When Computational Representation Meets Neuroscience: A Survey on Brain Encoding and Decoding

Lu Cao[1] , **Dandan Huang**[2,3] and **Yue Zhang**[2,3*]
[1]Singapore University of Technology and Design, Singapore
[2]School of Engineering, Westlake University, China
[3]Institute of Advanced Technology, Westlake Institute for Advanced Study, China
lu_cao@mymail.sutd.edu.sg, {huandandan, zhangyue}@westlake.edu.cn

Abstract

Real human language mechanisms and the artificial intelligent language processing methods are two independent systems. Exploring the relationship between the two can help develop human-like language models and is also beneficial to reveal the neuroscience of the reading brain. The flourishing research in this interdisciplinal research field calls for surveys to systemically study and analyze the recent successes. However, such a comprehensive review still cannot be found, which motivates our work. This article first briefly introduces the interdisciplinal research progress, then systematically discusses the task of brain decoding from the perspective of simple concepts and complete sentences, and also describes main limitations in this field and put forward with possible solutions. Finally, we conclude this survey with certain open research questions that will stimulate further studies.

1 Introduction

How the human brain encodes and processes languages attracts research attention from neuroscientists, computational linguists and psychologists. Studies have shown that distinct spatial pattern of neural activity is related to the viewing pictures of certain semantic categories [Mitchell *et al.*, 2008]. Mitchell *et al.* [2008] presents the first computational model that makes directly testable predictions of the fMRI activities from concrete nouns. As shown in Figure 1, the study presents concrete nouns to subjects, collects fMRI data during the word presentation, and trains a linear model to predict the fMRI activity from the word vector. The underlying theory is that the distributional properties of nouns in a large corpus is related to the neural basis of the semantic representation in the human brain. Based on this theory, subsequent studies extend the prediction tasks from concrete nouns to abstract nouns [Wang *et al.*, 2013], text fragments [Wehbe *et al.*, 2014a; Huth *et al.*, 2016], and sentences [Sun *et al.*, 2021].

In recent years, the rapid progress of deep learning and natural language processing (NLP) provide an opportunity for further exploration. Distributed representation of words or sentences can encode richer semantic information. Some studies attempt to use extensive word vectors to predict neuron activity patterns and evaluate the performance accordingly [Murphy *et al.*, 2012; Anderson *et al.*, 2013; Fernandino *et al.*, 2015; Bulat *et al.*, 2017; Abnar *et al.*, 2018; Wang *et al.*, 2020b; Pereira *et al.*, 2010]. Most recently, pre-trained language models (LMs) have achieved success in various fields of NLP. The network structure and rich training data enables the model to capture more information covering syntax, semantics and even common sense knowledge from the context. On this basis, much work attempts to deeply explore the relationship between pre-trained LMs and human language mechanisms [Schwartz *et al.*, 2019; Toneva and Wehbe, 2019; Jat *et al.*, 2019; Schwartz and Mitchell, 2019; Hale *et al.*, 2015; Søgaard, 2016; Minnema and Herbelot, 2019; Pereira *et al.*, 2010].

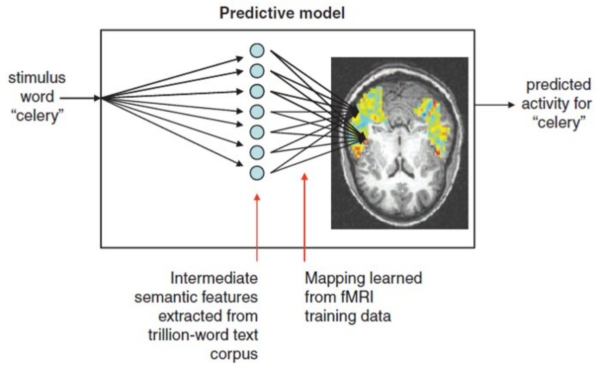
Predictive model
stimulus word "celery"
predicted activity for "celery"
Intermediate semantic features extracted from trillion-word text corpus
Mapping learned from fMRI training data

Figure 1: Predict fMRI from word stimuli [Mitchell *et al.*, 2008]

We give a survey of recent efforts on bridging natural language processing and neuroscience of language, aiming to provide a comprehensive overview of existing research, discussing the focuses, limitations and future directions.

2 Background

2.1 Motivation

By drawing correlation between computational representations and neural representations, we can benefit both the development of NLP algorithm and the understanding of brain activities. Some studies revealed the correlations between deep neural networks and human language process-

[*]Contact Author

Dataset	Type	Language	Stimulus	#Subject	Paradigm	Task
Pereira *et al.* [2001]	fMRI	English	20 sentences	6	reading	decoding
Cox and Savoy [2003]	fMRI	English	10 object pictures	4	reading	decoding
Mitchell *et al.* [2004]	fMRI	English	concrete noun	10	reading	decoding
Mitchell *et al.* [2008]	fMRI	English	60 concrete noun	9	reading	encoding
Sudre *et al.* [2012]	MEG	English	60 concrete noun	9	reading	decoding
Anderson *et al.* [2012]	fMRI	Italian	70 concrete&abstract noun	7	reading	decoding
Wehbe *et al.* [2014b]	fMRI	English	story	9	reading	encoding
Wehbe *et al.* [2014c]	MEG	English	story	3	reading	decoding
Rafidi [2014]	MEG	English	32 sentences	8	reading	decoding
Frank *et al.* [2015]	EEG	English	205 sentences	24	reading	decoding
Huth *et al.* [2016]	fMRI	English	story	7	listening	encoding
Brennan *et al.* [2016]	fMRI	English	story	26	listening	decoding
Hollenstein *et al.* [2018]	EEG	English	700 sentences	12	reading	/
Brennan and Hale [2019]	EEG	English	story	33	listening	decoding
Zinszer *et al.* [2017]	fNIRS	English	8 concrete noun	24	viewing & listening	decoding
Oseki and Asahara [2020]	EEG	Japanese	20 newspaper articles	40	reading	/
Cao *et al.* [2021]	fNIRS	Chinese	50 concrete nouns	7	viewing & listening	decoding

Table 1: Summary of available datasets for brain decoding and encoding.

ing. Supervised RNNGs [Dyer *et al.*, 2016] have been shown can better encode syntactic properties of language [Kuncoro *et al.*, 2017] and correlate with electrophysiological responses in the human brain [Hale *et al.*, 2018]. A recent study [Schrimpf *et al.*, 2020] tested 43 LMs on three neural datasets and found that the most powerful generative Transformer models [Radford *et al.*, 2019] accurately predict neural responses.

Brain activity can be leveraged to increase the performance of NLP models. For example, Bingel *et al.* [2016] found that the fMRI data contains a strong signal, enabling a 4% error reduction over a state-of-the-art unsupervised parts of speech (PoS) tagger. Hollenstein *et al.* [2019a] leveraged the EEG and eye-tracking data to improve the performance of named entity recognition, relation classification and sentiment analysis, yielding significant improvements over the baselines. Hollenstein *et al.* [2020] reviewed studies leveraging different cognitive processing signals, i.e., eye-tracking, MEG, EEG, and fMRI data recorded during language understanding, and proposed practical strategies for using cognitive signals to augment NLP models.

Another interesting usage of brain activity data is evaluating how much a word representation can reflect the semantic representation in the human brain. Xu *et al.* [2016] presented a fast and lightweight tool, *Brainbench*, which can evaluate word vectors of 60 concrete nouns with their fMRI brain images. Hollenstein *et al.* [2019b] presented the first multi-modal large-scale cognitive word embedding evaluation framework, *CogniVal*. The word vector can be evaluated by fitting 15 datasets of eye-tracking, EEG, and fMRI signals.

2.2 Task

Neuroscientists use **decoding** and **encoding** to analyze the information represented in brain activity. Decoding and encoding are two complementary processes: decoding uses neural activity to predict stimuli while encoding uses stimuli to predict brain activity. The pioneering work of decoding and encoding is Pereira *et al.* [2001] and Mitchell *et al.* [2008], respectively. Both approaches are used to establish a mapping between brain activity and feature space, with the only difference of the mapping direction. Thus we do not distinguish them in the discussion. We summarize the most used brain activity datasets in these tasks in Table 1. Among these researches, brain activity data including fMRI, EEG, MEG, and fNIRS are used. From the point of data usage, the main difference lies in the temporal and spatial resolutions. Given the space limit, we do not discuss the difference of data types indepth. Instead, below we categorize concept representations in detail and list the most representative works in Table 2.

3 Concept Representation

Decoding concepts from patterns of brain activity is also referred neurosemantics. It aims to learn the mapping between concepts and the neural activity patterns elicited during neuroimaging experiments. This task relies on three essential parts: semantic vectors, brain activity data, and mapping functions. Collecting brain activity data is subject to the complex experiment environment and the high expenses. The linear mapping can give the lower bound of the mapping. Thus the majority of neurosemantic studies focuses on the discussion of semantic vector construction.

3.1 Feature-based Representations

Initial studies mainly focused on constructing semantic features. The work of Mitchell *et al.* [2008] computed the noun representations based on their co-occurrence with a hand-designed set of 25 verbs (reflecting sensory-motor features) in a trillion-token corpus. Fernandino *et al.* [2015] instead focused on only five semantic attributes related to sensory-motor experience: sound, color, manipulation, visual motion and shape. Rather than the co-occurrence rate, the ratings for these attributes reflect the salience of each attribute to the meaning of the word on a 7-point Likert scale ranging from “not at all important” to “very important”. Babaeian Jelodar *et al.* [2010] used the same 25 features [Mitchell *et al.*, 2008] to represent the concrete nouns. But instead of corpus statistics, the value for the 25 features were computed

Type	Concept Representations	Representative Works
Feature-based	Tom 25 Sensor-motor	Mitchell *et al.* [2008] Devereux *et al.* [2010], Babaeian Jelodar *et al.* [2010], Fernandino *et al.* [2015]
Distributional	GloVe, Word2Vec, fastText, etc. Dependency-based	Ruan *et al.* [2016], Abnar *et al.* [2018], Wang *et al.* [2020b] Murphy *et al.* [2012], Abnar *et al.* [2018], Wang *et al.* [2020b]
Multi-modal	Text, Visual, Audio-derived	Anderson *et al.* [2013], Anderson *et al.* [2015], Anderson *et al.* [2017], Bulat *et al.* [2017], Cao *et al.* [2020]
Context-aware	LSTM, ELMo, BERT, GPT, RoBERTa, etc.	Jain and Huth [2018], Gauthier and Levy [2019], Schwartz *et al.* [2019], Jat *et al.* [2019], Sun *et al.* [2021], Schwartz and Mitchell [2019], Schrimpf *et al.* [2020], Wang *et al.* [2020a]

Table 2: Concept representations and representative works

by the WordNet similarity. They also combined the WordNet extracted features with corpus based semantic features of the nouns. And they found that the combined features gave better results in predicting brain activity patterns.

The above studies suggest that the meaning defined by sensor-motor verbs may have a distinctive role in predicting brain activity. However, Devereux *et al.* [2010] reached different conclusions. Rather than using sensor-motor features, they extracted general feature-based conceptual representation from four different sources of information available in corpora and evaluated performance individually. Surprisingly, the study did not find any significant difference in performance between the four models. This suggested that general feature-based conceptual representations were equally capable of predicting activation to conceptual stimuli and placed no priority distinction on sensory-motor properties.

3.2 Distributional Representations

Subsequently, distributional representation semantics [Mikolov *et al.*, 2013; Pennington *et al.*, 2014] became popular and achieved great success in the NLP community. The distributed representation can be automatically generalized to novel situations and is also tunable to changing environment [Hinton *et al.*, 1986]. Thus researchers attempt to use distributed concept representation to correlate with brain activity. Most studies evaluated the correlation of brain activity with various concept representations to explore the brain concept representation mechanism, and evaluated the similarity of human and distributed concept representation.

Murphy *et al.* [2012] examined different corpus-based models and concentrated on which types of basic corpus pattern perform well on the neurosemantic decoding task. The study found that dependency parse-based features were the most effective and achieving accuracy higher than any published corpus-based model. The study also found that simple word features rich in directional information provided a near-optimal solution at a much lower computational cost.

Pereira *et al.* [2013] examined whether it was possible to learn a semantic space from a relatively small corpus that reflected semantic representations of those concepts in the human mind. They produced semantic features by topic models on a corpus of a few thousand Wikipedia articles about concrete or visualizable concepts, and showed that the learned features can outperform those in other brain decoding tasks.

Subsequent studies [Murphy *et al.*, 2011; Anderson *et al.*, 2013; Bingel *et al.*, 2016; Ruan *et al.*, 2016; Abnar *et al.*, 2018; Wang *et al.*, 2020b] tested various concept representation methods, including Word2Vec [Mikolov *et al.*, 2013], GloVe [Pennington *et al.*, 2014], fastText [Joulin *et al.*, 2016], Dependency [Levy and Goldberg, 2014], Meta-word and RWSGwn [Goikoetxea *et al.*, 2015]. The overall conclusion is that different word classes can be decoded most effectively with different word embeddings, and syntactically informed models (dependency) give the overall best performance.

3.3 Multi-modal Representations

Humans obtain multi-modal information from the real world and form conceptual representations in the brain. During the data collection procedure, subjects can be given text, image, and audio stimuli simultaneously and were also required to think of the properties of the stimuli. Thus, in addition to text-based features, some studies attempted to use multi-modal features, mainly vision and auditory, to predict brain activity [Ruan *et al.*, 2016; Anderson *et al.*, 2017; Bulat *et al.*, 2017; Wang *et al.*, 2020b]. The general approach to construct the visual feature is to use the deep neural network, such as ResNet [He *et al.*, 2015], VGG [Simonyan and Zisserman, 2015], to extract image features of the same class from ImageNet. The extracted features are averaged across image class so that the salient features of that class will be retained. Finally, the resulting features are used to correlate with brain activity. The process of extracting auditory features is similar. But the acoustic features here are extracted from real sounds, such as the jingle of keys.

Some studies [Ruan *et al.*, 2016; Anderson *et al.*, 2017; Bulat *et al.*, 2017; Wang *et al.*, 2020b] compared the performance of text, images, auditory features and their combinations in predicting brain activity pattern. The conclusion is consistent among different studies: 1. Each modal of features can be used to predict brain activity among the chance level significantly. 2. Image features or multi-modal features including image features give a better performance than others.

3.4 Abstract Concept

The studies surveyed above mainly focused on the decoding of concrete concepts. Fernandino *et al.* [2015] demonstrated that the prediction by using sensor-motor features was successful for concrete concepts but not for abstract ones. To investigate how abstract knowledge is organized in the brain, many studies have repeated the process of decoding concrete concepts (i.e., presenting the abstract stimuli to the subject and then correlating the brain activity). It is worth noting that instructing the subject in data collection is different when the stimulus is an abstract or concrete concept. With concrete concepts, subjects are usually instructed to think actively about the properties of the object, but eliciting properties are not so easy for abstract concepts. Subjects appear able to produce situation-related objects [James A., 1981; McRae *et al.*, 2005]. Therefore, subjects are often instructed to "think about situations that exemplify the object the word refers to" when stimuli are abstract concepts.

Anderson *et al.* [2012] studied the decoding of abstract concepts from brain activity. In the study, subjects were presented with 60 abstract concepts belonging to 6 categories (location, social role, event, communication, attribute and urabstract). As a comparison, subjects were also presented with 10 concrete concepts of 1 category (tools). The concepts were selected from two different domains (music and law). This study came to two conclusions : "(1) WordNet-style taxonomic categories for abstract concepts are at least cognitively relevant in that they can be distinguished from neural data. (2) In contrast to previous findings for concrete concepts, it is unable to detect a relationship between inter-representation of abstract concept categories in fMRI data and inter-representations in popular distributed semantic models."

There has been extensive subsequent research in abstract concept decoding. Wang *et al.* [2013] found that single-trial-based brain activity was sufficient to distinguish abstract versus concrete representations. In addition to distinguishing the concepts, Anderson *et al.* [2014] investigated how concreteness of the concepts can affect the decoding performance by varying concreteness of the stimuli. They found a clear concreteness effect: concrete concepts can be reliably predicted for unseen participants, but less concrete concepts can only be reliably predicted withing subjects. They also found that domain (e.g., law vs. music) can be better predicted for less concrete categories. Thus the author concluded that "both taxonomic and domain class distinctions are relevant for interpreting neural structuring of concrete and abstract concepts".

In addition to textual features, some work [Anderson *et al.*, 2013; Anderson *et al.*, 2015; Anderson *et al.*, 2017; Wang *et al.*, 2020b] adopted visual features to decode brain activity. The conclusions are consistent: both visual and textual models decode the more concrete nouns. Anderson *et al.* [2017] investigated how visually grounded and textual semantic models decode abstract and concrete concepts. It decoded the concrete and abstract concepts from brain activity by using text or image-derived features separately. The results suggest that both models significantly decod the most concrete nouns and the decoding accuracy is significantly greater using the text-based models for the most abstract nouns. The observation is also in line with the dual-coding theory [Paivio, 1971]: the human brain represents concrete concepts in terms of visual and linguistic code, whereas it represents abstract concepts only in linguistic code.

There is an arguable fact when decoding abstract concepts by using image-derived features [Anderson *et al.*, 2017]. Most studies use Google Image to select images for abstract concepts. It may not perform as well as concrete concepts, thus can lead to decreased performance in decoding accuracy.

3.5 Associative Thinking

Subjects were often instructed to think of the properties of the stimuli during the experiments. So brain activity was not only stimulated by stimuli but also influenced by association concepts. Small World of Words [De Deyne *et al.*, 2018] is a word association dataset, where more than $100K$ fluent English speakers were asked to list three associations for each target word. For example, when subjects were given the word *apple*, most people can think of *fruit*, *red*, *orange*, *tree*, and *pear*. Inspired by this, some studies [Bulat *et al.*, 2017; Cao *et al.*, 2020] attempted to use the associate word decoding brain activity. The results suggested that cognitively-motivated association-based models performed on par with or better than other linguistic models.

3.6 Incorporating Context

Most previous studies focus on word-level representations and thus ignore the influence of context on language understanding. Relatively little early work considered decoding whole sentences from brain activation patterns. This is largely because the experimental process of using sentences as stimuli is more complicated, and sentence representation remained a challenge for NLP at the dawn of brain decoding. With NLP development in recent years, some studies pay attention to relating sentence encoding with brain activity patterns [Jain and Huth, 2018; Gauthier and Levy, 2019; Jat *et al.*, 2019; Sun *et al.*, 2021]. There are two main types of sentence encoding. The first is the bag-of-words encoding, which sums the word embedding together and then uses the average value as the sentence encoding. It is mainly used as a baseline. Such sentence encoding ignores the order of words, and consequently the influence of syntax and context. The second is sequence models, such as LSTM [Hochreiter and Schmidhuber, 1997], Transformer [Vaswani *et al.*, 2017], and pre-trained language models like BERT [Devlin *et al.*, 2019].

Jain and Huth [2018] used LSTM [Hochreiter and Schmidhuber, 1997] language models to encode sentences and then correlate the sentence encoding with brain activity. The results suggested that contextual models performed significantly better than word embedding models. They also compared the effectiveness of models that used multiple LSTM layers output and context lengths. The results revealed a hierarchy of brain areas sensitive to different types of contextual information and different temporal receptive field sizes.

Gauthier and Levy [2019] fine-tuned a pre-trained BERT [Devlin *et al.*, 2019] on various natural language understanding (NLU) tasks to find out what NLU tasks lead to improvements in brain-decoding performance. However, the results suggested that none of the sentence encoding tasks tested yielded significant increases in brain decoding performance. Further analysis revealed that fine-tuning BERT with scrambled sentences yielded significant improvements in performance. The authors named such sentence encoding as syntax-light sentence representation. But this does not mean that syntactic information is not retained in brain activity. It is more likely that the brain activity data used in the study is fMRI, which is too coarse to preserve the syntactic trace.

4 Discussion

Decoding language (words, phrases, sentences) from brain activity can benefit both NLP and neuroscience. Past decades of studies have promoted this field a big step forward, however, we currently have only very limited understanding. This section discusses the main factors that limit the progress, and promising future directions to take.

4.1 Limitations

Limitation on brain activity data. The most commonly used brain activity data are fMRI, EEG, MEG, and fNIRS. Each type of data has its advantages and disadvantages. For example, fMRI has a very high spatial resolution and is very suitable for source analysis. This means that we can accurately pinpoint the position in the brain. If we want to find a specific concept that will elicit a response to a specific location in the brain, fMRI is the best choice. However, the extremely low time resolution of fMRI makes it unsuitable for sentence-level analysis. fMRI takes about two seconds to complete a scan. This is far lower than the speed at which humans can process language. So this makes it difficult for fMRI to capture syntactic information. As mentioned in Gauthier and Levy [2019], when using fMRI to decode sentences, the accuracy is higher after removing the syntactic information from the text-derived sentence features. Other studies have produced conflicting results when using fMRI for syntactic research. EEG can preserve rich syntactic information [Hale *et al.*, 2018] due to the high temporal resolution, but the extremely low spatial resolution restricts it from performing source analysis. fNIRS is a compromise option, the time resolution of fNIRS is better than fMRI, and the spatial resolution is better than EEG. Some studies [Zinszer *et al.*, 2017; Cao *et al.*, 2021] demonstrated that the fNIRS preserves information for concept decoding tasks. However, some believe that the balance of spatial and temporal resolution is not enough to compensate for the loss in both.

Individual differences. Humans form representations of concepts based on their own experiences. Therefore, the same concept typically causes different brain activity among people (e.g. igloos may never have been experienced by people in the tropics). These differences in experiences have caused huge differences between individuals in the same experiment. We can observe the high variance in performance among subjects [Mitchell *et al.*, 2008; Gauthier and Levy, 2019; Pereira *et al.*, 2018; Cao *et al.*, 2021]. Despite this, there are some commonalities in concept representations among subjects. Some studies [Cao *et al.*, 2021] use data from multiple subjects for training and then apply the trained model to new subjects. The results suggested that the model can decode the brain activity of an unseen subject to some extent. However, the performance will not improve all the time as the number of training subjects increases. Learning the commonality of the brain activity among subjects is inevitable, and we believe that future work can seek breakthroughs from two directions. First, we can try to use adversarial training. Adversarial training can be used to learn the commonalities between different distributions. Suppose we treat the brain activity of subjects as different distributions. It is possible to learn more commonalities by using adversarial training. Second, we can emulate the idea of pre-training in NLP and computer vision. We train a model using data from multiple subjects, then use the new subject data to fine-tune the pre-trained model. In this way, the model retains the commonality of most subjects and reflects the differences among them.

Linking hypothesis. Most studies in decoding brain activity are using pure linear transformation. As Gauthier and Levy [2019] pointed out "It is likely that some syntactic information — among other features of the input — is conserved in the fMRI signal but not readable by a linear decoder." Cao and Zhang [2019] used a three-layer neural network to decode concrete nouns, and the effect is better than linear regression. But this may be because neural networks inherently have better fitting functions. It cannot be explained that this is because the human representation of concepts is nonlinear.

Limitation in brain research. The most fundamental limitation is that the neural mechanism of human language processing has not been fully revealed in the current research. Consider computer vision as an analogy. Current research has studied the neural mechanisms of human visual pathways, and we can imitate visual pathways to build visual networks. For example, the convolution region in the most classic neural network CNN imitates the receptive field in the human visual system. However, we understand less about the neural mechanism in the language process, making it difficult to directly use neural networks to simulate human brain language processing. For example, the most widely used NLP neural networks (such as RNN and Transformers [Vaswani *et al.*, 2017]) simulate humans' behavior-i.e., eye movement- rather than neural activity in reading sentences. In general, the network in computer vision is closer to the human visual system, while the model in NLP temporarily stays at imitating human behavior in reading. This may be the reason for the faster development of the computer vision research field.

4.2 Future Directions

The future direction of this field can be developed from two perspectives. From the NLP perspective, the existing research can detect the syntactic traces [Hale *et al.*, 2018] from the neural signal. Then we should be able to use neural signals to fine-tune LMs to make it more human-like. It should be noted that fine-tuning LMs requires at least thousands of data entries, and neural data is not easy to obtain. Jat *et al.* [2019] generate synthetic brain data and show that it helps in improving subsequent stimuli decoding task accuracy. Future research can attempt to fine-tune LM using synthetic brain data.

From the perspective of neuroscience, future research can try to use deep NLP models to explore the neural mechanisms of language processing. There is a successful study [Ratan Murty *et al.*, 2020] in the field of human vision research recently. There is an area in the human brain named fusiform face area (FFA) [Kanwisher *et al.*, 1997] responsible for processing face-related information. However, it is difficult to prove that FFA deals with face-related processing exclusively because it is impossible to present subjects with all the objects in the world. Instead, the researchers trained a visual network by using neural activity data from FFA. Then they tested the network with millions of images from ImageNet [Deng *et al.*, 2009] and found that this network only has a strong activation for face pictures. This largely proves that FFA only deals with facial information in the brain.

A recent study [Schrimpf *et al.*, 2020] tested state-of-the-art artificial neutral network (ANN) LMs to predict language processing in the brain. The results indicate a strong correlation between ANN LMs and the brain activity, opening a window for neuroscience. Although both the human brain and ANN LMs are black-box models, the variables of LMs can be altered to generate various outputs. This allows multiple outputs to be correlated with brain activity without the need to experiment with humans, providing an option to probe the neural mechanism of language processing in the human brain.

References

[Abnar *et al.*, 2018] Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proc. of the 8th Workshop on CMCL 2018*, pages 57–66. ACL, 2018.

[Anderson *et al.*, 2012] Andrew Anderson, Tao Yuan, Brian Murphy, and Massimo Poesio. On discriminating fMRI representations of abstract WordNet taxonomic categories. In *Proc. of the 3rd Workshop on CogALex*, pages 21–32. COLING, 2012.

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

[Anderson *et al.*, 2014] Andrew J. Anderson, Brian Murphy, and Massimo Poesio. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of cognitive neuroscience*, 26(3):658–681, 2014.

[Anderson *et al.*, 2015] Andrew James Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. 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.

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

[Babaeian Jelodar *et al.*, 2010] Ahmad Babaeian Jelodar, Mehrdad Alizadeh, and Shahram Khadivi. WordNet based features for predicting brain activity associated with meanings of nouns. In *Proc. of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 18–26. ACL, 2010.

[Bingel *et al.*, 2016] Joachim Bingel, Maria Barrett, and Anders Søgaard. Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proc. of the ACL*, pages 747–755. ACL, 2016.

[Brennan and Hale, 2019] Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

[Brennan *et al.*, 2016] Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94, 2016.

[Bulat *et al.*, 2017] Luana Bulat, Stephen Clark, and Ekaterina Shutova. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proc. of EMNLP*, pages 1081–1091. ACL, 2017.

[Cao and Zhang, 2019] Lu Cao and Yue Zhang. Investigating lexical and semantic cognition by using neural network to encode and decode brain imaging. In *HBAI*, pages 84–100. Springer, 2019.

[Cao *et al.*, 2020] Lu Cao, Yulong Chen, Dandan Huang, and Yue Zhang. Investigating rich feature sources for conceptual representation encoding. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 12–22. Association for Computational Linguistics, 2020.

[Cao *et al.*, 2021] Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. Brain decoding using fnirs. In *Proc. of the AAAI*, volume 33, pages 7047–7054, 2021.

[Cox and Savoy, 2003] David D Cox and Robert Savoy. fmri brain reading: detecting and classifying distributed patterns of fmri activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003.

[De Deyne *et al.*, 2018] Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. The ”small world of words” english word association norms for over 12000 cue words. *Behavior Research Methods*, 2018.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

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

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186. ACL, 2019.

[Dyer *et al.*, 2016] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proc. of NAACL*, pages 199–209. ACL, 2016.

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

[Frank *et al.*, 2015] Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11, 2015.

[Gauthier and Levy, 2019] Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. In *Proc. of EMNLP-IJCNLP*, pages 529–539. ACL, 2019.

[Goikoetxea *et al.*, 2015] Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. Random walks and neural network language models on knowledge bases. In *Proc. of NAACL*, pages 1434–1439. ACL, 2015.

[Hale *et al.*, 2015] John Hale, David Lutz, Wen-Ming Luh, and Jonathan Brennan. Modeling fMRI time courses with linguistic structure at various grain sizes. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 89–97, Denver, Colorado, June 2015. Association for Computational Linguistics.

[Hale *et al.*, 2018] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proc. of the ACL*, pages 2727–2736. ACL, 2018.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[Hinton *et al.*, 1986] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. *Distributed Representations*, page 77–109. MIT Press, 1986.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Hollenstein *et al.*, 2018] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

[Hollenstein *et al.*, 2019a] Nora Hollenstein, Maria Barrett, M. Troendle, Francesco Bigiolli, N. Langer, and Ce Zhang. Advancing nlp with cognitive language processing signals. *ArXiv*, abs/1904.02682, 2019.

[Hollenstein *et al.*, 2019b] Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. CogniVal: A framework for cognitive word embedding evaluation. In *Proc. of CoNLL*, pages 538–549. ACL, 2019.

[Hollenstein *et al.*, 2020] Nora Hollenstein, Maria Barrett, and Lisa Beinborn. Towards best practices for leveraging human language processing signals for natural language processing. In *Proc. of the Second Workshop on LiNCR*, pages 15–27. ELRA, 2020.

[Huth *et al.*, 2016] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[Jain and Huth, 2018] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *NeurIPS*, volume 31. Curran Associates, Inc., 2018.

[James A., 1981] Hampton James A. An investigation of the nature of abstract concepts. *Memory & Cognition*, 9:149–156, 1981.

[Jat *et al.*, 2019] Sharmistha Jat, Hao Tang, Partha Talukdar, and Tom Mitchell. Relating simple sentence representations in deep neural networks and the brain. In *Proc. of the ACL*, pages 5137–5154. ACL, 2019.

[Joulin *et al.*, 2016] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[Kanwisher *et al.*, 1997] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

[Kuncoro *et al.*, 2017] Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. What do recurrent neural network grammars learn about syntax? In *Proc. of EACL*, pages 1249–1258. ACL, 2017.

[Levy and Goldberg, 2014] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proc. of the ACL)*, pages 302–308. ACL, 2014.

[McRae *et al.*, 2005] K McRae, G S Cree, M S Seidenberg, and C McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behav Res Methods*, 37:547–559, 2005.

[Mikolov *et al.*, 2013] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[Minnema and Herbelot, 2019] Gosse Minnema and Aurélie Herbelot. From brain space to distributional space: The perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161, Florence, Italy, July 2019. Association for Computational Linguistics.

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

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

[Murphy *et al.*, 2011] Brian Murphy, Massimo Poesio, Francesca Bovolo, Lorenzo Bruzzone, Michele Dalponte, and Heba Lakany. Eeg decoding of semantic category reveals distributed representations for single concepts. *Brain and language*, 117:12–22, 2011.

[Murphy *et al.*, 2012] Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In **SEM 2012*, pages 114–123. ACL, 2012.

[Oseki and Asahara, 2020] Yohei Oseki and Masayuki Asahara. Design of BCCWJ-EEG: Balanced corpus with human electroencephalography. In *Proceedings of the 12th LREC*, pages 189–194. ELRA, 2020.

[Paivio, 1971] Allan. Paivio. Imagery and verbal processes. In *Imagery and verbal processes*. Holt, Rinehart & Winston., 1971.

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

[Pereira *et al.*, 2001] Francisco Pereira, Marcel Just, and Tom Mitchell. Distinguishing natural language processes on the basis of fmri-measured brain activation. In *European Conference on PKDD*, pages 374–385. Springer, 2001.

[Pereira *et al.*, 2010] Francisco Pereira, Matthew Botvinick, and Greg Detre. Learning semantic features for fMRI data from definitional text. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 1–9, Los Angeles, USA, June 2010. Association for Computational Linguistics.

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

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

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *bioRxiv*, 2019.

[Rafidi, 2014] Nicole Rafidi. The role of syntax in semantic processing: A study of active and passive sentences. In *Poster and Oral Session at HBM annual meeting 2015*, 2014.

[Ratan Murty *et al.*, 2020] N. Apurva Ratan Murty, Santani Teng, David Beeler, Anna Mynick, Aude Oliva, and Nancy Kanwisher. Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, 2020.

[Ruan *et al.*, 2016] Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. Exploring semantic representation in brain activity using word embeddings. In *Proc. of EMNLP*, pages 669–679. ACL, 2016.

[Schrimpf *et al.*, 2020] Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial neural networks accurately predict language processing in the brain. *bioRxiv*, 2020.

[Schwartz and Mitchell, 2019] Dan Schwartz and Tom Mitchell. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. In *Proc. of NAACL*, pages 43–57. ACL, 2019.

[Schwartz *et al.*, 2019] Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. In *NeurIPS*, pages 14100–14110, 2019.

[Simonyan and Zisserman, 2015] Karen. Simonyan and Andrew. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[Søgaard, 2016] Anders Søgaard. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Sudre *et al.*, 2012] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463, 2012.

[Sun *et al.*, 2021] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):589–603, 2021.

[Toneva and Wehbe, 2019] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *NeurIPS*, pages 14928–14938, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[Wang *et al.*, 2013] Jing Wang, Laura B. Baucom, and S. V. Shinkareva. Decoding abstract and concrete concept representations based on single-trial fmri data. *Human Brain Mapping*, 34, 2013.

[Wang *et al.*, 2020a] Shaonan Wang, Jiajun Zhang, Nan Lin, and Chengqing Zong. Probing brain activation patterns by dissociating semantics and syntax in sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9201–9208, 2020.

[Wang *et al.*, 2020b] Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256 – 272, 2020.

[Wehbe *et al.*, 2014a] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11), 2014.

[Wehbe *et al.*, 2014b] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

[Wehbe *et al.*, 2014c] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proc. of EMNLP*, pages 233–243, 2014.

[Xu *et al.*, 2016] Haoyan Xu, Brian Murphy, and Alona Fyshe. BrainBench: A brain-image test suite for distributional semantic models. In *Proc. of EMNLP*, pages 2017–2021. ACL, 2016.

[Zinszer *et al.*, 2017] Benjamin D Zinszer, Laurie Bayet, Lauren L Emberson, Rajeev DS Raizada, and Richard N Aslin. Decoding semantic representations from functional near-infrared spectroscopy signals. *Neurophotonics*, 5(1):011003, 2017.