Does the brain represent words?

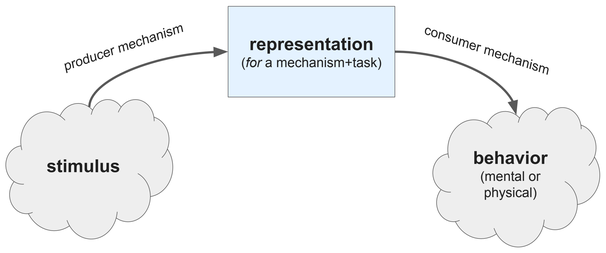
An evaluation of brain decoding studies of language understanding

Jon Gauthier and Anna Ivanova
jon@gauthiers.net annaiv@mit.edu
MIT Department of Brain and Cognitive Sciences

Introduction

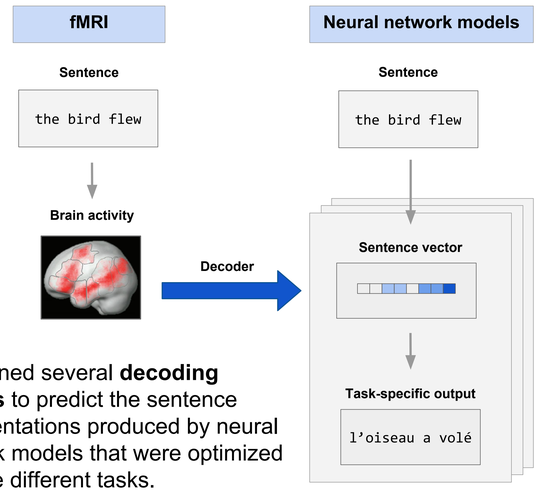
Recent language decoding studies claim to have discovered “meaning representations” and “semantic maps” in the brain [1,5].

We show that this style of evaluation greatly **underdetermines** the neural representation of language.

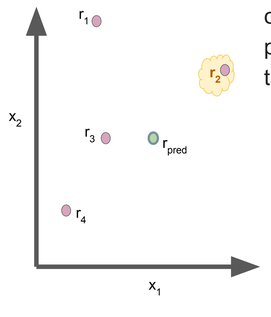


Representations do not exist in a vacuum. Claims of semantic representation without descriptions of their *consumers* and *producers* are dangerously underspecified.

Brain decoding



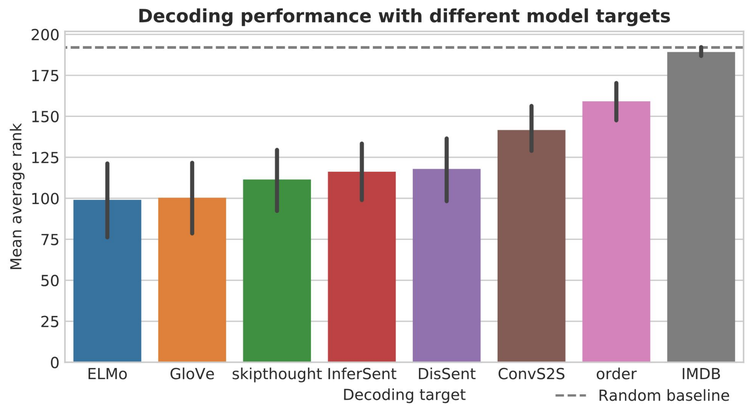
We trained several **decoding models** to predict the sentence representations produced by neural network models that were optimized to solve different tasks.



Our **mean average rank** evaluation compares the distance between a predicted representation r_{pred} and a target model representation r_2.

	Point	Distance from r_{pred}
1	r_3	0.1
2	r_4	0.3
3	r_2	0.35
4	r_1	0.45

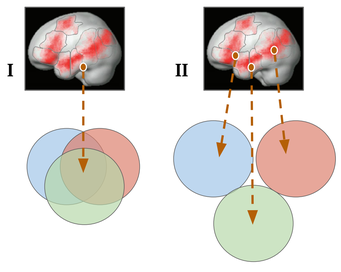
Results



Following Pereira et al. [5], we use neural activity to predict the representations of sentences produced by NLP sentence encoding models.

	Name	Task
1	ELMo	Language modeling
2	GloVe	Distributional modeling
3	skipthought	Language modeling
4	InferSent	Natural language inference
5	DisSent	Discourse understanding
6	ConvS2S	Machine translation
7	order	Image caption retrieval
8	IMDB	Sentiment analysis

Reverse inference on fMRI data fails to distinguish between several target models optimized to solve vastly different tasks.



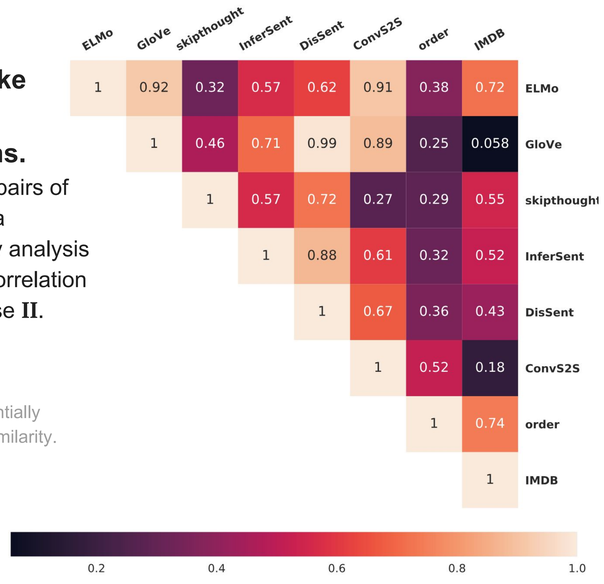
Why might decoding performance be similar across models?

Case I: Most models share some *core* representation; brain data can successfully predict this core.
Case II: Models do not share much *core* representation; brain data predicts all of their contents anyway.

These model representations make *different* sentence similarity predictions.

This heatmap compares pairs of model representations via representational similarity analysis [3]. The wide spread of correlation values here supports Case II.

Follow-up qualitative analysis (unpublished) suggests that the representations capture substantially different aspects of sentence similarity.



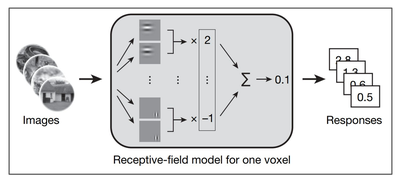
Brain decoding evaluations underdetermine

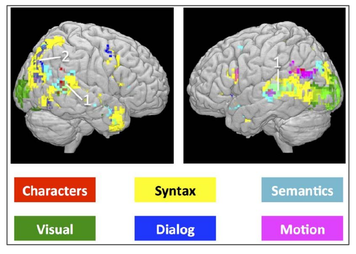
1. the **contents** of neural representations,
2. the **algorithms** which produce and consume them,
3. and the **tasks** which they are designed to solve.

Conclusions

Commit to a specific mechanism and task.

Kay et al. [2] propose an explicit encoding mechanism mapping from stimulus (natural image) to brain activity. They compare its performance to reasonable *baseline* models. The encoding mechanism is independently motivated by efficient coding arguments.



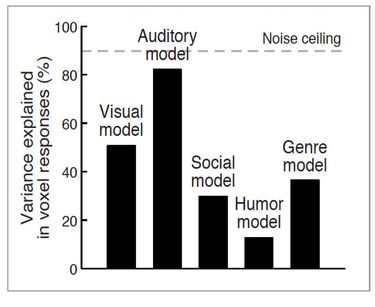


Decode/encode with interpretable representations.

Wehbe et al. [6] learn encoder models predicting brain activity from separate visual, syntactic, semantic, and discourse features. Their analysis reveals how different levels of representation are spatially arranged.
The same could be done with neural network representations — but *we first need to understand their contents.*

Explicitly measure explained variance in predicted neural activity.

Naselaris and Kay [4] advocate for a shift away from stimulus-based decoding models to explicit encoding models of representation. Encoding models can be directly compared by measuring the percentage of signal variance they explain. A model that accounts for 100% of the variance provides a perfect description of the system.



[1] Huth et al. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 2016.
[2] Kay et al. Identifying natural images from human brain activity. Nature Letters 2008.
[3] Kriegeskorte et al. Representational similarity analysis — connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2008.
[4] Naselaris & Kay. Resolving ambiguities of MVPA using explicit models of representation. TICS 2015.
[5] Pereira et al. Toward a universal decoder of linguistic meaning from brain activation. Nature Communications 2018.
[6] Wehbe et al. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLOS One 2014.