



HAL
open science

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

► **To cite this version:**

Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gael Jobard, Frédéric Alexandre, et al.. Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey). 2023. hal-04162064

HAL Id: hal-04162064

<https://hal.science/hal-04162064v1>

Preprint submitted on 14 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Subba Reddy Oota^{1,2}, Manish Gupta^{3,4}, Raju S. Bapi³, Gael Jobard²
Frederic Alexandre^{1,2}, Xavier Hinaut^{1,2}

¹INRIA, Bordeaux, France, ²University of Bordeaux, France, ³IIIT Hyderabad, India, ⁴Microsoft, 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

1 How does the brain represent different modes of
2 information? Can we design a system that auto-
3 matically understands what the user is think-
4 ing? Such questions can be answered by study-
5 ing brain recordings like functional magnetic re-
6 sonance imaging (fMRI). As a first step, the neu-
7 roscience community has contributed several large
8 cognitive neuroscience datasets related to passive
9 reading/listening/viewing of concept words, narra-
10 tives, pictures and movies. Encoding and decod-
11 ing models using these datasets have also been pro-
12 posed in the past two decades. These models serve
13 as additional tools for basic research in cognitive
14 science and neuroscience. Encoding models aim
15 at generating fMRI brain representations given a
16 stimulus automatically. They have several practi-
17 cal applications in evaluating and diagnosing neu-
18 rological conditions and thus also help design ther-
19 apies for brain damage. Decoding models solve
20 the inverse problem of reconstructing the stim-
21 uli given the fMRI. They are useful for designing
22 brain-machine or brain-computer interfaces. In-
23 spired by the effectiveness of deep learning mod-
24 els for natural language processing, computer vi-
25 sion, and speech, recently several neural encoding
26 and decoding models have been proposed. In this
27 survey, we will first discuss popular representations
28 of language, vision and speech stimuli, and present
29 a summary of neuroscience datasets. Further, we
30 will review popular deep learning based encoding
31 and decoding architectures and note their benefits
32 and limitations. Finally, we will conclude with a
33 brief summary and discussion about future trends.
34 Given the large amount of recently published work
35 in the ‘computational cognitive neuroscience’ com-
36 munity, we believe that this survey nicely organizes
37 the plethora of work and presents it as a coherent
38 story.

1 Introduction

40 Neuroscience is the field of science that studies the structure
41 and function of the nervous system of different species. It

involves answering interesting questions like the following¹.
(1) How learning occurs during adolescence, and how it dif-
fers 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 de-
velop speech and learn languages. Neuroscientists study di-
verse topics that help us understand how the brain and ner-
vous 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 neu-
ral networks (DNN) offer a computational medium to cap-
ture the unprecedented complexity and richness of brain ac-
tivity. *Encoding* and *decoding* stated as computational prob-
lems succinctly encapsulate this puzzle. As the previous sur-
veys systematically explore the brain encoding and decod-
ing 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 decod-
ing, as shown in Figure 1. Encoding is the process of learn-
ing 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 learn-
ing 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 se-
mantic 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 pro-
cess new stimuli and brain activations, respectively.

Techniques for recording brain activations: Popular tech-
niques for recording brain activations include single Micro-

¹<https://zuckermaninstitute.columbia.edu/file/5184/download?token=qzld8vyR>

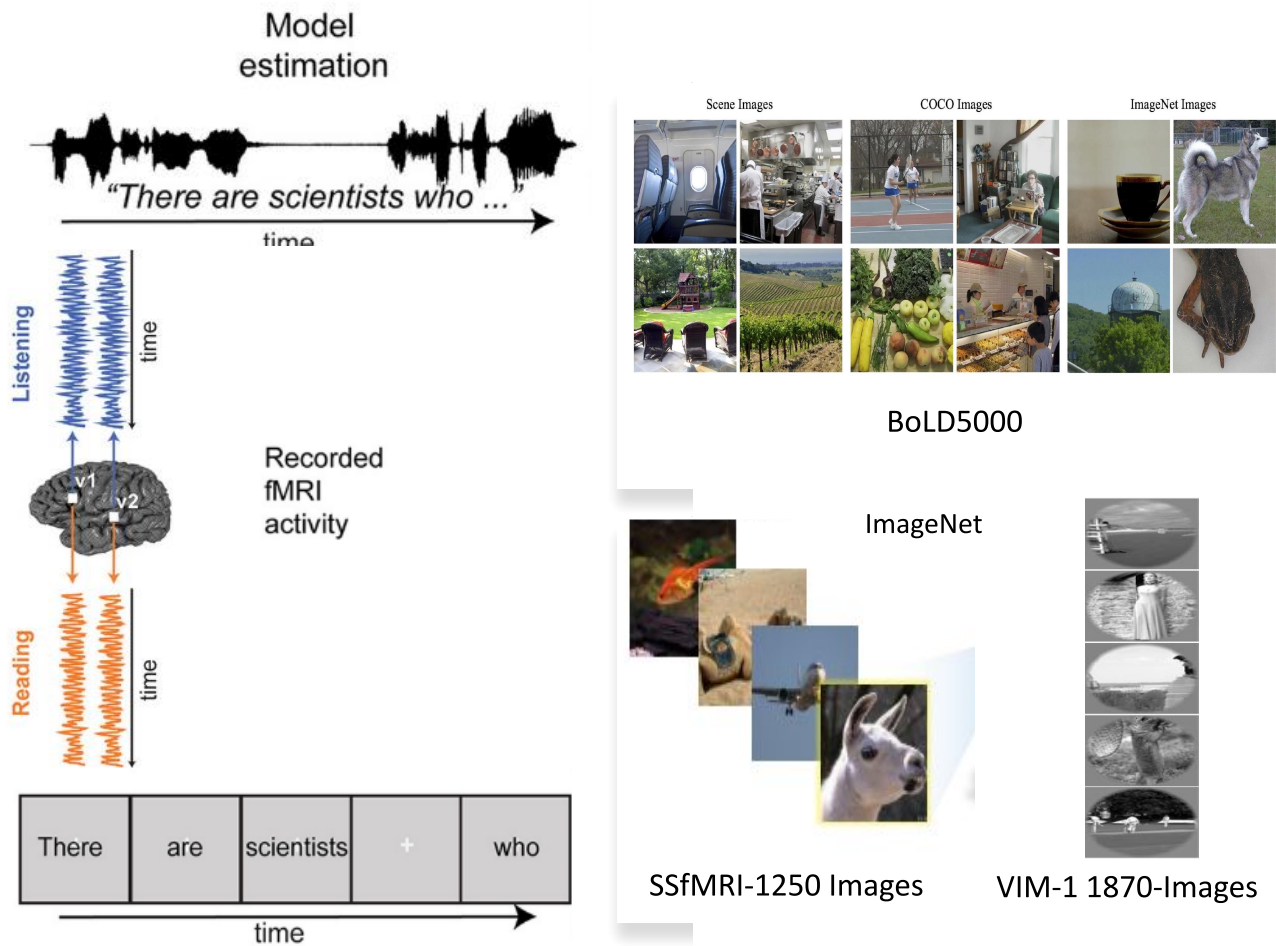


Figure 2: Representative Samples of Naturalistic Brain Dataset: (LEFT) Brain activity recorded when subjects are reading and listening to the same narrative (Deniz *et al.* 2019), and (RIGHT) example naturalistic image stimuli from various public repositories: BOLD5000 (Chang *et al.* 2019), SSfMRI (Beliy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

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

154 **Brain Decoding:** Ridge regression is the most popular brain
 155 decoder. Recently, a fully connected layer [Beliy *et al.*, 2019]
 156 or multi-layered perceptrons (MLPs) [Sun *et al.*, 2019] have
 157 also been used. While older methods attempted to decode to a
 158 vector representation using stimuli of a single mode, newer
 159 methods focus on multimodal stimuli decoding [Pereira *et al.*,
 160 2016; Oota *et al.*, 2022c]. Decoding using Transformers [Gauthier
 161 and Levy, 2019; Toneva and Wehbe, 2019; Défossez *et al.*,
 162 2022; Tang *et al.*, 2022], and decoding to actual stimuli (word,
 163 passage, image, dialogues) have also been explored. We discuss
 164 details of these decoding methods in Sec. 6.
 165

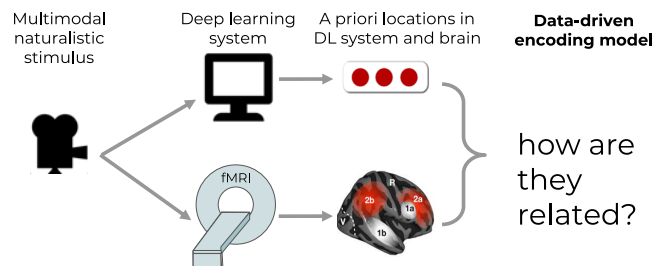


Figure 3: Alignment between deep learning systems and human brains [Toneva *et al.* 2019].

Computational Cognitive Science (CCS) Research goals: 166
 CCS researchers have primarily focused on two main areas 167
 [Doerig *et al.*, 2022] (also, see Figure 3). (1) Improving 168
 predictive Accuracy. In this area, the work is around the fol- 169
 lowing questions. (a) Compare feature sets: Which feature 170
 set provides the most faithful reflection of the neural repre- 171
 sentational space? (b) Test feature decodability: “Does neu- 172

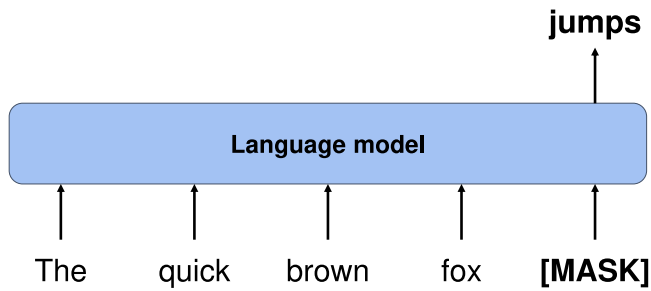


Figure 4: Language Model

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

186 2 Stimulus Representations

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

191 **Text Stimulus Representations:** Older methods for
 192 text-based stimuli representation include text corpus co-
 193 occurrence counts [Mitchell *et al.*, 2008; Pereira *et al.*, 2013;
 194 Huth *et al.*, 2016], topic models [Pereira *et al.*, 2013], syn-
 195 tactic features and discourse features [Wehbe *et al.*, 2014].
 196 In recent times, for text-based stimuli, both semantic mod-
 197 els as well as experiential attribute models have been ex-
 198 plored. Semantic representation models include word em-
 199 bedding methods [Pereira *et al.*, 2018; Wang *et al.*, 2020;
 200 Pereira *et al.*, 2016; Toneva and Wehbe, 2019; Anderson
 201 *et al.*, 2017a; Oota *et al.*, 2018], sentence representation
 202 models (see Figure 4) [Sun *et al.*, 2020; Sun *et al.*, 2019;
 203 Toneva and Wehbe, 2019], RNNs [Jain and Huth, 2018;
 204 Oota *et al.*, 2019] and Transformer methods [Gauthier and
 205 Levy, 2019; Toneva and Wehbe, 2019; Schwartz *et al.*, 2019;
 206 Schrimpf *et al.*, 2021a; Antonello *et al.*, 2021; Oota *et al.*,
 207 2022b; Aw and Toneva, 2022]. Popular word em-
 208 bedding methods include textual (i.e., Word2Vec, fastText,
 209 and GloVe), linguistic (i.e., dependency), conceptual (i.e.,
 210 RWSGwn and ConceptNet), contextual (i.e., ELMo). Pop-
 211 ular sentence embedding models include average, max, con-
 212 cat of avg and max, SIF, fairseq, skip, GenSen, InferSent,
 213 ELMo, BERT, RoBERTa, USE, QuickThoughts and GPT-
 214 2. Transformer-based methods include pretrained BERT with
 215 various NLU tasks, finetuned BERT, Transformer-XL, GPT-
 216 2, BART, BigBird, LED, and LongT5. Experiential attribute
 217 models represent words in terms of human ratings of their

218 degree of association with different attributes of experience,
 219 typically on a scale of 0-6 [Anderson *et al.*, 2019; Ander-
 220 son *et al.*, 2020; Berezutskaya *et al.*, 2020; Just *et al.*, 2010;
 221 Anderson *et al.*, 2017b] or binary [Handjaras *et al.*, 2016;
 222 Wang *et al.*, 2017].

223 **Visual Stimulus Representations:** For visual stimuli, older
 224 methods used visual field filter bank [Thirion *et al.*, 2006;
 225 Nishimoto *et al.*, 2011] and Gabor wavelet pyramid [Kay
 226 *et al.*, 2008; Naselaris *et al.*, 2009]. Recent methods use
 227 models like CNNs [Du *et al.*, 2020; Belyi *et al.*, 2019;
 228 Anderson *et al.*, 2017a; Yamins *et al.*, 2014; Nishida
 229 *et al.*, 2020] and concept recognition models [Anderson *et al.*,
 230 2020].

231 **Audio Stimuli Representations:** For audio stimuli, phoneme
 232 rate and presence of phonemes have been leveraged [Huth *et al.*,
 233 2016]. Recently, authors in [Nishida *et al.*, 2020] used
 234 features from an audio deep learning model called SoundNet
 235 for audio stimuli representation.

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

245 3 Naturalistic Neuroscience Datasets

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

252 **Text Datasets:** These datasets are created by presenting
 253 words, sentences, passages or chapters as stimuli. Some of
 254 the text datasets include Harry Potter Story [Wehbe *et al.*,
 255 2014], ZUCO EEG [Hollenstein *et al.*, 2018] and datasets
 256 proposed in [Handjaras *et al.*, 2016; Anderson *et al.*, 2017a;
 257 Anderson *et al.*, 2019; Wehbe *et al.*, 2014]. In [Handjaras
 258 *et al.*, 2016], participants were asked to verbally enumerate in
 259 one minute the properties (features) that describe the entities
 260 the words refer to. There were four groups of participants: 5
 261 sighted individuals were presented with a pictorial form of the
 262 nouns, 5 sighted individuals with a verbal-visual (i.e., written
 263 Italian words) form, 5 sighted individuals with a verbal audi-
 264 tory (i.e., spoken Italian words) form, and 5 congenitally
 265 blind with a verbal auditory form. Data proposed by [An-
 266 derson *et al.*, 2017a] contains 70 Italian words taken from
 267 seven taxonomic categories (abstract, attribute, communica-
 268 tion, event/action, person/social role, location, object/tool) in
 269 the law and music domain. The word list contains concrete
 270 as well as abstract words. ZUCO dataset [Hollenstein *et al.*,
 271 2018] contains sentences for which fMRIs were obtained for
 272 3 tasks: normal reading of movie reviews, normal reading of
 273 Wikipedia sentences and task-specific reading of Wikipedia
 274 sentences. For this dataset curation, sentences were presented

Table 1: Naturalistic Neuroscience Datasets

	Dataset	Authors	Type	Lang.	Stimulus	S	Task
Text	Harry Potter	[Wehbe <i>et al.</i> , 2014]	fMRI/MEG	English	Reading Chapter 9 of Harry Potter and the Sorcerer’s Stone	9	Story understanding
		[Handjaras <i>et al.</i> , 2016]	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns, four times	20	Property Generation
		[Anderson <i>et al.</i> , 2017a]	fMRI	Italian	Reading 70 concrete and abstract nouns from law/music, five times	7	Imagine a situation with noun
	ZuCo	[Hollenstein <i>et al.</i> , 2018]	EEG	English	Reading 1107 sentences with 21,629 words from movie reviews	12	Rate movie quality
	240 Sentences with Content Words	[Anderson <i>et al.</i> , 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
Visual	Subset Moth Radio Hour	[Deniz <i>et al.</i> , 2019]	fMRI	English	Reading 11 stories	9	Passive reading and Listening
		[Thirion <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 2019]	fMRI	-	Viewing 5254 images depicting real-world scenes	4	Passive viewing
	Algonauts	[Cichy <i>et al.</i> , 2019]	fMRI/MEG	-	Viewing 92 silhouette object images and 118 images of objects on natural background	15	Passive viewing
	NSD	[Allen <i>et al.</i> , 2022]	fMRI	-	Viewing 73000 natural scenes	8	Passive viewing
Audio	THINGS	[Hebart <i>et al.</i> , 2022]	fMRI/MEG	-	Viewing 31188 natural images	8	Passive viewing
		[Handjaras <i>et al.</i> , 2016]	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns, 4 times	20	Property Generation
	The Moth Radio Hour	[Huth <i>et al.</i> , 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 <i>et al.</i> , 2020]	fMRI	English	Listening one of 20 scenario names, 5 times	26	Imagine personal experiences
	Narratives	[Nastase <i>et al.</i> , 2021]	fMRI	English	Listening 27 diverse naturalistic spoken stories. 891 functional scans	345	Passive Listening
	Natural Stories	[Zhang <i>et al.</i> , 2020]	fMRI	English	Listening Moth-Radio-Hour naturalistic spoken stories.	19	Passive Listening
Video	The Little Prince	[Li <i>et al.</i> , 2021]	fMRI	English	Listening audiobook for about 100 minutes.	112	Passive Listening
	MEG-MASC	[Williams <i>et al.</i> , 2022]	MEG	English	Listening two hours of naturalistic stories. 208 MEG sensors	27	Passive Listening
	BBC’s Doctor Who	[Seeliger <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 2021]	fMRI	English	Viewing 1000 short video clips (3 sec each)	10	Passive viewing
	Natural Short Clips	[Huth <i>et al.</i> , 2022]	fMRI	English	Watching natural short movie clips	5	Passive viewing
Other Multimodal	Natural Short Clips	[Lahner <i>et al.</i> , 2023]	fMRI	English	Watching 1102 natural short video clips	10	Passive viewing
	60 Concrete Nouns	[Mitchell <i>et al.</i> , 2008]	fMRI	English	Viewing 60 different word-picture pairs from 12 categories, 6 times each	9	Passive viewing
		[Sudre <i>et al.</i> , 2012]	MEG	English	Reading 60 concrete nouns along with line drawings. 20 questions per noun lead to 1200 examples.	9	Question answering
		[Zinszer <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 2021]	fNIRS	Chinese	Viewing and listening 50 concrete nouns from 10 semantic categories.	7	Passive viewing and listening
	Neuromod	[Boyle <i>et al.</i> , 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

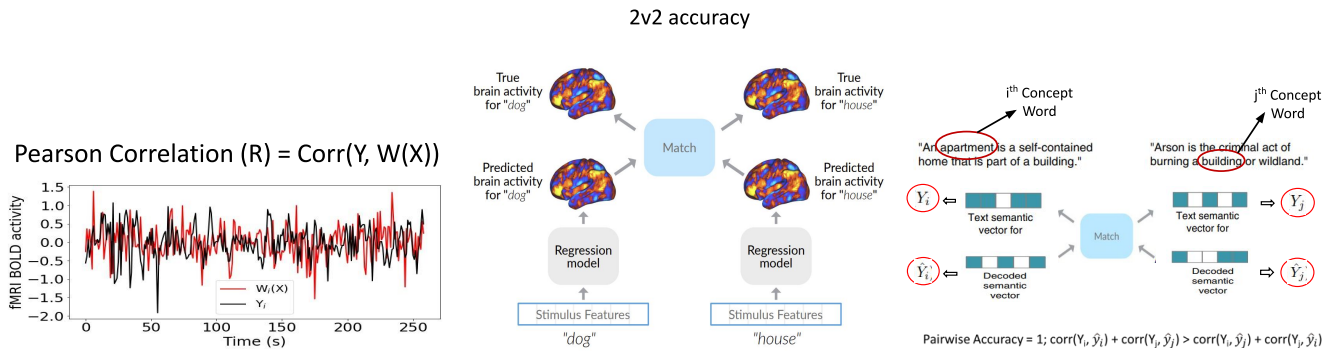


Figure 5: Evaluation Metrics for Brain Encoding and Decoding. (LEFT) Pearson Correlation, (MIDDLE) 2V2 Accuracy [Toneva et al. 2020], and (RIGHT) Pairwise Accuracy.

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 $\frac{1}{N C_2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N I[\{\cos D(Y_i, \hat{Y}_i) + \cos D(Y_j, \hat{Y}_j)\} < \{\cos D(Y_i, \hat{Y}_j) + \cos D(Y_j, \hat{Y}_i)\}]$ where $\cos D$ 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 $PC = \frac{1}{N} \sum_{i=1}^n \text{corr}[Y_i, \hat{Y}_i]$ where corr is the correla-

tion 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 $S = [S_0, S_1, \dots, S_n]$, $\hat{S} = [\hat{S}_0, \hat{S}_1, \dots, \hat{S}_n]$ 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 $\text{corr}(S_i, \hat{S}_i) + \text{corr}(S_j, \hat{S}_j) > \text{corr}(S_i, \hat{S}_j) + \text{corr}(S_j, \hat{S}_i)$, 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

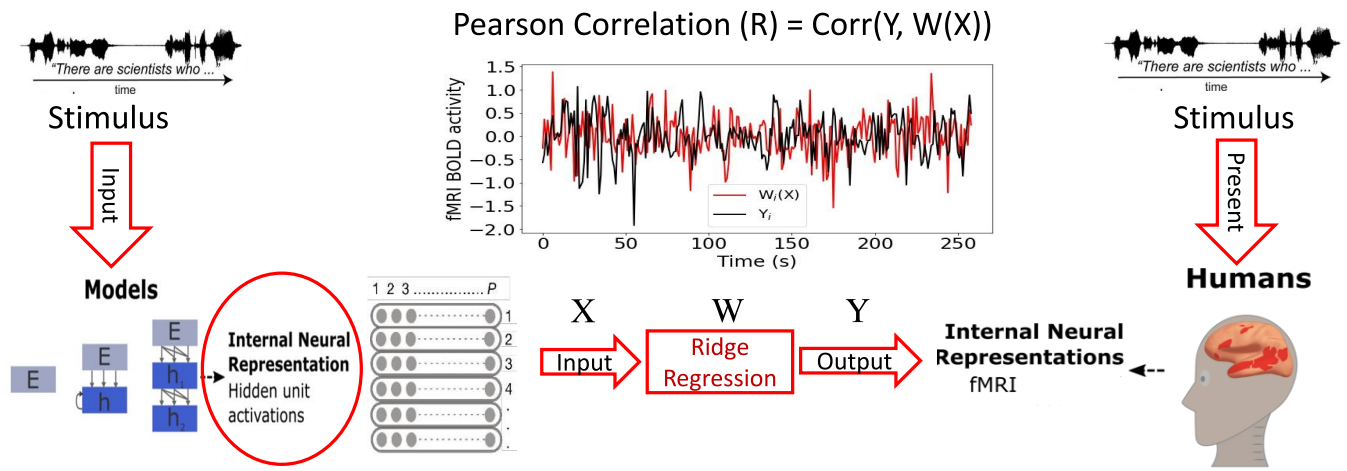


Figure 6: Schema for Brain Encoding

394 efforts are aimed at utilizing the internal representations
 395 extracted from transformer-based language models such as
 396 ELMo, BERT, GPT-2, etc for learning encoding models of
 397 brain activation [Jat *et al.*, 2020; Caucheteux *et al.*, 2021;
 398 Antonello *et al.*, 2021]. High-grain details such as lexical,
 399 compositional, syntactic, and semantic representations of nar-
 400 ratives are factorized from transformer-based models and uti-
 401 lized for training encoding models. The resulting models are
 402 better able to disentangle the corresponding brain responses
 403 in fMRI [Caucheteux *et al.*, 2021]. Finally, it has been found
 404 that the models that integrate task and stimulus representa-
 405 tions have significantly higher prediction performance than
 406 models that do not account for the task semantics [Toneva *et*
 407 *al.*, 2020; Schrimpf *et al.*, 2021a].

408 Similarly, in vision, early models focused on indepen-
 409 dent models of visual processing (object classification) us-
 410 ing CNNs [Yamins *et al.*, 2014]. Recent efforts in visual en-
 411 coding models focus on using richer visual representations
 412 derived from a variety of computer vision tasks [Wang *et*
 413 *al.*, 2019]. Instead of feed-forward deep CNN models, us-
 414 ing shallow recurrence enabled better capture of temporal dy-
 415 namics in the visual encoding models [Kubilius *et al.*, 2019;
 416 Schrimpf *et al.*, 2020].

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

422 **Linguistic Encoding:** A number of previous works have in-
 423 vestigated the alignment between pretrained language mod-
 424 els and brain recordings of people comprehending language.
 425 Huth *et al.* [2016] have been able to identify brain ROIs (Re-
 426 gions of Interest) that respond to words that have a similar
 427 meaning and have thus built a “semantic atlas” of how the
 428 human brain organizes language. Many studies have shown
 429 accurate results in mapping the brain activity using neural
 430 distributed word embeddings for linguistic stimuli [Ander-
 431 son *et al.*, 2017a; Pereira *et al.*, 2018; Oota *et al.*, 2018;
 432 Nishida and Nishimoto, 2018; Sun *et al.*, 2019]. Unlike ear-

433 lier models where each word is represented as an independ- 433
 434 ent vector in an embedding space, [Jain and Huth, 2018] 434
 435 built encoding models using rich contextual representations 435
 436 derived from an LSTM language model in a story listen- 436
 437 ing task. With these contextual representations, they demon- 437
 438 strated dissociation in brain activation – auditory cortex (AC) 438
 439 and Broca’s area in shorter context whereas left Temporo- 439
 440 Parietal junction (TPJ) in longer context. [Hollenstein *et al.*, 440
 441 2019] presents the first multimodal framework for evaluat- 441
 442 ing six types of word embedding (Word2Vec, WordNet2Vec, 442
 443 GloVe, FastText, ELMo, and BERT) on 15 datasets, includ- 443
 444 ing eye-tracking, EEG and fMRI signals recorded during lan- 444
 445 guage processing. With the recent advances in contextual rep- 445
 446 resentations in NLP, few studies incorporated them in relating 446
 447 sentence embeddings with brain activity patterns [Sun *et al.*, 447
 448 2020; Gauthier and Levy, 2019; Jat *et al.*, 2020]. 448

449 More recently, researchers have begun to study the align- 449
 450 ment of language regions of the brain with the layers of lan- 450
 451 guage models and found that the best alignment was achieved 451
 452 in the middle layers of these models [Jain and Huth, 2018; 452
 453 Toneva and Wehbe, 2019]. Schrimpf *et al.* [2021a] examined 453
 454 the relationship between 43 diverse state-of-the-art language 454
 455 models. They also studied the behavioral signatures of human 455
 456 language processing in the form of self-paced reading times, 456
 457 and a range of linguistic functions assessed via standard engi- 457
 458 neering tasks from NLP. They found that Transformer-based 458
 459 models perform better than RNNs or word-level embedding 459
 460 models. Larger-capacity models perform better than smaller 460
 461 models. Models initialized with random weights (prior to 461
 462 training) perform surprisingly similarly in neural predictiv- 462
 463 ity as compared to final trained models, suggesting that net- 463
 464 work architecture contributes as much or more than experi- 464
 465 ence dependent learning to a model’s match to the brain. 465
 466 Antonello *et al.* [2021] proposed a “language representation 466
 467 embedding space” and demonstrated the effectiveness of the 467
 468 features from this embedding in predicting fMRI responses 468
 469 to linguistic stimuli. 469

470 **Disentangling the Syntax and Semantics:** The represen- 470
 471 tations of transformer models like BERT, GPT-2 have been 471

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 <i>et al.</i> , 2020]	MEG	English	BERT	9	Question-Answering	Ridge
	[Schrimpf <i>et al.</i> , 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 <i>et al.</i> , 2019]	fMRI	English	GloVe	9	Subset Moth Radio Hour	Ridge
	[Jain <i>et al.</i> , 2020]	fMRI	English	LSTM	6	Subset Moth Radio Hour	Ridge
	[Caucheteux <i>et al.</i> , 2021]	fMRI	English	GPT-2, Basic syntax features	345	Narratives	Ridge
	[Antonello <i>et al.</i> , 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 <i>et al.</i> , 2022]	fMRI	English	GloVe, GPT-2 next word, pre-onset, post-onset word surprise	8	ECoG	
	[Oota <i>et al.</i> , 2022b]	fMRI	English	BERT and GLUE tasks	82	Pereira & Narratives	Ridge
	[Oota <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 2022b]	fMRI	English, Chinese	Node Count	19, 12	Zhang	Ridge
	[Oota <i>et al.</i> , 2023a]	fMRI	English	Constituency, Dependency trees, Basic syntax features and BERT	82	Narratives	Ridge
	[Oota <i>et al.</i> , 2023b]	MEG	English	Basic syntax features, GloVe and BERT	8	MEG-MASC	Ridge
	[Tuckute <i>et al.</i> , 2023]	fMRI	English	BERT-Large, GPT-2 XL	12	Reading Sentences	Ridge
[Kauf <i>et al.</i> , 2023]	fMRI	English	BERT-Large, GPT-2 XL	12	Pereira	Ridge	
[Singh <i>et al.</i> , 2023]	fMRI	English	BERT-Large, GPT-2 XL, Text Perturbations	5	Pereira	Ridge	
Visual	[Wang <i>et al.</i> , 2019]	fMRI		21 downstream vision tasks	4	BOLD 5000	Ridge
	[Kubilius <i>et al.</i> , 2019]	fMRI		CNN models AlexNet, ResNet, DenseNet	7	Algonauts	Ridge
	[Dwivedi <i>et al.</i> , 2021]	fMRI		21 downstream vision tasks	4	BOLD 5000	Ridge
	[Khosla and Wehbe, 2022]	fMRI		CNN models AlexNet	4	BOLD 5000	Ridge
	[Conwell <i>et al.</i> , 2023]	fMRI		CNN models AlexNet	4	BOLD 5000	Ridge
Audio	[Millet <i>et al.</i> , 2022]	fMRI	English	Wav2Vec2.0	345	Narratives	Ridge
	[Vaidya <i>et al.</i> , 2022]	fMRI	English	APC, AST, Wav2Vec2.0, and HuBERT	7	Moth Radio Hour	Ridge
	[Tuckute <i>et al.</i> , 2022]	fMRI	English	19 Speech Models (e.g. DeepSpeech, Wav2Vec2.0, VQ-VAE)	19	Passive listening	Ridge
	[Oota <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 2021]	fMRI	English	985D Semantic Vector	5	Moth Radio Hour & Short Movie Clips	Ridge
	[Oota <i>et al.</i> , 2022d]	fMRI	English	CLIP, VisualBERT, LXMERT, CNNs and BERT	5, 82	Pereira & Narratives	Ridge
	[Lu <i>et al.</i> , 2022]	fMRI	English	BriVL	5	Pereira & Short Movie Clips	Ridge
	[Tang <i>et al.</i> , 2023]	fMRI	English	BridgeTower	5	Moth Radio Hour & Short Movie Clips	Ridge

472 shown to linearly map onto brain activity during language
473 comprehension. Several studies have attempted to disentangle
474 the contributions of different types of information from
475 word representations to the alignment between brain recordings
476 and language models. Wang *et al.* [2020] proposed
477 a two-channel variational autoencoder model to dissociate
478 sentences into semantic and syntactic representations and
479 separately associate them with brain imaging data to find
480 feature-correlated brain regions. To separate each syntactic
481 feature, Zhang *et al.* [2022a] proposed a feature elimination
482 method, called Mean Vector Null space Projection.
483 Compared with word representations, word syntactic features
484 (parts-of-speech, named entities, semantic roles, dependencies)
485 seem to be distributed across brain networks instead of
486 a local brain region. In the previous two studies, we do not
487 know whether all or any of these representations effectively

488 drive the linear mapping between language models (LMs) and
489 the brain. Toneva *et al.* [2022] presented an approach to dis-
490 entangle supra-word meaning from lexical meaning in lan-
491 guage models and showed that supra-word meaning is pre-
492 dictive of fMRI recordings in two language regions (anterior
493 and posterior temporal lobes). Caucheteux *et al.* [2021] pro-
494 posed a taxonomy to factorize the high-dimensional activa-
495 tions of language models into four combinatorial classes: lex-
496 ical, compositional, syntactic, and semantic representations.
497 They found that (1) Compositional representations recruit a
498 more widespread cortical network than lexical ones, and en-
499 compass the bilateral temporal, parietal and prefrontal cor-
500 tices. (2) Contrary to previous claims, syntax and semantics
501 are not associated with separated modules, but, instead, ap-
502 pear to share a common and distributed neural substrate.

503 While previous works studied syntactic processing as cap-

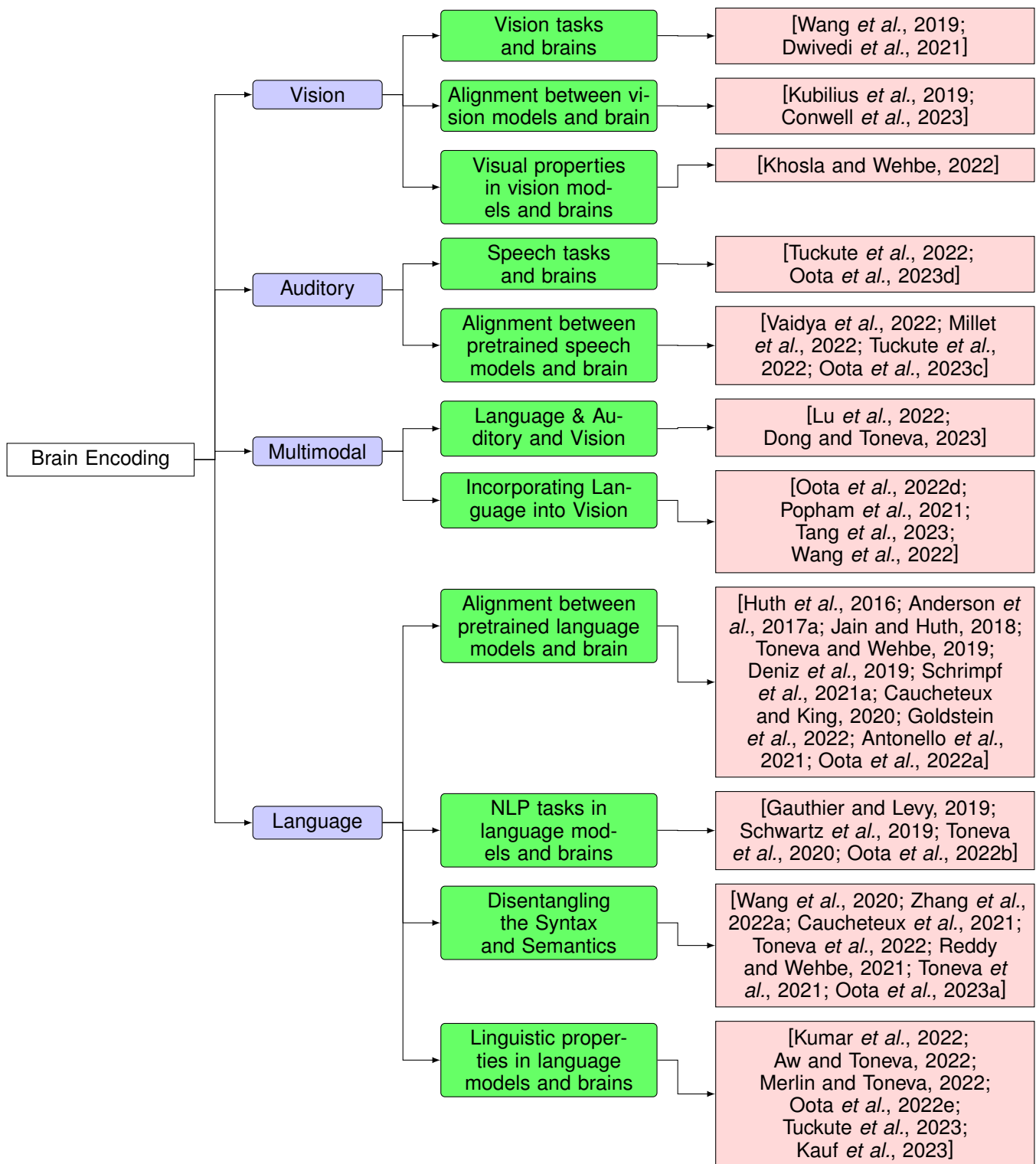


Figure 7: Brain Encoding Survey Tree

504 tured through complexity measures (syntactic surprisal, node
 505 count, word length, and word frequency), very few have stud-
 506 ied the syntactic representations themselves. Studying syn-

tactic representations using fMRI is difficult because: (1) 507
 representing syntactic structure in an embedding space is a 508
 non-trivial computational problem, and (2) the fMRI signal 509

510 is noisy. To overcome these limitations, Reddy et al. [2021]
511 proposed syntactic structure embeddings that encode the syn-
512 tactic information inherent in natural text that subjects read
513 in the scanner. The results reveal that syntactic structure-
514 based features explain additional variance in the brain activity
515 of various parts of the language system, even after control-
516 ling for complexity metrics that capture the processing load.
517 Toneva et al. [2021] further examined whether the represen-
518 tations obtained from a language model align with different
519 language processing regions in a similar or different way.

520 **Linguistic properties in LMs and brains:** Understanding
521 the reasons behind the observed similarities between lan-
522 guage comprehension in LMs and brains can lead to more
523 insights into both systems. Several works [Schwartz *et al.*,
524 2019; Kumar *et al.*, 2022; Aw and Toneva, 2022; Merlin and
525 Toneva, 2022; Oota *et al.*, 2022b] have found that using a
526 fine-tuned BERT leads to improved brain predictions. How-
527 ever, it is not clear what type of information in the fine-tuned
528 BERT model led to the improvement. It is unclear whether
529 and how the two systems align in their information processing
530 pipeline. Aw and Toneva [2022] used four pre-trained large
531 language models (BART, Longformer Encoder Decoder, Big-
532 Bird, and LongT5) and also trained them to improve their
533 narrative understanding, using the method detailed in Fig-
534 ure 8. However, it is not understood whether prediction of
535 the next word is necessary for the observed brain alignment
536 or simply sufficient, and whether there are other shared mech-
537 anisms or information that is similarly important. Merlin and
538 Toneva [2022] proposed two perturbations to pretrained lan-
539 guage models that, when used together, can control for the ef-
540 fects of next word prediction and word-level semantics on the
541 alignment with brain recordings. Specifically, they find that
542 improvements in alignment with brain recordings in two lan-
543 guage processing regions—Inferior Frontal Gyrus (IFG) and
544 Angular Gyrus (AG)—are due to next word prediction and
545 word-level semantics. However, what linguistic information
546 actually underlies the observed alignment between brains and
547 language models is not clear. Recently, Oota et al. [2022e]
548 tested the effect of a range of linguistic properties (surface,
549 syntactic and semantic) and found that the elimination of each
550 linguistic property results in a significant decrease in brain
551 alignment across all layers of BERT.

552 **Visual Encoding:** CNNs are currently the best class of mod-
553 els of the neural mechanisms of visual processing [Du *et al.*,
554 2020; Beliy *et al.*, 2019; Oota *et al.*, 2019; Nishida *et al.*,
555 2020]. How can we push these deeper CNN models to cap-
556 ture brain processing even more stringently? Continued ar-
557 chitectural optimization on ImageNet alone no longer seems
558 like a viable option. Kubilius et al. [2019] proposed a shal-
559 low recurrent anatomical network CORnet that follows neu-
560 roanatomy more closely than standard CNNs, and achieved
561 the state-of-the-art results on the Brain-score benchmark. It
562 has four computational areas, conceptualized as analogous to
563 the ventral visual areas V1, V2, V4, and IT, and a linear cate-
564 gory decoder that maps from the population of neurons in the
565 model’s last visual area to its behavioral choices.

566 Despite the effectiveness of CNNs, it is difficult to draw
567 specific inferences about neural information processing us-
568 ing CNN- derived representations from a generic object-

classification CNN. Hence, Wang et al. [2019] built encoding
569 models with individual feature spaces obtained from 21 com-
570 puter vision tasks. One of the main findings is that features
571 from 3D tasks, compared to those from 2D tasks, predict a
572 distinct part of visual cortex.
573

574 **Auditory Encoding:** Speech stimuli have mostly been rep-
575 resented using encodings of text transcriptions [Huth *et al.*,
576 2016] or using basic features like phoneme rate, the sum of
577 squared FFT coefficients [Pandey *et al.*, 2022], etc. Text
578 transcription-based methods ignore the raw audio-sensory in-
579 formation completely. The basic speech feature engineering
580 method misses the benefits of transfer learning from rigor-
581 ously pretrained speech DL models.

582 Recently, several researchers have used popular deep
583 learning models such as APC [Chung *et al.*, 2020],
584 Wav2Vec2.0 [Baevski *et al.*, 2020], HuBERT [Hsu *et al.*,
585 2021], and Data2Vec [Baevski *et al.*, 2022] for encoding
586 speech stimuli. Millet et al. [2022] used a self-supervised
587 learning model Wav2Vec2.0 to learn latent representations
588 of the speech waveform similar to those of the human brain.
589 They find that the functional hierarchy of its transformer lay-
590 ers aligns with the cortical hierarchy of speech in the brain,
591 and reveals the whole-brain organisation of speech processing
592 with an unprecedented clarity. This means that the first trans-
593 former layers map onto the low-level auditory cortices (A1
594 and A2), the deeper layers (orange and red) map onto brain
595 regions associated with higher-level processes (e.g. STS and
596 IFG). Vaidya et al. [2022] present the first systematic study
597 to bridge the gap between recent four self-supervised speech
598 representation methods (APC, Wav2Vec, Wav2Vec2.0, and
599 HuBERT) and computational models of the human auditory
600 system. Similar to [Millet *et al.*, 2022], they find that self-
601 supervised speech models are the best models of auditory ar-
602 eas. Lower layers best modeled low-level areas, and upper-
603 middle layers were most predictive of phonetic and semantic
604 areas, while layer representations follow the accepted hier-
605 archy of speech processing. Tuckute et al. [2022] analyzed
606 19 different speech models and find that some audio models
607 derived in engineering contexts (model applications ranged
608 from speech recognition and speech enhancement to audio
609 captioning and audio source separation) produce poor predic-
610 tions of auditory cortical responses, many task-optimized au-
611 dio speech deep learning models outpredict a standard spec-
612 trotemporal model of the auditory cortex and exhibit hierar-
613 chical layer-region correspondence with auditory cortex.

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

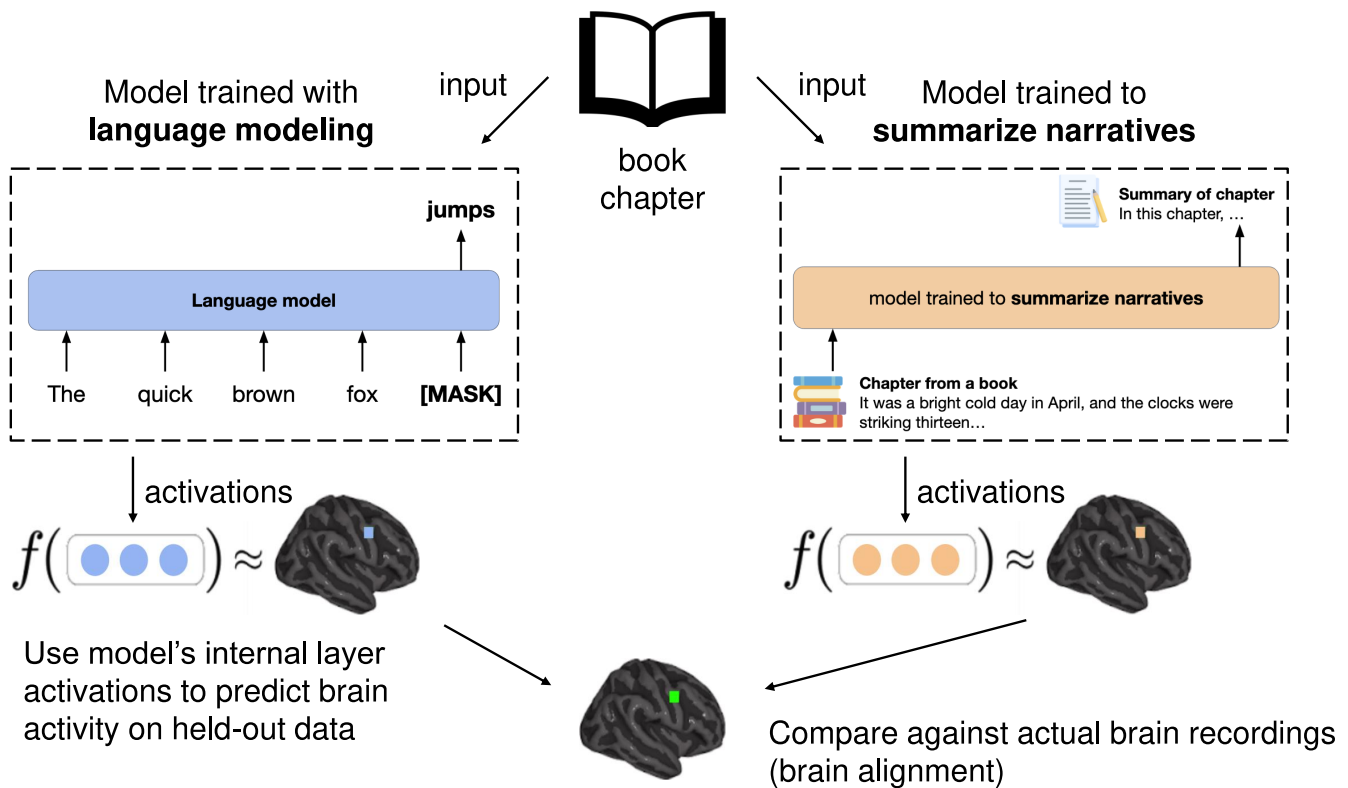


Figure 8: Comparison of brain recordings with language models trained on web corpora (LEFT) and language models trained on book stories (RIGHT) [Aw and Toneva, 2022].

628 intermediate layers of a multimodal video transformer are bet-
 629 ter at predicting multimodal brain activity than other layers,
 630 indicating that the intermediate layers encode the most brain-
 631 related properties of the video stimuli. Recently, [Tang *et al.*,
 632 2023] investigated a multimodal Transformer as the encoder
 633 architecture to extract the aligned concept representations for
 634 narrative stories and movies to model fMRI responses to nat-
 635 uralistic stories and movies, respectively. Since language and
 636 vision rely on similar concept representations, the authors
 637 perform a cross-modal experiment in which how well the lan-
 638 guage encoding models can predict movie-fMRI responses
 639 from narrative story features (story \rightarrow movie) and how well
 640 the vision encoding models can predict narrative story-fMRI
 641 responses from movie features (movie \rightarrow story). Overall, the
 642 authors find that cross-modality performance was higher for
 643 features extracted from multimodal transformers than for lin-
 644 early aligned features extracted from unimodal transformers.

645 6 Brain Decoding

646 Decoding is the learning of the mapping from neural activa-
 647 tions back to the stimulus domain. Figure 9 depicts the typical
 648 workflow for building an image/language decoder.

649 **Decoder Architectures:** In most cases, the stimulus repre-
 650 sentation is decoded using typical ridge regression models
 651 trained on each voxel and its 26 neighbors in 3D to pre-
 652 dict each dimension of the stimulus representation. Also,
 653 decoding is usually performed using the most informative

654 voxels [Pereira *et al.*, 2018]. In some cases, a fully con-
 655 nected layer [Beliy *et al.*, 2019] or a multi-layered percep-
 656 tron [Sun *et al.*, 2019] has been used. In some studies,
 657 when decoding is modeled as multi-class classification, Gaus-
 658 sian Naïve Bayes [Singh *et al.*, 2007; Just *et al.*, 2010] and
 659 SVMs [Thirion *et al.*, 2006] have also been used for decod-
 660 ing. Figure 10 summarizes the literature related to various
 661 decoding solutions proposed in vision, auditory, and language
 662 domains.

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

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

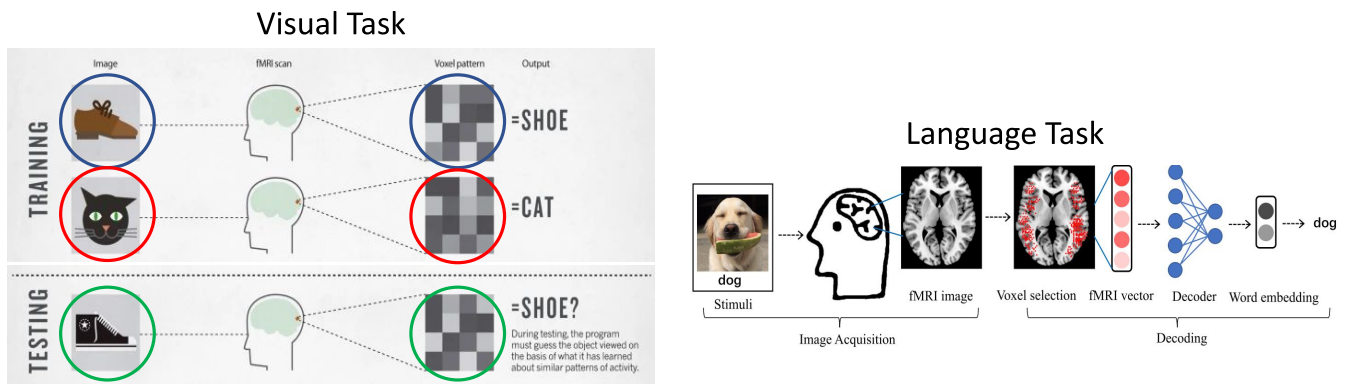


Figure 9: Schema for Brain Decoding. LEFT: Image decoder [Smith et al. 2011], RIGHT: Language Decoder [Wang et al. 2019]

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	

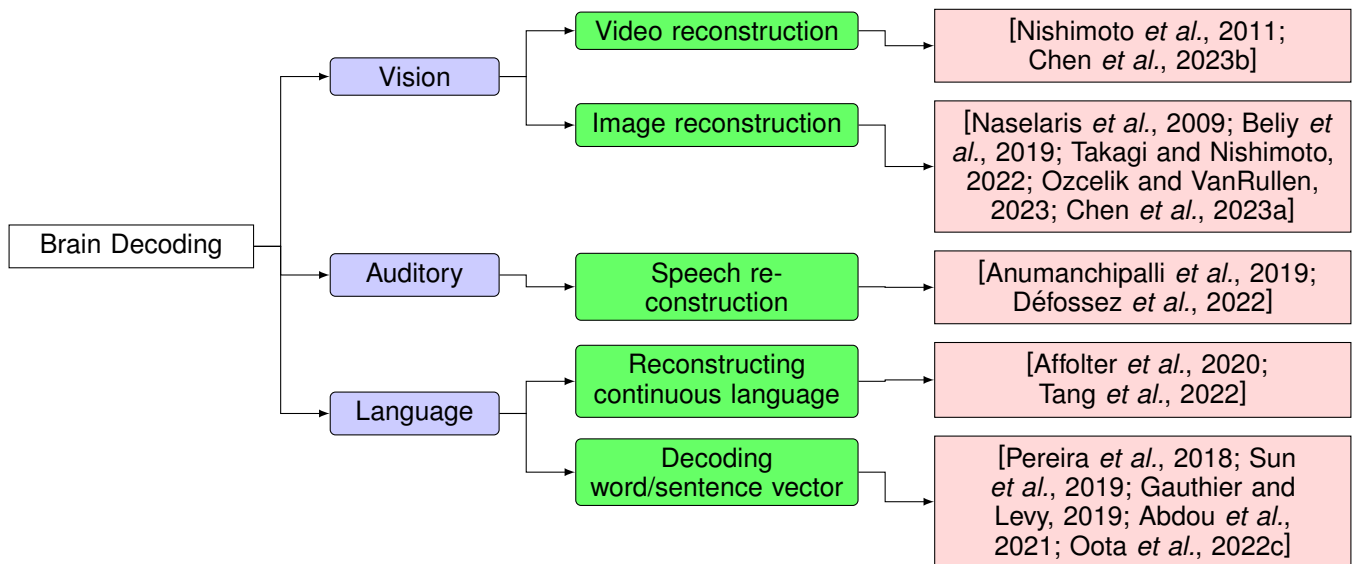


Figure 10: Brain Decoding Survey Tree

681 et al. [2022c] propose two novel brain decoding setups: (1)
 682 multi-view decoding (MVD) and (2) cross-view decoding
 683 (CVD). In MVD, the goal is to build an MV decoder that
 684 can take brain recordings for any view as input and predict
 685 the concept. In CVD, the goal is to train a model which takes
 686 brain recordings for one view as input and decodes a seman-

687 tic vector representation of another view. Specifically, they
 688 study practically useful CVD tasks like image captioning, im-
 689 age tagging, keyword extraction, and sentence formation.

690 To understand application of Transformer models for de-
 691 coding better, Gauthier et al. [2019] fine-tuned a pre-trained
 692 BERT on a variety of NLU tasks, asking which lead to im-

693 improvements in brain-decoding performance. They find that
694 tasks which produce syntax-light representations yield signif-
695 icant improvements in brain decoding performance. Toneva
696 et al. [2019] study how representations of various Trans-
697 former models differ across layer depth, context length, and
698 attention type.

699 Some studies have attempted to reconstruct words [Affolter
700 et al., 2020], continuous language [Tang et al., 2022], im-
701 ages [Du et al., 2020; Belyi et al., 2019; Fang et al., 2020;
702 Lin et al., 2022], speech [Défossez et al., 2022] or question-
703 answer speech dialogues [Moses et al., 2019] rather than just
704 predicting a semantic vector representation. Lastly, some
705 studies have focused on reconstructing personal imagined ex-
706 periences [Berezutskaya et al., 2020] or application-based
707 decoding like using brain activity scanned during a picture-
708 based mechanical engineering task to predict individuals’
709 physics/engineering exam results [Cetron et al., 2019] and
710 reflecting whether current thoughts are detailed, correspond
711 to the past or future, are verbal or in images [Smallwood and
712 Schooler, 2015]. Table 3 aggregates the brain decoding liter-
713 ature along different stimulus domains such as textual, visual,
714 and audio.

715 7 Conclusion, Limitations, and Future Trends

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

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

732 **Future Trends** Some of the future areas of work in this field
733 are as follows: (1) While there is work on the text, under-
734 standing the similarity in information processing between vi-
735 sual/speech/multimodal models versus natural brain systems
736 remains an open area. (2) Decoding to actual multimodal
737 stimuli seems feasible thanks to recent advances in generation
738 using deep learning models. (3) Deeper understanding of the
739 degree to which damage to different parts of the human brain
740 could lead to the degradation of cognitive skills. (4) How can
741 we train artificial neural networks in novel self-supervised
742 ways such that they compose word meanings or comprehend
743 images and speech like a human brain? (5) How can we lever-
744 age improved neuroscience understanding to suggest changes
745 in proposed artificial neural network architectures to make
746 them more robust and accurate? We hope that this survey
747 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. 748
- [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. 749
- [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. 750
- [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. 751
- [Anderson et al., 2017b] Andrew James Anderson, Jeffrey R Binder, Leonardo Ferdinando, 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. 752
- [Anderson et al., 2019] Andrew James Anderson, Jeffrey R Binder, Leonardo Ferdinando, 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. 753
- [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. 754
- [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. 755
- [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. 756
- [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. 757
- [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. 758
- [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. 759
- [Belyi et al., 2019] Roman Belyi, 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. 760
- [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. 761
- [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. 762
- [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. 763
- [Cao et al., 2021] Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. Brain decoding using fmris. In *AAAI*, volume 35, pages 12602–12611, 2021. 764
- [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. 765
- [Caucheteux et al., 2021] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Disentangling syntax and semantics in the brain with deep networks. In *ICML*, pages 1336–1348. PMLR, 2021. 766
- [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. 767
- [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. 768

- 818 [Chen *et al.*, 2023a] Xuhang Chen, Baiying Lei, Chi-Man Pun, and Shuqiang Wang. 819 Brain diffuser: An end-to-end brain image to brain network pipeline. *arXiv preprint* 820 *arXiv:2303.06410*, 2023.
- 821 [Chen *et al.*, 2023b] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic 822 mindscapes: High-quality video reconstruction from brain activity. *arXiv preprint* 823 *arXiv:2305.11675*, 2023.
- 824 [Chung *et al.*, 2020] Yu-An Chung, Hao Tang, and James Glass. Vector-quantized au- 825 toregressive predictive coding. *Interspeech*, pages 3760–3764, 2020.
- 826 [Cichy *et al.*, 2019] Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij 827 Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramak- 828 ishnan, and Aude Oliva. The algonauts project: A platform for communication 829 between the sciences of biological and artificial intelligence. *arXiv e-prints*, pages 830 *arXiv-1905*, 2019.
- 831 [Cichy *et al.*, 2021] Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex 832 Lascelles, Polina Iamshchinina, M Graumann, A Andonian, NAR Murty, K Kay, 833 Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain 834 makes sense of a world in motion. *arXiv preprint arXiv:2104.13714*, 2021.
- 835 [Conwell *et al.*, 2023] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. 836 Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pres- 837 sures shaping high-level visual representation in brains and machines? *bioRxiv*, 838 2023.
- 839 [Défosses *et al.*, 2022] Alexandre Défosses, Charlotte Caucheteux, Jérémy Rapin, Ori 840 Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings. 841 *arXiv preprint arXiv:2208.12266*, 2022.
- 842 [Deniz *et al.*, 2019] Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and 843 Jack L Gallant. The representation of semantic information across human cerebral 844 cortex during listening versus reading is invariant to stimulus modality. *Journal of* 845 *Neuroscience*, 39(39):7722–7736, 2019.
- 846 [Doerig *et al.*, 2022] Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, 847 Jenann Ismael, Grace Lindsay, Konrad Kording, Talia Konkle, Marcel AJ Van Ger- 848 ven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. 849 *arXiv preprint arXiv:2209.03718*, 2022.
- 850 [Dong and Toneva, 2023] Dota Tianai Dong and Mariya Toneva. Interpreting multi- 851 modal video transformers using brain recordings. In *ICLR 2023 Workshop on Multi-* 852 *modal Representation Learning: Perks and Pitfalls*, 2023.
- 853 [Du *et al.*, 2020] Changde Du, Changying Du, Lijie Huang, and Huiguang He. Con- 854 ditional generative neural decoding with structured cnn feature prediction. In *AAAI*, 855 pages 2629–2636, 2020.
- 856 [Dwivedi *et al.*, 2021] Kshitij Dwivedi, Michael F Bonner, Radoslaw Martin Cichy, 857 and Gemma Roig. Unveiling functions of the visual cortex using task-specific deep 858 neural networks. *PLoS computational biology*, 17(8):e1009267, 2021.
- 859 [Fang *et al.*, 2020] Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images 860 from brain activity by shape-semantic gan. *NeurIPS*, 33:13038–13048, 2020.
- 861 [Gauthier and Levy, 2019] Jon Gauthier and Roger Levy. Linking artificial and human 862 neural representations of language. *arXiv preprint arXiv:1910.01244*, 2019.
- 863 [Goldstein *et al.*, 2022] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, 864 Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon 865 Cohen, et al. Shared computational principles for language processing in humans 866 and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.
- 867 [Gwilliams *et al.*, 2022] Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkka- 868 nen, David Poeppel, and Jean-Rémi King. Meg-masc: a high-quality magneto- 869 encephalography dataset for evaluating natural speech processing. *arXiv preprint* 870 *arXiv:2208.11488*, 2022.
- 871 [Hale *et al.*, 2018] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 872 Finding syntax in human encephalography with beam search. In *ACL*, pages 2727– 873 2736, 2018.
- 874 [Handjaras *et al.*, 2016] Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, 875 Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, and Giovanna Marotta. How 876 concepts are encoded in the human brain: a modality independent, category-based 877 cortical organization of semantic knowledge. *Neuroimage*, 135:232–242, 2016.
- 878 [Hebart *et al.*, 2022] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam Rock- 879 ter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and 880 Chris I Baker. Things-data: A multimodal collection of large-scale datasets for in- 881 vestigating object representations in brain and behavior. *bioRxiv*, pages 2022–07, 882 2022.
- 883 [Hollenstein *et al.*, 2018] Nora Hollenstein, Jonathan Rotsztein, Marius Troendle, An- 884 dreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye- 885 tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.
- 886 [Hollenstein *et al.*, 2019] Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and 887 Ce Zhang. Cognival: A framework for cognitive word embedding evaluation. In 888 *CoNLL*, pages 538–549, 2019.
- [Horikawa and Kamitani, 2017] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic 889 decoding of seen and imagined objects using hierarchical visual features. *Nature* 890 *communications*, 8(1):1–15, 2017. 891
- [Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal 892 Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self- 893 supervised speech representation learning by masked prediction of hidden units. 894 *TASLP*, 29:3451–3460, 2021. 895
- [Huth *et al.*, 2016] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, 896 Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic 897 maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. 898
- [Huth *et al.*, 2022] Alexander G Huth, Shinji Nishimoto, An T Vu, Dupre la Tour T, 899 and Gallant JL. Gallant lab natural short clips 3t fmri data. *G-Node*, 2022. 900
- [Jain and Huth, 2018] Shailee Jain and Alexander G Huth. Incorporating context into 901 language encoding models for fmri. In *NIPS*, pages 6629–6638, 2018. 902
- [Jain *et al.*, 2020] Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S 903 Turek, and Alexander Huth. Interpretable multi-timescale models for predicting 904 fmri responses to continuous natural speech. *NeurIPS*, 33:13738–13749, 2020. 905
- [Jat *et al.*, 2020] S Jat, H Tang, P Talukdar, and T Mitchel. Relating simple sentence 906 representations in deep neural networks and the brain. In *ACL*, pages 5137–5154, 907 2020. 908
- [Just *et al.*, 2010] Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and 909 Tom M Mitchell. A neurosemantic theory of concrete noun representation based 910 on the underlying brain codes. *PLoS one*, 5(1):e8622, 2010. 911
- [Karamolegkou *et al.*, 2023] Antonia Karamolegkou, Mostafa Abdou, and Anders 912 Søgaard. Mapping brains with language models: A survey. *arXiv preprint* 913 *arXiv:2306.05126*, 2023. 914
- [Kauf *et al.*, 2023] Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and 915 Evelina Fedorenko. Lexical semantic content, not syntactic structure, is the main 916 contributor to ann-brain similarity of fmri responses in the language network. 917 *bioRxiv*, pages 2023–05, 2023. 918
- [Kay *et al.*, 2008] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gal- 919 lant. Identifying natural images from human brain activity. *Nature*, 452(7185):352– 920 355, 2008. 921
- [Khosla and Wehbe, 2022] Meenakshi Khosla and Leila Wehbe. High-level visual 922 areas act like domain-general filters with strong selectivity and functional specializa- 923 tion. *bioRxiv*, 2022. 924
- [Kubilius *et al.*, 2019] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajaling- 925 ham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, 926 Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow 927 recurrent ans. *NIPS*, 32:12805–12816, 2019. 928
- [Kumar *et al.*, 2022] Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel 929 Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, 930 and Samuel A Nastase. Reconstructing the cascade of language processing in 931 the brain using the internal computations of a transformer-based language model. 932 *BioRxiv*, pages 2022–06, 2022. 933
- [Lahner *et al.*, 2023] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika 934 Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, 935 SouYoung Jin, N Apurva Ratan Murty, et al. Bold moments: modeling short visual 936 events through a video fmri dataset and metadata. *bioRxiv*, pages 2023–03, 2023. 937
- [Li *et al.*, 2021] Jixing Li, Shohini Bhattachali, Shulin Zhang, Berta Franzluebbers, 938 Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe 939 Pallier, and John Hale. Le petit prince: A multilingual fmri corpus using ecological 940 stimuli. *Biorxiv*, pages 2021–10, 2021. 941
- [Lin *et al.*, 2022] Sikun Lin, Thomas Christopher Sprague, and Ambuj Singh. Mind 942 reader: Reconstructing complex images from brain activities. In *NeurIPS*, 2022. 943
- [Lu *et al.*, 2022] Haoyu Lu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, 944 Jingyuan Wen, Changde Du, Xin Zhao, Hao Sun, Huiguang He, et al. Multi- 945 modal foundation models are better simulators of the human brain. *arXiv preprint* 946 *arXiv:2208.08263*, 2022. 947
- [Malik-Moraleda *et al.*, 2022] Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, 948 Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina 949 Fedorenko. An investigation across 45 languages and 12 language families reveals 950 a universal language network. *Nature Neuroscience*, 25(8):1014–1019, 2022. 951
- [Merlin and Toneva, 2022] Gabriele Merlin and Mariya Toneva. Language models 952 and brain alignment: beyond word-level semantics and prediction. *arXiv preprint* 953 *arXiv:2212.00596*, 2022. 954
- [Millet *et al.*, 2022] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves 955 Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Rémi 956 King. Toward a realistic model of speech processing in the brain with self-supervised 957 learning. *arXiv:2206.01685*, 2022. 958

- [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 Zadbod, 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 Surampudi. 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 bccw-j-egg: 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, Umüt 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.

1099 [Tuckute *et al.*, 2022] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H
1100 McDermott. Many but not all deep neural network audio models capture brain re-
1101 sponses and exhibit hierarchical region correspondence. *bioRxiv*, 2022.

1102 [Tuckute *et al.*, 2023] Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro,
1103 Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving
1104 and suppressing the human language network using large language models. *bioRxiv*,
1105 2023.

1106 [Vaidya *et al.*, 2022] Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-
1107 supervised models of audio effectively explain human cortical responses to speech.
1108 *arXiv preprint arXiv:2205.14252*, 2022.

1109 [Wang *et al.*, 2017] Jing Wang, Vladimir L Cherkassky, and M Adam Just. Predicting
1110 the brain activation pattern associated with the propositional content of a sentence:
1111 Modeling neural representations of events and states. *HBM*, 10:4865–4881, 2017.

1112 [Wang *et al.*, 2019] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy:
1113 Inferring the similarity of task-derived representations from brain activity. *NeurIPS*,
1114 32:15501–15511, 2019.

1115 [Wang *et al.*, 2020] Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and
1116 Chengqing Zong. Fine-grained neural decoding with distributed word representa-
1117 tions. *Information Sciences*, 507:256–272, 2020.

1118 [Wang *et al.*, 2022] Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr,
1119 and Leila Wehbe. Incorporating natural language into vision models improves pre-
1120 diction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09, 2022.

1121 [Wehbe *et al.*, 2014] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aa-
1122 ditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain
1123 regions involved in different story reading subprocesses. *PLoS one*, 9(11):e112575,
1124 2014.

1125 [Yamins *et al.*, 2014] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A
1126 Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical
1127 models predict neural responses in higher visual cortex. *PNAS*, 111(23):8619–8624,
1128 2014.

1129 [Zhang *et al.*, 2020] Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu.
1130 Connecting concepts in the brain by mapping cortical representations of semantic
1131 relations. *Nature communications*, 11(1):1–13, 2020.

1132 [Zhang *et al.*, 2022a] Xiaohan Zhang, Shaonan Wang, Nan Lin, Jiajun Zhang, and
1133 Chengqing Zong. Probing word syntactic representations in the brain by a feature
1134 elimination method. *AAAI*, 2022.

1135 [Zhang *et al.*, 2022b] Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong.
1136 Is the brain mechanism for hierarchical structure building universal across lan-
1137 guages? an fmri study of chinese and english. In *Proceedings of the 2022 Con-
1138 ference on Empirical Methods in Natural Language Processing*, pages 7852–7861,
1139 2022.

1140 [Zinszer *et al.*, 2018] Benjamin D Zinszer, Laurie Bayet, Lauren L Emberson, Ra-
1141 jeev DS Raizada, and Richard N Aslin. Decoding semantic representations
1142 from functional near-infrared spectroscopy signals. *Neurophotonics*, 5(1):011003–
1143 011003, 2018.