Speaking, Seeing, Understanding: Correlating semantic models with conceptual representation in the brain

Luana Bulat, Stephen Clark and Ekaterina Shutova
Computer Laboratory
University of Cambridge
{luana.bulat,stephen.clark,ekaterina.shutova}@cl.cam.ac.uk

Abstract

Research in computational semantics is increasingly guided by our understanding of human semantic processing. However, semantic models are typically studied in the context of natural language processing system performance. In this paper, we present a systematic evaluation and comparison of a range of widely-used, state-of-the-art semantic models in their ability to predict patterns of conceptual representation in the human brain. Our results provide new insights both for the design of computational semantic models and for further research in cognitive neuroscience.

1 Introduction

Recent years have witnessed many breakthroughs in data-driven semantic modelling: from the log-linear skip-gram model of Mikolov et al. (2013a) to multi-modal meaning representations (Bruni et al., 2012; Kiela and Bottou, 2014; Kiela and Clark, 2015; Kiela et al., 2015a). These models boast of a higher performance accuracy in numerous semantic tasks, including modeling semantic similarity and relatedness (Silberer and Lapata, 2012), lexical entailment (Kiela et al., 2015b), analogy (Mikolov et al., 2013b) and metaphor (Shutova et al., 2016). However, less is known about the extent to which such models correlate with and reflect human conceptual representation.

Much research in the cognitive neuroscience community has been concerned with uncovering how the brain represents conceptual knowledge, by leveraging brain activation data associated with the meanings of concepts obtained during functional magnetic resonance imaging (fMRI) experiments. In the computational linguistics community, the availability of such fMRI data provides researchers with a benchmark for evaluating semantic model performance in terms of their ability to represent human semantic memory. Mitchell et al. (2008) were the first to demonstrate that distributional semantic models encode some of the patterns found in the fMRI data. Other researchers followed in their steps, evaluating traditional count-based distributional models (Devereux et al., 2010; Murphy et al., 2012), topic model-based semantic features (Pereira et al., 2013), psycholinguistic and behavioural features (Palatucci et al., 2009; Chang et al., 2010; Fernandino et al., 2015) and visual representations (Anderson et al., 2013, 2017). While all of these studies report correlation between the investigated semantic models and patterns found in the brain imaging data, their focus on individual models and the use of different datasets and prediction methods make their results difficult to compare and to integrate into a coherent evaluation landscape. The work of Murphy et al. (2012) is an exception, in that the authors systematically compare several distributional models with a range of parameters on the same brain imaging dataset. However, they focus on the traditional count-based distributional models only.

We take inspiration from the works of Mitchell et al. (2008) and Murphy et al. (2012); however, we conduct a more extensive study of the ability of different types of semantic models to predict the patterns of brain activity associated with conceptual representation. We evaluate and compare several kinds of semantic models, using different modalities and data sources: (1) traditional count-based distributional models (with word window-based and dependency-based contexts) learnt from text; (2) log-linear skip-gram models (with word window-based and dependency-based contexts); (3) behavioural models based on the free association task; (4) word representations learnt from

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1092–1102
Copenhagen, Denmark, September 7–11, 2017.

visual data; and (5) multi-modal word representations combining linguistic and visual information. Unlike previous studies, where evaluations were typically conducted using a single technique, we evaluate our models using several methods: ridge regression (Hoerl and Kennard, 1970), similarity-based encoding and similarity-based decoding (Anderson et al., 2016). Such an experimental setup allows for a comprehensive evaluation and comparison of the models.

To the best of our knowledge the dependency-based skip-gram model and the free association-based model, as well as their multimodal counterparts, have not been previously evaluated on the brain activity prediction task. Other models have been evaluated individually and have not yet been systematically compared within a single evaluation framework. Providing such a comparison, our experiments and results demonstrate that (1) visual information is a stronger predictor of brain activity than the linguistic information for concrete nouns; (2) sparse text-based models, whether dependency-based or built using linear bag-of-words context, tend to predict neural activity more accurately than dense models; (3) cognitively-motivated association-based models perform on par with or better than other linguistic models, which suggests that they provide an interesting avenue in computational semantics research.

2 Related work

The seminal work of Mitchell et al. (2008) introduced a new semantic model able to predict brain activation data associated with the meanings of concrete nouns from their corpus-harvested semantic representations. They chose a set of 25 verbs to act as semantic features in their distributional model, inspired by the importance of sensory-motor features in neural representations of concepts (Cree and McRae, 2003).

Since then, various studies have shown that distributional semantic models encode and are able to predict neural activation patterns associated with concepts (Devereux et al., 2010; Murphy et al., 2012; Pereira et al., 2013). Devereux et al. (2010) build on the work of Mitchell et al. (2008) and show that automatically acquired feature-norm like semantic representations can make equally powerful predictions about brain activity associated with the presentation of words. Pereira et al. (2013) use semantic features learnt from topic models on Wikipedia to predict neural activation patterns for unseen concepts.

Several other studies have demonstrated the fit of semantic models built from human behavioural data with regard to predicting neural activation patterns (Palatucci et al., 2009; Chang et al., 2010; Fernandino et al., 2015). Chang et al. (2010) use brain region encodings as well as detailed taxonomic encodings of McRae et al. (2005) feature norms to predict brain activation patterns using a linear regression model. They demonstrate that learned brain activity patterns can be used to decode mental states. Fernandino et al. (2015) use human elicited attribute salience scores based on five sensory-motor attributes (sound, color, visual motion, shape and manipulation) to derive fMRI brain activation patterns for concrete words, but are unsuccessful at modeling neural activation patterns for abstract words.

Recent advances in multi-modal semantics have shown that grounding semantic models in sensory modalities improves performance on a variety of tasks (Silberer and Lapata, 2012; Bruni et al., 2012; Kiela and Bottou, 2014; Bulat et al., 2016). Anderson et al. (2013) show that semantic models built from visual data correlate highly with fMRI-based brain activation patterns. Anderson et al. (2015) find that similarity in activity in the brain areas related to linguistic processing can be better predicted from text-based semantic representations, whilst image-based representations perform better at predicting similarity in the visual processing areas of the brain. In line with the dual coding theory, Anderson et al. (2017) demonstrate an advantage in decoding brain activity patterns of abstract words for text-based semantic models over the image-based ones. Contrary to previous findings, Anderson et al. (2017) find no advantage in decoding neural activity patterns associated with concrete words for image-based models.

Murphy et al. (2012) present the first study systematically comparing several text-based semantic models on the brain activity prediction task. They focus on the traditional count-based distributional models and achieve the best performance using dependency-based features. Our study is more extensive than that of Murphy et al. (2012), as we evaluate both the count-based models and the more recent skip-gram word embeddings, as well as comparing them to free association-based,

visual and multi-modal semantic representations. While Murphy and colleagues evaluate the models using one method only — linear regression, we compare predicted neural activation patterns obtained using both regression and the similarity-based encoding and decoding methods proposed by Anderson et al. (2016).

3 Brain imaging data

We use the dataset of fMRI neural activation patterns associated with the meanings of nouns, created by Mitchell et al. (2008) as described below.

3.1 fMRI experiment

Nine right-handed adults between the age of 18 and 32 (five female) participated in the study. They were presented with line drawings and noun labels for 60 concrete nouns from 12 semantic classes – animals, body parts, buildings, building parts, clothing, furniture, insects, kitchen items, tools, vegetables, vehicles and man-made objects – with five exemplars per class. The task for the participants during the scanning was to think about the properties of the noun stimuli they were presented with. The entire set of 60 stimulus words was presented six times to every participant, in a different order for each presentation.

The fMRI images were acquired on a Siemens Allegra 3.0T scanner. The initial data was corrected for slice timing, motion and linear trend; spatially normalised and resampled to 3x3x6mm^3 voxels. Only those voxels overlapping with the cortex were selected (approximately 20000 for every participant).

3.2 Voxel selection

We employ the same voxel selection procedure as Mitchell et al. (2008) for evaluating the similarity between actual fMRI images and model-predicted fMRI images. Similarity is computed by only taking into account 500 voxels with the most stable activation profile across words – with profiles compared across the six presentations. The evaluation is performed using leave-two-out cross validation. Voxel selection was performed independently for each of the cross validation folds, at training time. A voxel's stability score across the six presentations was approximated as the mean pairwise Pearson correlation between its activation profiles over the 58 training words in the cross-validation fold. The 500 voxels with the highest stability score were chosen.

3.3 Brain activity vectors

We evaluate our models on the data of each participant independently. Following Mitchell et al. (2008), we obtain a single fMRI image per concept (a representative image) by first computing the mean fMRI response over its six presentations, and then subtracting the mean of all 60 of these representative images from each. In the rest of this paper we will refer to these representations as *brain activity vectors*.

4 Semantic models

MITCHELL As a benchmark for all other semantic models, we use the publicly available[1] co-occurrence based semantic vectors developed in the Mitchell et al. (2008) study. The features of this semantic space are 25 sensory-motor verbs. Co-occurrence statistics were collected using a window size of 5 words either side of the target word, on a trillion-word corpus provided by Google.

4.1 Text-based semantic models

We train a variety[2] of context-counting and context-predicting text-based semantic models on the January 2016 dump of Wikipedia, which was tokenised using the Stanford NLP tools[3], lemmatised with the Morpha lemmatiser (Minnen et al., 2001), and parsed with the C&C parser (Clark and Curran, 2007).

DISTRIB We obtain count-based distributional semantic models, using the top 10K most frequent lemmatised words in the corpus (excluding stopwords) as contexts. The context window is defined as sentence boundaries. Counts are re-weighted using positive pointwise mutual information (PPMI) and vectors are L2-normalised.

SVD300 We also construct 300-dimensional dense semantic representations by applying singular value decomposition (SVD) (Deerwester et al., 1990) to DISTRIB.

[1] https://www.cs.cmu.edu/afs/cs/project/theo-73/www/science2008/data.html

[2] We have experimented with different parameter settings for each type of language-based semantic space (e.g. size of the vectors, number of iterations when learning the embeddings etc.) and found that the reported vectors with "standard' settings perform the best (or do not get significantly outperformed).

[3] https://nlp.stanford.edu/software/index.shtml

DEPS Following Murphy et al. (2012), who find that dependency-based semantic vectors perform best on a neurosemantic decoding task, we also include such a semantic space in our comparison. Vector representations are created by leveraging the dependency relations output by the C&C parser (Clark and Curran, 2007) as features. We use both the incoming and outgoing dependency relations as features; for example, given the dependency relation (RUN, DOBJ, MARATHON) we extract the tuple (DOBJ, MARATHON) as a feature for RUN and (!DOBJ, RUN) as a feature for MARATHON. The top 10K most frequent dependency features are used as contexts and counts are re-weighted using PPMI.

DEPS-SVD300 We also obtain 300-dimensional dense dependency-based semantic representations by applying SVD to DEPS.

EMBED-BOW We train 300-dimensional embeddings using the standard log-linear skipgram model with negative sampling of Mikolov et al. (2013a). The embeddings were trained using linear bag-of-words contexts, with the window defined as $k = 2$ (**EMBED-BOW2**) or $k = 5$ (**EMBED-BOW5**) words either side of the target word. We use 10 negative samples per word-context pair and 15 iterations over the corpus.

EMBED-DEPS In addition to the embeddings trained with linear bag-of-words contexts, we also obtain 300-dimensional dependency-based word embeddings using the Levy and Goldberg (2014) implementation of the generalised skip-gram with arbitrary contexts model. Using both incoming and outgoing dependency relations output by the C&C parser, we create word-context pairs using all words and contexts occurring more than 400 times in the corpus. This resulted in a vocabulary of about 92,000 words, with over 250,000 distinct syntactic contexts. We use 10 negative samples per word-context pair and 15 iterations over the corpus.

4.2 Association-based semantic model

Free word association datasets (Nelson et al., 2004; De Deyne et al., 2016) represent a rich source of semantic information and have been successfully used in NLP, including research on semantic memory (Steyvers et al., 2004) and multimodal semantics (Hill and Korhonen, 2014). Recent studies have shown the superiority of semantic models built using data collected from *multiple-response* free association tasks — where subjects are asked to list multiple associative cues for every target word rather than a single association — over the models built from single-response ones (De Deyne et al., 2013). Moreover, such association-based semantic models have been shown to outperform current state-of-the-art text-based language models on concept relatedness and similarity judgments (De Deyne et al., 2016).

We make use of the word association dataset collected as part of the Small World of Words[4] project, where more than 100K fluent English speakers were asked to list three associations for each target word. The dataset contains multiple-response association data for over 10K words. We use a subset of this dataset, where all target words have at least 50 primary, 50 secondary and 50 tertiary responses and all responses also appear as normed target words[5].

ASSOC We construct a count-based semantic model of word associations (henceforth ASSOC) similarly to a count-based distributional model: the responses are treated as semantic features, and counts are replaced by the sum of primary, secondary and tertiary association frequencies between the target word and the responses. Counts are re-weighted using PPMI and vectors are L2-normalised. The association-based representations obtained for the 60 target words in the Mitchell et al. (2008) dataset under this model are 9854-dimensional.

4.3 Image-based semantic model

We also build state-of-the-art deep visual semantic representations (henceforth VISUAL) for the 60 concepts in the Mitchell et al. (2008) dataset. Following previous work in multi-modal semantics (Bergsma and Goebel, 2011; Kiela and Bottou, 2014) and the findings of a recent study of system architectures and data sources for constructing visual representations (Kiela et al., 2016), we retrieve 10 images per concept from Google Images. We use the MMFeat toolkit[6] (Kiela, 2016) to build our image representations. We extract the 4096-dimensional pre-softmax layer from a for-

[4] https://smallworldofwords.org/

[5] Total of 9854 words (appearing as both target and responses) and 1092251 association pairs

[6] https://github.com/douwekiela/mmfeat

ward pass through a convolutional neural network (Krizhevsky et al., 2012), which has been pre-trained on the ImageNet classification task using Caffe (Jia et al., 2014). We obtain the visual representation for a given concept by taking the mean of the 10 resulting image representations.

4.4 Multi-modal semantic models

We also included multi-modal semantic spaces in our analysis, as these are currently widely used in NLP and have been previously shown to achieve the best performance at predicting conceptual encodings in the brain (Anderson et al., 2015). Multi-modal semantic spaces are constructed by combining the visual (VISUAL) and respective linguistic (e.g. MITCHELL, DISTRIB, DEPS) or association-based (ASSOC) representations into a multi-modal representation by concatenating their respective L2-normalized vectors.

5 Methods

In this study, we use two different ways of analysing the correlation between the semantic models described in Section 4 and the fMRI brain activation patterns used as a proxy for human conceptual representation. First, we compare these semantic models in their predictive power, by looking at how well they can synthesise, i.e. predict, brain activation patterns for unseen concepts (Section 5.1). Secondly, we look at how well they are able to decode neural activation patterns by measuring their success at predicting the stimulus that produced an unlabeled (unseen) fMRI pattern (Section 5.2).

5.1 Predicting brain activity patterns

The brain activity prediction task has been used in previous NLP research as a method of evaluating different semantic models in their ability to model conceptual representation. Most of these studies learn a mapping function between the semantic model of choice and the fMRI neural activity patterns using regression techniques (Mitchell et al., 2008; Devereux et al., 2010; Murphy et al., 2012). Recent work by Anderson et al. (2016) introduce a new method for synthesising fMRI activity patterns through similarity-based encoding that does not require model fitting. We compare the prediction performance of the semantic models detailed in Section 4 by implementing both a regression-based model and the similarity-based encoding algorithm of Anderson et al. (2016).

Regression-based learning Following previous work (Mitchell et al., 2008; Devereux et al., 2010; Murphy et al., 2012), for every participant, we learn a mapping function between semantic model features and brain activation vectors using linear regression. The learned weights are used to make predictions about brain activation vectors associated with concepts that were not seen during training. We implement Ridge regression (Hoerl and Kennard, 1970), a multiple linear regression model that uses a least squares loss function and L2 regularisation.

Similarity-based encoding We implement the similarity-based encoding method introduced by Anderson et al. (2016). This method predicts the brain activity vector for an unseen concept by exploiting its similarity (with respect to a particular semantic model) to words for which we have observed brain activity vectors.

The first step in predicting a brain activity vector for an unseen concept is to compute its *semantic model similarity code*. This is a N-dimensional[7] vector of similarity scores — computed using Pearson's correlation — between the unseen concept and the N words for which we have brain activation vectors[8]. The predicted brain activity vector for the unseen concept is then "synthesised" by using its semantic model similarity code to weight a superposition of brain activity vectors:

$$\vec{b\prime} = \frac{1}{C}\sum_{i=1}^{N} \vec{b}_i \cdot corr(\vec{v}_i, \vec{v}_{N+1}) \quad (1)$$

Assuming the unseen word is indexed $N+1$ and $\vec{v}_j$ is the semantic model representation of word j, C is a normalisation constant defined as the sum of absolute values of elements in the semantic model similarity code:

$$C = |\sum_{i=1}^{N} corr(\vec{v}_i, \vec{v}_{N+1})| \quad (2)$$

5.2 Decoding neural activity patterns

We then evaluate our semantic models in terms of their ability to decode unseen fMRI activation patterns. The analysis in this case does not involve

[7] Assuming that we have N words for which we have both semantic model representations (e.g. DISTRIB vectors) and observed brain activation vectors.

[8] The similarities are measured w.r.t. the semantic model we use as "predictor", e.g. DISTRIB, SVD300 or VISUAL

synthesising brain activation vectors for new concepts, but predicting the correct label (stimulus) associated with a given fMRI pattern.

We implement the similarity-based decoding procedure as detailed in Anderson et al. (2016). The first step is to obtain the *semantic model similarity matrix* — by computing the semantic model similarity codes for each of the 60 concepts in the Mitchell et al. (2008) dataset (as described above) — and the *brain activity similarity matrix* — by computing brain activity similarity codes.

At test time, two of the N words are chosen for decoding, together with their respective semantic model similarity codes ($\vec{s}_i, \vec{s}_j$) and brain activity similarity codes ($\vec{a}_i, \vec{a}_j$). Next, $\vec{s'_i}$, $\vec{s'_j}$, $\vec{a'_i}$ and $\vec{a'_j}$ are obtained by removing the i-th and j-th elements in $\vec{s}_i$, $\vec{s}_j$, $\vec{a}_i$ and $\vec{a}_j$ respectively, because entries in the similarity vectors corresponding to the test words would reveal the correct answer in the matching task. We will refer to $\vec{s'_i}$ and $\vec{s'_j}$ as *reduced semantic model similarity codes*, and by analogy to $\vec{a'_i}$ and $\vec{a'_j}$ as *reduced neural similarity codes*.

Decoding is considered a success if the sum of Pearson's correlations for the correct pairings ($corr(\vec{s'_i}, \vec{a'_i}) + corr(\vec{s'_j}, \vec{a'_j})$) is higher than the sum of Pearson's correlations for the incorrect pairings ($corr(\vec{s'_i}, \vec{a'_j}) + corr(\vec{s'_j}, \vec{a'_i})$).

6 Experiments

All semantic spaces presented in Section 4 have full coverage on the Mitchell et al. (2008) dataset. All experiments detailed in this section were performed separately for every participant and evaluated using leave-two-out cross validation.

6.1 Regression experiments

We repeatedly train a regression model to fit brain activation vectors for each of the semantic spaces described in Section 4, using only 58 of the 60 available concept representations (leave-two-out cross validation). This resulted in 1770 cross-validation folds.[9] The only hyperparameter in the regression is λ, which controls the degree of regularisation. The λ hyperparameter was optimised when training each cross-validation fold, by choosing from the range 0.0001 to 100 through generalised cross validation (i.e. λ was optimised by only looking at the training items during each cross-validation fold).

During each testing round, we used the learned mapping function to construct predicted brain activation vectors for the two held out words.We evaluated each of the semantic models by computing its accuracy of matching the two predicted brain activation vectors with the two observed ones. A matching score was computed by analysing the cosine similarity between the predicted and the observed brain activation vectors. If the sum of similarities for the correct pairing was higher than the one for the incorrect pairing the matching accuracy was set to 1 for this cross-validation fold, and otherwise it was set to 0. If the model was choosing the match at random, the expected accuracy is 0.50. The similarity between two brain activation vectors was computed by only taking into account the 500 most stable voxels (during each cross-validation fold) as detailed in Section 3.2. The cross-validated accuracies for each of our semantic models are presented in Table 1, with selected results also shown in Figure 1. We only report results on two multi-modal models (VISUAL+MITCHELL and VISUAL+ASSOC), as there was no significant difference in performance between any pair of multi-modal models.

All semantic models learn to predict neural activation patterns for unseen words significantly above chance level. Association-based semantic models (ASSOC) significantly[10] outperform all dense semantic representations (whether embedding-based or SVD-reduced), with $p < 0.05$. Sparse text-based representations with linear context (DISTRIB and DEPS) significantly outperform some dense semantic representations. However, no dense semantic models significantly outperform DISTRIB and DEPS. There is no significant difference between the performance of ASSOC, DISTRIB and DEPS. Contrary to the findings of Murphy et al. (2012), we do not find any advantage in predicting brain activation patterns from dependency-based text models.

Both VISUAL and multi-modal models significantly outperform text-based models overall ($p < 0.05$), excepting MITCHELL with $p < 0.11$ when comparing to VISUAL and $p < 0.09$ when comparing against multi-modal semantic models. These results support previous findings regarding the importance of grounding semantic models in perceptual input. These grounded semantic models per-

[9] There are (60 choose 2) ways to choose two test items from the 60 Mitchell et al. (2008) concepts.

[10] We used (pairwise) paired t-tests to judge the statistical significance of the difference in performance between any two models within the same experiment.

MODEL	P1	P2	P3	P4	P5	P6	P7	P8	P9	*mean*
MITCHELL	0.78	0.72	0.71	0.75	**0.76**	0.56	0.71	**0.63**	0.63	0.70
DISTRIB	0.85	0.67	0.73	0.84	0.72	0.55	0.70	0.54	0.69	0.70
SVD300	0.85	0.65	0.68	0.77	0.67	0.53	0.66	0.52	0.62	0.66
DEPS	0.85	0.70	0.77	0.86	0.74	0.40	0.70	0.59	**0.72**	0.70
DEPS-SVD300	0.80	0.68	0.74	0.81	0.70	0.32	0.68	0.61	0.66	0.67
EMBED-BOW2	0.85	0.65	0.70	0.78	0.64	0.55	0.60	0.57	0.65	0.66
EMBED-BOW5	0.83	0.62	0.72	0.74	0.66	0.56	0.70	0.56	0.58	0.66
EMBED-DEPS	0.82	0.60	0.67	0.81	0.67	0.49	0.63	0.62	**0.72**	0.67
ASSOC	**0.90**	0.65	0.78	0.87	0.74	0.51	0.75	0.60	0.67	0.72
VISUAL	**0.90**	**0.78**	**0.85**	**0.88**	0.69	0.56	0.75	0.57	0.69	0.74
VISUAL+ASSOC	**0.90**	**0.78**	0.84	0.86	0.70	**0.58**	**0.76**	0.56	0.70	0.74
VISUAL+MITCHELL	**0.90**	**0.78**	0.84	0.86	0.70	**0.58**	**0.76**	0.56	0.70	0.74

Table 1: *Regression results*. Cross-validated accuracies for models trained on participants P1 through P9, together with mean over participants.

form as well as models that encode mental representations through associations (ASSOC). There is no significant advantage for multi-modal models over VISUAL.

6.2 Similarity-based encoding experiments

We also compare performance of the semantic models when the predicted brain activation vector is computed using the Anderson et al. (2016) similarity based encoding method. We use a leave-two-out cross validation strategy, to match previous work and our experiments detailed in Section 6.1. The similarity-based encoding approach does not require any mapping function to be learned, hence is a robust and fast way to obtain synthesised brain activation vectors for unseen words.

During each cross-validation fold, semantic model similarity codes of the two test words were computed using the procedure outlined in Section 5.1. Predicted brain activation vectors were then synthesised for the two test words by weighting a superposition of brain activity vectors using their semantic model similarity codes. The matching score for each of the cross-validation folds was computed in the same way as in the case of the regression model (Section 6.1). The only difference was that we measured the similarity between the two brain activation vectors using Pearson's correlation coefficient, following Anderson et al. (2016). As in the previous experiment, the expected chance performance of this method is 0.5. The cross-validated accuracies for each of our semantic models are shown in Table 2, with selected results also shown in Figure 1.

All semantic models perform significantly above chance level. As in the case of the regression experiments, there is a clear advantage in synthesising brain activation vectors for visually grounded models (VISUAL and multi-modal models) over the language-based ones (this time including MITCHELL), as well as ASSOC. When looking at the performance of the text-based models in general, there is no difference in performance when comparing context-predicting models to count-based ones, or sparse semantic models to dense ones.

6.3 Brain activation pattern decoding

In the similarity-based decoding experiments, we assess the ability of semantic models to identify the correct stimulus for a given brain activation pattern, using the same leave-two-out cross-validation strategy. At test time, we obtain the reduced semantic model similarity codes and the reduced neural similarity codes for the two test items as described in Section 5.2. It is important to note that these similarity code vectors do not contain any information about the true labeling, since entries corresponding to the test items were removed. Decoding is considered successful if the matching score (computed as the sum of Pearson's correlations) is higher for the congruent pair than for the incorrect one. Again, the expected performance for a model decoding at random is 0.50. Table 3 shows the performance of our semantic models, with selected results also shown in Figure 1.

The performance of all semantic models in the decoding task is significantly above chance level. Grounded semantic models (visual and multi-modal) prove once again to have a significant advantage in decoding brain activation patterns over the text-based models and association-based model ($p < 0.05$). There is no signifi-

MODEL	P1	P2	P3	P4	P5	P6	P7	P8	P9	*mean*
MITCHELL	0.79	0.76	0.74	0.8	0.78	0.66	0.69	0.62	0.74	0.73
DISTRIB	0.87	0.69	0.79	0.89	0.79	0.75	0.75	0.52	0.79	0.76
SVD-300	0.89	0.72	0.79	0.90	0.79	0.74	0.78	0.56	0.83	0.78
DEPS	0.88	0.74	0.83	0.91	**0.81**	0.68	0.76	0.58	0.83	0.78
DEPS-SVD300	0.89	0.75	0.84	0.91	**0.81**	0.67	0.77	0.57	0.83	0.78
EMBED-BOW2	0.92	0.74	0.81	0.91	0.75	0.75	0.77	0.59	0.81	0.78
EMBED-BOW5	0.91	0.73	0.83	0.91	0.76	0.73	0.79	0.55	0.80	0.78
EMBED-DEPS	0.91	0.71	0.80	**0.92**	0.75	0.71	0.79	0.62	**0.85**	0.78
ASSOC	0.91	0.72	0.81	0.91	0.73	0.69	0.75	0.62	0.79	0.77
VISUAL	**0.94**	**0.82**	**0.88**	0.90	0.78	**0.76**	**0.83**	**0.65**	0.82	0.82
VISUAL+ASSOC	**0.94**	**0.82**	**0.88**	0.90	0.79	**0.76**	**0.83**	**0.65**	0.83	0.82
VISUAL+MITCHELL	**0.94**	**0.82**	**0.88**	0.90	0.78	**0.76**	**0.83**	**0.65**	0.82	0.82

Table 2: *Similarity based encoding results*: Cross-validated accuracies for models trained on participants P1 through P9, together with mean over participants.

MODEL	P1	P2	P3	P4	P5	P6	P7	P8	P9	*mean*
MITCHELL	0.80	0.76	0.75	0.82	0.77	0.7	0.71	**0.65**	0.75	0.75
DISTRIB	0.87	0.70	0.79	0.90	**0.80**	0.76	0.77	0.58	0.80	0.77
SVD300	0.88	0.73	0.79	0.89	**0.80**	0.76	0.79	0.61	**0.85**	0.79
DEPS	0.88	0.75	0.84	0.91	**0.80**	0.70	0.78	0.61	0.84	0.79
DEPS-SVD300	0.89	0.76	0.84	0.90	0.81	0.70	0.79	0.61	**0.85**	0.79
EMBED-BOW2	0.91	0.75	0.81	0.90	0.76	0.76	0.78	0.60	0.82	0.79
EMBED-BOW5	0.91	0.74	0.83	0.91	0.77	0.75	0.80	0.58	0.82	0.79
EMBED-DEPS	0.91	0.71	0.80	**0.92**	0.75	0.71	0.79	0.62	**0.85**	0.78
ASSOC	0.90	0.73	0.79	0.90	0.73	0.69	0.76	0.63	0.81	0.77
VISUAL	**0.94**	**0.83**	**0.89**	0.90	0.79	**0.78**	**0.84**	**0.65**	0.84	0.83
VISUAL+ASSOC	**0.94**	**0.83**	**0.89**	0.90	0.79	**0.78**	**0.84**	**0.65**	0.84	0.83
VISUAL+MITCHELL	**0.94**	**0.83**	**0.89**	0.90	0.79	**0.78**	**0.84**	**0.65**	0.84	0.83

Table 3: *Similarity based decoding results*: Cross-validated accuracies for models trained on participants P1 through P9, together with mean over participants.

cant difference in performance between any of the multi-modal models and VISUAL.

6.4 Discriminating between words in the same semantic class

Following Mitchell et al. (2008), we also compare the models in their ability to make accurate predictions when the two test words are exemplars of the same semantic category[11]. This formulation of the task is more difficult, since items in the same semantic class (e.g. *dog* and *cat*) are more similar than items from different semantic classes (e.g. *eye* and *desk*).

In order to measure the performance of our models in this task, we recompute the cross-validated accuracies for all three experiments (regression-based learning, encoding and decoding) by only taking into account the performance on the 120 cross-validation folds where the test items share the same semantic class. The results across models and experiments show very similar trends as the ones computed using all 1770 cross-validation folds. The majority of the models still perform above chance level, but as expected they perform worse than when evaluated using the entire dataset. Visually-grounded models still perform the best in all three experiments (mean performance across participants for multi-modal models in all three tasks is in the [0.61-0.63] range).

[11]The 60 concepts are exemplars of 12 semantic classes.

7 Conclusion and future work

We presented the first systematic comparison of a range of widely-used, state-of-the-art semantic models in their ability to predict patterns of conceptual representation in the human brain. Firstly, we demonstrated that visual information is a stronger predictor of brain activity than linguistic information for concrete nouns. These findings provide further support to the existing hypotheses about the interplay of linguistic, conceptual and perceptual systems in the human brain (Barsalou, 2008). These results also resonate with the success of the rapidly growing field of multimodal semantics (Kiela et al., 2016).

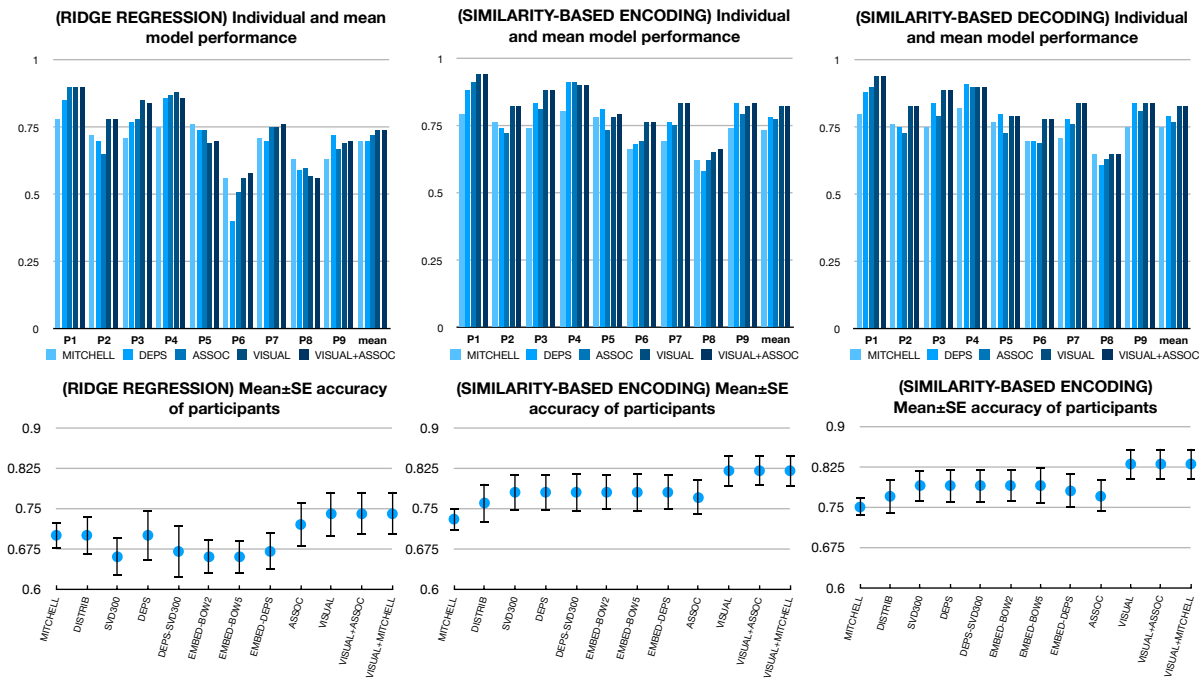

Figure 1: (*TOP*) Comparison of individual and mean model performance for five selected models (MITCHELL, DEPS, ASSOC, VISUAL, VISUAL+ASSOC), using results in Table 1 (*Ridge regression*), Table 2 (*Similarity-based encoding*) and Table 3 (*Similarity-based decoding*). (*BOTTOM*) Mean±SE accuracy of participants for all models.

Secondly, our results suggest that sparse text-based models, whether dependency-based or built using linear bag-of-words context, predict neural activity more accurately than dense models. We also show that the structure of the text-based semantic model (sparse vs. dense) has more influence on the performance than the type of information used to construct the context (linear bag-of-words vs. dependency-based).

Finally, we found that cognitively-motivated association-based models perform on par with or better than other linguistic models. These results are in line with the previous findings of behavioural research suggesting that humans represent the meanings of concepts through association with other concepts (Barsalou et al., 2008) which in turn endorses the association-based semantic models as a promising direction in computational semantics research.

An interesting avenue for future work would be to investigate the variance of results amongst individual participants (Figure 1). Previous studies that use fMRI data always report variation across participants (Devereux et al., 2010; Anderson et al., 2017) and most often attribute it to head motion. However, understanding how individual variations in participants can impact modeling decisions would be of great value to the computational semantics community.

Acknowledgments

Luana Bulat is supported by an EPSRC Doctoral Training Grant and ERC Proof of Concept Grant GroundForce (693579). Stephen Clark is supported by ERC Starting Grant DisCoTex (306920) and ERC Proof of Concept Grant GroundForce (693579). Ekaterina Shutova is supported by the Leverhulme Trust Early Career Fellowship. We thank the anonymous reviewers for their helpful comments.

References

Andrew J Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *EMNLP*, pages 1960–1970.

Andrew J Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. 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.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Andrew J Anderson, Benjamin D Zinszer, and Rajeev DS Raizada. 2016. Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Lawrence W Barsalou, Ava Santos, W Kyle Simmons, and Christine D Wilson. 2008. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pages 245–283.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *RANLP*, pages 399–405.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Luana Bulat, Douwe Kiela, and Stephen Clark. 2016. Vision and feature norms: Improving automatic feature norm learning through cross-modal maps. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 579–588, San Diego, California. Association for Computational Linguistics.

Kai-min Kevin Chang, Tom Mitchell, and Marcel Adam Just. 2010. Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *Neuroimage: Special Issue on Multi-variate Deciding and Brain Reading*, 56:716–727.

Stephen Clark and James R Curran. 2007. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552.

George S Cree and Ken McRae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163.

Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45(2):480–498.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Barry Devereux, Colin Kelly, and Anna Korhonen. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 70–78. Association for Computational Linguistics.

Leonardo Fernandino, Colin J Humphries, Mark S Seidenberg, William L Gross, Lisa L Conant, and Jeffrey R Binder. 2015. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia*, 76:17–26.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM.

Douwe Kiela. 2016. MMFEAT: A toolkit for extracting multi-modal features. In *Proceedings of ACL*.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, volume 2014.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015a. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal. Association for Computational Linguistics.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015b. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China.

Douwe Kiela, Anita L. Verő, and Stephen Clark. 2016. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*, pages 302–308. Citeseer.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering*, 7(03):207–223.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 114–123. Association for Computational Linguistics.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, pages 1410–1418.

Francisco Pereira, Matthew Botvinick, and Greg Detre. 2013. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194:240–252.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 160–170.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics.

Mark Steyvers, Richard M Shiffrin, and Douglas L Nelson. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, pages 237–249.