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

Clase 3: Transformers (chatGPT)

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

[ELBA]

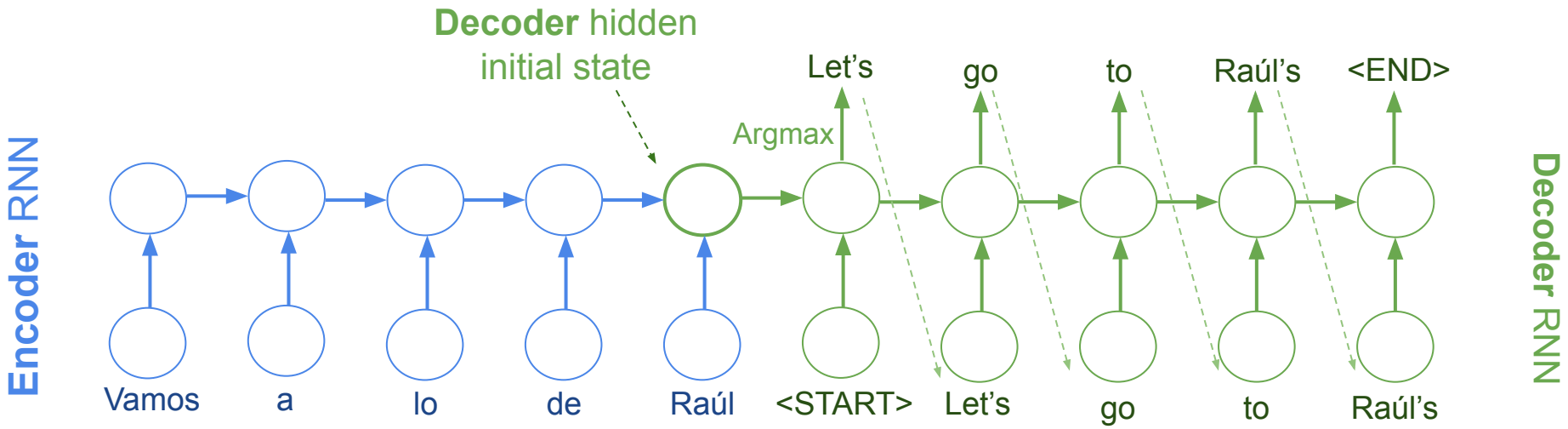
Julio 2024

Objetivos de esta clase

- Presentar la idea de auto-atención
- Transformers como sucesión de capas de atención
- Trabajos de neuro con transformers

Todo lo que necesitás es atención

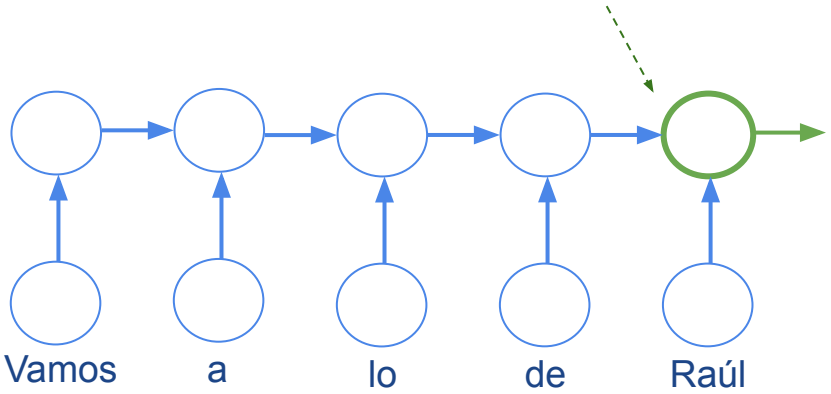
Todo lo que necesitás es atención



la red decoder es un modelo de lenguaje condicionado (por la red encoder)

Todo lo que necesitás es atención

Toda la información
de la secuencia
está en este vector

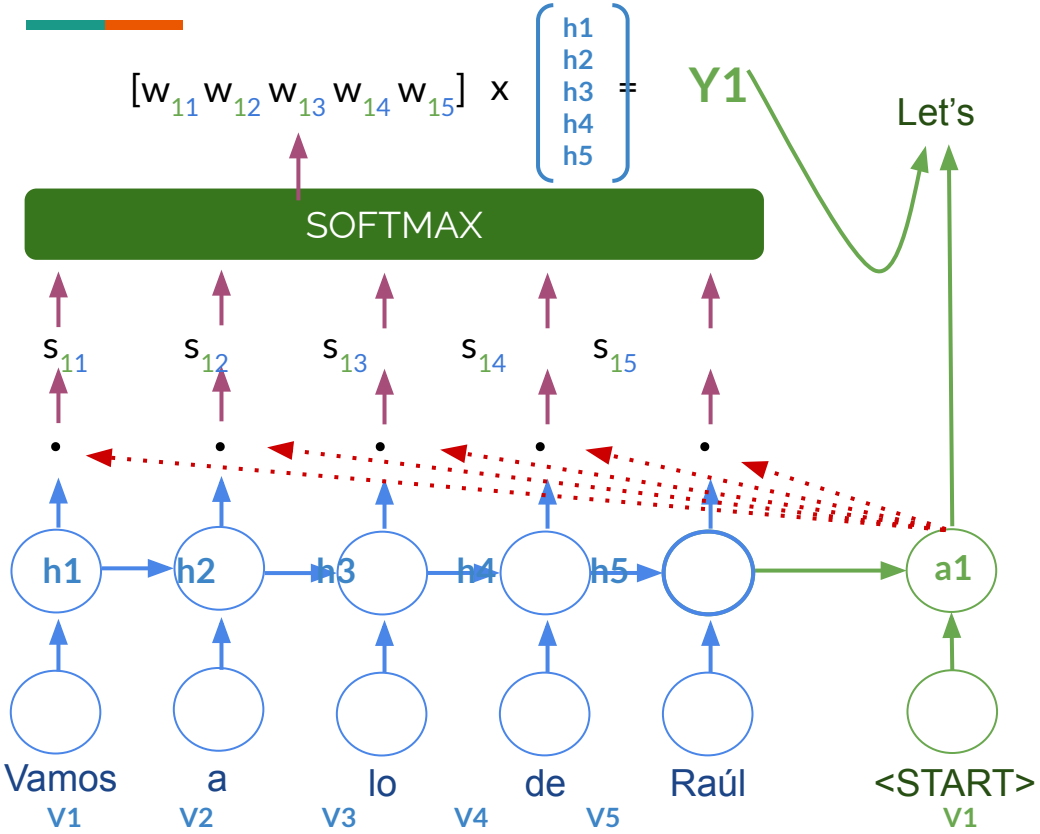


Encoder RNN

Esto es un cuello de botella

Si estamos en textos muy largos va a ser muy difícil codificar TODA la información importante en un solo vector

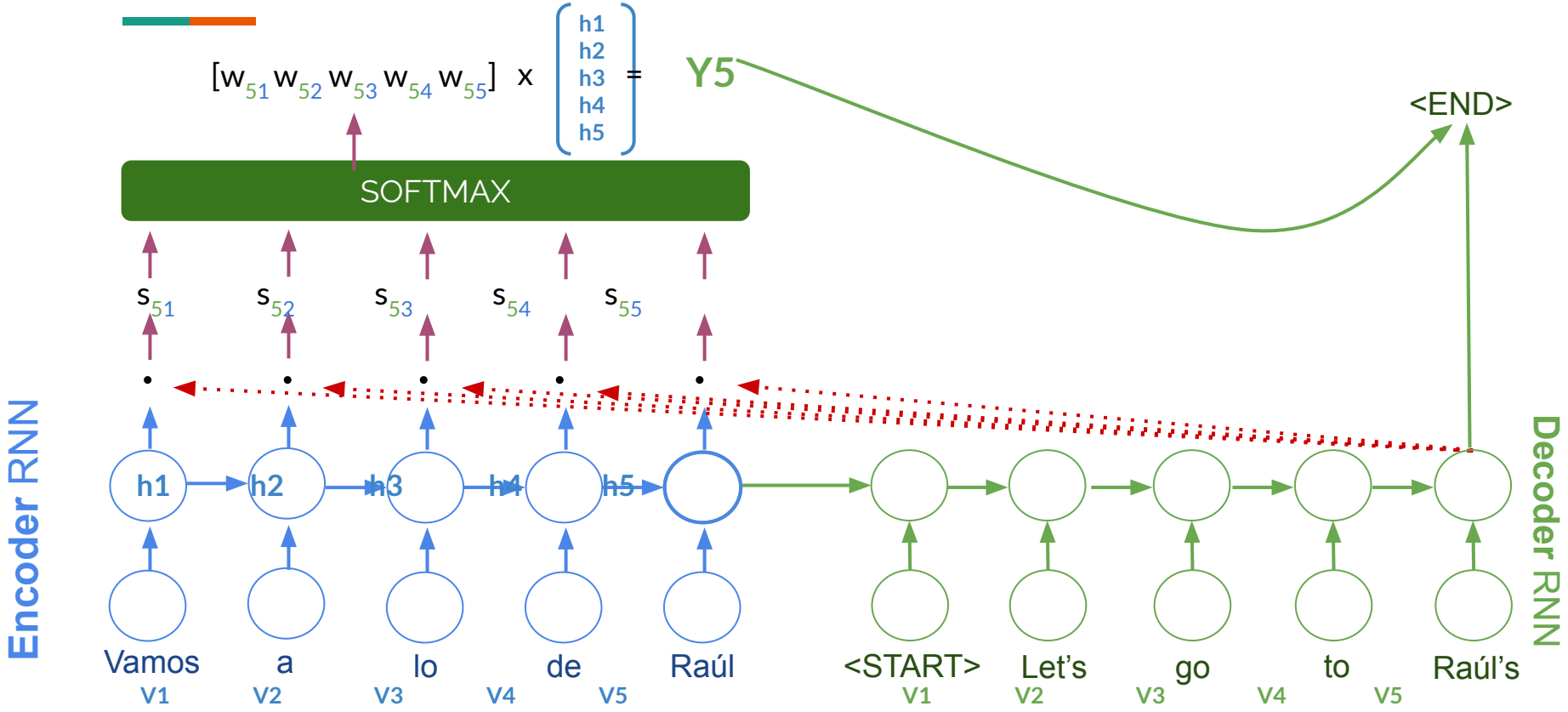
Todo lo que necesitás es atención



Encoder RNN

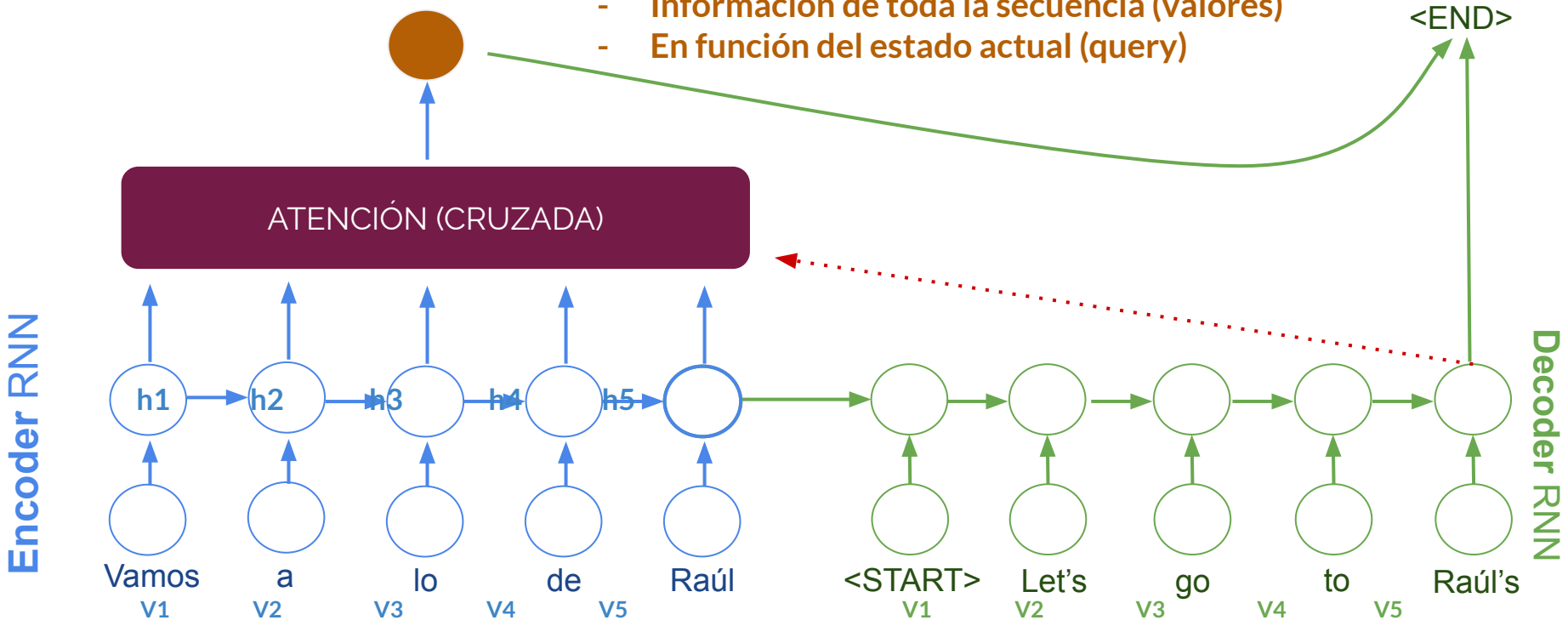
Decoder RNN

Todo lo que necesitás es atención



Todo lo que necesitás es atención

- Tamaño fijo
- Información de toda la secuencia (valores)
- En función del estado actual (query)



Todo lo que necesitás es atención

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

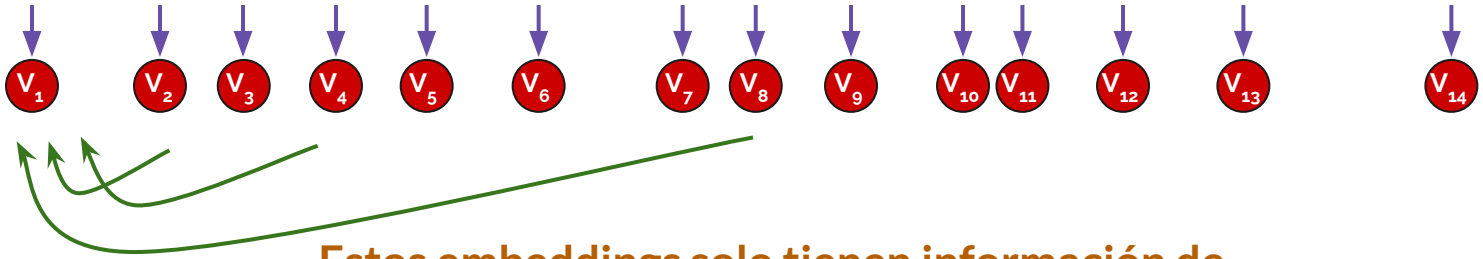
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT
to-German translation task, improving
ensembles, by

Self-Attention

River ganó y sumó 3 puntos que lo dejan en la punta del campeonato

Self-Attention

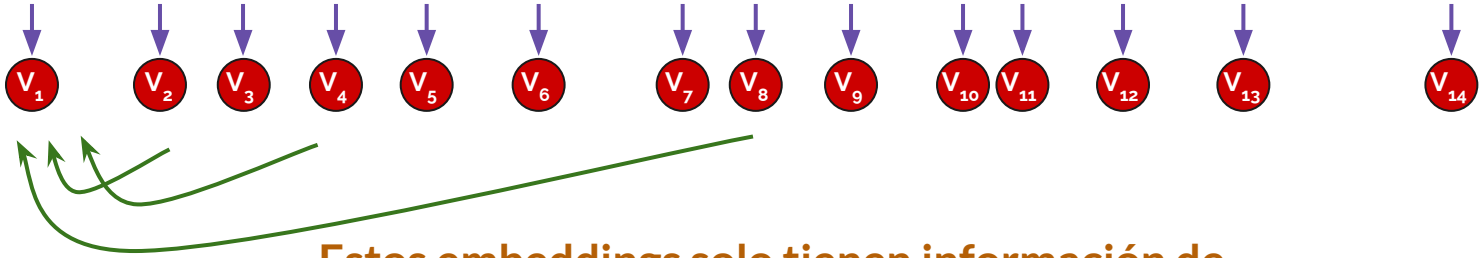
River ganó y sumó 3 puntos que lo dejan en la punta del campeonato



Estos embeddings solo tienen información de “la palabra” pero no de cómo se relacionan con el resto de las palabras de la oración

Self-Attention

River ganó y sumó 3 puntos que lo dejan en la punta del campeonato

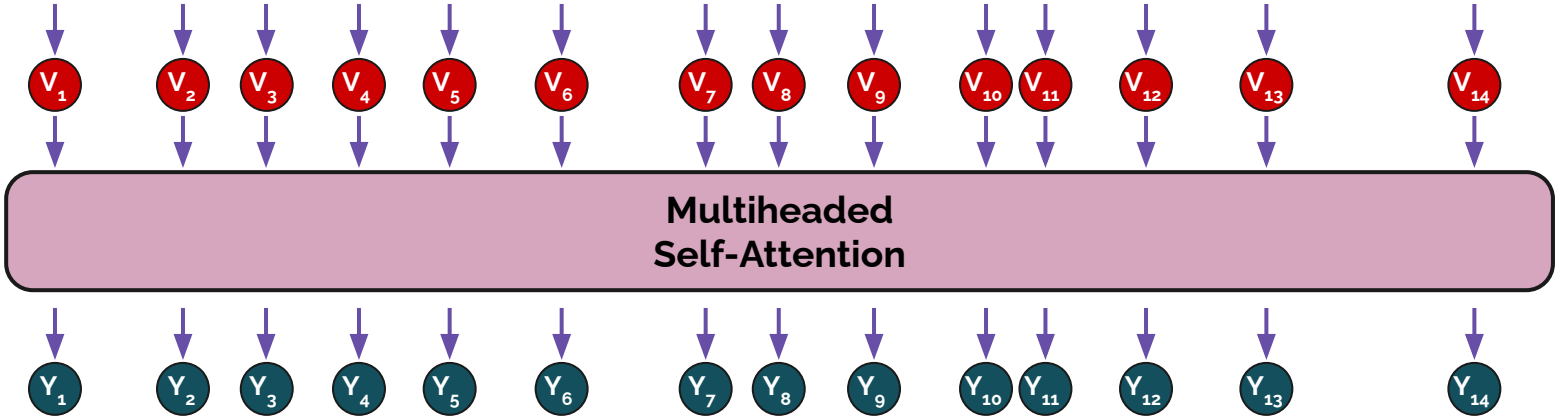


Estos embeddings solo tienen información de “la palabra” pero no de cómo se relacionan con el resto de las palabras de la oración

Apliquemos (auto) atención

Self-Attention

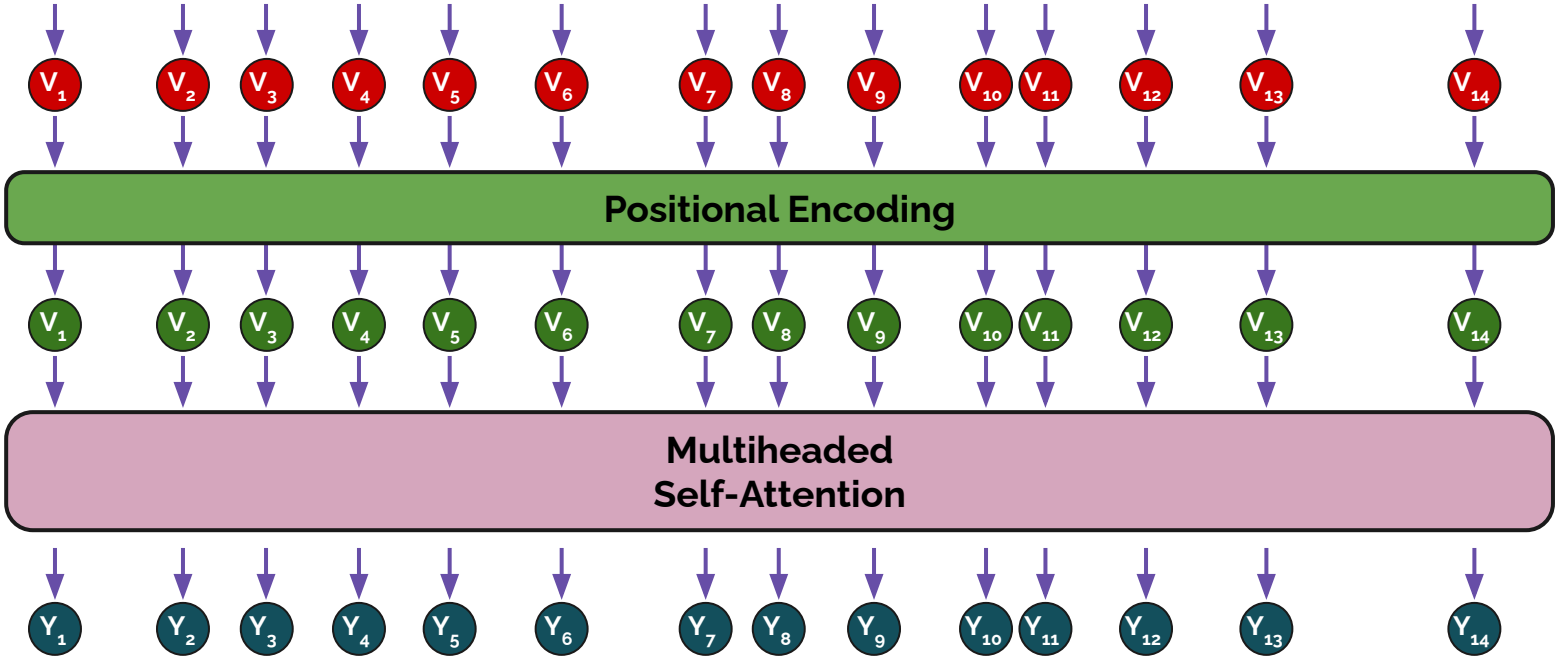
River ganó y sumó 3 puntos que lo dejan en la punta del campeonato



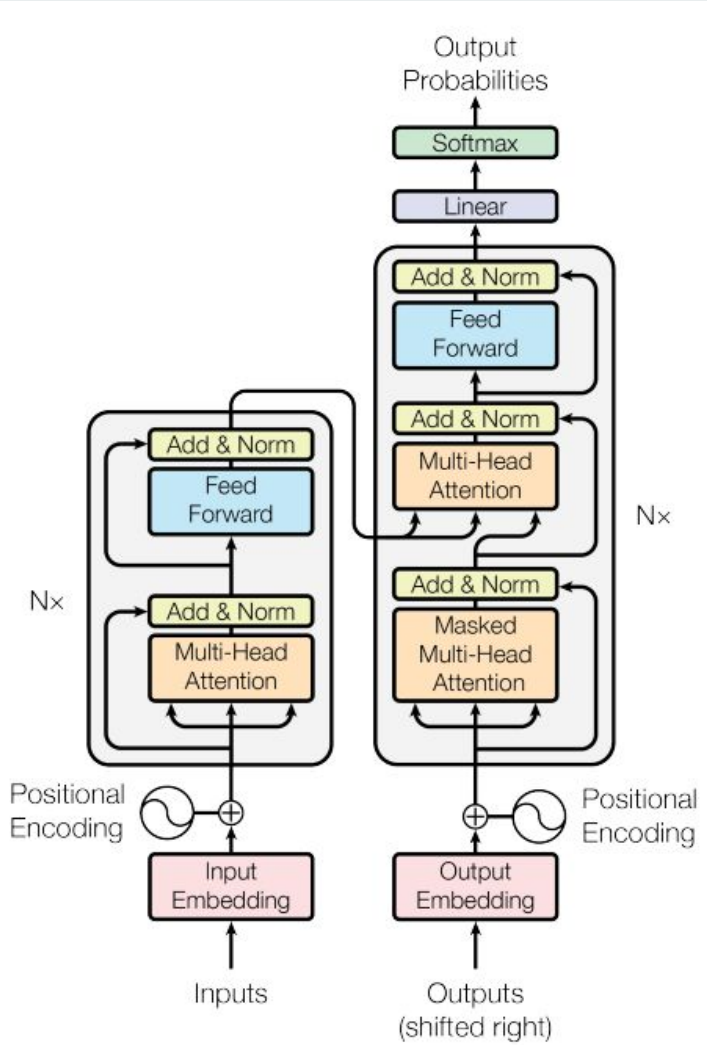
Ahora tenemos para cada palabra embeddings que tienen información del contexto de la oración

Positional Encoding

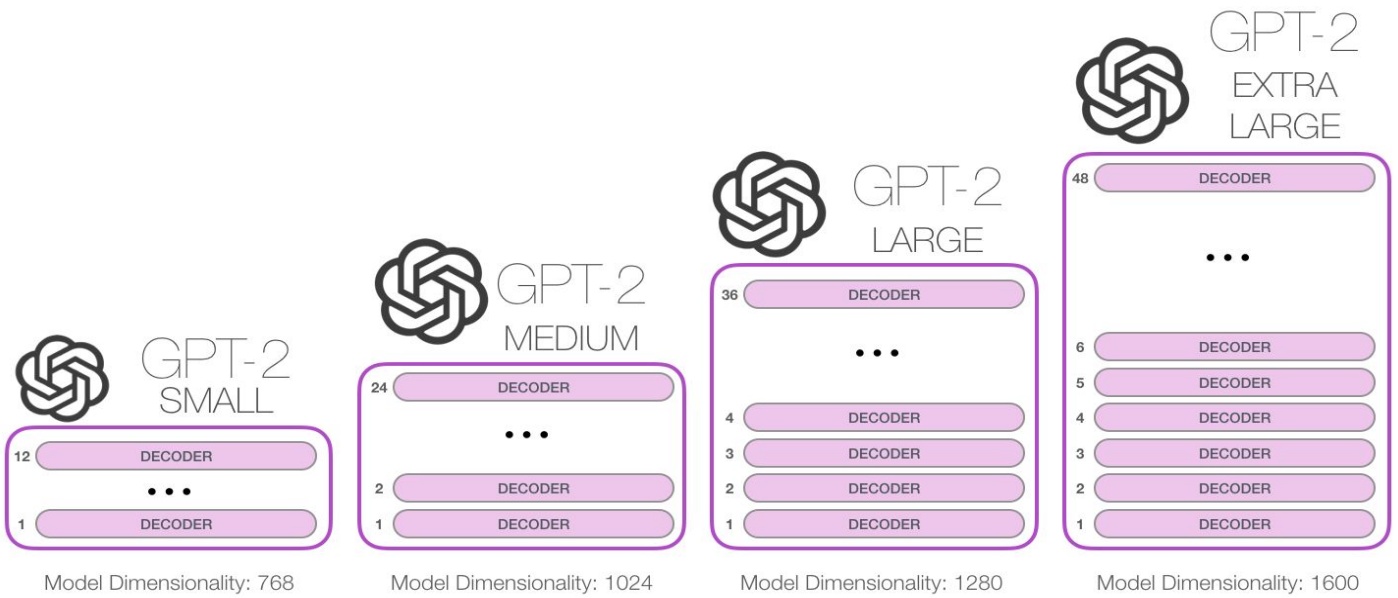
River ganó y sumó 3 puntos que lo dejan en la punta del campeonato



Transformer

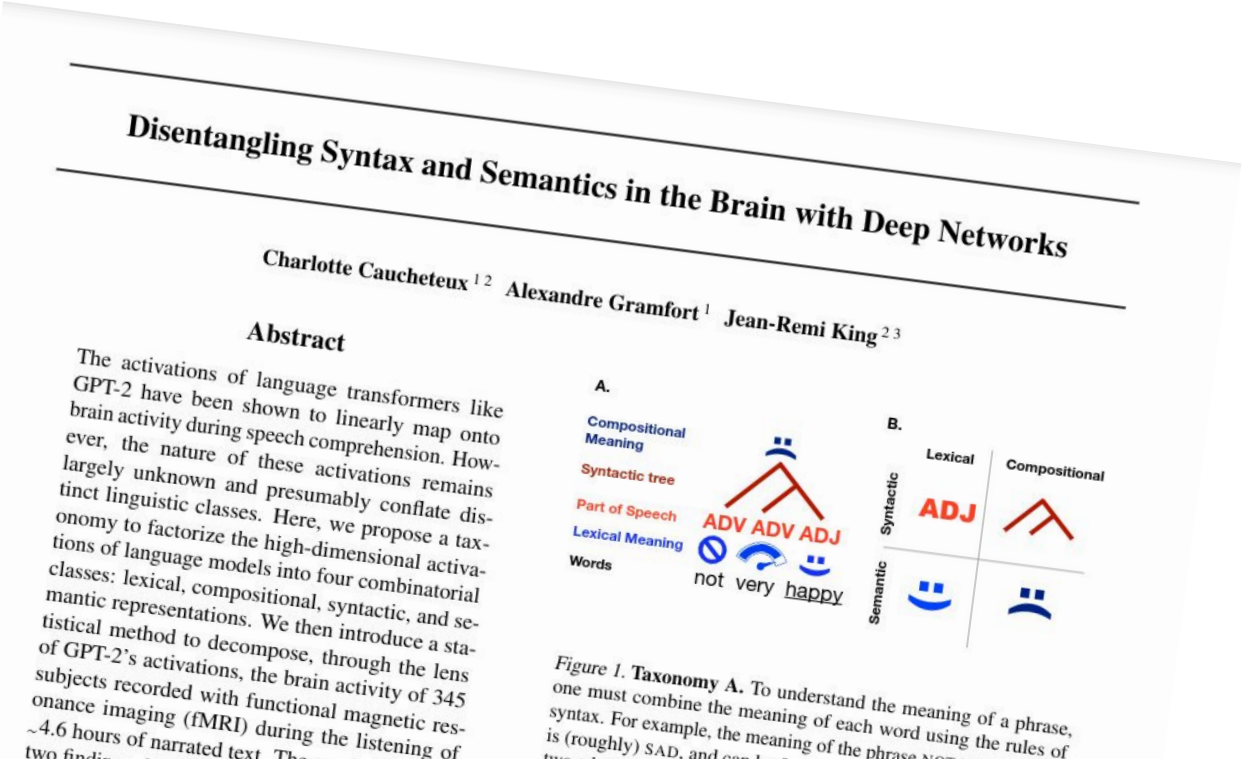


Transformer



A los papers!

Caucheteux and King, 2021



Disentangling Syntax and Semantics in the Brain with Deep Networks

Charlotte Caucheteux [1 2] Alexandre Gramfort [1] Jean-Remi King [2 3]

Abstract

The activations of language transformers like GPT-2 have been shown to linearly map onto brain activity during speech comprehension. However, the nature of these activations remains largely unknown and presumably conflate distinct linguistic classes. Here, we propose a taxonomy to factorize the high-dimensional activations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations. We then introduce a statistical method to decompose, through the lens of GPT-2's activations, the brain activity of 345 subjects recorded with functional magnetic resonance imaging (fMRI) during the listening of ~4.6 hours of narrated text. The

Figure 1. **Taxonomy A.** To understand the meaning of a phrase, one must combine the meaning of each word using the rules of syntax. For example, the meaning of the phrase NOT is (roughly) SAD, and can

Caucheteux and King, 2021

Dado que los embeddings actuales capturan más información además de la semántica, ¿podemos separar estas representaciones?

Propuesta:

- ➔ Tomar embeddings de GPT2 y separar:
  - ◆ Lexical
  - ◆ Compositional
  - ◆ Syntactic
  - ◆ Semantic
- ➔ Con las representaciones separadas
  - ◆ Hacen alignment sobre fMRI
    - 345 sujetos
    - 27 historias (4.6 hs)

Caucheteux and King, 2021

Definiciones:

- **Representación:** información que puede ser extraída de un vector de activaciones
  - **Lexical:** representaciones que son contexto independiente (embeddings de entrada).
  - **Composicional:** representaciones contextualizadas (embeddings del medio).
  - **Sintáctica:** representaciones asociadas solo a la estructura de la oración.
  - **Semántica:** todo lo que no sea sintáctico.

Proponen una forma de calcular esto

Puede ser Léxicos o Composicionales

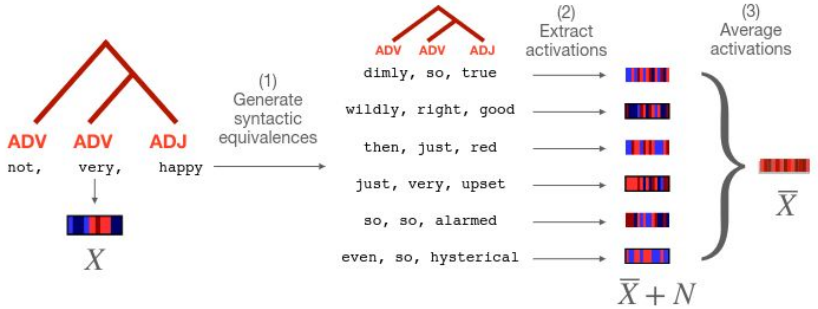
Caucheteux and King, 2021

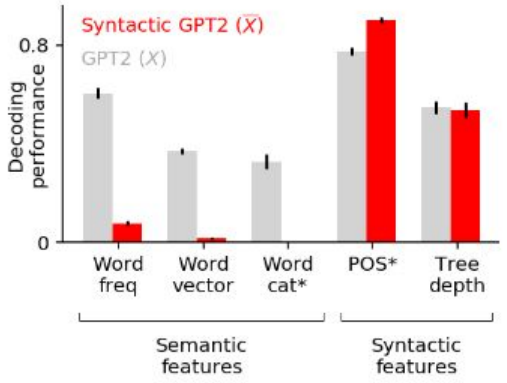
Representaciones Sintácticas:

- ➔ Para cada oración de su experimento general 10 oraciones con la misma estructura sintáctica
- ➔ Pasan todas estas oraciones por el modelo de lenguaje que vayan a usar
- ➔ El promedio de las activaciones para una determinada capa es la **presentación sintáctica**

Representaciones Semánticas:

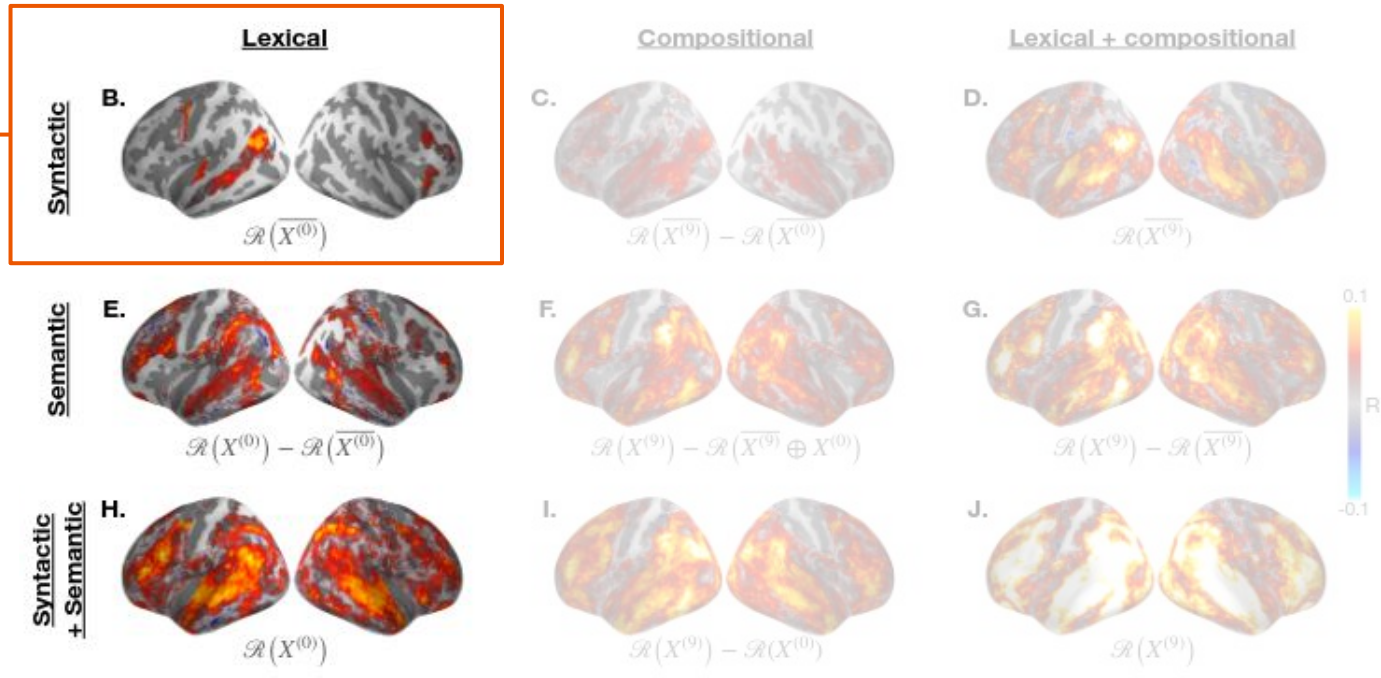
- ➔ Para una determinada palabra, es la resta de su representación sintáctica de su representación léxica o composicional





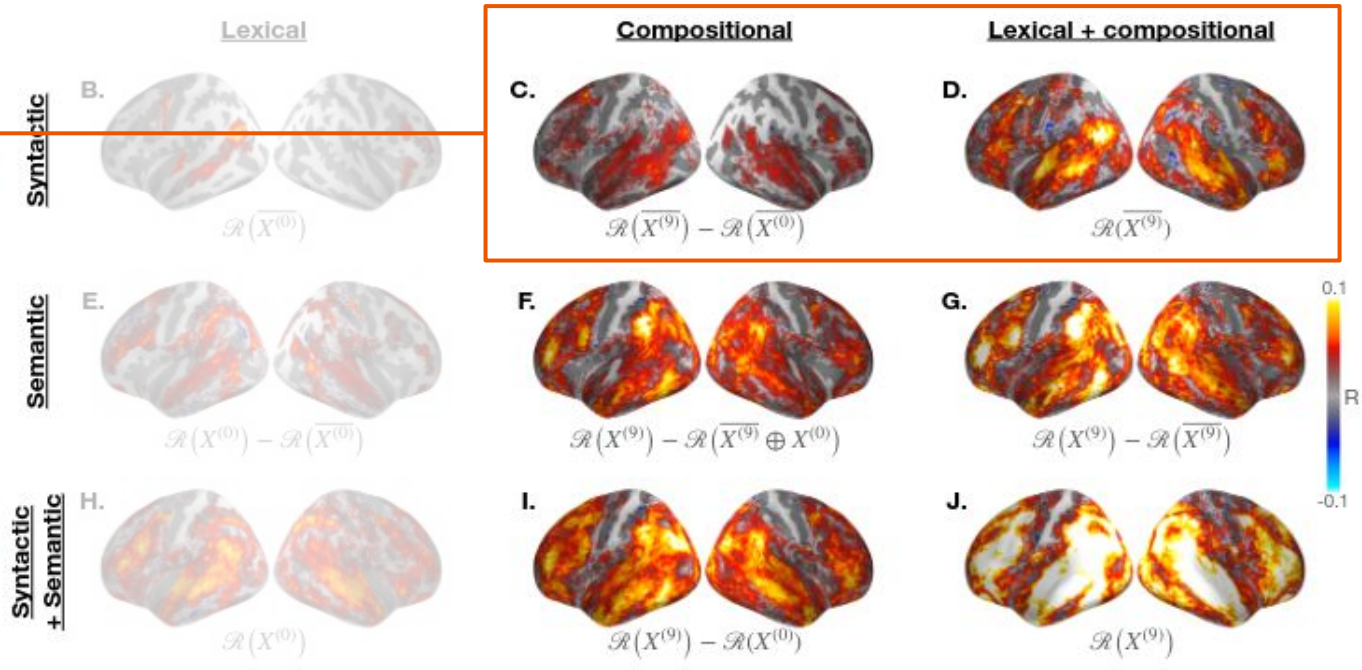
Caucheteux and King, 2021

Hay algo sintáctico en los embeddings lexicales



Caucheteux and King, 2021

El procesamiento sintáctico está distribuído



Caucheteux and King, 2022

communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-03036-1 OPEN

Check for updates

Brains and algorithms partially converge in natural language processing

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. Finally, we estimate how the architecture, training, and performance of these models independently account for the generation of brain-like representations. Our

Caucheteux and King, 2022

¿Qué embeddings correlacionan mejor con la actividad cerebral?

Caucheteux and King, 2022

¿Qué embeddings correlacionan mejor con la actividad cerebral?

Generación de embeddings:

- ➔ Entrenan 36 transformers
  - 18 CLM, 18 MLM
  - 3x largo de embeddings [128, 256, 512]
  - 3x cantidad de capas [4, 8, 12]
  - 2x cantidad de cabezas [4, 8]
- ➔ Analizan todos los pasos de entrenamiento (100)
- ➔ Analizan todas las capas (324)

Corpus de neuroimágenes:

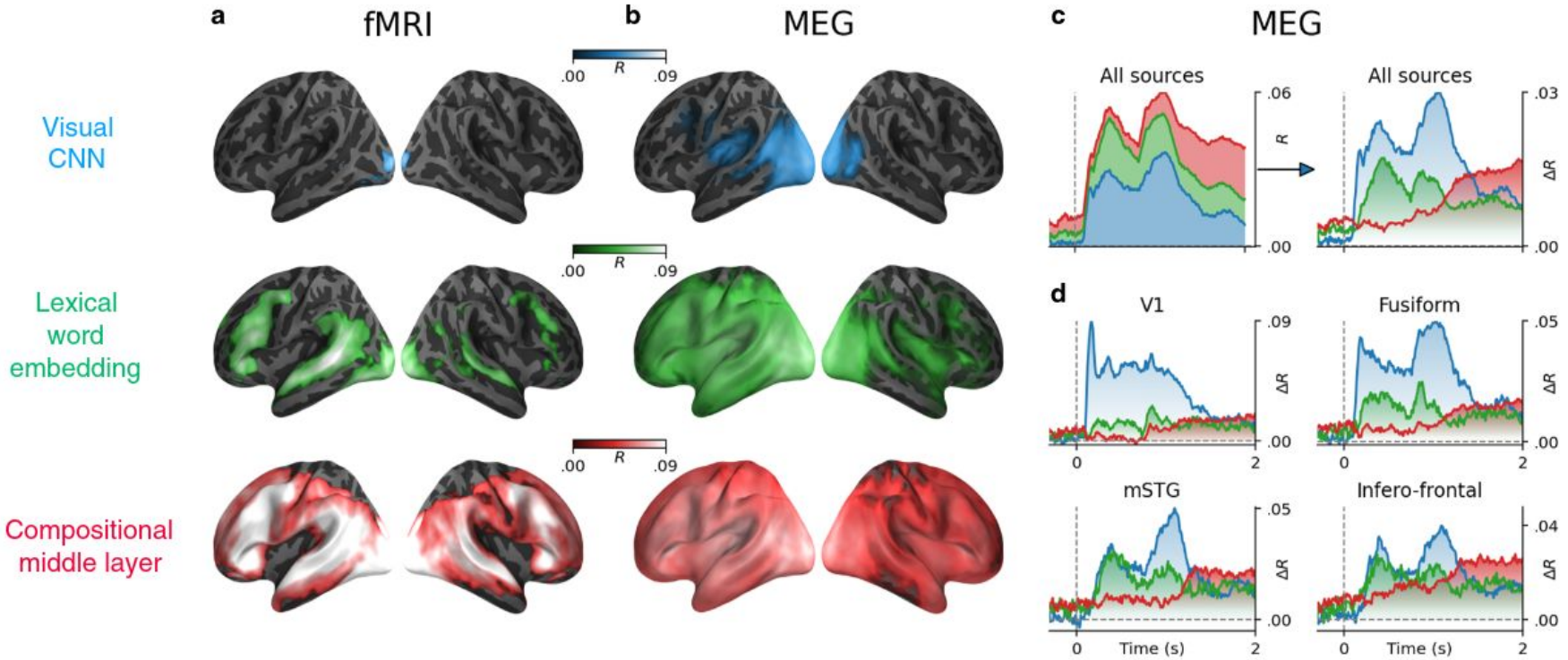
- ➔ fMRI y MEG
- ➔ 102 holandeses
- ➔ Lectura de 1 palabra por vez
- ➔ Oraciones de 9-15 palabras
- ➔ 2700 palabras por sujeto

Caucheteux and King, 2022

Hierarchy of neural responses originating in V1 around 100 ms and continuing within the left posterior fusiform gyrus around 200 ms, the superior and middle temporal gyri, as well as the pre-motor and infero-frontal cortices between 150 and 500 ms after word onset

Caucheteux and King, 2022

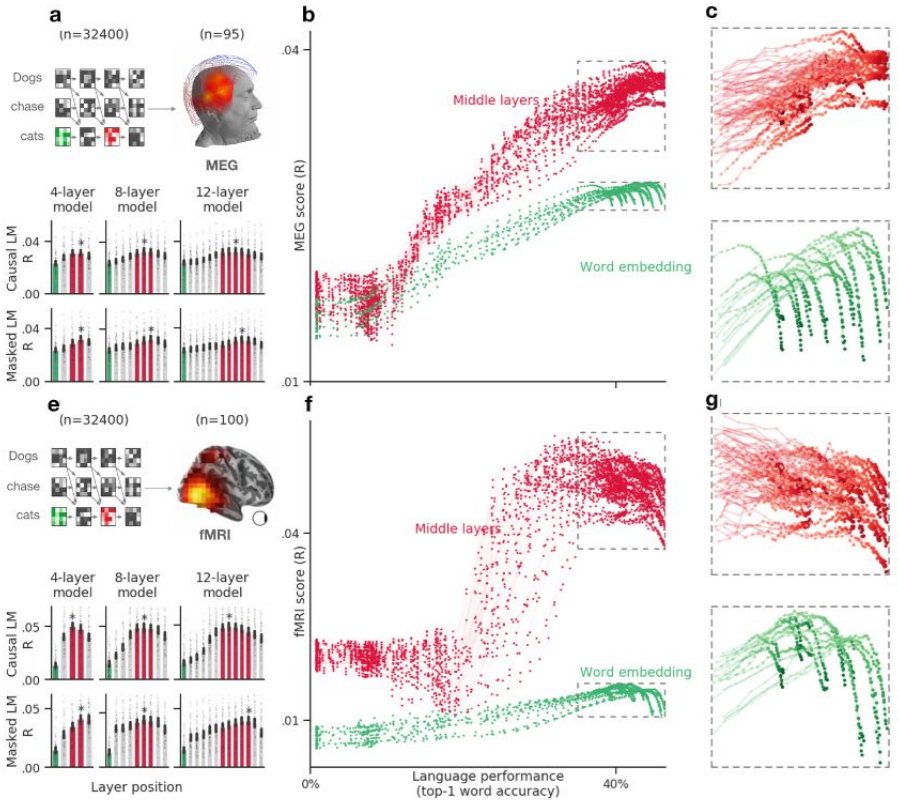
the brain-mapping of our three representative embeddings automatically recovers the hierarchy of visual, lexical, and compositional representations of language in each cortical region.



Caucheteux and King, 2022

Qué modelos/capas funcionan mejor:

- ➔ Las capas del medio son las mejores que las primeras y que las últimas
- ➔ Mejores modelos, mas parecidos al cerebro
- ➔ Los que tienen mayor performance de predicción, son un poquito peores



Hasta mañana!