Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey)

Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gael Jobard, Frédéric Alexandre, Xavier Hinaut

Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey)

Subba Reddy Oota[1,2] , **Manish Gupta**[3,4] , **Raju S. Bapi**[3] , **Gael Jobard**[2]
Frederic Alexandre[1,2] , **Xavier Hinaut**[1,2]
[1]INRIA, Bordeaux, France, [2]University of Bordeaux, France, [3]IIIT Hyderabad, India, [4]Microsoft, Hyderabad, India
subba-reddy.oota@inria.fr, gmanish@microsoft.com, raju.bapi@iiit.ac.in, gael.jobard@u-bordeaux.fr, frederic.alexandre@inria.fr, xavier.hinaut@inria.fr

Abstract

How does the brain represent different modes of information? Can we design a system that automatically understands what the user is thinking? Such questions can be answered by studying brain recordings like functional magnetic resonance imaging (fMRI). As a first step, the neuroscience community has contributed several large cognitive neuroscience datasets related to passive reading/listening/viewing of concept words, narratives, pictures and movies. Encoding and decoding models using these datasets have also been proposed in the past two decades. These models serve as additional tools for basic research in cognitive science and neuroscience. Encoding models aim at generating fMRI brain representations given a stimulus automatically. They have several practical applications in evaluating and diagnosing neurological conditions and thus also help design therapies for brain damage. Decoding models solve the inverse problem of reconstructing the stimuli given the fMRI. They are useful for designing brain-machine or brain-computer interfaces. Inspired by the effectiveness of deep learning models for natural language processing, computer vision, and speech, recently several neural encoding and decoding models have been proposed. In this survey, we will first discuss popular representations of language, vision and speech stimuli, and present a summary of neuroscience datasets. Further, we will review popular deep learning based encoding and decoding architectures and note their benefits and limitations. Finally, we will conclude with a brief summary and discussion about future trends. Given the large amount of recently published work in the 'computational cognitive neuroscience' community, we believe that this survey nicely organizes the plethora of work and presents it as a coherent story.

1 Introduction

Neuroscience is the field of science that studies the structure and function of the nervous system of different species. It involves answering interesting questions like the following[1]. (1) How learning occurs during adolescence, and how it differs from the way adults learn and form memories. (2) Which specific cells in the brain (and what connections they form with other cells), have a role in how memories are formed? (3) How animals cancel out irrelevant information arriving from the senses and focus only on information that matters. (4) How do humans make decisions? (5) How humans develop speech and learn languages. Neuroscientists study diverse topics that help us understand how the brain and nervous system work.

Motivation: The central aim of neuroscience is to unravel how the brain represents information and processes it to carry out various tasks (visual, linguistic, auditory, etc.). Deep neural networks (DNN) offer a computational medium to capture the unprecedented complexity and richness of brain activity. *Encoding* and *decoding* stated as computational problems succinctly encapsulate this puzzle. As the previous surveys systematically explore the brain encoding and decoding studies with respect to only language [Cao *et al.*, 2021; Karamolegkou *et al.*, 2023], this survey summarizes the latest efforts in how DNNs begin to solve these problems and thereby illuminate the computations that the unreachable brain accomplishes effortlessly.

Brain encoding and decoding: Two main tasks studied in cognitive neuroscience are brain encoding and brain decoding, as shown in Figure 1. Encoding is the process of learning the mapping e from the stimuli S to the neural activation F. The mapping can be learned using features engineering or deep learning. On the other hand, decoding constitutes learning mapping d, which predicts stimuli S back from the brain activation F. However, in most cases, brain decoding aims at predicting a stimulus representation R rather than actually reconstructing S. In both cases, the first step is to learn a semantic representation R of the stimuli S at the train time. Next, for encoding, a regression function $e : R \rightarrow F$ is trained. For decoding, a function $d : F \rightarrow R$ is trained. These functions e and d can then be used at test time to process new stimuli and brain activations, respectively.

Techniques for recording brain activations: Popular techniques for recording brain activations include single Micro-

[1]https://zuckermaninstitute.columbia.edu/file/5184/download?token=qzId8vyR

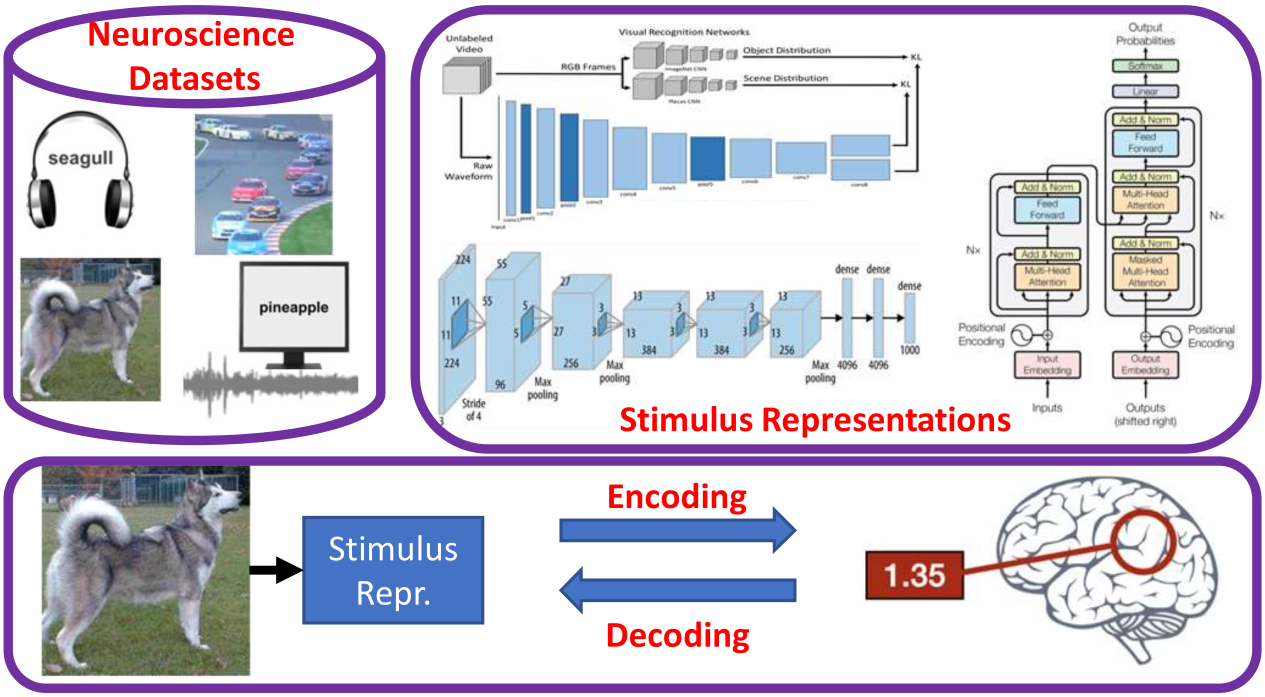

Figure 1: Computational Cognitive Neuroscience of Brain Encoding and Decoding: Datasets & Stimulus Representations

Electrode (ME), Micro-Electrode array (MEA), Electro-Cortico Graphy (ECoG), Positron emission tomography (PET), functional MRI (fMRI), Magneto-encephalography (MEG), Electro-encephalography (EEG) and Near-Infrared Spectroscopy (NIRS). These techniques differ in their spatial resolution of neural recording and temporal resolution.

fMRIs enable high spatial but low time resolution. Hence, they are good for examining which parts of the brain handle critical functions. fMRI takes 1-4 seconds to complete a scan. This is far lower than the speed at which humans can process language. On the other hand, both MEG and EEG have high time but low spatial resolution. They can preserve rich syntactic information [Hale *et al.*, 2018] but cannot be used for source analysis. fNIRS are a compromise option. Their time resolution is better than fMRI, and spatial resolution is better than EEG. However, this spatial and temporal resolution balance may not compensate for the loss in both.

Stimulus Representations: Neuroscience datasets contain stimuli across various modalities: text, visual, audio, video and other multimodal forms. Representations differ based on modality. Older methods for *text-based stimulus representation* include text corpus co-occurrence counts, topic models, syntactic, and discourse features. In recent times, both semantic and experiential attribute models have been explored for text-based stimuli. Semantic representation models include distributed word embeddings, sentence representation models, recurrent neural networks (RNNs), and Transformer-based language models. Experiential attribute models represent words in terms of human ratings of their degree of association with different attributes of experience, typically on a scale of 0-6 or binary. Older methods for *visual stimulus representation* used visual field filter bank and Gabor wavelet pyramid for visual stimuli, but recent methods use models like ImageNet-pretrained convolutional neural networks (CNNs) and concept recognition methods. For *audio stimuli*, phoneme rate and the presence of phonemes have been leveraged, besides deep learning models like SoundNet. Finally, for multimodal stimulus representations, researchers have used both early fusion and late fusion deep learning methods. In the early fusion methods, information across modalities is combined in the early steps of processing. While in late fusion, the combination is performed only at the end. We discuss stimulus representation methods in detail in Sec. 2.

Naturalistic Neuroscience Datasets: Several neuroscience datasets have been proposed across modalities (see Figure 2). These datasets differ in terms of the following criteria: (1) Method for recording activations: fMRI, EEG, MEG, etc. (2) Repetition time (TR), i.e. the sampling rate. (3) Characteristics of fixation points: location, color, shape. (4) Form of stimuli presentation: text, video, audio, images, or other multimodality. (5) Task that participant performs during recording sessions: question answering, property generation, rating quality, etc. (6) Time given to participants for the task, e.g., 1 minute to list properties. (7) Demography of participants: males/females, sighted/blind, etc. (8) Number of times the response to stimuli was recorded. (9) Natural language associated with the stimuli. We discuss details of proposed datasets in Sec. 3.

Brain Encoding: Other than using the standard stimuli representation architectures, brain encoding literature has focused on studying a few important aspects: (1) Which models lead

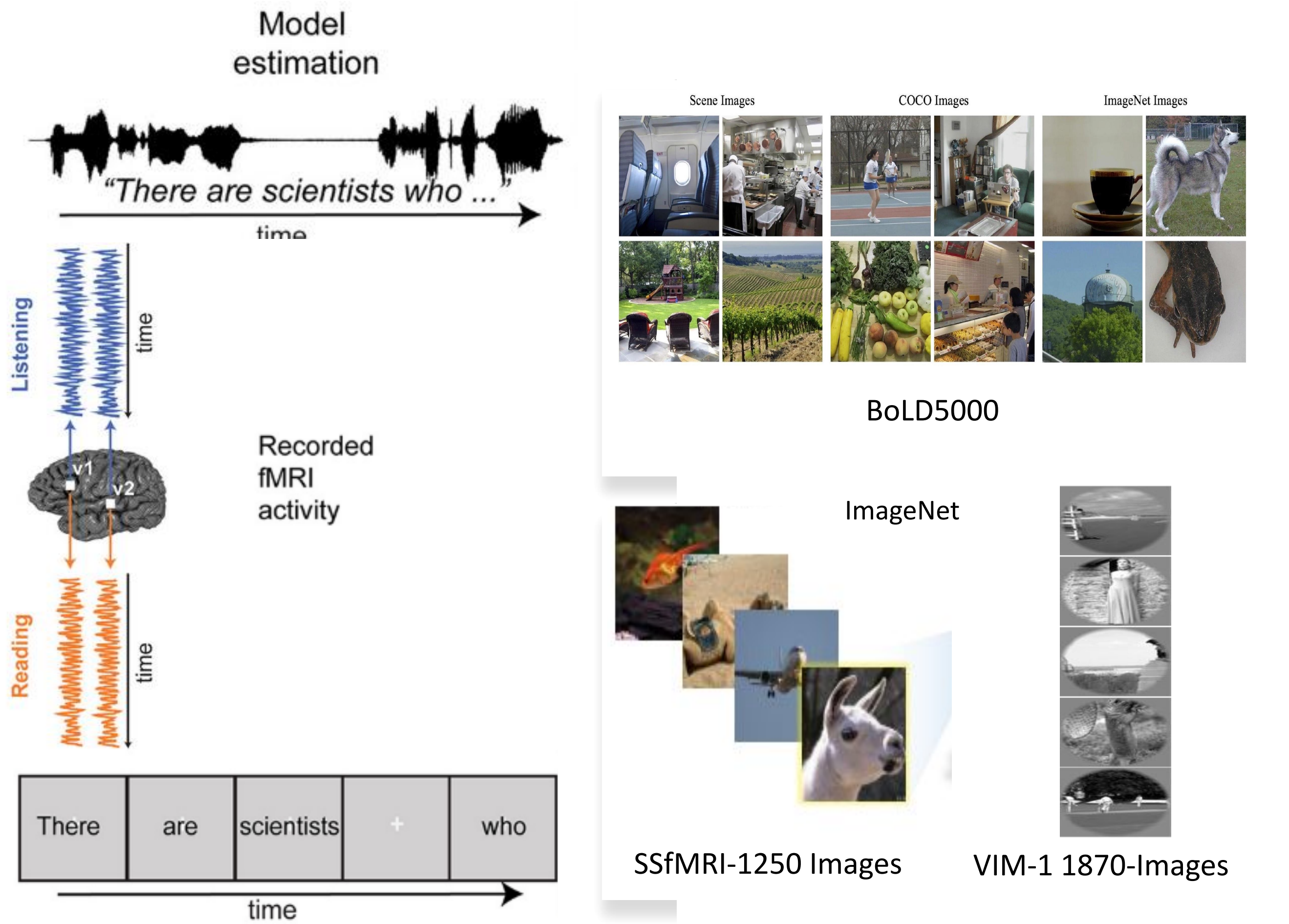

Figure 2: Representative Samples of Naturalistic Brain Dataset: (LEFT) Brain activity recorded when subjects are reading and listening to the same narrative (*Deniz et al. 2019*), and (RIGHT) example naturalistic image stimuli from various public repositories: BOLD5000 (*Chang et al. 2019*), SSfMRI (*Beliy et al., 2019*), and VIM-1 (*Kay et al., 2008*).

to better predictive accuracy across modalities? (2) How can we disentangle the contributions of syntax and semantics from language model representations to the alignment between brain recordings and language models? (3) Why do some representations lead to better brain predictions? How are deep learning models and brains aligned in terms of their information processing pipelines? (4) Does joint encoding of task and stimulus representations help? We discuss these details of encoding methods in Sec. 5.

Brain Decoding: Ridge regression is the most popular brain decoder. Recently, a fully connected layer [Beliy *et al.*, 2019] or multi-layered perceptrons (MLPs) [Sun *et al.*, 2019] have also been used. While older methods attempted to decode to a vector representation using stimuli of a single mode, newer methods focus on multimodal stimuli decoding [Pereira *et al.*, 2016; Oota *et al.*, 2022c]. Decoding using Transformers [Gauthier and Levy, 2019; Toneva and Wehbe, 2019; Défossez *et al.*, 2022; Tang *et al.*, 2022], and decoding to actual stimuli (word, passage, image, dialogues) have also been explored. We discuss details of these decoding methods in Sec. 6.

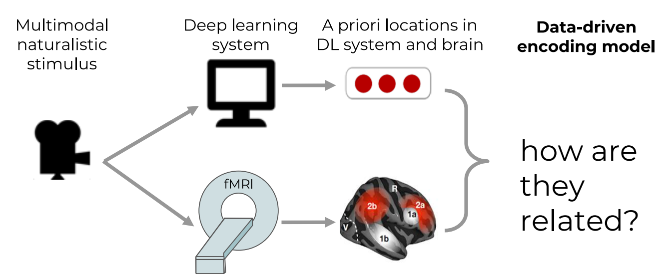

Figure 3: Alignment between deep learning systems and human brains *[Toneva et al. 2019]*.

Computational Cognitive Science (CCS) Research goals: CCS researchers have primarily focused on two main areas [Doerig *et al.*, 2022] (also, see Figure 3). (1) Improving predictive Accuracy. In this area, the work is around the following questions. (a) Compare feature sets: Which feature set provides the most faithful reflection of the neural representational space? (b) Test feature decodability: "Does neu-

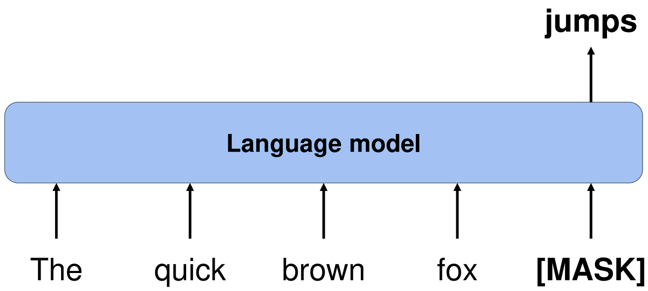

Figure 4: Language Model

ral data Y contain information about features X?" (c) Build accurate models of brain data: The aim is to enable simulation of neuroscience experiments. (2) Interpretability. In this area, the work is around the following questions. (a) Examine individual features: Which features contribute most to neural activity? (b) Test correspondences between representational spaces: "CNNs vs ventral visual stream" or "Two text representations". (c) Interpret feature sets: Do features X, generated by a known process, accurately describe the space of neural responses Y? Do voxels respond to a single feature or exhibit mixed selectivity? (d) How does the mapping relate to other models or theories of brain function? We discuss some of these questions in Sections 5 and 6.

2 Stimulus Representations

In this section, we discuss types of stimulus representations that have been proposed in the literature across different modalities: text, visual, audio, video and other multimodal stimuli.

Text Stimulus Representations: Older methods for text-based stimuli representation include text corpus co-occurrence counts [Mitchell *et al.*, 2008; Pereira *et al.*, 2013; Huth *et al.*, 2016], topic models [Pereira *et al.*, 2013], syntactic features and discourse features [Wehbe *et al.*, 2014]. In recent times, for text-based stimuli, both semantic models as well as experiential attribute models have been explored. Semantic representation models include word embedding methods [Pereira *et al.*, 2018; Wang *et al.*, 2020; Pereira *et al.*, 2016; Toneva and Wehbe, 2019; Anderson *et al.*, 2017a; Oota *et al.*, 2018], sentence representation models (see Figure 4) [Sun *et al.*, 2020; Sun *et al.*, 2019; Toneva and Wehbe, 2019], RNNs [Jain and Huth, 2018; Oota *et al.*, 2019] and Transformer methods [Gauthier and Levy, 2019; Toneva and Wehbe, 2019; Schwartz *et al.*, 2019; Schrimpf *et al.*, 2021a; Antonello *et al.*, 2021; Oota *et al.*, 2022b; Aw and Toneva, 2022]. Popular word embedding methods include textual (i.e., Word2Vec, fastText, and GloVe), linguistic (i.e., dependency), conceptual (i.e., RWSGwn and ConceptNet), contextual (i.e., ELMo). Popular sentence embedding models include average, max, concat of avg and max, SIF, fairseq, skip, GenSen, InferSent, ELMo, BERT, RoBERTa, USE, QuickThoughts and GPT-2. Transformer-based methods include pretrained BERT with various NLU tasks, finetuned BERT, Transformer-XL, GPT-2, BART, BigBird, LED, and LongT5. Experiential attribute models represent words in terms of human ratings of their degree of association with different attributes of experience, typically on a scale of 0-6 [Anderson *et al.*, 2019; Anderson *et al.*, 2020; Berezutskaya *et al.*, 2020; Just *et al.*, 2010; Anderson *et al.*, 2017b] or binary [Handjaras *et al.*, 2016; Wang *et al.*, 2017].

Visual Stimulus Representations: For visual stimuli, older methods used visual field filter bank [Thirion *et al.*, 2006; Nishimoto *et al.*, 2011] and Gabor wavelet pyramid [Kay *et al.*, 2008; Naselaris *et al.*, 2009]. Recent methods use models like CNNs [Du *et al.*, 2020; Beliy *et al.*, 2019; Anderson *et al.*, 2017a; Yamins *et al.*, 2014; Nishida *et al.*, 2020] and concept recognition models [Anderson *et al.*, 2020].

Audio Stimuli Representations: For audio stimuli, phoneme rate and presence of phonemes have been leveraged [Huth *et al.*, 2016]. Recently, authors in [Nishida *et al.*, 2020] used features from an audio deep learning model called SoundNet for audio stimuli representation.

Multimodal Stimulus Representations: To jointly model the information from multimodal stimuli, recently, various multimodal representations have been used. These include processing videos using audio+image representations like VGG+SoundNet [Nishida *et al.*, 2020] or using image+text combination models like GloVe+VGG and ELMo+VGG in [Wang *et al.*, 2020]. Recently, the usage of multimodal text+vision models like CLIP, LXMERT, and VisualBERT was proposed in [Oota *et al.*, 2022d].

3 Naturalistic Neuroscience Datasets

We discuss the popular text, visual, audio, video and other multimodal neuroscience datasets that have been proposed in the literature. Table 1 shows a detailed overview of brain recording type, language, stimulus, number of subjects (|S|) and the task across datasets of different modalities. Figure 2 shows examples from a few datasets.

Text Datasets: These datasets are created by presenting words, sentences, passages or chapters as stimuli. Some of the text datasets include Harry Potter Story [Wehbe *et al.*, 2014], ZUCO EEG [Hollenstein *et al.*, 2018] and datasets proposed in [Handjaras *et al.*, 2016; Anderson *et al.*, 2017a; Anderson *et al.*, 2019; Wehbe *et al.*, 2014]. In [Handjaras *et al.*, 2016], participants were asked to verbally enumerate in one minute the properties (features) that describe the entities the words refer to. There were four groups of participants: 5 sighted individuals were presented with a pictorial form of the nouns, 5 sighted individuals with a verbal-visual (i.e., written Italian words) form, 5 sighted individuals with a verbal auditory (i.e., spoken Italian words) form, and 5 congenitally blind with a verbal auditory form. Data proposed by [Anderson *et al.*, 2017a] contains 70 Italian words taken from seven taxonomic categories (abstract, attribute, communication, event/action, person/social role, location, object/tool) in the law and music domain. The word list contains concrete as well as abstract words. ZUCO dataset [Hollenstein *et al.*, 2018] contains sentences for which fMRIs were obtained for 3 tasks: normal reading of movie reviews, normal reading of Wikipedia sentences and task-specific reading of Wikipedia sentences. For this dataset curation, sentences were presented

Table 1: Naturalistic Neuroscience Datasets

	Dataset	Authors	Type	Lang.	Stimulus	|S|	Task
Text	Harry Potter	[Wehbe *et al.*, 2014]	fMRI/ MEG	English	Reading Chapter 9 of Harry Potter and the Sorcerer's Stone	9	Story understanding
		[Handjaras *et al.*, 2016]	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns, four times	20	Property Generation
		[Anderson *et al.*, 2017a]	fMRI	Italian	Reading 70 concrete and abstract nouns from law/music, five times	7	Imagine a situation with noun
	ZuCo	[Hollenstein *et al.*, 2018]	EEG	English	Reading 1107 sentences with 21,629 words from movie reviews	12	Rate movie quality
	240 Sentences with Content Words	[Anderson *et al.*, 2019]	fMRI	English	Reading 240 active voice sentences describing everyday situations	14	Passive reading
	BCCWJ-EEG	[Oseki and Asahara, 2020]	EEG	Japanese	Reading 20 newspaper articles for ∼30-40 minutes	40	Passive reading
	Subset Moth Radio Hour	[Deniz *et al.*, 2019]	fMRI	English	Reading 11 stories	9	Passive reading and Listening
Visual		[Thirion *et al.*, 2006]	fMRI	-	Viewing rotating wedges (8 times), expanding/contracting rings (8 times), rotating 36 Gabor filters (4 times), grid (36 times)	9	Passive viewing
	Vim-1	[Kay *et al.*, 2008]	fMRI	-	Viewing sequences of 1870 natural photos	2	Passive viewing
	Generic Object Decoder	[Horikawa and Kamitani, 2017]	fMRI	-	Viewing 1,200 images from 150 object categories; 50 images from 50 object categories; imagery 10 times	5	Repetition detection
	BOLD5000	[Chang *et al.*, 2019]	fMRI	-	Viewing 5254 images depicting real-world scenes	4	Passive viewing
	Algonauts	[Cichy *et al.*, 2019]	fMRI/ MEG	-	Viewing 92 silhouette object images and 118 images of objects on natural background	15	Passive viewing
	NSD	[Allen *et al.*, 2022]	fMRI	-	Viewing 73000 natural scenes	8	Passive viewing
	THINGS	[Hebart *et al.*, 2022]	fMRI/ MEG	-	Viewing 31188 natural images	8	Passive viewing
Audio		[Handjaras *et al.*, 2016]	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns, 4 times	20	Property Generation
	The Moth Radio Hour	[Huth *et al.*, 2016]	fMRI	English	Listening eleven 10-minute stories	7	Passive Listening
		[Brennan and Hale, 2019]	EEG	English	Listening Chapter one of Alice's Adventures in Wonderland (2,129 words in 84 sentences) as read by Kristen McQuillan	33	Question answering
		[Anderson *et al.*, 2020]	fMRI	English	Listening one of 20 scenario names, 5 times	26	Imagine personal experiencs
	Narratives	[Nastase *et al.*, 2021]	fMRI	English	Listening 27 diverse naturalistic spoken stories. 891 functional scans	345	Passive Listening
	Natural Stories	[Zhang *et al.*, 2020]	fMRI	English	Listening Moth-Radio-Hour naturalistic spoken stories.	19	Passive Listening
	The Little Prince	[Li *et al.*, 2021]	fMRI	English	Listening audiobook for about 100 minutes.	112	Passive Listening
	MEG-MASC	[Gwilliams *et al.*, 2022]	MEG	English	Listening two hours of naturalistic stories. 208 MEG sensors	27	Passive Listening
Video	BBC's Doctor Who	[Seeliger *et al.*, 2019]	fMRI	English	Viewing spatiotemporal visual and auditory videos (30 episodes). 120.8 whole-brain volumes (∼23 h) of single-presentation data, and 1.2 volumes (11 min) of repeated narrative short episodes. 22 repetitions	1	Passive viewing
	Japanese Ads	[Nishida *et al.*, 2020]	fMRI	Japanese	Viewing 368 web and 2452 TV Japanese ad movies (15-30s). 7200 train and 1200 test fMRIs for web; fMRIs from 420 ads.	52	Passive viewing
	Pippi Langkous	[Berezutskaya *et al.*, 2020]	ECoG	Swedish/ Dutch	Viewing 30 s excerpts of a feature film (in total, 6.5 min long), edited together for a coherent story	37	Passive viewing
	Algonauts	[Cichy *et al.*, 2021]	fMRI	English	Viewing 1000 short video clips (3 sec each)	10	Passive viewing
	Natual Short Clips	[Huth *et al.*, 2022]	fMRI	English	Watching natural short movie clips	5	Passive viewing
	Natual Short Clips	[Lahner *et al.*, 2023]	fMRI	English	Watching 1102 natural short video clips	10	Passive viewing
Other Multimodal	60 Concrete Nouns	[Mitchell *et al.*, 2008]	fMRI	English	Viewing 60 different word-picture pairs from 12 categories, 6 times each	9	Passive viewing
		[Sudre *et al.*, 2012]	MEG	English	Reading 60 concrete nouns along with line drawings. 20 questions per noun lead to 1200 examples.	9	Question answering
		[Zinszer *et al.*, 2018]	fNIRS	English	8 concrete nouns (audiovisual word and picture stimuli): bunny, bear, kitty, dog, mouth, foot, hand, and nose; 12 times repeated.	24	Passive viewing and listening
	Pereira	[Pereira *et al.*, 2018]	fMRI	English	Viewing 180 Words with Picture, Sentences, word clouds; reading 96 text passages; 72 passages. 3 times repeated.	16	Passive viewing and reading
		[Cao *et al.*, 2021]	fNIRS	Chinese	Viewing and listening 50 concrete nouns from 10 semantic categories.	7	Passive viewing and listening
	Neuromod	[Boyle *et al.*, 2020]	fMRI	English	Watching TV series (Friends, Movie10)	6	Passive viewing and listening

to the subjects in a naturalistic reading scenario. A complete sentence is presented on the screen. Subjects read each sentence at their own speed, i.e., the reader determines for how long each word is fixated and which word to fixate next.

Visual Datasets: Older visual datasets were based on binary visual patterns [Thirion *et al.*, 2006]. Recent datasets contain natural images. Examples include Vim-1 [Kay *et al.*, 2008], BOLD5000 [Chang *et al.*, 2019], Algonauts [Cichy *et al.*, 2019], NSD [Allen *et al.*, 2022], Things-data[Hebart *et al.*, 2022], and the dataset proposed in [Horikawa and Kamitani, 2017]. BOLD5000 includes ∼20 hours of MRI scans per each of the four participants. 4,916 unique images were used as stimuli from 3 image sources. Algonauts contains two sets of training data, each consisting of an image set and brain activity in RDM format (for fMRI and MEG). Training set 1 has 92 silhouette object images, and training set 2 has 118 object images with natural backgrounds. Testing data consists of 78 images of objects on natural backgrounds. Most of the visual datasets involve passive viewing, but the dataset in [Horikawa and Kamitani, 2017] involved the participant doing the one-back repetition detection task.

Audio Datasets: Most of the proposed audio datasets are in English [Huth *et al.*, 2016; Brennan and Hale, 2019; Anderson *et al.*, 2020; Nastase *et al.*, 2021], while there is one [Handjaras *et al.*, 2016] on Italian. The participants were involved in a variety of tasks while their brain activations were measured: Property generation [Handjaras *et al.*, 2016], passive listening [Huth *et al.*, 2016; Nastase *et al.*, 2021], question answering [Brennan and Hale, 2019] and imagining themselves personally experiencing common scenarios [Anderson *et al.*, 2020]. In the last one, participants underwent fMRI as they reimagined the scenarios (e.g., resting, reading, writing, bathing, etc.) when prompted by standardized cues. Narratives [Nastase *et al.*, 2021] used 17 different stories as stimuli. Across subjects, it is 6.4 days worth of recordings.

Video Datasets: Recently, video neuroscience datasets have also been proposed. These include BBC's Doctor Who [Seeliger *et al.*, 2019], Japanese Ads [Nishida *et al.*, 2020], Pippi

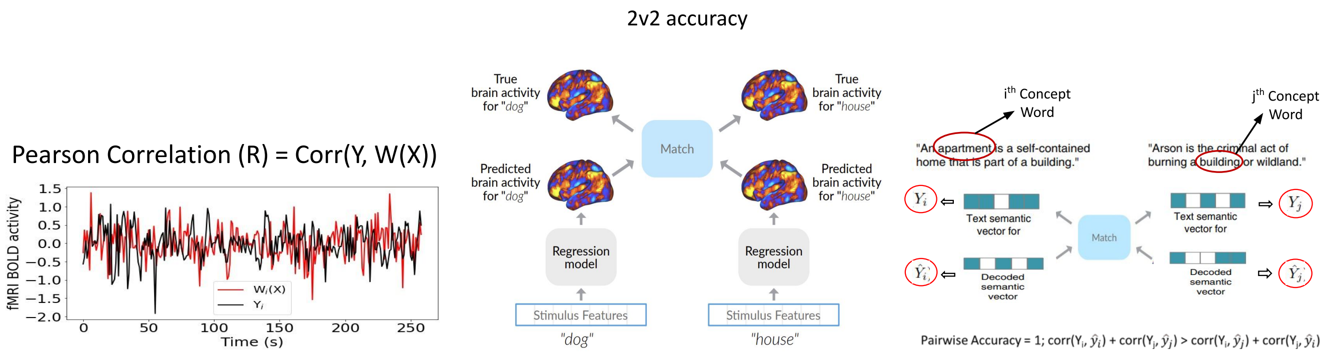

Figure 5: Evaluation Metrics for Brain Encoding and Decoding. (LEFT) Pearson Correlation, (MIDDLE) 2V2 Accuracy *[Toneva et al. 2020]*, and (RIGHT) Pairwise Accuracy.

Langkous [Anderson *et al.*, 2020] and Algonauts [Cichy *et al.*, 2021]. Japanese Ads data contains data for two sets of movies were provided by NTT DATA Corp: web and TV ads. There are also four types of cognitive labels associated with the movie datasets: scene descriptions, impression ratings, ad effectiveness indices, and ad preference votes. Algonauts 2021 contains fMRIs from 10 human subjects that watched over 1,000 short (3 sec) video clips.

Other Multimodal Datasets: Finally, beyond the video datasets, datasets have also been proposed with other kinds of multimodality. These datasets are audiovisual ([Zinszer *et al.*, 2018; Cao *et al.*, 2021]), words associated with line drawings [Mitchell *et al.*, 2008; Sudre *et al.*, 2012], pictures along with sentences and word clouds [Pereira *et al.*, 2018]. These datasets have been collected using a variety of methods like fMRIs [Mitchell *et al.*, 2008; Pereira *et al.*, 2018], MEG [Sudre *et al.*, 2012] and fNIRS [Zinszer *et al.*, 2018; Cao *et al.*, 2021]. Specifically, in [Sudre *et al.*, 2012], subjects were asked to perform a QA task, while their brain activity was recorded using MEG. Subjects were first presented with a question (e.g., "Is it manmade?"), followed by 60 concrete nouns, along with their line drawings, in a random order. For all other datasets, subjects performed passive viewing and/or listening.

4 Evaluation Metrics

Two metrics are popularly used to evaluate brain encoding models: 2V2 accuracy [Toneva *et al.*, 2020; Oota *et al.*, 2022b] and Pearson Correlation [Jain and Huth, 2018], as shown in Figure 5.

They are defined as follows. Given a subject and a brain region, let N be the number of samples. Let $\{Y_i\}_{i=1}^{N}$ and $\{\hat{Y}_i\}_{i=1}^{N}$ denote the actual and predicted voxel value vectors for the i^{th} sample. Thus, $Y \in R^{N\times V}$ and $\hat{Y} \in R^{N\times V}$ where V is the number of voxels in that region. **2V2 Accuracy** is computed as $\frac{1}{N_{C_2}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} I[\{\text{cosD}(Y_i,\hat{Y}_i) + \text{cosD}(Y_j,\hat{Y}_j)\} < \{\text{cosD}(Y_i,\hat{Y}_j) + \text{cosD}(Y_j,\hat{Y}_i)\}]$ where cosD is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if c is true, else it is 0. The higher the 2V2 accuracy, the better. **Pearson Correlation** is computed as PC=$\frac{1}{N}\sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$ where corr is the correlation function.

Brain decoding methods are evaluated using popular metrics like pairwise and rank accuracy [Pereira *et al.*, 2018; Oota *et al.*, 2022c]. Other metrics used for brain decoding evaluation include R^2 score, mean squared error, and using Representational Similarity Matrix [Cichy *et al.*, 2019; Cichy *et al.*, 2021].

Pairwise Accuracy To measure the pairwise accuracy, the first step is to predict all the test stimulus vector representations using a trained decoder model. Let S = [S_0, S_1,$\cdots$,S_n], $\hat{S}$ = [$\hat{S}_0$, $\hat{S}_1$,$\cdots$,$\hat{S}_n$] denote the "true" (stimuli-derived) and predicted stimulus representations for n test instances resp. Given a pair (i, j) such that $0 \leq i, j \leq n$, score is 1 if $corr(S_i,\hat{S}_i) + corr(S_j,\hat{S}_j) > corr(S_i,\hat{S}_j) + corr(S_j,\hat{S}_i)$, else 0. Here, *corr* denotes the Pearson correlation. Final pairwise matching accuracy per participant is the average of scores across all pairs of test instances. For computing rank accuracy, we first compare each decoded vector to all the "true" stimuli-derived semantic vectors and ranked them by their correlation. The classification performance reflects the rank r of the stimuli-derived vector for the correct word/picture/stimuli: $1 - \frac{r-1}{\#instances-1}$. The final accuracy value for each participant is the average rank accuracy across all instances.

5 Brain Encoding

Encoding is the learning of the mapping from the stimulus domain to the neural activation. The quest in brain encoding is for "reverse engineering" the algorithms that the brain uses for sensation, perception, and higher-level cognition. Recent breakthroughs in applied NLP enable reverse engineering the language function of the brain. Similarly, pioneering results have been obtained for reverse engineering the function of ventral visual stream in object recognition founded on the advances and remarkable success of deep CNNs. The overall schema of building a brain encoder is shown in Figure 6.

Initial studies on brain encoding focused on smaller data sets and single modality of brain responses. Early models used word representations [Hollenstein *et al.*, 2019]. Rich contextual representations derived from RNNs such as LSTMs resulted in superior encoding models [Jain and Huth, 2018; Oota *et al.*, 2019] of narratives. The recent

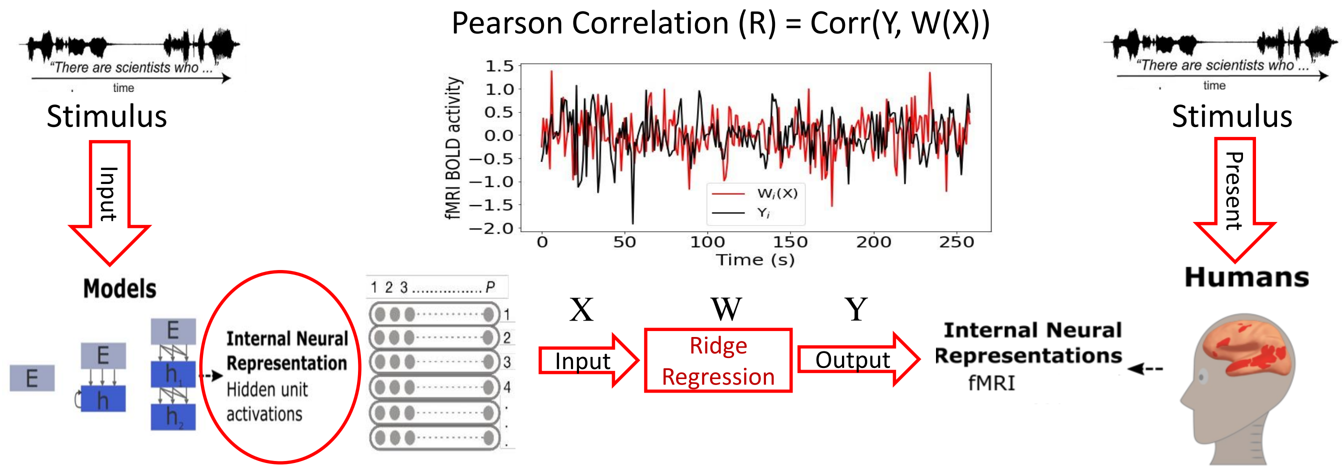

Figure 6: Schema for Brain Encoding

efforts are aimed at utilizing the internal representations extracted from transformer-based language models such as ELMo, BERT, GPT-2, etc for learning encoding models of brain activation [Jat *et al.*, 2020; Caucheteux *et al.*, 2021; Antonello *et al.*, 2021]. High-grain details such as lexical, compositional, syntactic, and semantic representations of narratives are factorized from transformer-based models and utilized for training encoding models. The resulting models are better able to disentangle the corresponding brain responses in fMRI [Caucheteux *et al.*, 2021]. Finally, is has been found that the models that integrate task and stimulus representations have significantly higher prediction performance than models that do not account for the task semantics [Toneva *et al.*, 2020; Schrimpf *et al.*, 2021a].

Similarly, in vision, early models focused on independent models of visual processing (object classification) using CNNs [Yamins *et al.*, 2014]. Recent efforts in visual encoding models focus on using richer visual representations derived from a variety of computer vision tasks [Wang *et al.*, 2019]. Instead of feed-forward deep CNN models, using shallow recurrence enabled better capture of temporal dynamics in the visual encoding models [Kubilius *et al.*, 2019; Schrimpf *et al.*, 2020].

Table 2 summarizes various encoding models proposed in the literature related to textual, audio, visual, and multimodal stimuli. Figure 7 classifies the encoding literature along various stimulus domains such as vision, auditory, multimodal, and language and the corresponding tasks in each domain.

Linguistic Encoding: A number of previous works have investigated the alignment between pretrained language models and brain recordings of people comprehending language. Huth et al. [2016] have been able to identify brain ROIs (Regions of Interest) that respond to words that have a similar meaning and have thus built a "semantic atlas" of how the human brain organizes language. Many studies have shown accurate results in mapping the brain activity using neural distributed word embeddings for linguistic stimuli [Anderson *et al.*, 2017a; Pereira *et al.*, 2018; Oota *et al.*, 2018; Nishida and Nishimoto, 2018; Sun *et al.*, 2019]. Unlike earlier models where each word is represented as an independent vector in an embedding space, [Jain and Huth, 2018] built encoding models using rich contextual representations derived from an LSTM language model in a story listening task. With these contextual representations, they demonstrated dissociation in brain activation – auditory cortex (AC) and Broca's area in shorter context whereas left Temporo-Parietal junction (TPJ) in longer context. [Hollenstein *et al.*, 2019] presents the first multimodal framework for evaluating six types of word embedding (Word2Vec, WordNet2Vec, GloVe, FastText, ELMo, and BERT) on 15 datasets, including eye-tracking, EEG and fMRI signals recorded during language processing. With the recent advances in contextual representations in NLP, few studies incorporated them in relating sentence embeddings with brain activity patterns [Sun *et al.*, 2020; Gauthier and Levy, 2019; Jat *et al.*, 2020].

More recently, researchers have begun to study the alignment of language regions of the brain with the layers of language models and found that the best alignment was achieved in the middle layers of these models [Jain and Huth, 2018; Toneva and Wehbe, 2019]. Schrimpf et al. [2021a] examined the relationship between 43 diverse state-of-the-art language models. They also studied the behavioral signatures of human language processing in the form of self-paced reading times, and a range of linguistic functions assessed via standard engineering tasks from NLP. They found that Transformer-based models perform better than RNNs or word-level embedding models. Larger-capacity models perform better than smaller models. Models initialized with random weights (prior to training) perform surprisingly similarly in neural predictivity as compared to final trained models, suggesting that network architecture contributes as much or more than experience dependent learning to a model's match to the brain. Antonello et al. [2021] proposed a "language representation embedding space" and demonstrated the effectiveness of the features from this embedding in predicting fMRI responses to linguistic stimuli.

Disentangling the Syntax and Semantics: The representations of transformer models like BERT, GPT-2 have been

Table 2: Summary of Representative Brain Encoding Studies

Stimuli	Authors	Dataset Type	Lang.	Stimulus Representations	|S|	Dataset	Model
Text	[Jain and Huth, 2018]	fMRI	English	LSTM	6	Subset Moth Radio Hour	Ridge
	[Toneva and Wehbe, 2019]	fMRI/ MEG	English	ELMo, BERT, Transformer-XL	9	Story understanding	Ridge
	[Toneva *et al.*, 2020]	MEG	English	BERT	9	Question-Answering	Ridge
	[Schrimpf *et al.*, 2021b]	fMRI/ECoG	English	43 language models (e.g. GloVe, ELMo, BERT, GPT-2, XLNET)	20	Neural architecture of language	Ridge
	[Gauthier and Levy, 2019]	fMRI	English	BERT, fine-tuned NLP tasks (Sentiment, Natural language inference), Scrambling language model	7	Imagine a situation with the noun	Ridge
	[Deniz *et al.*, 2019]	fMRI	English	GloVe	9	Subset Moth Radio Hour	Ridge
	[Jain *et al.*, 2020]	fMRI	English	LSTM	6	Subset Moth Radio Hour	Ridge
	[Caucheteux *et al.*, 2021]	fMRI	English	GPT-2, Basic syntax features	345	Narratives	Ridge
	[Antonello *et al.*, 2021]	fMRI	English	GloVe, BERT, GPT-2, Machine Translation, POS tasks	6	Moth Radio Hour	Ridge
	[Reddy and Wehbe, 2021]	fMRI	English	Constituency, Basic syntax features and BERT	8	Harry Potter	Ridge
	[Goldstein *et al.*, 2022]	fMRI	English	GloVe, GPT-2 next word, pre-onset, post-onset word surprise	8	ECoG	
	[Oota *et al.*, 2022b]	fMRI	English	BERT and GLUE tasks	82	Pereira & Narratives	Ridge
	[Oota *et al.*, 2022a]	fMRI	English	ESN, LSTM, ELMo, Longformer	82	Narratives	Ridge
	[Merlin and Toneva, 2022]	fMRI	English	BERT, Next word prediction, multi-word semantics, scrambling model	8	Harry Potter	Ridge
	[Toneva *et al.*, 2022]	fMRI / MEG	English	ELMo, BERT, Context Residuals	8	Harry Potter	Ridge
	[Aw and Toneva, 2022]	fMRI	English	BART, Longformer, Long-T5, BigBird, and corresponding Booksum models as well	8	Passive reading	Ridge
	[Zhang *et al.*, 2022b]	fMRI	English, Chinese	Node Count	19, 12	Zhang	Ridge
	[Oota *et al.*, 2023a]	fMRI	English	Constituency, Dependency trees, Basic syntax features and BERT	82	Narratives	Ridge
	[Oota *et al.*, 2023b]	MEG	English	Basic syntax features, GloVe and BERT	8	MEG-MASC	Ridge
	[Tuckute *et al.*, 2023]	fMRI	English	BERT-Large, GPT-2 XL	12	Reading Sentences	Ridge
	[Kauf *et al.*, 2023]	fMRI	English	BERT-Large, GPT-2 XL	12	Pereira	Ridge
	[Singh *et al.*, 2023]	fMRI	English	BERT-Large, GPT-2 XL, Text Perturbations	5	Pereira	Ridge
Visual	[Wang *et al.*, 2019]	fMRI		21 downstream vision tasks	4	BOLD 5000	Ridge
	[Kubilius *et al.*, 2019]	fMRI		CNN models AlexNet, ResNet, DenseNet	7	Algonauts	Ridge
	[Dwivedi *et al.*, 2021]	fMRI		21 downstream vision tasks	4	BOLD 5000	Ridge
	[Khosla and Wehbe, 2022]	fMRI		CNN models AlexNet	4	BOLD 5000	Ridge
	[Conwell *et al.*, 2023]	fMRI		CNN models AlexNet	4	BOLD 5000	Ridge
Audio	[Millet *et al.*, 2022]	fMRI	English	Wav2Vec2.0	345	Narratives	Ridge
	[Vaidya *et al.*, 2022]	fMRI	English	APC, AST, Wav2Vec2.0, and HuBERT	7	Moth Radio Hour	Ridge
	[Tuckute *et al.*, 2022]	fMRI	English	19 Speech Models (e.g. DeepSpeech, Wav2Vec2.0, VQ-VAE)	19	Passive listening	Ridge
	[Oota *et al.*, 2023c]	fMRI	English	5 basic and 25 deep learning based speech models (Tera, CPC, APC, Wav2Vec2.0, HuBERT, DistilHuBERT, Data2Vec	6	Moth Radio Hour	Ridge
	[Oota *et al.*, 2023d]	fMRI	English	Wav2Vec2.0 and SUPERB tasks	82	Narratives	Ridge
Multi Modal	[Dong and Toneva, 2023]	fMRI	English	Merlo Reseve	5	Neuromod	Ridge
	[Popham *et al.*, 2021]	fMRI	English	985D Semantic Vector	5	Moth Radio Hour & Short Movie Clips	Ridge
	[Oota *et al.*, 2022d]	fMRI	English	CLIP, VisualBERT, LXMERT, CNNs and BERT	5, 82	Periera & Narratives	Ridge
	[Lu *et al.*, 2022]	fMRI	English	BriVL	5	Pereira & Short Movie Clips	Ridge
	[Tang *et al.*, 2023]	fMRI	English	BridgeTower	5	Moth Radio Hour & Short Movie Clips	Ridge

shown to linearly map onto brain activity during language comprehension. Several studies have attempted to disentangle the contributions of different types of information from word representations to the alignment between brain recordings and language models. Wang et al. [2020] proposed a two-channel variational autoencoder model to dissociate sentences into semantic and syntactic representations and separately associate them with brain imaging data to find feature-correlated brain regions. To separate each syntactic feature, Zhang et al. [2022a] proposed a feature elimination method, called Mean Vector Null space Projection. Compared with word representations, word syntactic features (parts-of-speech, named entities, semantic roles, dependencies) seem to be distributed across brain networks instead of a local brain region. In the previous two studies, we do not know whether all or any of these representations effectively drive the linear mapping between language models (LMs) and the brain. Toneva et al. [2022] presented an approach to disentangle supra-word meaning from lexical meaning in language models and showed that supra-word meaning is predictive of fMRI recordings in two language regions (anterior and posterior temporal lobes). Caucheteux et al. [2021] proposed a taxonomy to factorize the high-dimensional activations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations. They found that (1) Compositional representations recruit a more widespread cortical network than lexical ones, and encompass the bilateral temporal, parietal and prefrontal cortices. (2) Contrary to previous claims, syntax and semantics are not associated with separated modules, but, instead, appear to share a common and distributed neural substrate.

While previous works studied syntactic processing as cap-

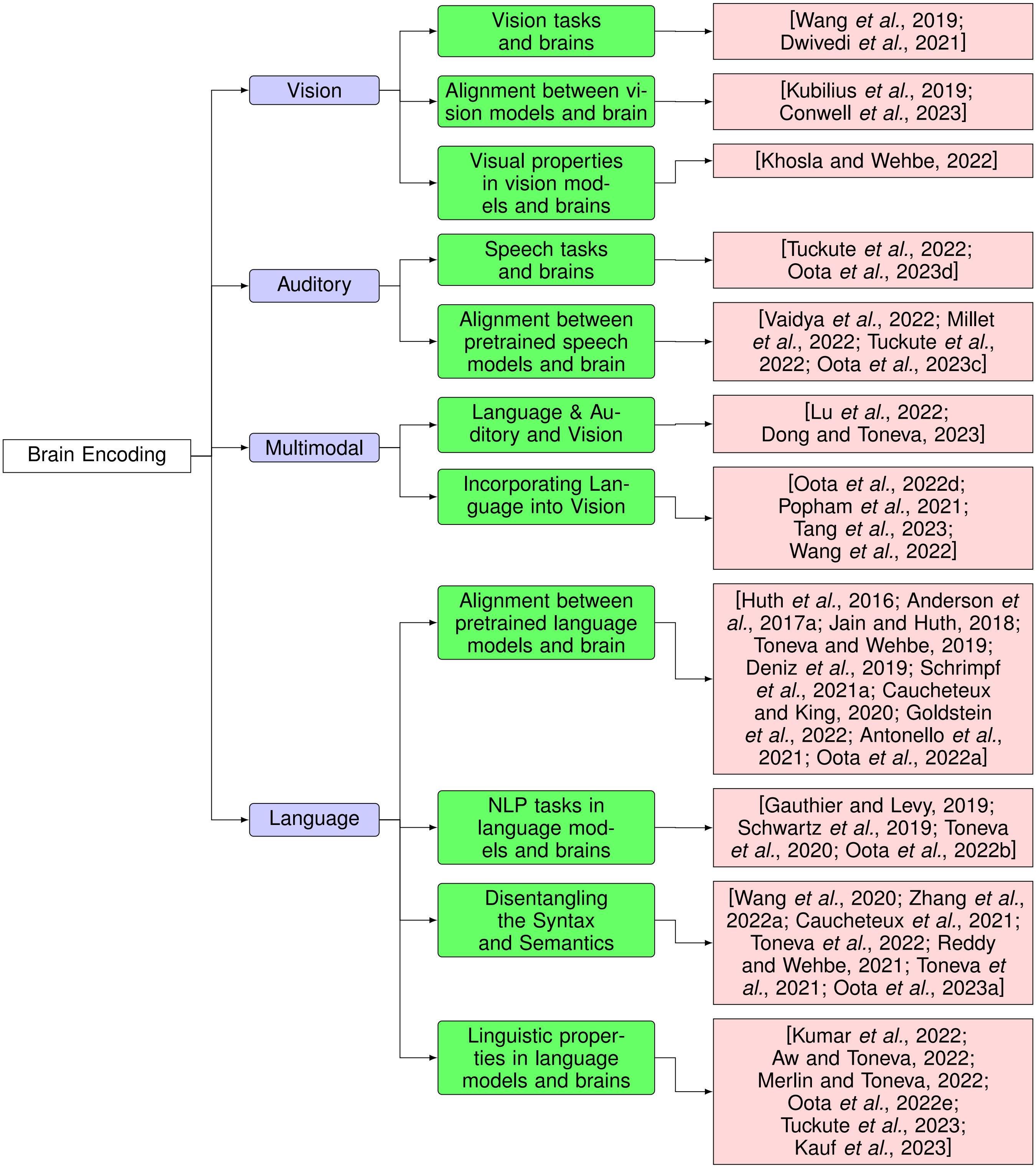

Figure 7: Brain Encoding Survey Tree

tured through complexity measures (syntactic surprisal, node count, word length, and word frequency), very few have studied the syntactic representations themselves. Studying syntactic representations using fMRI is difficult because: (1) representing syntactic structure in an embedding space is a non-trivial computational problem, and (2) the fMRI signal is noisy. To overcome these limitations, Reddy et al. [2021] proposed syntactic structure embeddings that encode the syntactic information inherent in natural text that subjects read in the scanner. The results reveal that syntactic structure-based features explain additional variance in the brain activity of various parts of the language system, even after controlling for complexity metrics that capture the processing load. Toneva et al. [2021] further examined whether the representations obtained from a language model align with different language processing regions in a similar or different way.

Linguistic properties in LMs and brains: Understanding the reasons behind the observed similarities between language comprehension in LMs and brains can lead to more insights into both systems. Several works [Schwartz *et al.*, 2019; Kumar *et al.*, 2022; Aw and Toneva, 2022; Merlin and Toneva, 2022; Oota *et al.*, 2022b] have found that using a fine-tuned BERT leads to improved brain predictions. However, it is not clear what type of information in the fine-tuned BERT model led to the improvement. It is unclear whether and how the two systems align in their information processing pipeline. Aw and Toneva [2022] used four pre-trained large language models (BART, Longformer Encoder Decoder, Big-Bird, and LongT5) and also trained them to improve their narrative understanding, using the method detailed in Figure 8. However, it is not understood whether prediction of the next word is necessary for the observed brain alignment or simply sufficient, and whether there are other shared mechanisms or information that is similarly important. Merlin and Toneva [2022] proposed two perturbations to pretrained language models that, when used together, can control for the effects of next word prediction and word-level semantics on the alignment with brain recordings. Specifically, they find that improvements in alignment with brain recordings in two language processing regions–Inferior Frontal Gyrus (IFG) and Angular Gyrus (AG)–are due to next word prediction and word-level semantics. However, what linguistic information actually underlies the observed alignment between brains and language models is not clear. Recently, Oota et al. [2022e] tested the effect of a range of linguistic properties (surface, syntactic and semantic) and found that the elimination of each linguistic property results in a significant decrease in brain alignment across all layers of BERT.

Visual Encoding: CNNs are currently the best class of models of the neural mechanisms of visual processing [Du *et al.*, 2020; Beliy *et al.*, 2019; Oota *et al.*, 2019; Nishida *et al.*, 2020]. How can we push these deeper CNN models to capture brain processing even more stringently? Continued architectural optimization on ImageNet alone no longer seems like a viable option. Kubilius et al. [2019] proposed a shallow recurrent anatomical network CORnet that follows neuroanatomy more closely than standard CNNs, and achieved the state-of-the-art results on the Brain-score benchmark. It has four computational areas, conceptualized as analogous to the ventral visual areas V1, V2, V4, and IT, and a linear category decoder that maps from the population of neurons in the model's last visual area to its behavioral choices.

Despite the effectiveness of CNNs, it is difficult to draw specific inferences about neural information processing using CNN- derived representations from a generic object-classification CNN. Hence, Wang et al. [2019] built encoding models with individual feature spaces obtained from 21 computer vision tasks. One of the main findings is that features from 3D tasks, compared to those from 2D tasks, predict a distinct part of visual cortex.

Auditory Encoding: Speech stimuli have mostly been represented using encodings of text transcriptions [Huth *et al.*, 2016] or using basic features like phoneme rate, the sum of squared FFT coefficients [Pandey *et al.*, 2022], etc. Text transcription-based methods ignore the raw audio-sensory information completely. The basic speech feature engineering method misses the benefits of transfer learning from rigorously pretrained speech DL models.

Recently, several researchers have used popular deep learning models such as APC [Chung *et al.*, 2020], Wav2Vec2.0 [Baevski *et al.*, 2020], HuBERT [Hsu *et al.*, 2021], and Data2Vec [Baevski *et al.*, 2022] for encoding speech stimuli. Millet et al. [2022] used a self-supervised learning model Wav2Vec2.0 to learn latent representations of the speech waveform similar to those of the human brain. They find that the functional hierarchy of its transformer layers aligns with the cortical hierarchy of speech in the brain, and reveals the whole-brain organisation of speech processing with an unprecedented clarity. This means that the first transformer layers map onto the low-level auditory cortices (A1 and A2), the deeper layers (orange and red) map onto brain regions associated with higher-level processes (e.g. STS and IFG). Vaidya et al. [2022] present the first systematic study to bridge the gap between recent four self-supervised speech representation methods (APC, Wav2Vec, Wav2Vec2.0, and HuBERT) and computational models of the human auditory system. Similar to [Millet *et al.*, 2022], they find that self-supervised speech models are the best models of auditory areas. Lower layers best modeled low-level areas, and upper-middle layers were most predictive of phonetic and semantic areas, while layer representations follow the accepted hierarchy of speech processing. Tuckute et al. [2022] analyzed 19 different speech models and find that some audio models derived in engineering contexts (model applications ranged from speech recognition and speech enhancement to audio captioning and audio source separation) produce poor predictions of auditory cortical responses, many task-optimized audio speech deep learning models outpredict a standard spectrotemporal model of the auditory cortex and exhibit hierarchical layer-region correspondence with auditory cortex.

Multimodal Brain Encoding: Multimodal stimuli can be best encoded using recently proposed deep learning based multimodal models. Oota et al. [2022d] experimented with multimodal models like Contrastive Language-Image Pre-training (CLIP), Learning Cross-Modality Encoder Representations from Transformers (LXMERT), and VisualBERT and found VisualBERT to the best. Similarly, Wang et al. [2022] find that multimdoal models like CLIP better predict neural responses in visual cortex, since image captions typically contain the most semantically relevant information in an image for humans. [Dong and Toneva, 2023] present a systematic approach to probe multi-modal video Transformer model by leveraging neuroscientific evidence of multimodal information processing in the brain. The authors find that in-

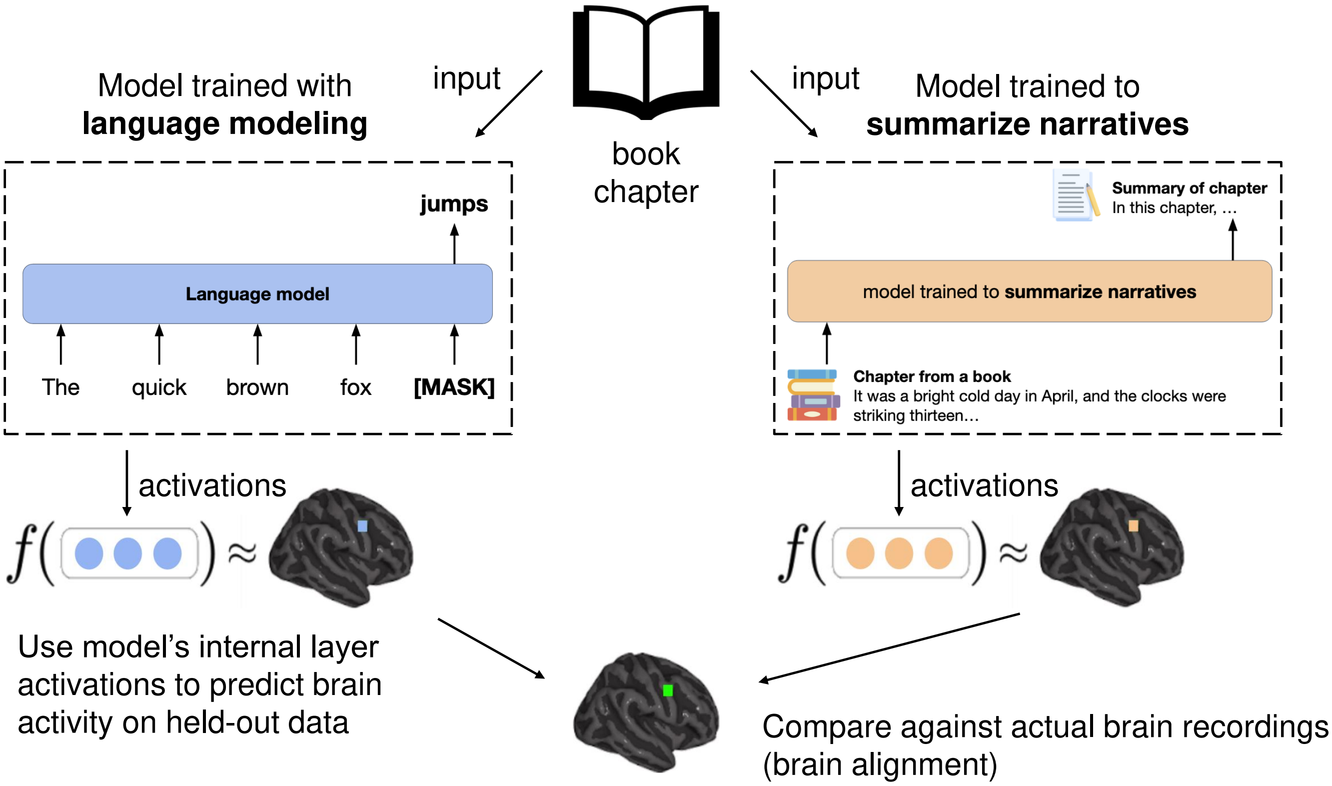

Figure 8: Comparison of brain recordings with language models trained on web corpora (LEFT) and language models trained on book stories (RIGHT) *[Aw and Toneva, 2022]*.

termediate layers of a multimodal video transformer are better at predicting multimodal brain activity than other layers, indicating that the intermediate layers encode the most brain-related properties of the video stimuli. Recently, [Tang *et al.*, 2023] investigated a multimodal Transformer as the encoder architecture to extract the aligned concept representations for narrative stories and movies to model fMRI responses to naturalistic stories and movies, respectively. Since language and vision rely on similar concept representations, the authors perform a cross-modal experiment in which how well the language encoding models can predict movie-fMRI responses from narrative story features (story $\rightarrow$ movie) and how well the vision encoding models can predict narrative story-fMRI responses from movie features (movie $\rightarrow$ story). Overall, the authors find that cross-modality performance was higher for features extracted from multimodal transformers than for linearly aligned features extracted from unimodal transformers.

6 Brain Decoding

Decoding is the learning of the mapping from neural activations back to the stimulus domain. Figure 9 depicts the typical workflow for building an image/language decoder.

Decoder Architectures: In most cases, the stimulus representation is decoded using typical ridge regression models trained on each voxel and its 26 neighbors in 3D to predict each dimension of the stimulus representation. Also, decoding is usually performed using the most informative voxels [Pereira *et al.*, 2018]. In some cases, a fully connected layer [Beliy *et al.*, 2019] or a multi-layered perceptron [Sun *et al.*, 2019] has been used. In some studies, when decoding is modeled as multi-class classification, Gaussian Naïve Bayes [Singh *et al.*, 2007; Just *et al.*, 2010] and SVMs [Thirion *et al.*, 2006] have also been used for decoding. Figure 10 summarizes the literature related to various decoding solutions proposed in vision, auditory, and language domains.

Decoding task settings: The most common setting is to perform decoding to a vector representation using a stimuli of a single mode (visual, text or audio). Initial brain decoding experiments studied the recovery of simple concrete nouns and verbs from fMRI brain activity [Nishimoto *et al.*, 2011] where the subject watches either a picture or a word. Sun et al. [2019] used several sentence representation models to associate brain activities with sentence stimulus, and found InferSent to perform the best. More work has focused on decoding the text passages instead of individual words [Wehbe *et al.*, 2014].

Some studies have focused on multimodal stimuli based decoding where the goal is still to decode the text representation vector. For example, Pereira et al. [2018] trained the decoder on imaging data of individual concepts, and showed that it can decode semantic vector representations from imaging data of sentences about a wide variety of both concrete and abstract topics from two separate datasets. Further, Oota

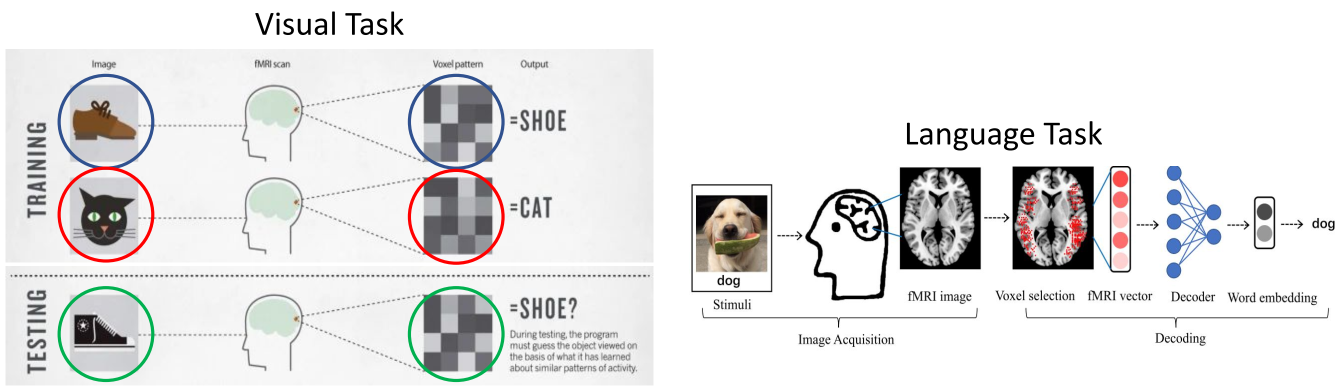

Figure 9: Schema for Brain Decoding. LEFT: Image decoder *[Smith et al. 2011]*, RIGHT: Language Decoder *[Wang et al. 2019]*

Table 3: Summary of Representative Brain Decoding Studies

Stimuli	**Authors**	**Dataset Type**	**Lang.**	**Stimulus Representations**	**|S|**	**Dataset**	**Model**
Text	[Pereira *et al.*, 2018]	fMRI	English	Word2Vec, GloVe, BERT	17	Pereira	Ridge
	[Wang *et al.*, 2020]	fMRI	English	BERT, RoBERTa	6	Pereira	Ridge
	[Oota *et al.*, 2022c]	fMRI	English	GloVe, BERT, RoBERTa	17	Pereira	Ridge
	[Tang *et al.*, 2022]	fMRI	English	GPT, fine-tuned GPT on Reddit comments and autobiographical stories	7	Moth Radio Hour	Ridge
Visual	[Beliy *et al.*, 2019]	fMRI		End-to-End Encoder-Decoder, Decoder-Encoder, AlexNet	5	Generic Object Decoding, ViM-1	
	[Takagi and Nishimoto, 2022]	fMRI		Latent Diffusion Model, CLIP	4	NSD	Ridge
	[Ozcelik and VanRullen, 2023]	fMRI		VDVAE, Latent Diffusion Model	7	NSD	
	[Chen *et al.*, 2023b]	fMRI		Latent Diffusion Model, CLIP	3	HCP fMRI-Video-Dataset	Ridge
Audio	[Défossez *et al.*, 2022]	MEG,EEG	English	MEL Spectrogram, Wav2Vec2.0	169	MEG-MASC	Ridge, CLIP
	[Gwilliams *et al.*, 2022]	MEG	English	Phonemes	7	MEG-MASC	

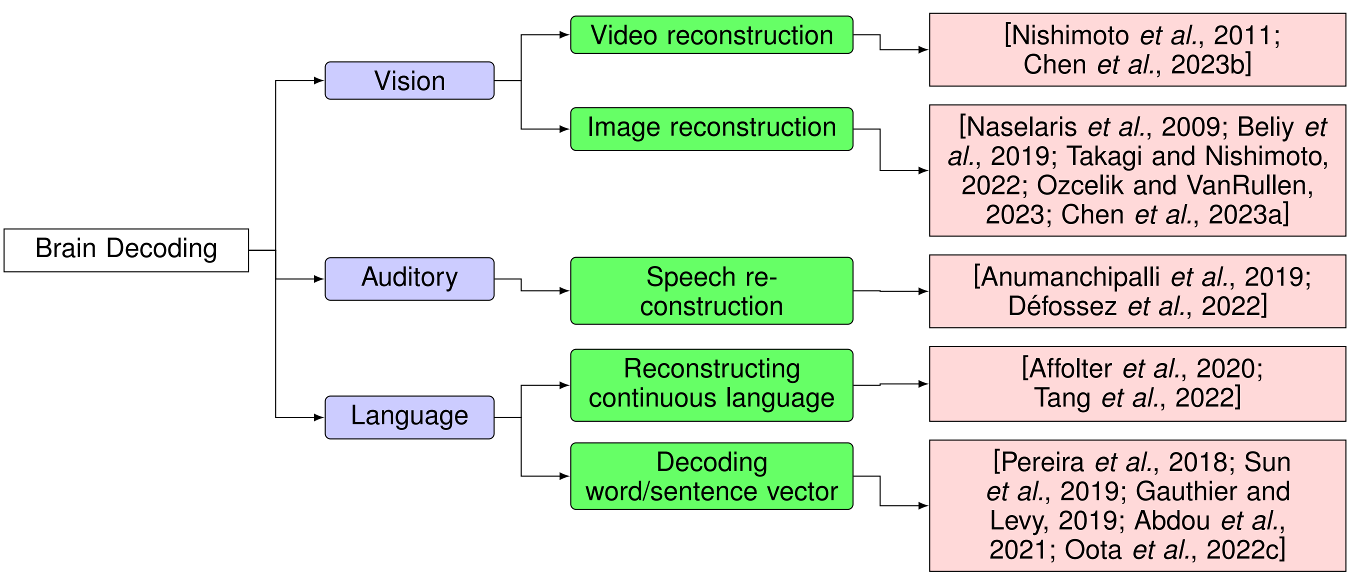

Figure 10: Brain Decoding Survey Tree

et al. [2022c] propose two novel brain decoding setups: (1) multi-view decoding (MVD) and (2) cross-view decoding (CVD). In MVD, the goal is to build an MV decoder that can take brain recordings for any view as input and predict the concept. In CVD, the goal is to train a model which takes brain recordings for one view as input and decodes a semantic vector representation of another view. Specifically, they study practically useful CVD tasks like image captioning, image tagging, keyword extraction, and sentence formation.

To understand application of Transformer models for decoding better, Gauthier et al. [2019] fine-tuned a pre-trained BERT on a variety of NLU tasks, asking which lead to improvements in brain-decoding performance. They find that tasks which produce syntax-light representations yield significant improvements in brain decoding performance. Toneva et al. [2019] study how representations of various Transformer models differ across layer depth, context length, and attention type.

Some studies have attempted to reconstruct words [Affolter *et al.*, 2020], continuous language [Tang *et al.*, 2022], images [Du *et al.*, 2020; Beliy *et al.*, 2019; Fang *et al.*, 2020; Lin *et al.*, 2022], speech [Défossez *et al.*, 2022] or question-answer speech dialogues [Moses *et al.*, 2019] rather than just predicting a semantic vector representation. Lastly, some studies have focused on reconstructing personal imagined experiences [Berezutskaya *et al.*, 2020] or application-based decoding like using brain activity scanned during a picture-based mechanical engineering task to predict individuals' physics/engineering exam results [Cetron *et al.*, 2019] and reflecting whether current thoughts are detailed, correspond to the past or future, are verbal or in images [Smallwood and Schooler, 2015]. Table 3 aggregates the brain decoding literature along different stimulus domains such as textual, visual, and audio.

7 Conclusion, Limitations, and Future Trends

Conclusion In this paper, we surveyed important datasets, stimulus representations, brain encoding and brain decoding methods across different modalities. A glimpse of how deep learning solutions throw light on putative brain computations is given.

Limitations Naturalistic datasets of passive reading/listening offer ecologically realistic settings for investigating brain function. However, the lack of a task (as in a controlled psycholinguistic experiment) that probes the participant's understanding of the narrative limits the inferences that can be made on what the participant's brain is actually engaged in while passively following the stimuli. This becomes even more important when multi-lingual, multiscriptal participants process stimuli in L2 language or script – it is unclear if the brain activity reflects the processing of L2 or active suppression L1 while focusing on L2 [Malik-Moraleda *et al.*, 2022].

Future Trends Some of the future areas of work in this field are as follows: (1) While there is work on the text, understanding the similarity in information processing between visual/speech/multimodal models versus natural brain systems remains an open area. (2) Decoding to actual multimodal stimuli seems feasible thanks to recent advances in generation using deep learning models. (3) Deeper understanding of the degree to which damage to different parts of the human brain could lead to the degradation of cognitive skills. (4) How can we train artificial neural networks in novel self-supervised ways such that they compose word meanings or comprehend images and speech like a human brain? (5) How can we leverage improved neuroscience understanding to suggest changes in proposed artificial neural network architectures to make them more robust and accurate? We hope that this survey motivates research along the above directions.

References

[Abdou *et al.*, 2021] Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *arXiv preprint arXiv:2101.12608*, 2021.

[Affolter *et al.*, 2020] Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

[Allen *et al.*, 2022] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[Anderson *et al.*, 2017a] 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. *TACL*, 5:17–30, 2017.

[Anderson *et al.*, 2017b] Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395, 2017.

[Anderson *et al.*, 2019] Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Rajeev DS Raizada, Feng Lin, and Edmund C Lalor. An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, 39(45):8969–8987, 2019.

[Anderson *et al.*, 2020] Andrew James Anderson, Kelsey McDermott, Brian Rooks, Kathi L Heffner, David Dodell-Feder, and Feng V Lin. Decoding individual identity from brain activity elicited in imagining common experiences. *Nature communications*, 11(1):1–14, 2020.

[Antonello *et al.*, 2021] Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *NeurIPS*, 34:8332–8344, 2021.

[Anumanchipalli *et al.*, 2019] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[Aw and Toneva, 2022] Khai Loong Aw and Mariya Toneva. Training language models for deeper understanding improves brain alignment. *arXiv preprint arXiv:2212.10898*, 2022.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460, 2020.

[Baevski *et al.*, 2022] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pages 1298–1312. PMLR, 2022.

[Beliy *et al.*, 2019] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *arXiv preprint arXiv:1907.02431*, 2019.

[Berezutskaya *et al.*, 2020] Julia Berezutskaya, Zachary V Freudenburg, Luca Ambrogioni, Umut Güçlü, Marcel AJ van Gerven, and Nick F Ramsey. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *Scientific reports*, 10(1):1–21, 2020.

[Boyle *et al.*, 2020] Julie A Boyle, Basile Pinsard, A Boukhdhir, S Belleville, S Brambatti, J Chen, J Cohen-Adad, A Cyr, A Fuente, P Rainville, et al. The courtois project on neuronal modelling: 2020 data release. In *OHBM*, 2020.

[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.

[Cao *et al.*, 2021] Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. Brain decoding using fnirs. In *AAAI*, volume 35, pages 12602–12611, 2021.

[Caucheteux and King, 2020] Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.

[Caucheteux *et al.*, 2021] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling syntax and semantics in the brain with deep networks. In *ICML*, pages 1336–1348. PMLR, 2021.

[Cetron *et al.*, 2019] Joshua S Cetron, Andrew C Connolly, Solomon G Diamond, Vicki V May, and James V Haxby. Decoding individual differences in stem learning from functional mri data. *Nature communications*, 10(1):1–10, 2019.

[Chang *et al.*, 2019] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18, 2019.

[Chen *et al.*, 2023a] Xuhang Chen, Baiying Lei, Chi-Man Pun, and Shuqiang Wang. Brain diffuser: An end-to-end brain image to brain network pipeline. *arXiv preprint arXiv:2303.06410*, 2023.

[Chen *et al.*, 2023b] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *arXiv preprint arXiv:2305.11675*, 2023.

[Chung *et al.*, 2020] Yu-An Chung, Hao Tang, and James Glass. Vector-quantized autoregressive predictive coding. *Interspeech*, pages 3760–3764, 2020.

[Cichy *et al.*, 2019] Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv e-prints*, pages arXiv–1905, 2019.

[Cichy *et al.*, 2021] Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Polina Iamshchinina, M Graumann, A Andonian, NAR Murty, K Kay, Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv preprint arXiv:2104.13714*, 2021.

[Conwell *et al.*, 2023] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023.

[Défossez *et al.*, 2022] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings. *arXiv preprint arXiv:2208.12266*, 2022.

[Deniz *et al.*, 2019] Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

[Doerig *et al.*, 2022] Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace Lindsay, Konrad Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*, 2022.

[Dong and Toneva, 2023] Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

[Du *et al.*, 2020] Changde Du, Changying Du, Lijie Huang, and Huiguang He. Conditional generative neural decoding with structured cnn feature prediction. In *AAAI*, pages 2629–2636, 2020.

[Dwivedi *et al.*, 2021] Kshitij Dwivedi, Michael F Bonner, Radoslaw Martin Cichy, and Gemma Roig. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS computational biology*, 17(8):e1009267, 2021.

[Fang *et al.*, 2020] Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *NeurIPS*, 33:13038–13048, 2020.

[Gauthier and Levy, 2019] Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. *arXiv preprint arXiv:1910.01244*, 2019.

[Goldstein *et al.*, 2022] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

[Gwilliams *et al.*, 2022] Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkanen, David Poeppel, and Jean-Remi King. Meg-masc: a high-quality magneto-encephalography dataset for evaluating natural speech processing. *arXiv preprint arXiv:2208.11488*, 2022.

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

[Handjaras *et al.*, 2016] Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, and Giovanna Marotta. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, 135:232–242, 2016.

[Hebart *et al.*, 2022] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data: A multimodal collection of large-scale datasets for investigating object representations in brain and behavior. *bioRxiv*, pages 2022–07, 2022.

[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.*, 2019] Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. Cognival: A framework for cognitive word embedding evaluation. In *CoNLL*, pages 538–549, 2019.

[Horikawa and Kamitani, 2017] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 29:3451–3460, 2021.

[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.

[Huth *et al.*, 2022] Alexander G Huth, Shinji Nishimoto, An T Vu, Dupre la Tour T, and Gallant JL. Gallant lab natural short clips 3t fmri data. *G-Node*, 2022.

[Jain and Huth, 2018] Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fmri. In *NIPS*, pages 6629–6638, 2018.

[Jain *et al.*, 2020] Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *NeurIPS*, 33:13738–13749, 2020.

[Jat *et al.*, 2020] S Jat, H Tang, P Talukdar, and T Mitchel. Relating simple sentence representations in deep neural networks and the brain. In *ACL*, pages 5137–5154, 2020.

[Just *et al.*, 2010] Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, 5(1):e8622, 2010.

[Karamolegkou *et al.*, 2023] Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping brains with language models: A survey. *arXiv preprint arXiv:2306.05126*, 2023.

[Kauf *et al.*, 2023] Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *bioRxiv*, pages 2023–05, 2023.

[Kay *et al.*, 2008] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[Khosla and Wehbe, 2022] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, 2022.

[Kubilius *et al.*, 2019] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *NIPS*, 32:12805–12816, 2019.

[Kumar *et al.*, 2022] Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, pages 2022–06, 2022.

[Lahner *et al.*, 2023] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Bold moments: modeling short visual events through a video fmri dataset and metadata. *bioRxiv*, pages 2023–03, 2023.

[Li *et al.*, 2021] Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince: A multilingual fmri corpus using ecological stimuli. *Biorxiv*, pages 2021–10, 2021.

[Lin *et al.*, 2022] Sikun Lin, Thomas Christopher Sprague, and Ambuj Singh. Mind reader: Reconstructing complex images from brain activities. In *NeurIPS*, 2022.

[Lu *et al.*, 2022] Haoyu Lu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, Jingyuan Wen, Changde Du, Xin Zhao, Hao Sun, Huiguang He, et al. Multimodal foundation models are better simulators of the human brain. *arXiv preprint arXiv:2208.08263*, 2022.

[Malik-Moraleda *et al.*, 2022] Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019, 2022.

[Merlin and Toneva, 2022] Gabriele Merlin and Mariya Toneva. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv preprint arXiv:2212.00596*, 2022.

[Millet *et al.*, 2022] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv:2206.01685*, 2022.

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

[Moses *et al.*, 2019] David A Moses, Matthew K Leonard, Joseph G Makin, and Edward F Chang. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications*, 10(1):1–14, 2019.

[Naselaris *et al.*, 2009] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[Nastase *et al.*, 2021] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. Narratives: fmri data for evaluating models of naturalistic language comprehension. *bioRxiv*, pages 2020–12, 2021.

[Nishida and Nishimoto, 2018] Satoshi Nishida and Shinji Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *Neuroimage*, 180:232–242, 2018.

[Nishida *et al.*, 2020] Satoshi Nishida, Yusuke Nakano, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. Brain-mediated transfer learning of convolutional neural networks. In *AAAI*, pages 5281–5288, 2020.

[Nishimoto *et al.*, 2011] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

[Oota *et al.*, 2018] Subba Reddy Oota, Naresh Manwani, and Raju S Bapi. fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings. In *ICONIP*, pages 3–15. Springer, 2018.

[Oota *et al.*, 2019] Subba Reddy Oota, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Stepencog: A convolutional lstm autoencoder for near-perfect fmri encoding. In *IJCNN*, pages 1–8. IEEE, 2019.

[Oota *et al.*, 2022a] Subba Reddy Oota, Frederic Alexandre, and Xavier Hinaut. Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

[Oota *et al.*, 2022b] Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? *arXiv preprint arXiv:2205.01404*, 2022.

[Oota *et al.*, 2022c] Subba Reddy Oota, Jashn Arora, Manish Gupta, and Raju S Bapi. Multi-view and cross-view brain decoding. In *COLING*, pages 105–115, 2022.

[Oota *et al.*, 2022d] Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. In *COLING*, pages 116–133, 2022.

[Oota *et al.*, 2022e] Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *arXiv preprint arXiv:2212.08094*, 2022.

[Oota *et al.*, 2023a] Subba Reddy Oota, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampud. Syntactic structure processing in the brain while listening. *arXiv preprint arXiv:2302.08589*, 2023.

[Oota *et al.*, 2023b] Subba Reddy Oota, Trouvain Nathan, Frederic Alexandre, and Xavier Hinaut. Meg encoding using word context semantics in listening stories. In *Interspeech*, 2023.

[Oota *et al.*, 2023c] Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Raju Surampudi Bapi. Neural architecture of speech. In *ICASSP*, 2023.

[Oota *et al.*, 2023d] Subba Reddy Oota, Agarwal Veeral, Marreddy Mounika, Gupta Manish, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *24th INTERSPEECH Conference*, 2023.

[Oseki and Asahara, 2020] Yohei Oseki and M Asahara. Design of bccwj-eeg: Balanced corpus with human electroencephalography. In *LREC*, pages 189–194, 2020.

[Ozcelik and VanRullen, 2023] Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*, 2023.

[Pandey *et al.*, 2022] Pankaj Pandey, Gulshan Sharma, Krishna P Miyapuram, Ramanathan Subramanian, and Derek Lomas. Music identification using brain responses to initial snippets. In *ICASSP*, pages 1246–1250, 2022.

[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.*, 2016] Francisco Pereira, Bin Lou, Brianna Pritchett, Nancy Kanwisher, Matthew Botvinick, and Ev Fedorenko. Decoding of generic mental representations from functional mri data using word embeddings. *bioRxiv*, page 057216, 2016.

[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):1–13, 2018.

[Popham *et al.*, 2021] Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636, 2021.

[Reddy and Wehbe, 2021] Aniketh Janardhan Reddy and Leila Wehbe. Can fmri reveal the representation of syntactic structure in the brain? *NeurIPS*, 34:9843–9856, 2021.

[Schrimpf *et al.*, 2020] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020.

[Schrimpf *et al.*, 2021a] Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *PNAS*, Vol:To appear, 2021.

[Schrimpf *et al.*, 2021b] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *PNAS*, 118(45), 2021.

[Schwartz *et al.*, 2019] Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *NIPS*, 32:14123–14133, 2019.

[Seeliger *et al.*, 2019] K Seeliger, RP Sommers, Umut Güçlü, Sander E Bosch, and MAJ Van Gerven. A large single-participant fmri dataset for probing brain responses to naturalistic stimuli in space and time. *bioRxiv*, page 687681, 2019.

[Singh *et al.*, 2007] Vishwajeet Singh, Krishna P. Miyapuram, and Raju S. Bapi. Detection of cognitive states from fmri data using machine learning techniques. In Manuela M. Veloso, editor, *IJCAI*, pages 587–592, 2007.

[Singh *et al.*, 2023] Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.

[Smallwood and Schooler, 2015] Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015.

[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.*, 2019] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Towards sentence-level brain decoding with distributed representations. In *AAAI*, pages 7047–7054, 2019.

[Sun *et al.*, 2020] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Neural encoding and decoding with distributed sentence representations. *IEEE TNNLS*, 32(2):589–603, 2020.

[Takagi and Nishimoto, 2022] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, pages 2022–11, 2022.

[Tang *et al.*, 2022] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *bioRxiv*, pages 2022–09, 2022.

[Tang *et al.*, 2023] Jerry Tang, Meng Du, Vy A Vo, Vasudev Lal, and Alexander G Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv preprint arXiv:2305.12248*, 2023.

[Thirion *et al.*, 2006] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

[Toneva and Wehbe, 2019] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*, 2019.

[Toneva *et al.*, 2020] Mariya Toneva, Otilia Stretcu, Barnabás Póczos, Leila Wehbe, and Tom M Mitchell. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *NIPS*, 33, 2020.

[Toneva *et al.*, 2021] Mariya Toneva, Jennifer Williams, Anand B, Christoph Dann, and Leila Wehbe. Same cause; different effects in the brain. In *CLeaR*, 2021.

[Toneva *et al.*, 2022] Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022.

[Tuckute *et al.*, 2022] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence. *bioRxiv*, 2022.

[Tuckute *et al.*, 2023] Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *bioRxiv*, 2023.

[Vaidya *et al.*, 2022] Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*, 2022.

[Wang *et al.*, 2017] Jing Wang, Vladimir L Cherkassky, and M Adam Just. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *HBM*, 10:4865–4881, 2017.

[Wang *et al.*, 2019] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *NeurIPS*, 32:15501–15511, 2019.

[Wang *et al.*, 2020] 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.

[Wang *et al.*, 2022] Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09, 2022.

[Wehbe *et al.*, 2014] 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.

[Yamins *et al.*, 2014] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, 111(23):8619–8624, 2014.

[Zhang *et al.*, 2020] Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1–13, 2020.

[Zhang *et al.*, 2022a] Xiaohan Zhang, Shaonan Wang, Nan Lin, Jiajun Zhang, and Chengqing Zong. Probing word syntactic representations in the brain by a feature elimination method. *AAAI*, 2022.

[Zhang *et al.*, 2022b] Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. Is the brain mechanism for hierarchical structure building universal across languages? an fmri study of chinese and english. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7852–7861, 2022.

[Zinszer *et al.*, 2018] 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):011003–011003, 2018.