to predict image memorability, covering over half of the explainable variance. Within this model, we find that semantic features have a stronger influence than visual features on what people will remember. Finally, we examined whether memorability could be accounted for fully by the atypicality of the objects, by comparing three complementary measures using human behavioral data, object feature dimensions, and deep neural network features. We discover, surprisingly, that the relationship between memorability and typicality is more complex than a simple positive or negative association, however, generally, prototypical objects are the most memorable. Taken together, our findings reveal important structural features underlying the organization of information in memory.

SIGNIFICANCE STATEMENT

Why is it that we seem to remember and forget the same things? Our lived experiences differ, but we observe remarkable consistency in what is remembered across people. Here, we collected memory performance scores for a comprehensive and diverse collection of natural object images to identify which properties determine our ability to remember. We create one of the best performing models for predicting memory from object features. We observe that semantic information contributes primarily to memorability and that the most typical items are remembered best. Our findings challenge decades of prior research that suggest that the most distinct items are most memorable and inform our understanding of the features and organizational principles of memory.

INTRODUCTION

What is it that makes something memorable? Researchers have been struggling for decades to understand the determinants of memory and how information is encoded, processed, and retrieved in the brain. The majority of research in memory uses a subject-centric framework, attempting to understand the underlying processes of memory and individual differences across people. This subject-centric framework is motivated by the highly personal nature of memory, as everyone has their own experiences that influence what they will later remember. However, an alternative stimulus-centric framework has arisen out of the surprising finding that, despite our individual experiences, we largely remember and forget the same images (Isola et al., 2011; Bainbridge et al., 2013). This new stimulus-driven perspective allows for a targeted examination of *what* we remember, and *why*.

This stimulus-driven perspective has revealed that images have an intrinsic *memorability*, defined for a stimulus as the likelihood that any given person will remember that stimulus later (Bainbridge et al., 2019). By using aggregated task scores for each stimulus rather than individual participant responses, memorability for a given stimulus can be quantified, repeatedly demonstrating a high degree of consistency in what people remember (Isola et al., 2011; Bainbridge et al., 2019) across stimulus types (see Isola et al., 2011; Bainbridge et al., 2013; Borkin et al., 2013; Xie et al., 2020). These memorability scores can account for upwards of 50% of variance in memory task performance (Bainbridge et al., 2013) and demonstrate remarkable resiliency across tasks and robustness to attention and priming (Bainbridge, 2020). This high consistency allows one to make honed predictions about what people will remember, which could have far-reaching implications for fields including advertising, marketing, public safety (Bainbridge et al., 2019), patient care (Bainbridge, Berron, et al., 2019), and computer vision (Needell & Bainbridge, in press). However, in spite of these high consistencies in what individuals remember, what specific factors determine the memorability of an image is still largely unknown.

Prior research has sought to explain memorability as either a proxy for a given stimulus feature like attractiveness or brightness, while others have attempted to reduce memorability to a linear combination of features in a constrained stimulus set. These studies mostly utilize faces (Bainbridge et al., 2013) or scenes (Isola et al., 2014) as stimuli, and none of them have explained the majority of variance in memorability using these models. More recently, researchers have emphasized the importance of considering items in a multidimensional representational space, with memorability arising from the relative location of an item within that space (Lukavský & Děchtěrenko, 2017; Bainbridge, 2019; Koch et al, 2020). This theoretical framework has sparked debate about the roles of low-level visual features such as color and shape and semantic information such as animacy in determining what we remember and what we forget (Khosla, 2015; Jaegle et al, 2019; Madan, 2020; Xie et al, 2020). Additionally, researchers have disagreed on whether the most memorable items are the most prototypical items (Bainbridge, Dilks, & Oliva, 2017; Bainbridge & Rissman, 2018) or the most atypical items (Bylinskii et al, 2015; Lukavský & Děchtěrenko, 2017; Mosenzadeh et al, 2019). It is clear that

there is a lack of consensus surrounding the roles of visual and semantic features as well as typicality with regards to what we remember, necessitating further analysis.

Here, we provide a comprehensive characterization of visual memorability across an exhaustive set of picturable object concepts in the English language (THINGS database, Hebart et al., 2019). Specifically, we determine the object features and their organizational principles that drive our memories. In the largest study of memorability to date, we collected over 1 million memory scores for all 26,107 images in the THINGS database, which we have made publicly available (https://osf.io/5a7z6/?view_only=675e901c176c4bec9c2540fc4981e5fe). We then leveraged three complementary measures—human judgments, multidimensional object features, and predictions from a deep convolutional neural network (CNN)—to examine the relationship of memorability to object typicality. We discover a feature model that is able to significantly predict a majority of the variance in image memorability. Among those features, our results uncover a primacy of semantic information over visual information in what we remember. Further, while we find evidence of the most typical items being best remembered, the high variance across categories suggests that the relationship between memorability and typicality is more complex than prior work would suggest.

RESULTS

Our analyses characterize memorability across object concepts, addressing whether memorability is more visually or semantically driven and whether it is the most prototypical or atypical objects that are best remembered. To explore memorability across concrete objects, we collected memorability scores for the entire image corpus of the THINGS database of object images (Hebart et al., 2019) and uncovered a dispersion of memorability across the hierarchical levels of THINGS. We examined the roles of semantic and visual information by attempting to predict memorability from semantic and visual features using multivariate regression, where we revealed that semantic information contributes primarily to object memorability. We then analyzed multiple measures of object typicality along with the memorability scores and found a small but robust effect of the most prototypical items being best remembered.

THINGS is a hierarchically structured dataset containing 26,107 images representing 1,854 object *concepts* (such as aardvark, tank, and zucchini) derived from a lexical database of picturable objects in the English language (see Methods), 1,619 of which are assigned to 27 higher *categories* (such as animal, weapon, and food). The concepts were assigned to categories in prior work through a two-stage process where one group of participants proposed categories for a given concept while a second group narrowed the potential categories further, with the most consistently chosen category becoming the assigned category for the concept (Hebart et al., 2020). The concepts and images are also characterized by a set of 49 dimensions that capture 92.25% of the variance in human behavioral similarity judgments of the objects (Hebart et al., 2020). Each concept and each image thus can be described by a 49-dimensional embedding that corresponds to the representation of that item in object feature space. This overall dataset structure enables the analysis of memorability at the image, concept, category, and dimensional levels.

Memorability is Diffuse Across Objects

In order to quantify memorability for all 26,107 images in THINGS, we conducted a continuous recognition memory task (N = 13,946) administered over the online experiment platform Amazon Mechanical Turk (AMT) wherein participants viewed a stream of images and were asked to press a key when they recognized a repeated image that occurred after a delay. Memorability was quantified as the corrected recognition (CR) score for a given image, calculated as the proportion of correct identifications of the image minus the proportion of false alarms on that image (Bainbridge & Rissman, 2018). However, all results replicate when corrected recognition is instead substituted with hit rate or false alarm rate (Supplemental Information). To test if we observe consistency across people in what they remember and forget, we conducted a split-half consistency analysis across 1,000 iterations and found significant consistency in what split halves of participants remembered (Spearman-Brown corrected split-half rank correlation, mean $\rho = 0.449$, $p < .001$), which is surprising given the diversity of the THINGS images. This consistency in memory performance implies that memorability can be considered an intrinsic property of these stimuli.

When assessing memorability at the concept level, we observe that memorability varied strongly across the concepts (Figure 1a). This dispersion of CR suggests that not all concepts in THINGS are equally memorable. For example, *candy bars* were highly memorable overall with a maximum CR of 1, a mean of 0.873, and a minimum of 0.756 (range = 0.127), while *windshields* were less memorable with a maximum CR of 0.756, a mean of 0.649, and a minimum of 0.404 (range = 0.352). We observe a similar diversity of memorability patterns at the higher category level (Figure 1b). The average CR across the THINGS categories is 0.793, with some categories demonstrating a higher average memorability than others; *body parts* attained the highest average memorability at 0.855 while *car parts* had the lowest average memorability of 0.753. These measures highlight the rich variation present within the THINGS database as it relates to memorability.

The embeddings along 49 dimensions for each of the object concepts allow us to determine if certain dimensions are more strongly reflected in memorable stimuli (Figure 1c). Specifically, we examined Spearman rank correlations between the memorability of the THINGS concepts and the concepts' embedding values for each of the 49 dimensions. We found that 36 dimensions showed a significant relationship to memorability (FDR-corrected $q < 0.01$), of which 9 were positive and the remaining 27 were negative. These correlations reveal that some properties used to characterize an object do show a relationship to memorability. For example, the positive relationship for the *body / body part* dimension ($\rho = 0.257$, $p = 1.873 \times 10^{-29}$) indicates that memorable stimuli tend to be more related to body parts, while a negative correlation like *metal / tools* ($\rho = -0.323$, $p = 1.689 \times 10^{-15}$) implies that more memorable stimuli tend to not be made of metal. These patterns of diffusion persist when examining hit rate and false alarm rate separately, rather than the combined measure of corrected recognition (see supplement).

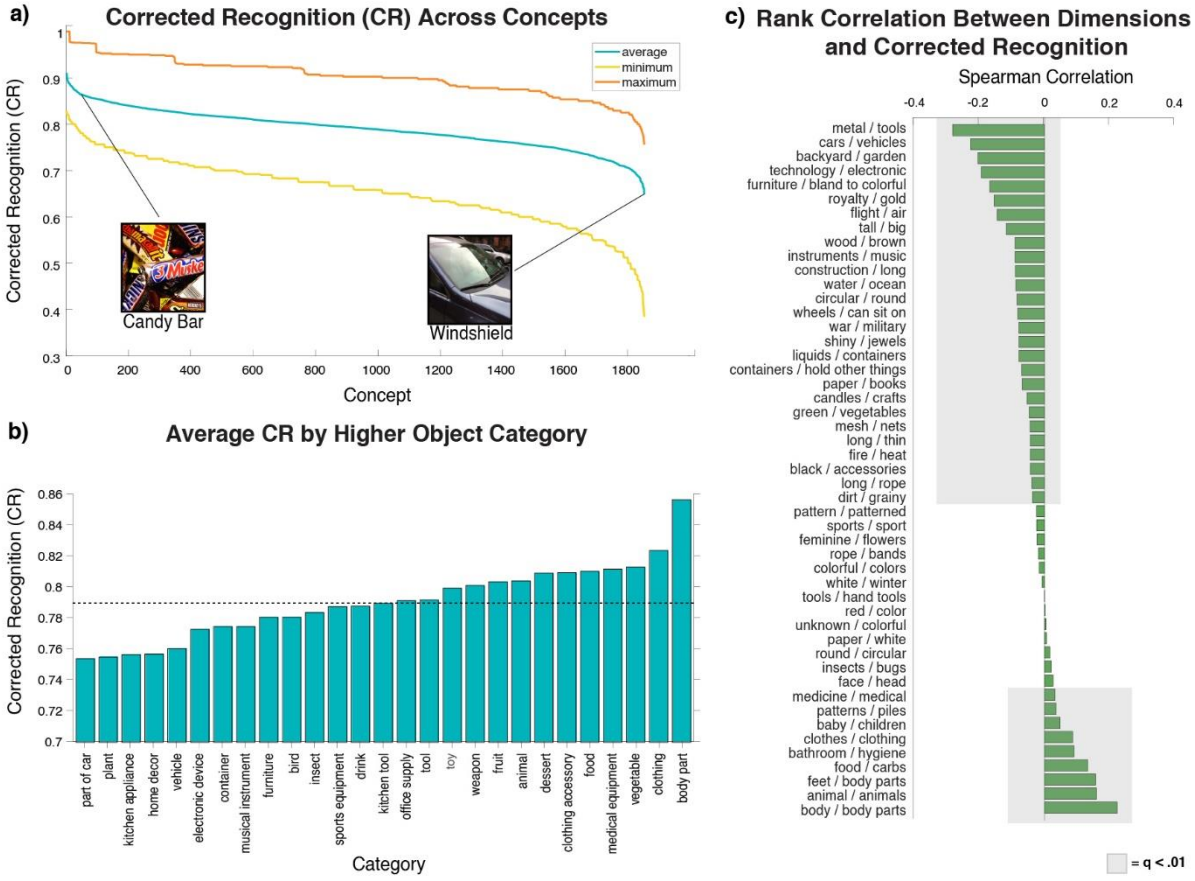


Figure 1. Descriptive analyses of memorability across the concept and category levels of the THINGS database as well as the 49 object dimensions. (A) The spread of corrected recognition (CR) across the 1,854 object concepts revealed that not all concepts are equally memorable. For concepts like *candy bars*, the entire range of component image memorability values were contained above the average value for a concept like *windshields*. (B) Visualizing the same spread across higher order categories revealed variation in average memorability across the 27 categories, with some categories including *car parts* displaying a CR score below the overall average memorability of 0.793 represented by the dotted horizontal line while others like *body parts* displayed a score above the average. (C) This trend of a spread of relationships continues when examining the correlation between memorability and embeddings along the object dimensions. 36 out of 49 dimensions displayed a significant association with memorability (shaded bars, FDR-corrected $q < 0.01$), with 9 showing a positive relationship (i.e., *body / body parts* being more memorable), and 27 showing a negative relationship (i.e., *metal / tools* being less memorable).

Having explored memorability across the structure of THINGS, we can readily observe that memorability varies at the exemplar, concept, higher category, and dimensional levels. With this understanding, the question becomes: what causes some concepts/categories/dimensions to be more memorable than others?

Semantic Information Contributes Most to Memorability

To examine which object features are most important for explaining what is remembered and what is forgotten, we predict the average memorability scores of the THINGS concepts using the object space dimensions (Table 1). Our regression model utilized the 49-dimensional

embedding of each concept in the object space to predict the average CR score for the concept. Overall, the model explained 38.52% of the variance in memorability (Figure 2B). Because memorability scores contain some noise, we also calculated performance of this model in comparison to a noise ceiling estimated by predicting split halves of the memory data across 100 iterations (see Methods). We found our model explained 61.66% of the variance given the noise ceiling, implying that these dimensions capture the majority of variance in memorability.

The explanatory power of our model serves as a strong starting point for an analysis of the types of dimensions that contribute most to memorability. We sorted the dimensions into two main categories: visual and semantic dimensions. Dimension names were determined in a prior study, as the top two-word phrases selected by naïve observers for sets of the most heavily weighted images on those dimensions (see Methods; Hebart et al., 2020). We defined visual dimensions of an image to be those concerned primarily with color and shape information, such as "red / color", "long / thin", "round / circular", and "pattern / patterned" (Table 1). We defined semantic dimensions as categorical information that did not include references to color or shape, such as "food / carbs", "technology / electronic", and "body / body parts". Any dimensions that contained both semantic and visual information as defined above were classified as mixed, such as "green / vegetables", "black / accessories", and "white / winter".

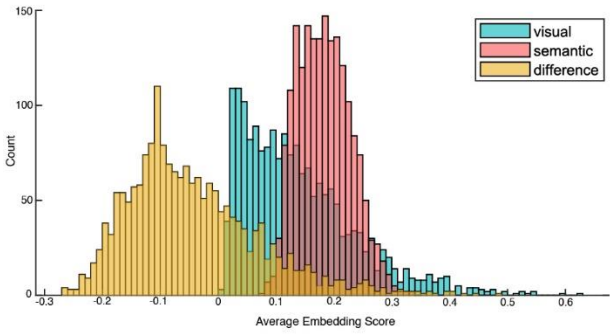
Table 1. Categorization of THINGS object space dimensions across semantic, visual, and mixed dimensions. Dimension names were derived from naïve observers viewing the highest weighted images on each dimension. Dimensions are listed in order of highest to lowest correlation with memorability score.

Semantic	Visual	Mixed
Metal / Tools	Colorful / Colors	Furniture / Bland to Colorful
Food / Carbs	Circular / Round	Green / Vegetables
Animal / Animals	Patterns / Piles	Wood / Brown
Clothes / Clothing	Long / Thin	Royalty / Gold
Backyard / Garden	Red / Color	Dirt / Grainy
Cars / Vehicles	Round / Circular	Black / Accessories
Body / Body Parts	Pattern / Patterned	Long / Rope
Technology / Electronic	Tall / Big	Paper / White
Sports / Sport	Mesh / Nets	Rope / Bands
Tools / Hand Tools		Construction / Long
Paper / Books		Unknown / Colorful
Liquids / Containers		White / Winter
Water / Ocean		Shiny / Jewels
Feminine / Flowers		
Bathroom / Hygiene		
War / Military		
Instruments / Music		
Flight / Air		
Insects / Bugs		
Feet / Body Parts		
Fire / Heat		
Face / Head		
Wheels / Can Sit On		
Containers / Hold Other Things		
Baby / Children		
Medicine / Medical		
Candles / Crafts		

With these categorized dimensions, we can differentiate the contributions of primarily semantic and primarily visual dimensions to memorability. By analyzing the embeddings of each concept in the multidimensional object space, we revealed that 70.44% of the concepts were more heavily embedded in dimensions classified as semantic than dimensions classified as visual (Figure 2a). We ran a regression model that predicted memorability only from the dimensions strictly classified as either semantic or visual (excluding mixed dimensions). The resulting 36-dimensional model (27 semantic, 9 visual) explained 35.16% of the variance in memorability, and the semantic dimensions contributed 31.22% of the variance while visual dimensions only accounted for 1.62% with a shared variance of 2.32% (Figure 2c). This result suggests a clear dominance of semantic over visual information in memorability. To examine the effects of dimensions labelled as mixed, we also break down the unique and shared variance contributions from semantic, visual, and mixed dimensions in the full 49-dimensional model, where we see that mixed dimensions contributed 1.03% of variance in memorability (see supplement).

However, since there are also more semantic dimensions than visual dimensions in that model, we conducted a follow-up analysis with a model using just the top 9 highest weighted semantic dimensions and top 9 highest weighted visual dimensions. This model accounted for 19.15% of variance in memorability, with the top 10 semantic dimensions contributing 15.21% of variance while the top 10 visual dimensions contributed 1.87% of variance with a shared variance of 2.07% (Figure 2d). A summary of all regression results is displayed in Figure 2b.

a) Embedding scores in 49 dimensional object space at concept level

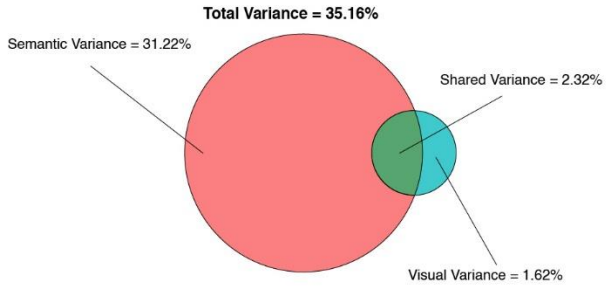


b) Summary of Regression Results

Model	R^2	F	p	error
All dimensions	0.3852	23.0505	2.13e-154***	0.0011
Semantic	0.3354	34.113	9.37e-141***	0.0012
Visual	0.0394	8.3998	2.25e-12***	0.0017
Top Semantic	0.1728	42.7848	5.25e-70***	0.0014
Top Visual	0.0394	8.3998	2.25e-12***	0.0017

*** = p < .001

c) Variance in Memorability Explained: 36 Dimensional Model



d) Variance in Memorability Explained: Top 18 Model

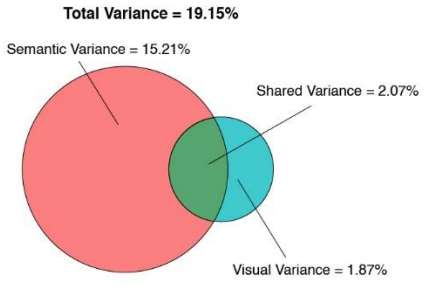


Figure 2. Analyses of relative contributions of semantic and visual information to memorability. (A) Histogram of averaged embedding values in semantic (red) and visual (blue) dimensions across concepts. The yellow histogram represents the difference between the visual and semantic embeddings (blue - red). The embeddings of the 1,854 concepts in the object space reveal that 70.44% of the concepts are more heavily embedded in semantic dimensions than in visual dimensions. (B) Table of regression models. The semantic and visual models utilize all 27 semantic and 9 visual dimensions respectively to predict memorability and captured 38.52% of the variance in memorability. The top models utilized only the 9 most heavily embedded semantic and visual dimensions, to balance the number of semantic and visual dimensions in the model. Across models, the majority of variance was captured by semantic dimensions. (C) Venn diagram displaying the unique contributions to memorability from semantic and visual dimensions. For the model using all non-mixed dimensions, the majority of variance is captured by the 27 semantic dimensions, with a smaller contribution from the 9 visual dimensions. Note the larger shared variance than visual variance, suggesting that most of the contribution of visual information may be contained in shared variance with semantic information. (D) The same type of Venn diagram as in (C) but with a model including equal numbers of semantic and visual dimensions (9 regressors each). Again, the majority of explained variance comes from semantic dimensions.

Taken together, our results indicate that semantic information contributes far more than visual information towards the memorability of an image. While the results reveal contributions of visual information, these contributions are largely captured by shared variance with semantic information. We observe a similar pattern of results when examining hit rate and false alarm rate as dependent variables in place of corrected recognition (see supplement).

Memorability is More than Typicality

While we have determined that semantic features are the most predictive dimensions of the object space for memorability, there is still the question of whether it is the most prototypical or most atypical items that are best remembered along these dimensions. In terms of the object feature space, items that are clustered closely together are the most prototypical items, while items spaced further apart are the most atypical items. The relationship between typicality and memory has been studied extensively in face processing, scene recognition, and related fields (Lee et al., 2000; Bylinskii et al., 2015; Lukavský & Děchtěrenko, 2017), and suggest three different hypotheses, where the relationship between typicality and memorability is either always negative (Lukavský & Děchtěrenko, 2017), always positive (Bainbridge & Rissman, 2018), or a specific combination of the two (Koch et al., 2020). Here, we leverage the scale of THINGS to determine this relationship utilizing converging methods for defining typicality based on behavioral ratings, the multidimensional object space derived from human similarity judgments, and a deep neural network for object recognition.

In order to assess whether the most prototypical or most atypical items are best remembered, we employ three complementary measures of object typicality. Our first measure of typicality we dub “object space typicality”, and it is derived from the object space employed in the previous analyses (Figure 3a). The 49-dimensional space has been demonstrated to capture human behavior in excess of 90% of ceiling (Hebart, Zheng, & Perreira, 2020) and predict memorability with high accuracy. Typicality scores calculated from the object space dimensions capture information about both visual and semantic features, which allows for testing hypotheses

relating object typicality to visual and semantic content. We term our second measure of typicality "CNN-based typicality", as it employs the VGG-F deep CNN to compute similarity ratings across the 22 layers of the network (Figure 3b). Deep neural network models have demonstrated success in predicting the neural responses of different regions in the visual system (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). A critical insight from these studies suggests that earlier layers in the network represent low-level visual information such as edges, while later layers represent more complex and semantic features like categorical information (Güçlü & van Gerven, 2015). Unlike the object space derived scores, these typicality values are directly computed from image features, rather than based off of behavioral similarity judgments in response to the images themselves. Finally, our third measure of typicality, referred to as "behavioral typicality", consists of behavioral ratings derived from a concept to category matching task (Hebart, Zheng, & Perreira, 2020) to capture human intuition regarding typicality (Figure 3a). In this prior study, participants on Amazon Mechanical Turk used a 0-10 Likert scale to assess the degree to which a given concept was typical of a category (e.g., how typical is a snake of animals?). These three complementary approaches allow for testing a wide range of hypotheses concerning whether the most prototypical or atypical items are most often remembered.

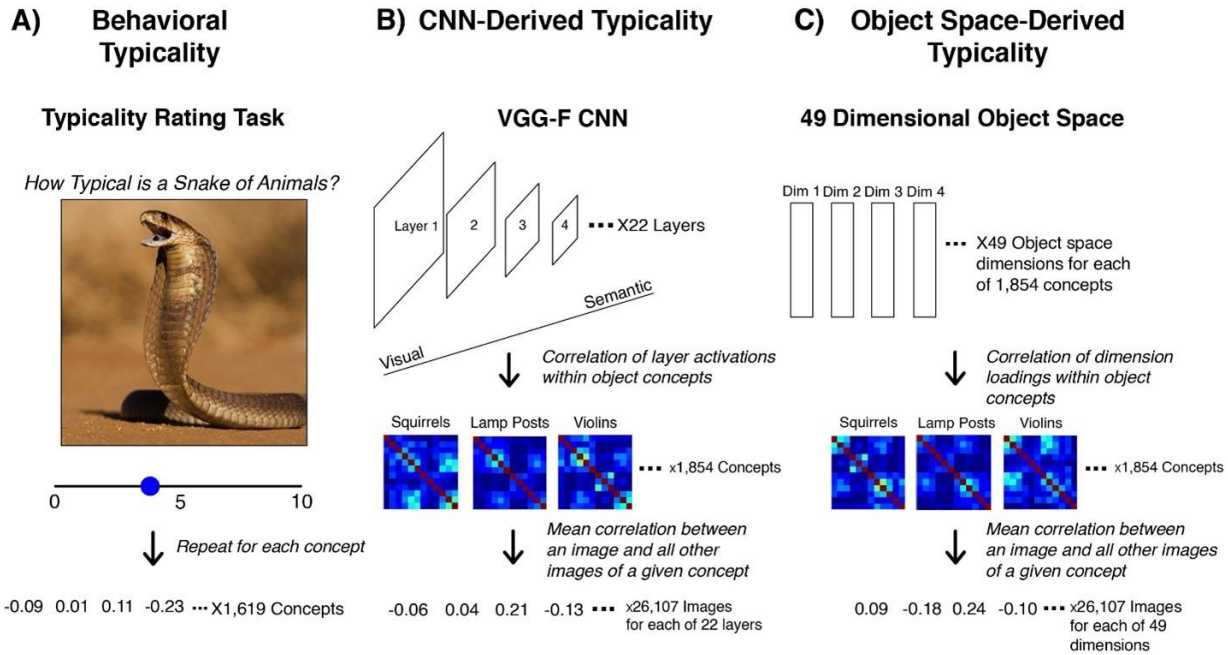


Figure 3. Generating typicality scores from behavior, object space dimensions, and CNN activations. (A) For behavioral typicality, participants on Amazon Mechanical Turk used a 0-10 Likert scale to assess the typicality of a given object concept (snake) to its higher category (animals). These typicality scores were then aggregated across all of the concepts under a given higher category to generate a typicality score for that category. (B) To generate the CNN-computed typicality score, each of the 26,107 images was input to the VGG-F network and had layer activations extracted at each of the 22 layers. Correlating the resulting layer values within each of the image concepts allowed for the generation of similarity matrices for each object concept. From these matrices, we compute the typicality of each image as the mean correlation between the image and all other images of a given object concept, resulting in a typicality score for every image in relation to its concept. (C) The procedure for generating typicality scores from the object space dimensions is largely the same as the process for the CNN but relying instead

on embeddings of images in the object space as the representation for each image, which was then correlated to form similarity matrices.

Our first assessment of the relationship between memorability and typicality was to examine the overall correlation between the corrected recognition scores and object space typicality scores for the 26,107 image corpus of THINGS. This typicality score reflects the typicality of a given example image (e.g. a particular example of a squirrel) relative to all other examples of that image's concept (e.g. all images of *squirrels* in THINGS). Using the typicality scores derived from the object space dimensions, we found a significant positive relationship between image typicality and memorability across the THINGS dataset ($r = 0.309$, $p = 6.131 \times 10^{-7}$). This suggests that more memorable images tend to be more prototypical of their concept in their representations across these dimensions, arguing against a general primacy of atypicality in memorability. We also analyzed the relationship between object space typicality and memorability within each of the 1,854 concepts in THINGS by correlating memorability and typicality values across the exemplar images of each concept. In other words, within each concept, what is the relationship between typicality and memorability? We again employed the typicality scores derived from the object space dimensions and produced a distribution of correlations between exemplar corrected recognition and exemplar typicality scores. Overall, the concepts were more likely to display a relationship where more prototypical images tended to be more memorable (one sample t-test: $t(1852) = 2.074$, $p = 0.038$).

Recent analyses have suggested that the relationship between typicality and memorability may differentially depend on similarity across semantic and visual features; for example, for a set of scene images, the most visually atypical but semantically prototypical images tended to be most memorable (Koch et al., 2020). Based on this hypothesis, it may be that differential contributions of semantic and visual features influence whether the most prototypical or atypical items are best remembered. To test this, we leveraged the CNN-derived typicality scores and the heuristic that early layers represent more visual information and late layers represent more semantic features. We visualize the correlation of typicality at both an early layer (2) and a late layer (20) with corrected recognition (Figure 4a) and segment the resulting figure into quadrants based on correlation magnitude. We observe a significant correlation between the relationship of memorability and typicality at early layers and late layers ($r = 0.253$, $p = 2.504 \times 10^{-28}$), suggesting that in general visual and semantic features show a similar relationship of typicality to memorability. A chi-square analysis on each quadrant revealed that significantly more concepts than chance showed a pattern where the most memorable items were prototypical in terms of both early and late layer features ($\chi^2 = 38.046$, $p = 6.909 \times 10^{-10}$). In contrast, we find significantly fewer concepts than chance show a mixed pattern, where memorable items were determined by early layer prototypicality and late layer atypicality ($\chi^2 = 8.454$, $p = 0.004$), or the opposite pattern of early layer atypicality and late layer prototypicality ($\chi^2 = 20.286$, $p = 6.668 \times 10^{-6}$). Finally, there was no difference from chance in the proportion of concepts that showed a pattern where the most memorable items were the most atypical items for both early and late CNN layers ($\chi^2 = 8.3993$, $p = 0.553$). These results suggest that in general, memorable images tend to be those that are both visually and semantically prototypical of their object concept,

although there are also concepts for which memorable images may tend to be either visually or semantically atypical. We find largely similar patterns of results when examining hit rate and false alarm rate in place of corrected recognition (see supplement).

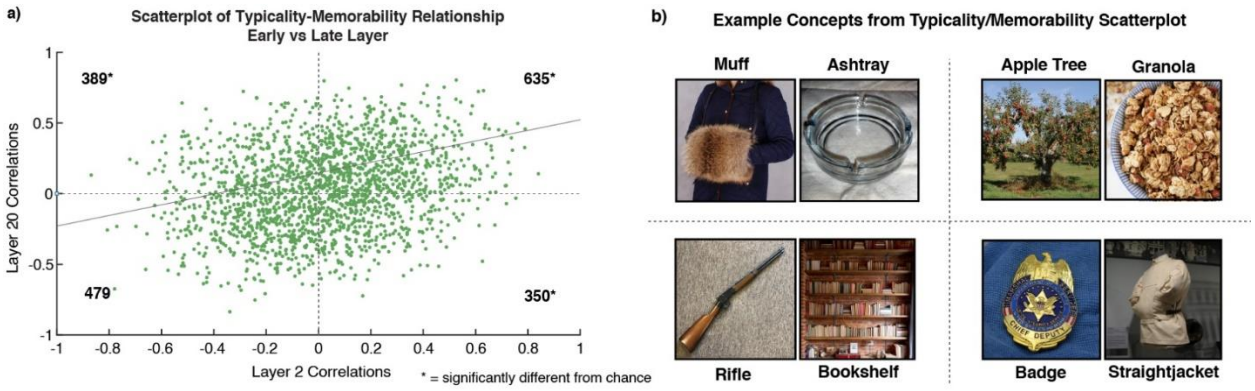


Figure 4. Examining relationships between typicality, memorability, and semantic and visual content. (A) Visualizing the correlation of CNN-based typicality and memorability for all 1,854 concepts in terms of an early layer (layer 2) and late layer (layer 20) allows for the observation of an overall positive relationship between early and late layer typicality scores across the concepts ($r = 0.253$, $p = 2.504 \times 10^{-28}$). A chi square analysis of the four quadrants of the scatterplot demonstrated significantly more concepts than chance showed a pattern where the most memorable items were prototypical in terms of both early and late layer features ($\chi^2 = 38.046$, $p = 6.909 \times 10^{-10}$). Contrastingly, we find significantly fewer concepts that demonstrate "mixed" patterns where more memorable items demonstrated early layer prototypicality and late layer atypicality ($\chi^2 = 8.454$, $p = 0.004$), or the opposite pattern ($\chi^2 = 20.286$, $p = 6.668 \times 10^{-6}$). We found no significant difference from chance for concepts where the most memorable items were atypical across both early and late layer features ($\chi^2 = 8.3993$, $p = 0.553$). This suggests that, in general, memorable concepts tend to be both visually and semantically prototypical. (B) Example concepts that fell into each quadrant of the scatterplot seen in C.

The previous findings demonstrate converging evidence for memorability corresponding to object prototypicality, however, there are also several counterexamples across the THINGS dataset. While, as a whole, a majority of object concepts showed a positive relationship between typicality and memorability, still many object concepts (917) show an opposite relationship, where more atypical images are more memorable. For example, for *coats*, more prototypical images were more memorable ($r = 0.857$, $p = 3.66 \times 10^{-4}$), but for other concepts such as *handles*, more atypical images were more memorable ($r = -0.798$, $p = 0.001$).

This mixed evidence is also apparent in analyses relating the typicality of concepts to the 27 higher categories present in THINGS, in contrast to the previously described analyses that tested the typicality of images in relation to their concepts. For any given concept, the category typicality score reflects the typicality of that concept (e.g. *squirrels*) relative to all other concepts of its higher category (e.g. *animals*). A correlation between CR scores and behavioral typicality scores across all higher categories showed no significant relationship between typicality and memorability ($r = 0.139$, $p = 0.576$). When examining the distribution of correlations between typicality and memorability, we observed a marginal effect of more atypical (rather than

prototypical) concepts being more memorable ($t(26) = -2.022$, $p = 0.054$). When examining the correlations for each of the 27 categories separately (see supplement), we found that *home décor* ($r = -0.384$, $p = 0.009$), *office supplies* ($r = -0.430$, $p = 0.032$), and *plants* ($r = -0.429$, $p = 0.003$) showed significant negative relationships, implying that more memorable examples of each category were more atypical. In contrast, *animals* ($r = 0.176$, $p = 0.020$), *food* ($r = 0.115$, $p = 0.050$) and *vegetables* ($r = 0.317$, $p = 0.041$) had positive relationships, implying that more memorable examples were more prototypical. A similar set of trends are observed when examining these relationships using object space typicality scores rather than behavioral scores (see supplement), where *containers* ($r = -0.213$, $p = 0.029$) and *electronic devices* ($r = -0.232$, $p = 0.047$) showed negative relationships (e.g., more atypical containers are more memorable), while *animals* ($r = 0.159$, $p = 0.034$) and *body parts* ($r = 0.473$, $p = 0.005$) demonstrated more positive relationships. Overall, across all high-level categories, there were an equal number of positive and negative significant relationships, demonstrating further mixed evidence within the THINGS dataset. As with our other analyses, we observed similar patterns of results when using hit rate and false alarm rate in place of corrected recognition (see supplement).

Taking all findings into account, it is clear that memorability cannot be considered synonymous with either prototypicality or atypicality, as has been suggested in previous studies (e.g., Valentine et al., 1991; Bylinskii et al, 2015; Bainbridge, Dilks, & Oliva, 2017). Certain results collected using both object space derived and CNN derived typicality scores suggest a trend towards more prototypical stimuli being more often remembered, but the large number of counterexamples present across the different typicality scores and levels of analysis suggest that the relationship between memorability and typicality is likely more complex than a simple positive or negative association, with a strong variance from concept to concept.

DISCUSSION

We analyzed a large, representative object image database to uncover what makes certain objects more memorable than others. We analyzed the roles of semantic and visual features and determined that semantic information more strongly influences what is remembered than visual information. We leveraged three complementary measures of object typicality to determine whether the most prototypical or most atypical images are best remembered and uncovered some evidence suggesting more prototypical items are more memorable, but also a high degree of variance across concepts and categories, suggesting that memorability is not just a measure of the typicality of an object or image. These findings shed new light on the determinants of what we remember and stand in contrast to previous studies that have claimed both that semantic information is not required to determine memorability (Lin et al., 2021) and that it is the most atypical items that are best remembered (Mohsenzadeh et al., 2019).

Semantic Primacy of Memorability

We analyzed the contributions of semantic and visual information to memorability to determine if the two types of information contribute differentially to the THINGS stimuli. Our results reveal a primacy of semantic information in explaining memorability, based on multiple regressions comparing the relationship of the entire object dimensional space to memorability. Even after equalizing the number of semantic and visual dimensions inputted to the model, 88.02% of the variance in memorability captured by the space was exclusively from the top 9 semantic dimensions.

Previous findings of the ability of CNNs (Khosla et al., 2015) and monkeys (Jaegle et al., 2019) to predict human performance on memorability tasks and examples of memory performance robust to semantic degradation (Lin et al., 2021) have led to the assertion that semantic knowledge is not required to make an image memorable. However, recent research has demonstrated that semantic similarity is predictive of memorability and lexical stimuli also display intrinsic memorability despite a lack of rich visual information (Xie et al., 2020; Madan et al., 2021). More recently, other studies have demonstrated that both visual and semantic information contribute differentially with regards to the typicality-memorability relationship, where visually atypical but semantically prototypical scene images may be the most memorable (Koch et al., 2020). Additionally, recent research in memorability prediction suggests that adding semantic information to a deep neural network improves the prediction of memorability scores (Needell & Bainbridge, in press). Our results demonstrate a strong semantic primacy in memory which lends additional support to recent findings demonstrating the importance of semantic information in determining what we remember.

Beyond behavior, our findings align with the results from recent neuroimaging studies that have examined the neural correlates of memorability. One such study found a lack of memorability-related activation in the Early Visual Cortex (EVC), suggesting that areas involved in lower-level perception may not be sensitive to memorability (Bainbridge et al., 2017). This result, coupled with a study demonstrating faster neural reinstatement for highly memorable stimuli in the Anterior Temporal Lobe (ATL), an area typically associated with semantic processing (Xie et al., 2020), could potentially reflect a neural signature of the observed outsize influence of semantic information in determining what is best remembered. In this study, memorability for word stimuli could be significantly predicted by the semantic connectedness of these words, where words that exist at the roots of a semantic structure tended to be more memorable (Xie et al., 2020). This suggests that memorability could reflect our semantic organization of items in a memory network. Other work has also found sensitivity to memorability in late perceptual areas, such as the Fusiform Face Area (FFA) and the Parahippocampal Place Area (PPA) (Bainbridge et al., 2017; Bainbridge & Rissman, 2018), often associated with the patterns seen in late CNN layers (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014).

Our findings are particularly surprising given the fact that the object space dimensions explained 61.66% of the variance in memorability. Unlike previous studies of memorability using single attributes (Bainbridge et al., 2017; Isola et al., 2014) or linear combination models

with constrained stimulus sets (Bainbridge et al., 2013), we are able to explain a large degree of the variance in memorability, further highlighting the importance of semantic properties. The success of this model also means that this same model can be applied to selecting stimulus sets intended to drive memory in specific ways; given an object's feature space, we can predict which items are likely to be remembered or forgotten. However, given the remaining unexplained variance, it is clear that there are still lingering questions about the determinants of what we remember and what we forget.

Typicality as it Relates to Memorability

Here, we observe that across our images, concepts, and categories, there are some by which the most prototypical are the most memorable, while there are others where the most atypical are the most memorable. These results suggest that memorability does not just reflect an object's typicality, and it is not merely that memorable items are the most distinctive, atypical items. In fact, across multiple levels of analysis, we observe the opposite, where in general more prototypical items tend to be the most memorable.

This is surprising, given that atypicality has long been thought to encapsulate the effect of memorability based on evidence from faces (Valentine, 1991) and scenes (Bylinskii et al., 2015), whereby more atypical items are thought to be easier to remember. Other studies have rebutted this claim by demonstrating that semantic similarity is predictive of memorability (Xie et al., 2020). Furthermore, late visual areas regions show neural patterns reflective of our current behavioral findings, where memorable face and scene images show more similar neural patterns to each other (i.e., have more prototypical patterns), while forgettable images have more dissimilar neural patterns (i.e., more atypical patterns; Bainbridge et al., 2017; Bainbridge & Rissman, 2018). Further, Koch and colleagues (2020) found a complex relationship with typicality, where visually distinct and semantically similar images were most often remembered in an indoor-outdoor classification task. Our divergent findings could possibly be explained by the constrained stimulus sets utilized in prior studies. While prior work focused on narrow stimulus sets such as faces or a smaller sampling of scene images, our study examines a comprehensive, representative set of object images across the human experience. Our divergent findings from these earlier studies may suggest that while previous findings are reasonable extrapolations from the stimuli domains examined, they are not characteristic of memorability as a whole. When assessed at a global scale, it is neither prototypicality nor atypicality of an item that makes it memorable.

The observation of variability in the typicality-memorability relationship may have important ramifications for neuroimaging research examining the neural correlates of memorability and memory more broadly. Observations of prototypicality in neuroimaging research reference a phenomenon called pattern completion as a means by which the hippocampus retrieves a complex representation from a given cue (LaRocque et al., 2013). This process depends on another hippocampal phenomenon termed pattern separation, where similar inputs are assigned distinct representations to facilitate the mnemonic discrimination required in

memory (Ngo et al., 2020). Whole-brain fMRI analyses have revealed that different areas involved in memory utilize separated and overlapping information to facilitate memory (LaRocque et al., 2013), suggesting a potential role for both prototypicality (as represented by pattern completion) and atypicality (as represented by pattern separation) in facilitating memory. Future neuroimaging research could identify potential neural markers of prototypicality and atypicality and determine if the effects of semantic and visual information are dissociable at a neural level.

Future Directions

Recently, a large-scale functional magnetic resonance imaging (fMRI) dataset has been created using a subset of the THINGS images. This dataset, termed THINGS-fMRI, leverages the comprehensive sampling of picturable objects provided by THINGS to enable researchers to examine the neural bases of object perception. The dataset itself employs 8,740 of the 26,107 THINGS images to cover 720 of the object concepts (Contier et al., 2021). The dataset itself comes from three participants, who completed 12 sessions of an oddball detection task on the THINGS images, in addition to localizer scans. The whole-brain data was collected using a 3T fMRI scanner with a 2mm functional resolution, a 1.5 second TR.

We seek to extend our findings from behavior into the neural domain by leveraging the features present in the THINGS images to better understand a potential neural substrate of memorability. Of particular interest is the analysis of medial temporal lobe (MTL) regions and examining phenomena known as pattern separation and pattern completion as potential neural mechanisms underlying memorability effects and episodic memory more broadly. We intend to leverage the THINGS-fMRI dataset and the features employed in the current manuscript to assess whether a feature-based model can explain observed associations between pattern separation and completion with memorability in the MTL.

Prior findings have implicated the MTL as a region underlying recognition memory (Brewer et al., 1998; LaRocque et al., 2013). Recently, several studies have demonstrated that memorability related information is represented in multiple ventral temporal regions such as the parahippocampal cortex (PHC), the posterior frontal gyrus (pFG), and the perirhinal cortex (PRC) and dissociable from individual memory effects, which are represented in prefrontal regions (Bainbridge, Dilks, & Oliva, 2017; Bainbridge & Rissman, 2018). These results suggest candidate regions of interest including the hippocampus and perirhinal (PRC) and parahippocampal (PHC) cortices as targets for analyses examining the neural mechanisms underlying what we remember and what we forget.

The roles of the hippocampus and other medial temporal regions in declarative memory have been examined for years (Tulving et al., 1998; Eichenbaum, 1999; Eichenbaum, 2001). More recently, researchers have begun to examine the role of these regions in facilitating episodic memory, sparking debate surrounding the nature of the contributions of various MTL structures, including the hippocampus, PRC, and PHC (Eichenbaum and Cohen, 2001; Hannula et al., 2006). One account, known as the complementary learning systems or CLS theory states

that non-hippocampal MTL structures and the hippocampus employ different computational strategies for encoding and retrieving information that together support learning and memory (McClelland et al., 1995; Norman, 2010). These strategies are referred to as pattern separation and pattern completion, which serve to facilitate two critical functions of episodic memory: the ability to store a wide range of potentially overlapping memories while distinguishing between them (pattern separation), and the ability to holistically recall memories when a cue stimulus is perceived (pattern completion; Ngo et al., 2021).

Under the CLS model, MTL regions such as PRC and PHC are thought to build overlapping representations, while the hippocampus is thought to construct separated representations. A recent neural analysis suggests that projections from the dentate gyrus (DG) to the CA3 network of the hippocampus facilitate pattern separation, while entorhinal contributions and recurrent connections facilitate pattern completion (Ngo et al., 2021).

A 2013 study analyzing the CLS model to elucidate the role of medial temporal regions in episodic memory demonstrated a dissociation between the hippocampus and PRC and PHC through visualization of the representational similarity structure within each region of interest. By relating the similarity between image categories (bodies, faces, objects, and scenes) to memory performance on the images, the researchers concluded that memory was best predicted by greater cross-stimulus similarity in PRC/PHC and greater cross-stimulus distinctiveness in the hippocampus (LaRocque et al., 2013). The results of the study track with the CLS model's account that the hippocampus receives pattern separated inputs while the PRC/PHC instead rely on overlapping representations.

In the current manuscript, we demonstrate that a feature-based model accounts for approximately 62% of the variance in memorability across the object concepts, largely captured by semantic features. We also demonstrate that the relationship between memorability and typicality varies across object concepts. Taking these findings along with the discovery of memory performance being associated with higher similarity in PRC/PHC and greater distinctiveness in the hippocampus, we ask whether a feature-based model similar to the one employed in the current manuscript could explain the observed relationship between memory performance and differences in similarity/distinctiveness in these MTL regions. As memorability is quantified using memory performance scores across participants, this question essentially asks if a feature-based model could account for differences in pattern separation and pattern completion within these MTL regions that are sensitive to memorability effects.

This study will be one of the first to relate memorability to a specific set of stimulus features within the brain using a highly representative set of object images. By employing the THINGS-fMRI dataset and the features presented in our behavioral analyses (object space dimensions, DNN activations, behavioral typicality scores), we can test if the predictive power of these features in the behavioral space will translate to a neural model of memorability.

To attempt to answer these questions, we propose a series of analyses on the THINGS-fMRI dataset. We begin with the generation of a per-stimulus Generalized Linear Model (GLM)

to produce an activation map across the whole brain for each of the 8,740 images used in the fMRI task. This would allow for analyses at the individual image, object concept, and higher category levels as implemented within the main manuscript. This is accomplished using AFNI, a toolkit for the analysis of fMRI data. After preprocessing the data (slice time correction, motion correction, alignment of anatomical to functional data), we create an individual model for each of the 8,740 images, essentially creating a map of the response to a given image across the brain. Repeating this process for all images gives us a matrix containing a three-dimensional map across the volume of the brain for each stimulus.

With our whole brain data, we can implement certain quality checks to ensure no problems occurred in acquisition or preprocessing. Prior literature has demonstrated functionally defined regions within the brain that are preferentially active for certain categories of visual stimuli, such as the Fusiform Face Area (FFA, preferentially active for faces) and the Parahippocampal Place Area (PPA, preferentially active for scenes) (Kanwisher et al., 1997). Given that our dataset contains images of faces, scenes, and many other stimulus categories, we can run linear contrasts examining whether these regions display preferential selectivity for face stimuli / scene stimuli. If we are unable to find these effects, it is possible that issues during acquisition or preprocessing have corrupted the data.

Assuming the quality checks do not reveal any major issues with the whole-brain data, we can extract our regions of interest: the PRC, PHC, and Hippocampus. By employing FreeSurfer's recon-all pipeline, we can automatically calculate the coordinates of each of these regions of interest based on anatomical images of each participant. With the coordinates for each region, we can then extract the beta values corresponding to each ROI for analysis in MATLAB.

Before we test what features may explain the association between similarity and distinctiveness in PRC/PHC and the hippocampus with memorability, we must confirm that this pattern holds true within the THINGS-fMRI data. We accomplish this through the implementation of Representational Similarity Analysis (RSA, Kriegeskorte et al., 2008). RSA is a method for analyzing the similarity in representations of information between multiple sources, such as a brain and a computational model, or two brains. In RSA, a matrix of pairwise similarity or dissimilarity scores referred to as a representational similarity matrix (RSM) is created for each source of interest. These matrices can then be correlated together to gauge the degree of similarity in representational structure between the sources. Using RSA, we can examine the similarity structure in each region of interest, as well as generate similarity scores for each stimulus by taking the mean of the correlations between each stimulus and all other stimuli of its object concept. This is the same method as employed in the manuscript to generate the object space derived typicality scores.

If our data also displays an association between similarity in PRC/PHC and distinctiveness in the hippocampus with memorability, we can then generate RSMs for each region of interest by correlating the activity in response to each image within the PRC, PHC, and hippocampus. These RSMs allow for the calculation of similarity scores for each image relative to all others within a given ROI. These RSMs can be correlated with RSMs generated from the

visual and semantic object space dimension loadings used in the current manuscript, which allows for us to test if our prior finding of semantic features exerting stronger influence on memorability is reflected in their influence on the representations in the MTL.

To test if a feature-based model can predict differences in similarity and distinctiveness in PRC/PHC and hippocampus, we begin by taking the mean Pearson correlation between the activity pattern in response to a given image and the activity patterns in response to all other images of its concept, which produces a similarity score for that image. We can then produce difference scores between the ROIs by subtracting similarity scores across ROIs for a given image. These difference scores serve as the dependent variable in our model. For predictor variables, we employ the corrected recognition scores collected for the current manuscript as well as the object space dimension loadings and DNN activations for each image in THINGS-fMRI. This model will allow us to see if our feature-based model for predicting behavioral memorability scores is also able to explain the differences in PRC/PHC and hippocampal representations.

This analysis could represent a significant step forward in our understanding of the neural mechanisms underlying episodic memory, as well as suggesting a potential neural substrate of memorability in pattern separation and completion in medial temporal regions. This work integrates behavioral, neuroimaging, and computational methodologies to tackle a fundamental question in cognitive neuroscience: how we encode, navigate, and retrieve information within the brain.

Conclusion

Here, we have created the best performing model to date of the object features that are predictive of image memorability. From this model, we have observed a primacy of semantic information in determining what we remember. This underscores recent findings of the important role of semantic information in memory (Xie et al., 2020) and emerging work with CNNs that demonstrate a classification performance benefit when including semantic information into their models (Needell & Bainbridge, in press).

Beyond highlighting the roles of semantic and visual information, our results demonstrate that neither prototypicality nor atypicality fully explains what makes something memorable, and if anything, prototypical items tend to be the most memorable. Our findings challenge decades of prior research suggesting we best remember more atypical items (Valentine, 1991; Vokey & Read, 1992; Lee, Byatt, & Rhodes, 2001; Bylinskii et al, 2015; Lukavský & Děchtěrenko, 2017). This trend towards prototypicality is reflected in recent neuroimaging studies (Bainbridge et al., 2017; Bainbridge & Rissman, 2018; Xie et al., 2020), suggesting that prototypicality may be related to the underlying neural mechanisms governing memory.

Our findings shed new light on the features and organizational principles of memory, opening up a wide variety of potential follow-up studies. In fact, with this large-scale analysis, we have identified the stimulus features that govern memorability within and across a comprehensive set of objects, and make this data publicly available for use

(https://osf.io/5a7z6/?view_only=675e901c176c4bec9c2540fc4981e5fe). This will allow researchers to make honed predictions of memory within these categories, or use these dimensions to design ideal stimulus sets. For example, our analysis found that animal images are highly memorable, while manmade, metal images are highly forgettable, and so memorability is an important factor to consider in studies looking at visual perception of animacy (Konkle & Caramazza, 2013). Further, given the success of our feature model in predicting memorability, this model could be potentially used to identify memorable images in other image datasets. While THINGS representatively samples concrete object concepts, there are additional stimulus domains beyond objects including dynamic stimuli such as movies, scenes, and non-visual stimuli that could be analyzed in the context of our results. With the understanding that neither prototypicality nor atypicality alone fully characterizes the relationship between typicality and memorability, there is the question of what biases certain stimuli towards one or the other.

We uncover both a semantic primacy in explaining memorability and determine that the relationship between typicality and memorability is more complex than either prototypicality or atypicality alone. We provide this comprehensive characterization in pursuit of a nuanced understanding of the underlying determinants of memorability, and memory more broadly. Developing this understanding further will have implications far beyond cognitive neuroscience in realms such as advertising, patient care, and computer vision. With the development of generative models of stimulus memorability, it is more important than ever before to ground these models in an empirical understanding of what makes something memorable.

METHODS

Participants

13,946 unique participants completed a continuous recognition repetition detection task on the THINGS images over AMT (see "*Obtaining Memorability Scores for THINGS"*). All online participants acknowledged their participation and were compensated for their time, following the guidelines of the National Institute of Health Office for Human Subjects Research Protections (OHSRP). Participants had to be located within the United States and have participated in at least 100 tasks previously on AMT with at least a 98% approval rating overall to be recruited for the experiment. Participants who made no responses on the task were removed from the data sample.

Stimuli: THINGS

To examine memorability across a broad range of object concepts, we utilized the entire 26,107 image corpus of the THINGS database (Hebart et al., 2019, https://osf.io/jum2f/) for all of our experiments. The THINGS concepts span the wide range of concrete objects, including animate and inanimate, as well as manmade and natural concepts, such as *aardvarks*, *goalposts*, *tanks*, and *boulders*. These 1,854 concepts were generated from the WordNet lexical database through a multilevel web scraping process (Hebart et al., 2019). Each concept has a minimum of

12 exemplar images, though some have as many as 35. These concepts were sorted into 27 overarching categories including *animal-related*, *food-related*, and *body parts*. These higher categories were generated using a two-stage AMT experiment.

At the concept level, we utilized the representational embedding of each concept supplied by THINGS as the multidimensional space for our analyses (Hebart et al., 2020). The original 49-dimensional behavioral similarity embeddings (Hebart et al., 2020) had been generated based on the 1,854 object concepts. Dimension names were generated by two pools of naïve observers in a categorization task (Hebart et al., 2020). The first pool of observers viewed the most heavily reflected dimensions along a given dimension of the space and generated potential labels from the images. The second pool of observers then narrowed down the list of labels until the top two labels remained for each dimension, which was then assigned as the name for that dimension. To derive 49-dimensional embeddings for each of the 26,107 images in the THINGS database, we used predictions from a deep neural network as a proxy. The prediction was carried out for each dimension separately using Elastic Net regression based on the activations of object images in the penultimate layer of the CLIP Vision Transformer (ViT, Radford et al., 2021), which has been shown to yield the most human-like behavior of all available CNN models in a range of tests (Geirhos et al., 2021). The Elastic Net hyperparameters were tuned and evaluated using nested 10-fold cross-validation, yielding high predictive performance in most dimensions (mean Pearson correlation between predicted and true dimension scores: $r > 0.8$ in 20 dimensions, $r > 0.7$ is 32 dimensions, $r > 0.6$ in 44/49 dimensions). We then tuned the hyperparameters on all available data using 10-fold cross-validation and applied the regression weights to the CNN representations of THINGS images, yielding 49-dimension scores for all 26,107 images.

Obtaining Memorability Scores for THINGS

In order to examine memorability in the context of the THINGS space, we collected memorability scores for all 26,107 images (publicly available in an online repository: https://osf.io/5a7z6/?view_only=675e901c176c4bec9c2540fc4981e5fe). To quantify the memorability of each stimulus, each participant viewed a stream of images on their screen and was instructed to press the R key whenever they saw a repeated image. Each image was presented for 500ms, and the interstimulus interval was 800ms. For each repeated stimulus, there was a minimum 60-second delay between the 1st and 2nd presentation of that image, although this delay was jittered so that repetitions could not be predicted based on timing. The task also included easier "vigilance repeats" of 1-5 images apart, to ensure participants were paying attention to the task. The presentation of images was such that approximately 40 participant responses were gathered per image. Of the 1,854 concepts in THINGS, each concept was either represented with a single exemplar or not represented at all during a HIT in order to control for within-concept competition effects on memory performance. To avoid familiarity effects, participants were only allowed to participate again after a minimum delay of 2 weeks.

Memorability was quantified in THINGS using corrected recognition (CR) scores for each image. Corrected recognition is calculated by subtracting the false alarm rate for a given

stimulus from the hit rate for the same stimulus. Hit rate is defined as the proportion of correct repetition detections, whereas false alarm rate is defined as the proportion of incorrect detections. CR allows for a single metric that integrates information about both hit rate and false alarm rate. However, we also replicate all results using hit rate and false alarm rate separately (Supplemental Information).

We ran a split-half consistency analysis to determine if participants were consistent in what they remembered. The analysis randomly partitioned participants into two halves and calculated a Spearman rank correlation between the CR scores for all images, as defined by the two random halves of participants. In other words, this analysis determines how similar the memory performance is for each image between these two independent halves of participants. This process was repeated across 1,000 iterations and an average correlation *rho* was calculated. This *rho* was then corrected using the Spearman-Brown correction formula for split-half correlations. If there is no consistency in memory performance across participants, we would expect a zero value for *rho*, whereas a high value would suggest that what one-half of participants remembered, so did the other. To estimate chance, we correlated one half of participants' scores with those for a shuffled image order of the other participant half, across 1,000 iterations. The p-value was calculated as the proportion of shuffled correlations higher than the mean consistency between halves.

Semantic/Visual Contribution and Regression Model Analyses

With memorability scores at the image level available, we can relate the memorability of THINGS stimuli with the associated representational space and determine the relative contributions of semantic and visual information to memorability. To accomplish this, we analyzed the embeddings of the 1,854 concepts in the 49 dimensions and separated them into semantic and visual dimensions. Of the 49 dimensions, 27 were identified as semantic, 9 as visual, and the remaining 13 as mixed (Table 1).

To determine the effects of semantic and visual dimensions on memorability, we ran a series of multiple regression models. We began with an omnibus model predicting average memorability for each of the 1,854 concepts using the full set of 49 dimensions. This model assessed the total variance in memorability explained by the dimensions. We then utilized a model predicting memorability from the 36 dimensions classified as either semantic or visual to determine the differential contributions of each type of information. As there were more semantic dimensions than visual dimensions, we also ran a model that only used the 9 most heavily reflected semantic and 9 most heavily reflected visual dimensions to control for the overrepresentation of semantic information. In order to assess the potential variance explained by dimensions classified as mixed, we also break down the unique variance contributed by mixed dimensions to the full 49-dimensional model (see supplement). In all models we also analyzed the unique and shared variance contributions of the two types of information to memorability using variance partitioning. Unique semantic variance was calculated as the overall R^2 value for the full model minus the R^2 value for a model containing only the visual dimensions and vice

versa for visual variance. The shared variance was calculated as the overall model R^2 minus both the unique semantic and unique visual variance.

In order to compare the performance of the omnibus model (all 49 dimensions) to the noise ceiling, we conducted a split-half regression analysis. Across 100 iterations, the participant sample was split into two random halves, and we ran two models. For the first model, we looked at the ability of the 49 dimensions to predict the memorability scores derived from the first half of participants. For the second model, we included an additional 50^{th} predictor which was the memorability scores derived from the second half of participants, for the same images. This second model serves as a noise ceiling of memorability from which we can compare the first model. To see the proportion of variance explained in comparison to this noise ceiling, we then averaged the ratio of the R^2 of the first model to the second model, across iterations.

Memorability-Typicality Relationship Analyses

To determine if memorability is highly correlated with prototypicality or atypicality, we assessed the relationship between typicality and memorability of the THINGS images. We conducted these analyses at two levels: the image level, mapping images to concepts, and the concept level, mapping concepts to categories. We utilized typicality scores from behavioral data, the object space dimensions, and the VGG-F convolutional neural network.

For behavioral ratings, we employed the ratings collected as part of the THINGS database (Hebart et al, 2020). These ratings were collected for each of the 1,854 THINGS concepts and represent the typicality of the concept in relation to its higher category on a scale of 0 to 10. For example, the typicality rating for *stomach* under the higher category *body parts* reflects how typical a stomach is as a body part (considering other body parts like legs or shoulders).

We also utilized the object space dimensions to generate typicality scores for each image in relation to its concept. For each concept, we generated a similarity matrix containing the embedding values of the component images of that concept along all 49 dimensions. From that matrix, we can extract a single value for each image that is the average similarity (Pearson correlation) between that image's dimensional embeddings and those of the other images of that concept, which we define as the typicality of that image. In other words, a low mean correlation would imply a highly atypical stimulus (distinct from other exemplars of the same concept), while a high mean correlation would imply a highly prototypical stimulus (very similar to exemplars of the same concept). We utilize the same paradigm to generate typicality values for each concept in relation to other concepts under a given category using an embedding of each concept in the object space and comparing its similarity to the embedding of all other concepts within the same category.

Beyond behavior and the object space, we leveraged the VGG-F object classification CNN to synthesize typicality values for each of the 26,107 images in the THINGS dataset. Early layers of CNNs are more sensitive to low-level image features, such as edges, while later layers

are more sensitive to higher-level and semantic features, such as animacy (Güçlü & van Gerven, 2015). We can therefore extract information at these various points in the network to test the separate contributions of visual and semantic typicality.

The paradigm for extracting typicality values was similar to the object space derived values: for each concept, similarity matrices were generated based on the flattened layer output values for all component images. The typicality for each exemplar was then calculated as the mean of its similarity (Pearson correlations) with all other exemplars in the concept. This measure tells us how similar a given exemplar is to all other exemplars in terms of its CNN-predicted features. This procedure is repeated for every layer in VGG-16, resulting in 21 typicality values for each image in relation to its object concept, one for each layer of VGG-16.

To analyze the relationships between typicality and memorability across the THINGS dataset, we use behavioral, object dimension based, and CNN based typicality values at two different levels of analysis: image level and concept level. At the image level, we analyze the object dimension-derived and CNN-derived typicality values to examine their relationship to memorability across all 26,107 images in THINGS, which gives a single value for the overall typicality-memorability relationship of the THINGS images. Beyond the overall trend, we also examine the relationship within each of the 1,854 image concepts by correlating the typicality scores and memorability scores of their component images. This allows for the visualization of more nuanced relationships between the THINGS concepts. At the concept level, we perform a correlation between the behavioral typicality scores and CR scores and examine the resulting distribution of the relationships for each of the 27 higher categories.

ACKNOWLEDGEMENTS

The researchers would like to thank Coen Needell and Deepasri Prasad for their helpful comments on the manuscript and Sara Hedberg for assistance in generating figures. This research was funded by the Intramural Research Program of the National Institutes of Health (ZIA-MH-002909), under National Institute of Mental Health Clinical Study Protocol 93-M-1070 (NCT00001360).

REFERENCES

1. Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? Journal of Vision, 11(11), 1282–1282. https://doi.org/10.1167/11.11.1282
2. 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
3. Bainbridge, W. A. (2019). Memorability: How what we see influences what we remember. Psychology of Learning and Motivation Knowledge and Vision, 1–27. https://doi.org/10.1016/bs.plm.2019.02.001

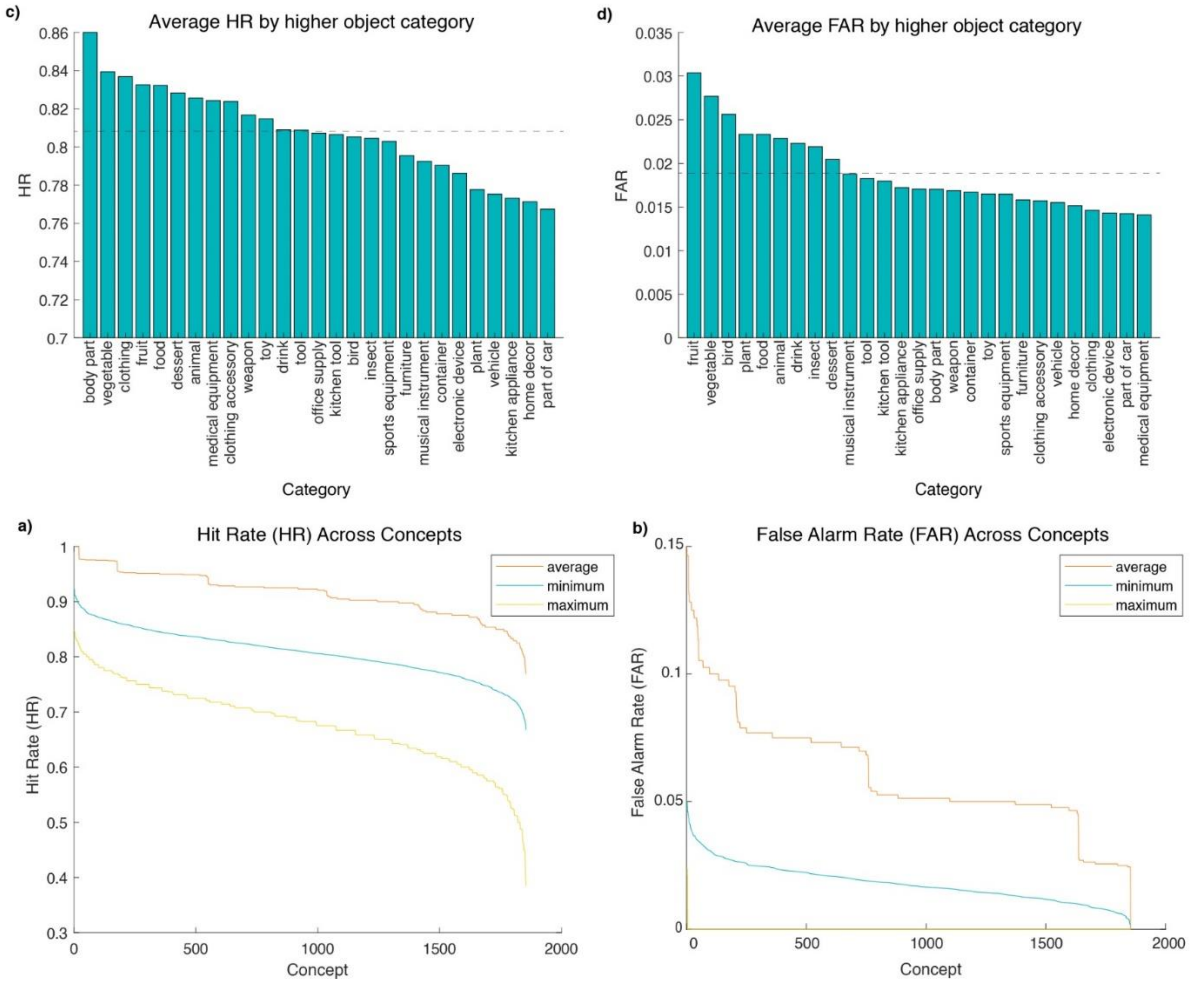
4. Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? IEEE Transactions on Visualization and Computer Graphics, 19(12), 2306–2315. https://doi.org/10.1109/tvcg.2013.234
5. Xie, W., Bainbridge, W. A., Inati, S. K., Baker, C. I., & Zaghloul, K. A. (2020). Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. Nature Human Behaviour, 4(9), 937–948. https://doi.org/10.1038/s41562-020-0901-2
6. Bainbridge, W. A. (2020). The resiliency of image memorability: A predictor of memory separate from attention and priming. Neuropsychologia, 141, 107408. https://doi.org/10.1016/j.neuropsychologia.2020.107408
7. Bainbridge, W. A., Berron, D., Schütze, H., Cardenas-Blanco, A., Metzger, C., Dobisch, L., … Düzel, E. (2019). Memorability of photographs in subjective cognitive decline and mild cognitive impairment: implications for cognitive assessment. https://doi.org/10.1101/660365
8. Needell, C. D., & Bainbridge, W. A. (2021). Embracing New Techniques in Deep Learning for Estimating Image Memorability. ArXiv:2105.10598 [Cs].
9. 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
10. 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
11. Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. The Quarterly Journal of Experimental Psychology Section A, 43(2), 161–204. https://doi.org/10.1080/14640749108400966
12. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1162
13. 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). https://doi.org/10.1109/iccv.2015.275
14. Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. ELife, 8. https://doi.org/10.7554/elife.47596
15. Lin, Q., Yousif, S. R., Chun, M. M., & Scholl, B. J. (2021). Visual memorability in the absence of semantic content. Cognition, 212, 104714. https://doi.org/10.1016/j.cognition.2021.104714
16. Madan, C. R. (2020). Exploring word memorability: How well do different word properties explain item free-recall probability? Psychonomic Bulletin & Review, 28(2), 583–595. https://doi.org/10.3758/s13423-020-01820-w

17. Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature Effects, Distinctiveness, and Identification: Testing the Face-Space Framework. Psychological Science, 11(5), 379–385. https://doi.org/10.1111/1467-9280.00274
18. 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
19. 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
20. Mohsenzadeh, Y., Mullin, C., Oliva, A., & Pantazis, D. (2019). The perceptual neural trace of memorable unseen scenes. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-42429-x
21. Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Wicklin, C. V., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. https://doi.org/10.1101/545954
22. Koch, G. E., Akpan, E., & Coutanche, M. N. (2020). Image memorability is predicted by discriminability and similarity in different stages of a convolutional neural network. Learning & Memory, 27(12), 503–509. https://doi.org/10.1101/lm.051649.120
23. Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-26467-5
24. LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. Journal of Neuroscience, 33(13), 5466–5474. https://doi.org/10.1523/jneurosci.4293-12.2013
25. Ngo, C. T., Michelmann, S., Olson, I. R., & Newcombe, N. S. (2020). Pattern separation and pattern completion: Behaviorally separable processes? Memory & Cognition, 49(1), 193–205. https://doi.org/10.3758/s13421-020-01072-y
26. Contier, O., Hebart, M. N., Dickter, A. H., Teichmann, L., Kidder, A., Corriveau, A., Zheng, C., Vaziri-Pashkam, M., Baker, C. I. (2021). THINGS-fMRI/MEG: A large-scale multimodal neuroimaging dataset of responses to natural object images.
27. Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., … Gorgolewski, K. J. (2017, September). *poldracklab/fmriprep: 1.0.0-rc5*. doi:10.5281/zenodo.996169
28. Hebart, M. N., Zhang, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature Human Behavior. https://doi.org/https://doi.org/10.1038/s41562-020-00951-3
29. Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *Proceedings of the British Machine Vision Conference 2014*. https://doi.org/10.5244/c.28.6

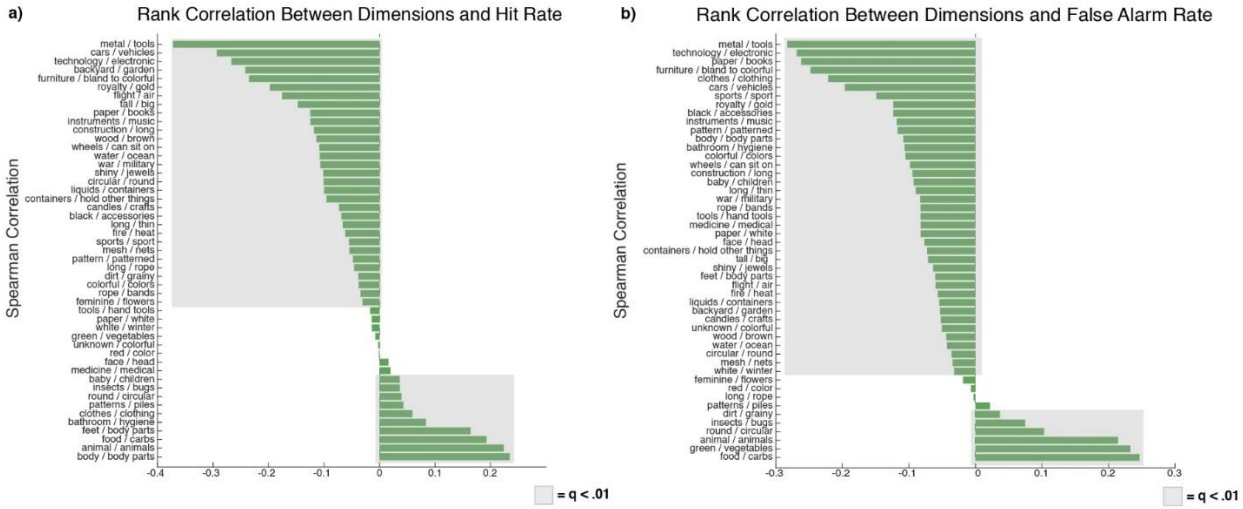
30. Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences, 111(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111
31. Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Computational Biology, 10(11). https://doi.org/10.1371/journal.pcbi.1003915
32. Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/jneurosci.5023-14.2015
33. Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. Memory & Cognition, 20(3), 291–302. https://doi.org/10.3758/bf03199666
34. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. http://arxiv.org/abs/2103.00020
35. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. http://arxiv.org/abs/2106.07411
36. Madan, C.R. (2021). Exploring word memorability: How well do different word properties explain item free-recall probability?. *Psychon Bull Rev* 28**,** 583–595 https://doi.org/10.3758/s13423-020-01820-w
37. Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/jneurosci.0983-13.2013

Supplemental Information for

The Features that Drive the Memorability of Objects



Supplemental Figure 1. This figure was calculated exactly as Figure 1 in the manuscript except using hit rate (HR) and false alarm rate (FAR) rather than corrected recognition (CR). The left side corresponds to HR while the right side corresponds to FAR. As seen in Figure 1 in the main manuscript, we observe a diffusion of memorability across the concepts (A, B) and categories (C, D) of the object space regardless of whether CR, HR (A,C), or FAR (B,D) is used in place of corrected recognition.

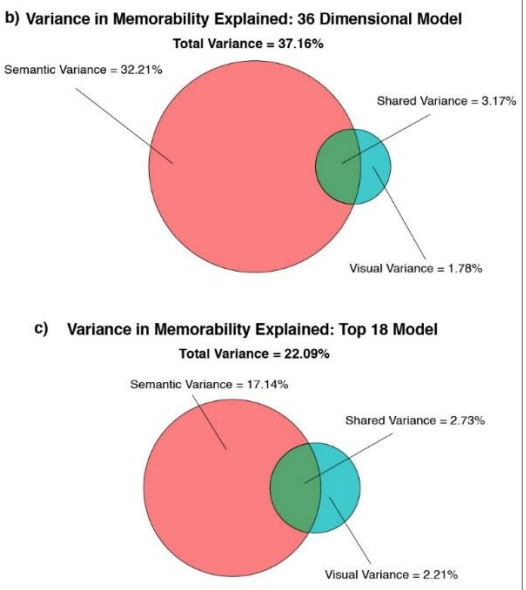


Supplemental Figure 2. This figure was calculated exactly as Figure 1c in the main manuscript except using hit rate (HR) and false alarm rate (FAR) rather than corrected recognition (CR). As in Figure 1c in the main manuscript, we observe that most of the object space dimensions are significantly correlated with (A) HR and (B) FAR, even when accounting for multiple comparisons.

a) Summary of Regression Results: Hit Rate

Model	R^2	F	p	error
All Dimensions	0.4014	24.6786	1.7417E-164***	0.0011
Semantic	0.3538	37.0049	1.4119E-151***	0.0012
Visual	0.0495	10.6754	2.9651E-16***	0.0017
Top Semantic	0.1988	50.7966	1.5291E-82***	0.0014
Top Visual	0.0495	10.6754	2.9651E-16***	0.0017

*** = p < .001

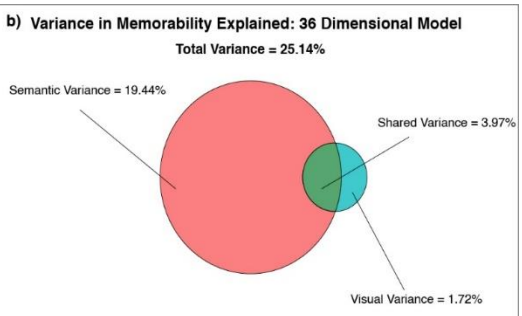


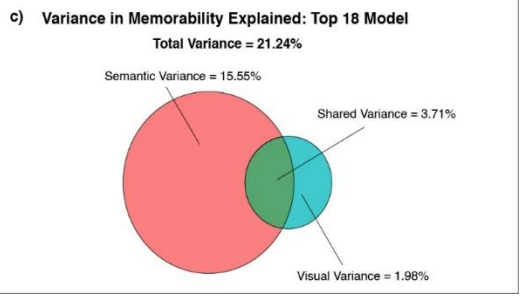
Supplemental Figure 3. This figure was calculated exactly as Figure 2 in the main manuscript but using hit rate rather than corrected recognition. (A) Regression output from all models. Utilizing HR as the dependent variable results in slightly higher performance across all models, leading to a 49-dimensional model capturing 40.14% of the variance in hit rate. (B) Venn diagram for the model excluding mixed dimensions. (C) The same type of Venn diagram as (B) but for the top 9 semantic and visual dimensions, leading to an 18-dimensional model.

a) Summary of Regression Results: False Alarm Rate

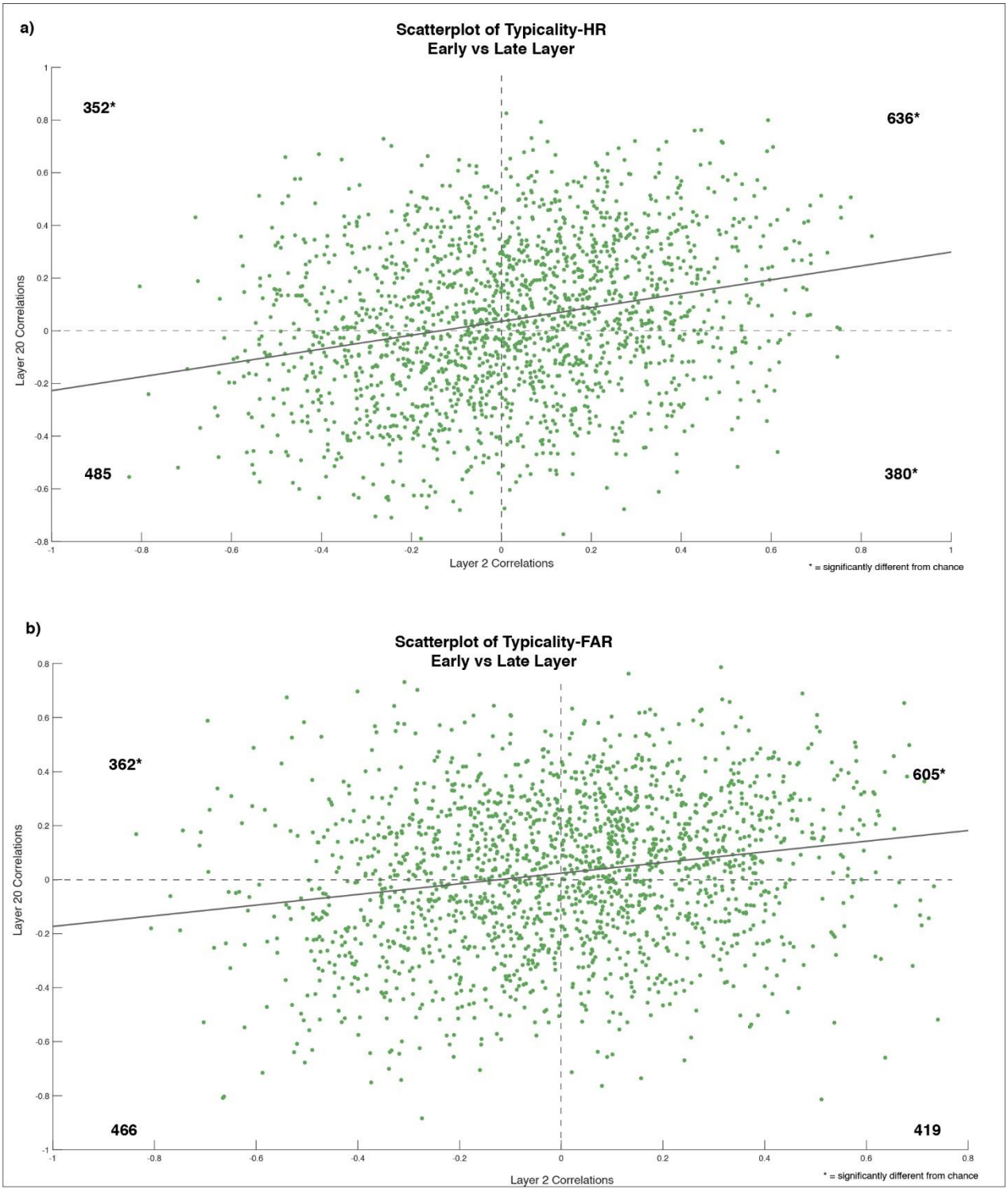
Model	R^2	F	p	error
All Dimensions	0.2748	13.9514	2.8881E-93***	3.93E-05
Semantic	0.2342	20.6806	1.4412E-86***	4.10E-05
Visual	0.0571	12.3786	3.1474E-19***	5.00E-05
Top Semantic	0.1926	48.8823	1.3876E-79***	4.28E-05
Top Visual	0.0571	12.3786	3.1474E-19***	5.00E-05

*** = $p < .001$



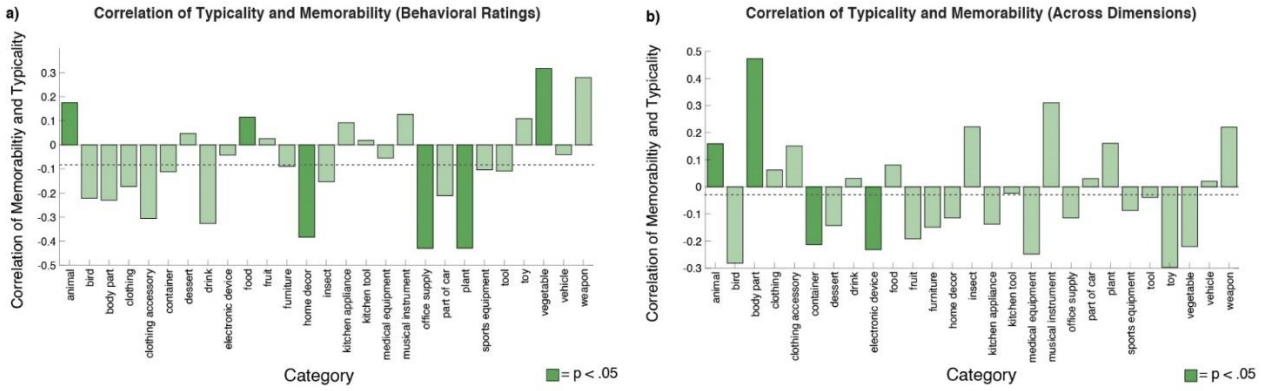


Supplemental Figure 4. This figure was calculated exactly as Figure 2 in the main manuscript but using false alarm rate rather than corrected recognition. (A) Regression output from all models. Utilizing FAR as the dependent variable results in slightly lower performance across all models, leading to a 49-dimensional model capturing 27.48% of the variance in false alarm rate. (B) Venn diagram for the model excluding mixed dimensions. (C) The same type of Venn diagram as (B) but for the top 9 semantic and visual dimensions, leading to an 18-dimensional model.



Supplemental Figure 5. This figure is calculated the same as Figure 4a in the main manuscript except for the use of hit rate and false alarm rate rather than corrected recognition. (A) When testing with hit rate, a chi square analysis of the four quadrants of the scatterplot demonstrated significantly more concepts than chance showed a pattern where the most memorable items were prototypical in terms of both early and late layer features ($\chi^2 = 38.588$, $p = 5.235 \times 10^{-10}$). Contrastingly, we find significantly fewer concepts that demonstrate "mixed" patterns where more

memorable items demonstrated early layer prototypicality and late layer atypicality ($\chi^2 = 10.638$, $p = 0.001$), or the opposite pattern ($\chi^2 = 19.460$, $p = 1.027 \times 10^{-5}$). We found no significant difference from chance for concepts where the most memorable items were atypical across both early and late layer features ($\chi^2 = 0.670$, $p = 0.413$). (B) when testing false alarm rate, we also observed significantly more concepts than chance showed a pattern where the most false-alarmable items were prototypical in terms of both early and late layer features ($\chi^2 = 26.421$, $p = 2.745 \times 10^{-7}$). We do not find a significant difference from chance for concepts where more false-alarmable items demonstrated early layer prototypicality and late layer atypicality ($\chi^2 = 2.912$, $p = 0.088$), or early layer atypicality and late layer prototypicality ($\chi^2 = 0.011$, $p = 0.917$). We also found no significant difference from chance for concepts where the most false-alarmable items were atypical across both early and late layer features ($\chi^2 = 15.979$, $p = 6.406 \times 10^{-5}$).



Supplemental Figure 6. Examples of mixed typicality-memorability relationships across categories. (A) The correlation between behavioral ratings of typicality and memorability across the categories was strong for *home décor, office supplies, and plants* (where atypical concepts are more memorable) as well as *animals, food,* and *vegetables* (where prototypical concepts are more memorable). (B) The correlation between dimension-based scores of typicality and memorability across the categories. *Containers* and *electronic devices* display negative relationships, while a*nimals* and *body parts* demonstrate positive relationships. For both behavioral and dimension-based visualizations (A & B), the overall average correlation between typicality and memorability is visualized as a dotted line.

Supplemental Table 1. Summary of typicality results. This table reproduces the correlations between object-space derived typicality and corrected recognition, hit rate, and false alarm rate, as well as t-tests across all concepts for each metric. Compared to corrected recognition, we observe a similar pattern of results for hit rate, with an overall positive significant relationship between typicality and memorability, which is also reflected in a general trend towards more concepts displaying positive relationships. We also observe a positive overall relationship when examining the correlation between false alarm rate and object space-derived typicality, but we do not find a significant difference from a normal distribution with a mean of 0 when testing the concepts for relationships between false alarm rate and typicality.

Metric	Full Dataset Correlation (r)	*p*	Distribution of Correlations (t)	df	*p*
CR	0.309	6.13E-07***	2.074	1852	0.038*
HR	0.0526	1.90E-17***	2.6687	1852	0.008***
FAR	0.077	1.24E-35***	1.1375	1852	0.256

* = $p < .05$ *** = $p < .001$

Supplemental Results

Impact of "mixed" labelled dimensions on memorability

To assess the proportion of variance in memorability explained by dimensions classified as mixed (Table 1), we examine the unique and shared variance contributions of mixed, semantic, and visual dimensions in the omnibus 49-dimensional model. We see that mixed dimensions uniquely contribute 1.03% of the variance in corrected recognition captured by the model.

Hit rate and false alarm rate analyses on CNN-derived typicality

When using hit rate rather than corrected recognition, we observe a similar pattern of results as in Figure 4a in the main manuscript. As with Figure 4a, we find that significantly more concepts than chance showed a pattern where the most memorable items were prototypical in terms of both early and late layer features ($\chi^2 = 38.588$, $p = 5.235 \times 10^{-10}$). We also find significantly fewer concepts than chance show a mixed pattern, where memorable items were determined by early layer prototypicality and late layer atypicality ($\chi^2 = 10.638$, $p = 0.001$), or the opposite pattern of early layer atypicality and late layer prototypicality ($\chi^2 = 19.460$, $p = 1.027 \times 10^{-5}$). We found there was no difference from chance in the proportion of concepts that showed a pattern where the most memorable items were the most atypical items for both early and late CNN layers ($\chi^2 = 0.670$, $p = 0.413$).

Using false alarm rate changes the pattern of results slightly from what is visible in Figure 4a and supplemental figure 4a. As with CR and HR analyses, we observed significantly more concepts than chance displayed a pattern where the most false-alarmable items were prototypical across early and late layer features ($\chi^2 = 26.421$, $p = 2.745 \times 10^{-7}$), but we did not see a significant difference from chance for either the concepts where the most false-alarmable items were visually prototypical and semantically atypical ($\chi^2 = 2.912$, $p = 0.088$) or where the most false-alarmable items were visually atypical and semantically prototypical ($\chi^2 = 0.011$, $p = 0.917$). As before, we found no significant difference from chance for the quadrant where the most false-alarmable items were atypical across early and late layers ($\chi^2 = 15.979$, $p = 6.406 \times 10^{-5}$).

Relating CNN-derived typicality and memorability within CNN layers

When examining the CNN derived typicality scores as they relate to memorability, we found no significant relationship ($p > 0.985$) between the typicality scores derived across the 21 layers of the network and memorability. All 21 of the observed correlations failed to exceed a maximum magnitude of 0.05, suggesting that this image-computed measure of typicality is not a strong predictor of memorability.