No todo lo que necesitas es atención

(Attention is not all you need)

Trabajo Fin de Grado

Grado en Matemáticas

Diego Rivera López-Brea

20 de septiembre de 2022

Contenidos:

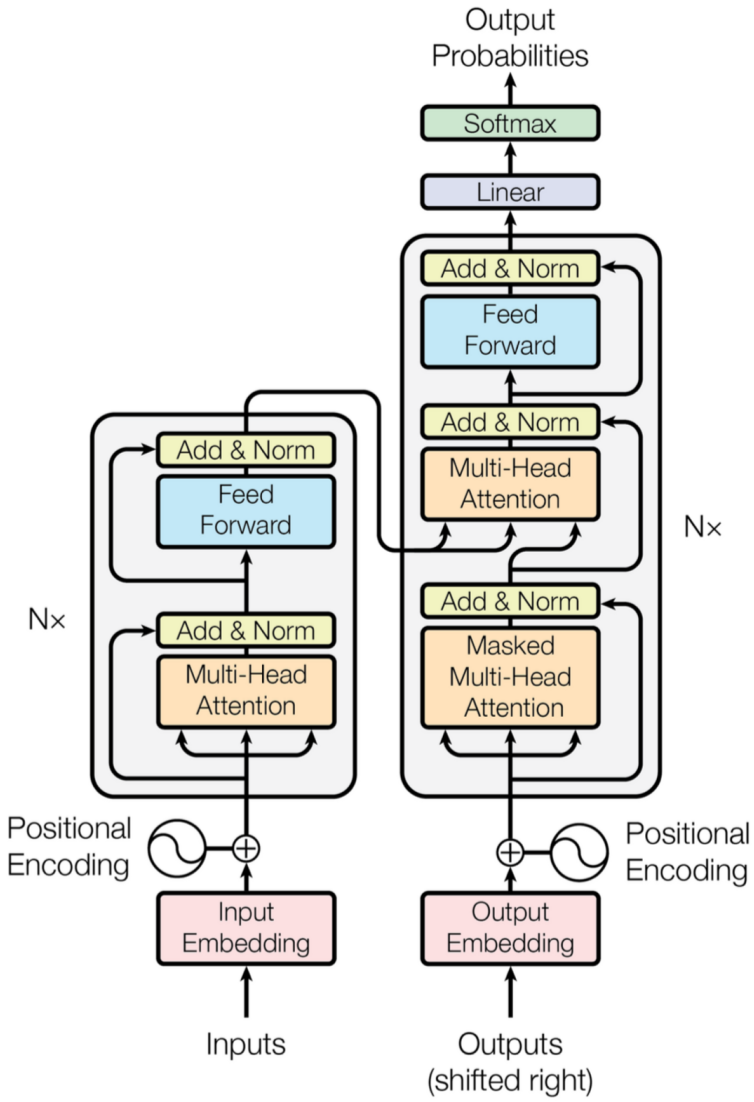
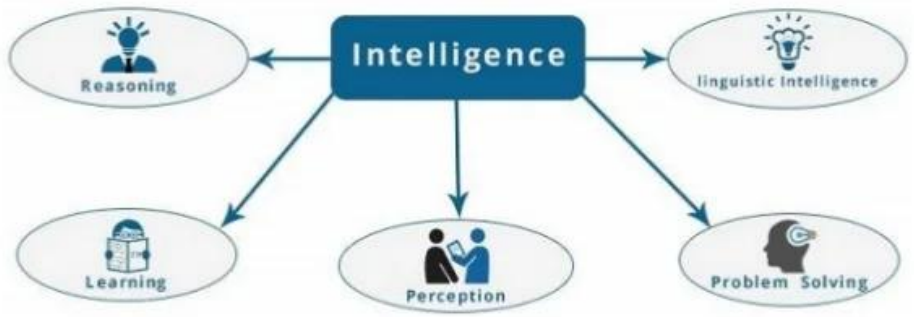
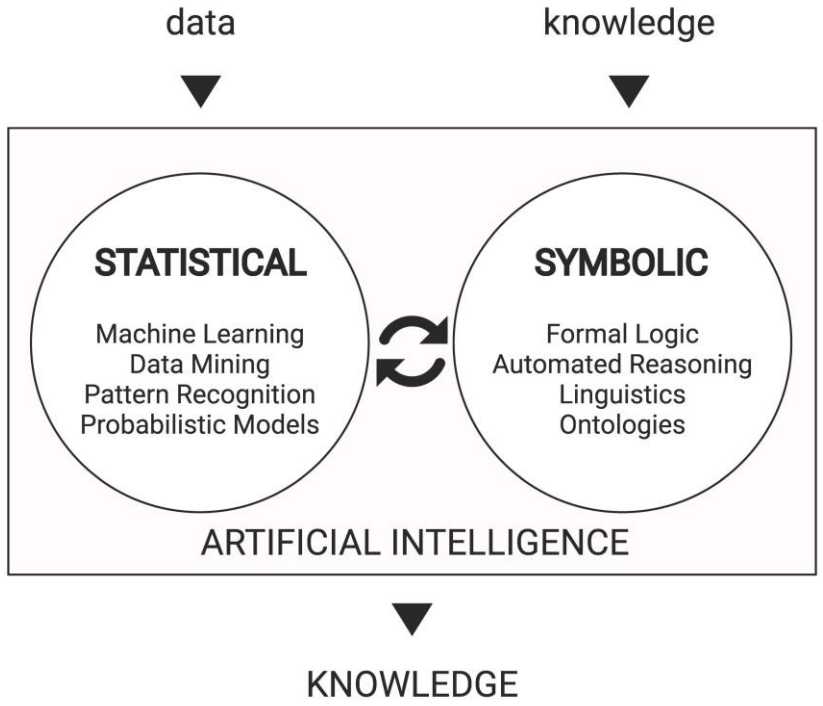


Figure 1: The Transformer - model architecture.

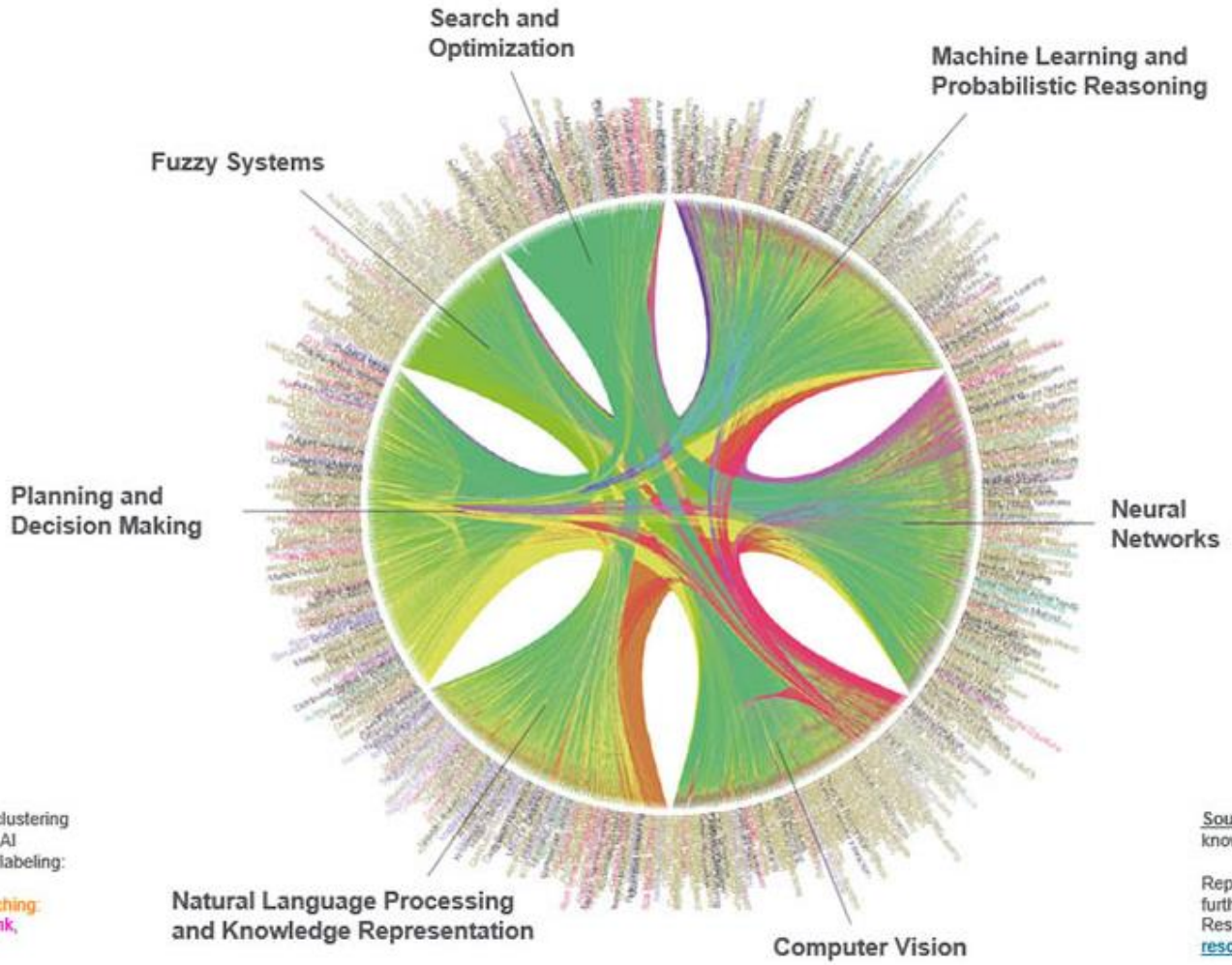
Artificial Intelligence Composed

- **The intelligence is intangible. It is composed of :-**
  - **Reasoning**
  - **Learning**
  - **Problem Solving**
  - **Perception**
  - **Linguistic Intelligence**





The AI research field clusters around seven main research areas



Using keyword co-occurrences with unsupervised clustering on article-level, based on keywords extracted from AI documentation of different actors and using expert labeling:

The color of the keyword represents its origin: Teaching: orange, Industry: green, Research: blue, Media: pink, multiple: black

Source: Elsevier, 2018, Artificial Intelligence: how knowledge is created, transferred, and used

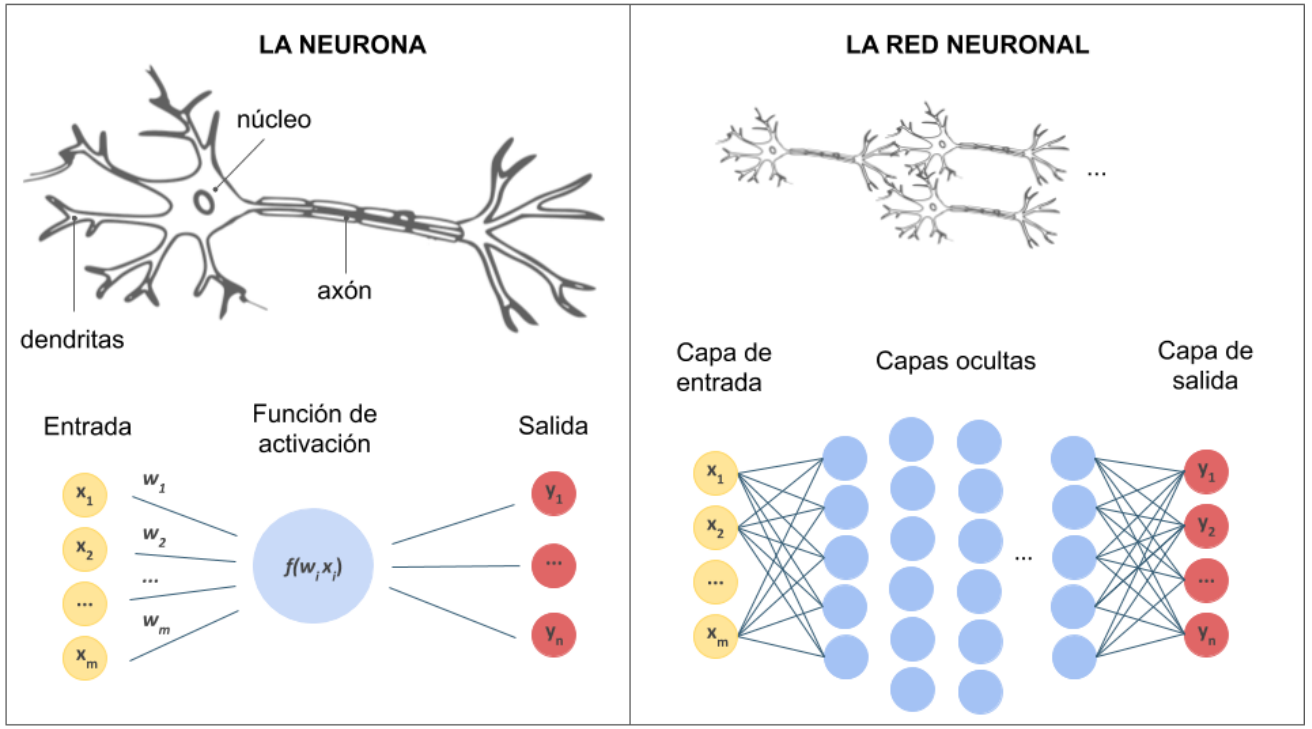
Report will be launched: Dec. 11th, 2018. Download and further information from Elsevier's Artificial Intelligence Resource Center: https://www.elsevier.com/connect/ai-resource-center

Modus operandi

1) Se propone modelo hipótesis*:*
   - Definido por unos **hiperparámetros** (ej. Grado polinomio).
   - Acompañado por unos **parámetros** a entrenar.
2) **Estimador de Máxima Verosimilitud** sobre subconjunto de **entrenamiento***.*
3) Subconjunto de **validación cruzada**: ¿mala generalización? → ajuste hiperparámetros.
4) Subconjunto de **test**: ¿mala generalización? → cambio de hipótesis.

Redes neuronales (NN)

1) Modelo hipótesis basado (por lo menos remotamente) en la estructura del cerebro, NO en la naturaleza de los datos.
2) Teoremas de Aproximación Universal.
3) Fáciles de paralelizar → *Deep learning.*
4) Funcionamiento de la **memoria**.



Procesamiento de Lenguaje Natural (NLP)

- Chomsky:
  - Distinción entre sintáctica y semántica.
  - Gramática innata vs teorías ambientalistas (marco para las NN).
- NN buenos resultados reproduciendo gramática, pero sin “sentido común” para la semántica.

Procesamiento de Lenguaje Natural (NLP)

- **Tokenización**: palabras, caracteres, o subpalabras.
- Representaciones vectoriales de tókenes.
- *Bags of words.*
- Representaciones contextuales: modelos secuenciales.
  - *RNNs*: comprobación manual, corta memoria y difícil paralelización.
  - *Transformers*.

Transformers. Mecanismo de atención

- Atención por token.
- Vector *query* → para el token sobre el que se realiza la atención.
- Vectores *key* y *value* → para cada token de contexto.

Algoritmo 1: Atención básica para un único token.

Input: $\boldsymbol{e} \in \mathbb{R}^{d_{in}}$, *input embedding* del token a representar x_o.
Input: $\boldsymbol{e}_t \in \mathbb{R}^{d_{in}}$, *input embedding* del token de contexto $z_t \in T$.
Output: $\tilde{\boldsymbol{v}} \in \mathbb{R}^{d_{out}}$, la representación vectorial conjunta del token x_o y su contexto T.
Parámetros: $\boldsymbol{W_q}, \boldsymbol{W_k} \in \mathbb{R}^{d_{attn} \times d_{in}}$, $\boldsymbol{b_q}, \boldsymbol{b_k} \in \mathbb{R}^{d_{attn}}$; proyecciones lineales asociadas a *query* y a *key*.
Parámetros: $\boldsymbol{W_v} \in \mathbb{R}^{d_{out} \times d_{in}}$, $\boldsymbol{b_v} \in \mathbb{R}^{d_{out}}$; la proyección lineal asociada a *value*

1 $\boldsymbol{q} \longleftarrow \boldsymbol{W_q e} + \boldsymbol{b_q}$
2 $\forall t \in [N_T] : \boldsymbol{k}_t \longleftarrow \boldsymbol{W_k e}_t + \boldsymbol{b_k}$
3 $\forall t \in [N_T] : \boldsymbol{v}_t \longleftarrow \boldsymbol{W_v e}_t + \boldsymbol{b_v}$
4 $\forall t \in [N_T] : \alpha_t = \frac{\exp\left(\boldsymbol{q}^T \boldsymbol{k}_t / \sqrt{d_{attn}}\right)}{\sum_u \exp\left(\boldsymbol{q}^T \boldsymbol{k}_u / \sqrt{d_{attn}}\right)}$
5 return $\tilde{\boldsymbol{v}} = \sum_{t=1}^{T} \alpha_t \boldsymbol{v}_t$

Transformers. Mecanismc

Algoritmo 2: $\tilde{\boldsymbol{V}} \longleftarrow$ Atención$(\boldsymbol{X}, \boldsymbol{Z}|\mathcal{W}_{qkv}, \text{Mask})$

/* Calcula una única *head* de *self-* o *cross-attention* (*con máscara*). */

Input: $\boldsymbol{X} \in \mathbb{R}^{d_X \times l_X}$, $\boldsymbol{Z} \in \mathbb{R}^{d_Z \times l_Z}$ los *input embeddings* de la secuencias primaria y de contexto.

Output: $\tilde{\boldsymbol{V}} \in \mathbb{R}^{d_{out} \times l_X}$ la representación vectorial conjunta de los tókenes en $\boldsymbol{X}$ y su contexto $\boldsymbol{Z}$.

Parámetros: $\mathcal{W}_{qkv}$ formada por: $\boldsymbol{W_q} \in \mathbb{R}^{d_{attn} \times d_X}$, $\boldsymbol{b_q} \in \mathbb{R}^{d_{attn}}$.
$\boldsymbol{W_k} \in \mathbb{R}^{d_{attn} \times d_Z}$, $\boldsymbol{b_k} \in \mathbb{R}^{d_{attn}}$.
$\boldsymbol{W_v} \in \mathbb{R}^{d_{out} \times d_Z}$, $\boldsymbol{b_v} \in \mathbb{R}^{d_{out}}$.

Hiperparámetros: Mask $\in \{0,1\}^{l_Z \times l_X}$

1 $\boldsymbol{Q} \longleftarrow \boldsymbol{W_q X} + \boldsymbol{b_q} \boldsymbol{1}^T$, ▷ $\boldsymbol{Query} \in \mathbb{R}^{d_{attn} \times l_X}$

2 $\boldsymbol{K} \longleftarrow \boldsymbol{W_k Z} + \boldsymbol{b_k} \boldsymbol{1}^T$, ▷ $\boldsymbol{Key} \in \mathbb{R}^{d_{attn} \times l_Z}$

3 $\boldsymbol{V} \longleftarrow \boldsymbol{W_v Z} + \boldsymbol{b_v} \boldsymbol{1}^T$, ▷ $\boldsymbol{Value} \in \mathbb{R}^{d_{out} \times l_Z}$

4 $\boldsymbol{S} \longleftarrow \boldsymbol{K}^T \boldsymbol{Q}$, ▷ $\boldsymbol{Score} \in \mathbb{R}^{l_Z \times l_X}$

5 $\forall t_Z, t_X$ si $\text{Mask}[t_X, t_Z] \neq 1$ entonces $\boldsymbol{S}[t_X, t_Z] \longleftarrow -\infty$

6 **return** $\tilde{\boldsymbol{V}} = \boldsymbol{V} \cdot \text{softmax}(\boldsymbol{S}/\sqrt{d_{attn}})$

- Atención para una secuencia de tókenes.
  - Secuencia primaria ***X,*** y de contexto ***Z.***
  - Softmax y Mask.
- Tipos de atención:
  - Atención unidireccional (*Unidirectional/masked self-attention*) → ***X=Z,*** *Mask=*$[[t_Z \leq t_X]]$.
  - Atención bidireccional (*Bidirectional/Unmasked self-attention*)→ ***X=Z,*** *Mask=1.*
  - Atención cruzada (*Cross-attention*) → ***X≠Z,*** *Mask=1.*
  - Atención multi-head (*Multi-head attention*).

Transformers: Encoder-Decoder original

- **Objetivo**: entrenar en predecir secuencia de tókenes ***x*** dada otra secuencia de tókenes original ***z***.

- **Núcleo**: varios bloques de atención distribuidos en dos módulos: un *encoder* y un *decoder*.

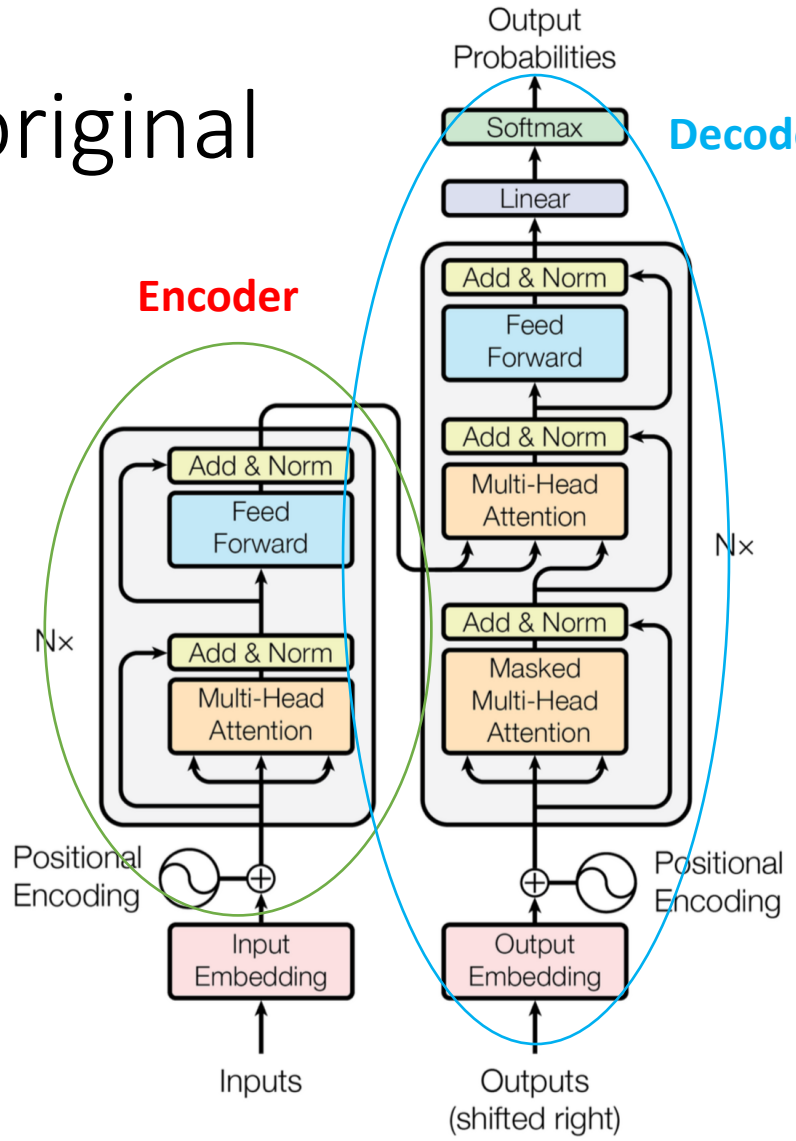


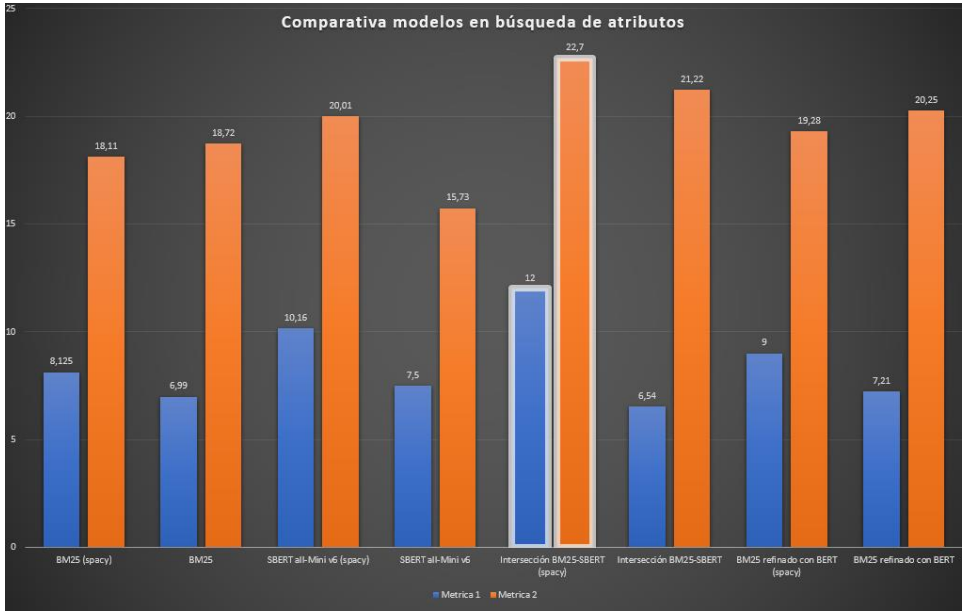
Figure 1: The Transformer - model architecture.

BERT y GPT

- El entrenamiento de un *transformer* se centra en predecir un solo token.
- *Bidirectional encoders* → BERT, RoBERTa
- *Autoregressive decoders* → GPT, Gopher

SBERT y aplicación práctica

- *Sentence similarity: cross-encoders vs bi-encoders.*
- Búsqueda de atributos más relevantes para una lista de productos.



Valoración

1) Muchas más información lingüística que cualquier ser humano pero peores resultados.

2) Los *transformers* SÍ que están basados en la naturaleza del lenguaje, además de la estructura neuronal del cerebro. Pero no es suficiente.

3) Posible vía para reflejar "sentido común"→ Causalidad.

Causalidad

- Una causa no es una “correlación fuerte”.
- Los datos son resultados de una jerarquía causal invisible, pero necesaria para una descripción correcta de la realidad.
- Jerarquía causal representada por grafos acíclicos: diagramas causales.

Escalera de causalidad

1) Estadística: observación de datos.
2) Intervención: nuevo operador en probabilidad: *do()*
3) Razonamiento contrafáctico: ¿qué pasaría si en vez de A fuese B? → Imaginación (información desde datos no existentes)

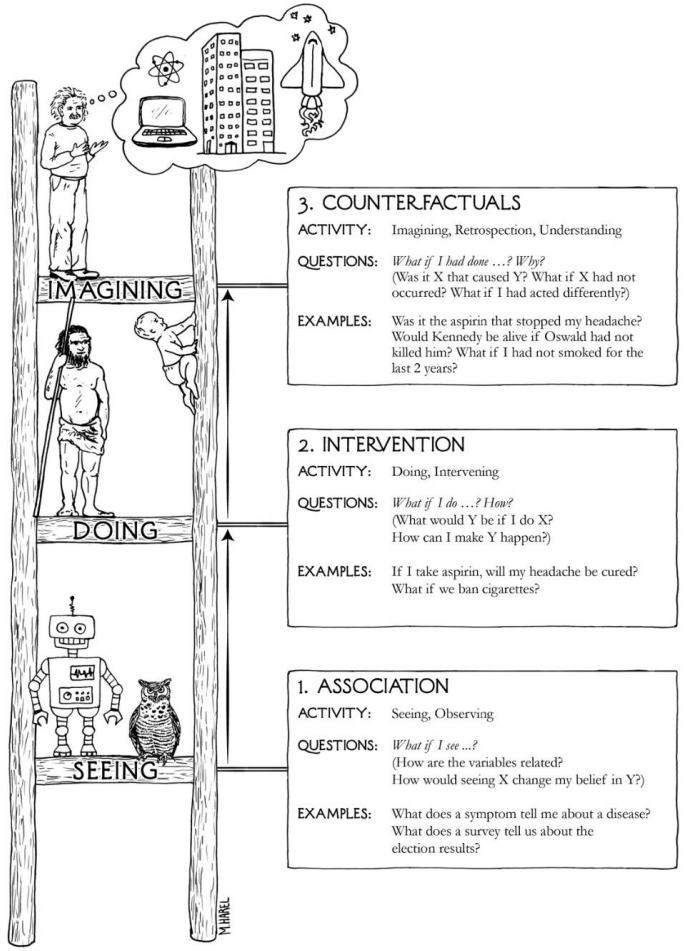


FIGURE 1.2. The Ladder of Causation, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first rung, learning from association. Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge. Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena. (*Source:* Drawing by Maayan Harel.)

Modelo de razonamiento contrafáctico[1]

- No construye estadística nueva.
- Aprende razonamiento contrafáctico de ejemplos contrafácticos.
- Como ocurría con los *transformers*, su estructura responde a la naturaleza del problema que intenta resolver.

[1] F. Feng, J. Zhang, X. He, H. Zhang, and T.-S. Chua, "Empowering Language Understanding with Counterfactual Reasoning," vol. abs/2106.03046, 2021.

Conclusiones y líneas futuras.

- Distinción entre acercamiento simbólico y estadístico es ambigua con los *Transformers.*
- Obtienen mejores resultados que los métodos tradicionales en un problema de búsqueda y recuperación de la información.
- Consiguen reproducir de forma adecuada la gramática del lenguaje desde un marco ambientalista, pero no así su semántica.
- La causalidad es una buena baza para línea de investigación.
    - Construcción con inferencia causal en vez de con estadística clásica.
    - Aprendizaje estadístico a partir de ejemplos contrafácticos.

Gracias por su “atención”