Task-sensitive retrieval from semantic memory

Andrew Z. Flores (azf2@illinois.edu)
Department of Psychology, 603 E Daniel St
Champaign, IL 61820 USA

Jon A.Willits (jwillits@illinois.edu)
Department of Psychology, 603 E Daniel St
Champaign, IL 61820 USA

Abstract

This study investigates the interaction between semantic relatedness and goals or task on memory retrieval. We used varied tasks and concepts to explore how task influences how different kinds of semantic relatedness influences semantic processing. Our findings reveal a task-dependent interaction with semantic relatedness. Specifically, in similarity judgement tasks (experiments 1a and 1b), participants' ratings closely aligned with taxonomic relatedness, influenced by abstract visual and linguistic similarity dimensions. In discrimination tasks (experiments 2a and 2b), where participants distinguished a target from a semantically related distractor, visual characteristics explained a greater amount of variance. These results suggest semantic memory representations are dynamic and task-dependent, supporting theories of a distributed semantic memory system.

Keywords: semantic relatedness; memory retrieval; distributional learning; eye-tracking

Introduction

The structure of semantic memory has been a long-standing question in cognitive science. A primary means by which this has been studied is the presentation of stimuli varying in semantic relatedness, and using response accuracy and reaction time to make inferences about the system's representations and processes. The underlying assumption of this approach is that related concepts are represented in a manner that affects retrieval.

For example, semantic priming (Hutchison, 2003) involves presenting participants with pairs of words (prime and target) and observing responses to the target word. Reaction times are often faster for related prime-target pairs, and by manipulating the prime-target relationship, this can be used to study the types of relationships encoded in semantic memory. In addition to priming, other tasks like self-paced reading (Jegerski, 2013), and list memory tasks (Bower, Clark, Lesgold, & Winzenz, 1969; Roediger & McDermott, 1995), and semantic feature verification judgments (McRae, Cree, Seidenberg, & McNorgan, 2005) have all been used to study the structure and function of semantic memory.

However, using these techniques to analyze the semantic relationship between two words is not straightforward. Consider *banana* and a *strawberry*, which share many relationships. They are both members of the same taxonomic category, share many semantic features, and have overlapping sets of associations and roles they can fill. The fact that strawberry might prime banana could be explained by any of these factors, and attempts to determine which kinds of relationships do and do not lead to priming have not met with success (Hutchison, 2003). An additional problem has been that there has been inconsistencies in which studies find facilitated semantic processing for different kinds of relationships (Willits, Amato, & MacDonald, 2015).

One possible explanation for the problems above is that the picture is overly simplistic. Perhaps semantic memory is not organized in a static structure with a set of semantic relationships that are always activated and always lead to facilitated processing. This simple picture does not take into account the context, goals, or other top-down constraints on a person interacting with those objects or words. They may be engaged in a task that requires the activation of specific features (e.g., shape, flavor, color), specific functions (e.g., juiciness, tastiness), specific roles (e.g., for eating, or as a decoration), or specific associative relationships (e.g., strawberry jam, where the bananas are in a grocery store). Different features or relations may be activated differentially as a function of the top-down context, leading to disparate patterns of facilitation if these factors are not taken into account.

Rather than viewing semantic memory as a static set of relationships, an alternative perspective is to treat semantic memory as a dynamical system that is sensitive to the task being performed. The idea of semantic memory structure as a dynamical system that integrates and differentially activates various modality specific information has long been examined in the context of neuroscience research (Binder & Desai, 2011; Ralph, Jefferies, Patterson, & Rogers, 2017). The perspective also has historical support within cognitive psychology. For example, in the classic "Transfer Appropriate Processing" experiment, Morris, Bransford, and Franks (1977) showed that retrieval from memory could be better explained by whether a memory retrieval task matched the memory encoding task, rather than how "deeply" the stimulus was processed during encoding. Barsalou (1983) demonstrated that human categorization abilities are extremely dynamic, and that the features used for categorization change dramatically depending on the context. In related research, Nosofsky (1986) demonstrated in category learning experiments that different similarity structures are for exemplar recall and categorization. Willits et al. (2015) showed that whether participants were activating linguistic knowledge or world event knowledge could be shifted dramatically based on whether the task was better performed with one kind of information or the other.

Current Research

Our current research builds on the previous research suggesting task sensitive retrieval from semantic memory. Central to our investigation is the hypothesis that the nature of the task and semantic relatedness between concepts interact during semantic memory retrieval. In order to test this notion, we use a set of concepts (described below) that are paired so as to vary their degree of semantic relatedness. Using these pairs of

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

concepts, we quantify their semantic relatedness along three primary dimensions: 1) low-level visual features, 2) high-level visual features used for object recognition and classification, and 3) linguistic distributional semantic features. We use these three types of features to investigate the extent to which they predict unique or overlapping information in different behavioral tasks. Our two main experiments are a semantic similarity rating task (Experiment 1), and a visual object identification task (Experiment 2). This manipulation was designed to elicit the activation of different semantic features. Within each experiment, we also performed additional task manipulations designed to shift the relative importance of different features.

I: Stimuli

Our stimulus set consisted of 128 concepts from two top-level categories (natural and human-made), four superordinate categories (animals, foods, room objects, tools), and four subordinate categories from each superordinate categories(e.g., from animals, birds, insects, mammals, and sea creatures). . The items were used to create sets of pairings such that words were always paired with a word matched in frequency (high vs. low), and each concept participated in each of the following pairings that manipulated strength of relatedness. **Level 4** (most related) pairs were the most similar pairs from the same subordinate categories, with high visual and linguistic feature overlap. **Level 3** pairs were less similar concepts from the same subordinate category. **Level 2** pairs were from the same superordinate category but from different subordinate categories. **Level 1** (least related) pairs were from different superordinate categories.

The pairings were counterbalanced such that each concept occurred in all four conditions. For each concept, simple images were selected that showed the object on a white background. Examples of the images and their pairings are shown in Figure 1 below.

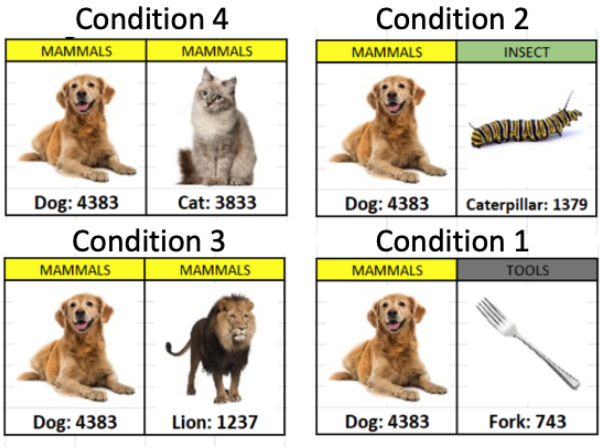

Figure 1: Example pairings from the four semantic relatedness conditions and their corresponding word frequencies.

II: Measures of Semantic Relatedness

This section focuses on deriving metrics of semantic relatedness in terms of low level visual similarity, high level visual features useful for object categorization, and distributional linguistic similarity. The hypothesis was that different tasks would engage these features to different degrees, and thus the similarity of the concepts in terms of the different kinds of features would have different degrees of predictive power on different experiments.

Low-Level Visual Features

Lower level features of images such as shape and color have previously been shown to impact memory retrieval. For instance color relations have been shown to influence allocation of attention in visual world paradigms (Huettig & Altmann, 2011) (i.e., hearing the word *lime* diverts looks toward a similarly colored *frog*). Thus, we examined three measures emphasizing lower-level perceptual information:

Histogram of Oriented Gradients (Edge-Detection): Utilized for object detection in computer vision. It calculates gradient orientation in image parts, capturing local shape characteristics (Dalal & Triggs, 2005).

Histogram of Color Similarity (RGB Histogram): Analyzes the color distribution in images to assess color similarity (Stricker & Orengo, 1995).

Structural Similarity Index (Struct-Similarity): Measures perceived image quality compared to a reference, focusing on structural changes, luminance, and texture, rather than pixel differences (Wang, Bovik, Sheikh, & Simoncelli, 2004).

High Level Visual Features for Object Classification

There are many ways one could go about identifying visual features useful for object recognition and categorization. We chose to use the internal representations of convolutional neural networks, which are both very good at object classification, and have been shown to be correlated with human behaviors (Van Dyck, Kwitt, Denzler, & Gruber, 2021). We used ResNet-50 v1.5 (He, Zhang, Ren, & Sun, 2016) to transform visual stimuli into high-dimensional feature vectors. In particular we were interested in contrasting initial and deeper layers in the CNN, due to research which has found that earlier layers (L0) primarily encode basic visual elements such as brightness, hue, and contours, and later layers (L7 ResNet) encode abstract object features and relationships useful in object categorization (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015; Zeiler & Fergus, 2014).

Distributional Linguistic Similarity

A popular approach to quantifying the semantic similarity between two words has been the extent to which the two words share linguistic contexts (Burgess & Atchley, n.d.). Distributional semantic models have yielded many insights into the kinds of concept representations which are predictive of semantic memory retrieval performance (McDonald & Lowe, 2022). In this study, we used the similarity of two words according to the Word2Vec Skip-Gram model

(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This model constructs a high-dimensional semantic space by minimizing the error predicting words near the target word.

Comparing Semantic Relatedness Measures

Our first analysis examined the similarity distribution of the different relatedness measures across our four stimulus relatedness conditions. For each of our concepts, we created a feature vector based on each of the six above-described measures of semantic similarity. We then computed the similarity scores of these measures for all the pairs in our four relatedness conditions. These results are shown in Figure 2

We found that similarity in terms of lower level visual measures (RGB Histogram, Structural Similarity and ResNet Layer 0) had overlapping distributions and did not vary significantly across relatedness conditions. In contrast Word2Vec and ResNet Layer 7 better captured the graded levels of relatedness across our conditions. This is in some respects expected given how both Word2Vec and ResNet Layer 7 each encode information that has been used for semantic categorization of images and words respectively, which is embodied in the nature of our different conditions. But it is notable and important for interpreting the later results, that the lower level features did not vary.

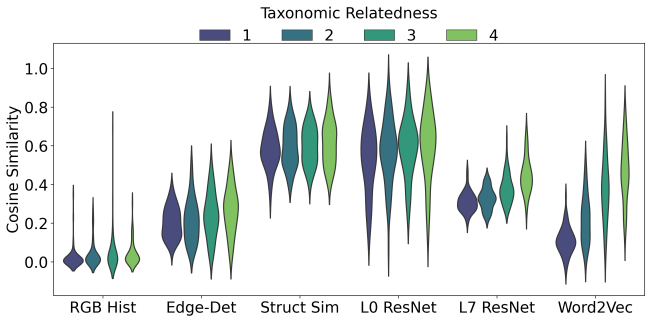

Figure 2: Violin plots showing the distribution of cosine similarity scores across all item-pairs and taxonomic relatedness.

We were also interested in knowing the intercorrelations of our similarity measures. Because the nature of the distributions of the different measures varied considerably, we computed a Spearman rank correlations for each measure (e.g., did the similarity rank of pairs in terms of their Resnet Layer 7similarity predict their similarity rank in terms of Word2Vec similarity). Overall the pattern of results show greater correlations within lower-level visual features and higher level (linguistic and visual) features, rather than between high and low level features. As shown in Figure 2, the strongest correlation was between Structural Similarity and ResNet layer 0, both measures that encode low level features of images. Similarly we observed a negative correlation between Structural Similarity and Edge Detection (-0.28), which may be attributable to the nature of each method. Structural Similarity encodes aspects of image quality that focus on broader textural features while Edge Detection focuses on pixel-wise properties. The higher level features were also correlated with each other. The distributional linguistic semantic measures of Word2Vec were correlated with Resnet layer 7 (0.45).

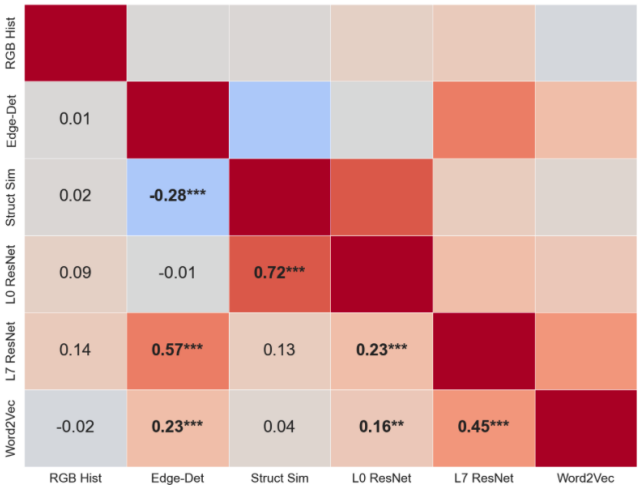

Figure 3: Correlation matrix of semantic relatedness measures, bolded values indicate significance after Bonferroni correction. $p<0.05$, ** $p<0.01$, *** $p<0.001$

However, exceptions to this pattern emerged. For instance, the Spearman rank correlation between Edge-Detection and Word2Vec were low but significant (0.23). One possibility may regard the prominent role that shape plays in conceptual representations (Landau, Smith, & Jones, 1988). Edge-Detection was also highly correlated with ResNet layer 7 (0.57), also not surprising how important shape might be to object categorization.

To summarize, our concept pairs varied considerably along all six feature measures. This provide ample variance along visual and linguistic dimensions that may be useful for predicting behavioral data in the following experiments. Some of the similarity measures (Word2Vec, Resnet Layer 7) correlated strongly with the differences in semantic relatedness defined in our design. Others (low level visual features) did not. In the following we examine how these disparate measures are recruited during semantic memory retrieval.

Experiment 1: Semantic Similarity Judgement

Across two sub-experiments (1a and 1b) we asked participants to provide semantic similarity judgements. Participants in both experiments saw all the same items, but their similarity distribution varied across conditions. In experiment 1a participants were shown a series of images with an equal distribution across all four levels of taxonomic relatedness (*i.e.*, one quarter of their items from relatedness conditions 1, 2, 3, *and* 4). In experiment 1b, participants saw only image pairs from a single relatedness condition (*i.e.*, all of their items from either condition 1, 2, 3, *or* 4).

Hypotheses. We hypothesize that the nature of the task will significantly impact which dimensions of meaning matter.

Specifically, we expect that lower level visual features will account for a significant portion of the variance, due to the nature of the tasks and the use of visual stimuli across all experiments. In tasks that rely on complex evaluations of meaning (like similarity ratings) we predicted that Word2Vec would account for more variance than Resnet Layer 7. In tasks that are more similar to object recognition tasks, we predicted that Resnet Layer 7 would predict more variance than Word2Vec.

For experiment 1a in which all four relatedness conditions are present, we hypothesized participants will closely align to these categories in their judgements, further we expect semantic relatedness measures which correlate highly with taxonomic relatedness (Resnet Layer 7 and Word2Vec) to account for most variance in this task, with the linguistic features predicting more variance than the visual features. In contrast, we hypothesize that experiment 1b will require participants to adjust the kinds of relationships which are diagnostic of similarly paired items. That is if participants continually see pairs that are highly semantically related (e.g *dog-cat, tiger-lion, guitar-violin, clarinet-flute*) in which broad category memberships are less useful we may see a more prominent role for visual information. This is largely due to the visual features in Resnet Layer 7 being highly biased towards features that help discriminate easily confusable items.

Methods

We recruited 70 participants for experiment 1a and 120 for experiment 1b from the University of Illinois student participation pool. We asked the participants to rate the semantic relatedness of two presented images using a 5 point likert scale. Participants were provided the following definition of semantic relatedness: "Semantic relatedness refers to the degree to which the concepts represented by the images are similar or share meaning. Consider factors such as similarity in category, function, or appearance". We intended these instructions to be broad so as to not bias the participant to consider any one aspect of meaning to heavily. During the testing phase each trial initially began with the slider positioned in the neutral center position of the likert scale, participants were then free to indicate each image pairs judged semantic relatedness. All image positions were randomized across all participants and all stimuli pairs.

Results

For each of the experiments, we aim to characterize the interaction between semantic relatedness and memory retrieval processes. Using the measures of semantic relatedness from the previous section, we aim to predict the performance in each behavioral experiment using a statistical modeling approach that allows us to quantify the unique contribution of each measure to each task. For each experiment we build a progression of hierarchical models in which the response variable (e.g., similarity rating, reaction times, etc.) are predicted by each measure of semantic relatedness from section I. Notably we took an approach that allows us to measure how the addition of each predictor influences the overall model fit using the Akaike Information Criterion (AIC). This involves first building a model with a single predictor (i.e RGB Hist) and successively adding each one while measuring model fit.

In each experiment, we always first built a model that added the low-level visual features one at a time, resulting in a model with all low level visual features. We then created one model that first added Resnet L7 (high level visual features) and then Word2Vec (linguistic features), and a second model that added Word2Vec first and Resnet Layer 7 second. Because these two measures were correlated, and we were particular interested in the relative importance of visual and linguistisc features across tasks, we wanted to see if they explained different and more variance on different tasks.

Our initial analysis examined the relationship between the similarity ratings given by participants and taxonomic relatedness. A high correlation between these ratings and relatedness condition would suggest that participants' judgments are influenced by these semantic relationships. Conversely, low correlations, where ratings vary significantly and don't align with taxonomic groups, could imply that other factors influence how participants perceive similarity. We also considered how different conditions might affect these ratings. In experiment 1a, participants encountered a variety of relatedness levels. In experiment 1b, they were exposed to only one level of relatedness. Our hypothesis was that the dynamic nature of the task – encountering different levels of relatedness – would impact the ratings. Specifically, we expected that in experiment 1b, where participants experience less variation in taxonomic relatedness, their judgments might rely more on individual interpretations and less on clear taxonomic categories. This could lead to a greater focus on visual aspects or other non taxonomic factors. As depicted in Figure 4 in experiment

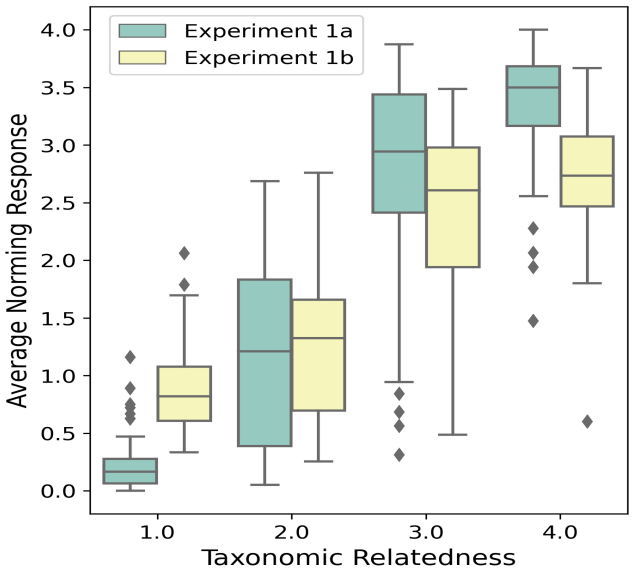

Figure 4: Average norming across taxonomic relatedness condition for all item pairs.

1a, where participants saw a full range of relatedness levels, their ratings more closely aligned with relatedness condition (r=.88). In experiment 1b, with limited exposure to relatedness variety, the ratings were more diverse and less aligned with semantic relatedness(r=.78). This suggests that exposure to a wider range of relationships might make those relationships more prominent in participants' judgments, and may have shifted which features they were attending to.

To further investigate this possibility we built a series of hierarchical models predicting a given similarity judgement response using each relatedness measure. These modeling results are displayed in Table 1, where the Δ AIC indicates the difference in terms of raw AIC scores between each model and the model with the smallest overall AIC value. The χ^2 values reflect the model comparison of each simple model with the next more complex model in sequence (i.e M0 vs M1, M1 vs M2, etc). For experiment 1a, each predictor explained additional significant variance, even the lowest level visual features. In the critical contrast where Word2Vec and Resnet Layer 7 were added in different orders, Resnet Layer 7 and Word2Vec predicted some of the same variance, but more was explained by Word2Vec than Resnet Layer 7. When Word2Vec was added first, little AIC change was seen from then adding Resnet Layer 7, whereas in the other order, adding Word2Vec changed AIC by a large amount even after Resnet Layer 7 was added.

In contrast for experiment 1b, in which participants saw a consistent range of similarity (some saw all highly related, some medium, some all low related), we found that the inclusion of lower level feature predictors including RGB Histogram and Edge-Detection did not improve model fit. Like 1a, we found that both Resnet Layer 7and Word2Vec explained larger amounts of variance. But the relative importance of Resnet Layer 7 was even lower in 1b than in 1a.

Discussion

The patterns of results found in both experiments 1a and 1b, showcase the multiple dimensions of meaning which underlie semantic similarity judgements. As expected, higher level linguistic semantic features dominated the models in terms of what predicted similarity ratings. Visual features mattered to a small extent when participants where shifting from high similarity to low similarity comparisons across trials, and varied much less when they were fixated on similarity comparisons of consistent similarity. One potential reason we saw a reduced role of lower level visual features across both tasks may be due to the nature of the task, in which participants are allowed to consider the extent to which two items are related with no time constraints. Given unlimited time, participants initial considerations of these lower level features may be replaced by more abstract and retrievable kinds of meaning which are more categorical in nature . Overall, in both experiment 1a and 1b we found a consistent set of results that highlight the role of more conceptually driven, abstract measures compared to lower level visual feature information.

Experiment 2: Semantic Discrimination

In Experiment 2 we asked participants to identify a referent target when in competition with a competitor that varied in terms of its semantic relatedness. Experiment 2a was a reaction time task where participants where asked to click on the identified object. Experiment 2b was an eye-tracking task where they were asked to look at one of the objects. We investigated how the distinct measures of semantic relatedness impacted performance in these task. Both tasks utilized the same paradigm, pairs of images varying in similarity and an outcome that captures the cognitive processing of semantic memory retrieval. What differed between these tasks was the difference in the nature of the outcome measures. Experiment 2a measures reaction times, a summary measure. In turn, experiment 2b captures moment by moment cognitive processing allowing us to measure the relative importance of the varied semantic relatedness measures over time.

Methods

The experimental paradigm was the same for both 2a and 2b. Each involved the presentation of pairs of pictures, along with an auditory cue asking the person to select via button press (Experiment 2a) or "look at" (Experiment 2b) one of the two pictures. The image pairs shown were the same as those previously described in our stimuli section. Each participant saw 64 total image pairs distributed evenly across the four taxonomic relatedness levels. Across participants the presentation (i.e left or right) of the target image, as well as the overall order of trials was randomized. The items were put onto counterbalanced lists such that each picture in each pair appeared as a target and a distractor for different participants.

Experiment 2a (Button Press) Design. We recruited 480 participants from the University of Illinois from a student subject pool. Participants were instructed to access an online version of the experiment (https://run.pavlovia.org/azf2/visual/semantics), which implemented the task using the JavaScript library JSpsych (De Leeuw, 2015). Precautions to ensure data quality included providing participants with an opportunity to check whether their audio was working, providing instructions that specified how to enter each choice using the keyboard and practice trials prior to starting the experiment. During a single trial participants were able to input a decision via button press at any point during the trial, with trials ending once a response was recorded.

Experiment 2a Results and Discussion. As predicted, reaction times varied significantly as a function of relatedness, with slower reaction times observed in higher relatedness conditions (Table 2). To understand how distinct similarity measures predicted behavior in experiment 2a, we again applied our model comparison approach. As shown in Table 1, we found that low level visual features explained relatively little variance. In contrast, the higher level features Resnet Layer 7and Word2Vec explained considerably more

Model	Experiment 1a		Experiment 1b		Experiment 2a		Experiment 2b	
	Δ AIC	$\chi^2_{(sig.)}$	Δ AIC	$\chi^2_{(sig.)}$	Δ AIC	$\chi^2_{(sig.)}$	Δ AIC	$\chi^2_{(sig.)}$
M0: + Rand.Eff	8864		1560		149		11740076	
Time: M0 + Time							7611	**895055**
M1: Time + RGB Hist	8733	**139**	1567	2	150	0	7440	**1357**
M2: M1 + Struct Sim	8636	**105**	1576	0	149	**3**	6169	**93**
M3: M2 + L0 ResNet	8397	**246**	1549	**36**	151	0	6154	**19**
M4: M3 + Edge-Det	7639	**767**	1490	**68**	129	**24**	5357	**800**
M5a: M4 + L7 ResNet	4600	**3048**	1253	**246**	29	**102**	960	**4401**
M6a: M5a + Word2Vec	0	**4612**	0	**1263**	0	**31**	0	**964**
M5b: M4 + Word2Vec	836	**6814**	150	**1349**	33	**98**	1939	**3423**
M6b: M5b + L7 ResNet	0	**845**	0	**160**	0	**35**	0	**1943**

Table 1: Note—Δ AIC = [AICi – min(AIC)]. χ^2 values represent the comparison of the simpler vs complex model in succession. Bolded χ^2 values are significant at the $p < 0.001$.

and highly overlapping variance. Whichever one was put in the model first explained the most (ΔAIC near 100, vs. 30).

Experiment 2b (Eye Tracking) Design We recruited 320 participants from the University of Illinois student participant pool. All participants were required to report English as their first language. During the experiment, participants were seated in front of a 16-inch LCD monitor with an SR Research EyeLink 1000 eye tracker. The monitor was positioned a 600 from the participant's forehead, and the auditory stimuli were presented through a speaker below the monitor. Before starting the experiment, a calibration procedure was performed. Each trial started with a 50x50 central fixation point that disappeared once fixated for 50 ms. This was followed by the presentation of the target and distractor image for 2000 milliseconds in silence. Then, the auditory stimulus (variable length) was played, and both images remained on the screen. After the audio offset, the target and distractor remained on the screen for another 4,000 milliseconds.

Experiment 2b Results and Discussion The data was divided into 8ms intervals for analysis. The areas of interest were defined as two 300x300 pixel regions surrounding the target and distractor images. All fixations to the target or distractor were coded as either 1 (fixation toward target) and 0 (fixation toward distractor). All statistical analysis is derived from eye movements that occurred between 150 milliseconds and 2000 milliseconds after the onset of the target label. As we predicted based on previous research using the visual world paradigm, participants were faster to look at objects the more semantically unrelated the target (Table 2). To understand how distinct similarity measures predicted behavior in experiment 2b, we once again applied our model comparison approach, this time putting in each predictor as an interaction with time since target onset. Here we found that all the predictors accounted for significant variance (the first being very high because of the time interaction). For models contrasting Resnet 7 and Word2Vec, both models still predicted relatively high variance when entered first, but this time, in the most visual-feature sensitive task, Resnet 7 was the clear winner compared to Word2Vec.

Relat.	Experiment 2a	Experiment 2b
	Mean RT [95% CI]	Mean Acc. [95% CI]
1	835 [828, 842]	0.753 [0.753, 0.754]
2	842 [835, 849]	0.778 [0.778, 0.779]
3	859 [851, 866]	0.782 [0.782, 0.783]
4	873 [865, 881]	0.781 [0.780, 0.781]

Table 2: Reaction times (Exp 2a) and target fixation (Exp 2b) across semantic relatedness conditions. Note: *RT* is the time required to identify target after noun onset. *Accuracy* reflects the proportion of fixations towards target vs distractor item.

General Discussion

Utilizing a wide range of concepts and a variety of naturalistic semantic memory tasks, we examined how semantic relatedness, defined multiple ways, impacted memory retrieval, and how the task affected how those different kinds of relatedness affected semantic processing. When asked to judge the similarity between concepts (Exp 1a and 1b), participants judgments were highly correlated with measures of relatedness reflecting the taxonomic organization of the stimuli, with linguistic features predicting the most variance. In experiments where participants were asked to discriminate a target referent from a semantically related distractor (Exp 2a and 2b), we found an overall pattern of increasing interference with increased semantic relatedness. These interference effects were characterized in our modeling work, which showed that the more "visual" the task, the more visual rather than linguistic semantic features predicted the behavior.

These findings that retrieving concepts from memory activates qualitatively distinct sets of features across distinct dimensions of meaning supports the notion of a highly dynamic representational system that is sensitive to the relations highlighted or available within the current task demands.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, *11*, 211–227.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in cognitive sciences*, *15*(11), 527–536.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of verbal Learning and verbal Behavior*, *8*(3), 323–343.

Burgess, C., & Atchley, R. A. (n.d.). Semantic and associative priming in high-dimensional semantic space.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 1, pp. 886–893).

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*, 1–12.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Huettig, F., & Altmann, G. T. (2011). Looking at anything that is green when hearing "frog": How object surface colour and stored object colour knowledge influence language-mediated overt attention. *Quarterly Journal of Experimental Psychology*, *64*(1), 122–145.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic bulletin & review*, *10*, 785–813.

Jegerski, J. (2013). Self-paced reading. In *Research methods in second language psycholinguistics* (pp. 36–65). Routledge.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

McDonald, S., & Lowe, W. (2022). Modelling functional priming and the associative boost. In *Proceedings of the twentieth annual conference of the cognitive science society* (pp. 675–680).

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

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*.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behavior*, *16*(5), 519–533.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, *18*(1), 42–55.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Stricker, M. A., & Orengo, M. (1995). Similarity of color images. In *Storage and retrieval for image and video databases iii* (Vol. 2420, pp. 381–392).

Van Dyck, L. E., Kwitt, R., Denzler, S. J., & Gruber, W. R. (2021). Comparing object recognition in humans and deep convolutional neural networks—an eye tracking study. *Frontiers in Neuroscience*, *15*, 750639.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive psychology*, *78*, 1–27.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i 13* (pp. 818–833).