The neural architecture of language: Integrative modeling converges on predictive processing
The neural architecture of language: Integrative modeling converges on predictive processing
Integrative Modeling: link neural mechanisms, behavior, and computation

Schrimpf et al. Neuron 2020
Modeling Sensory Systems: Vision

V1

- edges
  - Hubel & Wiesel 1962

V2

- textures?
  - Freeman*, Ziemba*, Heeger, Simoncelli, Movshon 2015

V4

- curvatures??
  - Pasupathy & Connor 1999

IT

- objects??
  - Hung*, Kreiman*, Poggio, DiCarlo 2005

- purely descriptive

- very little predictive power

- often fall short for real-world stimuli

Necessary first steps! But …
Enter artificial neural networks

highly accurate at predicting neural responses

work well for real-world stimuli

make predictions on any new input

$\text{IT Explained Variance (\%)}$

$\text{Task performance (4,500 images of 36 objects)}$

$r = 0.87 \pm 0.15$
Recent progress in sensory systems: increasingly accurate neurally-mechanistic models


Artificial neural networks have worked well in modeling sensory cortex – could they also predict higher cognition?
The human language network

working definition (Fedorenko et al.): a set of left-lateralized regions on the lateral surfaces of frontal and temporal cortex that support high-level language processing.

Language > Perceptually matched control
Sentences > Lists of nonwords
The human language network

Key signature: stronger response to sentences than lists of unconnected words
What are the mechanisms underlying human language comprehension?

the dog is taking a bath

"meaning"
What are the mechanisms underlying human language comprehension?

Want to find the neural mechanisms that predict our data

Computational understanding of human language processing

Guide and constraint next-generation NLP models

Future clinical applications (BMI)
The data target: human neural recordings

courtesy of Idan Blank
The data target: human neural recordings

Pereira et al. 2018
627 sentences x 13,517 voxels in 10 subjects
Beekeeping encourages the conservation of local habitats. | It is in every beekeeper’s linguistic interest ...

Fedorenko et al. 2016
416 words x 97 electrodes in 5 subjects
ALEX | WAS | TIRED | SO | HE | TOOK | A | NAP

Blank et al. 2014
1,317 story fragments x 60 fROIs in 5 subjects
If you were to journey to the | North of England, you would come to a valley | that is surrounded by moors as high as | mountains. It is in this | valley where you would find the city of Bradford, | ...
Quantifying match-to-brain: Benchmarking

EXPERIMENTAL PARADIGM

stimulus

present

E₁ E₂ ... Eₙ

h₁,₁ h₂,₁ ... hₙ,₁

... ... ... ...

h₁,l h₂,l ... hₙ,l

record

observed data

SIMILARITY METRIC

continuum of scores

“how close is this model in comparison to others”

stimulus

present

record

observed data
Quantifying match-to-brain: Benchmarking

Goal 1: differentiate models
We only care about best-matching model (for now)

Goal 2: how close are we?
Goal 2 “how close are we” – ceiling

compute how well a pool of subjects predicts a held-out subject

\[
predictivity_{\text{normalized}} = \frac{\text{predictivity}}{\text{ceiling}}
\]
Integrative Modeling: one model to predict them all

Pereira2018
"Beekeeping encourages the conservation of local habitats. It is in every beekeeper's interest..."

Fedorenko2016
"Alex was tired so he took a nap."

Blank2014
"If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains. It is in this valley where you..."
Models considered (n=43)

**Embedding** type models: GloVe, word2vec, topicETM
Models considered (n=43)

**Embedding** type models: GloVe, word2vec, topicETM

**Recurrent networks**: skip-thoughts, LSTM lm_1b

Language Modeling

- Alaska is
- Alaska is about
- Alaska is about twelve
- Alaska is about twelve times
- Alaska is about twelve times larger
- Alaska is about twelve times larger than
- Alaska is about twelve times larger than New York

Image from https://www.quora.com/What-is-a-masked-language-model-and-how-is-it-related-to-BERT

Jozefowicz, vinyals, Schuster, Shazeer, Wu 2016
Models considered (n=43)

**Embedding** type models: GloVe, word2vec, topicETM

**Recurrent networks**: skip-thoughts, LSTM lm_1b

---

**Transformers**
- BERTs
- RoBERTas
- XLMs
- Transformer-XLs
- XLNets
- CTRL
- T5s
- ALBERTs
- GPTs

---

Treating models as experimental subjects

Beekeeping encourages the conservation of local habitats.

It is in every beekeeper's interest to conserve local plants that produce pollen.
Similarity metric: Neural Predictivity / Encoding

Model activations

Brain recordings

sentences
Similarity metric: Neural Predictivity / Encoding

- Model activations
  - 80% of model activations
  - 20% of model activations

- Brain recordings
  - 80% of brain recordings
  - 20% of brain recordings

*Fit* regression weights

*Predict* held-out

*Measure* correlation

*5-fold cross-validation*
GloVe voxel-wise predictivity scores

GloVe →

Aggregate scores: median over voxels and subjects
Certain language models predict human language recordings

gpt2-xl hits our estimated ceiling for this benchmark

Small differences can lead to very different brain predictivities, warranting a full survey

Jain & Huth 2018
Gauthier & Ivanova 2018
Jat et al. 2019
Toneva & Wehbe 2019
Gauthier & Levy 2020
Wang et al. 2020
GPT2-xl accurately predicts a large portion of voxels.
Language Models predict human language recordings

Pereira2018

Lack of long-range semantic context?

Fedorenko2016

Blank2014

Normalized Predictivity

Normalized Predictivity
Model scores across benchmarks are correlated, although differences exist.

Scores generalize to a good extent within Pereira2018 (some subject overlap, similar experiments).

Scores generalize to a good extent:
- Pereira2018 → Fedorenko2016: $r = .94$
- Pereira2018 → Blank2014: $r = .61$

But there are also differences, making each individual benchmark valuable.
What explains the model differences?

Goal 1: possible explanation why some models are better than others (hinting at optimization in the brain)

Goal 2: if x-axis is easier to optimize than y-axis, we can more efficiently improve models
= Gold dollar =
The gold dollar or gold one @@ dollar piece was a coin struck as a regular issue by the United States Bureau of the Mint from 1849 to 1889. The coin had three types over its lifetime, all designed by Mint Chief Engraver James B. Longacre. The Type 1 issue had …

Surprisal of seeing actual next word: \textit{perplexity} = \exp(\text{NLL Loss})

Merity et al. 2016
Next-Word Prediction performance correlates with neural predictivity

$r = .44$
What about other language tasks?

9 “General Language Understanding Evaluation” tasks:

- Sentence grammaticality (CoLa)
- Sentence sentiment (SST-2)
- Semantic similarity (QQP, MRPC, STS-B)
- Entailment (MNLT, RTE)
- Question-answer coherence (QNLI)
- Winograd (WNLI; ignored due to known issues)
Next-Word Prediction performance selectively correlates with neural predictivity.

Online prediction may fundamentally shape language processing in the brain.
Is any of this behaviorally relevant?
Is any of this behaviorally relevant?

Behavioral predictivity

Neural predictivity
Futrell et al. 2018

10256 words x 179 subjects

If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains. It is in this valley where you would find the city of Bradford, where once a thousand spinning...
Behavioral scores

Futrell2018

Normalized Predictivity

Emb.  BERT  XLM  T5  AIBERT  GPT
Neural scores correlate with Behavioral scores

Task scores correlate with Behavioral scores
What is the relative importance of evolutionary and learning-based optimization?

Evolution $\approx$ community optimization over architectural properties

Experience-dependent learning $\approx$ updating of weights over training
Architecture substantially contributes to models’ brain predictivity

Training generally improves scores by ~53%

Inherent structure might be a key driver of brain-like language representations

Large feature size without structure is insufficient

$r = .74$
Summary

Certain artificial neural networks are viable **mechanistic hypotheses** for how **predictive language processing** is implemented in human neural tissue.

1. Across three datasets, specific models such as gpt2-xl consistently predict human recordings

2. Model performance on a **next-word prediction** task is **correlated** with its brain-predictivity

3. Neural-predictivity is correlated with behavior-predictivity

4. **Architecture** substantially contributes to brain-predictive representations