Investigating Rich Feature Sources for Conceptual Representation Encoding

Lu Cao[♣], **Yulong Chen**[♠♡], **Dandan Huang**[♡], **Yue Zhang**[♡♢*]
[♣] Singapore University of Technology and Design, Singapore
[♠] Zhejiang University, China
[♡]School of Engineering, Westlake University, China
[♢]Institute of Advanced Technology, Westlake Institute for Advanced Study, China

Abstract

Functional Magnetic Resonance Imaging (fMRI) provides a means to investigate human conceptual representation in cognitive and neuroscience studies, where researchers predict the fMRI activations with elicited stimuli inputs. Previous work mainly uses a single source of features, particularly linguistic features, to predict fMRI activations. However, relatively little work has been done on investigating rich-source features for conceptual representation. In this paper, we systematically compare the linguistic, visual as well as auditory input features in conceptual representation, and further introduce associative conceptual features, which are obtained from Small World of Words game, to predict fMRI activations. Our experimental results show that those rich-source features can enhance performance in predicting the fMRI activations. Our analysis indicates that information from rich sources is present in the conceptual representation of human brains. In particular, the visual feature weights the most on conceptual representation, which is consistent with the recent cognitive science study.

1 Introduction

How a simple concept is represented and organized by human brain has been of long research interest in cognitive science and natural language processing (NLP) (Ishai et al., 1999; Martin, 2007; Fernandino et al., 2016). The rise of brain imaging methods such as fMRI technology has now made it feasible to investigate conceptual representation within human brain. In particular, fMRI is a technique that allows for the visualization of neuron activity in brain regions, which has become an essential tool for analyzing the neural correlates of brain activity in recent decades (Mitchell et al., 2004; Mitchell et al., 2008; Pereira et al., 2009; Pereira et al., 2011; Just et al., 2010).

Neuroscientists have shown that distinct patterns of neural activation are associated with both encoding and decoding the concepts of different semantic categories in brains. Mitchell et al. (2008) first introduced the task of predicting fMRI activation and proposed a featured-based model which takes a semantic representation of a single noun to predict the fMRI activation elicited by that noun. Subsequent studies (Pereira et al., 2018) introduced distributed based methods to build correlations between distributed semantic representations and patterns of neural activation. However, previous work mostly focuses on a single source of input features, e.g. count-based word vectors (Devereux et al., 2010; Murphy et al., 2012; Pereira et al., 2013; Pereira et al., 2018) to explore the in brain encoding process, which builds correlation between neural signals and distributed representation, and thus can be useful for better understanding both the brain and the word representation. But there has been little work systematically investigating the effect of different modalities on predicting fMRI activations.

We address this limitation by empirically investigating two forms of rich source features: multimodal features and associative conceptual feature. First, we systematically compare input features that come from linguistic, visual and auditory sources into fMRI activation encoding. To investigate the influence of each source of information in the brain conceptual representation, we build and evaluate a multimodal

*Corresponding author.

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 12–22
Barcelona, Spain (Online), December 12, 2020

Reference	Stimuli	Presentation mode	Subj.
Mitchell et al. (2008)	60 concrete nouns	Word, Image	9
Pereira et al. (2018)	180 words	Word cloud, Sentence, Image	16

Table 1: fMRI datasets for language-brain encoding.

Categories	Words
animal	bear, cat, dog, horse, cow
vegetable	lettuce, carrot, corn, tomato, celery
body part	eye, arm, foot, leg, hand
man-made	telephone, key, bell, watch, refrigerator
building	igloo, barn, house, apartment, church
kitchen	spoon, bottle, cup, knife, glass
vehicle	truck, car, train, bicycle, airplane
clothing	dress, skirt, coat, pants, shirt
furniture	chair, dresser, desk, bed, table
build part	door, chimney, closet, arch, window
insect	fly, bee, butterfly, ant, beetle
tool	hammer, chisel, screwdriver, saw, pliers

Table 2: 60 nouns, organized by categories (Mitchell et al., 2008).

conceptual representation model with different modal input features and their combinations. Second, we investigate associative thinking of related concepts. We assume that associative thinking for concepts has individual difference, and it is insufficient to reflect such differences via distributed semantics representation. To verify this assumption, we propose an associative conceptual embedding that predicts brain activity by using associative conceptual words other than the concept presented to the subjects when collecting the brain activity data.

Experiments of multi-sense representation show that not only linguistic features, but also visual and auditory features, can be used to predict fMRI activations. It demonstrates that multimodal information is present in the conceptual representation in human brains, and we also observe that the weights of various modalities in brain conceptual representation are unequal. In particular, we find that performances of visual feature grounded multimodal models are overall improved compared with unimodal models, while the performances of auditory feature grounded models are not consistently improved. This observation leads to a conclusion that the visual information weights the most in brain conceptual representations. In addition, experiments of associative conceptual representation show that the associative conceptual words, which though are distinct in distributed semantic vector space, are related in conceptual representation in human brains.

2 Related Work

Previous studies on conceptual representation mainly focus on correlation between words and corresponding fMRI activations, including feature based methods and distributed representation based methods. Seminal work of Mitchell et al. (2008) pioneered the use of corpus-derived word representations to predict brain activation data associated with the meaning of nouns. This feature based method selected 25 verbs (*i.e., 'see', 'say', 'taste'.*), and calculated the co-occurrence frequency of the noun with each of 25 verbs. In this regard, a noun word is encoded into 25 sensor-motor features. Subsequent work including Jelodar et al. (2010) used WordNet (Miller, 1995) to compute the values of the features. Obviously, such feature based methods are constrained by corpora, and only focus on linguistic unimodal.

Pereira et al. (2013) proposed a distributed semantics based method using features learnt form Wikipeida to predict neural activations for unseen concepts. Since then, various studies have shown

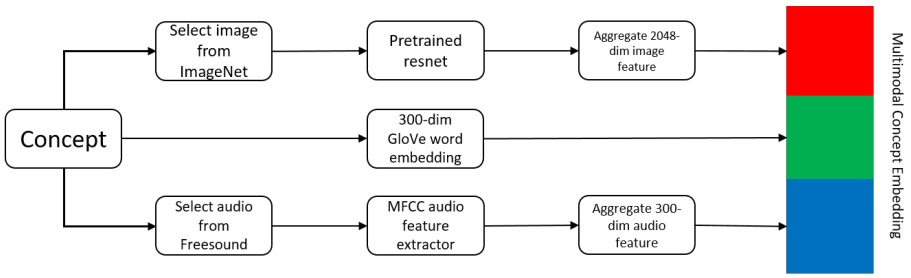


Figure 1: Compute multimodal embeddings.

that distributed semantic representations have correlations with brain concept representation (Devereux et al., 2010; Murphy et al., 2012; Pereira et al., 2013; Pereira et al., 2018; Bulat et al., 2017). However, though these methods outperform the feature based methods, they still ignore the fact that the information in the real world comes as different modalities. In contrast to their work, we investigate the human conceptual representation mechanism via evaluating the effects of multimodal features rather than only unimodal linguistic feature.

More closely related to our work, Bulat et al. (2017) presented a systematic evaluation and comparison of unimodal and multimodal semantic models in their ability to predict patterns of conceptual representation in the human brain. However, they only focused on the model level, contrasting unimodal representations and multimodal representations that involve linguistic and visual signals, but not the effect of each modality. While little previous work studied the influence of each source of information in the brain conceptual representation, our study is more extensive by evaluating multiple modalities data and their combinations. To our knowledge, we are the first to report auditory data in exploring human conceptual representations. More vitally, we explore their importance in concrete noun representations. Different from all work above, we are also the first to introduce associative conceptual words as input features to human conceptual representation.

3 Task: Predicting the fMRI Activation

The task is to predict the corresponding fMRI activations with elicited stimuli. The encoder operates by predicting fMRI activation given feature vectors. Each dimension (voxel) of fMRI activation is predicted by using a separate ridge regression estimator. More formally, given the matrix X and the matrix Z, we learn regression coefficients b and b_0 that minimize

$$\|Xb + b_0 - z\|^2 + \alpha\|b\|^2 \tag{1}$$

for each column of z of Z matrix. X is the semantic matrix, the dimension is the number of words (training set) by the dimension of semantic vector (300 for GloVe); and Z is the corresponding fMRI activation matrix, the dimension is the number of fMRI activation by the imaging dimension (amount of selected voxel, 500 for Mitchell et al. (2008) dataset and 5000 for Pereira et al. (2018) dataset).

We investigate three types of multi-sense inputs, namely, linguistic, visual and auditory sources. And further we use associative conceptual input, namely, the associative conceptual words which is obtained from Small World of Word game. In the next two sections, we will introduce how to obtain multi-sense representations and associative conceptual representations.

4 Multi-Sense Representations

Following Bruni et al. (2014) and Kiela and Bottou (2014), we construct multimodal semantic representation vector, V_m, by concatenating the linguistic, visual and auditory representations as shown in Figure 1:

$$V_m = V_{linguistic} \| V_{visual} \| V_{auditory}, \tag{2}$$

where $\|$ is the concatenation operator.

4.1 Linguistic Representations

The linguistic representation can be a dense vector that represents a word associated with a concept. Distributed word representations have been applied to statistical language modeling with considerable success (Bengio et al., 2003). This idea has enabled a substantial amount of progress in a wide range of NLP task, and was also shown useful for brain conceptual representation (Devereux et al., 2010; Murphy et al., 2012). The approach is based on the distributional hypothesis (Firth, 1957; Harris, 1954) which assumes that words with similar contexts tend to have similar semantic meaning. The intuition underlying the model is ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning. GloVe (Pennington et al., 2014) provides multiple versions of pre-trained word embeddings. In this paper, we use a 300-dimensional version of GloVe, which trained on a corpus consisting of Wikipedia 2014 and Gigaword 5.

4.2 Visual Representations

Visual representation is used to represent an image associated with a concept in a dense vector. Our approach to constructing the visual representations component is to utilize a collection of images associated with words representing a particular concept. For example, given a stimulus *'carrot'*, the associated images are a collection of *'carrot'* images that we retrieve from the dataset. In our implementation, we use Deep Residual Network (ResNet) (He et al., 2016) to produce the image feature map.

ResNet is widely used in image recognition as it is a deep neural network with many convolution layers stack together and can extract rich image features. The network is pre-trained on ImageNet (Deng et al., 2009), one of the largest image databases. Then, we chop the last layer of the network and use the remaining part as the feature extractor to compute the 2048-dimensional feature vector for each image. To represent a particular concept, we extract the image features of all images belong to that concept. Then, we directly compute the average of all image features as the visual representation.

4.3 Auditory Representations

Auditory representation is a dense vector used to present the acoustic properties of a concept. For example, given the concept *'key'*, correlated sounds are keys hitting or rubbing together; and for *'hand'*, correlated sounds can be applause. For the auditory representations, we retrieve 3 to 100 audios from Freesound (Font et al., 2013) for each concept. To generate the auditory representation for each noun, we first obtain Mel-scale Frequency Cepstral Coefficients (MFCCs) (O'Shaughnessy, 1987) features of each audio and then quantize the features into a bag of audio words (BoAW) (Foote, 1997) representations. MFCCs are commonly used as features in speech recognition, information retrieval, and music analysis. After obtaining a BoAW set, we take the mean of each BoAW as the auditory representation. In this paper, we use MMFeat (Kiela, 2016) to generate 300-dimensional auditory representations. The code is available at `https://github.com/douwekiela/mmfeat`.

5 Associative Conceptual Representation

Associative conceptual representation is a dense vector obtained from the associative conceptual words that are produced by humans in a game scene, and it is used to presented human's associative thinking related a concept. To investigate that whether associative thinking can be reflected in the fMRI activation, we fuse the word vectors linearly and use it as our associative conceptual representations. The linear fusion is represented as:

$$V_m = V_{stimuli} \| V_{associate}, \tag{3}$$

where $\|$ is the concatenation operator.

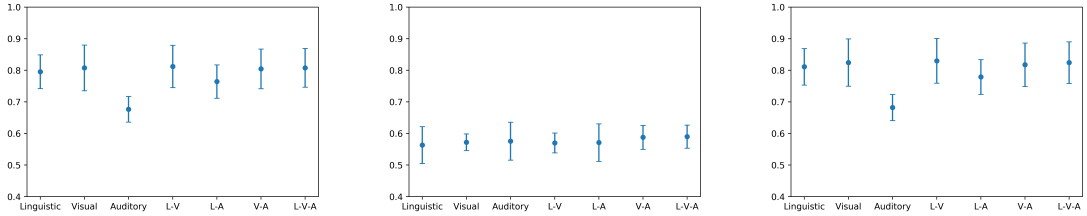
6 Experiments

We apply three sources of features to predict fMRI activations with unimodal model and multimodal models, and compare their performances. Further, we compare the performances of models with irrelevant words and associative conceptual words as inputs respectively.

	Linguistic		Visual		Auditory		L+V[1]		L+A[2]		V+A[3]		L+V+A[4]	
	W/I	B/W	W/I	B/W	W/I	B/W	W/I	B/W	W/I	B/W	W/I	B/W	W/I	B/W
P1	0.49	0.91	0.61	0.95	0.68	0.76	0.58	0.95	0.62	0.87	0.63	0.92	0.64	0.92
P2	0.47	0.75	0.60	0.81	0.55	0.67	0.52	0.80	0.54	0.70	0.54	0.77	0.52	0.78
P3	0.68	0.85	0.57	0.83	0.57	0.66	0.61	0.85	0.59	0.78	0.62	0.83	0.63	0.84
P4	0.55	0.89	0.57	0.92	0.51	0.71	0.57	0.92	0.51	0.85	0.58	0.91	0.55	0.92
P5	0.58	0.79	0.58	0.80	0.53	0.64	0.61	0.81	0.48	0.75	0.54	0.79	0.58	0.80
P6	0.55	0.77	0.59	0.80	0.53	0.65	0.55	0.80	0.57	0.77	0.62	0.79	0.60	0.78
P7	0.57	0.75	0.54	0.81	0.68	0.73	0.53	0.81	0.68	0.80	0.64	0.83	0.61	0.83
P8	0.61	0.76	0.52	0.67	0.54	0.63	0.56	0.69	0.62	0.70	0.55	0.68	0.61	0.70
P9	0.57	0.83	0.57	0.83	0.59	0.69	0.60	0.84	0.53	0.79	0.57	0.84	0.57	0.85
Mean	0.56	0.81	0.57	0.82	0.58	0.68	0.57	0.83	0.57	0.78	0.59	0.82	0.59	0.82

[1] LINGUISTIC+VISUAL [2] LINGUISTIC+AUDITORY [3] VISUAL+AUDITORY [4] LINGUISTIC+VISUAL+AUDITORY
W/I WITHIN CATEGORY B/W BETWEEN CATEGORY

Table 3: Accuracies of within and between-category examples for all participants (Pi). Within-category refers to stimuli coming from the same category (e.g. bear and cat come from the category of the animal) whereas between-category refers to stimuli coming from different categories.



(a) Mean ± SE accuracy of participants

(b) Mean ± SE accuracy of within-category

(c) Mean ± SE accuracy of between-category

Figure 2: Mean ± SE accuracies of participants for all modals of data, using results in Table 3.

6.1 Datasets

6.1.1 fMRI Datasets

In this paper, we use the fMRI activation datasets of Mitchell et al. (2008) and Pereira et al. (2018). The summary of the datasets is shown in Table 1.

Mitchell et al. (2008)'s fMRI activation dataset was collected from nine right-handed subjects (5 females and 4 males between 18 and 32 years old). Each time, every subject was presented with noun labels and line drawings of 60 concrete objects from 12 semantic categories with 5 exemplars per category and the corresponding fMRI activation was recorded. The 60 concrete nouns and categories are shown in Table 2. Each exemplar was presented six times with randomly permutation and each exemplar was presented 3 seconds followed by a 7 seconds rest period. During the exemplar presenting, subjects were required to think about the proprieties of it freely. For example, for the concept *'dog'*, the proprieties might be *'pet'*, *'fluffy'*, and *'labrador retrievers'*. It is not required to obtain consistency properties across subjects. Given an exemplar, the fMRI activation of each subject was recorded during the presenting each of the six times. In this paper, we create one representative fMRI activation for each exemplar by averaging six scans.

Pereira et al. (2018)'s fMRI activation dataset was collected from 16 subjects. Similarly to Mitchell et al. (2008), subjects were asked to think about the properties when they were presented with stimulus in form of words, pictures and sentences. But the exemplar words of Pereira et al. (2018) cover a broader semantic vector space and are more distinct in vector space. First, they applied 300-dimensional GloVe (Pennington et al., 2014) to obtain semantic vectors for all words in a vocabulary size of approximately 30,000 words (Brysbaert et al., 2013). They then utilized spectral clustering (Luxburg, 2007) to group the vectors into 180 regions, and hand-selected 180 representative words for each regions.

Categories	Linguistic	Visual	Auditory	L-V[1]	L-A[2]	V-A[3]	L-V-A[4]
man-made	27	24	27	26	25	26	27
building	38	31	38	31	33	32	30
build part	56	64	40	62	48	62	61
tool	44	56	40	62	44	46	50
furniture	36	47	50	47	40	44	45
animal	22	34	36	35	32	36	33
kitchen	16	17	19	12	13	12	11
vehicle	50	40	37	42	44	37	34
insect	38	34	38	34	42	33	36
vegetable	32	33	49	30	48	42	37
body part	58	30	48	33	50	28	32
clothing	44	52	39	51	44	45	47

[1] LINGUISTIC+VISUAL [2] LINGUISTIC+AUDITORY [3] VISUAL+AUDITORY [4] LINGUISTIC+VISUAL+AUDITORY

MOST ERROR LEAST ERROR

Table 4: Selected within-category error statistics.

6.1.2 Multi-Sense Dataset

We obtain **linguistic** features from the GloVe (Pennington et al., 2014), which is trained on Wikipedia 2014 and Gigaword 5. For **visual** features, We retrieve 300 to 1500 images for each concept noun from ImageNet, except human body word: *'hand'*, *'foot'*, *'arm'*, *'leg'* and *'eye'*, which are not included in the ImageNet. Thus, we retrieve these images from Google Image (Afifi, 2017). The retrieved images from ImageNet and Google are combined together as the image dataset for visual feature extraction. For **auditory** features, we use the Freesound dataset (Font et al., 2013), which is a huge collaborative database of audio snippets, samples, recordings, and bleeps.

6.1.3 Associative Word Dataset

In this paper, we use Small World of Words (SWW) (De Deyne et al., 2018) as the word association data source. SWW is a mental dictionary or lexicon in the major languages of the world. It collects associative words by inviting participants globally to play an online game of word associations[1]. The game is simple and easy to play: given a list of 18 cue words, participants are asked to give first three words that come to mind. It counts and demonstrates the human level word associations. For example, top ten forward associations of the cue word *'machine'* are *'robot'*, *'computer'*, *'engine'*, *'metal'*, *'gun'*, *'work'*, *'car'*, *'washing'*, *'factory'*, *'sewing'*; and top ten backward associations of it are *'slot'*, *'fax'*, *'pinball'*, *'mechanism'*, *'sewing'*, *'washing'*, *'xerox'*, *'contraption'*, *'cog'*, *'copier'*. Here, forward association refers to the words will come to mind when participants see the cue word *'machine'*; and backward association refers to the word *'machine'* will come to mind when participants view other cue words. And their rankings indicate the average order of the word that participants think of in the SWW game.

In our paper, we use 60 concrete words from Mitchell et al. (2008) and choose 175 words from Pereira et al. (2018) (we discard 5 words: *'argumentatively'*, *'deliberately'*, *'emotionally'*, *'tried'*, *'willingly'*, which do not present in the associative words data source) as the cue words.

6.2 Training

As mentioned in Section **Task: Predicting the fMRI Activation**, the task is to predict the fMRI activations. Following Mitchell et al. (2008), we train the encoder consisting of several estimators (500 for Mitchell et al. (2008) and 5000 for Pereira et al. (2018)). Each estimator predicts a fMRI activation value of a specific position in the brain. The estimator is trained by ridge regression where the loss function is the linear least squares function and is regularized by the L_2-norm (Eq. 1). The regularization strength α is chosen by cross-validation.

[1] https://smallworldofwords.org/en

6.3 Evaluation

We evaluate each encoder's performance by following the strategy of Mitchell et al. (2008) and Pereira et al. (2018). For each possible pair of fMRI activation, we compute the cosine similarity between predicted and actual one. If the predicted fMRI activation is more similar to its actual one than the alternative, we deem the classification correct. For the data of Mitchell et al. (2008), each encoder is trained on 58 words and tested on the 2 left out words. The training and testing procedure iterates 1770 times. For the data of Pereira et al. (2018), each encoder is trained within a cross-validation procedure. In each fold, the parameters are learned from 165 word vectors, and predicted fMRI activation from the 10 left out words. The overall classification accuracy is the fraction of correct pairs. The match score S is calculated as:

$$S(p_1 = i_1, p_2 = i_2) = cosine(p_1, i_1) + cosine(p_2, i_2). \tag{4}$$

6.4 Results and Discussion

6.4.1 Uni- and Multi- Modal in fMRI Prediction

The cross-validated prediction accuracies are presented in Table 3. The expected accuracy of matching the left-out words and images is 0.5 if the model was randomly matching. All learned models predict unseen words significantly above the chance level.

In terms of unimodal prediction, VISUAL based model overall outperforms others, which verifies the **picture superiority effect** — human brain is extremely sensitive to the symbolic modality of presentation. VISUAL and LINGUISTIC significantly outperform AUDITORY based model, with the mean between category accuracy drops from approximately 0.8 to 0.68.

In terms of multimodal prediction, adding visual features improves performance as LINGUISTIC+VISUAL outperforms LINGUISTIC, VISUAL+AUDITORY outperforms AUDITORY and LINGUISTIC+VISUAL+AUDITORY outperforms LINGUISTIC+AUDITORY. These results provide a new proof for the **interactive model** of brain in behaviour measures which holds that structural and semantic information interact immediately during comprehension at any point in time, and weaken the serial model which proposes that semantic aspects only come into play at later stage and do not allow overlap with previous stages. We also notice that AUDITORY weakens model's prediction ability except for $P6$ and $P7$. Together with the finding in unimodal experiments that auditory based model performs less significantly than the linguistic and visual based model, the result suggests that visual properties contribute the most in conceptual representation in conceptual representations of nouns in the human brain, while acoustic properties contribute less. The results from $P6$ and $P7$ also suggest there are individual differences in the effects of different modality data on conceptual representations in the brain.

Kiela and Clark (2015) indicate that multimodal representations enriched by auditory information perform well on relatedness and similarity on words that have auditory associations such as *instruments*. We explore if the fMRI activation can be predicted by sound features, which is generated by using the objects which do not have obvious acoustic properties such as *hand*, *foot*, etc. Although the prediction accuracy is lower when using auditory features than using linguistic and visual features, it is significantly above the chance level. The results suggest that acoustic properties play a less important role but are ubiquitous in cognitive processes. We may need to consider the sound factors in the conceptual representation in general.

Figure 2 shows the individual mean SE$\pm$accuracy and mean SE$\pm$accuracy of within-category and between category. From Figure 2, we can see that individual performances vary in prediction and also, the result of between category prediction is better than within category prediction. We assume that this is because the features are much different between a category but more similar within a category, which makes predictions within category more demanding. For example, for linguistic feature, *'dog'* has a very similar context with *'cat'*, such as *play*, *eat*, but a very different context from *'machine'*, of which the context might be *artificial*, *fix*. Previous research has suggested that brain may rely on enhanced perceptual processing in order to compensate for inefficient higher level semantic processing, thus the phenomena of high within-category error rate and low between category error rate reflects the **sensory compensation mechanism** of brain in language processing.

	Stimuli	Forward Association Word				
		1	2	3	4	5
s-random	0.80	0.73	0.73	0.74	0.74	0.74
s-linear		0.80	0.79	0.78	0.79	0.80

(a) Mean accuracy on Mitchell et al. (2008) dataset.

	Stimuli	Forward Association Word				
		1	2	3	4	5
s-random	0.73	0.67	0.68	0.68	0.67	0.68
s-linear		0.71	0.71	0.71	0.71	0.71

(b) Mean accuracy on Pereira et al. (2018) dataset.

Table 5: Mean **FORWARD** fMRI activation prediction accuracy on Mitchell et al. (2008) and Pereira et al. (2018) dataset.

	Stimuli	Backward Association Word				
		1	2	3	4	5
s-random	0.80	0.74	0.74	0.74	0.74	0.74
s-linear		0.77	0.78	0.80	0.79	0.78

(a) Mean accuracy on Mitchell et al. (2008) dataset.

	Stimuli	Backward Association Word				
		1	2	3	4	5
s-random	0.73	0.68	0.68	0.69	0.69	0.68
s-linear		0.71	0.71	0.71	0.71	0.71

(b) Mean accuracy on Pereira et al. (2018) dataset.

Table 6: Mean **BACKWARD** fMRI activation prediction accuracy on Mitchell et al. (2008) and Pereira et al. (2018) dataset.

Table 4 shows the within category error, and we observe that Auditory features reduce the error of some categories, for example, for *body part*, VISUAL+AUDITORY outperforms simply VISUAL, and for *building part*, LINGUISTIC+AUDITORY outperforms simply LINGUISTIC. It reflects that the brain does trigger auditory senses during the rapid visual analysis and the activation of semantic knowledge, and also supports behavioural neuroscientists on that semantic processes can strongly affect generation of **auditory imagery**.

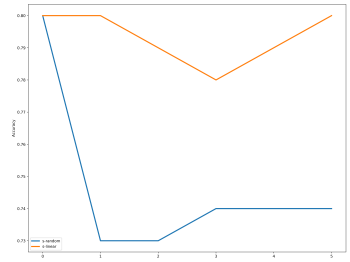
6.4.2 Associated Concept in fMRI Prediction

We choose the top 5 forward associate words and 5 backward words in our experiments. The concept of 'associate' and associative word dataset are introduced in section 6.1.3. For example, for the word *'invention'*, the associative words that people most likely to think of are *'new', 'light bulb', 'idea', 'innovation', 'creation', 'patent', 'Edison', 'Einstein', 'science', 'scientist', 'clever', 'smart', 'creative', 'create', 'Genius'*. We use the word *'invention'*, its associative words and their combinations to predict the fMRI activation separately.

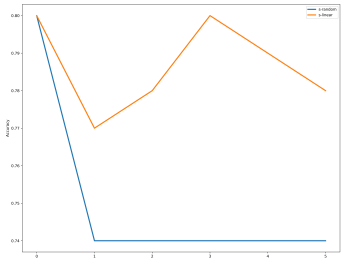
Table 5a and Table 5b are the prediction accuracy that we use stimuli and forward associative words as the input on both datasets. Tables 6a and 6b are the prediction accuracy that we use stimuli and backward associative words as input. **s-random** means using linear combination of stimuli and irrelevant word, which is randomly chosen. **s-linear** means using linear combination of stimuli and one correspondent associate word. It is important to note that, the irrelevant word is randomly chosen, and it is not associative to the stimuli. For example, for the stimuli *'invention'*, we may choose the word *'washing'*, which is not in the associative word pool of *'invention'*, as the irrelevant word. Figure 3 is the comparison of using various word association, where the original data is extracted from Table 5a, Table 5b, Table 6a and Table 6b.

Compared with (a), (b) in Figure 3, the prediction accuracy in (c), (d) is the average of 175 words. Thus, the lines in (c), (d) are more smooth. However, though the results in (a), (b) vary, they can still show the overall trend. Further, compared with using forward associative words (results from (a), (c)), using backward associative words has an equivalent performance, which means both forward and backward associative thinkings can reflect the associative conceptual representation.

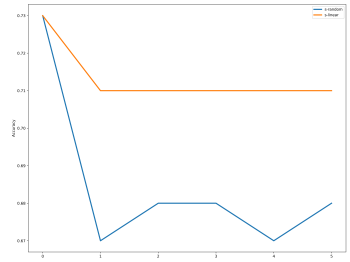
We observe that all models with associative conceptual features outperform above the chance level on both datasets. Compared with using only stimuli or associate word (bottom blue line in Figure 3), we also find that the model can better predict fMRI activation by using their linear combination (top yellow line in Figure 3). Particularly, by using stimuli and their associative words, the model has the best ability to predict fMRI activations (top yellow line in Figure 3). We also observe that after added the irrelevant word, the model's performance decreases. These results show that even though both associative words and irrelevant words are not directly associated with the stimuli words and are distinct from the stimuli



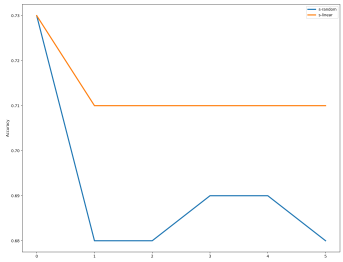
(a) 60 concrete noun forward



(b) 60 concrete noun backward



(c) 175 words forward



(d) 175 words backward

Figure 3: Comparison of various word association features. The top yellow line is corresponding to the results of *s-linear*, the below blue line is the result of *s-random*. For the point (x, y) in bottom blue line or top yellow line, x means using only the $x - th$ ranked associative word, or using linear combination of stimuli word and $x - th$ ranked associative word to predict the result. The rank tag of an associative word here means the average order of the word that participants think of in the SWW game.

words in distributed semantic representation in vector space, the associative words share some significant commonality with stimuli words in human conceptual representations while irrelevant words do not. It demonstrates that associative words serve as a complement to the stimuli words and accord with the brain activity, but the irrelevant words are noise to the conceptual representation.

In addition, there is a clear trend that the prediction accuracy decreases as the associative word rank decreases (bottom blue line in Figure 3). This result suggests that, given a stimuli, the higher ranked associate word can better reflect associative thinking related to a concept, and the subsequent associative words are less related. In other words, the rank of associative words can reflect the its weight of associative thinking in conceptual representations.

7 Conclusion and Future Work

We explored conceptual representation in human brains by evaluating the effect of multimodal data in predicting fMRI activation, observing a clear advantage in predicting brain activation for visually grounded models. This finding consistent with the neurological evidence that the word comprehension first involves activation of shallow language-based conceptual representation , which is then complemented by deeper simulation of visual properties of the concept (Louwerse and Hutchinson, 2012).

From the associative thinking perspective, we find that though the associative words might be far away in the distributed semantic vector space, we could still use them to better predict fMRI activation. We carried out more thorough and extensive work compare to the work of Bulat et al. (2017). The findings also support the hypotheses that the linguistic, conceptual and perceptual systems interplay in the human brain (Barsalou, 2008). The fMRI datasets used in our study are generated by presenting subjects with written words together with pictures. In other words, the fMRI representations are the participants' reactions to linguistic and visual input - but not acoustic. To further study human brain response representations to the acoustic stimuli, we plan to collect fMRI when presenting acoustic concepts.

References

Mahmoud Afifi. 2017. 11k hands: Gender recognition and biometric identification using a large dataset of hand images.

Lawrence Barsalou. 2008. Grounded cognition. *Annual review of psychology*, 59:617–45, 02.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46, 10.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on EMNLP*, pages 1081–1091, Copenhagen, Denmark, September. ACL.

Simon De Deyne, Danielle Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51, 10.

Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. 2009. Imagenet: A large-scale hierarchical image database. In *In CVPR*.

Barry Devereux, C Kelly, and A Korhonen. 2010. Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. *Proceedings of First Workshop On Computational Neurolinguistics, NAACL HLT*, pages 70–78, 01.

Leonardo Fernandino, Jeffrey R. Binder, Rutvik H. Desai, Suzanne L Pendl, Colin J Humphries, William L. Gross, Lisa L. Conant, and Mark S. Seidenberg. 2016. Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral cortex*, 26 5:2018–34.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-1955. *Special Volume of the Philological Society.*, page 11.

Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *ACM International Conference on Multimedia (MM'13)*, pages 411–412, Barcelona, Spain, 21/10/2013. ACM, ACM.

Jonathan T. Foote. 1997. Content-based retrieval of music and audio. In *MULTIMEDIA STORAGE AND ARCHIVING SYSTEMS II, PROC. OF SPIE*, pages 138–147.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on CVPR*, pages 770–778.

A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby. 1999. Distributed representation of objects in the human ventral visual pathway. *Proc Natl Acad Sci U S A*, 96(16):9379–9384, Aug. 10430951[pmid].

Ahmad Babaeian Jelodar, Mehrdad Alizadeh, and Shahram Khadivi. 2010. Wordnet based features for predicting brain activity associated with meanings of nouns. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, CN '10, pages 18–26, Stroudsburg, PA, USA. ACL.

Marcel Adam Just, Vladimir L. Cherkassky, Sandesh Aryal, and Tom M. Mitchell. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, 5(1):e8622, jan.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on EMNLP (EMNLP)*, pages 36–45, Doha, Qatar, October. ACL.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on EMNLP*, pages 2461–2470, Lisbon, Portugal, September. ACL.

Douwe Kiela. 2016. MMFeat: A toolkit for extracting multi-modal features. In *Proceedings of ACL-2016 System Demonstrations*, pages 55–60, Berlin, Germany, August. ACL.

Max Louwerse and Sterling Hutchinson. 2012. Neurological evidence linguistic processes precede perceptual simulation in conceptual processing. *Frontiers in Psychology*, 3, 10.

Ulrike Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December.

Alex Martin. 2007. The representation of object concepts in the brain. *Annual Review of Psychology*, 58(1):25–45. PMID: 16968210.

George A. Miller. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. 2004. Learning to decode cognitive states from brain images. *Mach. Learn.*, 57(1-2):145–175, October.

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 **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 114–123, Montréal, Canada, 7-8 June. ACL.

D. O'Shaughnessy. 1987. *Speech communication: human and machine*. Addison-Wesley series in electrical engineering: digital signal processing. Universities Press (India) Pvt. Limited.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP (EMNLP)*, pages 1532–1543.

Francisco Pereira, Tom M. Mitchell, and Matthew Botvinick. 2009. Machine learning classifiers and fmri: A tutorial overview. *NeuroImage*, 45 1 Suppl:S199–209.

Francisco Pereira, Greg Detre, and Matthew Botvinick. 2011. Generating text from functional brain images. *Frontiers in Human Neuroscience*, 5:72.

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

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.