Uso de modelos de NLP para el estudio del lenguaje en el cerebro

Clase 2: Neuroimágenes

Dr. Bruno Bianchi
Laboratorio de Inteligencia Artificial Aplicada
Dpto Computación - FCEN - UBA
Instituto Cs Computación - CONICET - UBA

[ELBA]

Julio 2024

Objetivos de esta clase

- Intro a la actividad cerebral
- Intro a la resonancia magnética funcional
- Paper de Huth 2016

Viendo el cerebro

Ya sabemos cómo se representan las palabras en la computadora

¿Y en el cerebro?

¿Cómo estudiamos cómo se representan?

Breve intro a neuroimágenes

Breve intro a neuroimágenes

¿De qué hablamos cuando hablamos de neuroimágenes?

- **MRI**: resonancia magnética
- **fMRI**: resonancia magnética funcional
- **DTI:** diffusion tensor imaging
- **fNIRS**: Functional near-infrared spectroscopy
- **PET**: Tomografía por emisión de positrones
- **EEG**: Electroencefalografía
- **MEG**: Magnetoencefalografía
- **iEEG**: Electroencefalografía Intracraneal
- **LFP**: Potenciales de campo local
- Single-Cell Recordings

Breve intro a neuroimágenes

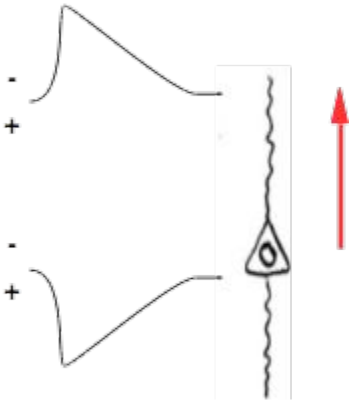
¿De qué hablamos cuando hablamos de neuroimágenes?

- **MRI**: resonancia magnética
- **fMRI**: resonancia magnética funcional
- **DTI:** diffusion tensor imaging
- **fNIRS**: Functional near-infrared spectroscopy
- **PET**: Tomografía por emisión de positrones
- **EEG**: Electroencefalografía
- **MEG**: Magnetoencefalografía
- **iEEG**: Electroencefalografía Intracraneal
- **LFP**: Potenciales de campo local
- Single-Cell Recordings

Breve intro a neuroimágenes: M|EEG

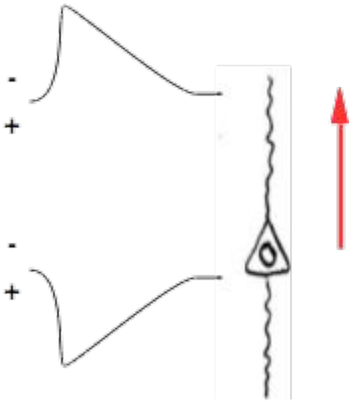
Breve intro a neuroimágenes: M|EEG

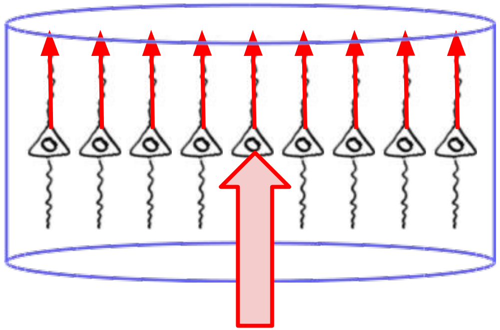
Medio
extracelular



Breve intro a neuroimágenes: M|EEG

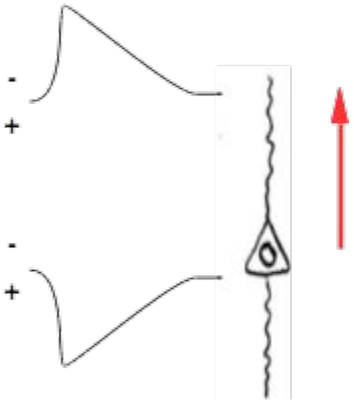
Medio extracelular

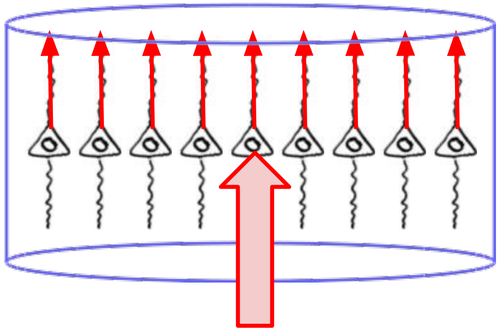


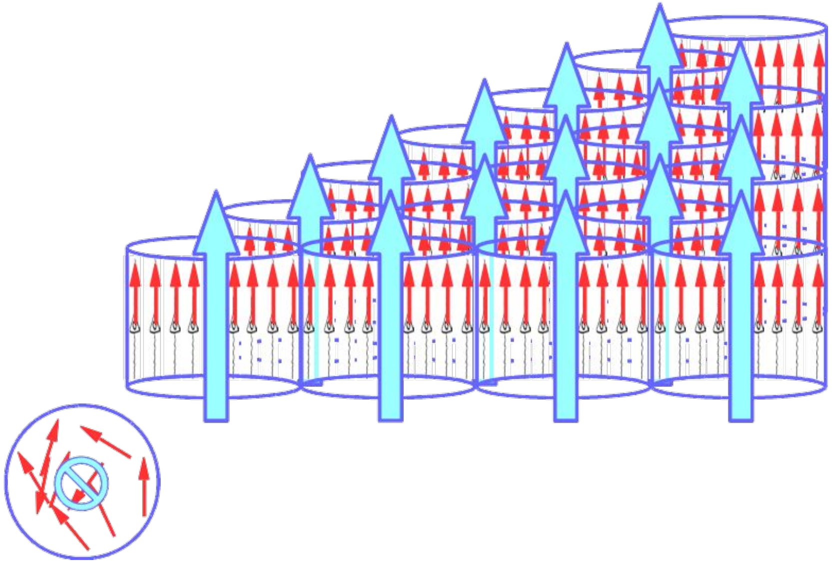


Breve intro a neuroimágenes: M|EEG

Medio extracelular

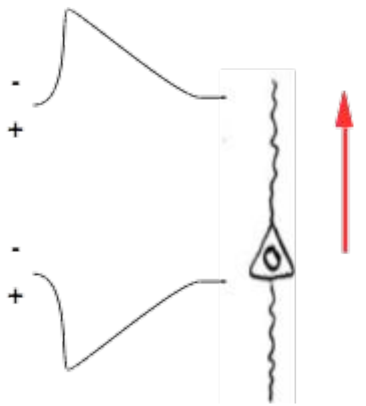


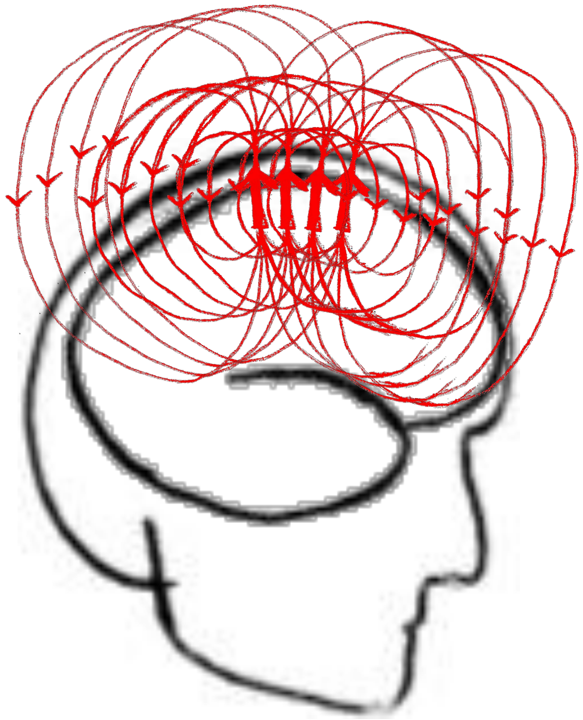




Breve intro a neuroimágenes: M|EEG

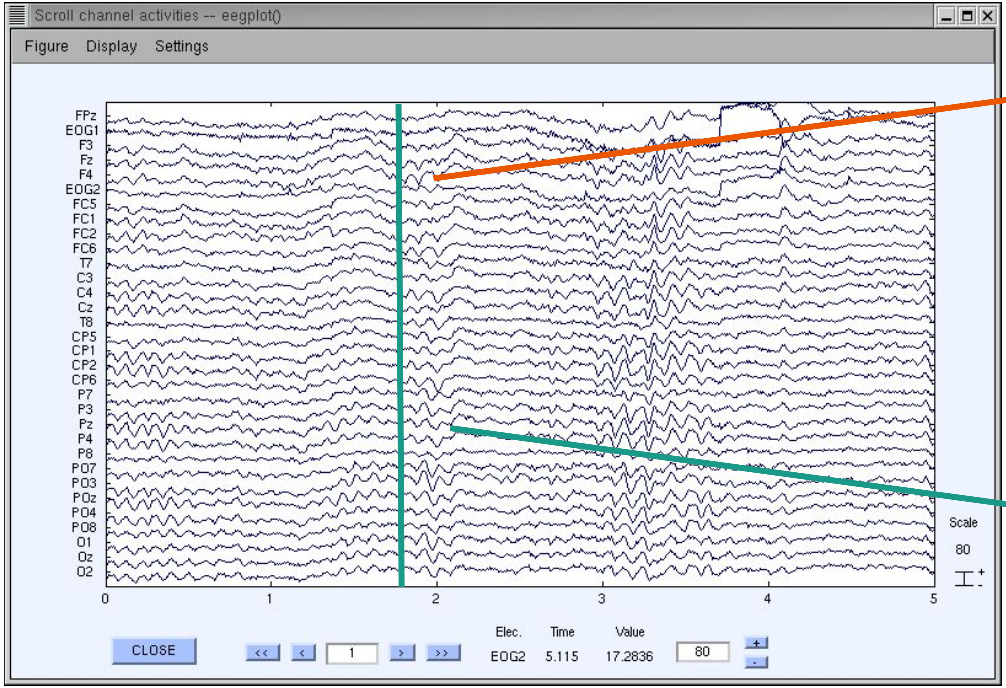
Medio extracelular



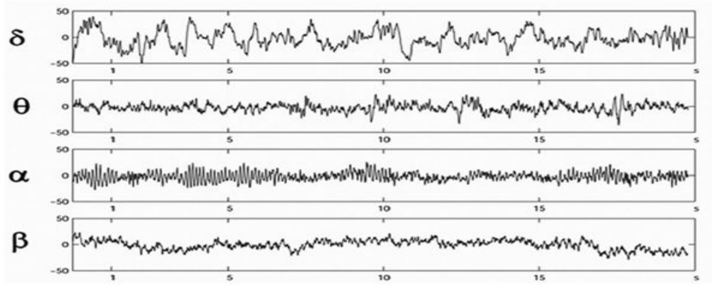


Breve intro a neuroimágenes: M|EEG

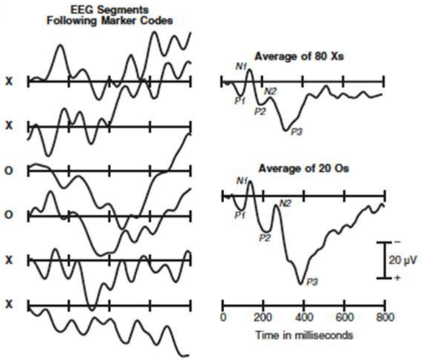
Breve intro a neuroimágenes: M|EEG



$f_x(t)$



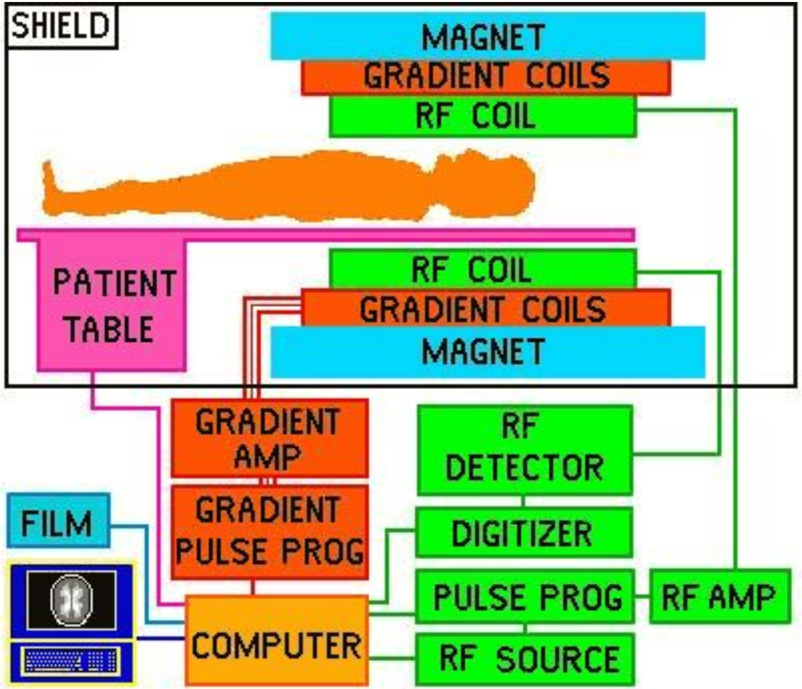
$x(t)$



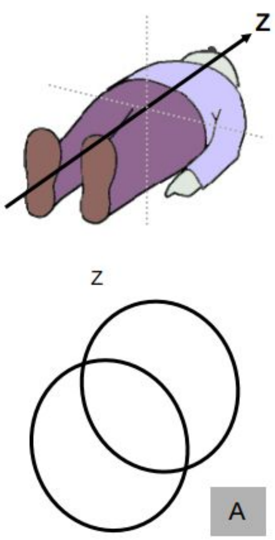
ERP
Components:
P = Positive
N = Negative
P1 = P100
P2 = P200
P3 = P300
etc.

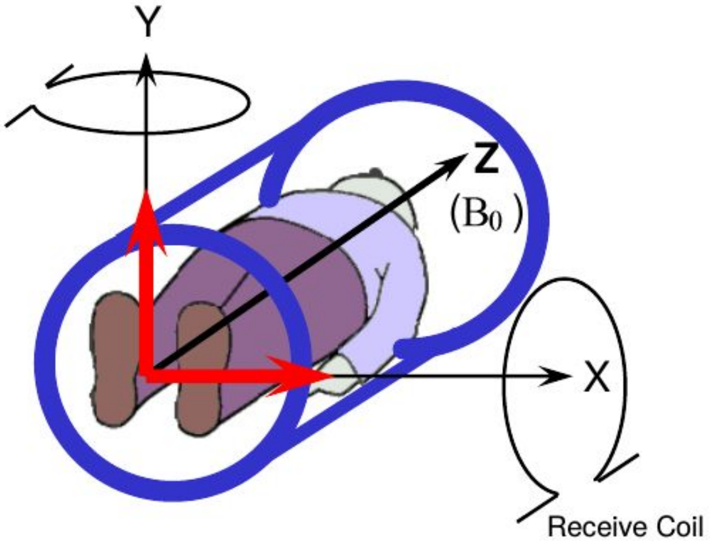
Breve intro a neuroimágenes: MRI/fMRI

Breve intro a neuroimágenes: MRI/fMRI

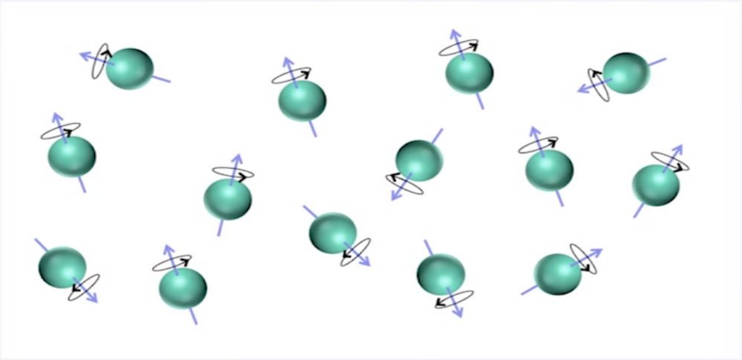


Breve intro a neuroimágenes: MRI/fMRI

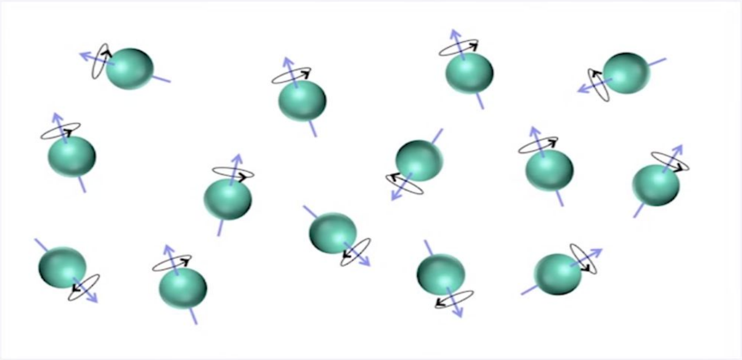


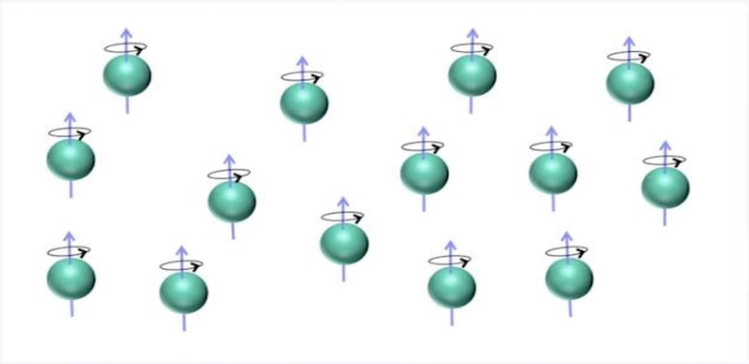


Breve intro a neuroimágenes: MRI/fMRI

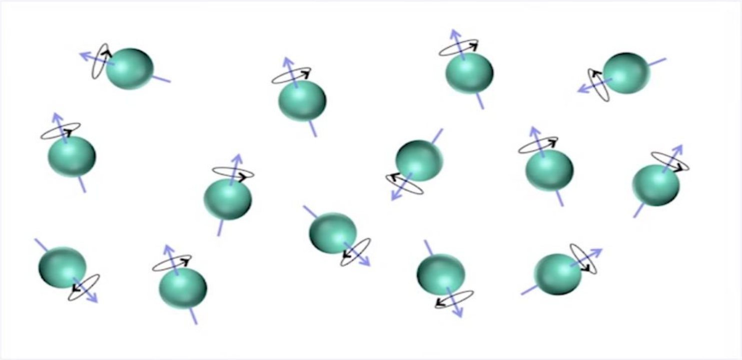


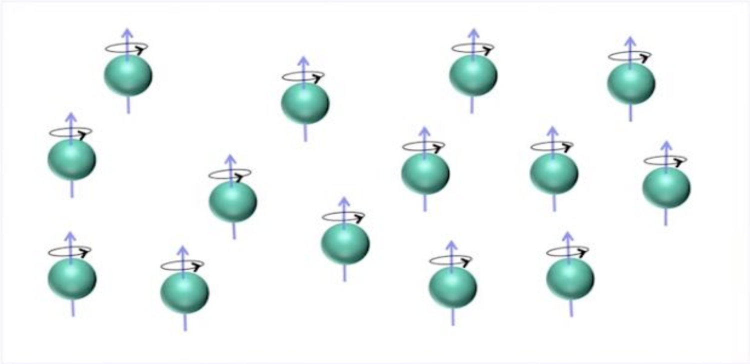
Breve intro a neuroimágenes: MRI/fMRI

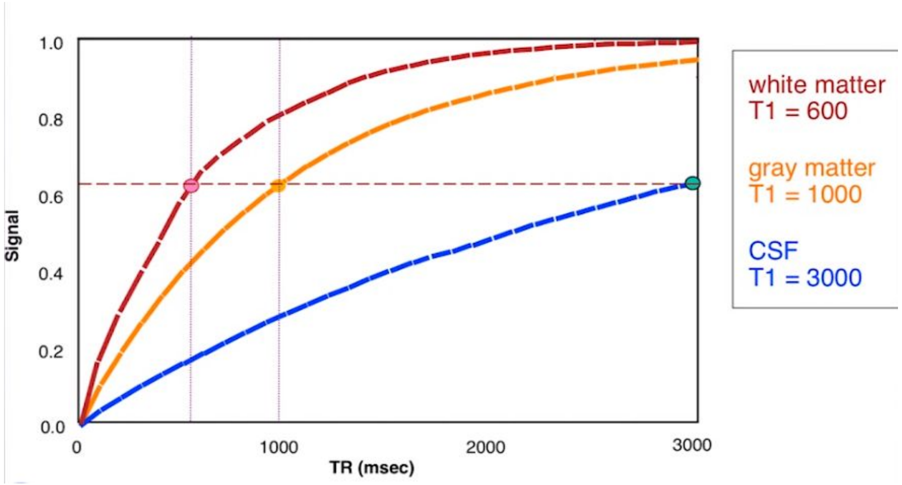




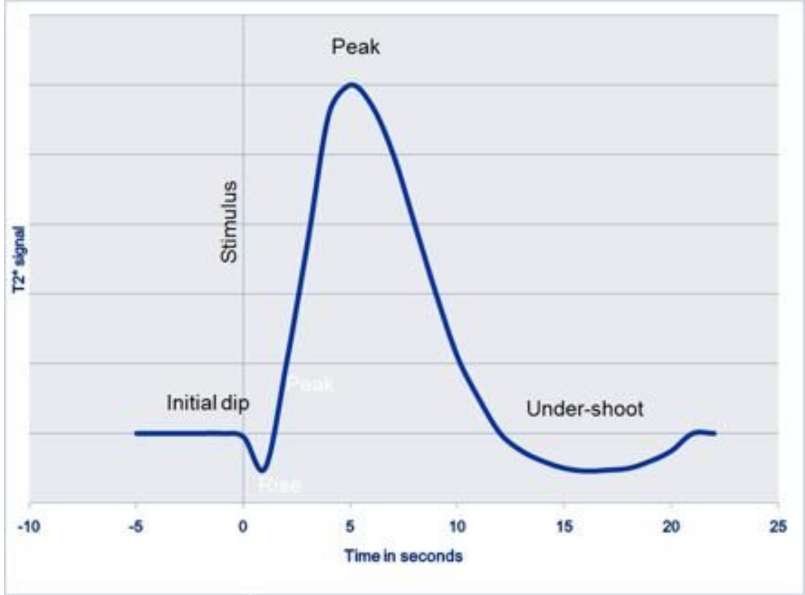
Breve intro a neuroimágenes: MRI/fMRI

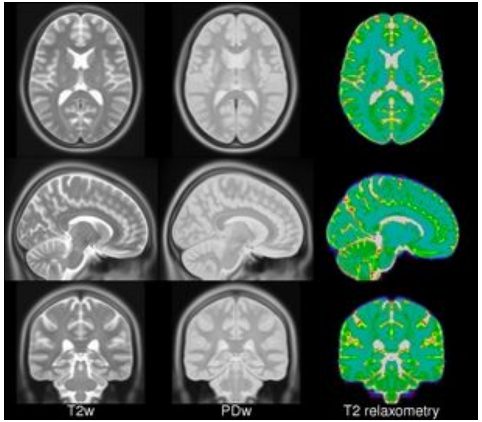


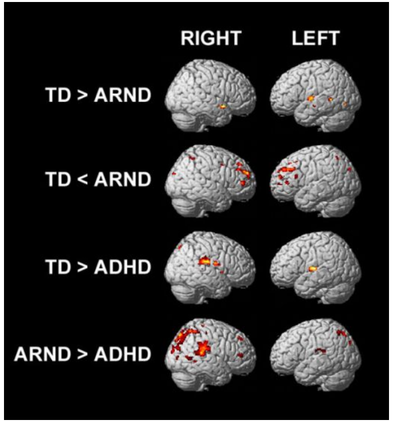


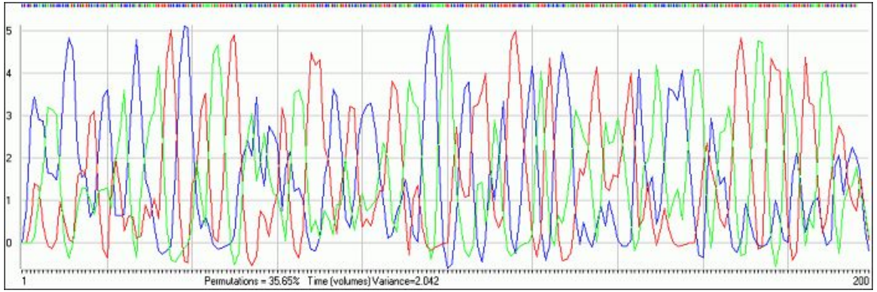


Breve intro a neuroimágenes: MRI/fMRI



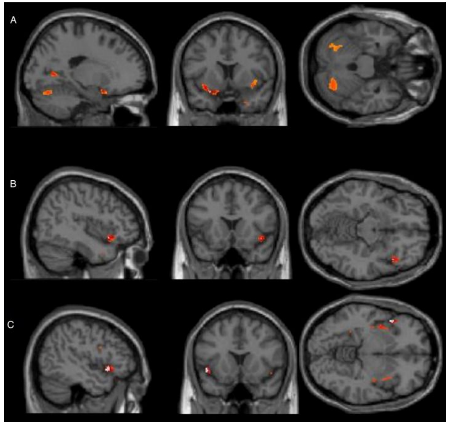






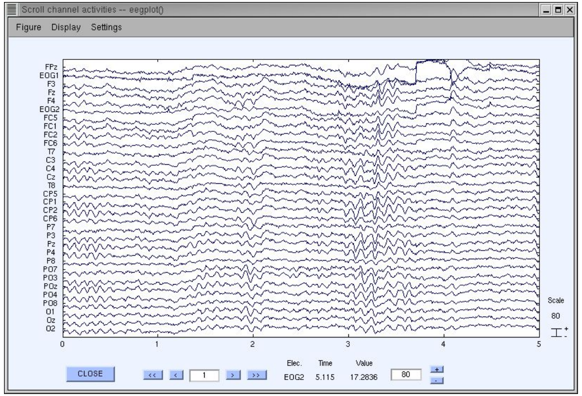
En resumen

fMRI



- Medida indirecta de la actividad neuronal
- Baja resolución temporal
- Excelente resolución espacial

EEG

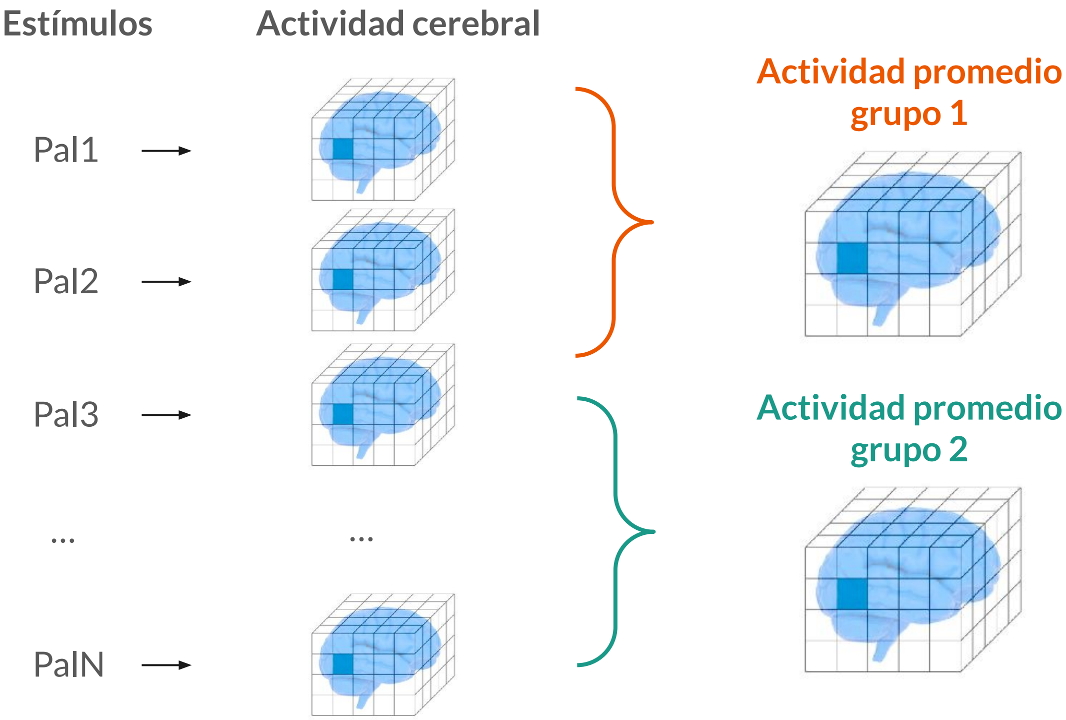


- Medida directa de la actividad neuronal
- Excelente resolución temporal
- Baja resolución espacial

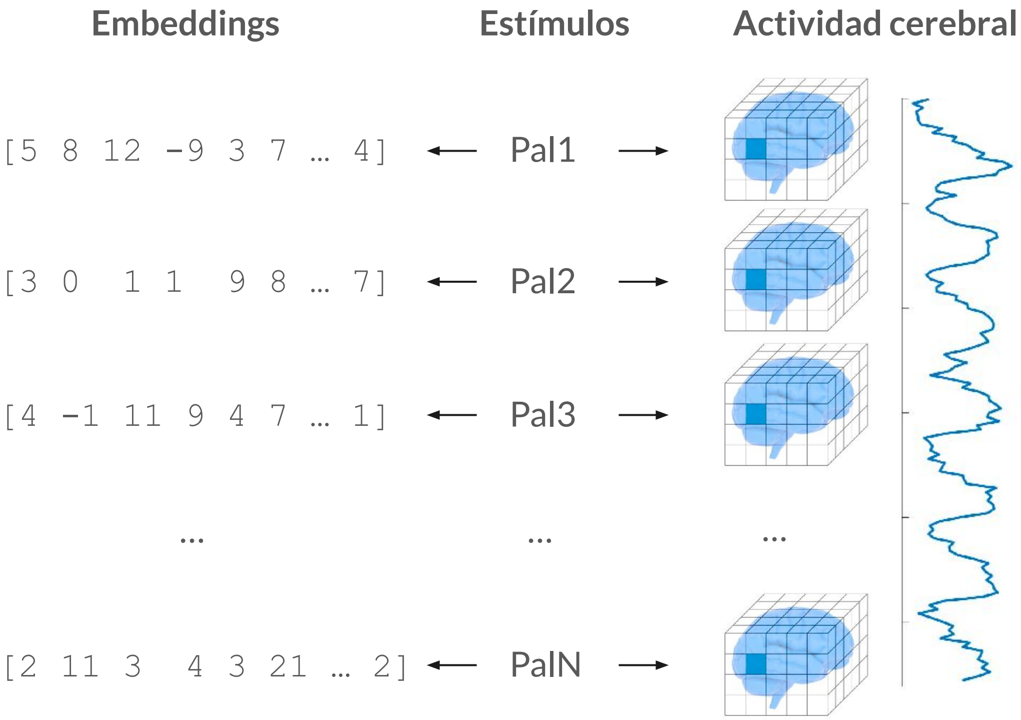
¿Y todo esto para qué?

Estudios clásicos de neuroimágenes

- ➔ Estímulos cuidadosamente seleccionados
- ➔ Presentaciones aisladas
- ➔ Análisis categóricos y por grupos
- ➔ Análisis de las diferencias entre los grupos



Alineamiento de embeddings con neuroimágenes



- Estímulos naturales
- Presentaciones continuas, contextualizadas
- Análisis más complejos
- Se analizan las relaciones entre variables continuas

Alineamiento de embeddings con neuroimágenes

Embeddings | **Estímulos** | **Actividad cerebral**

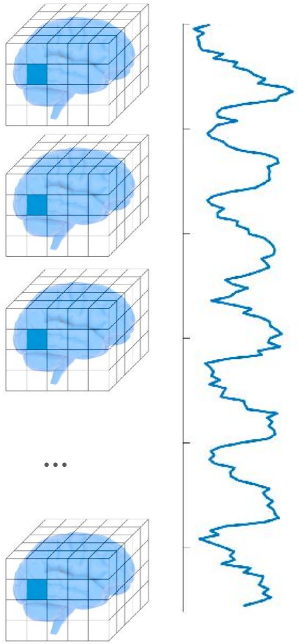
[5 8 12 -9 3 7 … 4] ⟵ Pal1 ⟶

[3 0 1 1 9 8 … 7] ⟵ Pal2 ⟶

[4 -1 11 9 4 7 … 1] ⟵ Pal3 ⟶

… … …

[2 11 3 4 3 21 … 2] ⟵ PalN ⟶

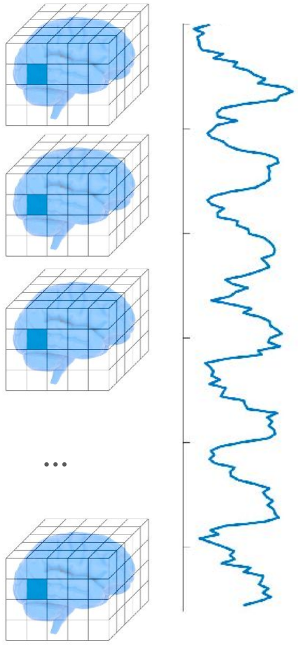


Regresión Lineal:

$$Act \sim \beta_1 d_1 + \beta_2 d_2 + \ldots \beta_D d_D$$

Alineamiento de embeddings con neuroimágenes

Embeddings		Estímulos		Actividad cerebral
[5 8 12 -9 3 7 … 4]	←	Pal1	→	
[3 0 1 1 9 8 … 7]	←	Pal2	→	
[4 -1 11 9 4 7 … 1]	←	Pal3	→	
…		…		…
[2 11 3 4 3 21 … 2]	←	PalN	→	



Regresión Lineal:

$$Act \sim \beta_1 d_1 + \beta_2 d_2 + \dots \beta_D d_D$$

- ➔ Qué dimensiones modelan mejor la activación?
- ➔ Qué información codifican esas dimensiones?
- ➔ Qué información codifican esos voxeles?
- ➔ Tipos de palabras?
- ➔ Qué capas de los modelos?

Alineamiento de embeddings con neuroimágenes

2008

Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell,[1*] Svetlana V. Shinkareva,[2] Andrew Carlson,[1] Kai-Min Chang,[3,4] Vicente L. Malave,[5] Robert A. Mason,[3] Marcel Adam Just[3]

The question of how the human brain represents conceptual knowledge has been debated in many scientific fields. Brain imaging studies have shown that different spatial patterns of neural activation are associated with thinking about different semantic categories of pictures and words (for example, tools, buildings, and animals). We present a computational model that predicts the functional magnetic resonance imaging (fMRI) neural activation associated with words for which fMRI data are not yet available. This model is trained with a combination of data from a trillion-word text corpus and observed fMRI data associated with viewing several dozen concrete nouns. Once trained, the model predicts fMRI activation for thousands of other concrete nouns in the text corpus, with highly significant accuracies over the 60 nouns for which we currently have fMRI data.

Evaluating word embeddings with fMRI and eye-tracking

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

2016

Abstract

The workshop CfP assumes that downstream evaluation of word embeddings is impractical, and that a valid evaluation metric for pairs of word embeddings can be found. I argue below that if so, the only meaningful evaluation procedure is comparison with measures of *human word processing in the wild*. Such evaluation is non-trivial, but I present a practical procedure here, evaluating word embeddings as features in a multi-dimensional regression model predicting brain imaging or eye-tracking word-level aggregate statistics.

Natural speech reveals the semantic maps that tile human cerebral cortex

Alexander G. Huth[a], Wendy A. de Heer[b], Thomas L. Griffiths[a,b], Frédéric E. Theunissen[a,b], and Jack L. Gallant[a,b]
[a]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA
[b]Department of Psychology, University of California, Berkeley, CA 94720, USA

2016

Abstract

The meaning of language is represented in regions of the cerebral cortex collectively known as the "semantic system". However, little of the semantic system has been mapped comprehensively, and the semantic selectivity of most regions is unknown. Here we systematically map semantic selectivity across the cortex using voxel-wise modeling of fMRI data collected while subjects listened to hours of narrative stories. We show that the semantic system is organized into intricate patterns that appear consistent across individuals. We then use a novel generative model to create a detailed semantic atlas. Our results suggest that most areas within the semantic system represent information about specific semantic domains, or groups of related concepts, and our atlas shows which domains are represented in each area. This study demonstrates that data-driven methods—commonplace in studies of human neuroanatomy and functional connectivity—provide a powerful and efficient means for mapping functional representations in the brain.

Decoding the Neural Representation of Story Meanings across Languages

2017

Morteza Dehghani,[1*] Reihane Boghrati,[1] Kingson Man,[1] Joe Hoover,[1] Sarah I. Gimbel,[1] Ashish Vaswani,[2] Jason D. Zevin,[1] Mary Helen Immordino-Yang,[1] Andrew S. Gordon,[1] Antonio Damasio,[1] and Jonas T. Kaplan[1]
[1]*University of Southern California, Los Angeles, CA*
[2]*Google Brain, Mountain View, California*

Alineamiento de embeddings con neuroimágenes

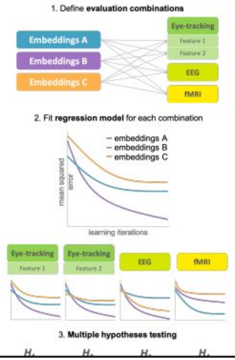
2019

CogniVal: A Framework for Cognitive Word Embedding Evaluation

Nora Hollenstein[1], Antonio de la Torre[1], Nicolas Langer[2], Ce Zhang[1]
[1] Department of Computer Science, ETH Zurich
{noraho,antonide,ce.zhang}@ethz.ch
[2] Department of Psychology, University of Zurich
n.langer@psychologie.uzh.ch

Abstract

An interesting method of evaluating word representations is by how much they reflect the semantic representations in the human brain. However, most, if not all, previous works only focus on small datasets and a single modality. In this paper, we present the first multi-modal framework for evaluating English word representations based on cognitive lexical semantics. Six types of word embeddings are evaluated by fitting them to 15 datasets of eye-tracking, EEG and fMRI signals recorded during language processing. To achieve a global score over all evaluation hypotheses, we apply statistical significance testing accounting for the multiple comparisons problem. This framework is easily extensible and available to include other intrinsic and extrinsic evaluation methods. We find strong correlations in the results between cognitive datasets, across recording modalities and to their performance on extrinsic NLP tasks.



Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains

Samira Abnar Lisa Beinborn Rochelle Choenni Willem Zuidema

Institute for Logic, Language and Computation
University of Amsterdam

{abnar,l.beinborn}@uva.nl,rochelle.choenni@student.uva.nl,zuidema@uva.nl

Robust Evaluation of Language–Brain Encoding Experiments

Lisa Beinborn, Samira Abnar, Rochelle Choenni

Institute for Logic, Language and Computation
University of Amsterdam
l.beinborn@uva.nl, s.abnar@uva.nl, rochelle.choenni@student.uva.nl

Abstract. Language-brain encoding experiments evaluate the ability of language models to predict brain responses elicited by language stimuli. The evaluation scenarios for this task have not yet been standardized which makes it difficult to compare and interpret results. We perform a series of evaluation experiments with a consistent encoding setup and compute the results for multiple fMRI datasets. In addition, we test the sensitivity of the evaluation measures to randomized data and analyze the effect of voxel selection methods. Our experimental framework is publicly available to make modelling decisions more transparent and support reproducibility for future comparisons.

Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)

Mariya Toneva
Neuroscience Institute
Department of Machine Learning
Carnegie Mellon University
mariya@cmu.edu

Leila Wehbe
Neuroscience Institute
Department of Machine Learning
Carnegie Mellon University
lwehbe@cmu.edu

Abstract

Neural networks models for NLP are typically implemented without the explicit encoding of language rules and yet they are able to break one performance record after another. This has generated a lot of research interest in interpreting the representations learned by these networks. We propose here a novel interpretation approach that relies on the only processing system we have that does understand language: the human brain. We use brain imaging recordings of subjects reading complex natural text to interpret word and sequence embeddings from 4 recent NLP models - ELMo, USE, BERT and Transformer-XL. We study how their representations differ across layer depth, context length, and attention type. Our results reveal differences in the context-related representations across these models. Further, in the transformer models, we find an interaction between layer depth and context length, and between layer depth and attention type. We finally hypothesize that altering BERT to better align with brain recordings would enable it to also better understand language. Probing the altered BERT using syntactic NLP tasks reveals that the model with increased brain-alignment outperforms the original model. Cognitive neuroscientists have already begun using NLP networks to study the brain, and this work closes the loop to allow the interaction between NLP and cognitive neuroscience to be a true cross-pollination.

Alineamiento de embeddings con neuroimágenes

Long-range and hierarchical language predictions in brains and algorithms

2021

Charlotte Caucheteux[1,2], Alexandre Gramfort[2], and Jean-Rémi King[1,3]

[1]Facebook AI Research, Paris, France; [2]Université Paris-Saclay, Inria, CEA, Palaiseau, France; [3]École normale supérieure, PSL University, CNRS, Paris, France

Deep learning has recently made remarkable progress in natural language processing. Yet, the resulting algorithms remain far from competing with the language abilities of the human brain. Predictive coding theory offers a potential explanation to this discrepancy: while deep language algorithms are optimized to predict adjacent words, the human brain would be tuned to make long-range and hierarchical predictions. To test this hypothesis, we analyze the fMRI brain signals of 304 subjects each listening to ≈70 min of short stories. After confirming that the activations of deep language algorithms linearly map onto those of the brain, we show that enhancing these models with long-range forecast representations improves their brain-mapping. The results further reveal a hierarchy of predictions in the brain, whereby the fronto-parietal cortices forecast more abstract and more distant representations than the temporal cortices. Overall, this study strengthens predictive coding theory and suggests a critical role of long-range and hierarchical predictions in natural language processing.

Here, we address these issues by analyzing the brain signals of 304 subjects listening to short stories, while their brain activity was recorded with fMRI (32). First, we confirm that deep language algorithms linearly map onto brain activity (6, 8, 33). Then, we show that adding long-range and hierarchical predictions improves such mapping. After confirming that the activations of deep language algorithms linearly map onto brain activity (6, 8, 33), we show that enhancing these models with long-range and hierarchical predictions improves their brain mapping. Critically, and in line with predictive coding theory, our results reveal a hierarchical organization of language prediction in the cortex, in which the highest stages forecast *(i)* the most distant and *(ii)* the most abstract representations.

Results

Deep language models map onto brain activity. First, we quantify the similarity between deep language models and the brain,

Brains and algorithms partially converge in natural language processing

2022

Charlotte Caucheteux[1,2✉] & Jean-Rémi King[1,3✉]

Deep learning algorithms trained to predict masked words from large amount of text have recently been shown to generate activations similar to those of the human brain. However, what drives this similarity remains currently unknown. Here, we systematically compare a variety of deep language models to identify the computational principles that lead them to generate brain-like representations of sentences. Specifically, we analyze the brain responses to 400 isolated sentences in a large cohort of 102 subjects, each recorded for two hours with functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). We then test where and when each of these algorithms maps onto the brain responses. Fina
we estimate how the architecture, training, and performance of these models independen
account for the generation of brain-like representations. Our analyses reveal two m
findings. First, the similarity between the algorithms and the brain primarily depends on th
ability to predict words from context. Second, this similarity reveals the rise and maintenan
of perceptual, lexical, and compositional representations within each cortical region. Over
this study shows that modern language algorithms partially converge towards brain-l
solutions, and thus delineates a promising path to unravel the foundations of natural langua
processing.

OPEN

Shared computational principles for language processing in humans and deep language models

2022

Ariel Goldstein[1,2✉], Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Lucia Melloni[4], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3], Liat Hasenfratz[1], Omer Levy[6], Avinatan Hassidim[2], Michael Brenner[2,7], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3] and Uri Hasson[1,2]

Departing from traditional linguistic models, advances in deep learning have resulted in a new type of predictive (autoregressive) deep language models (DLMs). Using a self-supervised next-word prediction task, these models generate appropriate linguistic responses in a given context. In the current study, nine participants listened to a 30-min podcast while their brain responses were recorded using electrocorticography (ECoG). We provide empirical evidence that the human brain and autoregressive DLMs share three fundamental computational principles as they process the same natural narrative: (1) both are engaged in continuous next-word prediction before word onset; (2) both match their pre-onset predictions to the incoming word to calculate post-onset surprise; (3) both rely on contextual embeddings to represent words in natural contexts. Together, our findings suggest that autoregressive DLMs provide a new and biologically feasible computational framework for studying the neural basis of language.

Alineamiento de embeddings con neuroimágenes

NeurIPS2021

Can fMRI reveal the representation of syntactic structure in the brain?

Aniketh Janardhan Reddy
Machine Learning Department
Carnegie Mellon University
ajreddy@cs.cmu.edu

Leila Wehbe
Machine Learning Department
Carnegie Mellon University
lwehbe@cmu.edu

Abstract

While studying semantics in the brain, neuroscientists use two approaches. One is to identify areas that are correlated with semantic processing load. Another is to find areas that are predicted by the semantic representation of the stimulus words. However, most studies of syntax have focused only on identifying areas correlated with syntactic processing load. One possible reason for this discrepancy is that representing syntactic structure in an embedding space such that it can be used to model brain activity is a non-trivial computational problem. Another possible reason is that it is unclear if the low signal-to-noise ratio of neuroimaging tools such as functional Magnetic Resonance Imaging (fMRI) can allow us to reveal the correlates of complex (and perhaps subtle) syntactic representations. In this study, we propose novel multi-dimensional features that encode information about the syntactic structure of sentences. Using these features and fMRI recordings of participants reading a natural text, we model the brain representation of syntax. First, we find that our syntactic structure-based features explain additional variance in the brain activity of various parts of the language system, even after controlling for complexity metrics that capture processing load. At the same time, we see that regions well-predicted by syntactic features are distributed in the language system and are not distinguishable from those processing semantics. Our code and data will be available at https://github.com/anikethjr/brain_syntactic_representations.

Low-Dimensional Structure in the Space of Language Representations is Reflected in Brain Responses

Richard Antonello
UT Austin
rjantonello@utexas.edu

Javier Turek
Intel Labs
javier.turek@intel.com

Vy Vo
Intel Labs
vy.vo@intel.com

Alexander Huth
UT Austin
huth@cs.utexas.edu

Abstract

How related are the representations learned by neural language models, translation models, and language tagging tasks? We answer this question by adapting an encoder-decoder transfer learning method from computer vision to investigate the structure among 100 different feature spaces extracted from hidden representations of various networks trained on language tasks. This method reveals a low-dimensional structure where language models and translation models smoothly interpolate between word embeddings, syntactic and semantic tasks, and future word embeddings. We call this low-dimensional structure a *language representation embedding* because it encodes the relationships between representations needed to process language for a variety of NLP (natural language processing) tasks. We find that this representation embedding can predict how well each individual feature space maps to human brain responses to natural language stimuli recorded using fMRI. Additionally, we find that the principal dimension of this structure can be used to create a metric which highlights the brain's natural language processing hierarchy. This suggests that the embedding captures some part of the brain's natural language representation structure.

Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics

Bhavin Choksi*
CerCo CNRS, UMR 5549 &
Université de Toulouse
bhavin.choksi@cnrs.fr

Milad Mozafari*
CerCo CNRS, UMR 5549 &
IRIT CNRS, UMR 5505
milad.mozafari@cnrs.fr

Callum Biggs O'May
CerCo CNRS
UMR 5549

Benjamin Ador
CerCo CNRS
UMR 5549

Andrea Alamia
CerCo CNRS
UMR 5549

Rufin VanRullen
CerCo CNRS, UMR 5549 &
ANITI, Université de Toulouse
rufin.vanrullen@cnrs.fr

Abstract

Deep neural networks excel at image classification, but their performance is far less robust to input perturbations than human perception. In this work we explore whether this shortcoming may be partly addressed by incorporating brain-inspired recurrent dynamics in deep convolutional networks. We take inspiration from a popular framework in neuroscience: "predictive coding". At each layer of the hierarchical model, generative feedback "predicts" (i.e., reconstructs) the pattern of activity in the previous layer. The reconstruction errors are used to iteratively update the network's representations across timesteps, and to optimize the network's feedback weights over the natural image dataset–a form of unsupervised training. We show that implementing this strategy into two popular networks, VGG16 and EfficientNetB0, improves their robustness against various corruptions and adversarial attacks. We hypothesize that other feedforward networks could similarly benefit from the proposed framework. To promote research in this direction, we provide an open-sourced PyTorch-based package called *Predify*, which can be used to implement and investigate the impacts of the predictive coding dynamics in any convolutional neural network.

Alineamiento de embeddings con neuroimágenes

Divergences between Language Models and Human Brains

Yuchen Zhou[1] Emmy Liu[1] Graham Neubig[1] Michael J. Tarr[1] Leila Wehbe[1]

Abstract

Do machines and humans process language in similar ways? Recent research has hinted in the affirmative, finding that brain signals can be effectively predicted using the internal representations of language models (LMs). Although such results are thought to reflect shared computational principles between LMs and human brains, there are also clear differences in how LMs and humans represent and use language. In this work, we systematically explore the divergences between human and machine language processing by examining the differences between LM representations and human brain responses to language as measured by Magnetoencephalography (MEG) across two datasets in which subjects read and listened to narrative stories. Using a data-driven approach, we identify two domains that are not captured well by LMs: social/emotional intelligence and physical commonsense. We then validate these domains with human behavioral experiments and show that fine-tuning LMs on these domains can improve their alignment with human brain responses [1].

Behavioral/Cognitive

Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual

Amanda LeBel,[1] Shailee Jain,[3] and Alexander G. Huth[2,3]
[1]Helen Wills Neuroscience Institute, University of California–Berkeley, Berkeley, California 94720, [2]Department of Neuroscience, University of Texas–Austin, Austin, Texas 78712, and [3]Department of Computer Science, University of Texas–Austin, Austin, Texas 78712

There is a growing body of research demonstrating that the cerebellum is involved in language understanding. Early theories assumed that the cerebellum is involved in low-level language processing. However, those theories are at odds with recent work demonstrating cerebellar activation during cognitive tasks. Using natural language stimuli and an encoding model framework, we performed an fMRI experiment on 3 men and 2 women, where subjects passively listened to 5 h of natural language stimuli, which allowed us to analyze language processing in the cerebellum with higher precision than previous work. We used these data to fit voxelwise encoding models with five different feature spaces that span the hierarchy of language processing from acoustic input to high-level conceptual processing. Examining the prediction performance of these models on separate BOLD data shows that cerebellar responses to language are almost entirely explained by high-level conceptual language features rather than low-level acoustic or phonemic features. Additionally, we found that the cerebellum has a higher proportion of voxels that represent social semantic categories, which include “social” and “people” words, and lower representations of all other semantic categories, including “mental,” “concrete,” and “place” words, than cortex. This suggests that the cerebellum is representing language at a conceptual level with a preference for social information.

Alineamiento de embeddings con neuroimágenes

A natural language fMRI dataset for voxelwise encoding models

Amanda LeBel[1], Lauren Wagner[2], Shailee Jain[3], Aneesh Adhikari-Desai[3,4], Bhavin Gupta[3], Allyson Morgenthal[4], Jerry Tang[3], Lixiang Xu[5] & Alexander G. Huth[3,4]

Speech comprehension is a complex process that draws on humans' abilities to extract lexical information, parse syntax, and form semantic understanding. These sub-processes have traditionally been studied using separate neuroimaging experiments that attempt to isolate specific effects of interest. More recently it has become possible to study all stages of language comprehension in a single neuroimaging experiment using narrative natural language stimuli. The resulting data are richly varied at every level, enabling analyses that can probe everything from spectral representations to high-level representations of semantic meaning. We provide a dataset containing BOLD fMRI responses recorded while 8 participants each listened to 27 complete, natural, narrative stories (~6 hours). This dataset includes pre-processed and raw MRIs, as well as hand-constructed 3D cortical surfaces for each participant. To address the challenges of analyzing naturalistic data, this dataset is accompanied by a python library containing basic code for creating voxelwise encoding models. Altogether, this dataset provides a large and novel resource for understanding speech and language processing in the human brain.

Scaling laws for language encoding models in fMRI

Richard J. Antonello
Department of Computer Science
The University of Texas at Austin
rjantonello@utexas.edu

Aditya R. Vaidya
Department of Computer Science
The University of Texas at Austin
avaidya@utexas.edu

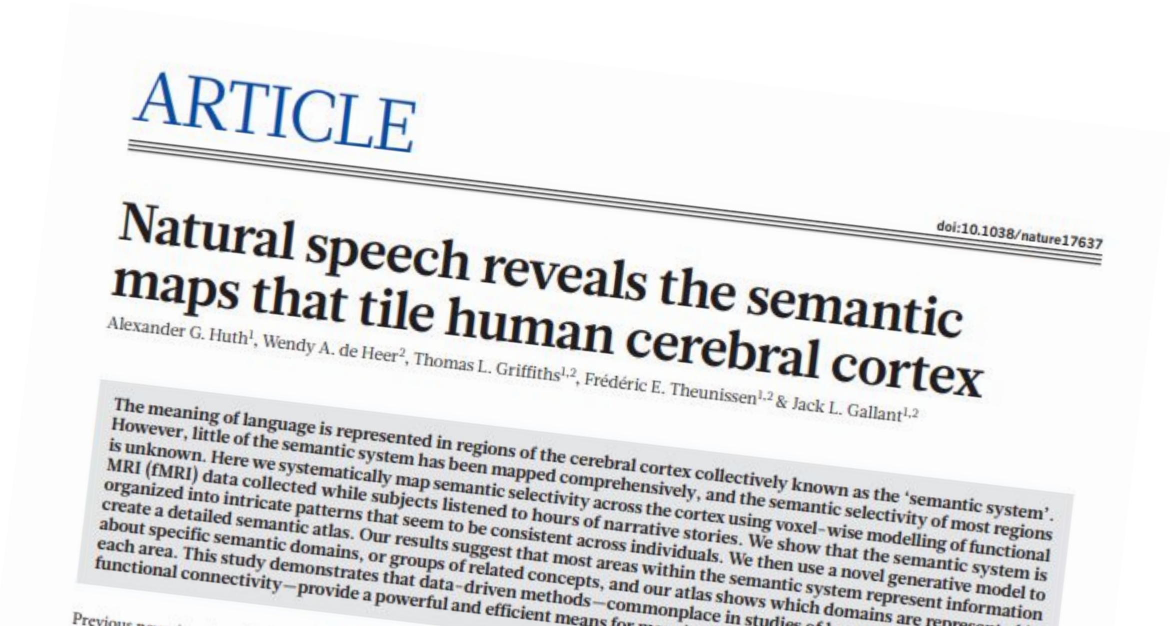
Alexander G. Huth
Departments of Computer Science and Neuroscience
The University of Texas at Austin
huth@cs.utexas.edu

Abstract

Representations from transformer-based unidirectional language models are known to be effective at predicting brain responses to natural language. However, most studies comparing language models to brains have used GPT-2 or similarly sized language models. Here we tested whether larger open-source models such as those from the OPT and LLaMA families are better at predicting brain responses recorded using fMRI. Mirroring scaling results from other contexts, we found that brain prediction performance scales logarithmically with model size from 125M to 30B parameter models, with $\sim$15% increased encoding performance as measured by correlation with a held-out test set across 3 subjects. Similar logarithmic behavior was observed when scaling the size of the fMRI training set. We also characterized scaling for acoustic encoding models that use HuBERT, WavLM, and Whisper, and we found comparable improvements with model size. A noise ceiling analysis of these large, high-performance encoding models showed that performance is nearing the theoretical maximum for brain areas such as the precuneus and higher auditory cortex. These results suggest that increasing scale in both models and data will yield incredibly effective models of language processing in the brain, enabling better scientific understanding as well as applications such as decoding.

Huth et al. (2016)

- Trabaja con estímulos naturales (podcasts)
- Alinea embeddings generados con term-by-term matrix
- Reduce la dimensionalidad para entender mejor los resultados



ARTICLE

doi:10.1038/nature17637

Natural speech reveals the semantic maps that tile human cerebral cortex

Alexander G. Huth[1], Wendy A. de Heer[2], Thomas L. Griffiths[1,2], Frédéric E. Theunissen[1,2] & Jack L. Gallant[1,2]

The meaning of language is represented in regions of the cerebral cortex collectively known as the 'semantic system'. However, little of the semantic system has been mapped comprehensively, and the semantic selectivity of most regions is unknown. Here we systematically map semantic selectivity across the cortex using voxel-wise modelling of functional MRI (fMRI) data collected while subjects listened to hours of narrative stories. We show that the semantic system is organized into intricate patterns that seem to be consistent across individuals. We then use a novel generative model to create a detailed semantic atlas. Our results suggest that most areas within the semantic system represent information about specific semantic domains, or groups of related concepts, and our atlas shows which domains are represented in each area. This study demonstrates that data-driven methods—commonplace in studies of

Huth et al. (2016)

Dataset de neuroimágenes:

- ➔ fMRI 3T
- ➔ 7 sujetos (5M, 2F)
- ➔ 4 horas por sujeto (2 sesiones)
- ➔ 11 caps de podcast (10-15 mins)
    - ◆ **Train:** 10 historias
    - ◆ **Test:** 1 historia (2 veces)

Huth et al. (2016)

Dataset de neuroimágenes:

- ➔ fMRI 3T
- ➔ 7 sujetos (5M, 2F)
- ➔ 4 horas por sujeto (2 sesiones)
- ➔ 11 caps de podcast (10-15 mins)
  - ◆ **Train:** 10 historias
  - ◆ **Test:** 1 historia (2 veces)

Construcción de embeddings:

- ➔ Matrix término-término
  - ◆ 10,470 filas
  - ◆ 985 columnas (pals más frecuentes)
- ➔ Co-Ocurrencia en ventana de 15 pals
- ➔ Agregan 41 dimensiones auditivas

Huth et al. (2016)

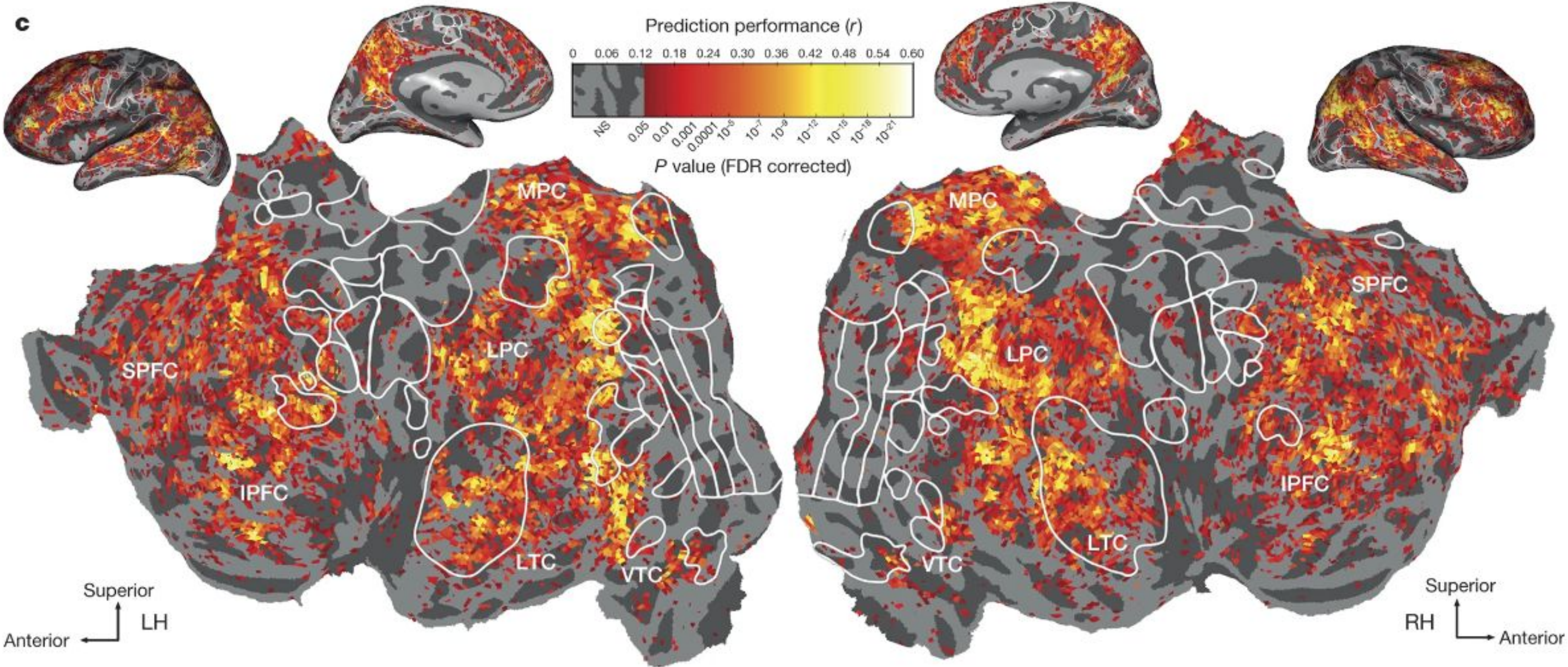
Estimación del modelo:

- ➔ Act ~ embedding
- ➔ Regularización Ridge
- ➔ Concatenación de palabras

Validación del modelo:

- ➔ Modelo entrenado prediciendo el capítulo de test
- ➔ Performance: pearson correlation entre las predicciones y la actividad neuronal

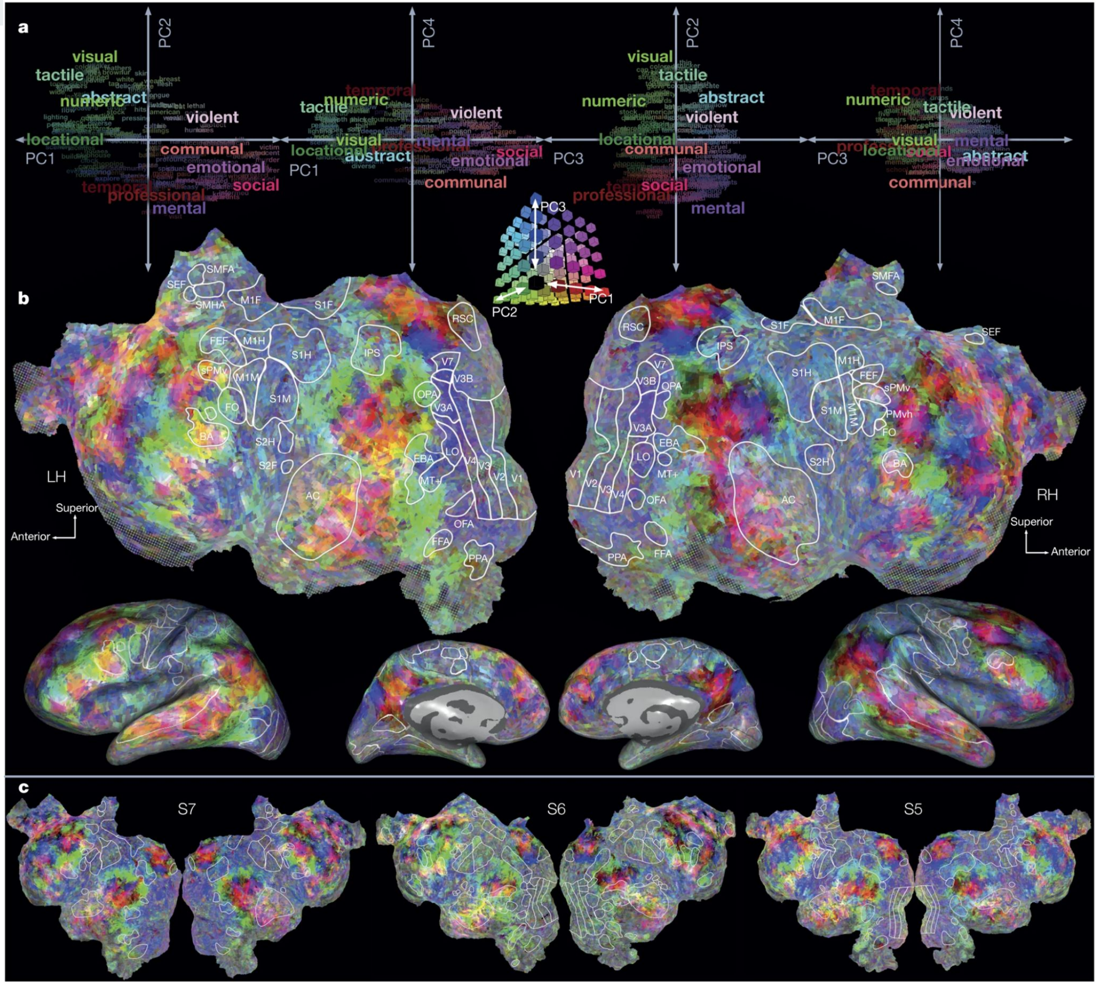
Huth et al. (2016)



Huth et al. (2016)

Explicabilidad:

- ➔ Análisis sobre los 10,000 mejores voxeles
- ➔ PCA sobre los 985 del modelo
- ➔ 4 dimensiones relevantes
- ➔ Mapeo de palabras a estas 4 dimensiones
- ➔ Clustering + labeling manual



Resumen

Hoy vimos:

- Introducción a Neuroimágenes
    - M/EEG
    - fMRI
- Técnica de alineamiento: Regresión lineal sobre embeddings
- Recapitulación de trabajos de los últimos años
- Huth et al. 2016

Hasta mañana!