Memory & Cognition (2024) 52:595–609
https://doi.org/10.3758/s13421-023-01485-5

Comparing memory capacity across stimuli requires maximally dissimilar foils: Using deep convolutional neural networks to understand visual working memory capacity for real-world objects

Timothy F. Brady[1] · **Viola S. Störmer**[2]

Accepted: 17 October 2023 / Published online: 16 November 2023

Abstract
The capacity of visual working and visual long-term memory plays a critical role in theories of cognitive architecture and the relationship between memory and other cognitive systems. Here, we argue that before asking the question of how capacity varies across different stimuli or what the upper bound of capacity is for a given memory system, it is necessary to establish a methodology that allows a fair comparison between distinct stimulus sets and conditions. One of the most important factors determining performance in a memory task is target/foil dissimilarity. We argue that only by maximizing the dissimilarity of the target and foil in each stimulus set can we provide a fair basis for memory comparisons between stimuli. In the current work we focus on a way to pick such foils objectively for complex, meaningful real-world objects by using deep convolutional neural networks, and we validate this using both memory tests and similarity metrics. Using this method, we then provide evidence that there is a greater capacity for real-world objects relative to simple colors in visual working memory; critically, we also show that this difference can be reduced or eliminated when non-comparable foils are used, potentially explaining why previous work has not always found such a difference. Our study thus demonstrates that working memory capacity depends on the type of information that is remembered and that assessing capacity depends critically on foil dissimilarity, especially when comparing memory performance and other cognitive systems across different stimulus sets.

Keywords Memory capacity · Visual working memory · Similarity · Deep learning · Convolutional neural networks

Introduction

One major focus of memory research is to examine, and quantify, how much information we can remember. For example, studies of visual long-term memory have shown that people can remember massive amounts of visual information (Brady et al., 2008; Standing, 1973), and attempts have been made to quantify the upper bounds on this capacity (e.g., Landauer, 1986) in order to understand how memory might limit or interact with other cognitive systems (e.g., object recognition; Palmeri & Tarr, 2008).

✉ Timothy F. Brady
tfbrady@ucsd.edu

[1] Department of Psychology, University of California San Diego, La Jolla, CA 92093, USA

[2] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

The domain of memory where capacity has been seen as most relevant, however, is visual working memory. In contrast to long-term memory, visual working memory is used to hold visual information actively in mind for relatively short periods of time and has a stark capacity limit (Baddeley, 2012; Cowan, 2001). Importantly, individual differences in this capacity limit are closely related to differences in fluid intelligence and academic achievement (Alloway & Alloway, 2010; Babic et al., 2019; Fukuda et al., 2010), which is one of the reasons why describing and understanding these limits is of broad general interest. Theories markedly differ in their claims about the nature of the capacity limits of this system (e.g., Adam et al., 2017; Bays et al., 2022; Luck & Vogel, 2013; Ma et al., 2014; Schurgin et al., 2020).

One particularly important question distinguishing theories of working memory limits has been whether the stimulus itself matters for working memory capacity: that is, does working memory capacity differ fundamentally for different stimuli, or is there a single fixed visual working memory capacity regardless of what we are remembering? Many

studies have asked this question, comparing, for example, working memory for simple unidimensional stimuli (e.g., colors, orientations) to more complex meaningless stimuli (Alvarez & Cavanagh, 2004; Awh et al., 2007) or to realistic objects (e.g., Brady & Störmer, 2022; Brady et al., 2016; Li et al., 2020; Quirk et al., 2020), and making claims about the fundamental nature of working memory capacity as a result (e.g., fixed numbers of objects or not; fixed capacity or not). In the current work we focus on one particularly critical aspect of making such comparisons – the similarity of foil or lure items to the remembered item during the memory test. We demonstrate how this seemingly simple aspect of such studies can fundamentally change our inferences about the underlying capacity of visual working memory for different stimuli.

Standard assessments of visual working memory – particularly for stimuli that cannot make use of continuous adjustment tasks – ask participants to remember several items and then, at test, to distinguish whether a given item is old or new ("change detection," where "new" items are the relevant "lure"), or which of two items has been previously seen ("two-alternative forced-choice" (2-AFC), where the "new" item is the "foil"). In either case, the similarity of the lure or foil item is critical to performance (e.g., Awh et al., 2007; Keshvari et al., 2013; Schurgin et al., 2020). For example, if remembering the color red, then performance will be higher when your memory is tested against a blue lure than if you are asked to distinguish between the red item you saw and an orange lure item in both forced-choice (e.g., Schurgin et al., 2020) and change detection (Keshvari et al., 2013).

In the case of color and other simple low-level stimuli often used in visual working memory tasks, foils that are maximally distinct are commonly used – at minimum, large cross-category color differences (e.g., Luck & Vogel, 1997). In many cases where continuous color spaces are used (e.g., Wilken & Ma, 2004), foils are even more distinct: for example, when probing memory with a 2-AFC, many studies have used colors that are 180° away on the chosen color circle (Brady & Störmer, 2022; Brady et al., 2016; Li et al., 2020; Quirk et al., 2020), very close to as far as possible apart in that feature space given the color circles are generally chosen to be approximately maximally large for a given luminance level (Zhang & Luck, 2008). This decision is, it turns out, critical: maximally distinct foils from within the feature space are the only appropriate way to measure performance in a feature space, since performance can arbitrarily be driven lower from that point but never higher: If participants were asked to distinguish two extremely similar colors at test (e.g., red vs. orange), performance would necessarily be lower, relative to when they are asked to distinguish between very distinct colors (e.g., red vs. blue); but within a given luminance level, performance can never be higher than when the foil is maximally far apart from the target color in hue. For color, then, performance on these commonly used tasks with maximally dissimilar foils is interpretable as approximately the upper bound of memory performance for that particular feature space, and thus is a valid measure of the "capacity" of the memory system for colors.

By contrast, there is considerably more variability in the way foils have been selected when assessing both long-term memory capacity (e.g., Brady et al., 2008) and working memory capacity for complex objects (e.g., Alvarez & Cavanagh, 2004, vs. Awh et al., 2007) and realistic, meaningful objects (e.g., Brady et al., 2016; Quirk et al., 2020), and this has significant potential implications for conclusions about the nature of memory capacity. As is the case for colors, memory performance is necessarily lower when participants are asked to make comparisons between more similar objects at test: for example, people perform worse with within-category foils (Awh et al., 2007; Brady et al., 2009, 2016; Schurgin & Brady, 2019), and the similarity of foils impacts performance for both realistic objects and complex but meaningless objects (Frank et al., 2020; Mate & Baqués, 2009). This variance – insofar as it produces differences in performance – is problematic for interpretation: Just like for more simple feature spaces like color, when trying to assess overall memory capacity for complex or realistic objects, it is necessary to choose maximally distinct foils, or else performance will be arbitrarily lower – and the more similar the object foils, the lower performance will be. This is particularly important if the goal is to compare memory capacity across different stimulus types: color versus orientation, color versus real-world objects, real-world objects versus scenes, etc. Only when foil distance at test is maximal within each stimulus space, it is possible to make inferences about the nature of capacity limits.

One example of this from visual long-term memory is work by Brady et al. (2008) that sought to assess and quantify visual long-term memory capacity. Although aiming to measure the upper bound of memory performance, Brady et al. (2008) did not use maximally dissimilar foils, but instead randomly chosen cross-category foils from their stimulus set. Despite the impressive performance they found, to the extent that randomly chosen foils are not maximally dissimilar, they likely underestimated the true performance level that can be achieved in visual long-term memory. Similar questions exist about working memory tasks, where again random cross- category foils have often been used (e.g., Li et al., 2020; Quirk et al., 2020). To determine whether this affects the capacity estimated in these long-term memory and working memory studies, assessing whether random cross-category objects are approximately "maximally dissimilar" is key, and this is what we probe in the current work.

Although this issue is of broad relevance to many findings (e.g., Brady et al., 2008, in long-term memory), rather than test it in all domains, we focus on one particular example of this general issue: the case of comparing working memory capacity for objects to other simple stimuli, like colors. Comparing memory performance across different stimuli is a case where the strategy used to choose foils is particularly critical, especially if the way the foils are chosen is different for different stimuli. In the case of meaningful objects versus colors in particular, most work has found significant working memory performance advantages for real objects when compared to simple stimuli like colors (e.g., Brady & Störmer, 2022; Brady et al., 2016; Torres et al., 2023) or scrambled stimuli (e.g., Brady & Störmer, 2022; Shoval & Makovski, 2022; Starr et al., 2020; Thibeault et al., 2023), where other work has not found differences between objects and colors (e.g., Li et al., 2020; Quirk et al., 2020). This has led to different assessments of the nature of working memory storage, and in particular the impact of knowledge and familiarity on memory capacity.

Interestingly, whereas the color foils were chosen to be maximally distinct in CIE *L*a*b* space, in all of these studies, the studies that found better memory capacity for objects than colors (Brady & Störmer, 2022; Brady et al., 2016; Thibeault et al., 2023; Torres et al., 2023) used approximately maximally distinct object foils, whereas the studies that did not (Li et al., 2020; Quirk et al., 2020) used random cross-category foils. Even though choosing cross-category foils ensures no very similar objects are presented as the foil object, there are still significant gradations of how similar the objects are semantically and visually (e.g., an office chair and a dining room chair are both included, and while they are dubbed distinct categories in the stimulus set, it is clear that they are similar; likewise, a tape measure and a ruler are both included; etc.) – so choosing foils randomly from other categories than the target stimulus may not maximize semantic or visual distinctiveness across all trials. Thus, this particular literature provides an interesting case study that motivates the question of whether random cross-category objects are approximately maximally dissimilar, or whether, in assessing object memory, foils should be chosen in a way that more directly maximizes such dissimilarity to ensure capacity is being accurately measured.

While creating maximally distinct foils is fairly straightforward for colors and other simple low-level features – because we can relatively directly measure perceptual distinctiveness, for example as distance in CIE *L*a*b* space for color – it is not obvious what a maximally distinct foil should be for a real-world object. Past work largely used intuition to do this (e.g., Brady et al., 2016). In the current study, to operationalize this, we formally assessed real-world object similarity using a deep convolutional neural network (CNN) that allows us to objectively identify distinct foils, similar to how foils for simple visual stimuli like color are chosen. In particular, we use the VGG16 pre-trained convolutional neural network to assess image similarity between each pair of objects from the thousands of objects in the Brady et al. (2008) stimulus set. This convolutional neural network is trained to go from images (pixels) to category labels (e.g., ball), and, in the past 10 years, such networks have become increasingly good not only at solving object recognition from images, but also increasingly good as models of visual processing in the primate visual system (e.g., Yamins et al., 2014). Thus, they provide a useful way to find approximately maximally dissimilar object images without requiring participants to give similarity ratings on millions of pairwise object sets. At the same time, using such networks provides additional information about the networks as models of human information processing: to the extent VGG16-chosen foils are indeed more distinct than random foils, this suggests convolutional networks of this type provide a useful model of human memory confusability. Note that we focus on confusability between pairs, as we are interested in issues of capacity, but similar work using deep nets has also been done looking for images that are not confusable with other images on average ("memorability"; Needell & Bainbridge, 2022).

Overall, in a series of high-powered, preregistered experiments, we find that (1) foils chosen to be dissimilar by VGG16 are, in fact, particularly dissimilar according to human participants; (2) this dissimilarity modulates memory performance, such that participants perform better at memory tasks with such dissimilar foils than the randomly chosen cross-category foils sometimes used in previous work, and (3) when we use equivalently chosen, maximally dissimilar foils for both colors and objects, we find a large benefit for objects relative to colors. Our results thus demonstrate that realistic objects result in better working memory performance than simple features once foil dissimilarity is matched. More broadly, our results emphasize the importance of maximizing foil dissimilarity when measuring capacity; comparing different sets of stimuli on common ground, a general challenge in many studies trying to compare cognitive operations across different stimulus sets; and point to a means of generating such foils objectively and automatically.

Experiment 1: Stimulus creation and validation

For objects, we used the Brady et al. (2008) object image database. This database contains 2,400 objects, with, as best as possible, one object per "basic-level category" (e.g., one office chair). In our first experiment, we ask whether, within this database, there is variability in how similar randomly

chosen pairs are, despite the objects being cross-category. In subsequent experiments we ask whether this variability is critical to explaining differences in visual working memory performance. Experiment 1 thus measures the similarity of different pairs of objects within the database, using both convolutional neural networks and human similarity ratings, as a precursor to the memory experiments (Experiments 2 and 3).

With 2,400 distinct objects there are potentially greater than 5 million possible pairs of stimuli that could serve as the studied item/foil. Thus, finding the maximally dissimilar pairs from this set is not straightforward. To address this, we make use of deep convolutional neural networks to achieve this goal. Recently deep convolutional neural networks have reached very high performance on object recognition tasks, reaching, and even surpassing in limited circumstances, human category-level recognition performance (He et al., 2016; Kietzmann et al., 2019; Lindsay, 2020; Russakovsky et al., 2015; Yamins & DiCarlo, 2016). Furthermore, evidence suggests that these networks capture some aspects of the neural representations of visual information, as they have been shown to account for neural data recorded in higher-level visual areas of the human and non-human primate brain (Güçlü & van Gerven, 2015; Kar et al., 2019; Khaligh-Razavi & Kriegeskorte, 2014; Kubilius et al., 2018; Yamins et al., 2014). While there is still some debate of how well CNNs capture human vision (e.g., Xu & Vaziri-Pashkam, 2020), and different model architectures result in different performance levels and how well they explain neural representations (e.g., Storrs et al., 2021), overall it is clear that they provide one of the best current computational tools to model human object recognition, and, of most importance to the current paper, are sensitive to visual and category-based similarity of objects (Peterson et al., 2018). Thus, they provide a useful tool for choosing object pairs that will provide a fair assessment of the upper bound on memory performance, allowing us to measure memory capacity in a fair way.

Method

Creating pairs

To choose maximally dissimilar and maximally similar foils from within the Brady et al. (2008) database and see how they compare to randomly selected foils, we used a VGG16 convolutional neural network (CNN) architecture (Simonyan & Zisserman, 2014), pretrained on ImageNet, to select images based on image-level similarity. This model is a deep convolutional neural network with 16 layers, and among the most well-known and well-cited models of its kind. It achieves 92.7% top-5 test accuracy (i.e., the correct label was in the top five suggestions by the network) on ImageNet, a dataset of over 14 million images.

Using this model allowed us to create "maximally dissimilar" and "maximally similar" object foils, at least with respect to the features of this network. To do so, we calculated the features of all of our object images using the CNN for all objects in our database, using the Keras implementation of ImageNet-pretrained VGG16 in Tensorflow. Then we used the CNN feature matrix from the top max-pooling layer (with images 256×256, $8 \times 8 \times 512 = 32{,}768$ features/image) to compute similarity between all pairs of objects (cosine similarity; i.e., length-normalized dot product), and chose both the 120 most dissimilar and most similar pairs with only the constraint that no object appeared in multiple pairs (Figs. 1 and 2). Creating the maximally similar pairs revealed that one image was duplicated in the Brady et al. (2008) set, and so we replaced this image. We used the top max-pooling layer as we were most interested in the features that are used for categorization/classification, as opposed to the lower layers that are more similar to earlier visual regions (e.g., Eickenberg et al., 2017; Yamins et al., 2014).

Note that despite being a model that works on image features, such networks are trained to do categorization, and so even though we use the network for feature extraction rather than categorization, the top layers of such models contain category-specific information, designed to be read-out by the fully connected categorization layers that are omitted when doing feature extraction. Thus, such deep nets trained on categorization are sensitive to some extent to both visual and semantic features (e.g., Jozwik et al., 2017; Peterson et al., 2018). Given that such deep convolutional neural networks are useful models of human recognition and the human visual system (e.g., Yamins et al., 2014), these pairs should be more dissimilar than randomly chosen pairs, although they may not be as dissimilar as pairs chosen specifically to avoid similarity by humans (as used in Brady et al., 2016). In particular, networks like VGG and ResNet have been criticized for not precisely matching humans in pivotal ways, for example, relying more on texture than global shape (unlike humans; Geirhos et al., 2018), which may affect their ability to be a completely accurate model of object similarity. However, they nonetheless provide a useful benchmark. We have publicly shared the code to extract features for the objects at: https://colab.research.google.com/drive/1vKpxABn0J9vi_vPleBb-XqdrIQmhgc4L.

Validation of pairs with another convolutional neural network

Our goal is not to fully assess convolutional neural nets as models of human behavior and neural processing (e.g., Storrs et al., 2021). We simply wish to use such convolutional nets as a tool to choose dissimilar foils. So we

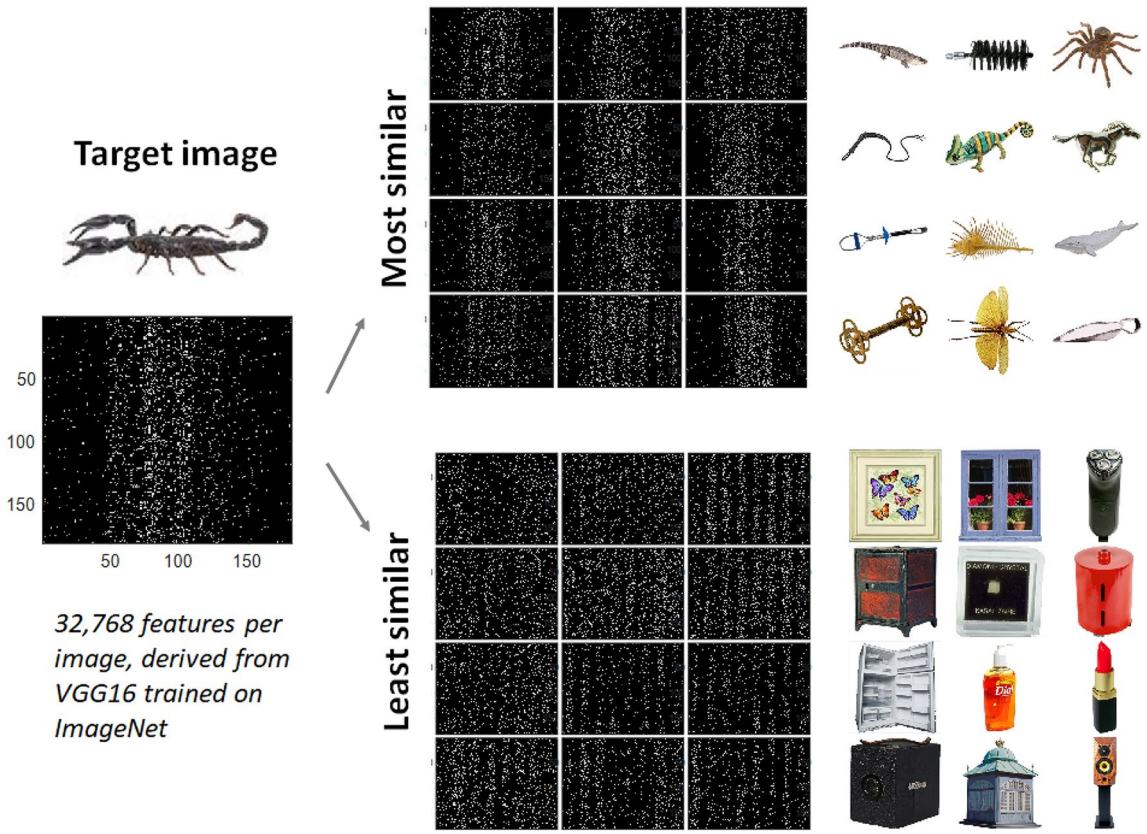

Fig. 1 Designing maximally distinct foils using a deep convolutional neural network to measure similarity. From each image, we extracted features from VGG16, a network trained to do object classification on ImageNet. The features for each object are visualized in the black/white plots, which show white for more activity for a given feature and black for less. We used these features to select, for each image, the most dissimilar foil. Here, we visualize the most similar and most dissimilar items for one particular target image. Deep neural networks are trained on categorization, and thus are sensitive to both visual and semantic features. For example, note that the most similar items to the example target image are largely animals, with similar shape and texture; whereas the least similar items are inanimate, with different shapes and textures

compared the similarity of VGG16-chosen-foils to only one other convolutional net – helping ensure the convolutional net similarities were not totally idiosyncratic to VGG16. In particular, we picked a distinct architecture, ResNet (He et al., 2016), and used ResNet-18 via the PyTorch implementation (Paszke et al., 2017), again focusing on the final max pooling layer. We assessed the similarity (again using cosine similarity) of all the pairs chosen by VGG16 in this network (see Fig. 3). Both ResNet-18 and VGG-16 consist of multiple convolutional layers that learn features from the input images. These convolutional layers are stacked to capture increasingly abstract and hierarchical features, and it has been found that lower layers of such networks are most similar to early visual cortex, i.e.,V1, and higher layers of these networks are more similar to higher-level brain regions like V4 and IT (e.g., Eickenberg et al., 2017; Yamins et al., 2014). However, these networks differ significantly in architecture: VGG-16 follows a simple and uniform architecture where each convolutional block consists of convolutional layers followed by max-pooling. In contrast, ResNet-18 uses a residual architecture with residual blocks. Residual blocks contain shortcut connections (skip connections) that allow the network to skip one or more layers, making it easier to train deep networks. We used versions of each network trained on ImageNet.

Validation of pairs with human similarity data

Participants All experiments were approved by the Institutional Review Board at UC San Diego. Fifty US-based participants (33 male, 16 female, one other/chose not to say) from Amazon's Mechanical Turk were recruited to perform a similarity task (all with $\geq 95\%$ previously accepted submissions in their previous tasks performed on MTurk ["HITs"]).

Stimuli and procedure Participants completed 120 trials overall, 40 from each of the three conditions (maximally distinct cross-category object-foils; random cross-category

Fig. 2 Representative pairs of objects that were derived from the convolutional neural net and that served as the study item and foil items in the memory experiments (i.e., one item from each pair would be studied and the other would serve as the foil at test for that item). The "maximally similar" and "maximally distinct" pairs chosen by the convolutional neural network features appear to be more similar, or more divergent, both semantically and visually than randomly chosen pairs despite all the pairs being "cross category." Also noteworthy is that despite being cross-category, there exist pairs in the dataset that are incredibly similar, both conceptually and visually

object-foils; maximally similar cross-category object foils). The 40 pairs in each condition for a given participant were chosen randomly from the available set of 120 pairs per condition. On each trial, participants were shown an object and asked "How similar are these objects?" on a scale from 1 (not similar) to 7 (maximally similar). We told participants we were interested in their intuitive visual judgments, and not the similarity of the words you might use to describe the objects, but otherwise left the task open-ended. Trials were not speeded and participants responded with the keyboard buttons 1–7. When participants responded, the pair of objects disappeared, followed by a 500-ms delay and then the next pair of objects. Pairs from all conditions were interleaved randomly.

Results

VGG16, naturally, predicts extremely low similarity for maximally dissimilar pairs, and quite high for maximally similar, as this is the network that was used to select the pairs and so these similarities are not independent of the condition they are in. Note, however, that in the maximally similar condition, similarities are nowhere near 1, as all of the images available in the Brady et al. (2008) dataset are distinct and cross-category.

The similarity of the two deep convolutional neural networks was extremely strongly related. Their measured similarities as a function of stimulus kind (VGG16-generated pairs and randomly paired objects) are plotted in Fig. 3.

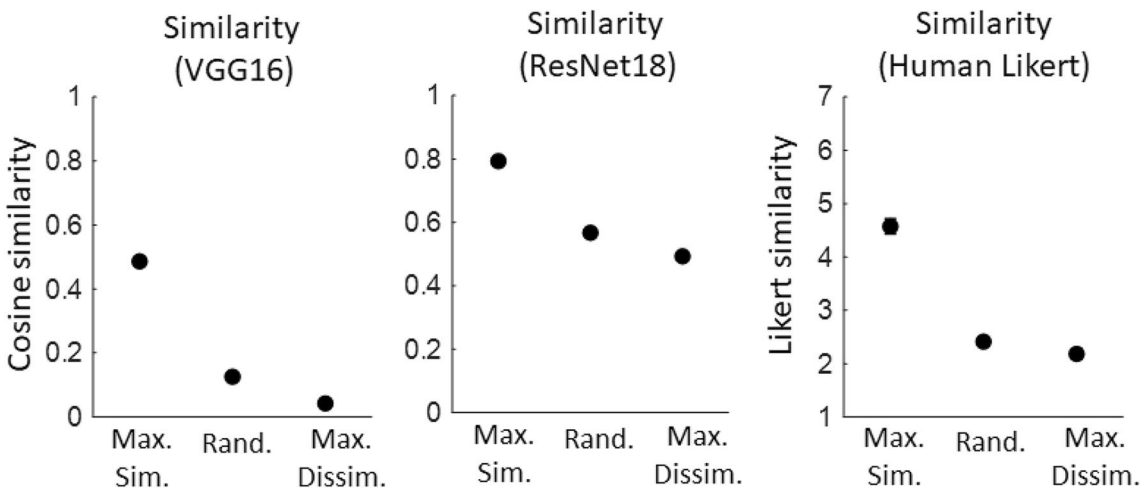

Fig. 3 Similarity of each set of pairs, from both two different deep convolutional neural networks and from human similarity ratings. The similarity measures indicate that the VGG16 network effectively chose maximally similar and maximally dissimilar pairs, which differed from randomly chosen pairs in the predicted directions. Error bars are present on all graphs (by pair for the networks; by participant for the human similarity) but are generally smaller than the dots themselves

ResNet-18 predicts nearly the same pattern as VGG16, despite not being used to select the stimuli, although there is an overall shift where all stimuli pairs are more similar in ResNet-18 than in VGG16. Across all 360 pairs, similarity derived from VGG16 and ResNet-18 had a correlation of $r = 0.93$ ($p < 0.0001$), suggesting both tap extremely similar features of real-world objects for the purpose of constructing such pairs.

Human Likert ratings agree that in aggregate, the maximally similar pairs chosen by VGG16 are more similar than randomly paired stimuli ($t(49) = 10.87$, $p < 0.001$, $d_z = 1.54$); and maximally dissimilar pairs are less similar than randomly paired stimuli ($t(49) = -4.90$, $p < 0.001$). Note that this later effect was small in absolute terms but highly reliable ($d_z = 0.69$), in part because many participants gave all object pairs that were clearly of a different category the lowest similarity level.

How well do the networks capture human similarity? Even though human similarity ratings for objects generally depend on many semantic and visual features (e.g., Hebart et al., 2020), both captured effectively all of the explainable variance in human similarity ratings for this set of 360 pairs. We took 1,000 random split half correlations from the human similarity data and corrected them using the Spearman-Brown formula to get an estimate of the reliability of the human ratings for each of the 360 pairs. This gives a noise ceiling (e.g., a maximum expected correlation given the measurement noise of the human data per pair) of approximately $r = 0.71$. The similarity ratings derived from ResNet18 correlate with the human data with $r = 0.76$ ($p < 0.0001$) and the similarity ratings derived from VGG16 correlate with the human data with $r = 0.75$ ($p < 0.00001$) – at the noise ceiling of the human similarity measure.

Overall then, stimulus selection and validation reveals that both VGG16 and ResNet-18 provide extremely good predictive power for choosing both minimally and maximally similar pairs for object images; and that, as expected, randomly chosen pairs from the Brady et al. (2008) stimulus set are not maximally dissimilar, making them not a valid measure of memory performance for comparing across stimulus sets (e.g., for comparing with maximally dissimilar color foils).

Experiment 2: Object foil similarity determines performance even within a set of cross-category objects

Experiment 1 reveals that within the stimulus set used by much prior work on visual working memory and visual long-term memory (Brady et al., 2008), randomly chosen pairs of objects, despite being putatively cross-category, are not maximally dissimilar. Does this have implications for memory performance? Or does any foil that is reasonably distinct from the target object give the same memory performance?

To assess this, in Experiment 2 we asked what the role of such similarity differences is for visual memory performance. In particular, we had participants remember real-world objects in visual working memory, and then at test, had them do a 2-AFC memory test. The study-test pairs in the 2-AFC were drawn from either the maximally similar, randomly paired, or maximally dissimilar stimulus pairs developed in Experiment 1 (see Fig. 2). If randomly chosen

foils are sufficiently dissimilar to provide a fair measure of the upper bound on memory performance there should be no difference between randomly chosen pairs and maximally dissimilar pairs in memory performance. By contrast, if memory performance is significantly impacted by similarity even for maximally dissimilar versus random pairs, then this would suggest that memory is underestimated for real-world objects when using randomly chosen foil pairs.

Methods

The study design, hypothesis, analysis plan, and exclusion criteria were preregistered at https://aspredicted.org/blind.php?x=c9nd9e. Materials, data, and analysis code are available at https://osf.io/axyqs/

Participants Fifty US-based participants (34 male, 16 female, 0 other/chose not to say) from Amazon's Mechanical Turk were included in the final data set (all with ≥ 95% previously accepted HITs). Eleven additional participants were excluded and replaced based on our preregistered exclusion criteria. We chose to pre-register a sample size of 50 participants because we expected a moderate effect of foil similarity and a sample size of $N = 50$ gives > 90% power to detect a standardized effect size of a Cohen's d of 0.5 in a t-test, which we took to be a reasonable expectation of a moderate effect we could expect for high versus low foil similarity.

Stimuli and procedure We used the Brady et al. (2008) object image database, as described above. In particular, in the random-pairs condition, we chose each pair of objects randomly from the full set of 2,400 objects without regard to similarity. For the maximally similar and dissimilar cross-category foils, we used the sets generated as described in Experiment 1: all the images were from the set of 2,400 cross-category objects, but chosen in such a way as to maximize or minimize similarity (Figs. 1 and 2).

Participants completed 180 trials overall, 60 in each of the three conditions (maximally distinct cross-category object foils; random cross-category object foils; maximally similar cross-category object foils). On each trial, participants were asked to remember six objects that were shown for 2,000 ms. We kept the placeholders that contained the objects continuously visible. They were also widely spaced. In previous work, this has resulted in almost no binding errors or swap errors (e.g., Brady & Alvarez, 2015; Schurgin et al., 2020), thus allowing a direct measure of capacity from a forced-choice task. Thus, the long encoding time, combined with fixed spatial positions with placeholders present during the delay, helped ensure there was little to no location noise that can cause misbinding. After a delay period of 700 ms, one of the items was probed in a 2-AFC format. One

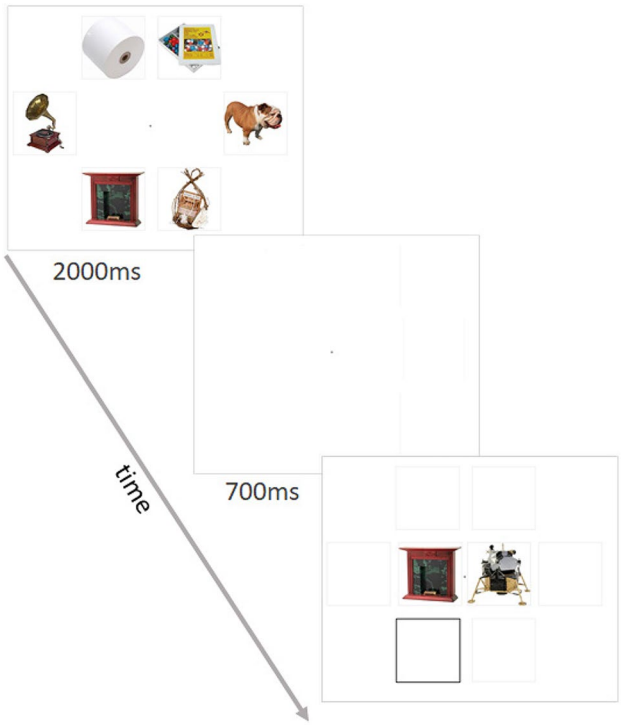

Fig. 4 Method of Experiment 2. Participants saw six objects, and remembered them over a short delay. They were then presented a two-alternative forced-choice (2-AFC) probe and needed to choose which item was presented at the cued location. The foil items in the 2-AFC memory test could be either maximally distinct cross-category foils, minimally distinct cross-category foils, or randomly chosen from a different category. Placeholders were visible throughout each trial

location was cued and two stimuli – one that was previously seen on the memory display and one foil – were shown in the center of the screen and participants had to indicate which of these two items (left vs. right) was part of the initial memory display (Fig. 4). All trial types were randomly interleaved.

Verbal overshadowing was performed while encoding the objects. In other studies we have shown that in-lab verbal interference (that is monitored continuously by an experimenter) and mental rehearsal of a single word during online studies results in similar visual working memory performance (Brady & Störmer, 2022). Thus, participants were instructed to mentally rehearse the word "the" for the entire duration of the encoding period. They were reminded of this on every trial, as well as the need to use purely visual memory and not use words to remember the stimuli.

Analysis Working memory performance was quantified using d′ for a 2-AFC task, $[zH - zFA]/\sqrt{2}$; where P is percent correct and Φ is the Gaussian cumulative distribution, $zH = \Phi(P)$ and $zFA = \Phi(1\text{-}P)$. Per the preregistration, data were excluded if d' averaged across all conditions was below

0, or if greater than 10% of individual trials were excluded. Individual trials were excluded if: (1) A response occurred less than 150 ms after the response screen appeared; (2) the response occurred more than 5 s after the response screen appeared.

Results

Participants overall showed higher memory performance for more distinct foils, even though all foils were cross-category foils (Fig. 5). An analysis of variance (ANOVA) with foil type (maximally similar, random, maximally distinct) as factors confirmed there was a main effect ($F(1,49) = 22.18$, $p < 0.0001$). Planned follow-up pairwise comparisons showed that maximally similar cross-category object pairs resulted in lower performance than random cross-category pairs ($t(49) = 4.79$, $p < 0.001$, $d_z = 0.68$) and random cross-category pairs resulted in lower performance than maximally dissimilar cross-category pairs ($t(49) = 2.26$, $p = 0.028$, $d_z = 0.32$).

Note that the difference in human similarity ratings for maximally dissimilar objects and randomly chosen objects, was rather small (see Fig. 3), though reliable; we noted this is in part because many participants gave all object pairs that were clearly of a different category the lowest similarity level. The memory data from this experiment – where the gain was considerable for maximally dissimilar foils relative to randomly chosen foils – suggests that this overall small difference in similarity rating likely obscures differences in their conceptual and perceptual relatedness that do, indeed, matter quite a bit for memory performance. This is perhaps what has driven some past researchers (including Brady et al., 2008, who created this stimulus set) to treat cross-category foils as "dissimilar enough" to judge memory capacity. Yet the current Experiment suggests that these differences matter substantially for memory performance. Thus, different levels of what – for human judgments – seem to be effectively floor levels of similarity may matter in important ways for human memory performance, making the convolutional neural network approach to stimulus generation particularly well-suited for this stimulus-choosing task.

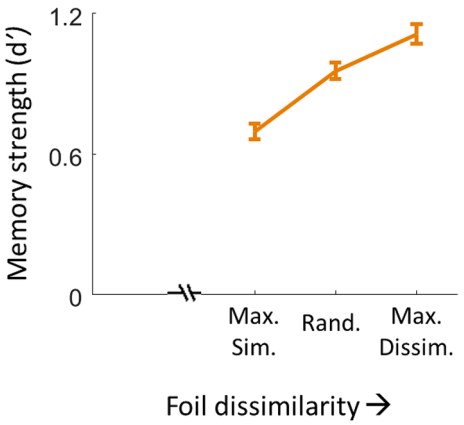

Fig. 5 Results of Experiment 2. We find that randomly chosen pairs are not in fact maximally dissimilar for the purposes of memory: participants perform better when the foils are maximally distinct than when they are randomly selected. They perform worst when foils are maximally similar within this across-category stimulus set. This means only using maximally dissimilar foils is a non-arbitrary way of measuring memory performance, providing the upper bound of memory strength

Experiment 3: Objects are better remembered than colors with matched foil similarities

We have shown that within the cross-category stimulus set of Brady et al. (2008) used in many previous working memory studies, there is room for significant variation in foil difficulty. Does this variation – and the use of random cross-category foils in some previous working memory studies rather than maximally dissimilar foils – affect why some studies did not find a benefit for real objects compared to colors in working memory? In Experiment 3, we examine that question directly by varying foil similarity not only for real-world objects but also colors, and compare memory performance across these different sets of stimuli and foil conditions. Specifically, we compare memory performance for randomly chosen foils and maximally dissimilar foils for both objects and colors in a working memory task.

Method

The study design, hypothesis, analysis plan, and exclusion criteria were preregistered at https://aspredicted.org/blind.php?x=s8bd3w. Materials, data, and analysis code are available at https://osf.io/axyqs/

Participants Fifty US-based participants (30 male, 19 female, one other/chose not to say) from Amazon's Mechanical Turk were included in the final data set (all with ≥ 95% previously accepted HITs). Nine additional participants were excluded and replaced based on our preregistered exclusion criteria. We chose to pre-register a sample size of 50 participants because we expected a moderate effect of foil similarity, and a sample size of $N = 50$ gives > 90% power to detect a standardized effect size of a Cohen's d of 0.5 in a t-test, which we took to be a reasonable expectation of a moderate effect we could expect for high versus low foil similarity.

Stimuli and procedure We contrasted randomly chosen foils and maximally dissimilar foils for both objects and colors in a standard long encoding time working memory task

modeled after Brady et al. (2016). For objects, we used the maximally dissimilar foils derived by VGG16 as described in Experiments 1 and 2, and for randomly chosen foils we chose each pair of objects randomly from the full set of 2,400 objects without regard to similarity, as described above and as done in recent papers (e.g., Li et al., 2020; Quirk et al., 2020). As in Experiment 2, one item from each pair was shown in a memory display, and the other served as the foil during the 2-AFC memory test. Which item served as the memory item versus foil in each pair was randomized across participants.

For colors, we used a standard color circle (Schurgin et al., 2020; Suchow et al., 2013) of radius 49 in the CIE *L*a*b* space (centered at L = 54, a = 21.5, b = 11.5). We created two analogous conditions to the object conditions: the maximally distinct color pairs condition, with foils 180° away from the target color (as is traditionally done); and a "random" foil condition, where the foil color could be any color that was > 30° from the target; thus, random foil colors were ranging from 30 to 180° away from the target, randomly chosen on each trial. We choose this limit to mimic the distinct object categories in the object set, as this limits foils to be different color categories (by carving out a wedge of 60° of the color wheel around the target). To minimize ensemble-based encoding (e.g., Brady & Alvarez, 2015), we chose all of the memory colors and the foil color subject to the constraint that no two colors could be less than 15° apart on the color wheel.

Participants completed 160 trials overall, 40 in each of the four conditions (maximally distinct object-foils; random object-foils; maximally distinct color foils; random color foils). On each trial, participants were asked to remember six stimuli – either colors or objects – that were shown for 2,000 ms. Just like in Experiment 2, placeholder remained visible during the delay. The long encoding time, combined with fixed spatial positions with placeholders present during the delay, helped ensure there was little to no location noise that can cause misbinding. After a delay period of 700 ms, one of the items was probed in a 2-AFC format. In particular, one location was cued and two stimuli – one that was previously seen on the memory display and one the pre-chosen foil – were shown in the center of the screen and participants had to indicate which of these two items (left vs. right) was part of the initial memory display (Fig. 6). All trial types were blocked.

Verbal overshadowing was again performed while encoding the objects, as participants were instructed to mentally rehearse the word "the" for the entire duration of the encoding period. They were reminded of this on every

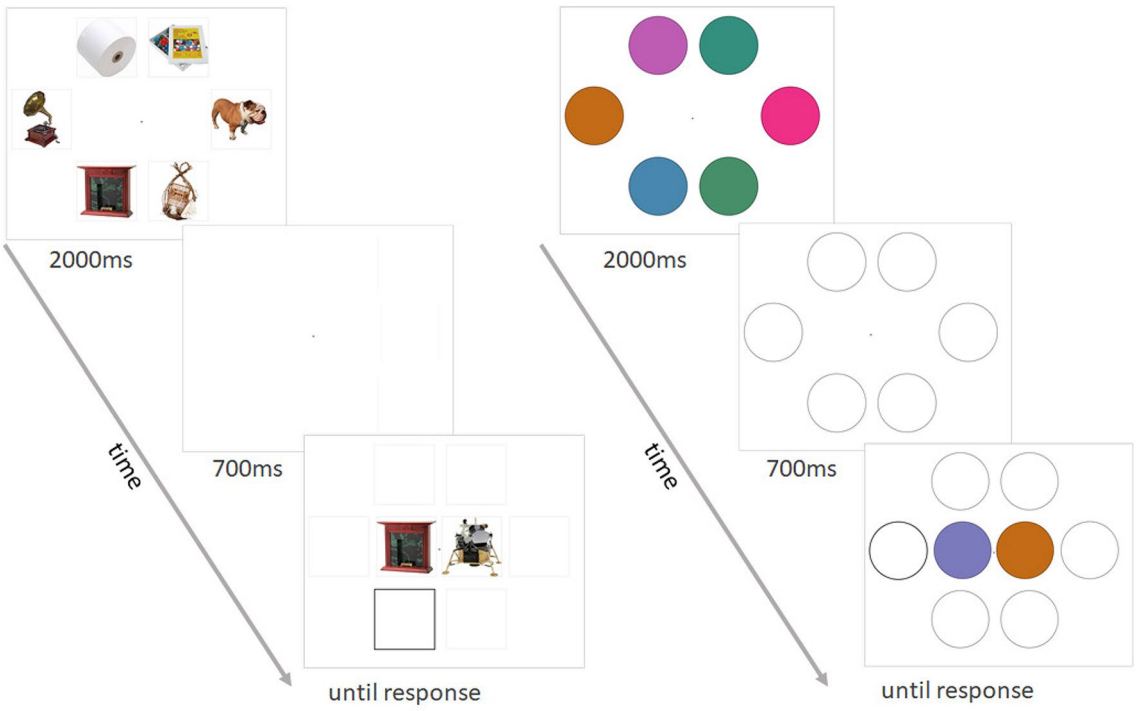

Fig. 6 Method of Experiment 3. Participants saw six items, either objects or colors, and remembered them over a short delay. They were then presented a two-alternative forced-choice (2-AFC) probe and needed to choose which item was presented at the cued location. The foil items in the 2-AFC memory test could be either maximally distinct or randomly chosen from a different category. Placeholders were present throughout each trial

trial, as well as the need to use purely visual memory and not use words to remember the stimuli.

Analysis Just like in Experiment 2, working memory performance was quantified using d′ for a 2-AFC task, $[zH - zFA]/\sqrt{2}$. Where P is percent correct and Φ is the Gaussian cumulative distribution, zH = Φ(P) and zFA = Φ(1-P). Per the preregistration, data were excluded if d' averaged across all conditions was below 0.5, or if greater than 10% of individual trials were excluded. Individual trials were excluded if: (1) A response occurred less than 150 ms after the response screen appeared; (2) the response occurred more than 5 s after the response screen appeared.

Results

Participants overall showed higher memory performance for real-world objects relative to colors, as well as higher memory performance for maximally dissimilar foils relative to randomly picked foils for both colors and objects (Fig. 7). An analysis of variance (ANOVA) with stimulus type (objects, colors) and foil type (distinct, random) as factors confirmed there were two main effects (objects > colors; $F(1,49) = 15.05$, $p = 0.0003$; distinct > random; $F(1,49) = 14.932$, $p = 0.0003$), and no interaction ($F(1,49) = 1.56$, $p = 0.22$). Planned follow-up pairwise comparisons showed that foil type affected performance for both objects ($t(49) = 4.02$, $p < 0.001$, $d_z = 0.57$) and to some extent for colors ($t(49) = 1.84$, $p = 0.071$, $d_z = 0.26$).

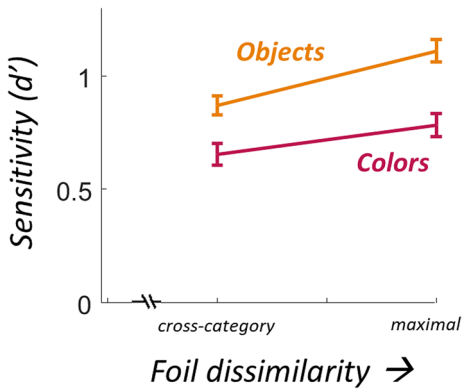

Fig. 7 Results of Experiment 3. We find that while cross-category changes (i.e., randomly selected foils) are indeed large changes – and thus result in well-above chance memory performance for both colors and objects – performance is best for both colors and real-world objects when foils are maximally distinct. Thus, to compare memory performance across stimulus sets in a fair way, the foils need to be chosen the same way. These data show that once test foils are matched in terms of their similarity to the target, there is a clear object benefit in working memory performance. Only when making the unfair comparison of cross-category object changes (chosen at random, left) vs. maximally dissimilar color changes (right) is there no benefit for objects. This means that for both objects and colors, only by using maximally dissimilar foils can we measure and compare memory performance in a non-arbitrary way

Overall, these results replicate the object-advantage previously reported by Brady et al. (2016) and replicated by Brady and Störmer (2022), and demonstrate the importance of choosing comparable test foils across stimulus sets when assessing memory performance. When comparing conditions in which color foils are chosen to be maximally distinct from the target color, but objects are tested against object foils chosen at random, we find no reliable performance difference between colors and objects ($t(49) = 1.06$, $p = 0.296$, $d_z = 0.15$). Thus, the comparison used in some previous work (e.g., Li et al., 2020; Quirk et al., 2020) – between randomly chosen objects and maximally distinct colors – may not be interpretable as a comparison of memory performance between colors and real-world objects.

General discussion

The present results demonstrate the importance of assessing memory capacity in a fair and consistent way across stimulus sets, and point to the significance of choosing appropriate foils in memory tasks more broadly. Previous work has quantified how foil similarity at test drives memory performance arbitrarily lower for simple feature spaces, like color, when more similar foils are chosen (e.g., Keshvari et al., 2013). Here we show the same is true for complex and realistic stimuli, such as pictures of real-world objects, and offer a solution for how to quantify similarity in real-world object spaces using CNNs. Indeed, while the similarity structure of the CNN matches human similarity ratings closely, if anything it appeared to show more sensitivity to similarity between otherwise quite dissimilar objects in a way that is important for memory performance, indicating that CNNs provide a particularly useful similarity measure in these circumstances.

We used this understanding of foil similarity to focus on a single case study, that when stimulus pairs are not matched in foil similarity across stimulus sets this can result in difficulty comparing across stimulus sets. Specifically, we find that when choosing foils in a way that is maximally dissimilar for colors but just cross-category for objects, the behavioral results of recent studies (Li et al., 2020; Quirk et al., 2020) that showed no visual working memory advantage for objects can be seen. However, when choosing foils in a way that is maximally dissimilar for both, objects are consistently better remembered than colors.

The necessity for using maximally dissimilar foils from a stimulus set to assess capacity has been repeatedly overlooked for real-world objects (e.g., by Brady et al., 2008). In the current work, we used a novel method to maximize foil dissimilarity, using deep convolutional neural networks. By contrast, in previous work, attempts to maximize dissimilarity among object pairs was done intuitively, solely by removing clearly similar target/foil pairs by hand (e.g., Brady et al., 2016). The current method provides a much more systematic alternative that is much more transparent and thus also highly replicable.

Even large, cross-category changes are not maximal

Our results reveal that even relatively large differences on the color wheel (e.g., categorically distinct colors, as used in Luck & Vogel, 1997, and many follow-up papers), and large cross-category object changes (e.g., as used in Brady et al., 2008), do not result in the highest possible performance level. In the case of objects, this suggests that Brady et al. (2008), for example, underestimate the "upper bound" of visual long-term memory performance. In the case of color, there are potentially broad implications as well. We here show that performance is better with maximally distinct color foils than with random cross-category colors, consistent with other recent data on this issue (e.g., Keshvari et al., 2013; Schurgin et al., 2020), which show that even foils > 70° from the target on the color wheel do not give rise to maximal performance (compared to foils 180° away). As noted by Keshvari et al. (2013), the continuous nature of this decrement in performance with foil distance is difficult to account for in item-limit models, but is consistent with models that propose continuous variation in precision (Keshvari et al., 2013), a continuous spreading of familiarity in the given feature space (Schurgin et al., 2020), or population-coding models that rely on shared neural representations to impose capacity constraints (Bays, 2015).

Convolutional neural networks as models of memory confusability

While only maximally dissimilar foils measure overall memory capacity non-arbitrarily for a given stimulus set, maximizing dissimilarity for objects is not straightforward. Here, we offer a novel, objective solution to how similarity of real-world objects can be taken into account. In particular, we use a deep convolutional neural network trained to categorize visual objects to maximize dissimilarity. In recent years, CNNs have achieved impressively human-like object categorization performance, and these models have been argued to resemble the human visual system, with early and late layers of these networks tracking the human early and later processing pathways in the visual system, respectively (Cichy et al., 2016; Eickenberg et al., 2017; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). Because such networks are trained to do categorization, they are sensitive to both visual and semantic features (e.g., Jozwik et al., 2017; Peterson et al., 2018).

Here, we used CNNs as a measure of object similarity, and showed that the CNN similarity ratings match the pattern of similarity ratings by humans (Exp. 1), and the CNN-chosen foils significantly affected memory performance (Exps. 2 and 3), showing that they provide a useful metric for judging memory confusability. If anything, the small divergences of the human similarity and CNN seemed to favor the CNN: While maximally dissimilar objects and randomly chosen objects were judged only as slightly different in similarity by human observers – both being largely at floor – the CNNs predicted relatively larger similarity differences between these sets of stimuli. Interestingly, in Experiments 2 and 3 we find that memory performance differs quite a bit between these stimuli – i.e., we find a large increase in memory performance for maximally dissimilar objects relative to randomly chosen objects. At least qualitatively, this suggests memory performance more closely tracked the pattern of CNN similarity, although we cannot quantitatively dissociate them here as human similarity and CNN similarity were very strongly related.

Overall, we believe CNNs show significant promise for being integrated into the study of memory for real-world objects (see also Needell & Bainbridge, 2022). Future research can examine the extent to which different CNN layers predict memory confusability (e.g., as a window in visual vs. semantic confusability), and how such metrics relate to different measures of human similarity.

Other differences that may be relevant for colors versus objects comparisons

In the current paper we demonstrate the importance of foil similarity when comparing memory performance across different stimulus sets, focusing in on the comparison of color versus objects in working memory. However, we note that foil similarity is far from the only factor important to consider when studying memory for real-world objects relative to colors.

In fact, while maximally dissimilar foils are a prerequisite to studying this issue, there are likely several other factors – for example, the way items are initially encoded – that play a critical role in eliciting benefits for real-world objects relative to colors (Brady & Störmer, 2022). Specifically, when participants are encouraged to attend each item individually using focused attention – processing each object in a deep way – memory performance for real-world objects increases, while color memory is impaired. By contrast, simple feature displays made out of only colors appear to benefit from

fast and parallel encoding – presumably because of the use of non-item-based strategies such as ensemble processing or grouping (Brady & Alvarez, 2015) – while real-world object memory is impaired in such circumstances (Brady & Störmer, 2022). Related to this, dissimilarity of the items at encoding is also an important factor, which can influence how multiple items on the initial display are encoded and remembered. For color memory, for example, it has been shown that similarity can lead to either repulsive or attractive biases in the memorized items, and that the strength of these biases vary with different encoding times (Chunharas & Brady, 2023; Chunharas et al., 2022). Encoding time also plays a role in eliciting the object benefit we observed in Experiment 3 and our previous work (Brady et al., 2016; Asp et al., 2021), as longer encoding time allows deeper processed of the items. However, in other work we have shown that while experiments using long encoding time result in a benefit of objects relative to colors, this is far from the most effective manipulation in promoting a deeper, item-based encoding strategy, and a more effective manipulation is to actually show objects sequentially at encoding (one at a time; Brady & Störmer, 2022). Overall, then, the evidence suggests that in addition to using proper foils, how the memory display is encoded initially is an important aspect of measuring memory capacity, and different strategies at encoding might lead to fundamentally different conclusions (see also, Chung et al., 2023a).

The role of meaning in visual working memory

When target/foil similarity was matched, we found a robust and clear object benefit, replicating other work showing higher memory capacity for objects than colors (e.g., Brady & Störmer, 2022; Thibeault et al., 2023; Torres et al., 2023). While comparing different stimulus sets in this way (e.g., color vs. objects) can provide insights into questions like the role of meaning or stimulus complexity in visual working memory, or the role of object complexity in memory performance, such comparisons are also difficult to interpret directly for conceptual reasons, in addition to the methodological issues that must be taken into account (like foil choice).

For example, although we find objects are better remembered than colored circles, colored circles and real-world objects differ in many ways: objects are visually more complex; they are familiar; and they connect to categorical and semantic knowledge. For example, the increased visual complexity of real-world objects compared to colors means that objects, but not colors, differ on a number of dimensions, such as color, shape, luminance, and orientation. Thus, the space of possible objects is far larger than that of simple features – which only differ in a singular dimension (e.g., color).

The current work does not seek to explain why objects are better remembered than colors. However, we believe the weight of the evidence suggests that it is the meaningfulness of objects that is critical, rather than these other factors (e.g., complexity). For example, previous studies have found that increased visual complexity – despite more complex objects having more dimensions they can differ on – actually result in lower working memory performance than simple features (e.g., Alvarez & Cavanagh, 2004), suggesting that visual complexity alone is not the underlying factor behind enhanced working memory for real-world objects.

The hypothesis that meaningfulness is critical is supported by other work that has focused on stimuli that are nearly perfectly matched except in the critical dimension of interest. Indeed, many studies have used such methods to show significant benefits to visual working memory from familiarity and meaning (e.g., Alvarez & Cavanagh, 2004; Asp et al., 2021; Brady et al., 2009; Curby et al., 2009; Jackson & Raymond, 2008; Ngiam et al., 2019; O'Donnell et al., 2018; Sahar et al., 2020; Starr et al., 2020). For example, Asp et al. (2021) showed a benefit of meaning on visual working memory by using ambiguous stimuli that could either be recognized as meaningful (i.e., a face) or not, thus matching visual input while varying the meaningfulness of the stimuli, and found enhanced working memory performance and increased neural delay activity for nearly visually identical recognizable versus non-recognizable stimuli. Results like these point to a clear role of meaningfulness in working memory and suggest that connections to knowledge, and not visual features – or the number of visual features – per se, improve working memory capacity. Other recent work has demonstrated that memory performance is improved for simple visual features, such as color, if these features are part of real-world objects compared to unrecognizable scrambled shapes (Chung et al., 2023a, b), suggesting an even broader role of meaningfulness in structuring visual working memory.

Thus, overall, we propose that the benefits for real-world objects relative to colors observed here is in large part due to real-world objects being conceptually meaningful, and meaningful stimuli recruiting additional working memory resources.

Conclusion

The capacity of visual working and visual long-term memory plays a critical role in theories of cognitive architecture and the relationship between memory and other cognitive systems. Here, we have shown that previous work in both visual long-term memory and visual working memory in particular has neglected to carefully consider one of the most important factors determining performance in a memory task, target/foil dissimilarity. Across three experiments, we

showed that only by maximizing the dissimilarity of the target and foil in each stimulus set can we provide a fair basis for memory comparisons between stimuli, and we introduced a new way to pick such foils objectively for complex, meaningful real-world objects by using deep convolutional neural networks. This work thus demonstrates not only that working memory capacity is not fixed capacity but depends critically on the type of information that is remembered, and also offers a solution of how to compare memory performance and other cognitive systems across different stimulus sets on common ground.

Acknowledgements Support by NSF (BCS-1829434) to TFB/VSS and NSF (BCS- 2141189) to TFB.

Data Availability Materials, data, and analysis code are available at https://osf.io/axyqs/.

References

Adam, K. C., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology, 97*, 79–97.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106–111.

Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater visual working memory capacity for visually matched stimuli when they are perceived as meaningful. *Journal of Cognitive Neuroscience, 33*(5), 902–918.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18*(7), 622–628.

Babic, Z., Schurgin, M. W., & Brady, T. F. (2019). Is short-term storage correlated with fluid intelligence? Strategy use explains the apparent relationship between""number of remembered item"" and fluid intelligence. *PsyArXiv*. https://doi.org/10.31234/osf.io/83ch4

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29.

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences, 19*(8), 431–438.

Bays, P., Schneegans, S., Ma, W. J., & Brady, T. (2022). Representation and computation in working memory. PsyArxiv preprint.

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 921.

Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 48*(7), 942–958. https://doi.org/10.1037/xlm0001014

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325–14329.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138*(4), 487.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences, 113*(27), 7459–7464.

Chung, Y. H., Brady, T., & Störmer, V. S. (2023a). *Sequential encoding aids working memory for meaningful objects' identities but not for their colors*. PsyArxiv preprint.

Chung, Y. H., Brady, T. F., & Störmer, V. S. (2023b). No fixed limit for storing simple visual features: Realistic objects provide an efficient scaffold for holding features in mind. *Psychological Science*, 09567976231171339.

Chunharas, C., & Brady, T. (2023). *Chunking, attraction, repulsion and ensemble effects are ubiquitous in visual working memory*. PsyArxiv preprint.

Chunharas, C., Rademaker, R. L., Brady, T. F., & Serences, J. T. (2022). An adaptive perspective on visual working memory distortions. *Journal of Experimental Psychology: General*.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 94.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage, 152*, 184–194.

Frank, D., Gray, O., & Montaldi, D. (2020). SOLID-Similar object and lure image database. *Behavior Research Methods, 52*(1), 151–161.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17*, 673–679.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). *ImageNet-trained CNNs are biased towards texture; Increasing shape bias improves accuracy and robustness*. arXiv preprint: 1811.12231.

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.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In Computer Vision–ECCV 2016. In: *14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14* (pp. 630–645). Springer International Publishing.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour, 4*(11), 1173–1185.

Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces. *Journal of Experimental Psychology: Human Perception and Performance, 34*(3), 556.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology, 8*, 1726.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream"s execution of core object recognition behavior. *Nature Neuroscience, 22*(6), 974–983.

Keshvari, S., Van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology, 9*(2), e1002927.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PloS Computational Biology, 10*(11), e1003915.

Kietzmann, T., McClure, P., & Kriegeskorte, N. (2019, January 25). Deep Neural Networks in Computational Neuroscience. Oxford Research Encyclopedia of Neuroscience. Retrieved 31 Oct. 2023, from https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. BioRxiv preprint. https://doi.org/10.1101/408385

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science, 10*(4), 477–493.

Li, X., Xiong, Z., Theeuwes, J., & Wang, B. (2020). Visual memory benefits from prolonged encoding time regardless of stimulus type. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(10), 1998.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience, 33*(10), 2017–2031.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17*(8), 391–400.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347.

Mate, J., & Baqués, J. (2009). Short article: Visual similarity at encoding and retrieval in an item recognition task. *Quarterly Journal of Experimental Psychology, 62*(7), 1277–1284.

Needell, C. D., & Bainbridge, W. A. (2022). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-022-00126-5

Ngiam, W. X., Khaw, K. L., Holcombe, A. O., & Goodbourn, P. T. (2019). Visual working memory for letters varies with familiarity but not complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(10), 1761.

O'Donnell, R. E., Clement, A., & Brockmole, J. R. (2018). Semantic and functional relationships among objects increase the capacity of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(7), 1151.

Palmeri, T. J., & Tarr, M. (2008). Visual object perception and long-term memory. In S.J. Luck & A. Hollingworth (Eds.) *Visual memory* (pp. 163–207). Oxford University Press.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42*(8), 2648–2669.

Quirk, C., Adam, K. C. S., & Vogel, E. K. (2020). No evidence for an object working memory capacity benefit with extended viewing time. *eNeuro, 7*(5). https://doi.org/10.1523/ENEURO.0150-20.2020

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sahar, T., Sidi, Y., & Makovski, T. (2020). A metacognitive perspective of visual working memory with rich complex objects. *Frontiers in Psychology, 11*, 179.

Schurgin, M. W., & Brady, T. F. (2019). When "capacity" changes with set size: Ensemble representations support the detection of across-category changes in visual working memory. *Journal of Vision, 19*(5), 3–3.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour, 4*, 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Shoval, R., & Makovski, T. (2022). Meaningful stimuli inflate the role of proactive interference in visual working memory. *Memory & Cognition, 50*(6), 1157–1168.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology, 25*(2), 207–222.

Starr, A., Srinivasan, M., & Bunge, S. A. (2020). Semantic knowledge influences visual working memory in adults and children. *PLoS ONE, 15*(11), e0241110.

Storrs, K. S., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience, 33*(10), 2044–2064. https://doi.org/10.1162/jocn_a_01755

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision, 13*(10), 9–9.

Thibeault, A., Stojanoski, B., & Emrich, S. M. (2023). Investigating the effects of perceptual complexity versus conceptual meaning on the object benefit in visual working memory. *PsyArxiv*. https://doi.org/10.31234/osf.io/3dmrq

Torres, R. E., Duprey, M., Campbell, K. L., & Emrich, S. M. (2023). Not all objects are created equal: the object benefit in visual working memory is supported by greater recollection, but only for some objects. https://doi.org/10.31234/osf.io/v2ta5

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 11–11.

Xu, Y., & Vaziri-Pashkam, M. (2020). Limited correspondence in visual representation between the human brain and convolutional neural networks. *BioRxiv*. https://doi.org/10.1101/2020.03.12.989376

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

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.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235.