Investigating Lexical and Semantic Cognition by Using Neural Network to Encode and Decode Brain Imaging

Lu Cao[1] and Yue Zhang[2]

[1] Singapore Univearsity of Technology and Design lu_cao@mymail.sutd.edu.sg
[2] West Lake University, China yue.zhang@wias.org.cn

Abstract. The question of how the human brain represents conceptual knowledge has received significant attention in many scientific fields. Over the last decade, there has been increasing interest in the use of deep learning methods for analyzing functional magnetic resonance imaging (fMRI) data. In this paper, we report a series of experiments with neural networks for fMRI encoding and decoding. Results show that by using neural networks, both encoding and decoding accuracies are improved compared to a linear model on the same input. To evaluate the contextual information influences in cognitive modeling, we also extend the stimuli dataset from single noun to description sentences. The experiments indicate the impact of context information varies from person to person. To illustrate the strong correlation between linguistic and visual representations in the human brain, we extend the stimuli from a single word to images which were not present to the participant during fMRI data collection.

Keywords: semantics · fmri · cognitive.

1 Introduction

How a simple concept is represented and organized by the human brain has been of long research interest [19,15,9]. The rising of brain imaging have now made it feasible to look at the neural representation of such concepts within the brain. In particular, the functional magnetic resonance imaging (fMRI) is a technique that allows for the visualization of activated brain regions. Neurons consume more oxygen as they become active. fMRI locates the activated neuron by measuring the blood-oxygen-level-dependent (BOLD) contrast. It has become an essential tool for analyzing the neural correlates of brain activity in recent decades.[42,33,30,31,38,37,22]

Many fMRI studies identify correlations between brain activity and a task the participant performs during the scan. To this end, most current understanding has been achieved by analyzing fMRI data from the viewpoint of encoding and decoding [32]. Here encoding refers to predict brain activity by using stimuli (Figure 1) and decoding refers to predict stimuli by using brain activity. When analyzing data from the encoding perspective, researchers aim to understand how activity in the brain varies when there is concurrent variation in the real world. Neuroscientists have shown that distinct patterns of fMRI activity are associated with the different stimulus, such as viewing semantic categories of pictures, including tools, animals, and buildings, reading an article which describes a concrete or abstract concepts. When analyzing data from the decoding perspective, researchers attempt to determine how much can be learned about the word. Generally speaking, decoding brain imaging is to tell what participant is thinking during the scan.

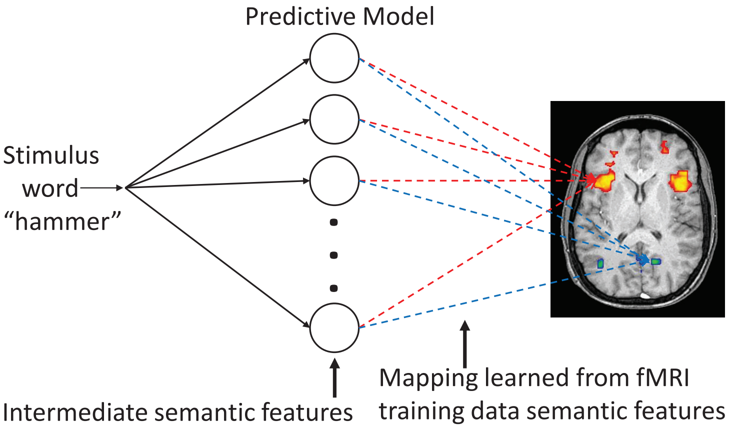
Predictive Model
Stimulus word "hammer"
Intermediate semantic features
Mapping learned from fMRI training data semantic features

Fig. 1. Predict fMRI from word stimuli

In NLP, the idea that a word's context gives its meaning dominates the approach of word representation. Consequently, a word can be represented as a high-dimensional real-valued distributional semantic vector, where the similarity among vectors reflects the similarity of contexts. These representations of linguistic meaning capture the human judgments in various tasks (*e.g.*, word analogy, word similarity judgment, and word categorization). To test whether such word representations are also neurally plausible, some studies have attempted to learn mappings between semantic vectors and brain activation patterns. If such mapping can be learned and a model can be used to predict brain activation to a new stimulus, the model can reflect some aspects of meaning representation in the human brain.

[31] first introduced the task of taking a semantic representation of a single noun and predicting the brain imaging elicited by that noun. Subsequent research [13] has attempted to fulfill this task by using various word embeddings. Most of the above work uses linear models, assuming that the concept representations are a linear combination in the human brain. For example, the word ”celery” is the linear combination of concept ”eat”, ”taste”, ”fill”, etc. However, intuitively, the human brain is complex, which apart from the linear combination, should also capture non-linear functions. This hypothesis can be verified if a non-linear model can give high accuracy compared to a linear model when fitting the same noun representation to the brain imaging output. In this paper, we apply neural networks to capture complex knowledge representations and compare the performances with a linear model. Results show that by using neural networks, both encoding and decoding accuracies are improved compared to a linear model on the

same input. This conforms our hypothesis that a non-linear function better simulates brain activation, which is also intuitive. In addition, rather than end-to-end mapping, associative thinking is also commonly believed as being involved in human cognition. For example, image thinking can be involved when giving a stimulus word such as ”celery”. As a result, it can be more accurate if a model is given a celery image as additional input when providing brain images. To address this hypothesis, we also investigate the impact of contextual and visual information of meaning representation in the human brain. Specifically, we do the following experiments:

First, human brain combines information about a word with its context. To evaluate how the context affects the meaning representation in mind, we extend the dataset [31] from single noun to some descriptive sentences. To this end, Wikipedia is used as a reference. For example, given the word ”celery”, the additional description extracted from Wikipedia, ”*Celery (Apium graveolens) is a marshland plant in the family Apiaceae that has been cultivated as a vegetable since antiquity. Celery has a long fibrous stalk tapering into leaves. Depending on location and cultivar, either its stalks, leaves, or hypocotyl are eaten and used in cooking. Celery seed is also used as a spice and its extracts have been used in herbal medicine.*”, is used together with the noun itself as the input.

Second, a growing body of evidence shows that visually embodied object representations elicited when participants are reading and contemplating object words [25,31,4]. More generally speaking, reading words evokes visual simulation, and viewing images evokes semantic representations [5,43]. The brain region of conceptual representations of linguistic is linked to visual perception. To illustrate this, we retrieve images from ImageNet [10] and using image features build the brain encoding and decoding models. Results show that out of 8 participants, the brain activation of most can be better predicted by adding visual inputs to the word, which strongly demonstrates that image associative thinking exist in human perceptron of noun semantics. Interestingly, the brain activation of 4 participants can also be better predicted by using the descriptive sentence as additional input information, which shows that some people can do analytically thinking when understanding a noun concept.

We investigate the cross-lingual brain encoding and decoding. The fMRI data are collected from English native speakers. We want to explore the brain semantic representation difference among different language speakers. To this end, we use Chinese word embeddings [45] to encode and decode brain imaging. Surprisingly, the accuracy is close to using English word embeddings. This result suggests that semantic representations can be irrelevant to the language itself.

Finally, although languages and cultures vary, perceptron of common concepts can be universal across people.

In summary, we use neural network models to investigate aspects of meaning representation in the human brain, showing that they better reflect cognitive functions compared with linear models used in prior research. We additionally find that image and descriptive sentences can help better predict fMRI as additional inputs to the noun stimuli, which reflects that associative thinking is likely involved in the semantic understanding process. Interestingly, cross-lingual embedding inputs can give similar results for predicting fMRI images, which shows that there can be cross-lingual common ground in semantic representations of words.

2 Related Work

In word embeddings, there are two main methods map words from vocabulary to vectors of the real numbers. The first is one-hot encoding, in which each word is represented as a vocabulary size vector of zeros except one position of one (*e.g.*, apple = [1, 0, 0 ... 0], fruit = [0, 0, 1, ..., 0]). The limitation of the one-hot representation is it fails to capture the relationship between words because the words are perpendicular in one-hot encoding. The other method to represent a word is distributed representation, in which word is mapped to a vector space of continues real number. [29] first efficiently trained the word embedding on Google News Corpus. The word to vector [29] is trained by language modeling by using skip-gram or continues bag of words (CBOW) algorithm. The CBOW is learning to maximize the probability of the target word by looking at the context while the skip-gram is to predict the context. Similarly, the GloVe [35] is trained on aggregated global word-word co-occurrence statistics from a corpus. The intuition underlying the model is ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning. There are many other versions word embeddings *e.g.*, [24,39,12,45,6,21], released by different institutions. The main advantage of the distributed word representation mentioned above is that it captures the context of a word. By distributed representation, word-word relation can be reflected, for example, the cosine similarity of word vector $apple$ and $fruit$ is close compare to $apple$ and car.

The brain encoding task was first introduced by [31], the task is to learn a mapping between word embedding and the human brain imaging. To this end, [31] created 25-dimension word vector trained on [7] and learned a mapping by using the linear model. In addition to using 25 semantic features [31], [20] incorporate the relatedness measures based on WordNet. [11] choose a set of verbs for semantic features. [13] use five semantic attributes directly related to sensory-motor experience-sound, color, visual motion, shape, and manipulation. The relevance of these attributes to each word is rating by the human.

On the decoding side, studies have shown that it is possible to accurately decode a participant's mental content for words [34,36,14,2], text snippets [47,17], or sentences [28,46,3].

3 Methods

We build a linear model and a non-linear model for all our experiments. For the former, we follow [31] and use linear regression as the main method. For the latter, a standard multi-layer neural network is used.

For the encoding task, the inputs of the model are semantic features such as a single word and the model is required to predict the corresponding fMRI activation. [31] created 25 dimension distributed embeddings based on the co-occurrence statistics for each stimulus. Each word vector gives the normalized co-occurrence frequency of the stimulus noun with each of 25 verbs (*i.e.*, see, say, taste, wear, open, run, neared, eat, hear, drive, ride, touch, break, enter, move, listen, approach, fill, clean, lift, rub, smell, fear, push, manipulate). In this regard, a word is encoded into 25 semantic features. In addition to the 25 distributed features, we use the GloVe word vectors [35] as distributed semantic features. Compared with Mitchell's word vector [31]; GloVe [35] has flexible dimensions. With higher dimensions, the word can preserve rich semantic features.

The predicted value of y_t a voxel t in brain for the word w can be written as

$$y_t = \sum_{i=1}^{n} c_{ti} f_i(w), \tag{1}$$

where $f_i(w)$ is the $i^t h$ semantic feature of w, n is the dimension of semantic feature and c_{ti} is learn-able parameter. A multi-layer neural networks can be written as

$$y_t = \sum_{i=1}^{M} c_{ti} h_i + c_{i_0} \tag{2}$$

$$h_i = max(0, \sum_{j=1}^{N} c_{ji} x_j + c_{j_0}), \tag{3}$$

where y_t is the predicted value at voxel t, M is the number of hidden unit in a layer and h_i is the value of the $i^t h$ hidden unit for word w. c_{ti} is the learned coffcient that specifies the degree to which the i^{th} semantic feature activates a voxel.

For the decoding task, the inputs are the fMRI images and outputs are predicted semantic representations. The same algorithms are used.

3.1 Training Objective

To train a multiple regression classifier, if the number of training examples is larger than the semantic feature dimension, a unique solution exists. Otherwise, a solution can be obtained by introducing a regularization term. To train a neural network, we employ the standard back-propagating [44] algorithm, which minimizes the mean square loss between the real fMRI and predicted one.

Following [31], we train and evaluate separate computational models for each of the nine participants by using a cross-validation approach. The model was trained repeatedly by a leave-two-out procedure, in which the model was trained using 58 of the 60 stimuli. The process iterates 1770 times. At test two, each trained model predicts the fMRI for two ”left-out” words and then match predicted fMIR to their corresponding left-out ones. Given a trained model, the two left-out words w_1, w_2 and their fMRI images i_1, i_2, the model predicts the fMRI p_1, p_2 for w_1, w_2, respectively. It then decide which is the better match ($i_1 = p_1$, $i_2 = p_2$) or ($i_1 = p_2$, $i_2 = p_1$) by computing the cosine similarity. The match score S is calculated as

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

Similarly, the match score for the decoder is the sum of cosine similarity between actual and predicted word vectors.

3.2 Incorporating Contextual Information in Encoding

We extend the dataset by retrieving an article for each noun manually from Wikipedia. Each noun is assigned 3 to 6 sentences which describe the properties of the word. For example, to describe the proprieties of the chisel, we retrieve the article: *'A chisel is a tool with a characteristically shaped cutting edge (such that wood chisels have lent part of their name to a particular grind) of blade on its end, for carving or cutting a hard material such as wood, stone, or metal by hand, struck with a mallet, or mechanical power. The handle and blade of some types of chisel are made of metal or of wood with a sharp edge in it. Chiselling use involves forcing the blade into some material to cut it. The driving force may be applied by pushing by hand, or by using a mallet or hammer. In industrial use, a hydraulic ram or falling weight ('trip hammer') may be used to drive a chisel into the material. A gouge (one type of chisel) serves to carve small pieces from the material, particularly in woodworking, wood-turning and sculpture. Gouges most frequently produce concave surfaces. A gouge typically has a 'U'-shaped cross-section'*. To preserve contextual information, we apply convolution neural networks (CNNs) to each article. The CNNs involves a filter $W \in R^h$, which is applied to a window of h words to produce new features [23]. For example, a feature s_i is produced from a window of words $x_{i:i+h-1}$ by

$$s_i = f(w \cdot x_{i:i+h-1} + b), \tag{5}$$

where $f(\cdot)$ is a non-linear function and b is the bias term. Then substitute s_i into Eq. 3 as x_j.

Since word embeddings are trained based on words that co-occur with one another in a given corpus, they already encodes the context to some extent, why do we use the article to encode/decode fMRI again? The answer can be the article describes the word exclusively, provides detail and precision information, which contributes the high-quality features for subsequent brain imaging encoding and decoding. Besides, a word can be different meaning in multiple contexts, *e.g.*, the word "monster" in describing a computer or a person. Using articles for encoding and decoding can eliminate ambiguity.

3.3 Image and Text Correlation

Conceptual representations of linguistic and visual perception are linked together in the human brain [5,43]. The word comprehensiveness first involves activation of shallow language-based conceptual representation and then complemented by deeper simulation of visual properties of the concept [26]. The latter can play a much more critical role.

Based on the above argument, we assume that the image augmented model should be able to encode and decode the brain activation elicited by a word. To this end, we retrieve 300 to 1500 images for each noun from ImageNet [10], one of the largest image databases, except 'hand', 'foot', 'arm', 'leg' and 'eye'. The ImageNet [10] does not include these human body words. We retrieve these images from Google [18] and [1].

Deep Residual Network (ResNet) [16] is widely used in image recognition. It is a deep neural network with many convolution layers stack together. With more convolution layers, the network can extract rich image features. In our implementation, we use ResNet to produce the image feature map. Concretely, we extract each feature of each image and then average all features of the same word. For example, an image feature map $Fmap_i$ of image x_i is produced by

$$Fmap_i = resnet(x_i), \tag{6}$$

where $resnet$ is the function that produces the feature map and $Fmap_i$ is a 2048 dimensional real value vector. Suppose there are N images of the word 'airplane', we produce the image feature map $F_airplane$ of 'airplane' by

$$F_{airplane} = \frac{\sum_{i=1}^{N} Fmap_i}{N}, \tag{7}$$

then substitute $F_{airplane}$ into Eq. 3 as x_j to train the neural network.

3.4 Cross-lingual Embeddings

Is the semantic representation identical or various from different language speakers? The question can be answered by decode the brain imaging collected from different language speakers when shown the same word stimuli. With the word embedding, we give a simple alternative to answer this question. Since the word embedding captures semantic reasons while producing the text, if we can encode and decode the English native speakers' brain imaging using word embedding from a different language with high accuracy, we gain evidence there is the common ground of semantic representation.

We choose Chinese word embeddings [45] in our work, which consists 200-dimension vector representations which are pre-trained on large-scale high-quality data. These vectors, capturing semantic meanings for Chinese words and phrases, can be widely applied in many downstream Chinese processing tasks.

4 Experiments

We use the fMRI data from [31], which were collected from nine healthy, college-age participants. The stimuli are line drawings and nouns labels of 60 concrete objects from 12 semantic categories as Table 1. During data collection, the 60 word-picture pairs were presented to each participant six times with randomly permuting. To create a representative fMRI for each stimulus, we compute the mean fRMI response over its six scans. The mean of all 60 representative images is then subtracted from each image.

4.1 Experimental Settings

We design four experiments to evaluate our models:

- In Experiment 1, we re-implemented the linear model with Micheal's word vector [31] and GloVe [35].
- In Experiment 2, we use multiple layer feed-forward neural networks to predict fMRI activation.
- In Experiment 3, we incorporate the context into the model.
- In Experiment 4, we implement the linear and non-linear models to decode fMRI and analyze the results, respectively. Also, we also use the image feature as input to predict the fMRI activation.

Table 1. The 60 noun and categories

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

4.2 Voxel Selected

The fMRI data is 3D image consists of voxels (volumetric pixels) covering millions of brain neurons. It comes from the machine as stacked 2D slices. The number of voxels is enormous, usually greater than 50000, it makes impossible to fit a standard linear model [41] because the number of observations are usually far less than the voxels. To deal with this problem, a simple way is to use a subset of the whole brain voxels. Depend on the specific task; there are many algorithms to choose a subset of the voxels, in our work, we follow [31]'s method to pick the 500 most stable voxels. During the training process, each voxel was assigned a ”stability score” by using the data from 6 scan sessions of the 58 training stimuli. The 500 most stable voxels ranked highest by this stability score were used in the experiments.

4.3 Visualization of fMRI

t-SNE [27] is a supervised dimensionality reduction tool for visualization high-dimensional data into two-dimensional space using stochastic neighbor embedding.

The fMRI data is high dimensional, to help understand the properties of the data, we use t-SNE [27] to reduce the fMRI dimension to 2D and visualize them as Figure 2.

The dataset [31] consists of 360 brain imaging (fMRI) of 12 categories as Table 1. In left side of the Figure 2, each point represents an reduced dimension fMRI data. We assign different colors to each category and the label is the center of the category. Ideally, the points in Figure 2 can be clustered to 12 categories and the distance between groups should be large. In our visualization, the groups are not ideally separable, but we still can find some valuable information from it. For example, the group of furniture, building, build-part, vehicle are close and the group of kitchen, vegetable, tool are close. This is accord to common sense since the vegetables and tools usually appear in a kitchen, and furniture, building, build-part, vehicle regularly appear in the same scenario. When we talk about the kitchen, it is very natural to think of tools such as cutlery and foods

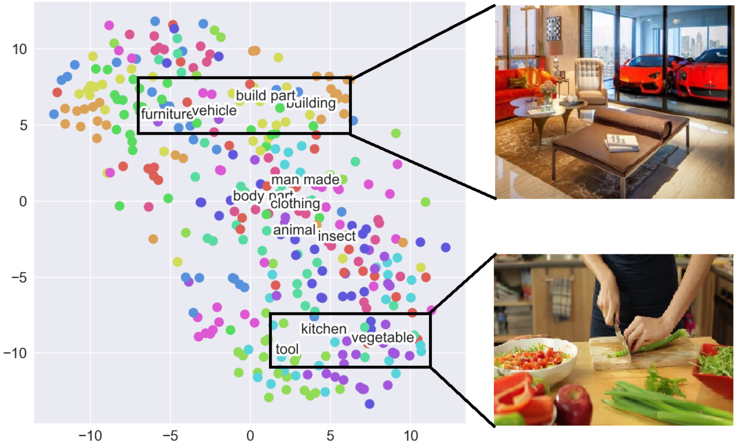

Fig. 2. Visualization of fMRI

Table 2. The average accuracy of the linear and non-linear ***encoders*** with different word vectors for participant P1 to P9

Word Embeddings	Linear	Non-Linear
[31]	**0.79**	0.77
GloVe50	0.69	**0.77**
GloVe100	0.74	**0.80**
GloVe200	0.74	**0.76**
GloVe300	0.77	**0.78**
Tencent200	0.75	**0.81**

such as vegetables. When we talk about the apartment, it is more likely to think about furniture, vehicle, building, and building part. The right side of the Figure 2 are images we retrieved from [8,40], as we can see, vegetables and knife appear in kitchen while furniture, building, build-part, vehicle appear together. But it is less likely think of car when talking about the kitchen. The visualization of the fMRI data gives us a preliminary conclusion that associative thinking is neurally plausible and can be reflected by fMRI categorization.

4.4 Encoding

The linear model's results of Experiment 1 are shown in the left column of Table 2. From the results, we observe that with linear regression, Mitchell's 25-dimensional word vector [31] achieves the best prediction results, better than GloVe vectors [35]. The high dimensional GloVe embedding [35] preserves rich semantic information compared with the low dimensional vectors, but the result does not accord with intuition. There can be two main reasons. First, Mitchell's 25-dimensional word vector [31] is based on the most

Table 3. The average accuracy of the linear and non-linear ***decoders*** with different word vectors for participant P1 to P9

WORD EMBEDDINGS	LINEAR	NON-LINEAR
[31]	0.75	**0.76**
GLOVE50	0.81	**0.84**
GLOVE100	0.84	**0.86**
GLOVE200	0.82	**0.85**
GLOVE300	0.82	**0.84**
TENCENT200	0.84	**0.87**

Table 4. The accuracy of ***encoders*** with or without contextual information

PARTICIPANTS	P1	P2	P3	P4	P5	P6	P7	P8	P9	AVG
WITHOUT CONTEXT	0.88	0.71	**0.79**	0.87	0.76	**0.74**	0.72	**0.70**	0.81	0.78
WITH CONTEXT	**0.90**	**0.77**	0.77	**0.88**	0.76	0.68	**0.78**	0.64	0.81	0.78

significant hand chosen semantic features which include the most common properties of a noun. For example, the word ”celery” is represented as $celery = 0.837 \times eat + 0.346 \times taste + 0.315 \times fill + ... + 0 \times move$, where the concept ”eat” contributes a 0.837 weight to celery, while to concept ”move” contributes a 0 to celery. This conforms to common sense because celery is edible. GloVe [35], on the other hand, contains distributed semantic features, which can be irrelevant. The second reason is that linear regression is too simple to capture complex representations. Even if the GloVe [35] preserves rich semantic information, where may not combine for the task. Hence a linear model may not be the best choice to learn the mapping between the human brain and the semantic representations. Instead, we need more expressive models to cope with the complexity.

The non-linear model results of Experiment 2 are shown in right column of Table 2. We observe that one hidden-layer neural network outweighs linear model on all GloVe [35] variants. The results support the view that the concept representation is not linearly combined in the human brain.

4.5 Deocding

We implement linear and non-linear decoders in Experiment 4. The results are summarized in Table 3. We observe that both linear and non-linear models can decode fMRI. The non-linear model outweighs the linear model on all word vectors. Both models have significant improvements by using GloVe [35]. This result implies that the human brain encodes semantic representation in a complicated function and a proper semantic space is essential to brain decoding.

The decoding is an illustration of *zero-shot learning* [34] of the human brain as the decoder can predict novel classes that were omitted from a training set with an accuracy

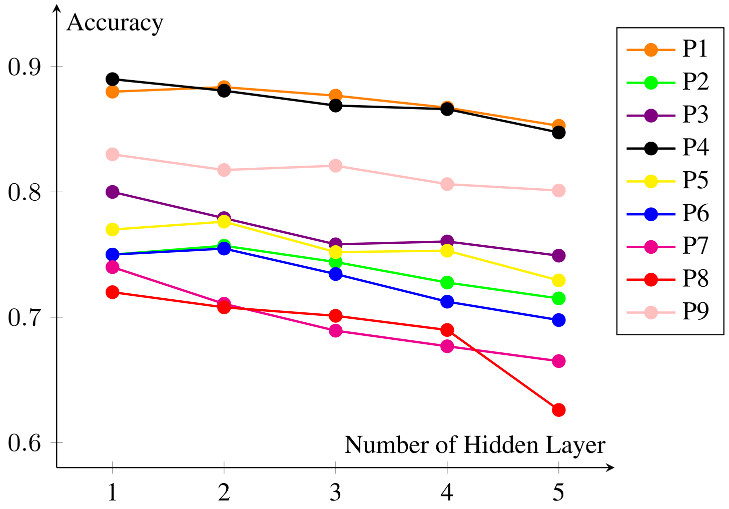

Fig. 3. The impact of number of hidden layers

much higher than chance. After training on a semantic feature space, the decoder can leverage a semantic knowledge base that encodes both training and test set features.

4.6 Impact of Hidden Layers

To evaluate the impact of network depth, we tested neural networks with one to five hidden layers and summarized the results as Figure 3, from which we observe that as the hidden layer increases, the accuracy decreases across all participants. Generally, the model can fit complex data with more hidden layers. We expect that the model could learn complex concepts from the semantic representations, but the result is the opposite. The reason can be over-fitting.

4.7 Impact of Contextual Information

In Experiment 3, we incorporate contextual information into the non-linear model. The results are summarized in Table 4. We observe that the impact of the contextual information varies from person to person. Compared to Experiment 2, the accuracy of participant P1, P2, P4, P7 increase, while that of P3 and P6 decrease. This result suggests that humans can understand the same concept from different perspectives. Through learning the mapping between large corpus-based semantic vectors and brain imaging, we may find out how the human brain associates one concept or thought to another.

4.8 Similarities Between Predicted and Actual Images

Figure 4 are the observed and predicted fMRI of the word ”hand” after training uses 58 nouns. The left image is the observed fMRI, and the right one is the predicted. Although

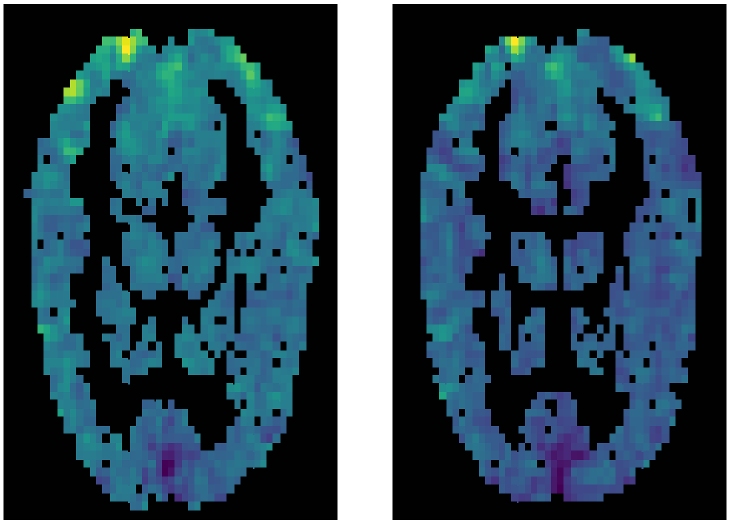

Fig. 4. Observed and predicted fMRI images for ”hand”

Table 5. The accuracy of ***encoders*** with word embedding and images

Participants	P1	P2	P3	P4	P5	P6	P7	P8	P9	AVG
GloVe300	0.88	0.71	0.82	0.83	**0.80**	0.76	0.69	**0.67**	0.77	0.77
Images	**0.93**	**0.79**	0.82	**0.89**	0.78	**0.79**	**0.81**	0.66	**0.81**	**0.81**

the predicted fMRI is not perfect, it captures main components of the ground truth. The yellow and green area near the top is the fusiform gyri. In neuroscience, the functionality of the fusiform gyrus has been linked to recognition (*e.g.*, processing of color information, face and body recognition, word recognition, within-category identification).

To provide more insight into the non-linear model. Figure 5 depicts for participant P1 the cosine similarity between each predicted images and real images.

4.9 Use Image Features to Encode Brain Imaging

We use the ResNet [16] extracted image features to predict the fMRI activation. Figure 6 is a part of celery images used in experiments. The result is obtained by using the linear regression model and is summarized in Table 5. Six out of nine participants' accuracy is improved, and average 4% increase across all participants.

The result supports the view that word comprehension involves both linguistic and visual processes in the human brain [26]. We use ResNet [16] to extract each image's feature as Eq. 6. The single image feature map contains noise, which leads to the model fail to encode the brain imaging. We extract multiple image features and average them on each dimension as Eq. 7. By means of sum up and average on all feature maps, the

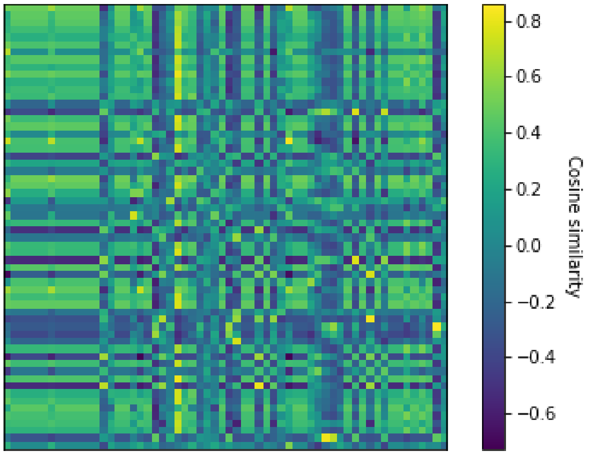

Fig. 5. Cosine similarity between actual and predicated images. The mean of the diagonal is **0.3371**, wheres the mean of entire matrix is **0.0720**. This indicates that the predicted image is similar to the real image than others in general.

salient features of the celery are kept such as color, shape, texture, contour, etc., and the noises are suppressed.

4.10 Cross-Lingual Analysis

The results of cross-lingual encoding and decoding are summarized in last line of Table 2 and Table 3. Intuitively, the difference of semantic representation among different language speakers' brain exists and can be reflected by using cross-lingual brain imaging encoding and decoding. However, the experimental result is the opposite. The result of encoding and decoding of English native speakers' brain imaging by using Chinese word embedding [45] are very close. Even though we cannot conclude the semantic representation is human-wide identical for sure, which needs more biological experiments, the result still suggests the brain semantic representation among different language speakers is highly correlated.

This finding brings many possibilities to incorporate the brain imaging data into NLP models (*e.g.*, leverage brain imaging to supervise the machine translation models).

5 Conclusion

We use a machine learning approach in understanding human cognition process of word semantics, finding that non-linear models with distributed semantic vectors give better accuracies compared to linear models with hand-chosen distributional vectors. This reflects that word embeddings are a good representation of lexical semantics and that human cognition involves non-linear activities, which is highly intuitive. We additionally

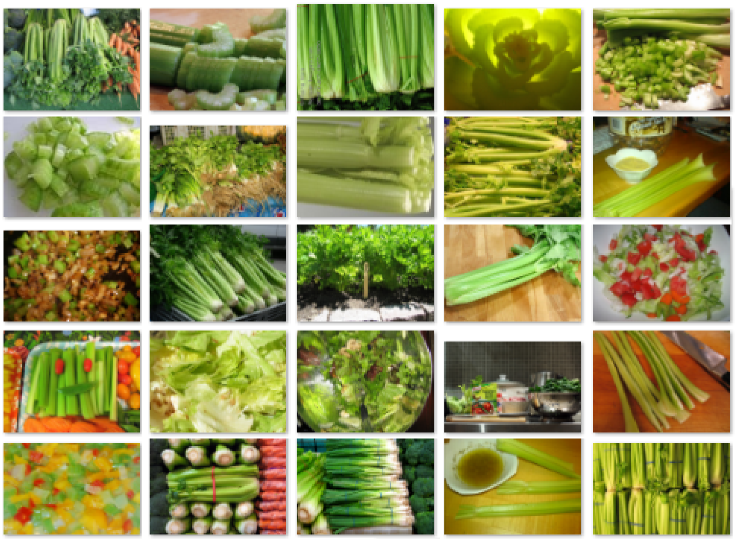

Fig. 6. A part of celery images used in experiments

investigated associative thinking by using additional image and descriptive sentence inputs to the non-linear model for predicting fMRI, finding that both are useful, with the former being more effective. Finally, we found that cross-lingual word representations can be equally useful in predicting brain images of English speakers, which shows that there can be strong cross-lingual associations in word vectors. The dataset in this work is relatively small, and the stimuli are all nouns, which limits our work in a small part of brain understanding. We will extend our work to a sizeable semantic space, which covers most of the human concepts in future work.

References

1. Afifi, M.: 11k hands: Gender recognition and biometric identification using a large dataset of hand images (2017)
2. Anderson, A.J., Kiela, D., Clark, S., Poesio, M.: 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 (2017)
3. Anderson, A.J., Binder, J.R., Fernandino, L., Humphries, C.J., Conant, L.L., Aguilar, M., Wang, X., Doko, D., Raizada, R.D.S.: Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cerebral Cortex **27**(9), 4379–4395 (2017)
4. Anderson, A.J., Bruni, E., Lopopolo, A., Poesio, M., Baroni, M.: Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. NeuroImage **120**, 309 – 322 (2015)
5. Binder, J.R., Desai, R.H.: The neurobiology of semantic memory. Trends Cogn Sci **15**(11), 527–536 (Nov 2011), 22001867[pmid]

6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
7. Brants, T., Franz, A., Consortium., L.D.: Web 1t 5-gram version 1 (2006)
8. ColourBox: Colourbox (2019)
9. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fmri) “brain reading”: detecting and classifying distributed patterns of fmri activity in human visual cortex. NeuroImage **19**(2), 261 – 270 (2003)
10. Deng, J., Dong, W., Socher, R., jia Li, L., Li, K., Fei-fei, L.: Imagenet: A large-scale hierarchical image database. In: In CVPR (2009)
11. Devereux, B., Kelly, C., Korhonen, A.: Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. Proceedings of First Workshop On Computational Neurolinguistics, NAACL HLT pp. 70–78 (01 2010)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Fernandino, L., J Humphries, C., Seidenberg, M., Gross, W., Conant, L., R Binder, J.: Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. Neuropsychologia **76** (04 2015)
14. Handjaras, G., Ricciardi, E., Leo, A., Lenci, A., Cecchetti, L., Cosottini, M., Marotta, G., Pietrini, P.: How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. NeuroImage **135**, 232 – 242 (2016)
15. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science **293**(5539), 2425–2430 (2001)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
17. Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural speech reveals the semantic maps that tile human cerebral cortex. Nature **532**, 453 EP – (Apr 2016), article
18. Images, G.: Google images (1998), `https://images.google.com`
19. Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., Haxby, J.V.: Distributed representation of objects in the human ventral visual pathway. Proc Natl Acad Sci U S A **96**(16), 9379–9384 (Aug 1999), 10430951[pmid]
20. Jelodar, A.B., Alizadeh, M., Khadivi, S.: Wordnet based features for predicting brain activity associated with meanings of nouns. In: Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics. pp. 18–26. CN ’10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
21. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification (2016)
22. Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M.: A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS ONE **5**(1), e8622 (jan 2010)
23. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1746–1751 (2014)
24. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 3294–3302. Curran Associates, Inc. (2015), `http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf`
25. Kriegeskorte, N., Mur, M., Bandettini, P.: Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience **2**, 4 (2008)

26. Louwerse, M., Hutchinson, S.: Neurological evidence linguistic processes precede perceptual simulation in conceptual processing. Frontiers in Psychology **3** (10 2012). https://doi.org/10.3389/fpsyg.2012.00385
27. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
28. Matsuo, E., Kobayashi, I., Nishimoto, S., Nishida, S., Asoh, H.: Describing semantic representations of brain activity evoked by visual stimuli (2018)
29. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
30. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to decode cognitive states from brain images. Mach. Learn. **57**(1-2), 145–175 (Oct 2004)
31. Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A.: Predicting human brain activity associated with the meanings of nouns. Science **320**(5880), 1191–1195 (2008)
32. Naselaris, T., Kay, K., Nishimoto, S., Gallant, J.: Encoding and decoding in fmri. NeuroImage **56**, 400–10 (05 2011)
33. Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fmri data. Trends in Cognitive Sciences **10**(9), 424 – 430 (2006)
34. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 1410–1418. Curran Associates, Inc. (2009)
35. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
36. Pereira, F., Botvinick, M., Detre, G.: Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. Artif. Intell. **194**, 240–252 (Jan 2013)
37. Pereira, F., Detre, G., Botvinick, M.: Generating text from functional brain images. Frontiers in Human Neuroscience **5**, 72 (2011)
38. Pereira, F., Mitchell, T.M., Botvinick, M.: Machine learning classifiers and fmri: A tutorial overview. NeuroImage **45 1 Suppl**, S199–209 (2009)
39. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. of NAACL (2018)
40. pinterest: pinterest (2019)
41. Poldrack, R.A., Mumford, J.A., Nichols, T.E.: Handbook of Functional MRI Data Analysis. Cambridge University Press (2011). https://doi.org/10.1017/CBO9780511895029
42. Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A.: Category-specific cortical activity precedes retrieval during memory search. Science **310**(5756), 1963–1966 (2005)
43. Pulvermüller, F.: How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. Trends in Cognitive Sciences **17**(9), 458 – 470 (2013)
44. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533– (Oct 1986)
45. Song, Y., Shi, S., Li, J., Zhang, H.: Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (2018)
46. Wang, J., Cherkassky, V.L., Just, M.A.: Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. Human brain mapping **38 10**, 4865–4881 (2017)

47. Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T.: Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS One **9**(11), e112575–e112575 (Nov 2014), 25426840[pmid]