# Representational magnitude as a geometric signature of image and word memorability

David Amadeus Vogelsang[a,b], Micha Heilbron[a,b]

[a]*Faculty of Social and Behavioural Sciences, University of Amsterdam, the Netherlands*
[b]*Programme group Brain and Cognition, University of Amsterdam*

**Abstract**

What makes some stimuli more memorable than others? While memory varies across individuals, research shows that some items are intrinsically more memorable, a property quantifiable as "memorability". Recently, a new representational signature for image memorability was identified: the magnitude of the population response in convolutional neural networks (CNNs) correlated with image memorability. However, it is unclear if this geometric principle was confined to the visual domain or whether it represents a broader computational phenomenon observable in other stimuli domains as well. Here we show that this representational magnitude effect not only replicates for images in an independent dataset but also generalizes to an entirely different cognitive domain and neural network architecture: lexical memorability and word embeddings. Across three large-scale lexical datasets, we found that the L2 norm (vector magnitude) of word embeddings reliably predicted recognition memorability, independent of word frequency, valence, or word length. This consistency suggests the effect reflects a general property of distributed representations, where representational magnitude may capture how strongly a stimulus projects onto dominant, conceptual/semantically meaningful features in a network's embedding space. At the same time, this effect does not appear to generalize to all domains, as our analysis of a novel voice memorability dataset showed no such relationship between representational magnitude and memorability. Together, these findings indicate that the geometry of distributed representations offers a useful lens for understanding memorability, suggesting an item's representational magnitude reflects its projection onto dominant dimensions of representation, in both artificial and biological systems.

*Corresponding author: d.vogelsang@uva.nl

## 1. Introduction

Why do we remember some stimuli better than others? Research on memory often takes a subject-specific approach, focusing on the cognitive processes behind memory and individual differences. This perspective is based on the idea that memory is highly personal and influenced by cognitive factors such as attention (Wakeland-Hart et al. (2022)). However, while factors like experimental conditions and individual differences influence memory performance, research over the past decade has found that certain stimuli are inherently more memorable than others across participants (Bainbridge et al. (2013, 2017, 2019); Isola et al. (2013); Wakeland-Hart et al. (2022); Aka et al. (2023); Kramer et al. (2023); Revsine et al. (2025)). By analyzing aggregated task scores for each stimulus instead of individual participant responses, memorability of a stimulus can be quantified, with a high level of agreement between participants across stimulus types and experimental contexts. This memorability effect has been observed across images (Isola et al. (2013); Kramer et al. (2023)), words (Aka et al. (2023); Madan (2021)), voices (Revsine et al. (2025) and even in art work (Davis and Bainbridge (2023)) and dance moves (Ongchoco et al. (2023)). To further support the idea that memorability is an intrinsic property of stimuli, a study by Needell and Bainbridge (2022) showed that a pretrained deep neural network, named ResMem, could predict memorability scores solely based on the perceptual features of an image. Furthermore, some studies have found that images that are more typical (e.g. how typical is this image of a banana?) are more memorable (Bainbridge and Rissman (2018); Bainbridge et al. (2019)), although other studies have found the opposite (Lukavskỳ and Děchtěrenko (2017)) or found a more complex relationship between typicality and memorability (Kramer et al. (2023)). Remarkably, memorability not only varies consistently across humans but also in rhesus monkeys (Jaegle et al. (2019)), suggesting memorability is independent of extensive object experience (e.g., vehicles, buildings etc.) and language ability.

Although the exact relationship between image content and memorability is still a matter of scientific scrutiny, the availability of image memorability scores has enabled research into the brain's representational correlates of memorability and the development of deep neural network (DNN) models that quantify and predict image memorability. Recently, a series of studies reported a dual coding scheme that predicts how well images will be discriminated and remembered (Jaegle et al. (2019); see Rust and Mehrpour (2020) and Rust and Jannuzi (2022) for reviews). Specifically, Jaegle and colleagues (2019) observed that whereas the angle of the response vector between two images quantifies how well they can be discriminated, the length of an image's response vector ("L2 norm" or response magnitude), significantly predicted its inherent memorability, as quantified across a large image dataset (LaMem) and their memorability scores (Khosla et al. (2015)). Strikingly, this dual coding scheme was present in both activation patterns in rhesus macaque monkey inferotemporal cortex (IT), and in deep convolutional neural networks (CNNs) trained for object recognition (Jaegle et al. (2019)). Thus, it appears that brain-like representations of stim-

ulus memorability emerge in artificial neural networks that are not explicitly trained to predict memorability.

However, these striking studies leave some questions unaddressed. First, it remains unclear whether the findings reported by Jaegle et al. (2019) can be replicated in other large datasets of image memorability. Second, and more importantly, it is unclear whether this coding scheme is inherent to visual perception and image memorability, or whether it reflects a more general principle, applicable to stimulus memorability (and stimulus encoding) more broadly. One possibility is that it is specific to visual memorability and visual representations because it derives from the representational convergence between the ventral visual stream and deep CNNs (DiCarlo and Cox (2007); Yamins and DiCarlo (2016)) – a neural network architecture that is, after all, explicitly modelled after the visual cortex (Lindsay (2021)). Alternatively, the emergent coding scheme (where L2 norm or representational magnitude predicts memorability) may be a more general principle of stimulus encoding in distributed representations, that would be applicable to all distributed representations, irrespective of stimulus modality or 'brain-likeness' of the network architecture used.

Here, we address both questions by first attempting to replicate the basic effects reported by Jaegle et al. (2019) to a different and independent large dataset of image memorability, spanning 26,107 naturalistic object images across 1,854 diverse categories tested in over 13,000 participants (THINGS; Hebart et al. (2020)). Secondly, we want to examine whether the same principle applies to different memorability domains, namely, word memorability. To test this, we analysed 3 datasets of word recognition memorability, covering a total of over 8500 memorability scores across 800 participants. To quantify representational magnitude, we computed the L2 norm of word embeddings, vector representations that captured a lot of information about words but are derived from neural networks that are not in any way inspired by or modelled after the human brain. Interestingly, we find that first, the L2 norm effect as reported by Jaegle et al. (2019) is robust and consistently replicates. Moreover, we find the exact same 'representational magnitude effect' for word recognition memorability, indicating that the effect is not specific to visual perception or sensory cortex, but reflects a more general property of stimulus encoding in distributed representations, perhaps related to the alignment between individual stimuli and dominant encoding 'axes' in both biological brains and neural networks. Finally, we evaluate the same effect in a recent voice memorability dataset covering a total of over 600 memorability scores across 2700 participants (Revsine et al. (2025)). However, this representational magnitude effect (using the L2 norm of waveform embeddings of wav2vec) was not found in the voice memorability dataset, which could relate to particularities of this dataset or reflect limitations in the generality of the phenomenon.

## 2. Methods

**Datasets** We analysed the effect of representational magnitude on recognition memory in six large-scale datasets – one image, three word memorability

datasets, and two voice memorability datasets – encompassing in total over 35,000 memorability ratings across visual, word and voice stimuli. We outline each in turn.

*Kramer et al. (2023; THINGS):* Memorability scores were obtained from the THINGS dataset as reported in (Kramer et al. (2023)). Details on the study design can be found in Kramer et al. (2023) but we provide a brief summary here. A total of 13,946 participants from Amazon Mechanical Turk (AMT) completed a continuous recognition repetition detection task using 26,107 images from the THINGS database (Hebart et al. (2020); see https://things-initiative.org/). The THINGS database includes 1,854 object concepts spanning 27 categories, each with at least 12 exemplar images. Memorability scores were collected for all 26,107 images in the THINGS database by having participants view a stream of images and press a key when they recognized a repeated image. Each participant saw a unique pseudorandom subset of 187 images, with approximately 40 responses gathered per image. Memorability was calculated as the average hit rate (HR) across participants for each image, adjusted for the false alarm rate (FA), which accounts for the tendency to incorrectly recognize novel images as familiar (see figure 1 for illustration). This correction is typically applied per image to reflect variations in how images may evoke a sense of familiarity, even when they have not been seen before (Rust and Mehrpour (2020)). The raw data can be obtained here: `https://osf.io/5a7z6/`.

*Aka et al. (2023):* For the Aka dataset, we used the word recognition and recall memorability scores which were calculated on datasets from the PEERS database. A detailed description of the PEERS database has been described by Kahana et al. (2024) and further details of the memorability scores has been reported by Aka et al. (2023). In short, the Penn Electrophysiology of Encoding and Retrieval Study dataset (PEERS; available at `http://memory.psych.upenn.edu/Data_Archive`) consists of recognition and recall memory data. The recognition memory dataset includes 171 participants (ages 18–30) who contributed 3120 sessions. Each session involved studying 12–16 lists of 16 words (from a pool of 1638 words) and completing a recognition test with 320 probe words. The recall memory dataset includes 98 participants (ages 18–30) who contributed 2254 sessions. Each session involved studying 24 lists of 24 words (from a pool of 576 nouns), followed by a recall test after a brief arithmetic distraction task. Recognition and recall probabilities were estimated as the percentage of participants who correctly recognized/recalled each word (see figure 1 for illustration). The Aka et al. (2023) dataset consists of 576 words with recognition and recall memorability scores. The Aka et al. (2023) data is available here: `https://osf.io/7phj6/files/osfstorage`.

*Cox et al. (2018):* We obtained recognition memorability scores from the Cox et al. (2018) dataset. Participants performed various memory tests (e.g. associative memory and cued recall test) of which only the recognition memory and the free recall were used in this study. During the study phase of each

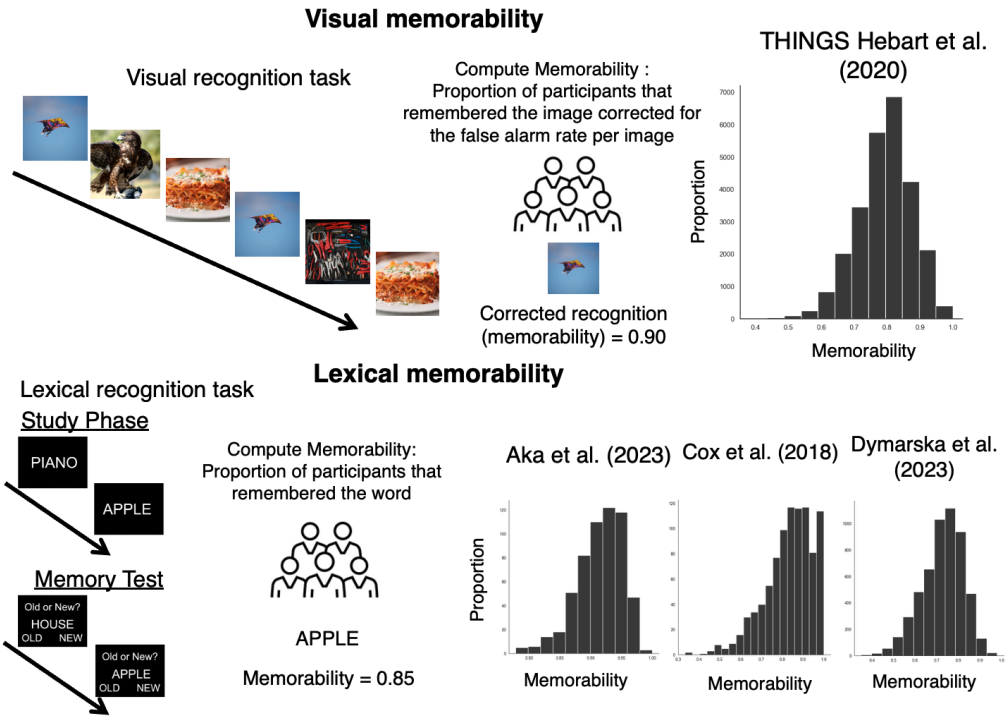



Figure 1: Overview of visual and word memorability paradigm. Images/words are studied and in a subsequent recognition memory test participants indicate whether a stimulus is old or new. For images, memorability is computed by calculating the proportion of participants that remembered a certain image corrected for the false alarm rate per image. For words, memorability is computed as the proportion of participants that correctly recognised a word in an old/new recognition memory test. Memorability score distributions are provided for the visual and word datasets in the graphs on the right.

task, 453 participants viewed 20 word pairs, each presented for two seconds. In the single-item recognition memory task, they identified which of 20 words had appeared on the studied list, with 10 words serving as new 'foils'. In the free recall task, participants attempted to recall as many words as possible from the study list. Cox et al. (2018) did not compute memorability scores themselves hence we computed recognition and recall memorability scores for each studied word by taking again the proportion of participants who correctly recognized/recalled each word. In total, 924 words with their corresponding recognition and recall memorability scores were computed in this dataset. The raw Cox et al. (2018) data files can be obtained here: `https://osf.io/dd8kp/`.

*Madan (2021):* Madan (2021) only reported the recall memorability scores that were obtained from the Penn Electrophysiology of Encoding and Retrieval Study (PEERS). In each session, participants studied 12–16 lists of 16 words, with each word displayed for 3 seconds followed by a short interval. After the final word, a brief delay preceded a free recall test, where participants had 75 seconds to vocally recall as many words as possible. Recall data were examined from 147 young adults (ages 16–30), each completing 20 sessions in the PEERS experiments. The word pool consisted of 1,638 words from the University of South Florida free association norms database, selected for their semantic relatedness and suitability for size and animacy judgments. In total, 1638 words and their corresponding recall memorability scores were used in the Madan dataset. The raw datafiles can be obtained here: `https://osf.io/spqjz/`

*Dymarska et al. (2023):* We used the data as reported by Dymarska et al. (2023) who analysed data from two large recognition memory studies for words (over 5300 words) as reported by Cortese and colleagues (Cortese et al. (2010, 2015)). We provide a brief summary of the Cortese studies here: Across two large studies, 207 participants (87 in Cortese et al. (2010); and 120 in Cortese et al. (2015)) were presented with 5301 words that were used a recognition memory test. In both studies, participants studied lists of 50 words at a time in preparation for a later recognition task. This task required them to make an old/new judgment on each presented word, with half of the target words being old and the other half new. Hit rate, false alarm rate and corrected hit rates per stimulus (i.e. word) were computed, of which for the purpose of this current study we only used the hit rate (i.e. memorability score) per word to make it comparable to the other word datasets used in the current study. The raw data used in the Dymarska study can be downloaded here: `https://osf.io/r8fmb/`.

*Revsine et al (2025):* We used the voice memorability data collected by Revsine and colleagues (2025), which comprised of three voice memorability experiments. In their study, 3,340 Amazon Mechanical Turk participants completed a 10-minute continuous recognition memory task, where they listened to the same sentence ("She had your dark suit in greasy wash water all year" in experiment 1, and "Don't ask me to carry an oily rag like that" in experiment 2) being spoken by different speakers and pressed a key when they heard a

repeated speaker. Stimuli were drawn from the TIMIT speech corpus. Memorability (measured with d') was collected for each voice clip in Experiments 1 and 2. All raw data can be downloaded here: `https://osf.io/pybwd/`.

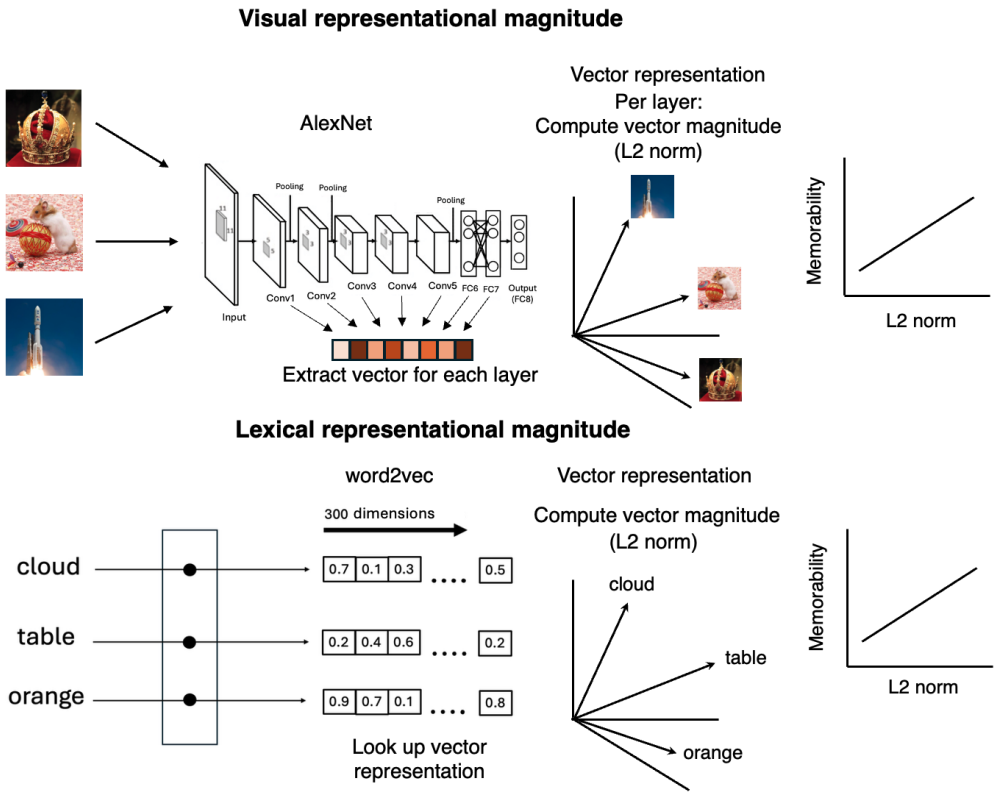



Figure 2: For each image in the THINGS dataset, we extracted the features (i.e. vector) representation of each layer of AlexNet and computed the vector magnitude (i.e. L2 norm) of each layer. Finally, we correlated the per image memorability score with the vector magnitude. For words, we extracted for each word the vector representation of the 300-word embeddings from word2vec. We again computed the L2 norm of each vector representation and correlated this with the recognition memorability score per word.

**Data Analysis** An overview of our analysis approach is provided in figure 2. We computed the correlation (i.e. Spearman correlation) between the response magnitude (after the ReLu activation function) at all layers of a convolutional neural network (i.e. AlexNet) and memorability scores for images in the THINGS dataset. We then investigated whether we could extend this analysis to the domain of word memorability by examining the correlation between the response magnitude of a simple neural network (i.e. Word2Vec) and recognition and recall memorability scores for words. For the images in the THINGS dataset, we used the memorability scores as reported by Kramer et al. (2023). For each image in the THINGS dataset, we extracted the features of each layer

in AlexNet. AlexNet is a deep convolutional neural network (CNN) designed for image recognition, and it contains eight layers of which five are convolutional and three are fully connected (Krizhevsky et al. (2017)). The response magnitude (L2 norm) for each layer was calculated using the entire output vector of the hidden layer. For every layer, we calculated Spearman correlation coefficients for all rectified linear unit (ReLu) layers. Correlations between the L2norm of each layer and memorability scores were computed across the entire set of 26,107 images from the THINGS dataset.

For the word memorability analysis, we utilized the pretrained Word2Vec model 'GoogleNews-vectors-negative300' (Mikolov et al. (2013)). Word2Vec is a neural network-based model that learns vector representations of words by analysing their usage in large text corpora. It captures semantic relationships by positioning words with similar meanings closer in a high-dimensional space (Mikolov et al., 2013). The model uses the Skip-gram architecture, which efficiently learns high-quality word vectors that capture a wide range of precise syntactic and semantic relationships and is therefore good at predicting surrounding words given a target word (Mikolov et al. (2013)). This model was trained by Google on approximately 100 billion words from the Google News dataset using the skip-gram architecture (Mikolov et al. (2013)). Here, we extracted for each word in each dataset (i.e. Aka, Cox and Madan) the 300-word embeddings and computed the L2 norm. As a final step, we computed the Spearman correlation between the L2 norm of each word and their corresponding recall/recognition memorability scores.

For the voice memorability analysis, we used wav2vec, which is a self-supervised deep learning model developed for speech representation learning from raw audio waveforms (Baevski et al. (2020)). Unlike traditional approaches that rely on handcrafted acoustic features, wav2vec learns feature representations directly from unlabeled audio data by predicting masked segments of the input waveform based on surrounding context. Its architecture consists of a multi-layer convolutional feature encoder that processes the raw waveform, Transformer-based context network that captures long-range temporal dependencies. This approach enables the model to capture phonetic and prosodic information that can be fine-tuned for various downstream tasks, including speech recognition. Its general purpose representations make the network well-suited to test the representational magnitude effect in the Revsine dataset. For each audiofile in the Revsine dataset (who used the TIMIT database), we extracted the features from all wav2vec layers and computed the L2 norm. We computed the d'prime score, which Revsine used as their measure of memorability, for each audiofile and conducted a Spearman correlation between the L2 norm and memorability scores.

To estimate the reliability of the observed correlation between representational magnitude and memorability scores, we employed a non-parametric bootstrapping procedure. Specifically, we computed the Spearman correlation coefficient between the two variables and then generated a 95% confidence interval via resampling. This was done by repeatedly sampling the dataset with replacement 10,000 times, preserving the pairing between each data point (i.e.,

paired resampling of the two variables). For each resampled dataset, we recalculated the correlation coefficient, resulting in a distribution of 10,000 correlation values. The 2.5th and 97.5th percentiles of this distribution were taken as the lower and upper bounds of the confidence interval. This approach allowed us to quantify the uncertainty in the estimated correlation without relying on assumptions of normality or homoscedasticity.

**Control Analysis** In our main analysis, we found a significant correlation between vector magnitude (L2 norm) and memorability. However, one potential alternative interpretation is that the observed relationship between L2 norm and memorability may be driven by confounding factors, such as image typicality or word frequency, valence or size. If higher L2norm values merely reflect more typical images or more frequently used words, it could be that the effect attributed to representational magnitude could instead be explained by image typicality/word frequency. To ensure that the positive L2norm effect on memorability was not driven by confounding factors such as typicality of images or word frequency, we conducted a follow-up control analysis. Specifically, for images in the THINGS dataset, Kramer et al. (2023) computed typicality based on both the 49 objects dimensions as obtained and described in Hebart et al. (2020) as well as from a deep convolutional neural network (DNN). In the 49 dimension object feature space, items that are closely clustered within a concept (i.e. within the concept of airplane, an image of an airplane is more similar to other images of airplanes) are the most prototypical, whereas those that are more widely spaced are the most atypical (e.g. an image of an airplane that looks very different from the other images of airplanes). To compute typicality from the DNN, Kramer et al. (2023) calculated the average similarity (Pearson correlation) between each image and all other images representing the same concept (e.g., the similarity of a given strawberry image to all other strawberry images in the THINGS dataset). This procedure was repeated across all layers of the DNN, yielding a typicality score for each layer. For the purpose of our regression analysis, we take those typicality values of the last layer of the DNN. The typicality values of the object space and DNN typicality were provided by the authors and can be obtained through the following link: `https://osf.io/5a7z6/files/osfstorage`. To examine possible confounding variables on the L2 norm effect on word memorability, we computed a Spearman correlation matrix of the most prominent word features that were provided in the Aka et al. (2023) dataset (see figure 7A). Crucially, based on this correlation matrix we wanted to examine which word features showed a correlation in the same direction as the L2 norm effect. As can be seen in the correlation matrix, three word features showed a significant correlation in the same direction as the L2 norm of word2vec, namely, word frequency, valence and size. These three features could therefore be a confounding variable explaining the positive association between the L2 norm and recognition memorability. To investigate the effect of frequency on the L2norm effect, we obtained frequency ratings per word using the SUBTLEX, which calculated word frequency based on television and film subtitles (Brysbaert and New (2009)). Valence and size ratings were

obtained from the Aka et al. (2023) dataset.

To ensure that the L2norm effect on image and word memorability was not confounded by image typicality or word frequency, valence or size, we conducted a two-step regression analysis. First, we regressed image L2norm on image typicality and word L2norm on word frequency, valence and size using ordinary least squares (OLS) regression. We then extracted the residuals, representing the component of vector magnitude independent of typicality (for images) and frequency, valence and size (for words). Finally, we tested whether these residualised L2 norm values still predicted memorability scores.

## 3. Results

In the primary analysis, we assess the effect of representational magnitude on recognition memory – as reported by Jaegle et al. (2019) – in 4 independent datasets, one of image memorability, three of word recognition memorability (see Figure 1). We first asked whether we could replicate the effect reported by Jaegle et al. (2019) of visual representational magnitude (based on CNN embeddings) on visual recognition memory in a different dataset. For this we turn to the THINGS-memorability dataset (Kramer et al. (2023)), which is a dataset of 26,107 images (see Methods). We computed the L2norm of each layer in AlexNet and correlated (Spearman correlation) it with the memorability scores of each of the 26107 images in the THINGS dataset. We ran this analysis for all 7 layers of the network. The results are displayed in Figure 3. As can be seen, the correlation between the L2norm and memorability is not significantly different from zero in early layers of the network but then increases significantly in later layers of the network (Layer 1: r = -.018, p = 0.004, 95% CI [-.03, -.006]; layer 2: r = .003, p = .62, 95% CI [-.009, .015]; layer 3: r = .007, p = .29, 95% CI [-.006, .019]; layer 4: r = .041, p = 2.5 e-11, 95% CI [.029, .053]; layer 5: r = .038, p = 7.9e-10, 95% CI [.026, .050]; layer 6: r = 0.05, p = 2.6e-16, 95% CI [.038, .061]; layer 7: r = .057, p = 2.7e-20, 95% CI [.045, 069] (see figure 3B)). This pattern of no/low correlation between the L2norm and image memorability in the early layers of a convolutional neural network and a higher and significant correlation in later layers of the network is a replication of the findings reported by Jaegle et al. (2019). However, the highest correlations we observe in the later CNN layers (r=0.057) are lower in magnitude than those reported by Jaegle (r=.30), which may be attributed to our dataset being derived from a large online crowd-sourced experiment that is inherently more noisy.

After confirming that the effect reported by Jaegle and colleagues (2019) indeed replicates, we next asked whether this effect is specific to visual recognition memory or whether it applies to other domains, such as word and voice memorability, as well. For this, we first turned to the non-sensory domain of verbal recognition memory, for which we analysed three datasets, encompassing 8500 memorability scores across over 800 participants in total. In these datasets, participants performed a very similar recognition memory task, but

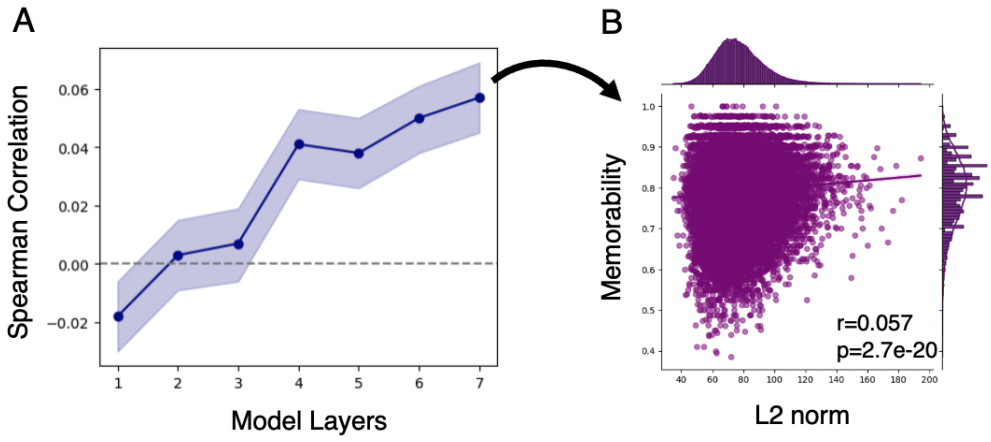



Figure 3: A) Spearman correlations for each layer of AlexNet (5 convolutional, 2 fully connected) between the magnitude of the representation at that layer (L2 norm), and the memorability score of each image in the THINGS dataset. B) Spearman correlation plotted for the last ReLU layer.

memorised words instead of objects (see Figure 1 and Methods). We computed the L2 norm of the static word embeddings from word2vec (see methods for details) for all words in every dataset and computed a Spearman correlation with the corresponding recognition memorability scores. As visible in Figure 3, this revealed a highly consistent and strong correlation for the Aka et al. (r = .32, p = 6.5e-15, 95% CI [.24, .39]), the Cox et al. (r = .22, p = 2.72e-11, 95% CI [.16, .28]) and Dymarska et al. datasets (r=0.47, p=1.5e-293, 95% CI [.45, .49]) between the embeddings representational magnitude and its recognition memorability. This result indicates that the L2norm effect as observed by Jaegle et al. (2019) may reflect a more general principle of stimulus encoding in distributed representations rather than a specific attribute of visual CNNs or visual cortex.

**Control Analysis** Given that embeddings are not inherently interpretable but encode a lot of stimulus information implicitly, it could be that the vector magnitude simply correlates with another stimulus attribute that is already known to co-vary with memorability. To rule this out, we consider several likely alternative explanations. For images, we considered the role of typicality. First, more "typical" object instances (e.g. in the visual case, typical cats) might be more similar to the instances seen during training (ImageNet), and hence more aligned to the feature detectors in AlexNet, leading to higher vector magnitudes. At the same time, typical images have also been found to be more memorable, although this relationship is complex and may depend on category-level expectations, distinctiveness and familiarity (Kramer et al. (2023)). To address this potential confound, we regressed out two different model-based

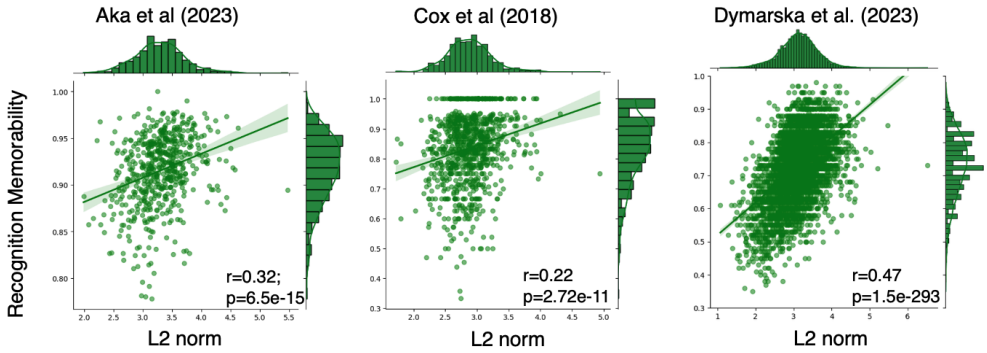
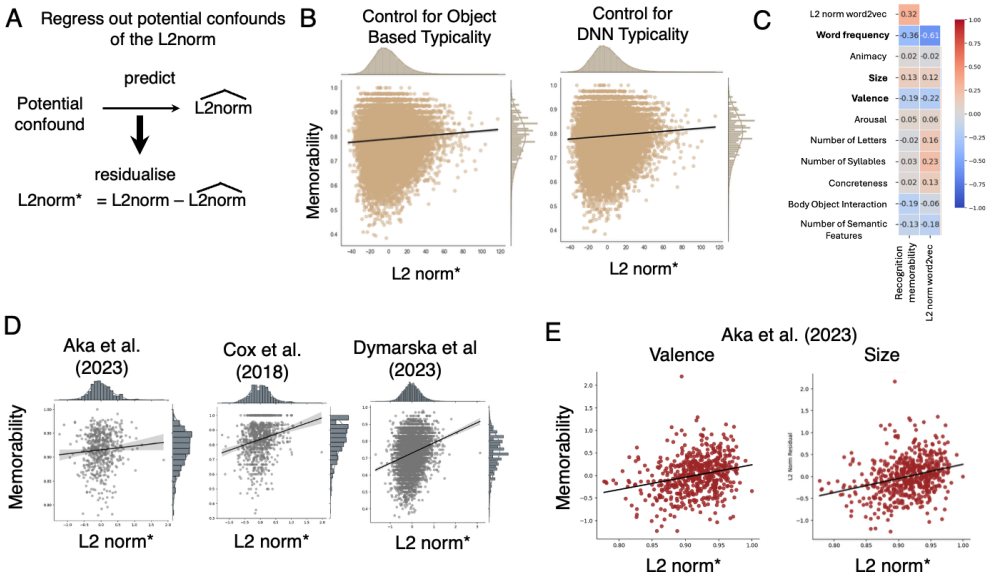



Figure 4: Spearman correlations between the L2norm and recognition memorability of each word across the three datasets.

image-typicality estimates (see methods for details) of the vector magnitude of the last ReLu layer of AlexNet and ran a regression with the residualised L2norm as predictor and image memorability as the outcome variable. The results are displayed in figure 5B. For both the object typicality ($\beta$ = .0003, t(26106) = 9.97, $p$ < .001, 95% CI [0, 0]) and DNN typicality ($\beta$ = .0003, t(26106) = 9.68, $p$ < .001, 95% CI [0, 0]) regressed out, the L2 norm still had a positive effect on memorability scores. This indicates that the effect of representational magnitude cannot be reduced to image typicality.

For words, we first identified potential confounding features by examining which word features reveal a similar correlation pattern with recognition memorability as the L2 norm (see figure 5C). In short, a potential confounding feature has to either correlate positively L2-norm and/or memorability, or correlate negatively with both. As can be seen, word frequency, valence and size show correlations in the same direction as the L2 norm thereby revealing the possible confound that the L2 norm effect reflect a word frequency, valence or size effect in disguise. To address this potential confound, we conducted a control analysis in which we examined whether the L2 norm effect is still observed when accounting for word frequency, valence and size in lexical stimuli.

For word recognition memorability, we regressed out word frequency of the vector magnitude and used the residualised L2norm as the predictor with recognition memorability as the outcome variable. For the Cox ($\beta$ = .074, t(922) = 6.75, $p$ < .001, 95% CI [.052, .095]) and Dymarska ($\beta$ = 0.06, t(5358) = 17.68, $p$ < .001, 95% CI [.053, .067]) datasets, the residualised L2norm had a significant positive effect on memorability scores and for the Aka dataset the residualised L2norm had a marginally significant positive effect on memorability scores ($\beta$ = .008, t(573) = 1.75, p = .08, 95% CI [-.001, .017], see figure 5D). Second, word-level variables such as valence and size were only available for all items in the Aka dataset, as the Cox and Dymarska datasets included a broader range of




Correcting for potential confounds.

Figure 5: A) Regressing out image typicality (both object and CNN based) of the L2 norm B) Regressing out object and CNN based typicality, we still find a positive relationship between the L2 norm and image memorability. C) Correlation matrix of word features from the Aka dataset reveals that word frequency, valence and size show a similar significant correlation patterns as the word2vec L2 norm. D)Regressing out word frequency, we will find a positive relationship between the L2 norm and recognition memorability in the Cox et al. (2018) and Dymarska et al. (2023) datasets, and a marginally positive effect in the Aka et al. (2023) dataset. E) Regressing out valence and size in the Aka dataset (only dataset with these features available) still revealed a positive relationship between L2 norm and recognition memorability.

word types beyond nouns. As a result, we were only able to control for valence and size in the Aka dataset. Importantly, even after statistically controlling for valence ($\beta$ = .023, t = 6.21, $p$ < .001, 95% CI [.016, .030]) and size ($\beta$ = .025, t(573) = 7.11, $p$ < .001, 95% CI [.018, .032]), the residualized L2 norm still showed a significant positive effect on recognition memorability (see figure 5E). All together, these results indicate that for word stimuli, the relationship between L2norm and memorability are unlikely to be explained by word frequency, valence or size.

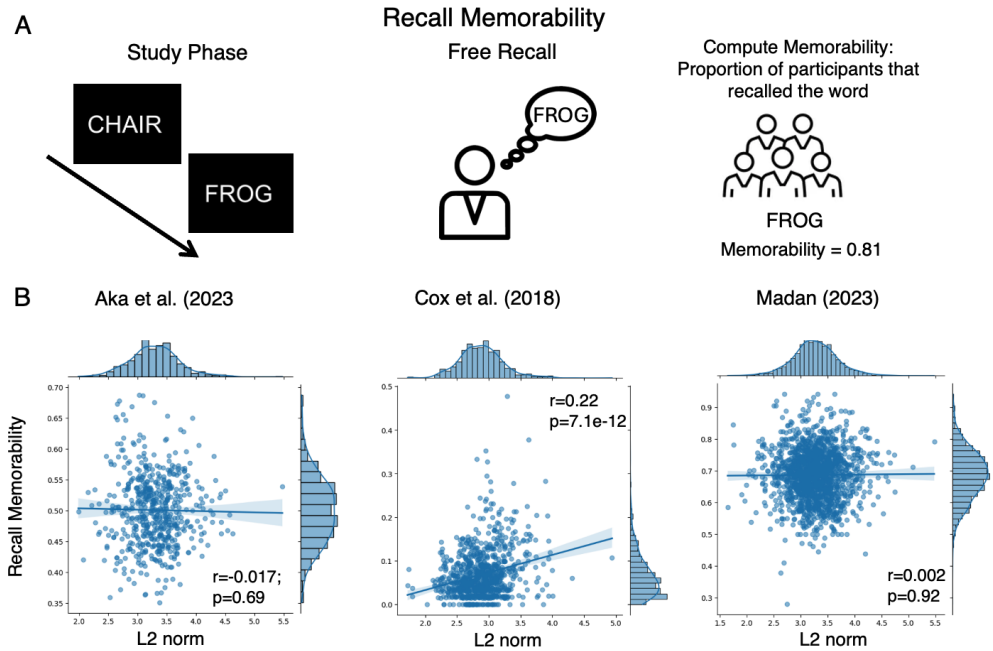



Figure 6: A) Recall memorability for words can be computed by calculating the proportion of participants that recalled a word. B) For the Aka et al. (2023) and Madan (2023) no significant correlation between the L2norm and memorability was found, however for the Cox et al. (2018) dataset there was a significant correlation.

**Word Recall Memorability** In verbal memory experiments, it is common to assess memory either through recognition memory or free recall. Jaegle et al. (2019) were specifically interested in recognition memory: they found their original effect in a continuous recognition task, and linked their effect to neural representations in the ventral visual cortex during object recognition. However, since we found the effect is not specific to visual object recognition or visual processing, a natural question is to ask if it is specific to recognition memory in the first place, or whether it also extends to memory (i.e. "free") recall. To address this, we repeated the same analysis, now analyzing the recall memorability scores in all three verbal datasets. The results are displayed in Figure

6. We found no significant correlation between the L2norm and recall memorability for the Aka (r = -0.017; p = .69, 95% CI [-.10, 0.07]) and Madan (r = .002, p = .92, 95% CI [-.046, .051]) datasets, however in the Cox dataset there was a significant positive correlation (r = .22, p = 7.1e-12, 95% CI [.16, .28]). These results reveal that the L2norm effect as observed for both visual and word memorability does not extend convincingly to recall memorability scores. We discuss the implications of these findings in the discussion.

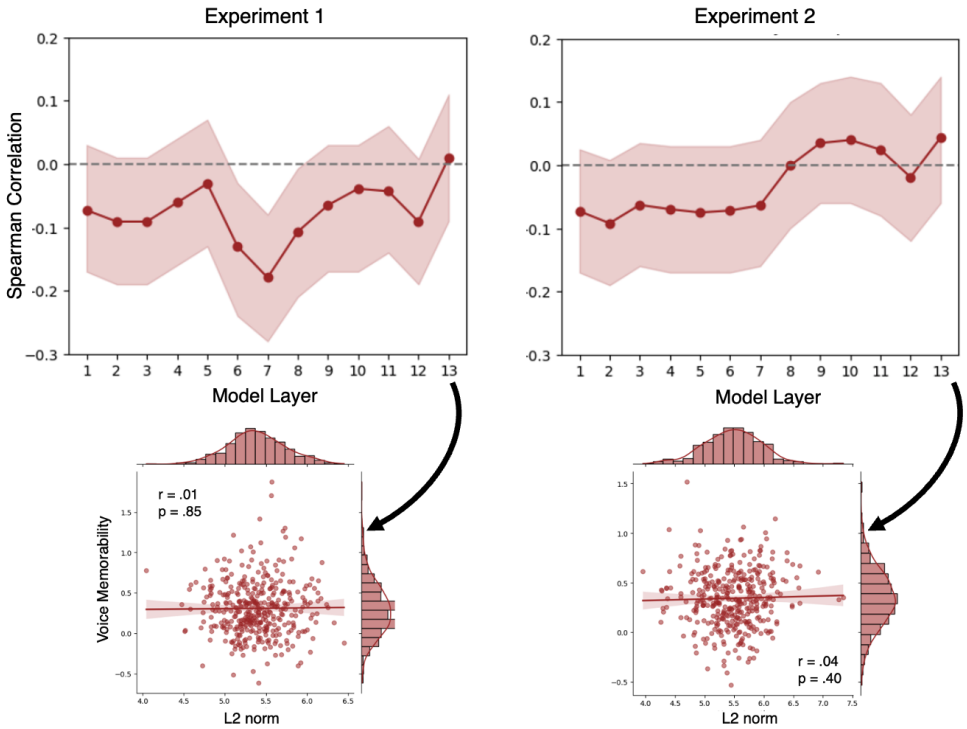



Figure 7: For both experiment 1 and 2 of the Revsine voice memorability datasets there was no reliable relationship between the L2 norm of any layer of wav2vec and voice memorability.

**Voice Memorability** After confirming that the vector magnitude and memorability relationship in the visual domain extends to the domain of word recognition memorability, we investigated whether this L2 norm can also be extended to the auditory domain of voice memorability. We analysed data from the first two experiments as reported by a recently published paper by Revsine et al. (2025), covering more than 600 voice memorability scores across over 2700 participants. We computed the L2 norm of all layers of Wav2Vec (Baevski et al. (2020)) and ran across the voice stimuli a Spearman correlation between the L2 norm and the corresponding memorability score (see methods for details). The results are displayed in figure 7 and all statistical results are reported in table

1. In contrast to the visual and word memorability domain, for wav2vec we did not observe a consistent significant relationship between the vector magnitude and memorability across experiment 1 and 2.

Table 1: Correlation between wav2vec vector magnitude and memorability across layers in experiment 1 and 2 of Revsine et al. (2025)

| **Layer** | **Experiment 1** | | | **Experiment 2** | | |
|---|---|---|---|---|---|---|
| | **r** | **p-value** | **95% CI** | **r** | **p-value** | **95% CI** |
| 1 | -0.07 | .15 | [-.17, .03] | -0.07 | .16 | [-.17, .025] |
| 2 | -0.09 | .08 | [-.19, .01] | -0.09 | .07 | [-.19, .008] |
| 3 | -0.09 | .08 | [-.19, .01] | -0.06 | .22 | [-.16, .035] |
| 4 | -0.06 | .23 | [-.16, .04] | -0.07 | .18 | [-.17, .03] |
| 5 | -0.03 | .54 | [-.13, .07] | -0.08 | .14 | [-.17, .03] |
| 6 | -0.13 | .009 | [-.24, -.03] | -0.07 | .16 | [-.17, .03] |
| 7 | -0.18 | .0004 | [-.28, -.08] | -0.06 | .22 | [-.16, .04] |
| 8 | -0.11 | .04 | [-.21, -.007] | -0.0008 | .99 | [-.1, .1] |
| 9 | -0.07 | .21 | [-.17, .03] | 0.04 | .49 | [-.06, .13] |
| 10 | -0.04 | .44 | [-.17, .03] | 0.04 | .44 | [-.06, .14] |
| 11 | -0.043 | .40 | [-.14, .06] | 0.02 | .64 | [-.08, .13] |
| 12 | -0.09 | .07 | [-.19, .008] | -0.02 | .71 | [-.12, .08] |
| 13 | 0.01 | .85 | [-.09, .11] | 0.04 | .40 | [-.06, .14] |

## 4. Discussion

In this paper, we investigated the relationship between the magnitude of DNN representations and memorability scores of linguistic and visual stimuli. We first replicated the findings of Jaegle et al. (2019), demonstrating that vector magnitude in later layers of a CNN (AlexNet) positively correlates with image memorability scores from the THINGS dataset, even after controlling for potential confounds such as image typicality. Extending this line of research to the linguistic domain, we found that the L2 norm of neural word embeddings was also positively correlated with recognition memorability across three independent datasets, demonstrating that the effect generalizes beyond visual recognition to the lexical domain. When we tested this effect in the auditory domain using voice memorability data from Revsine et al. (2025) and embeddings from wav2vec, however, we did not observe a significant relationship between vector magnitude and voice memorability. Together, these findings provide a new perspective on DNN stimulus encoding and stimulus memorability across sensory modalities. The presence of the L2 norm effect in both visual and lexical domains suggests that it may reflect a more general representational coding scheme, rather than one restricted to vision. This generalizability underscores the generality of the effect, and the potential of the geometry of distributed representations as a tool for studying understanding memorability.

We observed the L2 norm effect in both visual and lexical datasets which makes it a genuine effect that merits interpretation. A key question that stands

out based on our findings is: what does the L2 norm, as observed in the visual and lexical datasets, represent? One possible interpretation is that representational magnitude in these models reflects the richness or distinctiveness of features, which may in turn drive memorability, if these features are encoded by DNN representations and the brain. In the biological brain, larger neuronal amplitude is related to stimulus encoding: stimuli that tend to be recognised (high memorability) on average elicit a stronger neuronal amplitude during encoding, which may co-vary with activating more features and hence leaving a stronger memory trace (Vogelsang et al. (2018); Jaegle et al. (2019)). Equivalently, in DNNs, if a stimulus projects strongly onto learned feature dimensions (i.e., activates feature detectors in vision models or aligns with interpretable directions in language model embeddings), the L2 norm of that representation will be higher. And if there is an alignment of features or dimensions of the DNN and those of the brain (Chen and Bonner, 2025) then the neural network L2 norm will also correlate with the degree of activation and memory traces in a biological system. It is the distinctiveness of those features that results in a larger L2 norm and thus higher memorability. The importance of feature distinctiveness is well illustrated in the lexical domain: polysemous words, which have multiple meanings, tend to have smaller vector magnitudes because their usage is spread across diverse contexts (and thus there is more semantic ambiguity), pulling the representation in multiple directions Schakel and Wilson (2015). As a result, the L2 norm of polysemous words is lower. Supporting this idea, Tuckute et al. (2025) found that words with fewer synonyms (i.e., less semantic ambiguity) were more memorable, suggesting that semantic specificity leads to more distinctive feature representations, higher vector magnitudes, and ultimately greater memorability. This pattern is further supported by our observation that the L2 norm effect is weaker in the early layers of the DNN and becomes more pronounced in the later layers. These higher layers are known to encode increasingly abstract, conceptual features, aligning with neuroimaging findings showing that memorability effects emerge in higher-order regions of the medial temporal lobe, but not in early visual areas (Bainbridge et al., 2017). Similarly, in word2vec, the L2 norm appears to reflect projection onto interpretable semantic axes, further reinforcing the idea that representational magnitude is closely tied to features that encode conceptual/semantic richness in both artificial and biological systems.

We found a reliable L2 norm effect for visual and lexical memorability, but not for voice memorability. We see two potential explanations for this discrepancy. First, it may be that voice memorability, as compared to memorability of visual stimuli, is itself inherently less consistent across participants, making statistical relations of variables that co-vary with memorability more difficult to observe. Indeed, Revsine et al. (2025) reported that memorability of voices is generally less consistent across individuals than memory for faces, with lower inter-subject agreement ($p = 0.28$ vs. 0.82) and lower average d' scores (0.31 for voices vs. 1.11 for faces), which could be largely due to high number of false alarm rates in the voice memorability data. Second, it might be that for voices, there is less alignment between the features driving voice memorability

on the one hand, and voice representational magnitude on the other. Based on recent work, voice memorability appears to be primarily driven by relatively lower level acoustic features such as pitch and prosody (Revsine et al., 2025), which separates it from the relatively higher-level features driving visual and lexical memorability, such as category distinctiveness or conceptual/semantic richness (Bainbridge et al., 2017; Xie et al., 2020; Kramer et al., 2023). However, this difference does not explain why we see no consistent relation between voice memorability and representational magnitude, even for early layers of speech models such as wav2vec2, which should encode such low-level features. It could be that these features are encoded in a less aligned way in wav2vec, which, combined with the lower consistency of voice memorability, effectively masks this representational magnitude effect. Characterizing the mechanisms underlying this discrepancy between visual/lexical and voice memorability, and their relationship to representational magnitude, is an exciting avenue for future research.

In this study we focused on recognition memorability, following the original L2 norm effect reported by Jaegle et al. (2019). However, for words, memorability is often not only assessed through a recognition memory test but also through free recall. And while we observed a robust L2 norm effect for word recognition memorability, this effect did not consistently extend to recall memorability. The absence of a consistent effect in recall-based memorability may be due to fundamental differences between recognition and recall processes, which are known to rely on partially distinct neural systems (Cabeza et al. (1997)). Indeed, we found that overall the correlation between lexical recognition and recall memorability was weak (i.e. $r < 0.1$). Moreover, the dataset where this correlation was non-negligible (Cox et al., 2018 with r=0.2; p = 2.18e-1 versus r=0.04 and p = 0.08 in Aka et al., 2023), was also the only dataset in which we observed a relationship between L2 Norm and recognition, which further suggests that the representational magnitude effect is primarily related to recognition memory. This is likely explained by the fact that free recall places greater demands on strategic retrieval processes and may be influenced by additional cognitive and contextual factors (Yonelinas et al. (2022)). Additionally, word features such as word frequency seem to affect recognition versus free recall differently: whereas recognition memory tends to be negatively correlated with word frequency (i.e. less frequent words are more easily recognised in an old/new recognition memory test) the opposite tends to be true for free recall: items that are more frequently used in daily language are more easily recalled Aka et al. (2023). Thus, recognition memory versus free recall rely on different cognitive and neural mechanisms which could explain the absence of the representation magnitude effect for recall memorability.

A promising avenue for future research is to directly compare neural representations in the human brain with the representational magnitude of DNNs. While our results show that L2 norm predicts memorability in visual and lexical domains, and aligns with the more general finding that higher response magnitudes during encoding correlates with higher subsequent memorisation rates, it would be interesting to study this systematically, quantifying the joint neural

and DNN representational magnitudes for each item. Such studies could subsequently assess whether artificial networks and the brain share similar coding schemes for memorable information, and whether the L2 norm effect is uniform across representational directions or is pronounced for certain "universal dimensions of representation", as recently discovered in the visual modality Chen and Bonner (2025).

To conclude, our findings demonstrate that the magnitude of neural network representations, their L2 norm, reliably predicts memorability for visual and lexical stimuli, but not for auditory stimuli such as voices. This effect was robust across multiple datasets and could not be explained by confounding factors such as image typicality, word frequency, valence, or size. These results suggest that representational magnitude may serve as a marker of distinct features of representation, capturing the extent to which a stimulus is meaningfully embedded along distinct conceptual/semantic dimensions in space. More broadly, our findings contribute to a growing body of evidence showing that geometric properties of DNNs can offer insights into cognitive processes such as memory. By linking representational strength to memorability across stimulus domains, this work provides a step toward understanding how both artificial and biological systems encode information in ways that make certain experiences more likely to be remembered.

**Declaration of competing interest** The authors declare no competing financial interests

**Data and code availability statement** The datasets reported in this paper are all publicly available through the links that are mentioned in the main text. The code used for the analyses is publicly available via: `https://github.com/davidamadeusvogelsang/Memorability_L2norm_project` .

**CRediT authorship contribution statement** DAV: Conceptualization, data analysis, writing (original+final draft). MH: Conceptualization, writing (original+final draft).

**Declaration of generative AI and AI-assisted technologies in the writing process** During the preparation of this work the author(s) used ChatGPT4o in order to improve the readiblity of some sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

**Acknowledgements** M.H. is supported by a Veni grant obtained from the NWO (Dutch Research Council)

## References


Aka, A., Bhatia, S., McCoy, J., 2023. Semantic determinants of memorability. Cognition 239, 105497.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33, 12449–12460.

Bainbridge, W.A., Dilks, D.D., Oliva, A., 2017. Memorability: A stimulus-driven perceptual neural signature distinctive from memory. NeuroImage 149, 141–152.

Bainbridge, W.A., Hall, E.H., Baker, C.I., 2019. Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. Nature communications 10, 5.

Bainbridge, W.A., Isola, P., Oliva, A., 2013. The intrinsic memorability of face photographs. Journal of Experimental Psychology: General 142, 1323–1334.

Bainbridge, W.A., Rissman, J., 2018. Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. Scientific Reports 8, 8679.

Brysbaert, M., New, B., 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. Behavior research methods 41, 977–990.

Cabeza, R., Kapur, S., Craik, F.I., McIntosh, A.R., Houle, S., Tulving, E., 1997. Functional neuroanatomy of recall and recognition: A pet study of episodic memory. Journal of cognitive neuroscience 9, 254–265.

Chen, Z., Bonner, M.F., 2025. Universal dimensions of visual representation. Science Advances 11, eadw7697.

Cortese, M.J., Khanna, M.M., Hacker, S., 2010. Recognition memory for 2,578 monosyllabic words. Memory 18, 595–609.

Cortese, M.J., McCarty, D.P., Schock, J., 2015. A mega recognition memory study of 2897 disyllabic words. Quarterly Journal of Experimental Psychology 68, 1489–1501.

Cox, G.E., Hemmer, P., Aue, W.R., Criss, A.H., 2018. Information and processes underlying semantic and episodic memory across tasks, items, and individuals. Journal of Experimental Psychology: General 147, 545.

Davis, T.M., Bainbridge, W.A., 2023. Memory for artwork is predictable. Proceedings of the National Academy of Sciences 120, e2302389120.

DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends in cognitive sciences 11, 333–341.

Dymarska, A., Connell, L., Banks, B., 2023. More is not necessarily better: How different aspects of sensorimotor experience affect recognition memory for words. Journal of Experimental Psychology: Learning, Memory, and Cognition 49, 1572.

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, 1173–1185.

Isola, P., Xiao, J., Parikh, D., Torralba, A., Oliva, A., 2013. What makes a photograph memorable? IEEE transactions on pattern analysis and machine intelligence 36, 1469–1482.

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, e47596.

Khosla, A., Raju, A.S., Torralba, A., Oliva, A., 2015. Understanding and predicting image memorability at a large scale, in: Proceedings of the IEEE international conference on computer vision, pp. 2390–2398.

Kramer, M.A., Hebart, M.N., Baker, C.I., Bainbridge, W.A., 2023. The features underlying the memorability of objects. Science advances 9, eadd2981.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90.

Lindsay, G.W., 2021. Convolutional neural networks as a model of the visual system: Past, present, and future. Journal of cognitive neuroscience 33, 2017–2031.

Lukavskỳ, J., Děchtěrenko, F., 2017. Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. Attention, Perception, & Psychophysics 79, 2044–2054.

Madan, C.R., 2021. Exploring word memorability: How well do different word properties explain item free-recall probability? Psychonomic Bulletin & Review 28, 583–595.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26.

Needell, C.D., Bainbridge, W.A., 2022. Embracing new techniques in deep learning for estimating image memorability. Computational Brain & Behavior 5, 168–184.

Ongchoco, J.D.K., Chun, M.M., Bainbridge, W.A., 2023. What moves us? the intrinsic memorability of dance. Journal of Experimental Psychology: Learning, Memory, and Cognition 49, 889.

Revsine, C., Goldberg, E., Bainbridge, W.A., 2025. The memorability of voices is predictable and consistent across listeners. Nature Human Behaviour , 1–11.

Rust, N.C., Jannuzi, B.G., 2022. Identifying objects and remembering images: insights from deep neural networks. Current Directions in Psychological Science 31, 316–323.

Rust, N.C., Mehrpour, V., 2020. Understanding image memorability. Trends in cognitive sciences 24, 557–568.

Schakel, A.M., Wilson, B.J., 2015. Measuring word significance using distributed representations of words. arXiv preprint arXiv:1508.02297 .

Tuckute, G., Mahowald, K., Isola, P., Oliva, A., Gibson, E., Fedorenko, E., 2025. Intrinsically memorable words have unique associations with their meanings. Journal of Experimental Psychology: General .

Vogelsang, D.A., Gruber, M., Bergström, Z.M., Ranganath, C., Simons, J.S., 2018. Alpha oscillations during incidental encoding predict subsequent memory for new "foil" information. Journal of cognitive neuroscience 30, 667–679.

Wakeland-Hart, C.D., Cao, S.A., deBettencourt, M.T., Bainbridge, W.A., Rosenberg, M.D., 2022. Predicting visual memory across images and within individuals. Cognition 227, 105201.

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, 937–948.

Yamins, D.L., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience 19, 356–365.

Yonelinas, A.P., Ramey, M.M., Riddell, C., Kahana, M., Wagner, A., 2022. Recognition memory: The role of recollection and familiarity. AD Kahana, MJ, & Wagner (Ed.), The Oxford handbook of human memory .