Decoding and encoding models reveal how the depth of processing shapes the brain representation of meaning

David Soto[1,2], Usman Ayub Sheikh[1], Ning Mei[1], and Roberto Santana[3]

[1] *Basque Center on Cognition, Brain and Language, San Sebastian, Spain*
[2] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
[3] *Computer Science and Artificial Intelligence Department, University of Basque Country, Spain*

November 4, 2019

Correspondence to: d.soto@bcbl.eu, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor 20009 San Sebastian

Abstract

How the brain representation of conceptual knowledge vary as a function of processing goals remains unclear. We hypothesized that the brain representation of semantic categories is shaped by the depth of processing. Participants were presented with visual words during functional MRI. During shallow processing, participants had to read the items. During deep processing, they had to mentally simulate the features associated with the words. Multivariate classification, informational connectivity and encoding models were used to reveal how the depth of processing determines the brain representation of word meaning. Decoding accuracy in putative substrates of the semantic network was enhanced when the depth processing was high, and the brain representations were more generalizable in semantic space relative to shallow processing contexts. This pattern was observed even in association areas in inferior frontal and parietal cortex. Deep information processing during mental simulation also increased the informational connectivity within key substrates of the semantic network. To further examine the properties of the words encoded in brain activity, we compared computer vision models - associated with the image referents of the words - and word embeddings. Computer vision models explained more variance of the brain responses across multiple areas of the semantic network. These results indicate that the brain representation of word meaning is highly malleable by the depth of processing imposed by the task, rely on access to visual representations and are indeed highly distributed, including prefrontal areas thought to be primarily implicated in semantic control rather than representation.

1 Introduction

Grounded models of semantic cognition propose that knowledge about the world is re-enacted in the same modality-specific brain systems that are involved in perceptual or action processes. For instance, the concept of "guitar" comprises the way it looks, how it is played, and the sound it makes. Damasio's convergence zone theory proposes that mental re-enactment is mediated by brain association areas that integrate information from modality-specific systems (Damasio, 1989). The perceptual symbols theory (Barsalou, 1999) proposes that co-activation patterns in sensory-motor substrates are critical. On this account, conceptual knowledge involves an agent's brain simulating the different properties of the object in question (e.g. shape, color, texture, sound, action) in a way that resembles how the information is encoded in sensorimotor systems during overt behaviour. According to perceptual symbols theory, this simulation process may occur automatically without the need of committing the information into a conscious working memory system (Barsalou, 2008).

The anterior temporal lobe has been proposed as a key semantic hub that supports integration of modality (sensory) specific information and the formation of invariant conceptual representations (Ralph, Jefferies, Patterson, & Rogers, 2017; Patterson, Nestor, & Rogers, 2007). A meta-analysis of neuroimaging studies showed that activity in the anterior temporal lobe in semantic tasks is independent of the sensory modality of the stimuli (Visser, Jefferies, & Ralph, 2010). Other neurocognitive models propose however that multiple semantic hubs may exist (Damasio, 1989). For instance, the inferior parietal (Price, Peelle, Bonner, Grossman, & Hamilton, 2016), and lateral and ventral temporal regions (Binder & Desai, 2011) have been implicated in this regard. FMRI-based decoding studies (Simanova, Hagoort, Oostenveld, & van Gerven, 2012; Shinkareva, Malave, Mason, Mitchell, & Just, 2011) and computational modeling work (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; T. M. Mitchell et al., 2008) have also found that the representational space of meaning is highly distributed in the human brain.

However, how the brain representational spaces of conceptual knowledge vary as a function of internal processing goals and strategic factors remains unclear. For instance, previous research has not determined the role of the depth of processing on the brain representation of concepts. In some studies participants are required to actively process the meaning of experimental stimuli (T. M. Mitchell et al., 2008; Shinkareva et al., 2011; Buchweitz, Shinkareva, Mason, Mitchell, & Just, 2012), while in other studies, partici-

pants are exposed to the information in a more passive manner or were just required to pay attention to the information (Simanova et al., 2012; Huth et al., 2016).

Here we manipulated the depth of processing by asking participants to either engage in mental simulation of word items or by merely reading the words. Functional MRI, multivariate pattern (MVP) analyses and computational models addressed how mental simulation shapes the brain representation of meaning. If mental simulation influences the brain representational spaces of meaning then (i) the level of decoding accuracy of word category in substrates of the semantic network (Binder, Desai, Graves, & Conant, 2009) should be superior during mental simulation relative to when the words are merely attended and read; (ii) the corresponding MVP of a concept ought to be more generalizable with respect to new examples of the same category (i.e. MVP classifiers trained with a subset of the words should predict better out-of-sample words that were not used for training) (iii) we reasoned that if explicit simulation is a key process in the construction of meaning then the level of informational connectivity between key substrates of semantic network (i.e. the temporal correlations in decoding accuracy between areas) ought to increase with mental simulation. Lastly, to examine the properties of the semantic representations in the brain, we tested different encoding models based on word embeddings and computer vision models fitted with the image referents of the words. If mental simulation of the concepts is primarily driven by the re-enactment of the sensory properties of the stimulus then we would expect that computer vision models outperform word embedding models in explaining brain responses during semantic processing tasks. We tested these hypotheses in a set of left-lateralised regions involved in visual word processing based on a previous meta-analysis (Binder et al., 2009).

2 Methods

2.1 Participants

Following informed consent, twenty-seven participants (20-33 years, mean age: 24±3, 10 males) took part in return of monetary compensation. The study conformed to the Declaration of Helsinki and was approved by the BCBL Research Ethics Board.

2.2 Experimental Task and Procedure

The experiment was programmed using Psychopy (Peirce, 2007). It was comprised of 8 fMRI runs. Each trial began with a fixation period of 250 ms followed by a blank screen of 500 ms (see Figure 1) and then by the target visual word which was displayed for 1 s. The word was randomly selected from a pool of 18 living and 18 non-living words. Stimuli were centrally presented in white against black background in uppercase Arial font. The offset of the word was followed by a blank screen of 4 s. During these 4 s, depending on session instructions (shallow or deep processing), the participants were supposed to either read the word (shallow processing condition), or mentally simulate the properties associated with the word (e.g. its shape, its color etcetera), henceforth, the deep processing condition. A red asterisk was centrally presented during the inter-trial interval and participants were instructed to relax and wait for the next trial. To ensure that the participants focused on the stimuli and the task, a maximum of two catch trials were set to appear at random points in each of the sessions. These catch trials showed number words from among ZERO, ONE, and THREE in place of usual living/non-living words, and participants were supposed to respond by pressing any one of the four buttons on the fMRI response pad. The total number of catch trials was kept equal across conditions. To maximize the separation between the brain response for each of the trials, the time for which the asterisk stayed on the screen was jittered between 6 to 8 s. The jitter was based on a pseudo-exponential distribution resulting in 50% of trials with the inter-trial interval of 6 s, 25% with 6.5 s, 12.5% with 7 s and so on.

Each scanning run was comprised of 36 trials, one for each of the words, randomly selected. Living words included animal names presented in uppercase, i.e. tigre, gallo, perro, oveja, cerdo, gorila, burro, yegua, ardilla, conejo, gallina, caballo, ballena, tortuga, pantera, camello, elefante, canguro. Non-living words were comprised of tool names, i.e. llave, lapiz, tijera, aguja, pinza, sierra, clavo, pincel, alicate, tuerca, navaja, cepillo, taladro, soplete, tornillo, cuchara, martillo, cuchillo. Word classes were matched in lenght.

2.3 MRI Data Preprocessing

The preprocessing of fMRI data was performed using FSL FEAT (FMRIB Software Library; v5.0). The first 9 volumes were discarded to ensure steady state magnetisation; to remove non-brain tissue, brain extraction tool (BET) (Smith, 2002) was used; volume realignment was performed using MCFLIRT

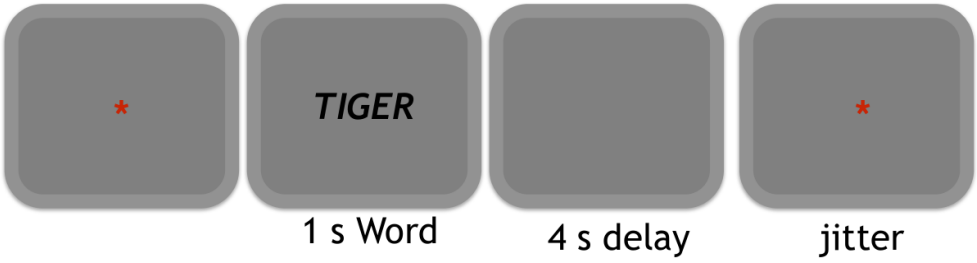

Figure 1: Illustration of the experimental protocol

(Jenkinson, Bannister, Brady, & Smith, 2002); minimal spatial smoothing was performed using a gaussian kernel with FWHM of 3 mm. Next, ICA based automatic removal of motion artifacts (ICA-AROMA) was used to remove motion-induced signal variations (Pruim et al., 2015) and this was followed by a high-pass filter with a cutoff of 60 s. The sessions aligned to a reference volume of the first session.

A set of 15 left-lateralized ROIs was pre-specified (see Figure 2) based on a meta-analysis of the semantic system (Binder et al., 2009) and one anterior temporal lobe (ATL) due to its role as a "semantic hub" (Patterson et al., 2007). Hence, the ROIs included: inferior parietal lobe (IPL), inferior temporal lobe (ITL), middle temporal lobe (MTL), precuneus, fusiform gyrus (FFG), parahippocampal gyrus (PHG), superior frontal gyrus (SPG), posterior cingulate gyrus (PCG), pars opercularis (POP), pars triangularis (PTR), pars orbitalis (POR), frontal pole (FP), medial orbitofrontal cortex (MOFC), laterial orbitofrontal cortex (LOFC), and anterior temporal lobe (ATL). First, automatic segmentation of the high-resolution structural scan was done with FreeSurfer's automated algorithm `recon-all`. The resulting masks were transformed to functional space using 7 DoF linear registrations implemented in FSL FLIRT (Jenkinson et al., 2002) and binarized. All further analyses were performed in native BOLD space.

2.4 Multivariate Pattern Analysis for Decoding

Multivariate pattern analysis was conducted using scikit-learn (Pedregosa et al., 2011) and PyMVPA (Hanke et al., 2009). Specifically, classification based on a supervised machine learning algorithm i.e. linear support vector machine (Fan, Chang, Hsieh, Wang, & Lin, 2008), was used to evaluate whether multi-voxel patterns in each of the ROIs contained information about the word semantic category (animal vs. tools) in each of the experimental

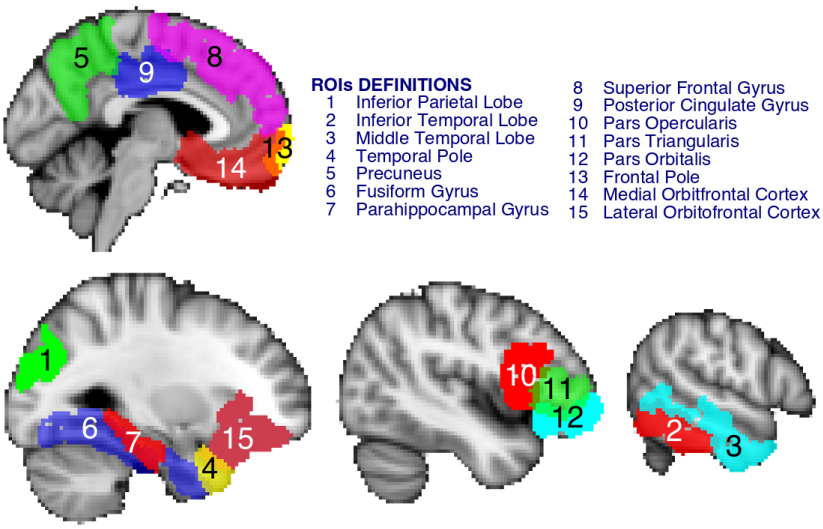

Figure 2: The figure shows the selected regions of interest projected on an MNI standard template image. The 15 left-lateralized areas were pre-specified and included regions: inferior parietal lobe, inferior temporal lobe, middle temporal lobe, precuneus, fusiform gyrus, parahippocampal gyrus, superior frontal gyrus, posterior cingulate gyrus, pars opercularis, pars triangularis, pars orbitalis, frontal pole, medial orbitofrontal cortex, laterial orbitofrontal cortex and anterior temporal lobe

contexts (i.e. deep and shallow processing).

2.4.1 Data Preparation

For each participant, the relevant time points or scans of the preprocessed fMRI data of each run were labeled with attributes such as word, category, and condition using the behavioural data files generated by Psychopy. Invariant voxels (or features) were removed. These were voxels whose value did not vary throughout the length of one session. If not removed, such features can cause numerical difficulties with procedures like z-scoring of features. Next, data from all sessions were stacked and each voxel's time series was run-wise z-scored (normalized) and linear detrended. Finally, to account for the hemodynamic lag, examples were created for each trial by averaging the 4 volumes between the interval of 3.4 s and 6.8 s after word onset.

2.4.2 Pattern Classification

Linear support vector machine (SVM) classifier, with all parameters set to default values as provided by the scikit-learn package ($l2$ regularization, $C = 1.0$, $tolerance = 0.0001$), was used. The following procedure was repeated for each ROI separately. To obtain an unbiased generalization estimate, following (Varoquaux et al., 2016) the data was randomly shuffled and split to create 300 sets of balanced train-test (80%-20%) splits with separate items for training and testing. Each example was represented by a single feature vector with each feature being the mean of voxel intensities across the sub-interval of 3.4 s and 6.8 s. Hence, the length of feature vector was equal to the number of voxels in the ROI. To further reduce the dimensionality of the data and thus reduce the chances of overfitting (Pereira, Mitchell, & Botvinick, 2009; T. Mitchell et al., 2004), Principal Component Analysis (PCA) with all parameters set to default values as provided by the scikit-learn was used. The number of components was equal to the number of examples thus resulting in all ROIs having an equal number of components. These components were linear combinations of the preprocessed voxel data and since none of the components was excluded, it was an information lossless change of the coordinate system to a subspace spanned by the examples (Mourão Miranda, Bokde, Born, Hampel, & Stetter, 2005). Features thus created were used to train the decoder, and its classification performance on the test set was recorded. This procedure was repeated separately for each of the 300 sets, and the mean of corresponding accuracies was collected for each of the participants.

2.4.3 Statistics

To determine whether the observed decoding accuracy in a given ROI is statistically significantly different from the chance-level of 0.5 (or 50%), a two-tailed t-test was performed with p-values corresponding to each of the ROIs corrected for multiple comparisons using false discovery rate (FDR).

2.4.4 Informational Connectivity Pipeline

Informational connectivity analysis is used to identify regions of the brain that display temporal correlation in the multivariate patterns regarding the key stimulus classes (Coutanche & Thompson-Schill, 2013). The purpose of this analysis here was to investigate how this informational connectivity between the 15 pre-specified ROIs varies across the deep and shallow processing conditions. The fMRI data was preprocessed and labeled as mentioned in § 2.3 and § 2.4.1. PCA was used for dimensionality reduction, and SVM

for classification (see § 2.4.2). A leave-one-trial out cross-validation was performed. Specifically, the classifier was trained on all the volumes between the sub-interval of 3.4 s to 6.8 s of all the trials except one, and was tested on all the volumes (not confined to any sub-interval) of the left-out trial. All the trials were used, one by one, as test trials starting with the first trial, and the corresponding probability of detecting a correct class was recorded for each of the volumes. In this way, a time-series of MVP discriminability values (one per timepoint) was obtained for each of the ROIs. To calculate the informational connectivity between the ROIs, these time-series were correlated using Pearson moment-to-moment correlation for each of the pairs of ROIs, and a matrix of informational connectivity was created. This procedure was performed separately for shallow and deep processing sessions resulting in two matrices for each of the participants.

2.4.5 Word Embedding Models

The word embedding models used in the analysis (Fast Text, GloVe and Word2vec) were pre-trained (Bravo-Marquez & Kubelka, 2018) using the Spanish Billion Word Corpus [a]. A common feature of these three models is that they are trained based on a corpus of text, and hence are sensitive to the characteristics of the text that capture both semantic but and syntactic information. For each word used in the current experiment, the corresponding vector representation of 300 dimensions (300-D) was extracted. 300 was a conventional choice that is commonly used in the natural language processing community. To visualize the representational patterns of the word embedding features, we computed the representational dissimilarity matrices (RDMs) of the models. Prior to the visualization, the representational feature of each word was normalized by subtracting the mean. The RDMs respect to Fast Text (see Figure S1), GloVe (Figure S2), and Word2Vec (Figure S3) are shown in the Supplementary materials.

2.4.6 Computer Vision Models

We applied computer vision models to extract abstract representations of the image referents associated with the words used in the present study. VGG19 (Simonyan & Zisserman, 2014), MobileNetV2 (Howard et al., 2017), and DenseNet169 (Huang, Liu, Van Der Maaten, & Weinberger, 2017) were the computer vision models used in the current analyses. These were pre-trained models provided by the Keras Python library (Tensorflow 2.0 edi-

[a] http://crscardellino.github.io/SBWCE/

tion) (Chollet, 2015) using the ImageNet dataset [b]. The pre-trained models, in general, were multi-layer convolutional neural networks. After several convolution-pooling blocks, a global pooling layer was applied to represent the image with a dense vector. Followed by several feedforward layers of fully-connected layers, the probability of each class was estimated. The training process involved predicting one of the one thousand classes of the ImageNet dataset given an image. However, the representational dimensionality of the computer vision models are different from the word embedding models (e.g. VGG19 represents images with 512-D, while DenseNet represents images with 1664-D). Thus, we performed fine-tuning (Mesnil et al., 2011; Yosinski, Clune, Bengio, & Lipson, 2014) to have each of the computer vision models to have 300-D as the word embedding models. The computer vision models were fine-tuned using 101 assorted object category images [c] (Fei-Fei, Fergus, & Perona, 2004). 'BACKGROUND Google', 'Faces', 'Faces easy', 'brain', 'stop sign', and 'trilobite' were removed from the categories due to the lack of conceptual relationship to the "living, non-living" categories of our experiment. In order to balance the instances of each category, 30 images were randomly selected from each image set. Six images of each category were randomly selected to form the validation set, while the rest formed the training set. We added an additional layer of 300 units, and a classification layer to decode the rest of 96 categories. The convolution pooling blocks of the pretrained model were frozen during training and validation, and only the newly added layer's weights were to be updated. In order to obtain robust performance, image augmentation such as rotation, shifting, zoom, and flipping was applied. Before training or validating, all images were normalized according to the selected models (see preprocessing steps in Keras documentation) and resized to 128 by 128 pixels, and the normalization procedure was applied to each image individually. The activation function of the layer with 300 units was Self-Normalizing (SELU) (Klambauer, Unterthiner, Mayr, & Hochreiter, 2017) and the weights were initialized with LeCun normal initialization procedure (Y. A. LeCun, Bottou, Orr, & Müller, 2012) as suggested in the Tensorflow documentation [d]. The optimizer was Adam and the learning rate was 0.0001. Loss function was categorical cross-entropy (Kingma & Ba, 2014). No other regularization procedures were used. During training, the fine-tuning model could reach to a maximum of 3000 epochs with batch size of 16 images per step of back-propagation. However, if the model's performance on the validation set did not improve for 5 consecutive epochs, the

[b] http://www.image-net.org/

[c] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[d] https://www.tensorflow.org/api_docs/python/tf/keras/activations/selu/

training would be terminated and the fine-tuned weights were saved for later use.

These fine-tuned computer vision models were applied to extract abstract representations of the image referents of the words. In particular, for each word, we sampled 10 images collected from the internet. Images were cropped and the object appeared at the center on white background. The values of each image were normalized according to the selected model. The output vector of the newly added layer of 300 units for a given image was the feature representation associated with the image referent of the word, with the 10 vector representations averaged. The feature extraction was done trial-wise accordingly for each participant. We selected the above 3 models on the following grounds. Evidence [e] has shown that the DenseNet169 (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Freeman et al., 2013; Majaj, Hong, Solomon, & DiCarlo, 2015) can explain on average more variance of the brain response than many other computer vision models. MobileNetV2 was chosen to represent a simpler computer vision model that had less parameters (VGG19 has 143,667,240, DenseNet169 has 14,307,880, and MobileNetV2 has 3,538,984). VGG19 was chosen to represent a baseline model that had shallower architeture (VGG19 has 26 layers, DenseNet169 has 169 layers, and MobileNetV2 has 88 layers), which has been well studied by previous studies (Ren, He, Girshick, & Sun, 2015; He, Zhang, Ren, & Sun, 2016; Russakovsky et al., 2015).

To visualize the representational patterns of the computer vision models, we computed the representational dissimilarity matrices (RDMs) of the models similar to the word embedding models. These are in shown in the Supplementary materials (DenseNet169, Figure S4; MobileNetV2, Figure S5; and VGG19,Figure S6).

2.4.7 Encoding Model Pipeline

The encoding model pipeline was the same as in Miyawaki et al. 2008 (Miyawaki et al., 2008) and implemented in Nilearn (Pedregosa et al., 2011; Buitinck et al., 2013). After standardizing the feature representations by subtracting the mean and dividing by standard deviation, the feature representations were mapped to the BOLD signal of a given ROI by means of L2 regularized regression (Ridge Regression) with the regularization term equal to 100, following Miyawaki et al.

[e] http://www.brain-score.org/

To estimate the performance of the regression, we partitioned the data into 300 folds to perform cross-validation by stratified random shuffling similar to the decoding pipeline above (Little et al., 2017; Varoquaux et al., 2016). It is important to note that labels of the BOLD signals were only used for cross-validation partitioning, but they were not used in the encoding model fitting nor testing procedure. In each fold, we randomly held out 20% of the data for testing, while the remaining 80% was used to fit the ridge regression model, using the feature representations from word embedding models or computer vision models as features and the BOLD signals as targets. Predictions were then derived for the held-out data. The proportion of variance explained in each voxel was computed for the predictions. An average estimate of the variance explained was calculated. The best possible score is 1. The score can be also negative if the model is worse than random guessing. Voxels that had positive variance explained values were identified for further analysis (Miyawaki et al., 2008; Holdgraf et al., 2017) for each participant, ROI and condition. To estimate the empirical chance level performance of the encoding models, a random shuffling was added to the training phase during cross-validation before model fitting. The random shuffling was applied to the order of the samples for the features in the encoding models while the order of the samples for the targets remained the same.

3 Results

3.1 Behavioral Results

To ensure that participants were attending to the items during the task, a few catch trials were randomly presented at different points in each session. These trials showed number words and required a response via button press. The total number of catch trials was kept the same in both shallow and deep processing sessions. Catch trial data from six initial participants could not be obtained due to a technical issue. The proportion of correct responses on catch trials was 0.80 ± 0.28 in the shallow processing, and 0.78 ± 0.37 in deep processing conditions, which did not differ ($t(20) = -0.40, p = 0.69$), hence showing that participants were equally engaged with the task in both conditions.

3.2 Multivariate Pattern Analyses: Decoding Results

Figure 3 presents the summary statistics of decoding accuracy across ROIs in both shallow and deep processing conditions (i.e. during mental simulation). During shallow processing condition, decoding accuracy of the semantic category (living/non-living) was significantly above chance in all pre-specified ROIs except the frontal pole and medial orbitofrontal cortex: frontal pole (50.74±4.38, $t(26) = 0.87$, $p = 0.39$), fusiform gyrus (54.17±3.95, $t(26) = 5.39$, $p = 0.00009$), inferior parietal lobe (53.20±3.75, $t(26) = 4.34$, $p = 0.0003$), inferior temporal lobe (54.70±4.19, $t(26) = 5.71$, $p = 0.00008$), lateral orbitofrontal cortex (53.42±3.96, $t(26) = 4.40$, $p = 0.0003$), medial orbitofrontal cortex (54.23±4.65, $t(26) = 4.64$, $p = 0.0003$), middle temporal lobe (53.79±3.78, $t(26) = 5.10$, $p = 0.0001$), parsopercularis (53.51±4.27, $t(26) = 4.18$, $p = 0.0004$), parsorbitalis (53.11±3.59, $t(26) = 4.42$, $p = 0.0003$), parstriangularis (53.22±4.48, $t(26) = 3.66$, $p = 0.001$), parahippocampal gyrus (52.63±4.11, $t(26) = 3.25$, $p = 0.004$), post-cingulate gyrus (52.78±4.17, $t(26) = 3.39$, $p = 0.003$), precuneus (53.61±3.95, $t(26) = 4.66$, $p = 0.0003$), superior frontal lobe (53.27±3.78, $t(26) = 4.41$, $p = 0.0003$), and temporal pole (51.64±3.73, $t(26) = 2.25$, $p = 0.036$).

In the deep processing condition, decoding accuracy was also significantly above chance in all pre-specified ROIs (frontal pole (52.62±4.36, $t(26) = 3.07$, $p = 0.005$), fusiform gyrus (63.75±9.22, $t(26) = 7.61$, $p = 2.23e - 7$), inferior parietal lobe (63.31±9.72, $t(26) = 6.98$, $p = 5.93e - 07$), inferior temporal lobe (64.18±9.50, $t(26) = 7.61$, $p = 2.23e - 07$), lateral orbitofrontal cortex (58.12±5.19, $t(26) = 7.98$, $p = 2.23e - 07$), medial orbitofrontal cortex (55.60±4.38, $t(26) = 6.52$, $p = 1.19e - 06$), middle temporal lobe (60.97±8.08, $t(26) = 6.92$, $p = 5.93e - 07$), parsopercularis (61.92±9.65, $t(26) = 6.30$, $p = 1.70e - 06$), parsorbitalis (56.15±4.84, $t(26) = 6.49$, $p = 1.19e - 06$), parstriangularis (58.68±4.48, $t(26) = 6.68$, $p = 9.45e - 07$), parahippocampal gyrus (57.30±5.23, $t(26) = 7.12$, $p = 5.47e - 07$), post-cingulate gyrus (57.06±7.24, $t(26) = 4.97$, $p = 3.88e - 05$), precuneus (58.55±7.41, $t(26) = 5.88$, $p = 4.16e - 06$), superior frontal lobe (60.10±8.23, $t(26) = 6.26$, $p = 1.74e - 06$), and temporal pole (54.51±4.44, $t(26) = 5.19$, $p = 0.2.36e - 05$)).

Notably, decoding in most of the above regions was higher during deep processing compared to the shallow processing condition (see Figure 3).

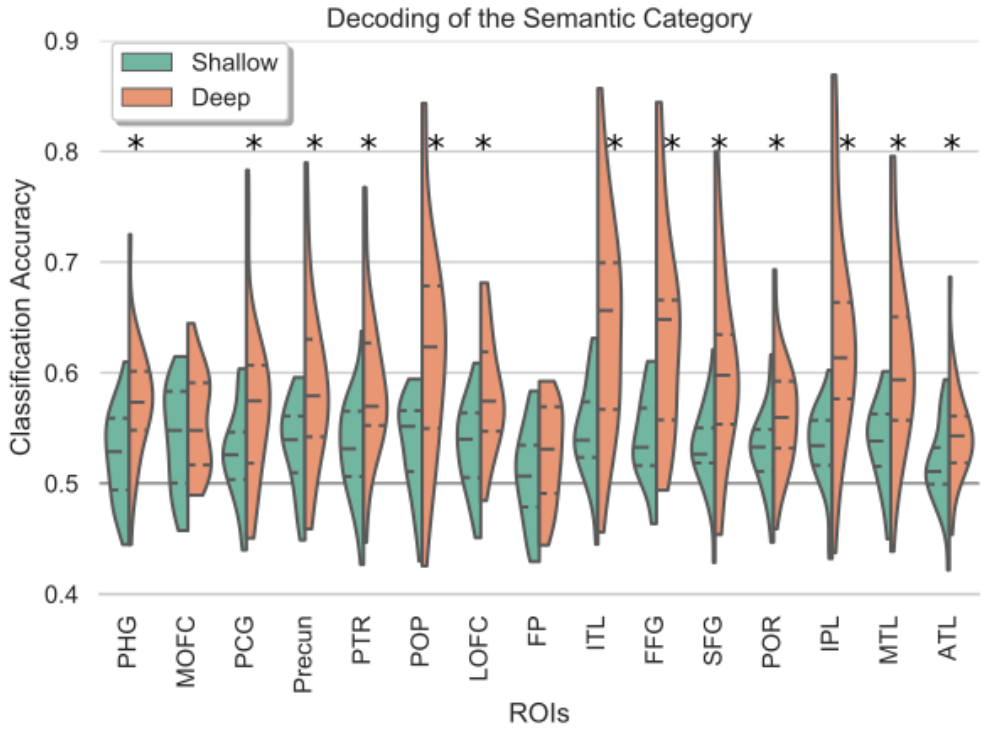

Figure 3: The figure shows summary statistics of decoding accuracy for the different ROIs. The three dotted lines inside each violin are the quartiles. The black asterisks mark ROIs that showed statistically significant improvement in decoding accuracy in deep as compared to shallow processing condition.

3.3 Out-of-sample Generalization

We then repeated the decoding analyses with a different cross-validation procedure that allowed testing the generalizability of the semantic representations and how this was modulated by the different task contexts. Specifically, the classifier was trained using all the words but leaving a pair of words out from each class. Then the classifier was tested on the left-out pair.

Figure 4 presents the summary statistics of the ROIs for out-of-sample generalization in both shallow and deep processing conditions. It can be seen that in the shallow processing condition, the decoding of the semantic category (living/non-living) was found to be at chance-level in all pre-specified ROIs including the frontal pole (50.23±2.60, $t(26) = 0.46$, $p = 0.82$), fusiform gyrus (50.70±2.02, $t(26) = 1.76$, $p = 0.28$), inferior parietal lobe (50.76±2.15,

$t(26) = 1.80$, $p = 0.27$), inferior temporal lobe (50.76±2.13, $t(26) = 1.81$, $p = 0.27$), lateral orbitofrontal cortex (50.17±2.44, $t(26) = 0.36$, $p = 0.82$), medial orbitofrontal cortex (50.92±2.85, $t(26) = 1.64$, $p = 0.28$), middle temporal lobe (50.74±1.91, $t(26) = 1.97$, $p = 0.27$), parsopercularis (50.44±2.36, $t(26) = 0.95$, $p = 0.59$), parsorbitalis (50.10±2.25, $t(26) = 0.23$, $p = 0.82$), parstriangularis (50.46±2.89, $t(26) = 0.82$, $p = 0.63$), parahippocampal gyrus (50.26±2.20, $t(26) = 0.59$, $p = 0.76$), post-cingulate gyrus (50.13±2.23, $t(26) = 0.30$, $p = 0.82$), precuneus (50.76±2.18, $t(26) = 1.77$, $p = 0.27$), superior frontal lobe (50.50±2.33, $t(26) = 1.09$, $p = 0.54$), and temporal pole (49.36±2.21, $t(26) = -1.48$, $p = 0.32$).

On the other hand, in the deep processing condition, the classification of the semantic category was found to be significantly above chance in all pre-specified ROIs (frontal pole (50.98±2.24, $t(26) = 2.24$, $p = 0.03$), fusiform gyrus (59.43±8.06, $t(26) = 5.97$, $p = 2.10e - 5$), inferior parietal lobe (59.25±8.64, $t(26) = 5.45$, $p = 2.10e-5$), inferior temporal lobe (59.17±8.35, $t(26) = 5.60$, $p = 2.10e - 05$), lateral orbitofrontal cortex (53.30±2.92, $t(26) = 5.78$, $p = 2.10e - 05$), medial orbitofrontal cortex (52.34±2.50, $t(26) = 4.77$, $p = 7.60e - 05$), middle temporal lobe (58.29±7.50, $t(26) = 5.64$, $p = 2.10e - 05$), parsopercularis (59.27±8.60, $t(26) = 5.50$, $p = 2.10e - 05$), parsorbitalis (53.18±3.13, $t(26) = 5.19$, $p = 3.40e - 05$), parstriangularis (55.78±5.09, $t(26) = 5.79$, $p = 2.10e - 05$), parahippocampal gyrus (54.50±4.22, $t(26) = 5.43$, $p = 2.10e - 05$), post-cingulate gyrus (54.59±6.04, $t(26) = 3.88$, $p = 0.0007$), precuneus (55.22±5.40 $t(26) = 4.93$, $p = 5.5e-05$), superior frontal lobe (56.67±6.83, $t(26) = 4.98$, $p = 5.3e - 05$), and temporal pole (51.57±2.97, $t(26) = 2.68$, $p = 0.01$).

Decoding accuracy in 13 out of 15 ROIs was also better than in the shallow processing condition (see Figure 4). This shows that generalization of brain representations of semantic knowledge is better when the depth of processing is higher during mental simulation.

3.4 Informational Connectivity Results

We then assessed changes in the informational connectivity across the different ROIs, during deep relative to shallow processing conditions. This approach assesses the temporal correlation of multivariate patterns of responses across different ROIs and can reveal whether different brain regions interact in terms of the specific information that they carry (Coutanche & Thompson-Schill, 2013).

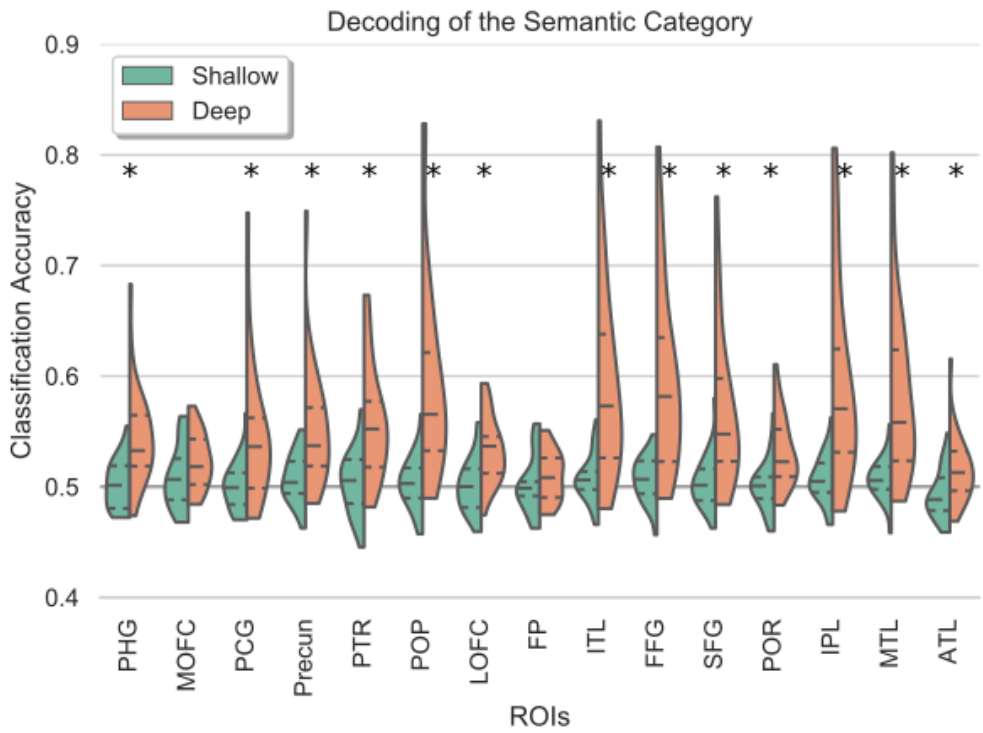

Figure 4: The figure shows summary statistics of the ROIs for out-of-sample generalization of the semantic category. The three dotted lines inside each violin are the quartiles. The black asterisks mark ROIs that showed statistically significant improvement in decoding accuracy in deep as compared to shallow processing condition.

Figure 5 illustrates the regions showing statistically higher correlations of decoding time-courses during mental simulation compared to the shallow processing. There were two substrates that showed the strongest changes in informational connectivity with the rest of the semantic network. First, informational connectivity between the left orbito-frontal cortex increased with inferior temporal, parahippocampal, fusiform, inferior temporal, anterior temporal lobe and inferior parietal. A similar effect involved the middle temporal cortex, which showed increased informational connectivity during mental simulation with pars triangularis, medial and lateral orbitofrontal, frontal pole, precuneus and posterior cingulate and anterior temporal lobe. Figure 5 also shows that the parsopercularis, and superior frontal lobe showed no changes in informational connectivity as function of the depth of processing.

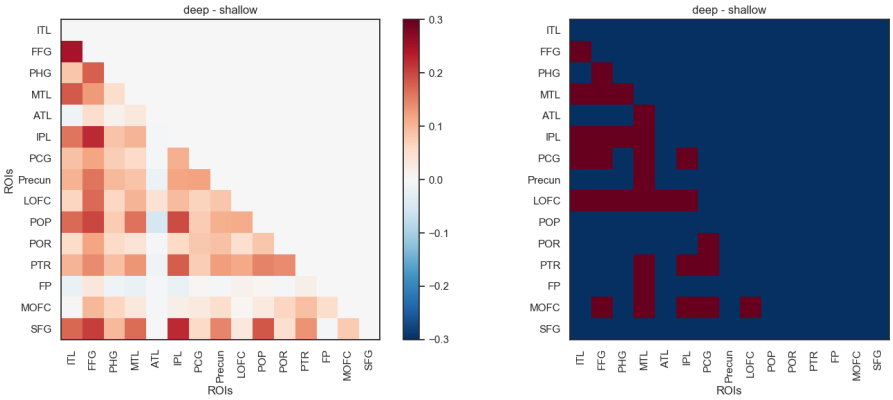

Figure 5: The figure shows the mean difference of informational connectivity between deep and shallow conditions. The right panel shows which of the pairs of ROIs were significantly more connected in deep as compared to the shallow condition.

3.5 Encoding Results

An encoding model predicts the brain activity patterns using a set of features that are (non)linearly transformed from the stimuli (Diedrichsen & Kriegeskorte, 2017; Kriegeskorte & Douglas, 2018). Word embedding models (i.e. Word2Vec-2013 (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)) provide a way to characterize the geometry of semantic spaces. Computer vision models (i.e. deep convolutional neural networks (Y. LeCun, Bottou, Bengio, & Haffner, 1998)can also reveal the structural organization of meaning in the ventral visual pathway (Simonyan & Zisserman, 2014). To examine the properties of the brain representations during word recognition, we tested both word embeddings and also computer vision models based on the image referents of the words used in the experiment (see § 2.4.5 and § 2.4.6).

Figure 6 shows the average variance explained by the computer vision (VGG19, Densent169, MobilenetV2) and the word embedding (Fast Text, GloVe, Word2Vec) models, averaging across 27 subjects. The error-bars represent 95% confidence interval of a bootstrapping of 1000 iterations. For each ROI, a one-way analysis of variance (ANOVA) was performed within the computer vision models and within the embedding models. The ANOVAs aimed to detect the difference in variance explained within a type of models. After all the ANOVAs were performed, FDR correction procedures were applied to the raw p-values to correct for the multiple comparison within each condition (deep vs. shallow). There was no difference among the different

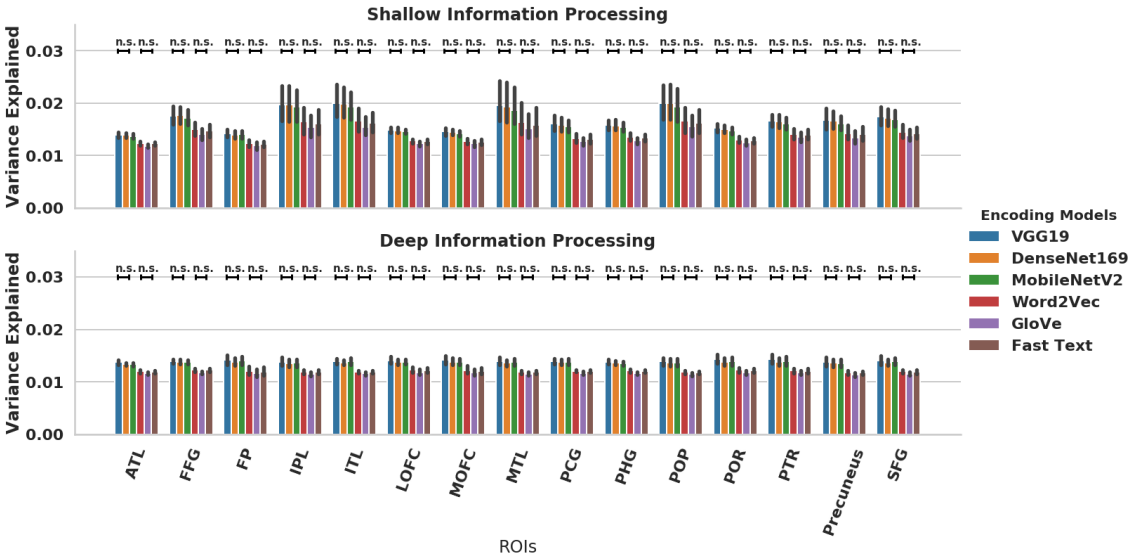

Figure 6: Average Variance Explained by each of the Word Embedding and Computer Vision Models.

word embedding models, and no difference among the computer vision models.

We then computed the difference between each of the computer vision model and each of the word embedding models within each ROI and condition, in order to assess whether word embedding models or computer vision models were better. One-sample t-tests against zero were conducted for each pair with FDR correction. All the computer vision models performed better than any of the word embedding models (see Figure 7).

Note the above results were obtained with pre-trained computer vision models that were fine tuned to ensure that the abstract layer had 300-D for comparison with the word embedding models. Otherwise, the results might not be comparable due to differences in the dimensionality of the models. However, we performed the same encoding analyses using the original dimensionality of the computer vision models and similar patterns were obtained. This indicates the robustness of the encoding results (see Supplementary Results and Supplementary Figures S7, S8, S9).

Finally, we computed the mean of the differences in variance explained between the computer vision models and the word embedding models and then assess the extent of this difference in the deep and shallow processing

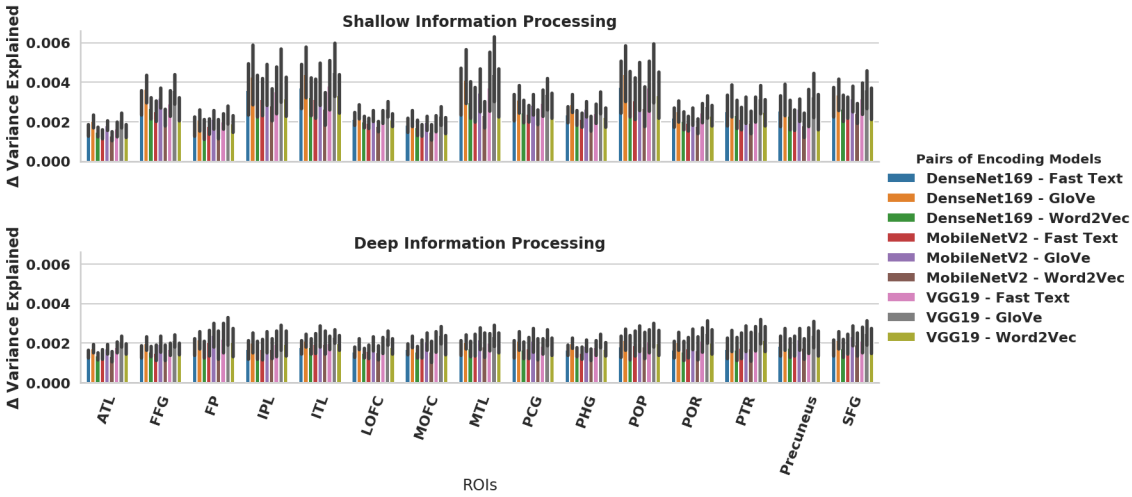

Figure 7: Differences between computer vision and word embedding models in variance explained. Computer vision models significantly explained more variance of the BOLD responses compared to word embedding models.

conditions by using paired t-tests with FDR correction. Figure 8 illustrates the pattern of results within each ROI. We found that the advantage of computer vision models over word embeddings models was higher in the deep processing condition relative to the shallow processing in PCG, PHG, and POP, while the opposite pattern was observed in FFG, IPL, and ITL (see Figure 7).

4 Discussion

We sought to understand how the depth of processing shapes the brain representations of conceptual knowledge by using both decoding and encoding models. The results clearly show that the decoding of word category information in most of the putative substrates of the semantic network (Binder et al., 2009) is consistently higher during the mental simulation condition relative to a shallow processing condition in which participants merely read the words. Semantic category was also decoded in the read condition, though decoding performance was far lower by comparison and the classifier did not reliably generalise to new examples that were not seen during training. By contrast, the generalizability of the classifier at predicting the semantic category of out-of-sample examples increased during mental simulation. Shallower processing modes are sufficient for in-sample classification but not enough for out-of-sample generalization.

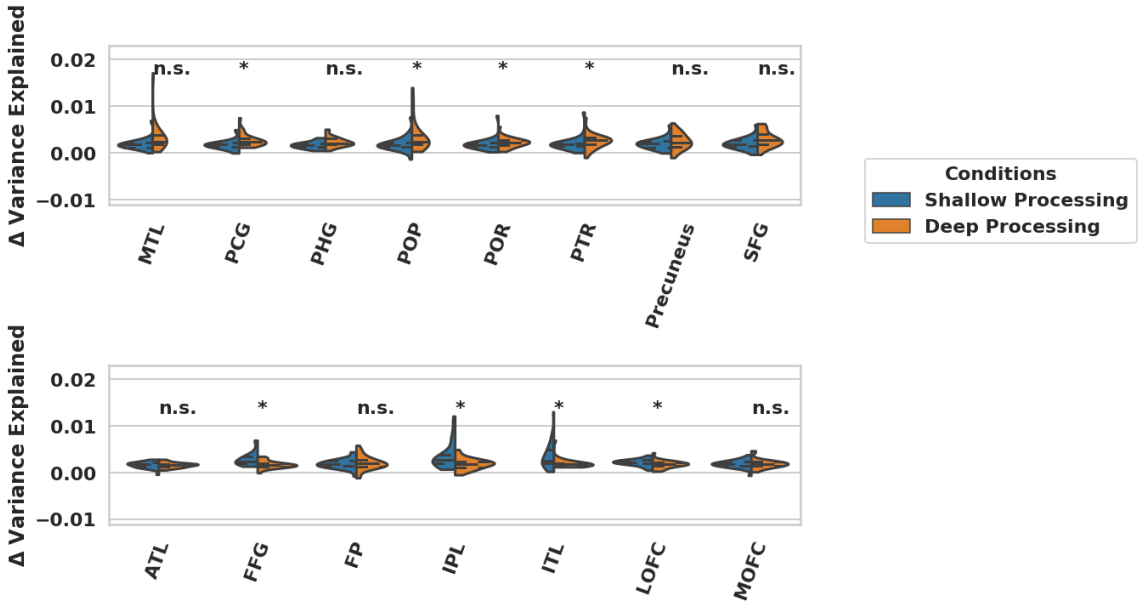

Figure 8: Overall Difference between Word Embedding and Computer Vision models per ROI.

Significant decoding of semantic category was observed in multiple areas of association (transmodal) cortex involving the middle temporal gyrus, anterior and inferior temporal, inferior parietal lobe and prefrontal substrates. Of particular relevance are the prefrontal areas, which are typically thought to be involved in semantic control, rather than representing semantic knowledge (Wagner, Paré-Blagoev, Clark, & Poldrack, 2001; Poldrack et al., 1999; Whitney, Kirk, O'Sullivan, Lambon Ralph, & Jefferies, 2010). The present multivariate classification results are consistent with a role of association cortex in the representation of semantic categories too. Interestingly, the level of informational connectivity (Anzellotti & Coutanche, 2018) observed during mental simulation also indicates that association transmodal cortices interact with multiple regions of the semantic network in terms of the specific information that multivoxel activity patterns carry across time. Relative to the shallow processing condition, informational connectivity increased during mental simulation between the left anterior prefrontal cortex, inferior parietal, temporal and occipital areas. Likewise, the middle temporal cortex - a region implicated in semantic control according to a prior meta-analyses (Noonan, Jefferies, Visser, & Lambon Ralph, 2013) - showed increased informational connectivity with inferior frontal and anterior prefrontal areas, the posterior cingulate and anterior temporal lobe. The depth of processing

associated with mental simulation therefore triggered a general broadcasting effect of the specific content being represented, recruiting a highly distributed set of areas of the semantic network in an integrated manner.

Encoding models further revealed the properties of the semantic representations encoded in brain activity. Intriguingly, we found that computer vision models outperformed word embedding models in explaining brain responses across the different regions of the semantic network. We found that the embedding layer of the computer vision model, which is likely to contain a condensed summarization of visual features that are frequent in the examples of the image referents for a given word, explained more variance of the brain responses. The results indicate that the semantic and syntactic similarity metrics provided by word embedding models may not capture the actual semantic knowledge that is represented during word processing. Rather, it appears that image based properties are necessary to account for how the brain represents word meaning. Interestingly, this occurred both during mental simulation and also during the shallow processing condition. This is consistent with the idea that simulation processes may occur, to some extent, automatically (Barsalou, 2008), however, it is clear from the superior decoding performance and the increased informational connectivity during mental simulation that top-down factors associated with the depth of processing play a key modulating role.

It may be argued that the advantage of computer vision models ought to be higher during mental simulation relative to the shallow processing case because access to visual representations seems more likely in the former case. We only found mixed evidence for this. While computer vision models explained more variance in the posterior cingulate, inferior frontal and anterior prefrontal cortex during deep compared to shallow processing, the opposite pattern was found in the fusiform, inferior parietal lobe and inferior temporal lobe, which are part of the ventral visual pathway. To account for this pattern of results, we propose that the presentation of a word during shallow processing automatically activates an image referent, but additional contextual features are not activated. By contrast, during mental simulation, additional contextual features associated with the word are likely to be activated. These contextual items may not be related to the image referents that were sampled for the encoding analyses, which only contained the object shape. This might lead to the computer vision models capturing less variance in the deep processing than in the shallow processing in the ventral visual pathway. This would also explain why, in contrast to the pattern of decoding results, encoding model performance was not affected by the depth

of processing. Either way, the results of the encoding models argue against strong versions of the view that conceptual knowledge relies on modal representations that need no access to visual representations (Pecher, van Dantzig, & Schifferstein, 2009).

Together, these results demonstrate that that the depth of processing is a key factor for triggering distributed, generalizable and integrated representations in the semantic network.

5 Competing Interests Statement

The authors declare no competing interests.

Acknowledgements

D.S. acknowledges support from the Basque Government through the BERC 2018-2021 program, from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R & D (SEV-2015-490) and also from project grants PSI2016-76443-P from MINECO and PI-2017-25 from the Basque Government. R.S. acknowledges support by the Basque Government (IT1244-19 and ELKARTEK programs), and the Spanish Ministry of Economy and Competitiveness MINECO (project TIN2016-78365-R).

References

Anzellotti, S., & Coutanche, M. N. (2018, mar). Beyond functional connectivity: Investigating networks of multivariate representations. *Trends in Cognitive Sciences*, *22*(3), 258–269. Retrieved from https://doi.org/10.1016/j.tics.2017.12.002 doi: 10.1016/j.tics.2017.12.002

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Barsalou, L. W. (2008, apr). Cognitive and neural contributions to understanding the conceptual system. *Current Directions in Psychological Science*, *17*(2), 91–95. Retrieved from https://doi.org/10.1111/j.1467-8721.2008.00555.x doi: 10.1111/j.1467-8721.2008.00555.x

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

from https://doi.org/10.1016/j.tics.2011.10.001 doi: 10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009, mar). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. Retrieved from https://doi.org/10.1093/cercor/bhp055 doi: 10.1093/cercor/bhp055

Bravo-Marquez, F., & Kubelka, J. (2018). *spanish-word-embeddings.* https://tinyurl.com/y47xzh6l. GitHub.

Buchweitz, A., Shinkareva, S. V., Mason, R. A., Mitchell, T. M., & Just, M. A. (2012, mar). Identifying bilingual semantic neural representations across languages. *Brain and Language*, *120*(3), 282–289. Retrieved from https://doi.org/10.1016/j.bandl.2011.09.003 doi: 10.1016/j.bandl.2011.09.003

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML pkdd workshop: Languages for data mining and machine learning* (pp. 108–122).

Chollet, F. (2015). *keras.* https://tinyurl.com/pna7m6p. GitHub.

Coutanche, M. N., & Thompson-Schill, S. L. (2013). Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in human neuroscience*, *7*, 15.

Damasio, A. R. (1989, nov). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25–62. Retrieved from https://doi.org/10.1016/0010-0277(89)90005-x doi: 10.1016/0010-0277(89)90005-x

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, *13*(4), e1005508.

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, *9*, 1871–1874.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop* (pp. 178–178).

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, *16*(7), 974.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V.,

& Pollmann, S. (2009). Pymvpa: A python toolbox for multivariate pattern analysis of fmri data. *Neuroinformatics*, *7*(1), 37–53.

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).

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in systems neuroscience*, *11*, 61.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016, apr). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. Retrieved from https://doi.org/10.1038/nature17637 doi: 10.1038/nature17637

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. M. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.*

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). *Self-normalizing neural networks.*

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 1.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9–48). Springer.

Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. perspectives on saeb et al. *GigaScience*, *6*(5), gix020.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I., ... others (2011). Unsupervised and transfer learning challenge: a deep

learning approach. In *Proceedings of the 2011 international conference on unsupervised and transfer learning workshop-volume 27* (pp. 97–111).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., MA, J., & Newman, S. (2004, 01). Learning to decode cognitive states from brain images. *Machine learning*, *13*, 667-668.

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 meanings of nouns. *science*, *320*(5880), 1191–1195.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., ... Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915–929.

Mourão Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005, Dec). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data. *NeuroImage*, *28*(4), 980–95.

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *Journal of cognitive neuroscience*, *25*(11), 1824–1850.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007, dec). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. Retrieved from https://doi.org/10.1038/nrn2277 doi: 10.1038/nrn2277

Pecher, D., van Dantzig, S., & Schifferstein, H. N. J. (2009, oct). Concepts are not represented by conscious imagery. *Psychonomic Bulletin & Review*, *16*(5), 914–919. Retrieved from https://doi.org/10.3758/pbr.16.5.914 doi: 10.3758/pbr.16.5.914

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J. W. (2007, May). Psychopy–psychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, *45*(1), S199–S209.

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage*, *10*(1), 15–35.

Price, A. R., Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016, mar). Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. *The Journal of Neuroscience*, *36*(13), 3829–3838. Retrieved from https://doi.org/10.1523/jneurosci.3120-15.2016 doi: 10.1523/jneurosci.3120-15.2016

Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *Neuroimage*, *112*, 267–277.

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.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011, feb). Commonality of neural representations of words and pictures. *NeuroImage*, *54*(3), 2418–2425. Retrieved from https://doi.org/10.1016/j.neuroimage.2010.10.042 doi: 10.1016/j.neuroimage.2010.10.042

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. J. (2012, oct). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex*, *24*(2), 426–434. Retrieved from https://doi.org/10.1093/cercor/bhs324 doi: 10.1093/cercor/bhs324

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

Smith, S. (2002). Fast robust automated brain extraction. *Human brain mapping*, *17*(3), 143–155.

Varoquaux, G., Raamana, P. R., A. Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016, 06). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*.

Visser, M., Jefferies, E., & Ralph, M. A. L. (2010, jun). Semantic processing in the anterior temporal lobes: A meta-analysis of the functional neu-

roimaging literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094. Retrieved from https://doi.org/10.1162/jocn.2009.21309 doi: 10.1162/jocn.2009.21309

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. *Neuron*, *31*(2), 329–338.

Whitney, C., Kirk, M., O'Sullivan, J., Lambon Ralph, M. A., & Jefferies, E. (2010). The neural organization of semantic control: Tms evidence for a distributed network in left inferior frontal and posterior middle temporal gyrus. *Cerebral Cortex*, *21*(5), 1066–1075.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Supplementary Figures

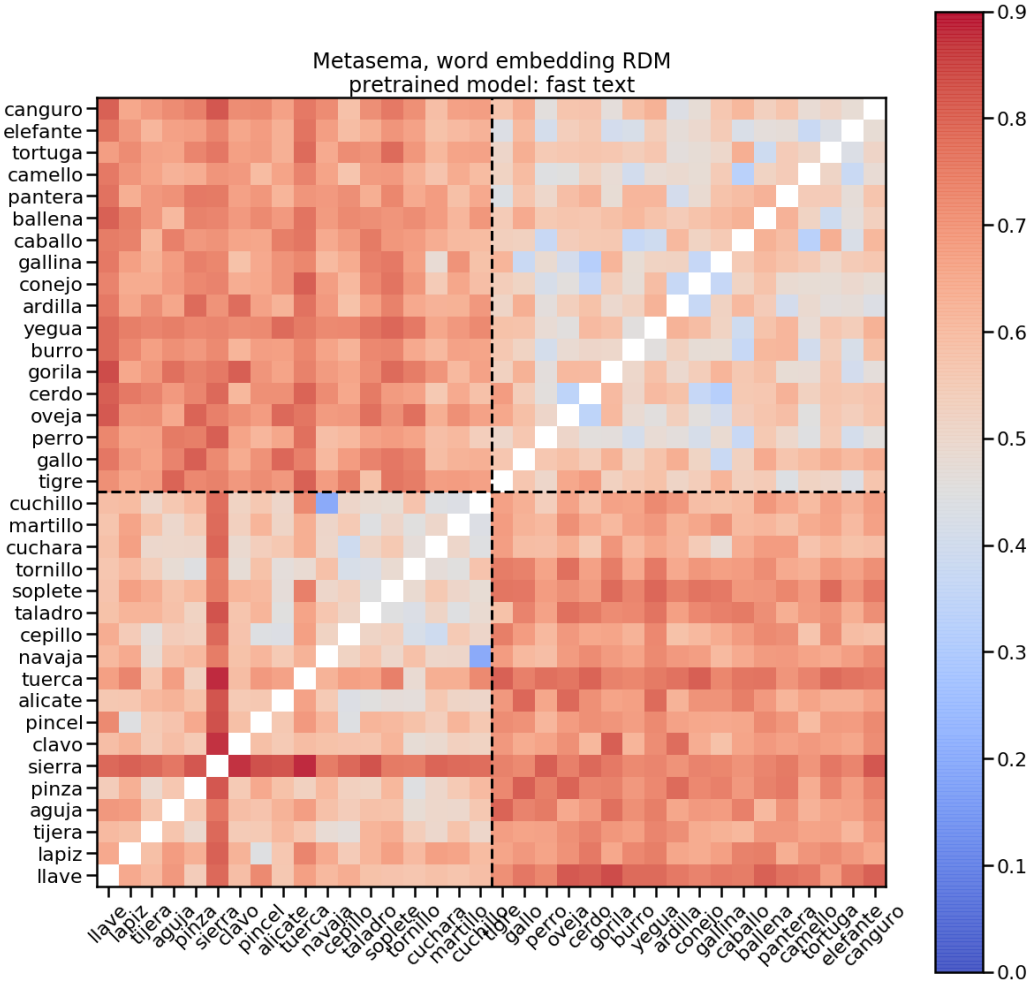

Figure S1: Representational Dissimilarity Matrix of Fast Text embedding features with cosine distance as metric.

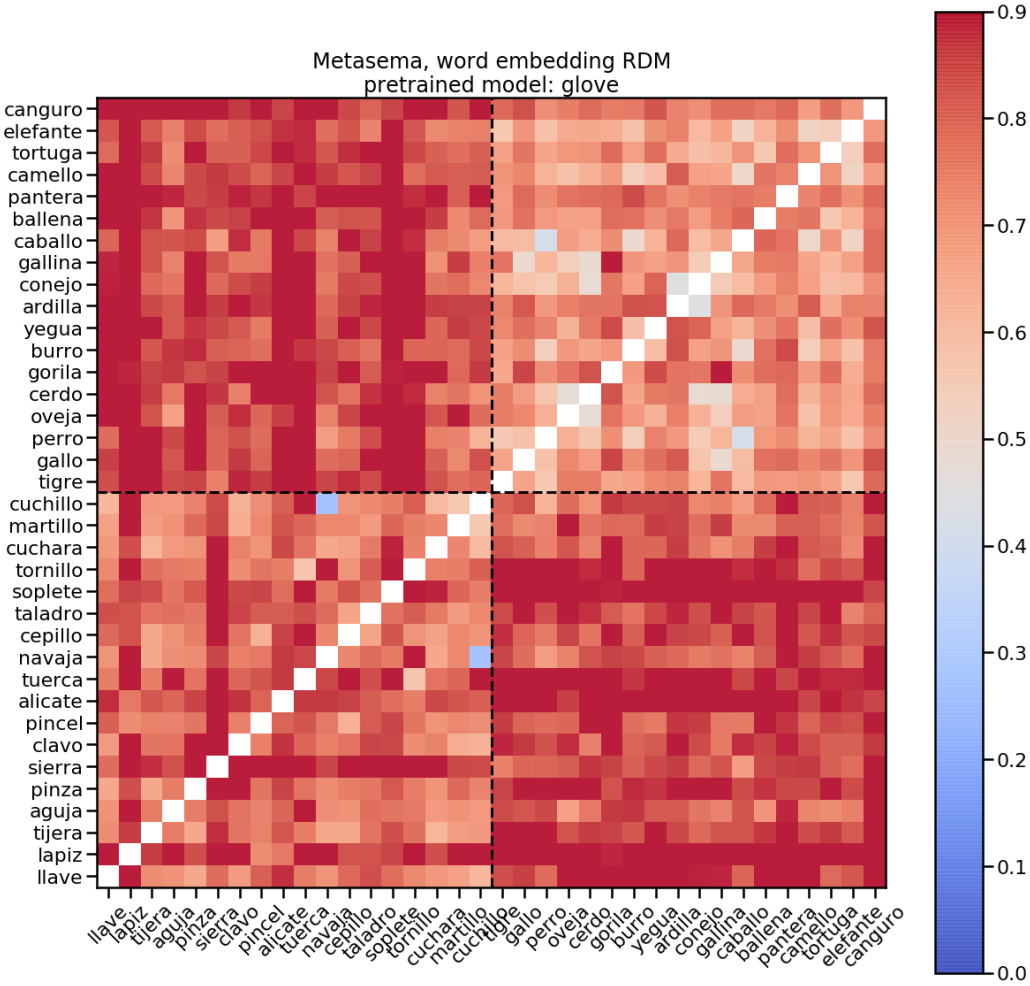

Figure S2: Representational Dissimilarity Matrix of GloVe embedding features with cosine distance as metric.

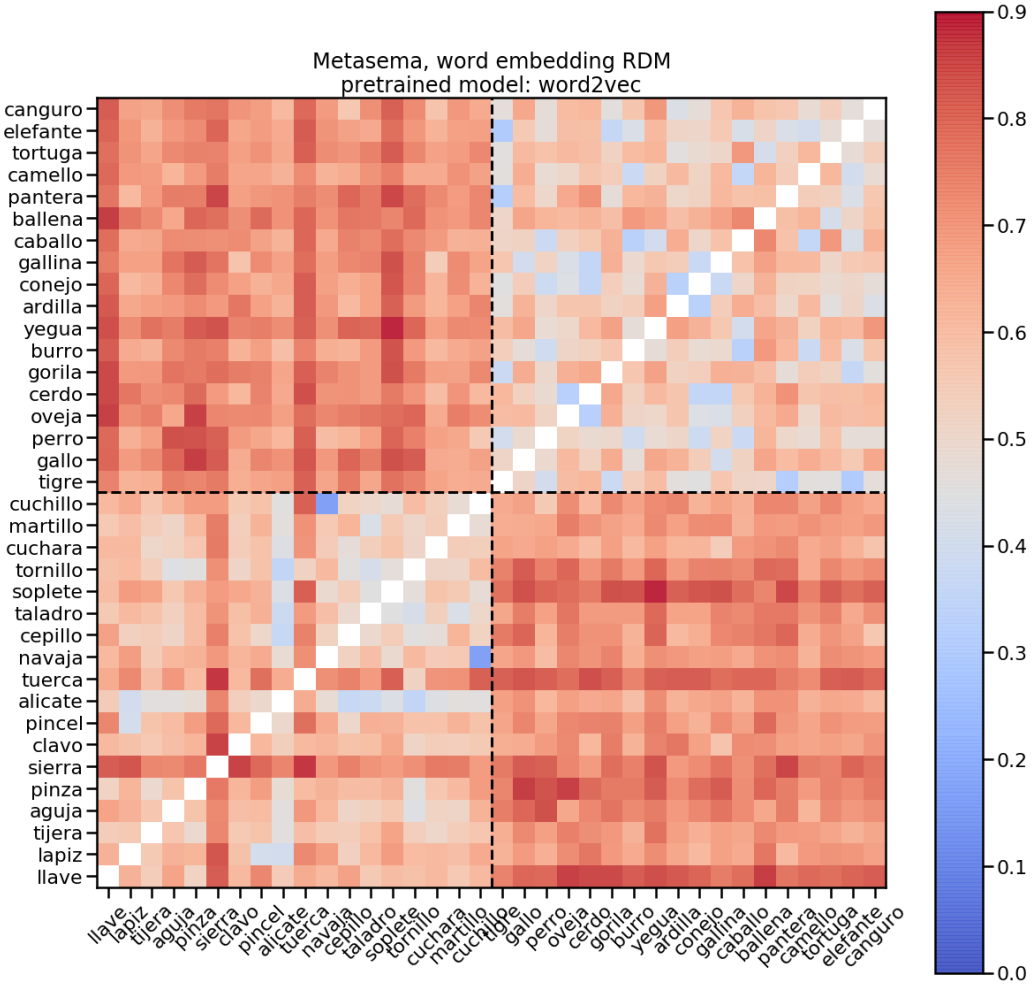

Figure S3: Representational Dissimilarity Matrix of Word2Vec embedding features with cosine distance as metric

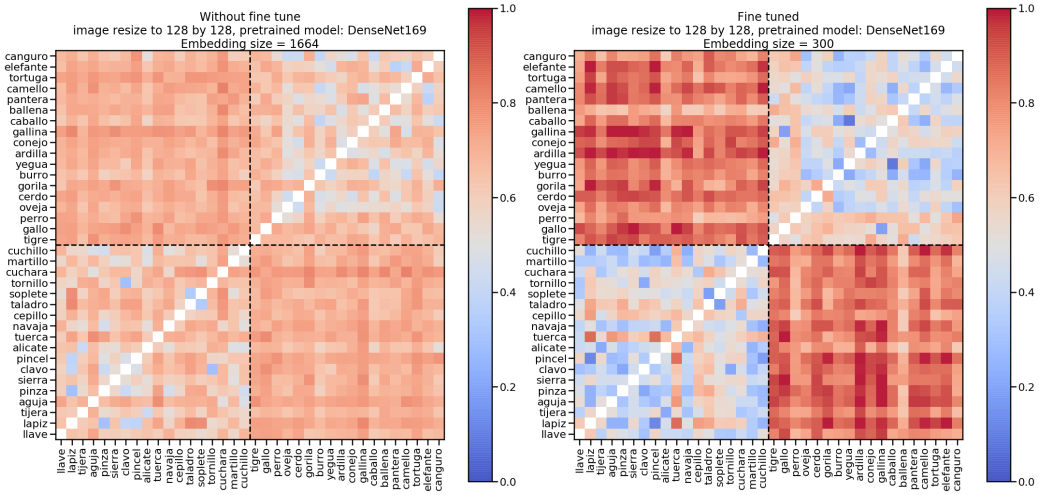

Figure S4: Representational Dissimilarity Matrix of DenseNet169 embedding features with cosine distance as metric.

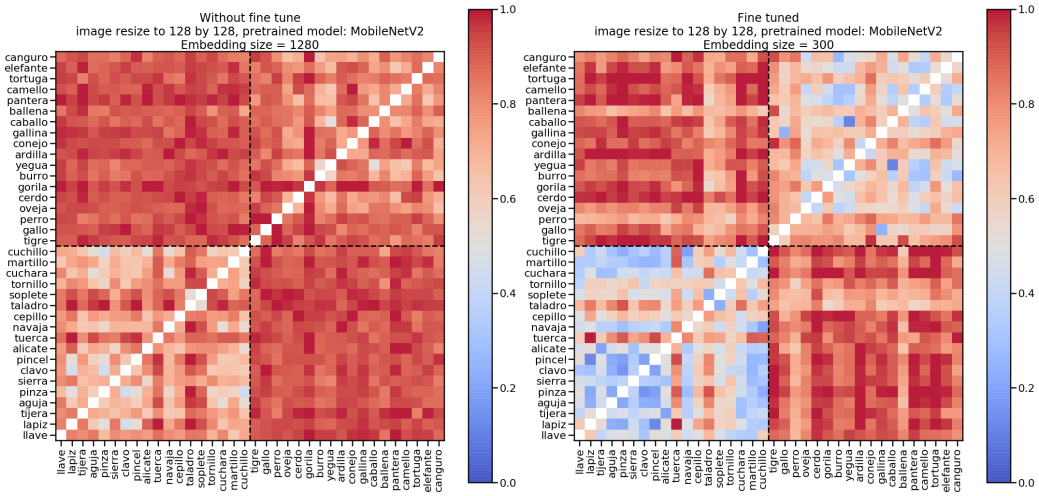

Figure S5: Representational Dissimilarity Matrix of MobileNetV2 embedding features with cosine distance as metric.

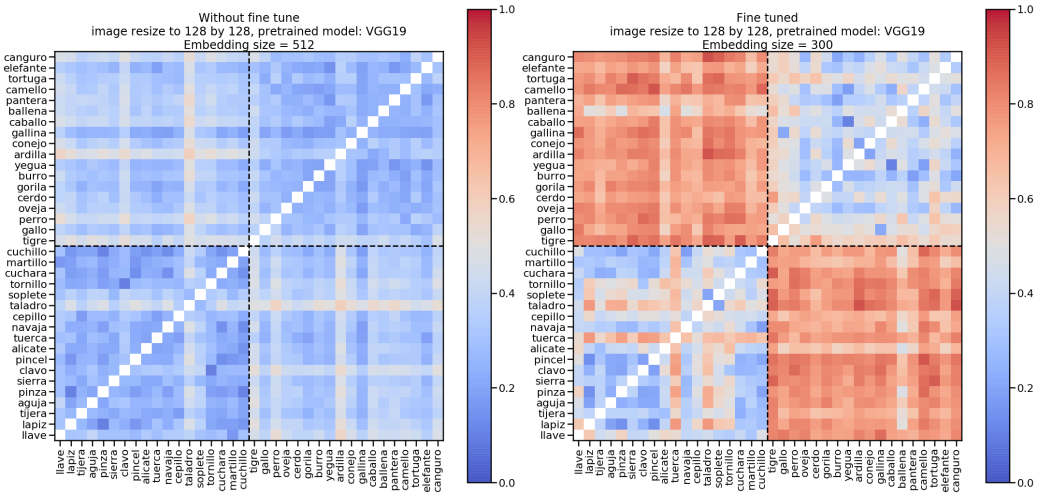

Figure S6: Representational Dissimilarity Matrix of VGG19 embedding features with cosine distance as metric.

Supplementary Results

Figure S7 shows the average variance explained by the computer vision (VGG19, Densent169, mobilenetV2) without fine-tuning and the word embedding (Fast Text, Glove, word2vec) models, averaged across 26 subjects. The errorbar represents 95% confidence interval of a bootstrapping with 1000 iterations. For each ROI, a one-way analysis of variance (ANOVA) was performed within the computer vision models and within the embedding models across subjects. The ANOVAs aimed to detect the difference in variance explained within a type of models. After all the ANOVAs were performed, FDR correction procedures were applied to the raw p-values to correct for the multiple comparison within each condition (deep v.s. shallow). There was no difference among different word embedding models, but there was a significant difference among the computer vision models in most of the ROIs for each condition. MobileNetV2, which was considered the simplest model within the computer vision models due to its smaller set of parameters, performed the best in explaining the variance in most of the ROIs for each condition.

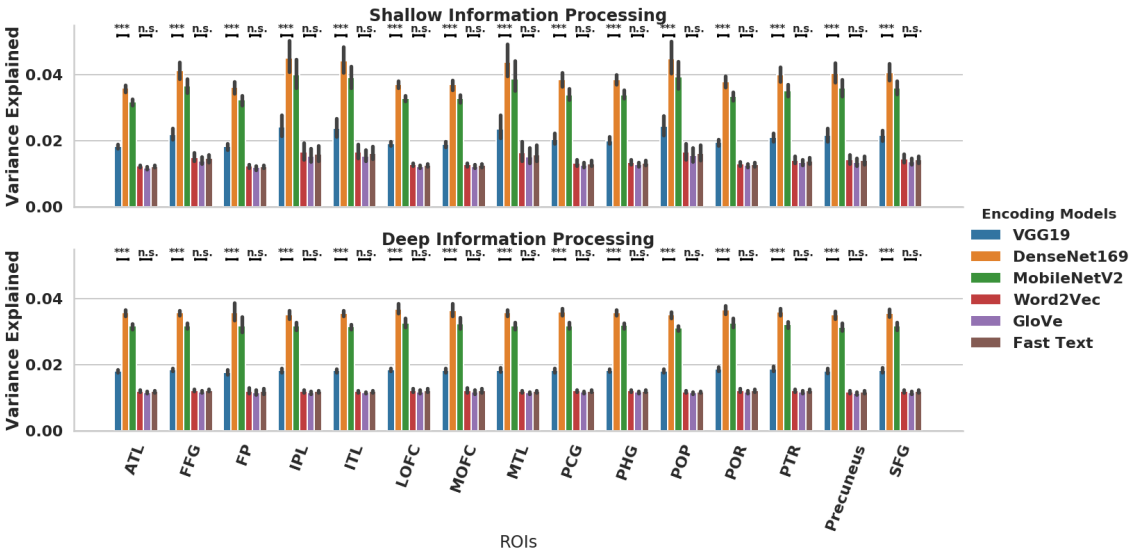

Figure S7: Average Variance Explained by each of the Word Embedding and Computer Vision Models. No fine-tuning was performed for the computer vision models.

We then computed the difference between each of the computer vision and word embedding models within each ROI and condition, in order to assess whether word embeddings or computer vision models were better.

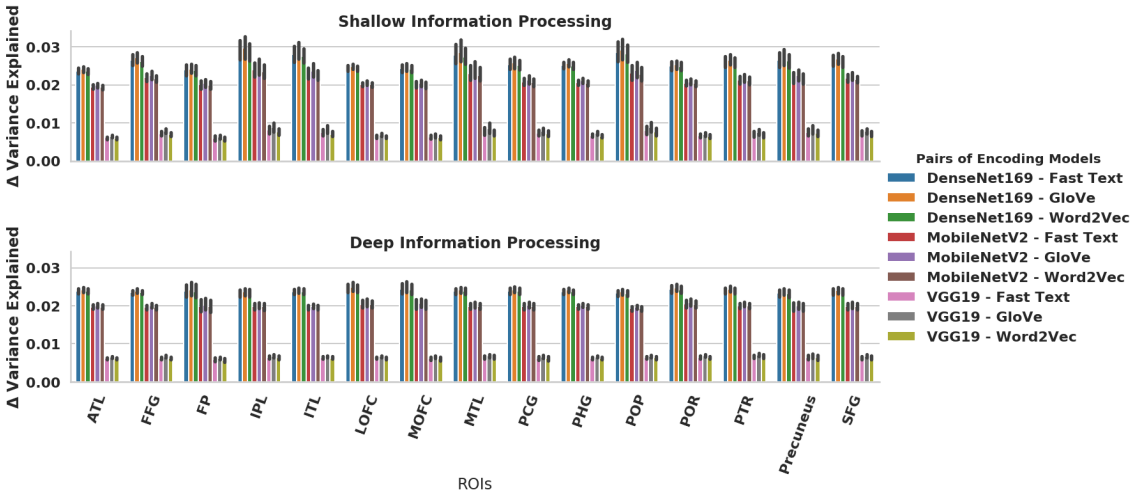

Figure S8: Differences between a Word Embedding and a Computer Vision model in variance explained. No fine-tuning was performed for the computer vision models.

One-sample t-tests against zero for each pair were conducted with FDR correction. All the computer vision models performed better than any of the word embedding models (see Figure S8).

Then, we computed the average difference between the computer vision models and the word embedding models. We then performed paired t-tests to compare the difference between the variance explained in the deep and shallow processing conditions, using FDR correction. Figure S9 illustrates the pattern of results within each ROI. We found that the advantage of computer vision models over word embeddings was higher in the deep processing condition relative the shallow processing in some of the ROIs (see Figure S8).

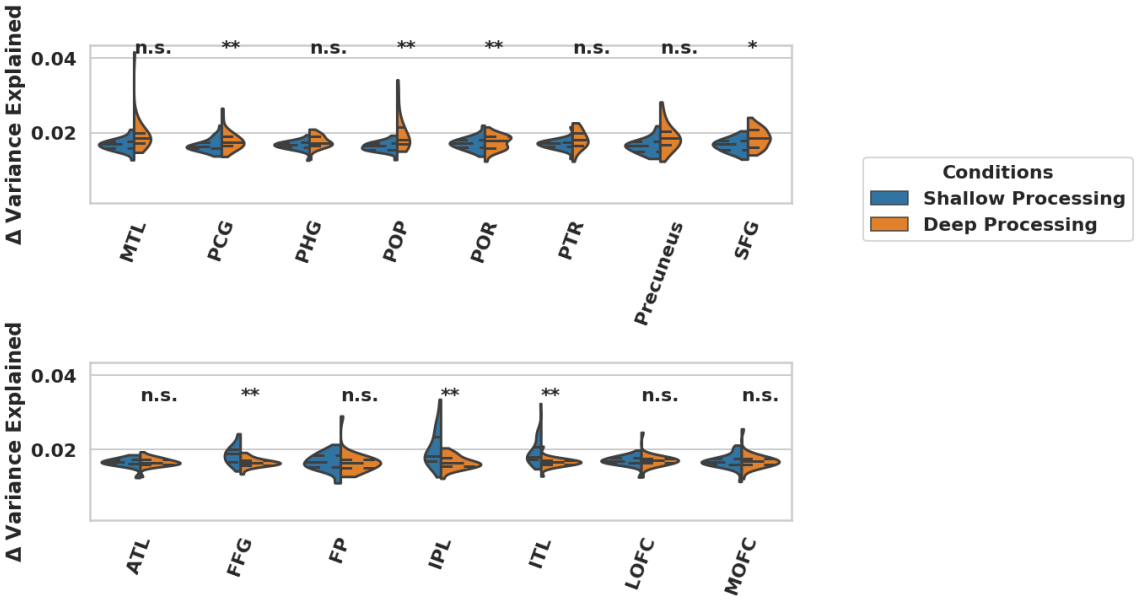

Figure S9: Overall Difference between Word Embedding and Computer Vision models per ROI. No fine-tuning was performed for the computer vision models.