NeuroImage 128 (2016) 44–53

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Full Length Articles

Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities

Andrew James Anderson *, Benjamin D. Zinszer, Rajeev D.S. Raizada

Brain and Cognitive Sciences, University of Rochester, NY 14627, USA

ARTICLE INFO

Article history:
Received 13 August 2015
Accepted 19 December 2015
Available online 28 December 2015

Keywords:
Representational Similarity Analysis
Encoding
Decoding
Semantic model
Semantic memory
fMRI

ABSTRACT

Patterns of neural activity are systematically elicited as the brain experiences categorical stimuli and a major challenge is to understand what these patterns represent. Two influential approaches, hitherto treated as separate analyses, have targeted this problem by using model-representations of stimuli to interpret the corresponding neural activity patterns. Stimulus-model-based-encoding synthesizes neural activity patterns by first training weights to map between stimulus-model features and voxels. This allows novel model-stimuli to be mapped into voxel space, and hence the strength of the model to be assessed by comparing predicted against observed neural activity. Representational Similarity Analysis (RSA) assesses models by testing how well the grand structure of pattern-similarities measured between all pairs of model-stimuli aligns with the same structure computed from neural activity patterns. RSA does not require model fitting, but also does not allow synthesis of neural activity patterns, thereby limiting its applicability. We introduce a new approach, representational similarity-encoding, that builds on the strengths of RSA and robustly enables stimulus-model-based neural encoding without model fitting. The approach therefore sidesteps problems associated with overfitting that notoriously confront any approach requiring parameter estimation (and is consequently low cost computationally), and importantly enables encoding analyses to be incorporated within the wider Representational Similarity Analysis framework. We illustrate this new approach by using it to synthesize and decode fMRI patterns representing the meanings of words, and discuss its potential biological relevance to encoding in semantic memory. Our new similarity-based encoding approach unites the two previously disparate methods of encoding models and RSA, capturing the strengths of both, and enabling similarity-based synthesis of predicted fMRI patterns.

Introduction

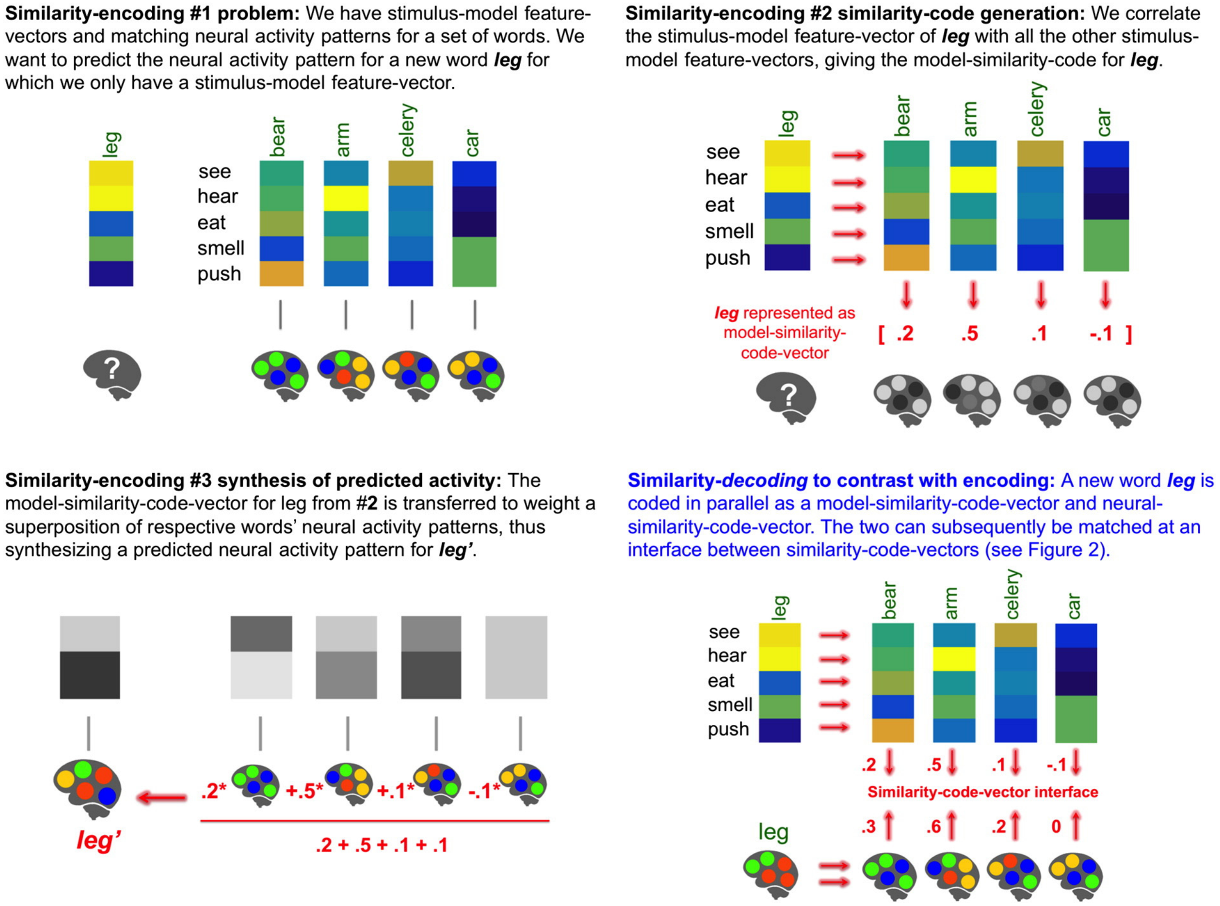
The brain represents different categories as spatially distributed and overlapping activity patterns, and a major challenge is to crack this representational code (Haxby et al., 2001; Haxby et al., 2014). Neural activity can be elicited by presenting participants with various stimuli (e.g. words, images, sounds) and recorded by neuroimaging techniques such as functional Magnetic Resonance Imaging (fMRI). Two approaches targeting the problem of explaining the resultant neural codes are stimulus-model-based-encoding and Representational Similarity Analysis (RSA). Stimulus-model-based-encoding forms models of stimuli as vectors of feature-weights. For pictorial stimuli, model-features may correspond to visual filters (e.g. Kay et al., 2008; Naselaris et al., 2009), for words, features may be the association of the word with senses used to experience the word's referent (e.g. Mitchell et al., 2008; Fernandino et al., 2015). Synthesized neural activity patterns corresponding to new model-stimuli are predicted by a mapping from model-features to voxels trained by fitting weights to features with supervised learning. In contrast, RSA assesses models by comparing the grand structure of similarities between all pairs of stimulus-model feature-vectors and neural activity patterns, and does not require model fitting but cannot synthesize predicted voxel-space activation patterns.

We present a new approach, similarity-encoding, that bridges between stimulus-model-based-encoding and RSA. The new method is illustrated in Fig. 1. This approach achieves similar accuracy in synthesizing predicted neural activity patterns to standard regression-based strategies, but without model fitting. Hence unlike standard regression we observe that similarity-encoding robustly manages situations where there are many more stimulus-model dimensions than stimuli. We also show how this new approach enables stimulus-model-based-decoding of novel fMRI data to be entirely abstracted to representational-similarity space (Fig. 2). Thus, like regression there is generalization from trained to untrained stimuli. However, the generalization here stems from exploiting the structure of similarity-space.

Encoding and decoding (discussed in detail in the context of fMRI by Naselaris et al., 2011) are of broad relevance to assess the value of models/and or neural data to making practical decisions, e.g., clinically in distinguishing healthy and unhealthy samples (e.g., Just et al.,

* Corresponding author.
E-mail address: andrewanderson@bcs.rochester.edu (A.J. Anderson).

http://dx.doi.org/10.1016/j.neuroimage.2015.12.035
1053-8119/© 2015 Elsevier Inc. All rights reserved.


Similarity-encoding #1 problem: We have stimulus-model feature-vectors and matching neural activity patterns for a set of words. We want to predict the neural activity pattern for a new word ***leg*** for which we only have a stimulus-model feature-vector.

Similarity-encoding #2 similarity-code generation: We correlate the stimulus-model feature-vector of ***leg*** with all the other stimulus-model feature-vectors, giving the model-similarity-code for ***leg***.

Similarity-encoding #3 synthesis of predicted activity: The model-similarity-code-vector for leg from **#2** is transferred to weight a superposition of respective words' neural activity patterns, thus synthesizing a predicted neural activity pattern for ***leg'***.

Similarity-*decoding* to contrast with encoding: A new word ***leg*** is coded in parallel as a model-similarity-code-vector and neural-similarity-code-vector. The two can subsequently be matched at an interface between similarity-code-vectors (see Figure 2).

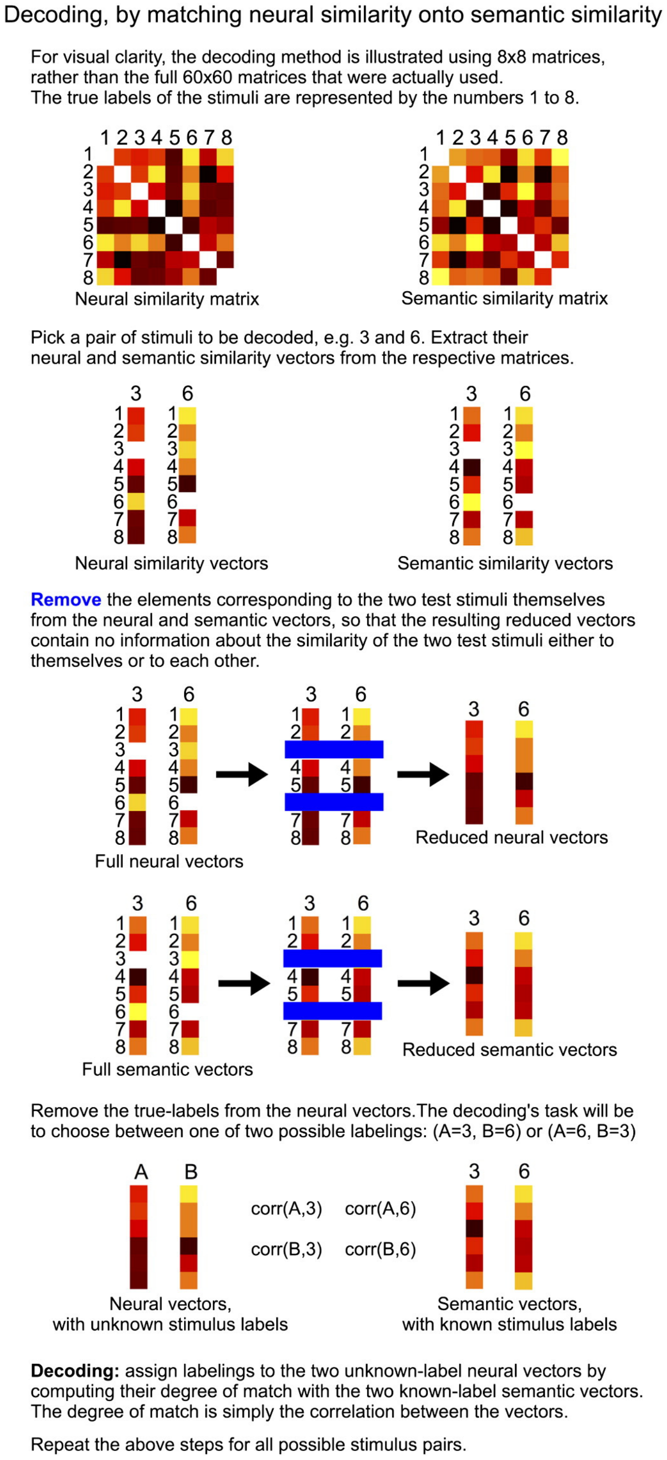
Fig. 1. The three stages of similarity-based neural-activity-pattern encoding. Separate to this the fourth panel illustrates similarity-based-decoding for contrast with encoding in the other three panels (see Fig. 2 for further details of the new similarity-based decoding algorithm).

2014; Matthews et al., 2006), in brain–computer-interfaces and neuroprosthetics (e.g. Sulzer et al., 2013; deCharms, 2008), or from an ecological perspective to estimate whether measured neural activity patterns could actually be the grounds of decision making within an individual. As such whilst RSA and neural encoding and decoding have tended to be treated as separate analyses with different properties and benefits (e.g. Haxby et al., 2014), the extension introduced here provides a means for all types of analyses to be easily undertaken within the same similarity based framework. Where previous analyses have decoded neural activity patterns using representational-similarity methods (e.g. Raizada and Connolly, 2012, Nili et al., 2014; Anderson et al., 2015; Zinszer et al., 2015), none have considered encoding (synthesis of predicted neural activity patterns from stimulus-models).

Methodologically, the new similarity-encoding strategy is a natural development to the Representational Similarity Analysis (RSA) framework (Kriegeskorte et al., 2008a,b; Kriegeskorte and Kievit, 2013; Nili et al., 2014), building on theories that visual-object categories are partially represented in terms of similarities in the brain (Edelman, 1998, Edelman et al., 1998) and (as we will return to in the Discussion) follows a computational architecture reminiscent of distributed associative memory neural networks (e.g. Willshaw et al., 1969). RSA takes a matching set of stimulus-feature-vectors and neural activity patterns and measures the degree of association between the stimulus models and neural modalities by (1) inter-correlating all pairs of stimulus-feature-vectors to produce a square model-correlation matrix; (2) likewise inter-correlating all pairs of neural activity patterns to produce an equivalent square neural-correlation matrix. (3) Quantifying the association between the model-correlation matrix and the neural-correlation matrix by extracting the lower below diagonal triangle (or upper) of unique pairwise comparisons from each matrix, vectorizing both to produce similarity-structure-vectors, and correlating model and neural-similarity-structure-vectors to quantify the association. By vectorizing the similarity-structure, conventional RSA treats an entire data set holistically. This strategy has proved extremely successful e.g. in interpreting pictorially induced representations in the brain, as in Kriegeskorte et al. (2008a,b) and Connolly et al. (2012), and demonstrating that the semantic structure embedded within neural activity patterns associated with comprehending concrete nouns matches sets of semantic models of those nouns (e.g. Bruffaerts et al., 2013; Carlson et al., 2014; Anderson et al., 2013, 2015). However this holistic comparison does not allow synthesis of predicted voxel-space activation patterns, and it is here that our approach introduces new capabilities.

As opposed to manipulating the representational similarity-structure holistically, we use inter-correlations between stimulus-model feature-vectors as a secondary code to represent stimuli. Therefore under our approach a stimulus is modeled with two codes, the first is the standard stimulus-model feature-vector, the second – the similarity-code – is a vector of correlations with other stimulus-model feature-vectors. The similarity-code thus defines the similarity between one stimulus and other stimuli and adheres to theories that consider similarities to underpin object categories in the brain (Edelman, 1998; Edelman et al., 1998).

Encoding – the synthesis of a predicted neural activity pattern – is achieved by: taking a new stimulus-model feature-vector for which we would like to predict the associated neural activity; generating a new similarity-code for that stimulus-model feature-vector; transferring that similarity code to a matching data set of stored neural



Decoding, by matching neural similarity onto semantic similarity

For visual clarity, the decoding method is illustrated using 8x8 matrices, rather than the full 60x60 matrices that were actually used.
The true labels of the stimuli are represented by the numbers 1 to 8.

Pick a pair of stimuli to be decoded, e.g. 3 and 6. Extract their neural and semantic similarity vectors from the respective matrices.

Remove the elements corresponding to the two test stimuli themselves from the neural and semantic vectors, so that the resulting reduced vectors contain no information about the similarity of the two test stimuli either to themselves or to each other.

Remove the true-labels from the neural vectors.The decoding's task will be to choose between one of two possible labelings: (A=3, B=6) or (A=6, B=3)

A
B
corr(A,3)
corr(A,6)
corr(B,3)
corr(B,6)
3
6
Neural vectors, with unknown stimulus labels
Semantic vectors, with known stimulus labels

Decoding: assign labelings to the two unknown-label neural vectors by computing their degree of match with the two known-label semantic vectors. The degree of match is simply the correlation between the vectors.

Repeat the above steps for all possible stimulus pairs.

Fig. 2. Visualization of the new leave-2-out similarity-decoding algorithm.

activity patterns; synthesizing the new predicted neural activity pattern by applying the code as weights in a superposition of the stored neural activity patterns.

In other stimulus-model-based-encoding approaches the encoding is encapsulated in a fixed mapping from stimulus-model feature-vectors to voxel-activity learnt in regression. In predicting novel neural activity patterns the stimulus-model features mapped into neural-activity-pattern-space are the basis functions and feature-vector-values are the weights applied to basis functions. In our new similarity-encoding approach, the neural activity patterns are basis functions, and the similarity-code derived from the stimulus-feature-vectors defines the weights. No mapping between stimulus-model feature-vectors and neural activity patterns needs to be learnt to synthesize a predicted neural activity pattern. All that is necessary is the similarity-code. This makes the similarity-encoding approach low cost because there is no need to fit a model, and robust because it is parameter free.

Given novel neural activity patterns without labels and labeled stimulus-model feature-vectors (on top of a stored set of different stimulus-model feature-vectors matched with neural activity patterns), we go on to demonstrate how building neural-similarity-codes and model-similarity-codes allows us to match neural-similarities to model-similarities and thus assign labels to the neural-similarity-codes to decode the neural activity patterns without the encoding phase.

We demonstrate our approach in a reanalysis of Mitchell et al. (2008) fMRI data set of neural activity elicited as participants viewed line drawings of objects presented alongside their names. Mitchell et al. (2008) built stimulus-model feature-vectors from which they synthesized predicted neural activity patterns associated with the objects using multiple-regression. We show how similarity-codes estimated using Mitchell et al. (2008)'s original semantic-models and state of the art computational-semantic-models (Baroni et al., 2014) can be easily applied to predict and decode neural activity patterns without model fitting. This process capitalizes both on RSA's power in relating high-dimensional data with few exemplars across modalities, and in these cases of high dimensional data is simple and fast because it does not involve training a mapping between semantic features and voxels (and setting/tuning learning parameters), it therefore sidesteps problems associated with overfitting. Because similarity-codes can be computed in a piecemeal fashion, it means that the approach is flexible to the acquisition of new training data, unlike regression where a model must be refit to the new data. Given the connections drawn between object similarities and object representation in the brain (Edelman, 1998; Edelman et al., 1998), and the power and simplicity of our approach, we close in the Discussion by considering the possible implications of similarity-encoding for knowledge representation in the brain.

Methods

Brief summary of the Mitchell et al. (2008) methods

We reanalyze Mitchell et al. (2008) fMRI data, available at http://www.cs.cmu.edu/~tom/science2008. Mitchell et al. scanned nine right-handed adult participants (5 female, age between 18 and 32) as they were presented with stimuli showing a particular concrete-object noun and also a picture of that object. The participants' task was to think about the properties of the object. There were 60 nouns in all, five each from twelve different classes, such as animals, furniture, tools and vehicles. Each noun was presented six times to each participant, in a randomly interleaved order. Beyond a few minor differences in preprocessing identified below, the fMRI data taken to analysis were the same as in Mitchell et al. (2008).

Scanning protocol and pre-processing

Mitchell et al. (2008) acquired functional images on a Siemens Allegra 3.0 T scanner using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 30 ms and a 60° angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1-mm between slices. The acquisition matrix was 64 × 64 with 3.125 × 3.125 × 5-mm voxels. They subsequently corrected data for slice timing, motion, linear trend, and performed temporal smoothing with a high-pass filter at 190 s cut-off. The data were spatially normalized to the MNI template brain image, and resampled to 3 × 3 × 6 mm^3 voxels. The voxel-wise percent signal change relative to the fixation condition was computed for each object presentation.

The voxelwise mean of the four volumes acquired 4 s after stimulus presentation was used to represent that noun presentation. To create a single representation per noun per participant, we took the voxel-wise mean of all six presentations of each word. We normalized voxel activity by transforming the nouns'values per voxel to z-scores (Mitchell et al. who predicted each voxel individually did not perform this standardization). Voxels estimated to have good signal were selected using the same criteria as Mitchell et al. (2008), who picked the 500 voxels with the most stable activation profile over words, with profiles compared across sessions: Pearson's correlation of each voxel's activity between matched word lists in all scanning session pairs (15 unique session pairs giving 15 correlation coefficients of voxel activity for the list of *training* words) was computed and the mean coefficient used as stability measure. The voxels with the 500 largest correlations were chosen. Voxel selection in the similarity-encoding analysis was conducted in a cross-validated fashion: In each test iteration, fMRI-activity patterns corresponding to two test words were held-out, and consequently voxel selection was based on the remaining 58/60 'training words' to ensure independence of training and test data. In similarity-decoding, where both target (fMRI similarity) and predictor (semantic-model-similarity) remain entirely separated (and are thus independent), voxel selection was conducted on all 60 fMRI-words. The difference between similarity-encoding versus decoding is detailed as the analyses are described in the Results.

Stimulus-model feature-vectors

In analysis we used Mitchell et al.'s original set of semantic-feature-vectors and also a set of state of the art vectors from computational linguistics. In building semantic models for the stimulus nouns, Mitchell et al. took inspiration from theories that sensorimotor features are important for representation and manually selected a set of 25 sensorimotor verbs (e.g. '*touch*', '*see*', and '*manipulate*'), and counted the co-occurrence frequencies of each of the 60 noun stimuli with each of the 25 verbs throughout Google's publicly available trillion-word corpus (called the Web 1T 5-gram, because co-occurrences were counted within a five-word window). This yielded a vector of 25 frequencies for each of the 60 nouns, each subsequently normalized to unit length. These models are referred to as *Mitchell-verbs*.

We sourced leading edge computational semantic models from Baroni et al. (2014) who compared a selection of state-of-the-art computational-semantic-models in a variety of benchmark tasks. For simplicity we focus on a model based on co-occurrence counts whose derivation follows much the same procedure as the Mitchell-verb semantic-models on a grander scale (despite new neural-network semantic models outperforming others in a number of the benchmarks, in preliminary tests that we do not report they did not afford a performance gain here). Baroni et al. built semantic models, subsequently referred to as Text-win2, by counting co-occurrences within a window of a fixed size of 2 to left and right of each target word in a corpus of about 2.8 billion tokens constructed by concatenating ukWaC, the English Wikipedia and the British National Corpus. The top 300K most frequent words (which included the 60 stimulus-nouns) in the combined-corpus were counted both as target and context elements. They transformed the co-occurrence matrix into nonnegative Pointwise Mutual Information and reduced it by Singular Value Decomposition to 500 dimensions.

Results

Similarity-based encoding: synthesizing predicted neural activity patterns for novel words using stimulus-model-similarity-codes

Neural activity associated with a novel word's meaning is predicted by a process of first coding the new word as a semantic-model similarity-code — a vector of model similarities to other words (calculated by inter-correlating semantic-model feature-vectors) and using these as weights in an average of respective fMRI words. This process is illustrated in Fig. 1, where stage #1 displays the stored 'training' set of semantic-feature-vectors for nouns linked to matching recordings of neural activity (to the right). Note that the more task specific term semantic-model feature-vector is used in place of stimulus-model feature-vector in the following text. We also have an extra semantic-feature-vector for a new word that we would like to predict the neural-activity-pattern for, displayed to the left. In stage #2 the semantic-model similarity-code is estimated by correlating the semantic-model feature-vector of the new word with all of the stored semantic-model feature-vectors using Pearson's correlation. The semantic-model similarity-code is therefore a vector of similarity values in the range [−1 1] that are tied to each noun we have neural coverage for. We synthesize the predicted neural activity pattern of the new word in stage #3 by simply transferring the semantic-model similarity-code across modalities to serve as weights in a similarity-weighted average of the corresponding neural activity patterns for respective words. In this case neural activity patterns are stored as long column vectors of voxel activities, so the weighted average firstly involves scaling each noun's fMRI-vector with the corresponding similarity-code value, and then summing the weighted fMRI-vectors (voxelwise). The summed vector is then normalized by dividing by the sum of the similarity-codes, with them first being converted into absolute values, as is standard for normalization quotients. This can be expressed formally as:

$$\vec{b}'_{N+1} = \frac{1}{C} \cdot \sum_{i=1}^{N} \vec{b}_i \cdot corr\left(\vec{s}_{N+1}, \vec{s}_i\right)$$

$$C = \sum_{i=1}^{N} \left| corr\left(\vec{s}_{N+1}, \vec{s}_i\right) \right|$$

where there are *N* words for which we have stored neural activity patterns, each neural-activity-pattern is stored in a vector *b* that is linked to a semantic-feature-vector (*s*). The new word we would like to predict is indexed $N+1$ and b'_{N+1} is the synthesized predicted neural-activity-pattern for the new word *s*. The normalizing constant *C* is the sum of absolute correlation values in the semantic-model-similarity-code for the new word. The relative magnitude of *C* potentially serves as a measure of confidence in the prediction made, following the intuition that if the new word's meaning is not similar to any other stored words, then the prediction is liable to be weak (and vice versa, if the new word is similar to known nouns, predictive power is liable to be strong). However with the current data set of concrete nouns that are equally distributed amongst semantic classes (i.e. there are no extreme semantic outliers) this measure was not found to be revealing in unreported analyses.

Visualization of predictions of neural activity made across all voxels in the brain for the nouns 'celery' and 'airplane' using similarities with the other 58 nouns computed from Mitchell et al.'s sensorimotor verb semantic models are in Fig. 3 (to match the visualization in Mitchell et al., 2008). A second visualization that displays each of the Mitchell-verb semantic-feature-vectors, auto-reconstructed using semantic-model-similarity-codes calculated by correlating that noun's semantic-feature-vector and each of the other 59 nouns, and then using this to weight an average of the semantic-feature-vectors for the other 59 nouns is in Fig. 4. Inspecting Fig. 4 reveals that prominent features in the pattern are reconstructed however this process shifts the measurement scale which although inconsequential for the subsequent correlation based analyses we undertake, may be undesirable in different circumstances and we identify ways to ameliorate the effect of this in the Discussion.

Quantitative evaluation of the synthesized predicted-neural-activity patterns can be undertaken using the leave-2-out pairwise matching strategy introduced by Mitchell et al. (2008): Two words at a time are selected for testing; semantic-model-similarity-codes for each of these two words are created by correlating each of their semantic-feature-

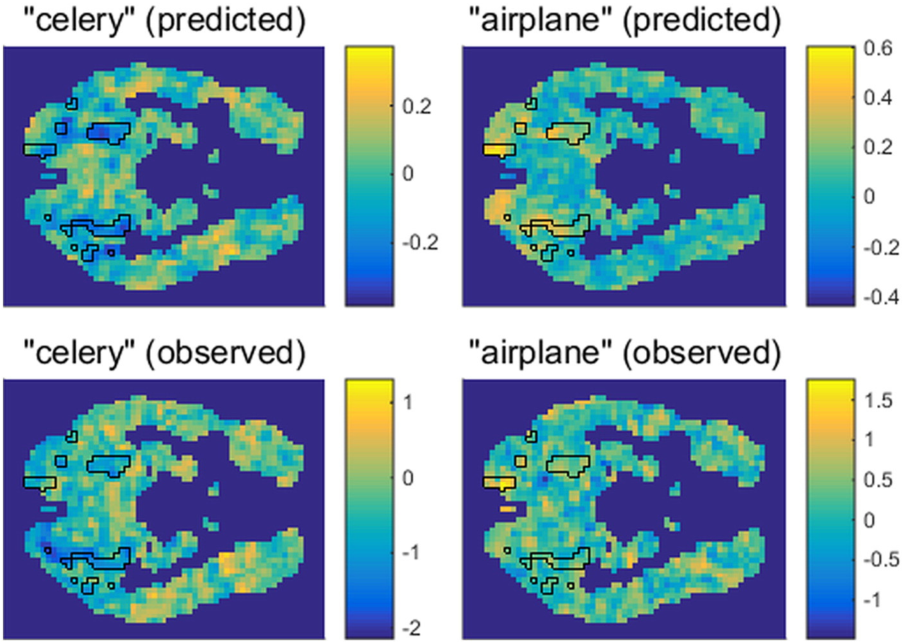


Fig. 3. Neural activity patterns predicted using the similarity-encoding method (Fig. 1) with the Mitchell-verb-semantic models as compared to observed neural activity patterns. The chosen words "celery" and "airplane" and slice $z = -12$ mm (MNI coordinates), match those displayed in Mitchell et al. (2008). The most stable voxels identified in voxel selection are bounded by black boxes.

vectors with those of the remaining 58 words; neural activity for each of the two held out-words is predicted by applying the words' semantic-model-similarity-codes to weight the superposition of neural activity patterns for the 58 other words; prediction accuracy is evaluated by correlating the predicted-neural activity patterns for the two words with the observed-neural activity patterns (which gives four correlation values), if the sum of correlations corresponding to the correctly matched predicted/observed pair exceeds the sum for the incongruent pair, decoding is a success, otherwise a failure. This process is repeated for all possible word pairs, with the mean accuracy giving a metric of success for each participant.

Significance was estimated empirically by permutation testing. Word-labels across both stimulus-models and fMRI-data were held fixed, and remained correctly assigned to the fMRI-vectors. Leave-2-out cross-validated voxel selection was repeated for all 1770 unique word-pairs, to produce 1770 different lists of selected voxels (note that if the word-labels were shuffled this would result in the same 1770 selected voxel-lists, just in a different order). The connection between word-labels and the semantic-model-vectors was then systematically jumbled to simulate random assignment of word-labels to semantic-content as follows.

To expedite computation a semantic-model correlation matrix was created (containing a stack of similarity-vectors). A vector of word indices was created [1,2, …, 60] and randomly shuffled. The shuffled indices were applied to reorder both rows and also columns of the semantic-model correlation matrix. This meant that word-labels were now misaligned to the semantic-model similarity-vector contents. An entire leave-2-out encoding analysis was rerun, drawing pairs of similarity-vectors, now mismatched to word-labels, from the correlation matrix, deleting entries for the two test words in the semantic-model similarity-codes such that the vectors contain 58 correlations) and using these to synthesize predicted fMRI-activity from the 58 word-label matched fMRI-vectors. This resulted in a list of 1770 decoding scores, each corresponding to a unique pair of word-labels. This list of scores was averaged to give a summary statistic of accuracy. Repeating this shuffling process 10,000 times allowed us to generate a null distribution of mean accuracies arising when the assignment of semantic-content to word-labels is governed by chance. Taking the proportion of times the mean decoding accuracy arising from randomly shuffled semantic-model correlation matrices was greater than the actual accuracy with the unshuffled semantic-model correlation matrix gave a permutation p-value.

Results for all nine participants (P1–P9) using similarity-encoding are displayed in Fig. 5. For comparison decoding accuracies from Mitchell et al. (2008) original analysis are displayed in light blue (Mitchell-verb-regress). We also display the outcome of a rerun of the analysis using standard multiple regression on the semantically richer and larger Text-win2 vectors (Text-win2-regress) displayed in dark blue. Here it is obvious that decoding accuracy is compromised by regression overfitting the Text-win2-vectors. Using similarity-encoding with the Mitchell-verb semantic-feature-vectors (orange), all results are significant ($p < .01$), and accuracy across participants is mean +/− sd = 76% +/− 4% correct (where 50% is chance-level).

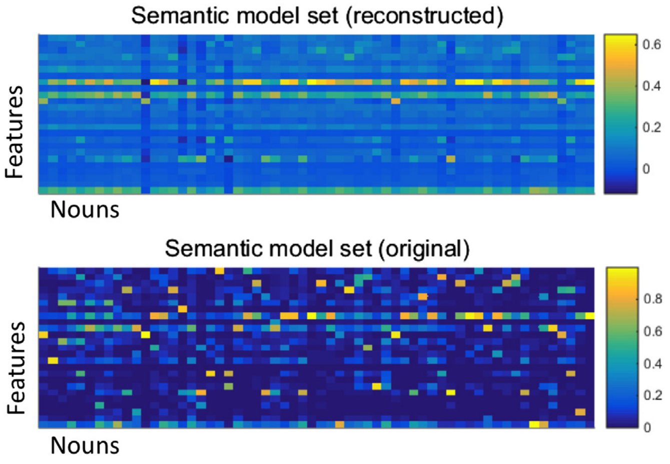


Fig. 4. Auto-reconstruction of each Mitchell-verb semantic-feature-vector based on its similarities to the other semantic-feature-vectors. Features are in rows and nouns are in columns (row/column names are not displayed to avoid cluttering the diagram).

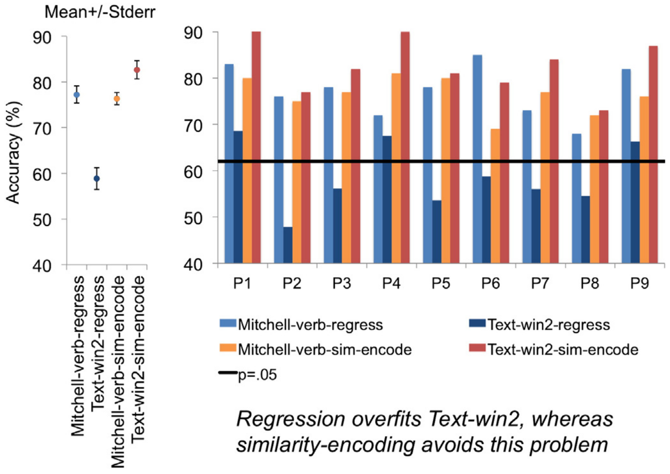


Fig. 5. Comparison of leave-2-out decoding accuracies using the Mitchell-verb and Text-win2 semantic models with regression and similarity-encoding (as per Fig. 1). The left most plot is mean +/− std-error accuracy across the 9 participants, the right plot displays accuracy per participant.

These are equivalent (signed rank = 22, p = .98) to the accuracies reported by Mitchell et al. (2008) which were mean +/− sd 77% +/− 6%. In contrast to regression which overfit Text-win2, mean accuracies using similarity-encoding with the Text-win2 model (dark red) had a mean +/− sd accuracy of 83% +/− 6% and were significantly higher (signed-rank = 45, p = .0039) than accuracies using similarity-encoding with the Mitchell-verb model. The similarity-encoding approach gave unanimously higher accuracy for all participants on Text-win2 than regression (signed-rank = 45, p = .0039).

Similarity-based decoding: decoding novel neural activity patterns by matching neural-similarity-codes to semantic-model-similarity-codes

We demonstrate how decoding can be entirely abstracted to representational similarity space, without synthesizing the predicted neural activity patterns beforehand. Here the problem differs because we require both the new but unlabeled neural activity patterns for the new words as well as labeled semantic-feature-vectors. The task is to estimate the labels for the neural data. This contrasts with similarity-encoding when only the semantic-feature-vectors were available and neural activity patterns for the new words were synthesized (this difference is illustrated in Fig. 1).

The algorithm we present for representational-similarity-decoding in overview operates as follows. First, both semantic-model-correlation matrices and neural-activity-correlation matrices (i.e. stacks of similarity-code-vectors) are calculated. Then, two words are chosen to serve as the test-stimuli to be decoded. The word-labels for those test stimuli are obtained by matching neural-activity-similarity-codes onto semantic-model-similarity-codes. This is repeated for all possible pairs of words. The algorithm is visualized in detail in Fig. 2.

For computational efficiency the first step is to calculate representational-similarity matrices for all 60 words for both semantic-models and neural activity patterns. However, if labels for two of the neural activity patterns are unknown, we only know the correspondence between neural activity patterns and semantic-feature-vectors for 58 nouns, so we will need to delete two entries in the 60 $*$ 60 correlation matrices to simulate this situation. For the leave-two-out testing, all possible pairs of words are chosen, in turn, to serve as the test-stimuli to be decoded. The neural-similarity-code-vectors for the two test words are extracted from the 60 $*$ 60 neural-similarity-matrix, and the words' semantic-model-similarity-code-vectors are correspondingly extracted from the semantic-model-similarity-matrix. The elements corresponding to the two test stimuli themselves are removed from the neural and semantic similarity-code-vectors (to simulate the case that labels are unknown) and the resulting reduced vectors contain no information about the similarity of the two test stimuli either to themselves or to each other.

With those reduced neural and semantic-similarity-code-vectors in hand, the actual decoding can now be performed. The decoding proceeds by matching neural-similarity-codes onto semantic-model-similarity-codes: the true labels of the two test-words' neural-similarity-vectors are unknown to the decoder, and the decoder's task is to assign labels to the neural-similarity-vectors (and hence the associated neural activity patterns) by choosing a labeling that produces the best match to the semantic-model-similarity-codes, whose labels are known. In the example illustrated in Fig. 2, the two test stimuli are the words 3 and 6 out of the set of 60. The actual labels get removed from the neural-similarity-vectors, so that they now have the unknown labels A and B. One possible labeling is (A = 3, B = 6), and the other possible labeling is (A = 6, B = 3). The decoding proceeds simply by calculating the correlations between the neural-similarity-vectors A and B and the semantic-similarity-vectors 3 and 6, and picking the labeling corresponding to the highest correlations.

The null hypothesis tested for similarity-decoding is that there will be no relationship between semantic-model similarity-vectors and neural-similarity-vectors, and importantly this is different to similarity-encoding (that there will be no relationship between observed-fMRI activity and model-predicted-fMRI activity). Differently for a similarity-decoding analysis it is essential that the semantic-model correlation matrix is strictly independent from the fMRI correlation matrix (for similarity-encoding it is essential that the held-out target fMRI activity is strictly independent from the predicted fMRI activity). The practical impact of this difference is in voxel selection. For similarity-decoding voxel selection can be undertaken a single time on the entire set of 60 fMRI-words because this has no knock on effect on the semantic-model correlation matrix. In contrast for similarity-encoding the two test fMRI-words need to be held out from voxel selection to avoid them contaminating the prediction process. Permutation testing with randomly shuffled data (following the procedure detailed in the following paragraph), empirically confirms that the null distribution is centered on 50% when voxel selection is conducted a single time on each participant before the similarity-decoding analysis. For each of the 9 participants, 10,000 permutations with shuffled data were run, This resulted in mean +/− sd accuracy across participants of 50.01 +/− .07, that was not significantly different to 50% p = .58, t = .58, df = 8, 2-tail.

The statistical significance of the accuracies achieved using similarity-decoding was empirically tested via permutation testing. Rows and columns of the semantic-model similarity matrix were shuffled, relative to the row and column word-labels (as described for similarity-encoding), whilst the fMRI similarity matrix was held fixed. Evaluation compared pairs of shuffled semantic-model similarity-vectors to the observed fMRI-similarity-vectors. Repeating this shuffling process 10,000 times generated a null distribution of mean accuracies arising when the assignment of semantic-similarity-vectors to word-labels is governed by chance. Taking the proportion of times the mean decoding accuracy arising from randomly shuffled semantic-model correlation matrices was greater than the actual accuracy achieved with the unshuffled semantic-model correlation matrix yielded the permutation p-value.

Leave-two-out decoding accuracies using the Mitchell-verb models are on average slightly improved over the previous similarity-encoding approach at mean +/− sd of 78% +/− 4% but this difference is not statistically significant (signed rank = 30, p = .13, 2-tail), where mean accuracy across participants is 78% +/− 4%. Using the Text-win2 vectors accuracies were equivalent to the similarity-encoding strategy previously reported (signed-rank = 4, p = .5), mean 84% +/− 7. Individual's results in both tests are plotted in Fig. 6.

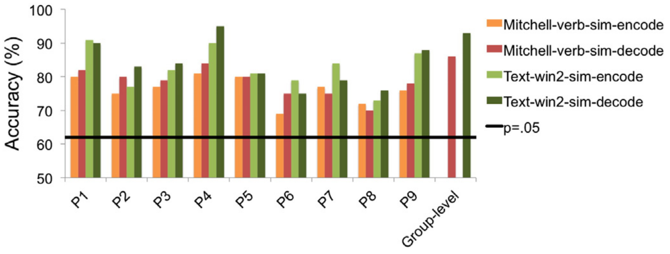


Fig. 6. Comparison of leave-2-out accuracies achieved with the similarity-encoding (see Fig. 1) and similarity-decoding (Fig. 2) approaches using the Mitchell-verb and Text-win2 semantic models.

Group-level decoding: decoding group-level-neural-similarity-codes with semantic-model-similarity-codes

The computational semantic models used are population-level linguistic models in the sense that they accumulate word co-occurrence statistics across documents written by many different authors. Group-level neural-similarity-codes can easily be estimated by taking the mean of individuals' similarity-codes, which is beneficial only if there are group-level commonalities in representational similarity, in which case averaging will serve to cancel out noise in individual-level data. Caution should be taken in interpreting group-level results, as significance tests are testing an inference on those particular participants (treating them as a fixed rather than random effect), meaning that results do not necessarily generalize to individuals randomly selected from the population. We apply the decoding algorithm (Fig. 2) to match the group-level-neural-similarity-codes onto the semantic-model-similarity-codes. This strategy is highly accurate. Using the leave-two-out similarity-decoding procedure, mean decoding accuracies of 86% and 93% are returned for the Mitchell-verb semantic vectors and Text-win2 respectively (both results are plotted in Fig. 6).

As a final test we examined which words were best and worst discriminated for the different models in the test using group-level neural similarities to see whether there were any obvious patterns. Discrimination accuracies for the 60 words (maximum score per word of 59) were significantly correlated across the two semantic models (Spearman's $\rho = .3$, $p = .02$). For the Mitchell-verb models the best five discriminated words were: car (59); door (59); glass (59); horse (58); airplane (58), and the worst were: key (32); saw (36); table (36); lettuce (36); hand (37). For Text-win2 the best five were: airplane (59); barn (59); house (59); glass (59); bicycle (58) whereas the worst were: saw (40); corn (45); eye (46); chisel (48); igloo (48). We observe that “saw”, ”key” and “table” which were all weakly distinguished have distinct senses of meaning (e.g. “saw” as the tool and vision related verb), however “lettuce”, “eye” and “igloo” do not, as such it is difficult to discern whether there is any systematic pattern.

Discussion

The key contribution of this paper is to unite the two previously disparate methods of encoding models and RSA, capturing the strengths of both, and enabling similarity-based synthesis of predicted fMRI patterns. Our new similarity-encoding method quickly and accurately predicts the neural representation of concrete-nouns based on using computational-semantic-models to measure how similar those nouns' meaning is to other nouns, and we have observed it to robustly scale to situations when there are many more stimulus-model features than stimulus-models. We have also demonstrated how re-representing both semantic-models for words and neural activity patterns for words as similarity-code-vectors allows semantic-model and neural-similarity-vectors to be matched to each other, and therefore neural activity patterns to be matched to computational-semantic-models (and decoded). We discuss: how our results and approach compare to previous results that have used semantic-model-based-encoding methods to decode the same brain data and then identify practical, architectural and computational differences to regression-based approaches. Finally, motivated by theories that consider similarities to be central to object representation in the brain (Edelman, 1998; Edelman et al., 1998), we close by identifying connections between our new similarity-encoding approach and analyses to the existing literature. This includes the potential relevance of similarity-codes to the organization of thematic and taxonomic knowledge in the brain, and the overlap in architecture between our approach and biologically plausible artificial neural networks.

Comparison of decoding accuracy to other approaches

Representational Similarity based decoding achieved equivalent accuracy (mean 78%) to Mitchell et al. (2008) original regression-based encoding analysis (mean 77%) using the same semantic-models, and a group-level accuracy of 86% (group-level decoding of this data has not previously been attempted). In subsequent work, Mitchell and colleagues (e.g. Palatucci et al., 2009; Murphy et al., 2012) and other research groups (e.g. Devereux et al., 2010; Jelodar et al., 2010; Pereira et al., 2013; Levy and Bullinaria, 2012; Akama et al., 2015) have explored using different semantic models to decode the same neural data. Successful models have tended to incorporate more semantic features e.g. Palatucci et al. (2009) who achieved a mean accuracy of 81% by using a human-generated set of 218 semantic features, and by Levy and Bullinaria (2012), who achieved a mean accuracy of 85% correct using 10,000 semantic features and tuned learning parameters in regularized regression. Our results using contemporary computational-semantic-models (with 500 features), resulted in a mean accuracy of 84% (group-level 93%) which is at least competitive with the previous approaches, without need to fit a model or tune optimization parameters.

Practical differences to other approaches

The similarity-based approach has advantages in its simplicity, as there is no model that needs to be fit an analysis can be run at high speed, and that it fits within the conventional RSA framework. The flipside of these benefits, is that unlike regression, because there is no mapping between semantic features and individual voxels the similarity method does not predict how specific voxels contribute to semantic representation (e.g. as valuable to test whether regions active in color/motion/acoustic perception are recruited in representing color/motion/acoustic related concepts Mitchell et al., 2008; Fernandino et al., 2015), however this could be compensated for using searchlight analyses to confine analyses to local brain regions (Kriegeskorte et al., 2006). Also without modifying the similarity-encoding algorithm presented here, the synthesized neural activity patterns predicted are constrained to be interpolations between the existing store of neural activity patterns. Extrapolation outside the space spanned by the stored patterns would require introducing non-linear scaling of the stored patterns.

A potentially undesirable consequence of the weighted average in similarity-encoding is that it is prone to shift the scale of the encoded vectors (as observed in Figs. 3 and 4). In the analyses reported here this was not an issue, and we opted to present the technique using Pearson's correlation because it is commonplace in the Representational Similarity Analysis literature and parameter free. However in cases where the shift is problematic, it may be possible to ameliorate problems using alternative similarity metrics that have tunable parameters. One alternative is Gaussian similarity:

$$\text{sim} = \exp\left(-\frac{1}{2}\cdot\frac{d^2}{\sigma^2}\right)$$

where d is the Euclidian distance and σ is a free parameter that can be appropriately tuned to the situation. As an example Fig. 7 replots Fig. 4 using Gaussian similarity and illustrates how modifying σ modulates the visual quality of the match to the original data.

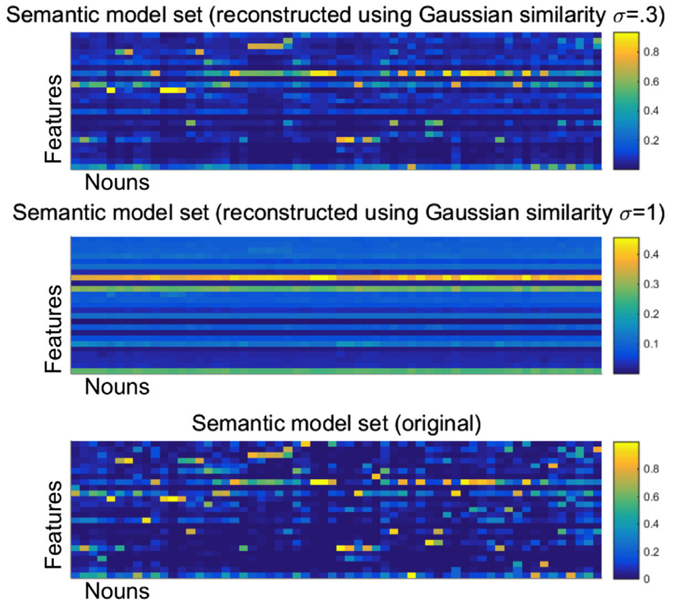
Differences in computational architecture to regression approaches

Regression approaches learn a mapping between each voxel and all semantic-model-features. The interface between semantic-model-features and all voxels is therefore a number-of-features $*$ number-of-voxels matrix of weights ($25 * 500 = 12{,}500$ using Mitchell et al.'s semantic-models) that needs to be learnt. Architecturally the similarity approach differs because semantic features and voxels are not directly linked to each other, but instead the link between model and fMRI data is between similarity-codes, where similarity-codes are vectors of correlations that are number-of-words long.

What are the computational differences between similarity-encoding and standard regression?

The most notable computational difference between similarity-encoding and regression is that the similarity-encoding contains zero tunable parameters, and thus there is no process of fitting weights. Therefore no values need to be adjusted in order to reduce any model-fit error. This is in contrast to regression, in which each regression weight (sometimes called a beta coefficient) is a free parameter which must be tuned in order to reduce the overall sum-of-squares error.

A useful distinction to draw here is between calculating and fitting. If one is asked to add up ten numbers, then a calculation is performed which involves ten operations. However, this is very different from the fitting of a model with ten free parameters when there may be many candidate solutions. When adding up ten numbers, the calculation can only proceed in one way. Likewise the correlation between feature vectors used in similarity-encoding is a direct calculation and overfitting is impossible because there are no free parameters to fit. To recap each weight i of a similarity-vector is calculated as $w_i = \text{corr}(s_{N+1}, s_i)$, where s_{N+1} is the semantic-model vector for the word to be estimated and s_i is a semantic-model vector for a stored word. Our similarity-decoding approach constructs a $60 * 60$ matrix consisting of the correlations between the words' semantic vectors. However, the number of free parameters in that $60 * 60$ matrix is also zero. In contrast, a regression based encoding model has a number of free parameters, equal to the number of semantic features multiplied by the number of voxels.


Semantic model set (reconstructed using Gaussian similarity σ=.3)

Semantic model set (reconstructed using Gaussian similarity σ=1)

Semantic model set (original)

Fig. 7. Auto reconstruction of each Mitchell-verb feature-vector using a Gaussian similarity metric to calculate the similarity to the other feature-vectors. The top row corresponds to reconstruction with a smaller value of sigma (.3), which produces a visibly better match to the original vectors (bottom row) than sigma = 1 (middle row).

Another difference which follows from the above, is that for the similarity-encoding approach, individual similarity-vectors, and synthesis of predicted neural activity patterns, can be computed one by one. For regression the entire regression model needs to be fit before a single neural activity pattern can be synthesized. This also entails that as new data becomes available to train on, the entire regression needs to be refit incorporating the new training word(s) into the calculation. However once the regression mapping has been learnt, the store of feature-vectors and fMRI data are redundant (and unlike the similarity-encoding which requires the 'training data' to be permanently stored, they can be erased).

A third difference is that the similarity-encoding approach computes weights (the similarity-vector) using inter-correlations between the feature-vectors and therefore the process is entirely disconnected from the voxel activities in the fMRI data. As it turns out, this disconnection between the model and the fMRI data is actually helpful here as can be seen from the fact that our similarity encoding model performs with higher accuracy than regression when using the rich semantic vectors of TextWin2, as shown in Fig. 5. In contrast for regression, the covariation between each individual-feature-value across words and each individual-voxel's-activity across words is measured and contributes towards the resulting weights that map between features and voxels.

Is there a similarity-metric that could make multiple regression and similarity-encoding identical?

Without a radical reconfiguration of the similarity-encoding architecture there is not a similarity metric that would make similarity and regression based approaches identical. The simple reason for this is that the computation of weights for the similarity-encoding approach is entirely disconnected from the voxel activities in the fMRI data (as per the previous section). For the two approaches to be equivalent there would need to be some mixing of semantic model and fMRI data in weight calculation.

Does the brain use similarity-encoding in semantic memory?

Having observed how similarity-encoding can efficiently generate robust and accurate predictions online (without having to train a model) and given theories that object categories are represented in the brain in terms of similarities (Edelman, 1998; Edelman et al., 1998) it naturally follows to consider if similarities play a part biologically as an encoding strategy in semantic memory. We close by considering the function that similarity-encoding could have biologically, and drawing connections to empirical studies of knowledge representation and computational aspects of biologically plausible artificial neural network architectures.

Everyday experience tells us that meaning can be rapidly assigned to words that have never been experienced (e.g. unicorns). It is also intuitive that meaning can be assigned to a new word-label by knowing which other word-labels are similar and dissimilar in meaning to the new word-label (this information is found in a thesaurus). In the context of this analysis the similarity-code could be considered as a variant of the type of information extracted from a thesaurus, and the encoding – the synthesis of predicted patterns of meaning – is analogous to synthesizing a memory trace grounded in a prediction of what the experience would have been like. In this case the similarity-code is sourced externally (e.g. from a thesaurus), however an alternative scenario is where the similarity-code is derived internally, as could play a role in cross-modal pattern synthesis. A completely new item is sensed only in one modality, and a prediction of that item's features is synthesized in a second modality based on merging experience with items that were judged to be similar in the first modality. For instance the taste of a berry that has not previously been experienced, but is now seen, might be synthesized based on merging past experience of the tastes of other berries that look similar.

Relevant to the question of whether similarity-codes have a role in the representation of semantic knowledge is the distinction between thematically related knowledge - things that occur together in space and time (thus musicians, instruments and music are associated with one another despite being intrinsically different entities), and taxonomic category relationships based on similarity between category members' (where cats and tigers would be similar despite almost never occurring in the same context in the world). The distinction between thematic/taxonomic relations has been extensively studied behaviorally (e.g. Lin and Murphy, 2001) and there is conflicting evidence as to whether thematic/taxonomically organized knowledge representations can be systematically spatially dissociated on the neural substrate (e.g. Kalénine et al., 2009; Schwartz et al., 2011; Anderson et al., 2014; Jackson et al., 2015). Even though the difference between thematic and taxonomic knowledge is not always clear-cut, semantic-model-vectors based on word co-occurrence frequencies in large text corpora (such as Text-win2) will reliably accumulate both aspects of knowledge (e.g. that dogs co-occur with leash and bone / fur and mammal respectively). Similarity-codes derived from these same semantic-feature-vectors visibly distill taxonomic-category structure in Mitchell et al.'s selection of classes (as can be seen from the block-diagonal structure of the matrix in Fig. 8 where each bright block along the diagonal indicates a group of objects from the same category whose Text-win2 semantic similarities with each other are high). As such this hints that similarity-codes resulting from comparisons made between experience based concrete object representations in the brain could provide a route to the emergence of taxonomic category related activity patterns, and as we discuss next we might expect similarity-codes to be a common byproduct of parallel-distributed computation.

Computationally similarity-codes are a fundamental component of biologically plausible artificial neural network models including self organizing map neural networks (Kohonen, 1997) and correlation matrix

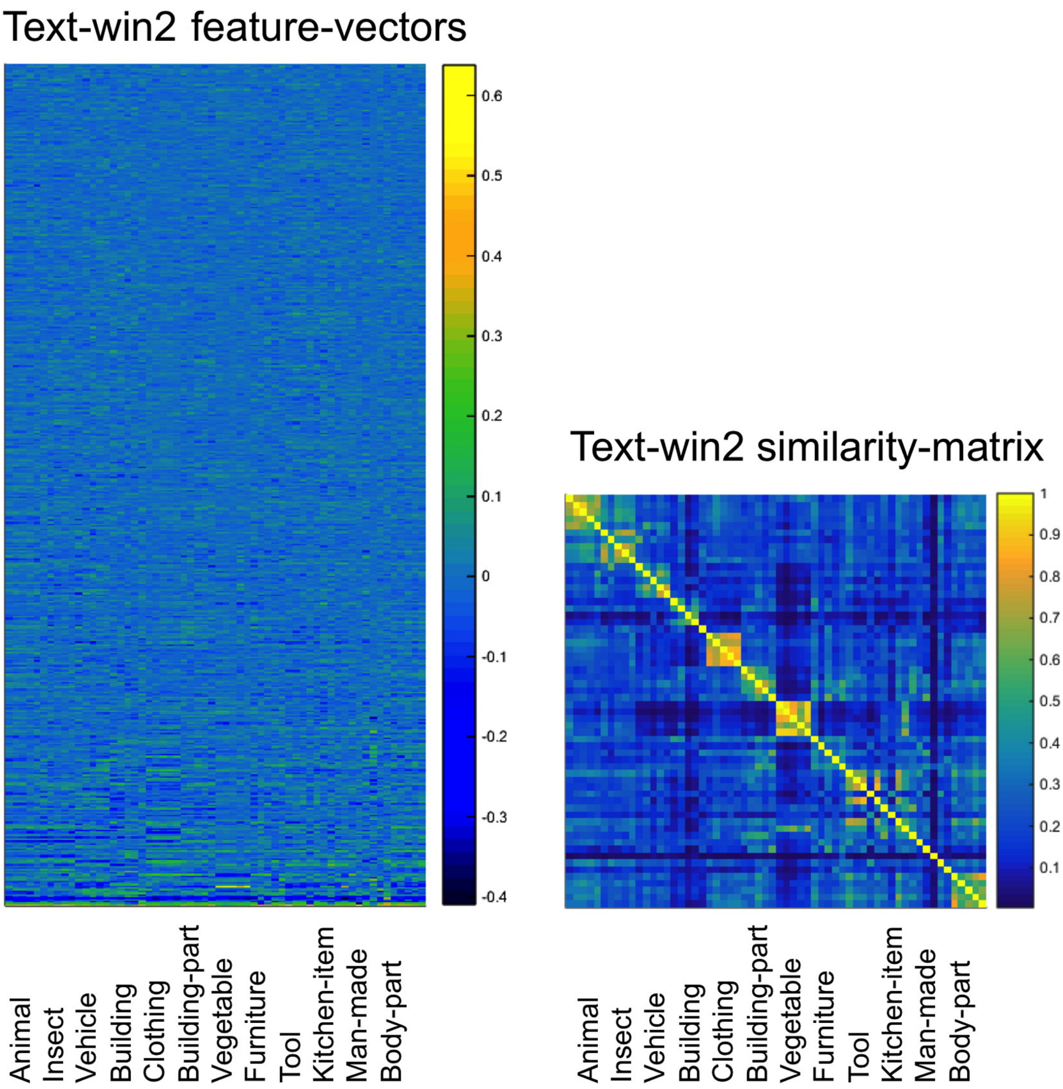


Fig. 8. Visualization of Text-win2 semantic models arranged according to Mitchell et al.'s manually selected classes and correlation-matrix that distills taxonomic-structure from this task domain (i.e. Mitchell et al.'s classes) as observed by the squares on the matrix diagonal (note that 'man-made items' in this context is not a well defined class in the sense that instances of this class could reasonably be assigned to other classes).

memory based neural networks introduced by Willshaw et al. (1969); Kohonen (1972); Austin and Stonham (1987). The procedure we have described is agnostic of topography and generating the similarity-code in stage #2 of Fig. 1 follows fundamentally the same procedure as the recall-phase of correlation matrix memories. More generally speaking we expect similarity-codes to emerge from any computational process that involves a parallel match of an input pattern to prototypical-template-patterns stored in memory (Edelman et al., 1998; Peelen et al., 2009 for evidence), where the strength of pattern match to each template is synonymous with a similarity measure. There is therefore good reason to expect similarity-codes to exist in the brain and therefore the brain has at least the potential to employ similarity-encoding.

Acknowledgments

Shimon Edelman and Russ Poldrack for comments on earlier versions of the manuscript, and Tom Mitchell for permission to use the dataset from his seminal 2008 Science paper. We thank our anonymous reviewers for their helpful comments that have improved the quality of this manuscript. This work was supported in part by NSF Award 1228261 (P.I. Raizada).

References

Akama, H., Miyake, M., Jung, J., Murphy, B., 2015. Using graph components derived from an associative concept dictionary to predict fMRI neural activation patterns that represent the meaning of nouns. PLoS ONE http://dx.doi.org/10.1371/journal.pone.0125725.

Anderson, A.J., Bruni, E., Lopopolo, A., Poesio, M., Baroni, M., 2015. Reading visually embodied meaning from the brain: visually grounded computational models decode visual-object mental imagery induced by written text. NeuroImage 120, 309–322.

Anderson, A.J., Murphy, B., Poesio, M., 2014. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. J. Cogn. Neurosci. 26 (3), 658–681.

Anderson, A.J., Bruni, E., Bordignon, U., Poesio, M., Baroni, M., 2013. Of words, eyes and brains: correlating image-based distributional semantic models with neural representations of concepts. Proceedings of EMNLP 2013 1960–1970 Association for Computational Linguistics, Seattle, WA.

Austin, J., Stonham, T.J., 1987. Distributed associative memory for use in scene analysis. Image Vis. Comput. 5 (4), 251–260.

Baroni, M., Dinu, G., Kruszewski, G., 2014. Don't count, predict! A systematic comparison of context-counting vs.context-predicting semantic vectors. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, USA.

Bruffaerts, R., Dupont, P., Peeters, R., De Deyne, S., Storms, G., Vandenberghe, R., 2013. Similarity of fMRI activity patterns in left perirhinal cortex reflects similarity between words. J. Neurosci. 33 (47), 18597–18607.

Carlson, T.A., Simmons, R.A., Kriegeskorte, N., Slevc, L.,R., 2014. The emergence of semantic meaning in the ventral temporal pathway. J. Cogn. Neurosci. 26 (1), 120–131.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., et al., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.

deCharms, R.C., 2008. Applications of real-time fMRI. Nat. Rev. Neurosci. 9 (9), 720–729.

Devereux, B., Kelly, C., Korhonen, A., 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In: Murphy, B., Chang, K.K., Korhonen, A. (Eds.), Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics. Association for Computational Linguistics, Los Angeles, USA, pp. 70–78.

Edelman, S., 1998. Representation is representation of similarities. Behav. Brain Sci. 21, 449–467.

Edelman, S., Grill-Spector, K., Kushnir, T., Malach, R., 1998. Towards direct visualization of the internal shape space by fMRI. Psychobiology 26, 309–321.

Fernandino, L., Humphries, C.J., Seidenberg, M.S., Gross, W.L., Conant, L.L., Binder, J.R., 2015. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. Neuropsycholigia http://dx.doi.org/10.1016/j.neuropsychologia.2015.04.009.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–436. http://dx.doi.org/10.1146/annurev-neuro-062012-170325.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Jackson, R.L., Hoffman, P., Pobric, G., Lambon Ralph, M.A., 2015. The nature and neural correlates of semantic association versus conceptual similarity. Cereb. Cortex 25 (11), 4319–4333.

Jelodar, A.B., Alizadeh, M., Khadivi, S., 2010. Wordnet based features for predicting brain activity associated with meanings of nouns. Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, W10, p. 0603.

Just, M.A., Cherkassky, V.L., Buchweitz, A., Keller, T.A., Mitchell, T.M., 2014. Identifying autism from neural representations of social interactions: neurocognitive markers of autism. PLoS ONE 9, e113879.

Kalénine, S., Peyrin, C., Pichat, C., Segebarth, C., Bonthoux, F., Baciu, M., 2009. The sensory-motor specificity of taxonomic and thematic conceptual relations: a behavioral and fMRI study. NeuroImage 44, 1152–1162.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355.

Kohonen, T., 1972. Correlation matrix memories. IEEE Trans. Comput. 21, 353–359.

Kohonen, T., 1997. Self-Organizing Maps. Springer, New York.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412. http://dx.doi.org/10.1016/j.tics.2013.06.007.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008a. Representational similarity analysis—connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2 (4), 1–28.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.

Levy, J.P., Bullinaria, J.A., 2012. Using enriched semantic representations in predictions of human brain activity. In: Davelaar, E.J. (Ed.), Connectionist Models of Neurocognition and Emergent Behavior: From Theory to Applications. World Scientific, Singapore, pp. 292–308.

Lin, E.L., Murphy, G.L., 2001. Thematic relations in adults' concepts. J. Exp. Psychol. Gen. 130, 3–28.

Matthews, P.M., Honey, G.D., Bullmore, E.T., 2006. Nat. Rev. Neurosci. 7, 732–744.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meaning of nouns. Science 320, 1191–1195.

Murphy, B., Talukdar, P., Mitchell, T., 2012. Selecting corpus-semantic models for neurolinguistic decoding. Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM), p. 114.

Naselaris, T., Prenger, R.J., Kay, K.,.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902–915.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56, 400–410.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10 (4), e1003553. http://dx.doi.org/10.1371/journal.pcbi.1003553.

Palatucci, M., Pomerleau, D., Hinton, G., Mitchell, T., 2009. Zero-shot learning with semantic output codes. Neural Inf. Process. Syst. 22, 1410–1418.

Peelen, M.V., Fei-fei, L., Kastner, S., 2009. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460, 94–97.

Pereira, F., Botvinick, M., Detre, G., 2013. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. Artif. Intell. 194, 240–252.

Raizada, R.D.S., Connolly, A.C., 2012. What makes different people's representations alike: neural similarity-space solves the problem of across-subject fMRI decoding. J. Cogn. Neurosci. 24 (4), 868–877.

Schwartz, M.F., Kimberg, D.Y., Walker, G.M., Brecher, A., Faseyitan, O.K., Dell, G.S., Mirman, D., Coslett, H.B., 2011. Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. Proc. Natl. Acad. Sci. U. S. A. 108, 8520–8524.

Sulzer, J., Haller, S., Scharnowski, F., Weiskopf, N., Birbaumer, N., Blefari, M.L., Bruehl, A.B., Cohen, L.G., deCharms, R.C., Gassert, R., Goebel, R., Herwig, U., LaConte, S., Linden, D., Luft, A., Seifritz, E., Sitaram, R., 2013. Real-time fMRI neurofeedback: progress and challenges. NeuroImage 76, 386–399.

Willshaw, D.J., Buneman, O.P., Longuet-Higgins, H.C., 1969. Non-holographic associative memory. Nature 222, 960–962. http://dx.doi.org/10.1038/222960a0.

Zinszer, B.D., Anderson, A.J., Kang, O., Wheatley, T., Raizada, R.D.S., 2015. You say potato, I say tudou: how speakers of different languages can share the same concept. Proceedings of the 37th Annual Conference of the Cognitive Science Society.