Language Models and Brain Alignment: Brain Encoding and Decoding

Subba Reddy Oota[1], S. Bapi Raju[2]

[1]TU Berlin, Germany; [2]IIIT Hyderabad, India

Subba.reddy.oota<AT>tu-berlin.de; raju.bapi<AT>iiit.ac.in

Thanks to Dr. Manish Gupta, Microsoft India for the slides

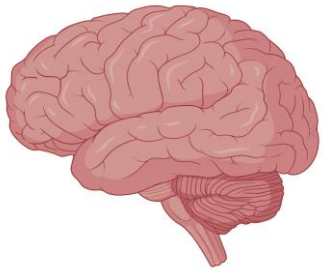
Agenda

- **09.00 AM – 10.30 AM Bapi Raju Neuro-AI alignment: Introduction**
- 10.30 AM – 11.00 AM Coffee Break
- 11.30 AM – 1.00 PM Subba Reddy Language and the Brain:
  DL for Brain Encoding and Decoding

Agenda

- Neuro-AI Alignment: Introduction
  - **Introduction to Brain Encoding & Decoding**
  - Types of Brain Recording & Popular Text Datasets
  - Types of Stimulus Representation
  - Methodology

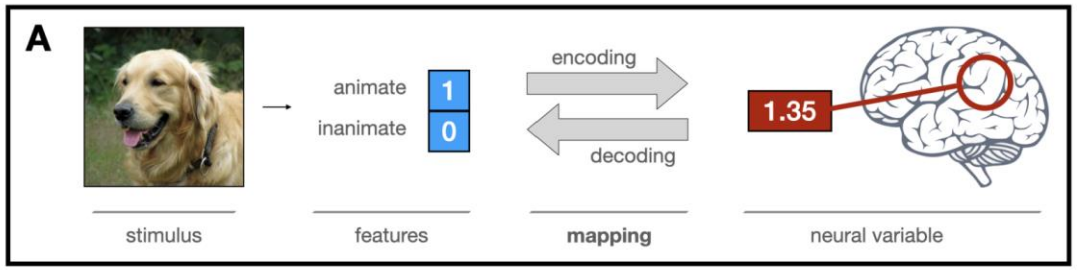
Neuroscience

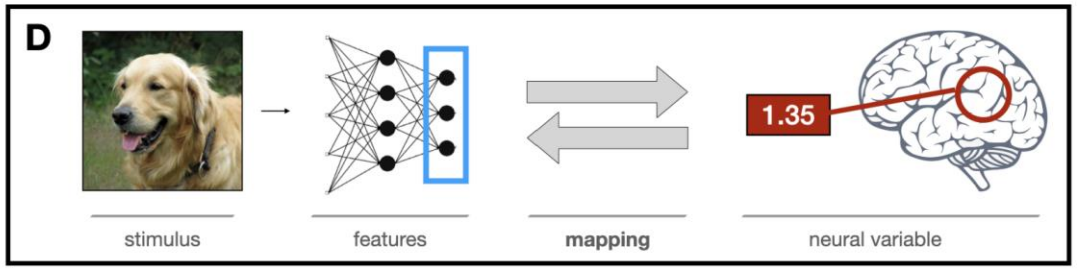


- Field of science that studies the structure and function of the nervous system of different species.
- Involves answering interesting questions
  - How learning occurs during adolescence, and how it differs from the way adults learn and form memories.
  - Which specific cells in the brain (and what connections they form with other cells), have a role in how memories are formed.
  - How do humans cancel out irrelevant information arriving from the senses and focus only on information that matters.
  - How do humans make decisions.
  - How humans develop speech and learn languages.
- Neuroscientists study diverse topics that help us understand how the brain and nervous system work.

NeuroAI: Brain encoding and decoding

- Encoding is the process of learning the mapping e from the stimuli S to the neural activation F.
  - Using feature engg (A) or deep learning (D)
- Decoding constitutes learning mapping d, which predicts stimuli S back from the brain activation F.
  - Oftentimes, we predict a stimulus representation R rather than actually reconstructing S.
- Other forms of encoding/decoding
  - (B): Map participants' behaviour to neural variables.
  - (C): Mapping between activity in different brain regions.





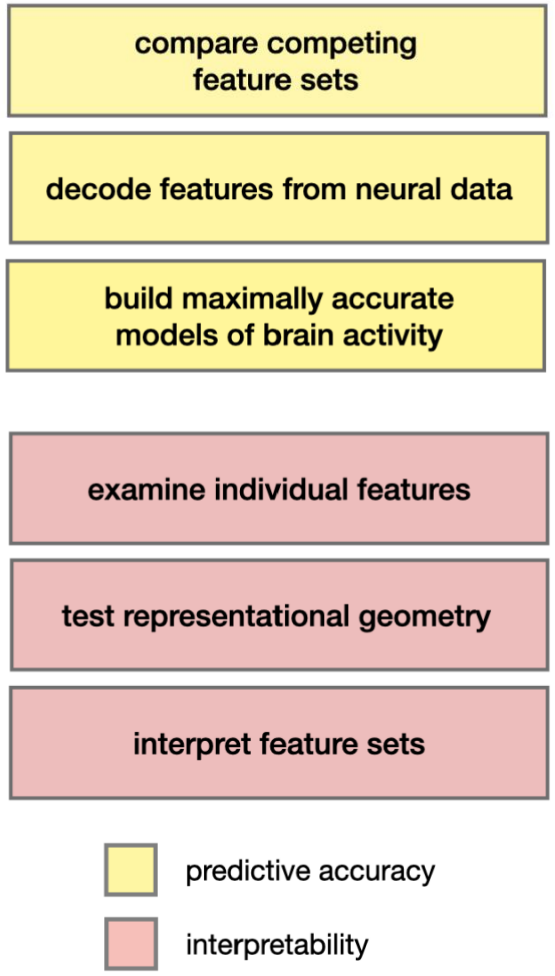
Ivanova, Anna A., Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. "Is it that simple? Linear mapping models in cognitive neuroscience." *bioRxiv* (2021).

Brain encoding and decoding

- For both encoding and decoding, the first step is to learn a stimulus representation R of the stimuli S at the **train time**.
- F is the brain response.
- 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.

Computational Cognitive Science Research goals

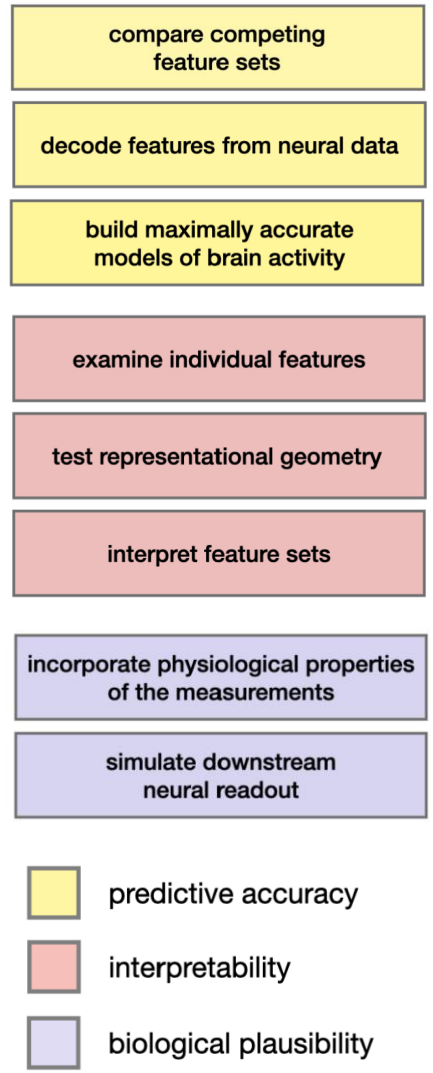
- Predictive Accuracy
  - Compare feature sets: Which feature set provides the most faithful reflection of the neural representational space?
  - Test feature decodability: “Does neural data Y contain information about features X?”
  - Build accurate models of brain data: Aim is to enable simulations of neuroscience experiments. (In-silico neuroscience)
- Interpretability
  - Examine individual features: Which features contribute the most to neural activity?
  - Test correspondences between representational spaces
    - “CNNs vs ventral visual stream” or “Two text representations”
  - 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?
  - How does the mapping relate to other models or theories of brain function?



Ivanova, Anna A., Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. "Is it that simple? Linear mapping models in cognitive neuroscience." *bioRxiv* (2021).

Computational Cognitive Science Research goals

- Biological plausibility
  - Simulate linear readout
    - If the features can be extracted with a linear mapping model, it means that they require few additional computations in order to be used downstream.
  - Incorporate measurement-related considerations
    - Rather than assuming a fixed HRF across voxels and/or conditions, what are better ways?



Ivanova, Anna A., Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. "Is it that simple? Linear mapping models in cognitive neuroscience." *bioRxiv* (2021).

Agenda

- Neuro-AI Alignment: Introduction
  - Introduction to Brain Encoding & Decoding
  - **Types of Brain Recording & Popular Text Datasets**
  - Types of Stimulus Representation
  - Methodology

DNN Model Representations

Stimulus

Language | Audio | Visual | Multimodal

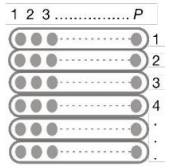
DNN Models

Language Model | Audio Model | Vision Model | Multimodal Model

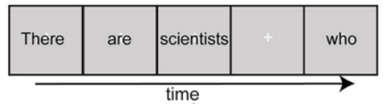
Stimulus Feature Space

Down sampling: Interpolation of the feature matrix

HRF estimation: use of FIR model, different delays



Human Brain Recordings



Reading

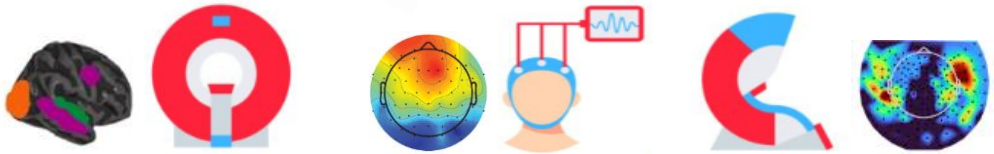
Listening

Static Image

Video clip (with or without audio)

Visual

Auditory



fMRI | EEG | MEG

Evaluation Metrics

PCC, R^2, 2V2 Accuracy, RDM, CKA, Noise Ceiling, Normalized brain alignment

fMRI: Whole brain, ROI level, Sub-ROI level, task-specific voxels

MEG: Sensor recordings over time points

EEG: Electrode signals recorded over time

Encoding Models → | ← **Decoding Models**

Linear

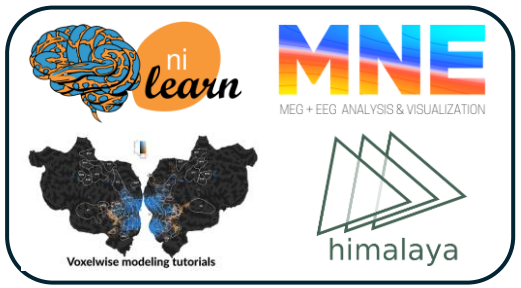
Ridge | Bootstrap Ridge | Banded Ridge

Lasso | PLS | Kernel Ridge

Non-Linear

Multi-Layer Perceptron | DNN Models

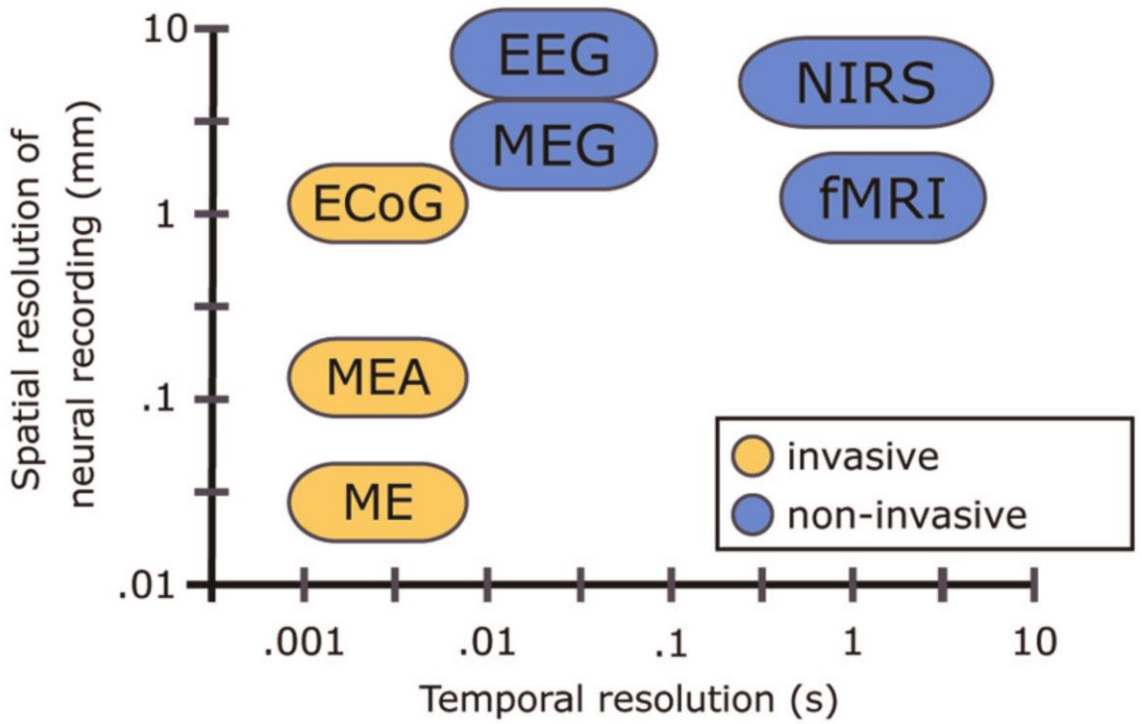
Visualization Tools



Oota et al (2025). Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey) [accepted TMLR]

https://openreview.net/pdf?id=YxKJihRcby

Techniques for studying the brain function

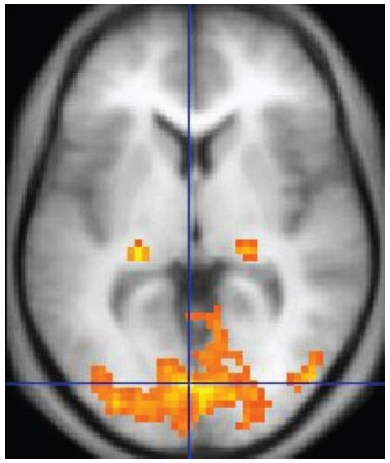


Single Micro-Electrode (ME), Micro-Electrode array (MEA), Electro-Cortico Graphy (ECoG), Positron emission tomography (PET), functional MRI (fMRI), Magneto-encephalography (MEG), Electro-encephalography (EEG), Near-Infrared Spectroscopy (NIRS)

- fMRI: high spatial but low time resolution.
  - Good to study a specific location in the brain
  - Unsuitable for sentence-level analysis. fMRI takes about two seconds to complete a scan. This is far lower than the speed at which humans can process language.
  - Cannot capture syntactic information (Gauthier and Levy, 2019)
- EEG: high time but low spatial resolution.
  - Can preserve rich syntactic information (Hale et al., 2018)
  - But cannot use for source analysis.
- fNIRS: compromise option
  - Time resolution better than fMRI
  - Spatial resolution better than EEG
  - Balance of spatial and temporal resolution may not be enough to compensate for the loss in both.

Vogel, Jörn, Sami Haddadin, Beata Jarosiewicz, John D. Simeral, Daniel Bacher, Leigh R. Hochberg, John P. Donoghue, and Patrick van der Smagt. "An assistive decision-and-control architecture for force-sensitive hand–arm systems driven by human–machine interfaces." *The International Journal of Robotics Research* 34, no. 6 (2015): 763-780.

fMRI



An fMRI image with yellow areas showing increased activity compared with a control condition

- No injections, surgery, the ingestion of substances, or exposure to ionizing radiation.
- The primary form of fMRI uses the blood-oxygen-level dependent (BOLD) contrast, discovered by Seiji Ogawa in 1990.
  - Measures brain activity by detecting changes associated with blood flow.
  - When an area of the brain is in use, blood flow to that region also increases.
- Hemodynamic response (HRF)
  - It takes a while for the vascular system to respond to the brain's need for glucose.
  - Blood flow lags the neuronal events triggering it by about 5 seconds.

Forms of stimulus presentation and data collection

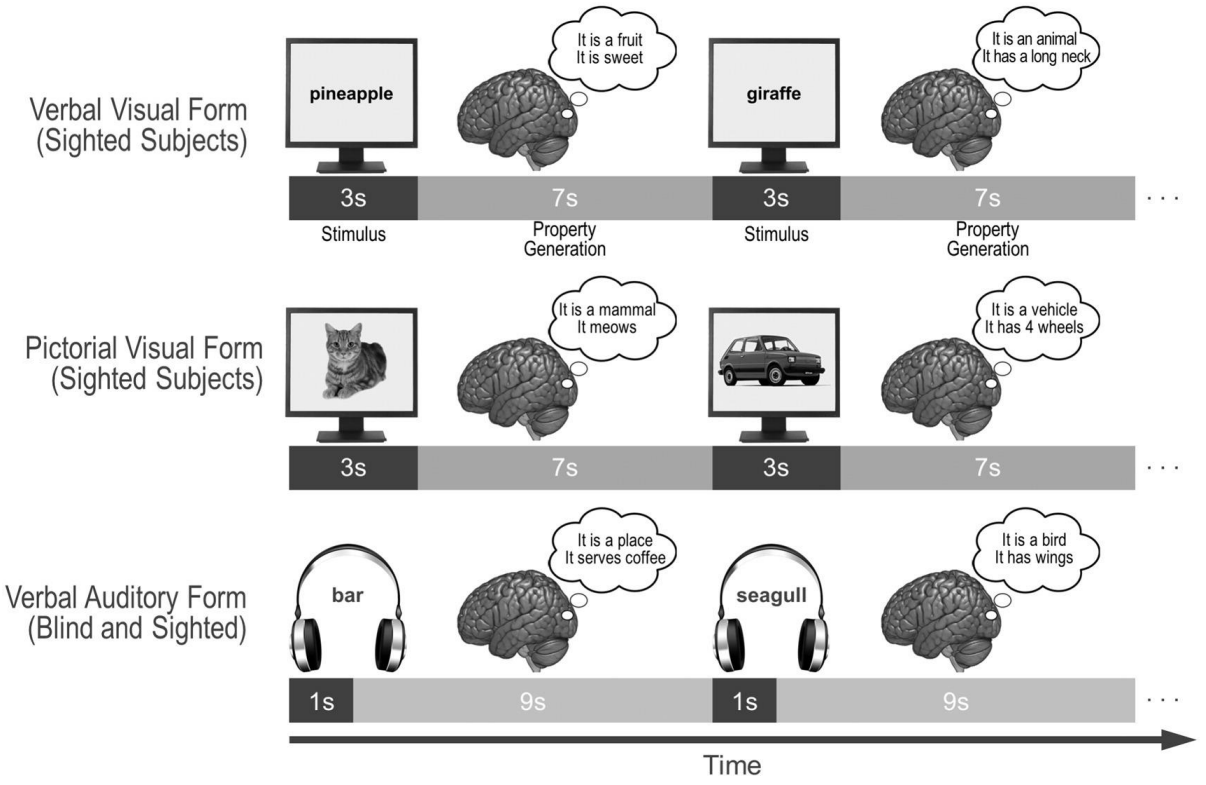
- Type: fMRI, EEG, MEG, …
- TR: Sampling time.
- Fixation points: location, color, shape.
- Form of stimuli presentation: text, video, audio, images.
- Task: question answering, property generation, understanding, …
- Time given to participants: 1 minute to list properties, …
- Type of participants: males/females, sighted/blind, …
- Number of times the response to stimuli was recorded.
- Language

Text Stimulus Datasets

Dataset	Type	Language	Stimulus	#Subjects	Paradigm	Size	Task
Wehbe et al., 2014	fMRI	English	Chapter 9 of *Harry Potter and the Sorcerer's Stone*	9	Reading stories	5000 word chapter was presented in 45 minutes.	Story understanding
Handjaras et al., 2016	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns	20	Reading, viewing or listening	40 nouns * 4 times.	Property Generation
Anderson et al., 2017	fMRI	Italian	70 concrete and abstract nouns from law/music.	7	Reading	70 nouns * 5 times.	Imagine a situation that they personally associate with the noun
Zurich Cognitive Language Processing Corpus (ZuCo): Hollenstein et al., 2018	EEG and eye-tracking	English	Sentences from movie reviews or Wikipedia	12	Reading natural sentences	21,629 words in 1107 sentences and 154,173 fixations	Rate movie quality, answer control questions, check for existence of a relation
Anderson et al., 2019	fMRI	English	240 active voice sentences describing everyday situations	14	Reading	240 sentences seen 12 times (by 10 subjects) and 6 times (by 4 subjects)	Passive reading
BCCWJ-EEG: Oseki and Asahara, 2020	EEG	Japanese	20 newspaper articles	40	Reading	1 time reading for ~30-40 minutes	Passive reading

Data for concrete nouns from sighted/blind subjects

- Participants were asked to verbally enumerate in one minute the properties (features) that describe the entities the words refer to.
- 4 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
  - 5 congenitally blind with a verbal auditory form.



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

70 - Italian word stimuli fMRI data

- Taxonomic categories in law and music domain
  - Ur-abstract: that are classified as abstract in WordNet
  - Attribute: A construct whereby objects or individuals can be distinguished
  - Communication: Something that is communicated by, to or between groups
  - Event/action: Something that happens at a given place and time
  - Person/Social role: Individual, someone, somebody, mortal
  - Location: Points or extents in space
  - Object/Tool: A class of unambiguously concrete nouns

	LAW		MUSIC	
Ur-abstracts	giustizia	justice	musica	music
	liberta'	liberty	blues	blues
	legge	law	jazz	jazz
	corruzione	corruption	canto	singing
	refurtiva	**loot**	punk	punk
Attribute	giurisdizione	jurisdiction	sonorita'	sonority
	cittadinanza	citizenship	ritmo	rhythm
	impunita'	impunity	~~melodia~~	~~melody~~
	legalita'	legality	tonality'	tonality
	illegalita	illegality	intonazione	pitch
Communication	divieto	prohibition	canzone	song
	verdetto	verdict	**pentagramma**	**stave**
	ordinanza	decree	ballata	ballad
	addebito	accusation	ritornello	refrain
	ingiunzione	injunction	sinfonia	symphony
Event/action	arresto	arrest	**concerto**	**concert**
	processo	trial	recital	recital
	reato	crime	assolo	solo
	furto	theft	**festival**	**festival**
	assoluzione	acquital	**spettacolo**	**show**
Person/Social-role	**giudice**	**judge**	**musicista**	**musician**
	ladro	**thief**	**cantante**	**singer**
	imputato	**defendant**	**compositore**	**composer**
	testimone	**witness**	**chitarrista**	**guitarist**
	avvocato	**lawyer**	**tenore**	**tenor**
Location	**tribunale**	**court/tribunal**	**palco**	**stage**
	carcere	**prison**	**auditorium**	**auditorium**
	~~**questura**~~	~~**police-station**~~	**discoteca**	**disco**
	penitenziario	**penitentiary**	**conservatorio**	**conservatory**
	patibolo	**gallows**	**teatro**	**theatre**
Object/Tool	**manette**	**handcuffs**	**violino**	**violin**
	toga	**robe**	**tamburo**	**drum**
	manganello	**truncheon**	**tromba**	**trumpet**
	cappio	**noose**	**metronomo**	**metronome**
	~~**grimaldello**~~	~~**skeleton-key**~~	**radio**	**radio**

Participants asked to imagine a situation that they personally associate with the noun

Anderson, Andrew J., Douwe Kiela, Stephen Clark, and Massimo Poesio. "Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns." *Transactions of the Association for Computational Linguistics* 5 (2017): 17-30.

Zurich Cognitive Language Processing Corpus (ZuCo)

	Task 1 Normal reading (Sentiment)	Task 2 Normal reading (Wikipedia)	Task 3 Task-specific reading (Wikipedia)
Material	Positive, negative or neutral sentences from movie reviews	Wikipedia sentences containing specific relations	Wikipedia sentences containing specific relations
Example	*"The film often achieves a mesmerizing poetry."* (positive)	*"Talia Shire (born April 25, 1946) is an American actress of Italian descent."* (relations: *nationality, job title*)	*"Lincoln was the first Republican president."* (relation: *political affiliation*)
Task	Read the sentences, rating the quality of the movie based on the sentence read	Read the sentences, answer control questions	Mark whether a specific relation occurs in the given sentence or not
Control question	*"Based on the previous sentence, how would you rate this movie from 1 (very bad) to 5 (very good)?"*	*"Talia Shire was a ...1) singer 2) actress 3) director"*	*"Does this sentence contain the* political affiliation *relation? 1) Yes 2) No"*

- Personal reading speed.
  - Sentences were presented to the subjects in a naturalistic reading scenario
  - 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.
  - EEG and eye tracking data acquired.

Hollenstein, Nora, 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, no. 1 (2018): 1-13.

Agenda

- Neuro-AI Alignment: Introduction
  - Introduction to Brain Encoding & Decoding
  - Types of Brain Recording & Popular Text Datasets
  - **Types of Stimulus Representation**
  - Methodology

DNN Model Representations

Stimulus

Language | Audio | Visual | Multimodal

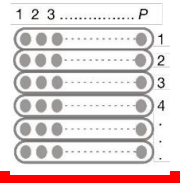
DNN Models

Language Model | Audio Model | Vision Model | Multimodal Model

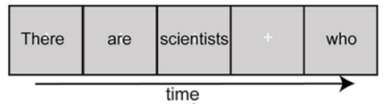
Stimulus Feature Space

Down sampling: Interpolation of the feature matrix

HRF estimation: use of FIR model, different delays



Human Brain Recordings



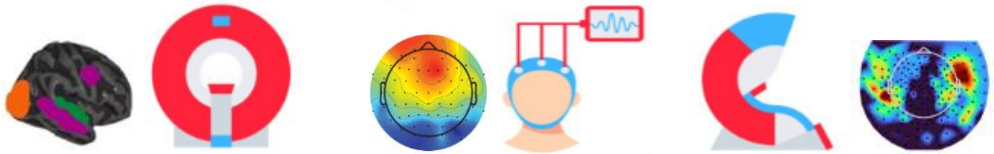
Reading

Listening

Static Image

Video clip (with or without audio)

Visual | Auditory



fMRI | EEG | MEG

Evaluation Metrics

PCC, R^2, 2V2 Accuracy, RDM, CKA, Noise Ceiling, Normalized brain alignment

fMRI: Whole brain, ROI level, Sub-ROI level, task-specific voxels

MEG: Sensor recordings over time points

EEG: Electrode signals recorded over time

Encoding Models → | ← **Decoding Models**

Linear

Ridge | Bootstrap Ridge | Banded Ridge

Lasso | PLS | Kernel Ridge

Non-Linear

Multi-Layer Perceptron | DNN Models

Visualization Tools

nilearn | MNE (MEG + EEG ANALYSIS & VISUALIZATION) | Voxelwise modeling tutorials | himalaya

Oota et al (2025). Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey) [accepted TMLR]

https://openreview.net/pdf?id=YxKJihRcby

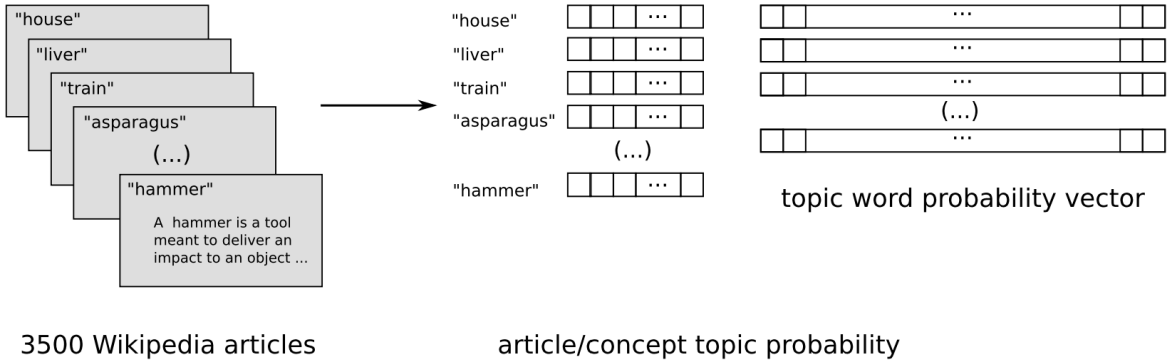
Text Stimulus Representations

- Basic NLP Representations
  - Corpus co-occurrence counts
  - Topic models
  - Linguistic: POS, dependencies, roles.
- Discourse
  - Characters, motion, speech, emotions, non-motion verbs
- Deep Learning based Representations
  - Embeddings
  - Longer context using LSTMs
  - Transformers
- Experiential attributes
  - Rated on 0-6 scale
  - Binary

Basic NLP Representations for Word Stimuli

- Corpus co-occurrence counts
  - 25 verbs (Mitchell et al., 2008; Pereira et al., 2013)
    - Verbs: see, hear, listen, taste, smell, eat, touch, nib, lift, manipulate, run, push, fill, move, ride, say, fear, open, approach, near, enter, drive, wear, break, and clean.
    - These verbs generally correspond to basic sensory and motor activities, actions performed on objects, and actions involving changes to spatial relationships.
    - For each (verb, stimulus word *w*), feature value = normalized co-occurrence count of *w* with any of three forms of the verb (e.g., taste, tastes, or tasted) over the text corpus.
  - 985 common English words (such as above, worry, and mother) in (Huth et al., 2016).

- Topic models (Pereira et al., 2013)
  - Get relevant Wiki pages (e.g., "airplane" is "Fixed-Wing Aircraft") and other linked pages (e.g. "Aircraft cabin")
  - LDA topic modelling on 3500 pages with #topics from 10 to 100, in increments of 5, setting the α parameter to 25/#topics.
  - LSA topic modelling (Wang et al., 2017)



Basic NLP Representations for Word Stimuli

- Word length
- Is the word related to one of the 28 unique parts of speech and 17 unique dependency relationships?
- Position of word in the sentence
- Roles
  - Main verb
  - Agent or experiencer
  - Patient or recipient
  - Predicate of a sentence (The window was dusty)
  - Modifier (The angry activist broke the chair)
  - Complement in adjunct and propositional phrase, including direction, location, and time (The restaurant was loud at night).

Wehbe, Leila, 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, no. 11 (2014): e112575.

Wang, Jing, Vladimir L. Cherkassky, and Marcel Adam Just. "Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states." *Human brain mapping* 38, no. 10 (2017): 4865-4881.

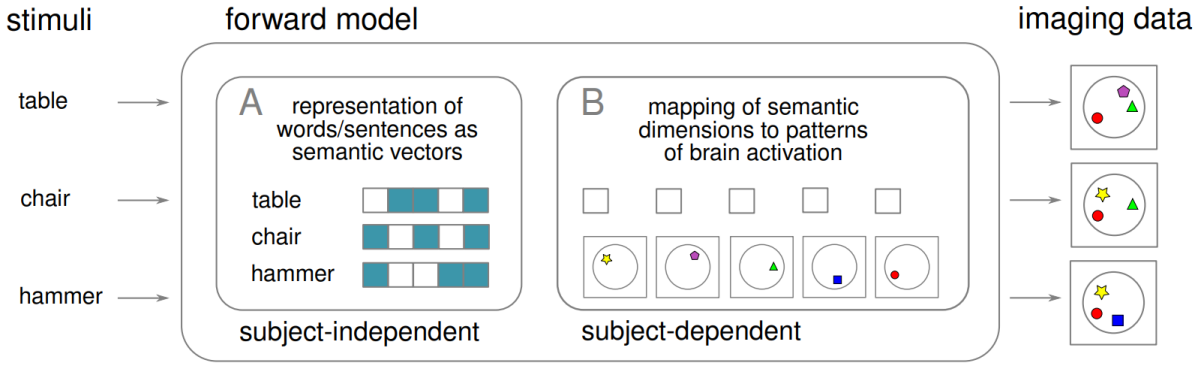
Discourse features (for Harry Potter dataset)

- Characters: Resolve all pronouns to the character to whom they refer, and make binary features to signal which of the 10 characters are mentioned.
- Motions: Identify a set of motions that occurred frequently in the chapter (e.g. fly, manipulate, collide physically, etc.).
- Speech: Indicate the parts of the story that correspond to direct speech between the characters. Used the presence of dialog as a feature.
- Emotions: Identified a set of emotions that were felt by the characters in the chapter (e.g. annoyance, nervousness, pride, etc.).
- Verbs: Identified a set of actions that occurred frequently in the chapter that were distinct from motion (e.g. hear, know, see, etc.).

Wehbe, Leila, 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, no. 11 (2014): e112575.

Wang, Jing, Vladimir L. Cherkassky, and Marcel Adam Just. "Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states." *Human brain mapping* 38, no. 10 (2017): 4865-4881.

DL Representations: Using embeddings for word stimuli

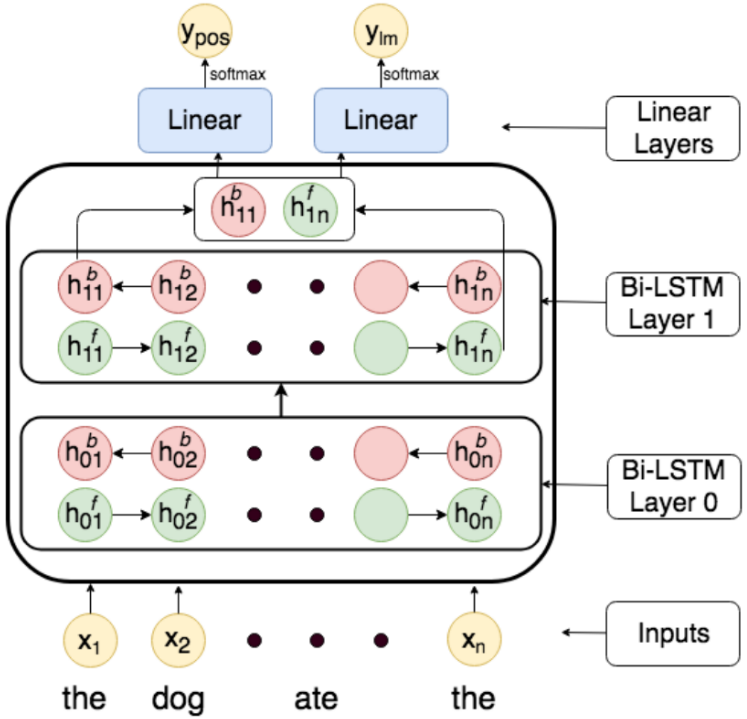


	Noun	Verb	Adjective
GloVe	**0.8768**(0.0792)	0.8544(0.0713)	0.8337(0.1081)
Word2Vec	**0.8386**(0.0942)	0.8309(0.0636)	0.8210(0.1028)
Fasttext	**0.8407**(0.0676)	0.8235(0.0766)	0.8077(0.0996)
RWSGwn	**0.8123**(0.0886)	0.7453(0.0771)	0.7425(0.1032)
ELMo	**0.9088**(0.0632)	0.8520(0.0797)	0.7993(0.1244)
ConceptNet	0.8646(0.0875)	**0.8702**(0.0695)	0.8249(0.0925)
Dependency	**0.8554**(0.0731)	0.8137(0.0755)	0.7891(0.0808)

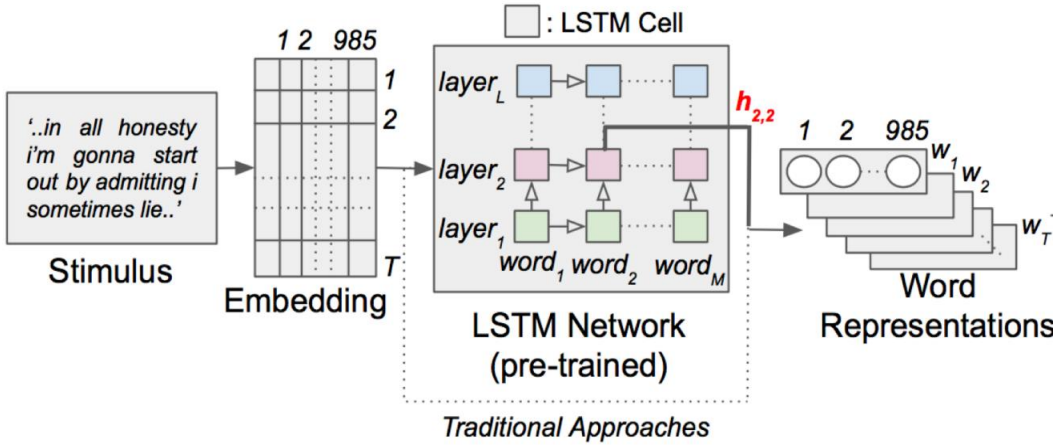
- GloVe 300D vectors (Pereira et al., 2016; Wang et al., 2017; Pereira et al., 2018; Anderson et al., 2019)
- 1000D Non-negative sparse embeddings (Wehbe et al., 2014).
- 300D embeddings by training a skip-gram model using negative sampling (SGNS) on Italian and English Wikipedia dumps using Gensim. (Anderson et al., 2017a)
- FastText (Berezutskaya et al., 2020)
- Comparison across multiple embedding methods
  - GloVe, word2vec, WordNet2Vec, FastText, ELMo (Hollenstein et al., 2019)
  - word2Vec, fastText, GloVe, Dependency-based word2vec, RWSGwn, ConceptNet, ELMo, averaged and concatenated combinations (Wang et al., 2020)

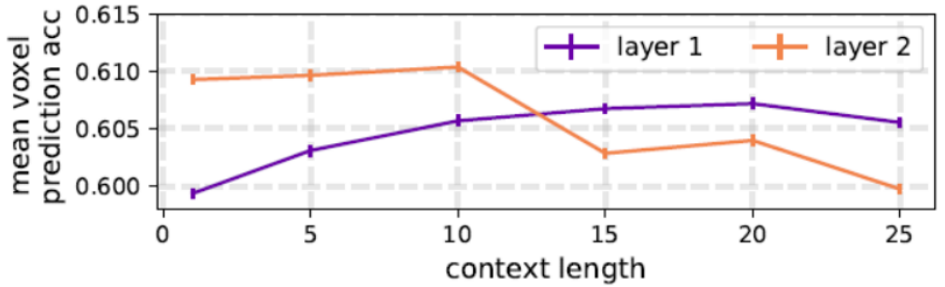
DL Representations: Using longer context for word stimuli

- Multi-task LSTMs
  - Predict next word and POS of next word.



- ELMo embeddings: LSTM based pretrained language model





(a) ELMo

Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in Neural Information Processing Systems 32* (2019).

Jain, Shailee, and Alexander Huth. "Incorporating context into language encoding models for fMRI." *Advances in neural information processing systems* 31 (2018).

Jat, Sharmistha, Hao Tang, Partha Talukdar, and Tom Mitchell. "Relating simple sentence representations in deep neural networks and the brain." *arXiv preprint arXiv:1906.11861* (2019).

DL Representations: Using sentence embeddings

- Unstructured Models: Ignore sentence structure
  - Simple Pooling Methods
    - Average/max/concat(max, avg) pooling over word embeddings.
  - Advanced Pooling Methods
    - FastSent (Hill, Cho, and Korhonen 2016) sums word embeddings in a sentence as its representation to predict the surrounding sentences.
    - SIF (Arora, Liang, and Ma 2016) adapts the naïve averaging of word embeddings to weighted averaging.
- Structured Models
  - Unsupervised Methods: Skip-thought, QuickThought.
  - Supervised Methods: InferSent, GenSen (Subramanian et al. 2018), Universal Sentence Encoder

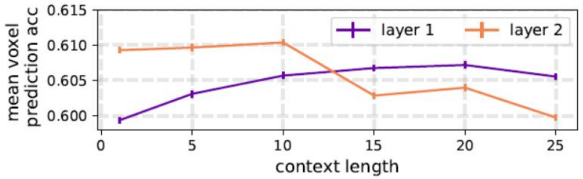
Topic	Passage	Sentence [b]
Musical Instruments	Piano	1. The piano is a popular musical instrument… 2. Pressing a piano key causes a felt-tipped hammer... 3. The piano has an enormous note range.
	Accordion	1. A clarinet is a woodwind musical instrument… 2. It is a long black tube with a flare at the bottom 3. The player chooses notes by pressing keys and holes.
	Clarinet	1. An accordion is a portable musical instrument. 2. One keyboard is used for individual notes 3. Accordions produce sound with bellow that blow air
…	…	…

[a]	Ridge			Lasso			MLP		
	topic	passa.	sente.	topic	passa.	sente.	topic	passa.	sente.
Max	0.88	0.76	0.65	0.88	0.75	0.70	0.83	0.70	0.63
Avg	0.90	0.83	0.73	0.92	0.81	0.78	0.89	0.78	0.67
Cat	0.92	0.83	0.74	0.90	0.81	0.80	0.86	0.74	0.66
Sif	0.89	0.84	0.69	0.91	0.77	0.72	0.84	0.73	0.65
Fast	0.92	0.81	0.74	0.90	0.79	0.77	0.88	0.76	0.67
Skip	0.90	0.82	0.75	0.91	0.80	0.79	0.86	0.81	0.73
Quik	0.91	0.84	0.75	0.91	0.81	0.79	0.90	0.82	0.77
Gen	0.91	0.84	0.78	0.92	0.84	**0.84**	0.91	**0.84**	**0.80**
Inf	**0.94**	**0.90**	**0.83**	**0.93**	**0.86**	**0.84**	**0.92**	**0.84**	0.79

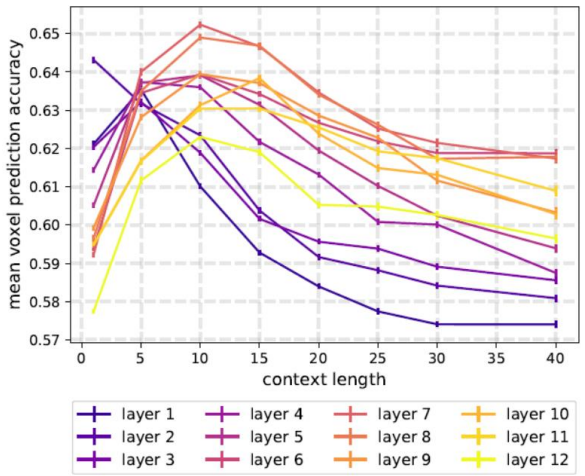
Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in Neural Information Processing Systems* 32 (2019).

Sun, Jingyuan, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. "Towards sentence-level brain decoding with distributed representations." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7047-7054. 2019.

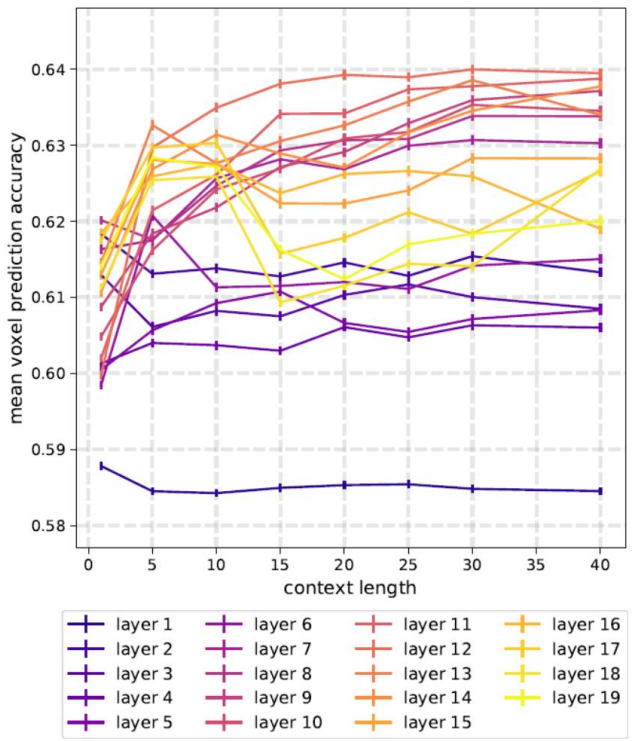
DL Representations: Transformer-based methods for text stimuli (Layer #, context length, architecture)



(a) ELMo

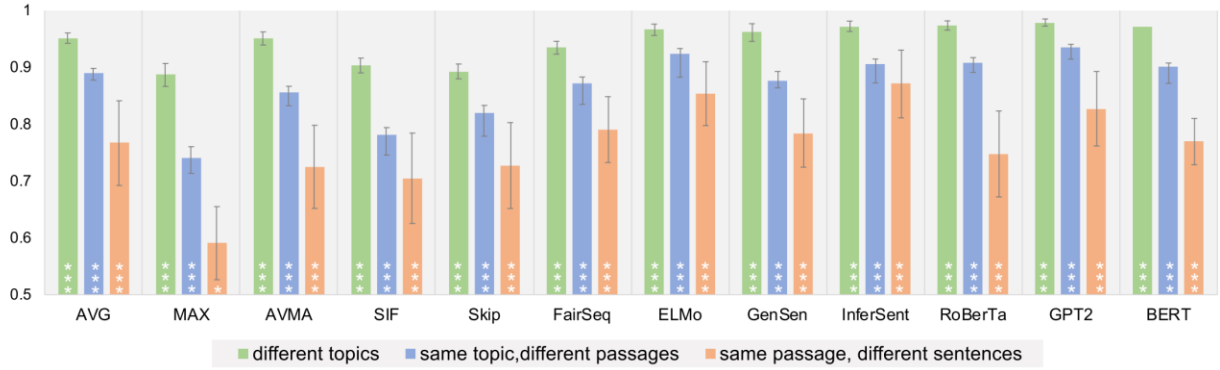


(b) BERT



(c) T-XL

DSM	Name	Structure and Training Task
Unstructured	AVG	Average Pooling
	MAX	Max Pooling
	AVMA	Concatenation of AVG and Max
	SIF	Weighted Average Pooling
Structured	FairSeq	CNN (language model)
	Skip	LSTM (language model)
	GenSen	BiLSTM (multi-task learning)
	InferSent	CNN-BiLSTM (natural language inference)
	ELMo	CNN-BiLSTM (language model)
	BERT	Transformer (language model)
	RoBerTa	
	GPT2	



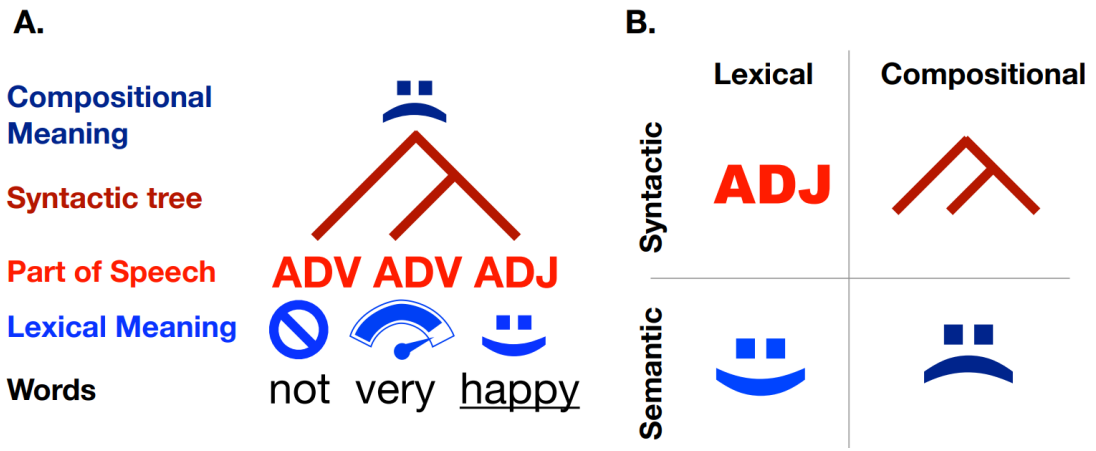
Transformer-XL is the only model that continues to increase performance as the context length is increased. In all networks, the middle layers perform the best for contexts longer than 15 words. The deepest layers across all networks show a sharp increase in performance at short-range context (fewer than 10 words), followed by a decrease in performance. [Toneva and Wehbe, 2019]

Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in Neural Information Processing Systems* 32 (2019).

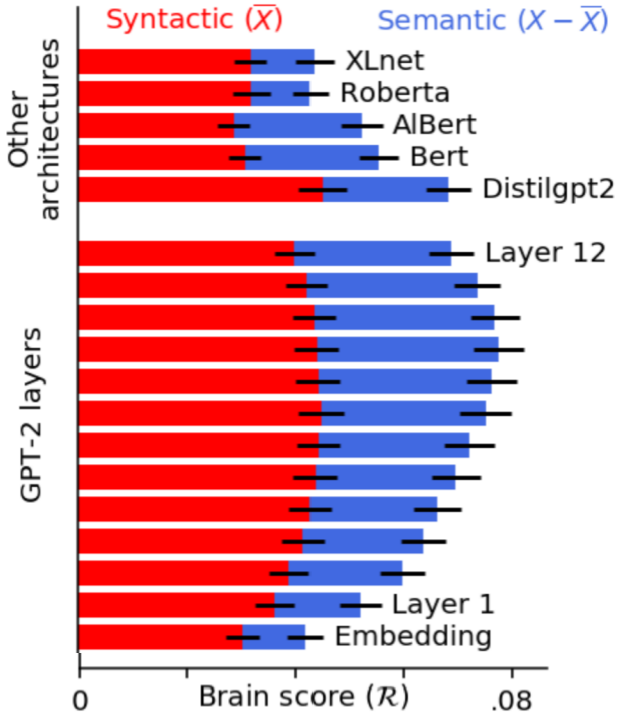
Sun, Jingyuan, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. "Neural encoding and decoding with distributed sentence representations." *IEEE Transactions on Neural Networks and Learning Systems* 32, no. 2 (2020): 589-603.

DL Representations: Comparing Transformers and extracting syntax vs semantics

- Representations:
  - Lexical: representation that is context-invariant. E.g., word embeddings.
  - Compositional: "contextualized" representation generated by a system combining multiples words. E.g., parse trees
  - Syntax: representation associated with the structure of sentences independently of their meaning
  - Semantics: representation of a language system that are not syntactic.



- If $X^{(l)}$ is activation of l^{th} layer, $\overline{X^{(l)}}$ is average activation across similar syntax inputs
  - Lexical: $X^{(0)}$
  - Compositional: $X^{(l)}; l > 0$
  - Syntax: $\overline{X^{(l)}}, l \geq 0$
  - Semantic: $X^{(l)} - \overline{X^{(l)}}$



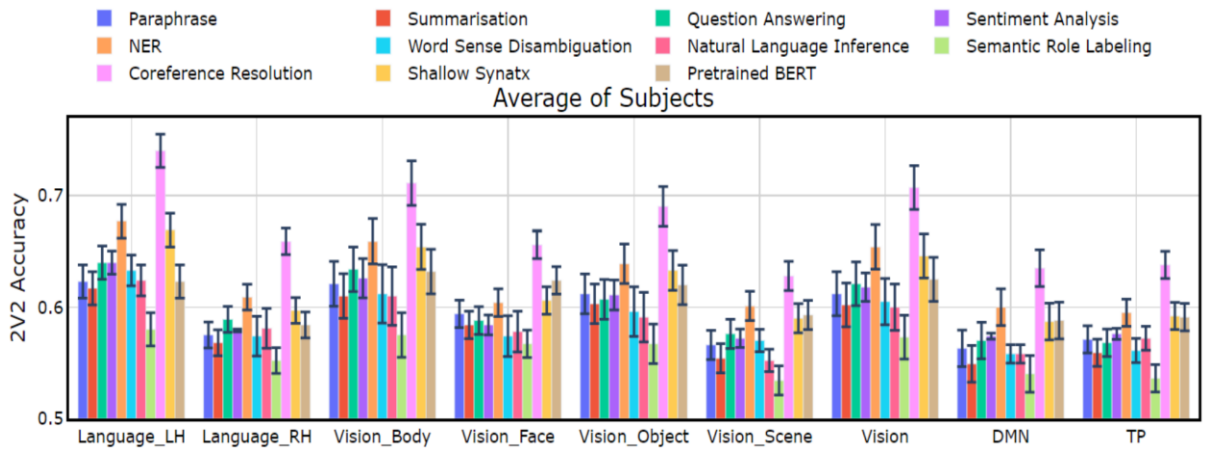
Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Remi King. "Disentangling syntax and semantics in the brain with deep networks." In *International Conference on Machine Learning*, pp. 1336-1348. PMLR, 2021.

DL Representations: Transformer-based methods for text stimuli (NLP task finetuning)

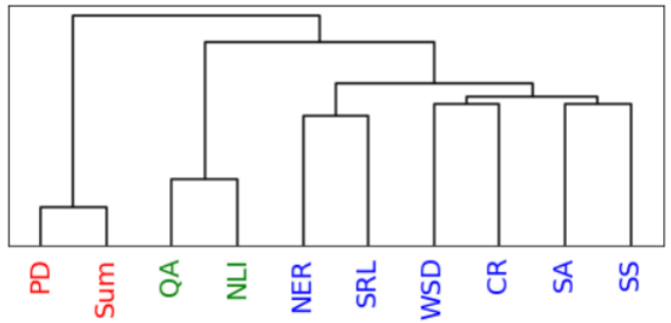
Task	HuggingFace Model Name	Dataset
NLI	bert-base-nli-mean-tokens	Stanford Natural Language Inference (SNLI), MultiNLI
PD	bert-base-cased-finetuned-mrpc	Microsoft Research Paraphrase Corpus (MRPC)
SS	bert-base-chunl	CoNLL-2003
Sum	bart-base-samsum	SAMSum
WSD	bert-base-baseline	English all-words
CR	bert_coreference_base	OntoNotes and GAP
NER	bert-base-NER	CoNLL-2003
QA	bert-base-qa	SQUAD
SA	bert-base-sst	Stanford Sentiment Treebank (SST)
SRL	bert-base-srl	English PropBank SRL

Tasks

Paraphrase, Summarization, Question Answering, Sentiment Analysis, NER, Word Sense Disambiguation, Natural Language Inference, Semantic Role Labeling, Coreference Resolution, Shallow Syntax Parsing



Pereira dataset: CR, NER, and SS perform the best.



Dendrogram constructed using similarity on representations from task-specific Transformer encoder models with stimuli from the dataset passed as input.

Oota, Subba Reddy, 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).

Experiential attributes model for text stimuli

- Represents words in terms of human (Amazon Mechanical Turk) ratings of their degree of association with different attributes of experience
  - "On a scale of 0 to 6, to what degree do you think of a banana as having a characteristic or defining color?"
  - Anderson et al., 2019: 65 attributes spanning sensory, motor, affective, spatial, temporal, causal, social, and abstract cognitive experiences.
- Value-add on top of text models: a lot of experiential information goes unstated in natural verbal communication.
  - E.g., it is rarely useful to communicate the color of bananas because it is obvious to all those with experience of bananas.
  - E.g., it would be unusual to specify that dropping things involves movement.
- Nishida et al., 2020 use a subset of 20 attributes.

Table 1 List of attributes first arranged by modality, and then subdivided into individual attributes

Dominant modality	Attribute
Vision	vision, bright, dark, color, pattern, large, small, motion, biomotion, fast, slow, shape, complexity, face, body.
Auditory	audition, loud, low, high, sound, music, speech.
Somatosensory	touch, temperature, texture, weight, pain.
Gustatory +Smell	taste, smell.
Motor	head, upper limb, lower limb, practice.
Attention	attention, arousal.
Event	duration, long, short, caused, consequential, social, time.
Evaluation	benefit, harm, pleasant, unpleasant.
Cognition	human, communication, self, cognition, number.
Emotion	happy, sad, angry, disgusted, fearful, surprised.
Drive	drive, needs.
Spatial	landmark, path, scene, near, toward, away.

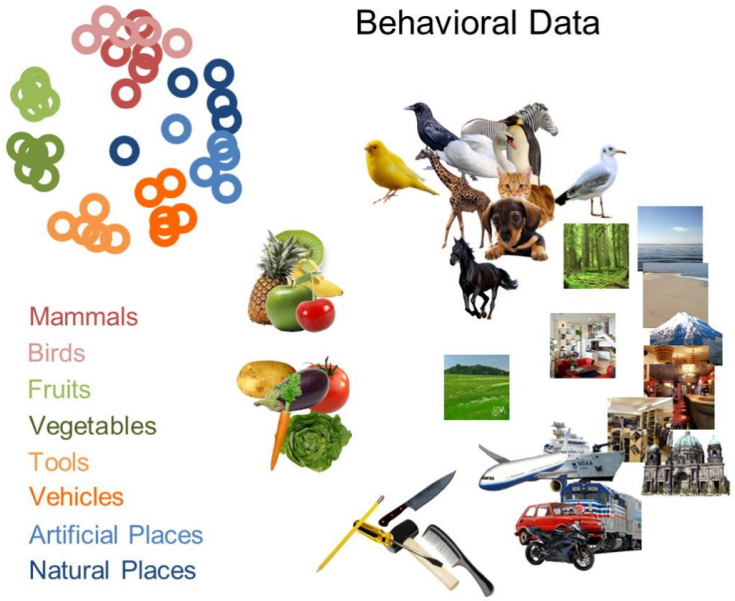
Anderson, Andrew James, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev DS Raizada, Feng Lin, and Edmund C. Lalor. "An integrated neural decoder of linguistic and experiential meaning." *Journal of Neuroscience* 39, no. 45 (2019): 8969-8987.

Anderson, Andrew James, 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, no. 9 (2017): 4379-4395.

Anderson, Andrew James, 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, no. 1 (2020): 1-14.

Binary attribute representations

- Each stimulus is represented using a binary vector capturing membership to one of the eight semantic categories.



- 42 neurally plausible semantic features (NPSFs)
  - Perceptual and affective characteristics of an entity (10 NPSFs coded such features, such as man-made, size, color, temperature, positive affective valence, high affective arousal), animate beings (person, human-group, animal), and time and space properties (e.g. unenclosed setting, change of location)

Word	NPSF features
Interview	Social, Mental action, Knowledge, Communication, Abstraction
Walk	Physical action, Change of location
Hurricane	Event, Change of physical state, Health, Natural, Negative affective valence, High affective arousal
Cellphone	Social action, Communication, Man-made, Inanimate
Judge	Social norms, Knowledge, Communication, Person
Clever	Attribute, Mental action, Knowledge, Positive affective valence, Abstraction

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

Wang, Jing, Vladimir L. Cherkassky, and Marcel Adam Just. "Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states." *Human brain mapping* 38, no. 10 (2017): 4865-4881.

Agenda

- Neuro-AI Alignment: Introduction
  - Introduction to Brain Encoding & Decoding
  - Types of Brain Recording & Popular Text Datasets
  - Types of Stimulus Representation
  - **Methodology**

DNN Model Representations

Stimulus

Language | Audio | Visual | Multimodal

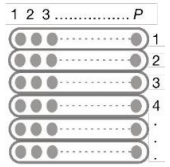
DNN Models

Language Model | Audio Model | Vision Model | Multimodal Model

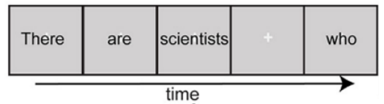
Stimulus Feature Space

Down sampling: Interpolation of the feature matrix

HRF estimation: use of FIR model, different delays



Human Brain Recordings



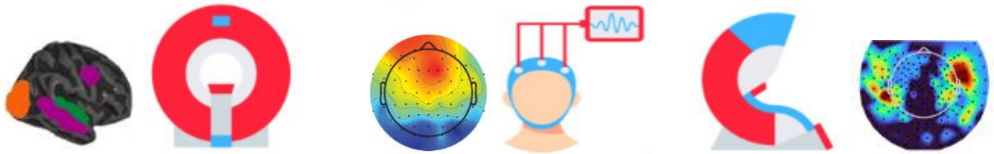
Reading

Listening

Static Image

Video clip (with or without audio)

Visual | Auditory



fMRI | EEG | MEG

Evaluation Metrics

PCC, R^2, 2V2 Accuracy, RDM, CKA, Noise Ceiling, Normalized brain alignment

fMRI: Whole brain, ROI level, Sub-ROI level, task-specific voxels

MEG: Sensor recordings over time points

EEG: Electrode signals recorded over time

Encoding Models → | ← **Decoding Models**

Linear

Ridge | Bootstrap Ridge | Banded Ridge

Lasso | PLS | Kernel Ridge

Non-Linear

Multi-Layer Perceptron | DNN Models

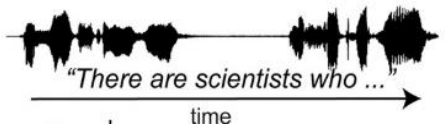
Visualization Tools

nilearn | MNE (MEG + EEG ANALYSIS & VISUALIZATION) | Voxelwise modeling tutorials | himalaya

Oota et al (2025). Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey) [accepted TMLR]

https://openreview.net/pdf?id=YxKJihRcby

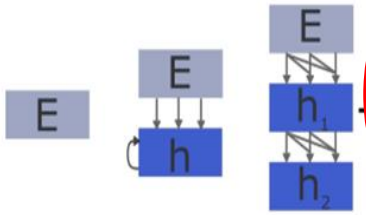
Encoding schema



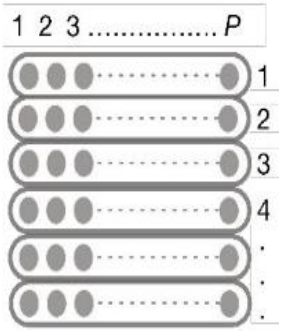
Stimulus

input

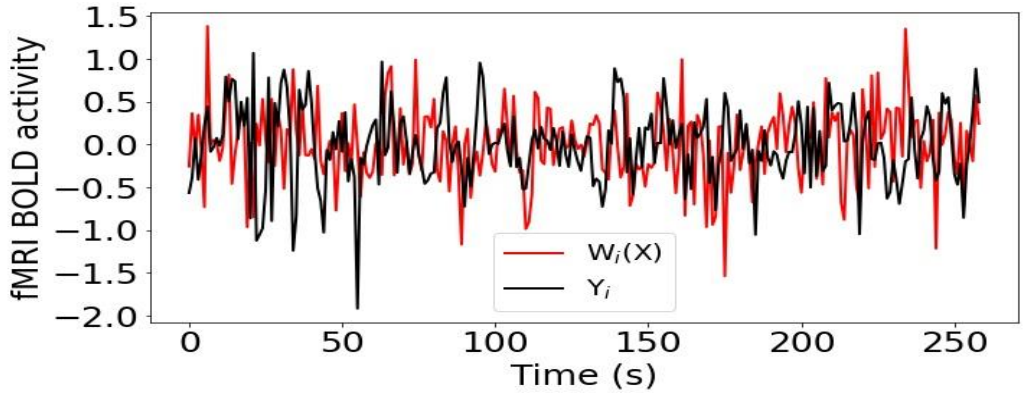
Models



Internal Neural Representation
Hidden unit activations



Pearson Correlation (R) = Corr(Y, W(X))



X

Input

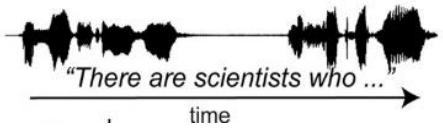
W

Ridge Regression

Y

Output

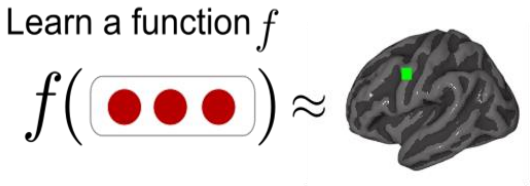
Internal Neural Representations
fMRI



Stimulus

Present

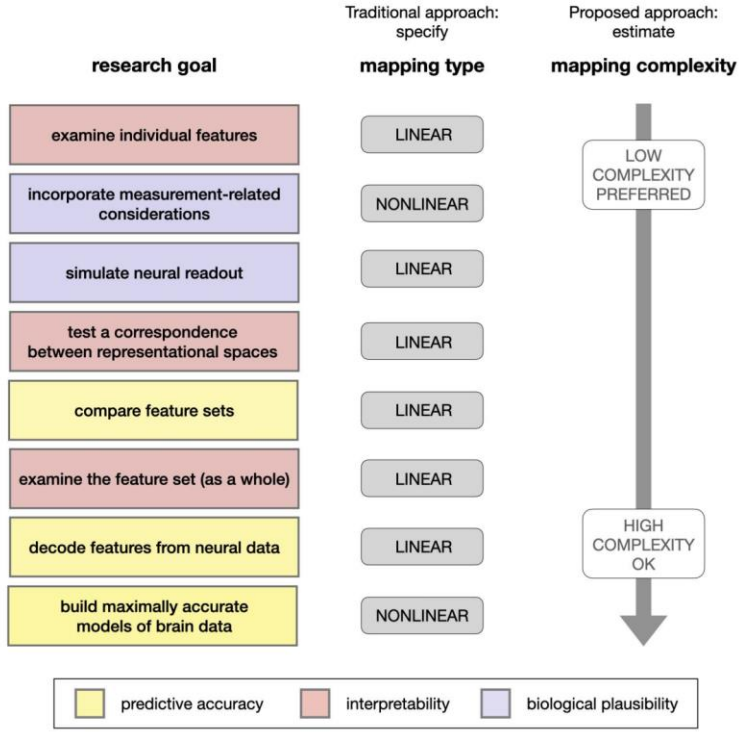
Encoding: **training and evaluation**



function f often modeled as linear

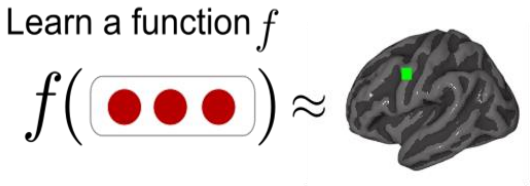
[Mitchell et al. 2008, Nishimoto et al., 2011; Sudre et al., 2012; Wehbe et al., 2014]

Considerations for f
Linear vs non-linear



Ivanova, Anna A., Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. "Is it that simple? Linear mapping models in cognitive neuroscience." bioRxiv (2021).

Encoding: **training and evaluation**



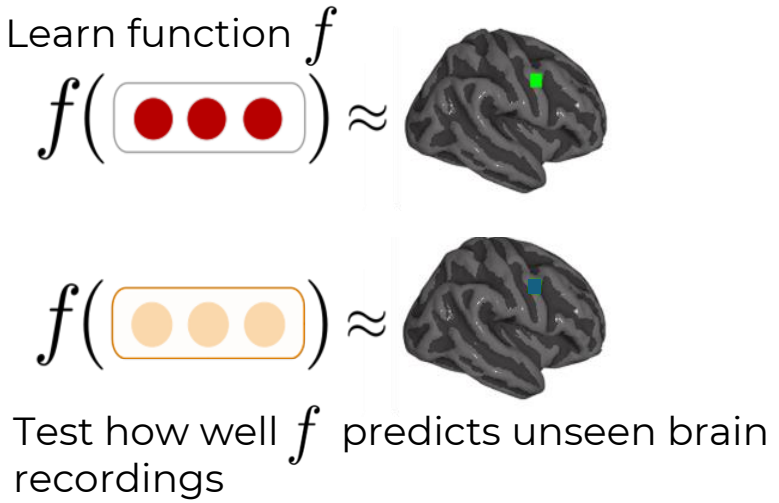
function f often modeled as linear

[Mitchell et al. 2008, Nishimoto et al., 2011; Sudre et al., 2012; Wehbe et al., 2014]

Training: cross validation (CV), regularization parameter chosen via nested CV

Evaluation: 1) make predictions for heldout data
2) compare predictions with true brain data
3) stringent statistical testing

Encoding: training **setup**



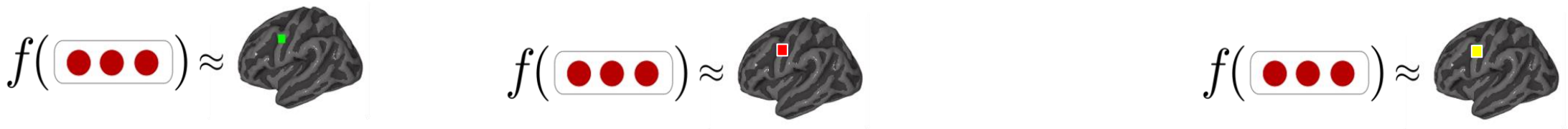
- Goal: find a mapping from stimulus representation to brain data that **generalizes** to new brain data

- Method:
  - Split dataset into train, validation, and test
  - Employ cross-validation to select model parameters based on validation dataset
  - Reduce overfitting by using regularization
    - Ridge regularization

Encoding: training **independent** models

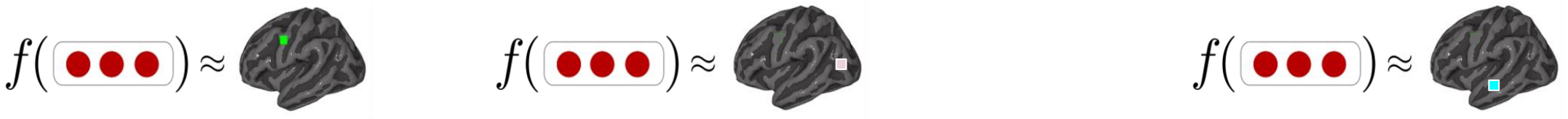
- Independent model per participant

P1 **P2** ... **PN**



- Independent model per voxel / sensor-timepoint

P1, v1 **P1, v2** ... **P1, vm**



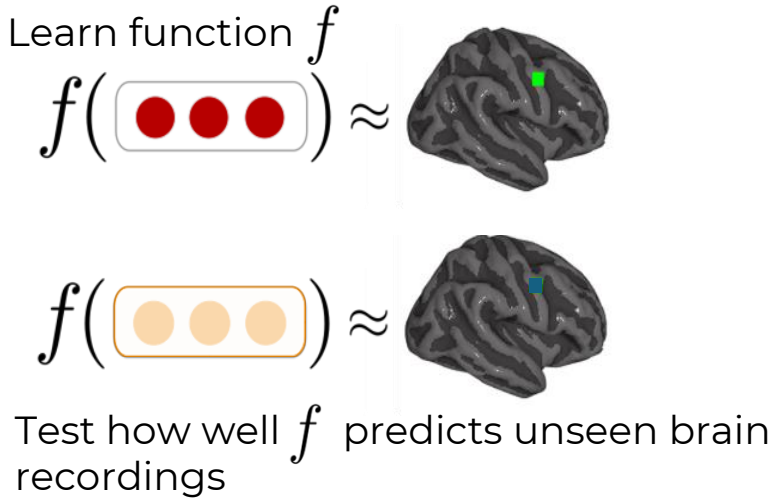
Encoding: fMRI specifics

Jain, Shailee, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S. Turek, and Alexander Huth. "Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech." Advances in Neural Information Processing Systems 33 (2020): 13738-13749.

Encoding: evaluation setup

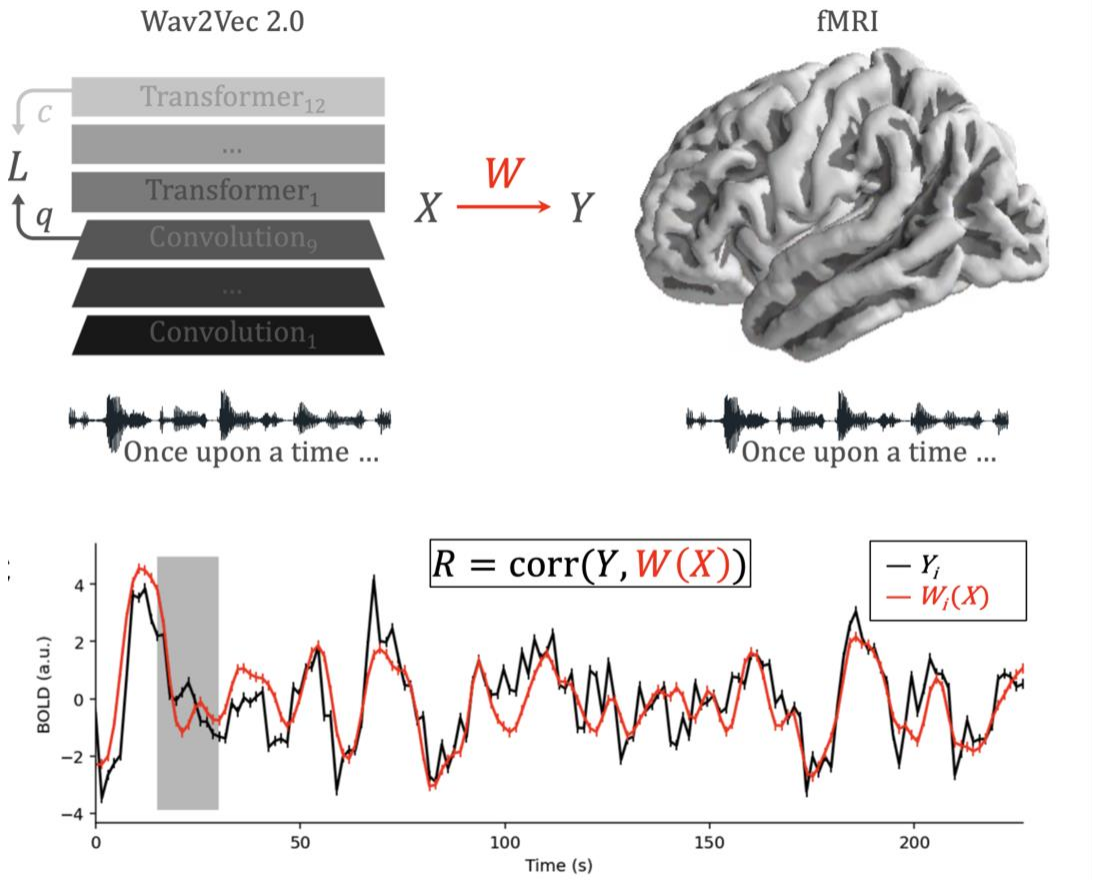
- Predict data heldout from training by applying learned function to corresponding stimulus representations

- Compare predictions of brain data to true brain data:
  - Evaluation metrics



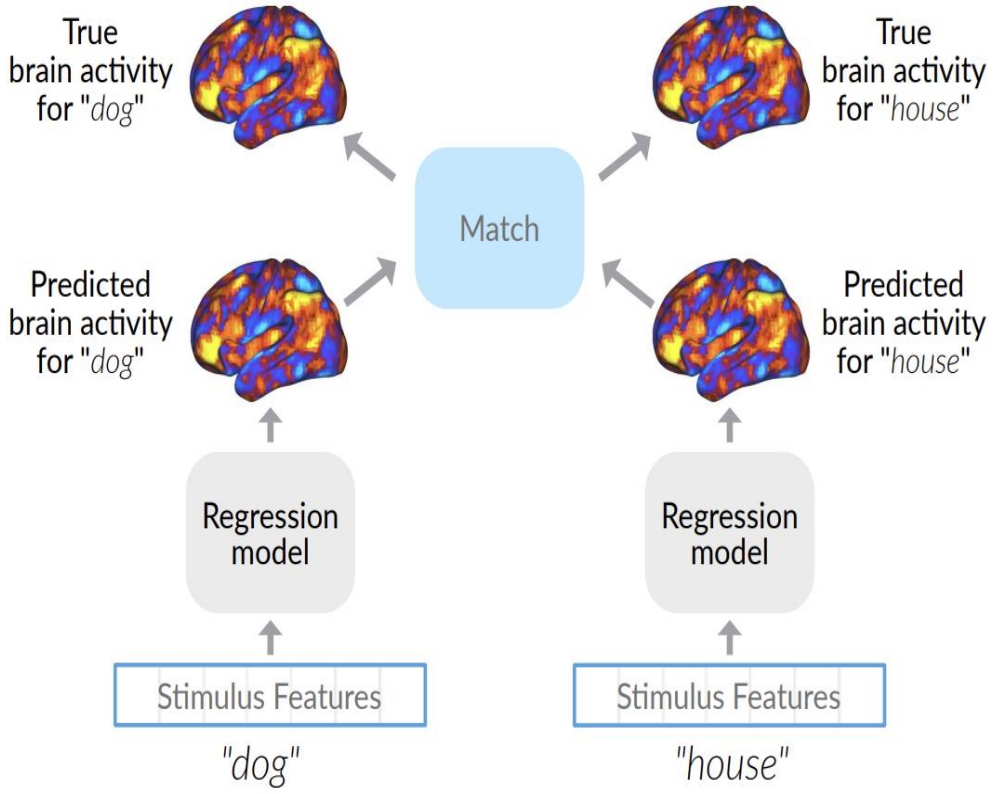
Encoding: evaluation metrics

Pearson correlation



Millet, Juliette, 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 preprint arXiv:2206.01685 (2022).

2v2 accuracy



Toneva, Mariya, Otilia Stretcu, Barnabás Póczos, Leila Wehbe, and Tom M. Mitchell. "Modeling task effects on meaning representation in the brain via zero-shot meg prediction." Advances in Neural Information Processing Systems 33 (2020): 5284-5295.

Agenda

- Neuro-AI Alignment: Introduction
  - Introduction to Brain Encoding & Decoding
  - Types of Brain Recording & Popular Text Datasets
  - Types of Stimulus Representation
  - Methodology

Agenda

- 09.00 AM – 10.30 AM Bapi Raju Neuro-AI alignment: Introduction
- 10.30 AM – 11.00 AM Coffee Break
- 11.00 AM – 12.30 PM Subba Reddy Language and the Brain:
  DL for Brain Encoding and Decoding

Thanks!

- Questions
  - Subba.reddy.oota<AT>tu-berlin.de
  - raju.bapi<AT>iiit.ac.in
- Connect with us:
  - https://www.linkedin.com/in/subba-reddy-oota-11a91254/
  - https://sites.google.com/view/bccl-iiith/home