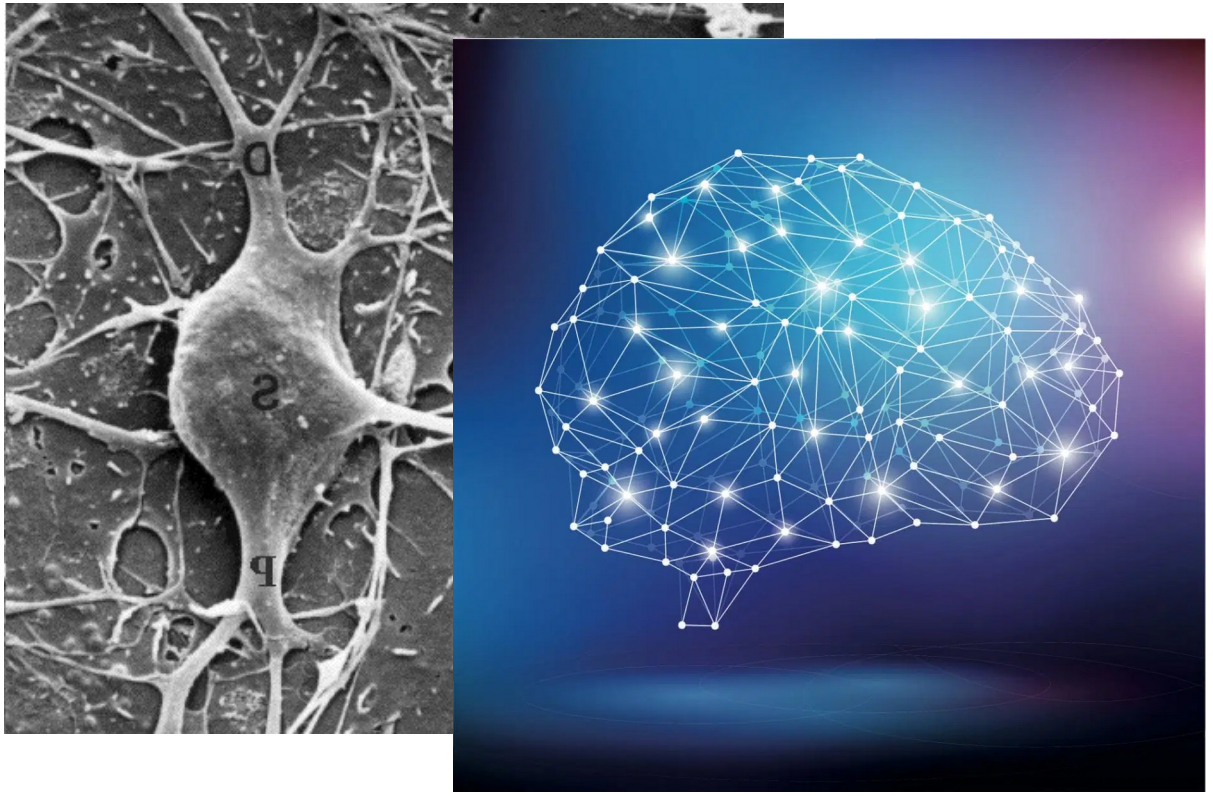
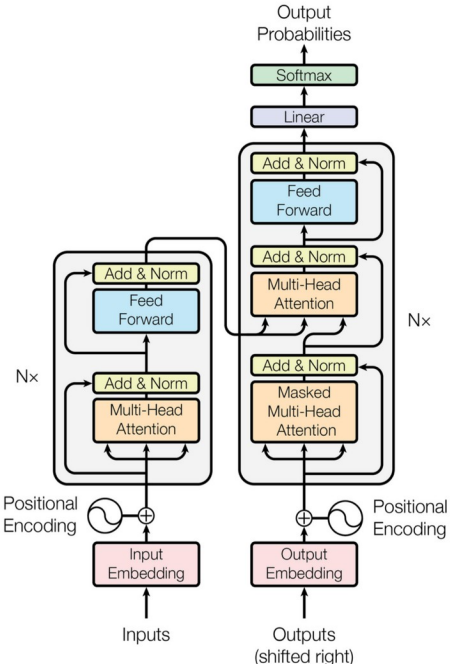
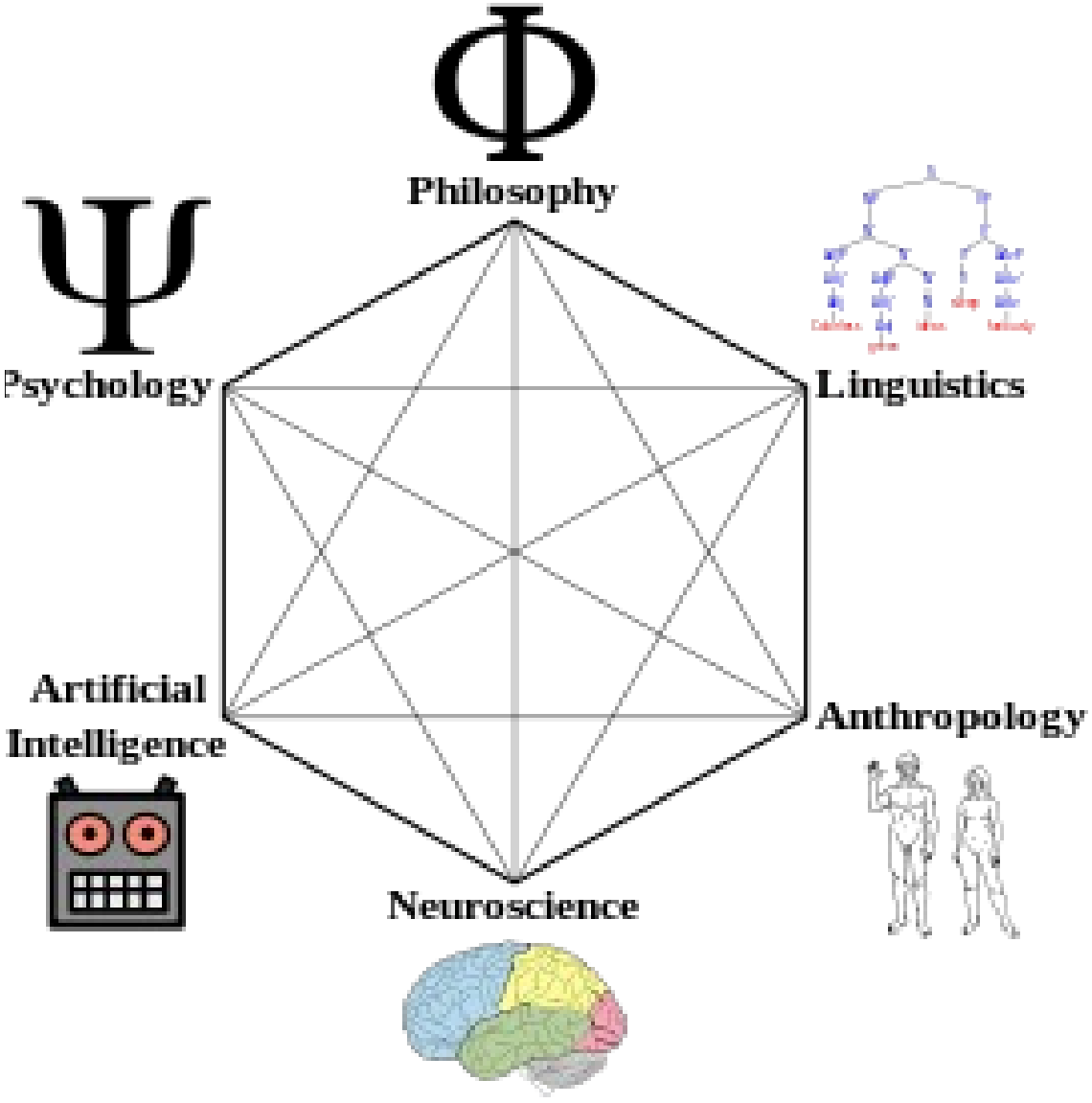
Una breve introducción a las Teorías de Redes Neuronales: 3



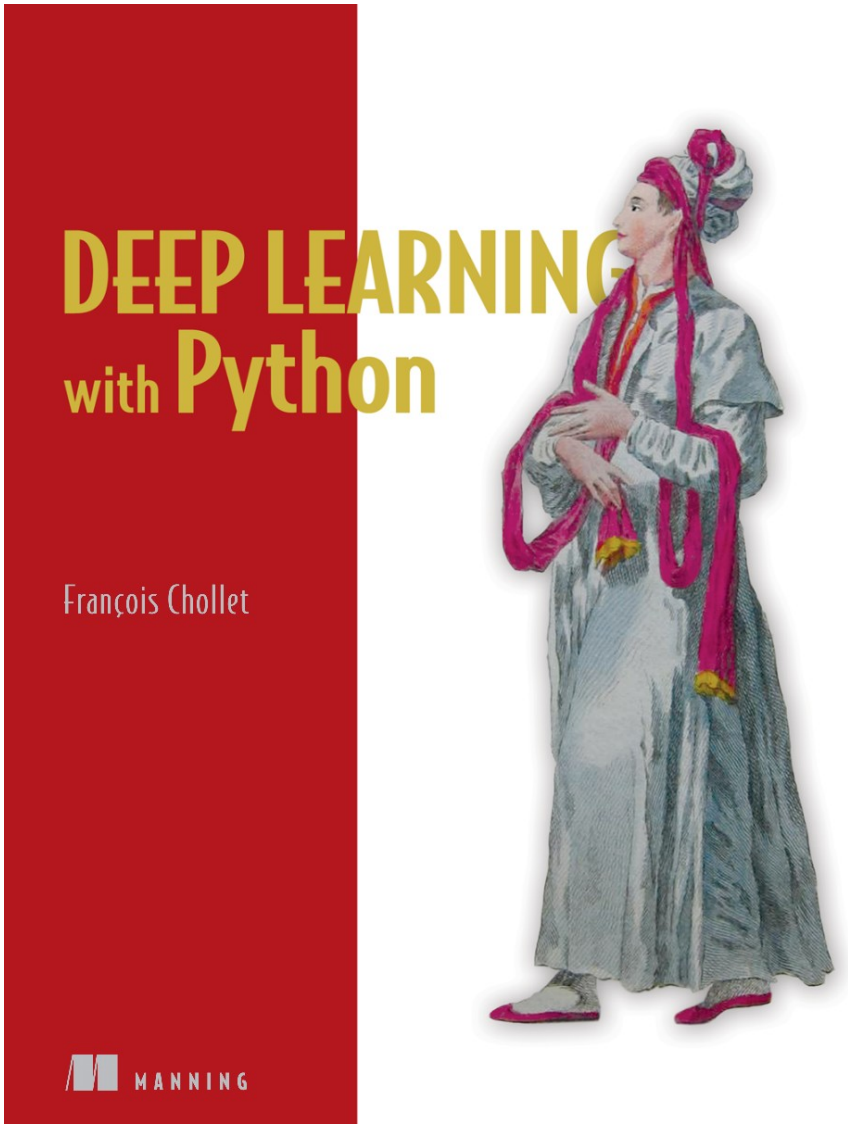


Juan Valle Lisboa,
Sección Biofísica, 2022

Neurociencia, IA y Ciencias Cognitivas



¿Sólo un tema de ingenieros?



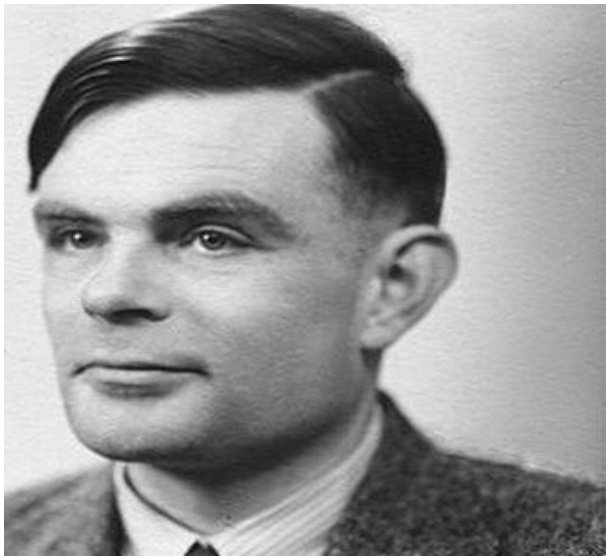
The term neural network is a reference to neurobiology, but although some of the central concepts in deep learning were developed in part by drawing inspiration from our understanding of the brain, deep-learning models are not models of the brain. There's no evidence that the brain implements anything like the learning mechanisms used in modern deep-learning models. You may come across pop-science articles proclaiming that deep learning works like the brain or was modeled after the brain, but that isn't the case. It would be confusing and counterproductive for newcomers to the field to think of deep learning as being in any way related to neurobiology; you don't need that shroud of "just like our minds" mystique and mystery, and you may as well forget anything you may have read about hypothetical links between deep learning and biology. For our purposes, deep learning is a mathematical framework for learning representations from data.

Computación es cognición

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]



Samuel N Alexander

First Draft of a Report
on the EDVAC

by

John von Neumann

Contract No. W-670-ORD-4926

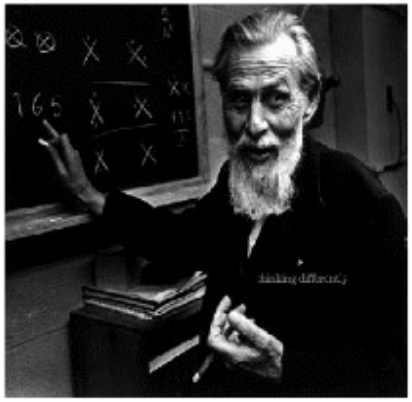
Between the

United States Army Ordnance Department

La primera red neuronal es Biofísica

Las primeras redes neuronales: McCullochs & Pitts,

(1943, Bulletin of Mathematical Biophysics)



activity. Certainly for the psychiatrist it is more to the point that in such systems "Mind" no longer "goes more ghostly than a ghost." Instead, diseased mentality can be understood without loss of scope or rigor, in the scientific terms of neurophysiology. For

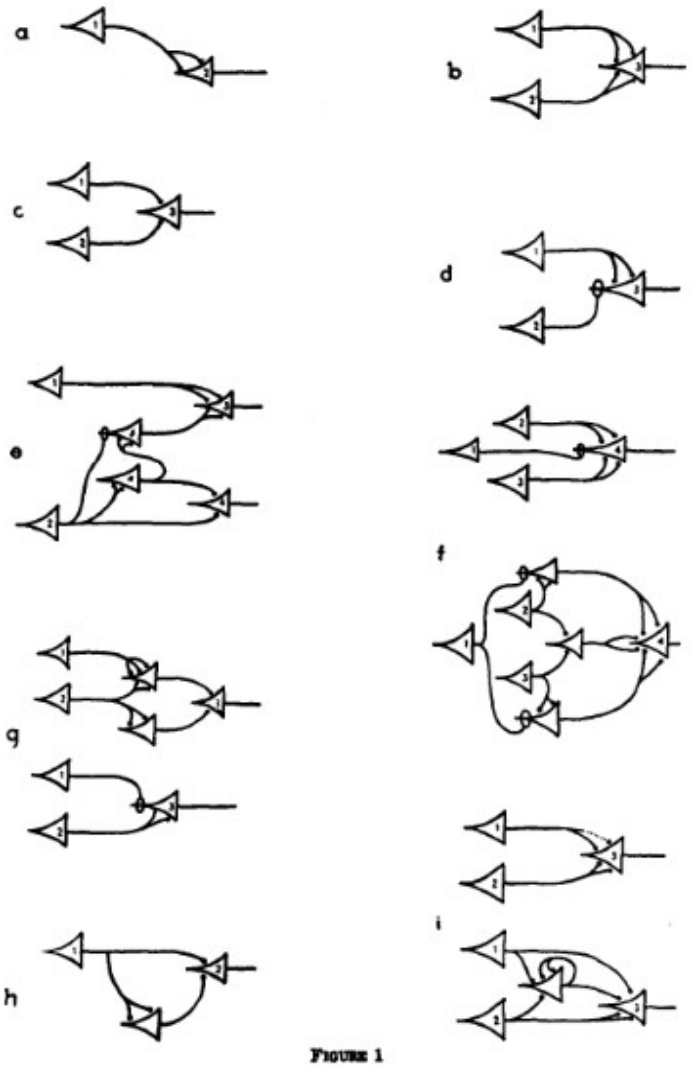
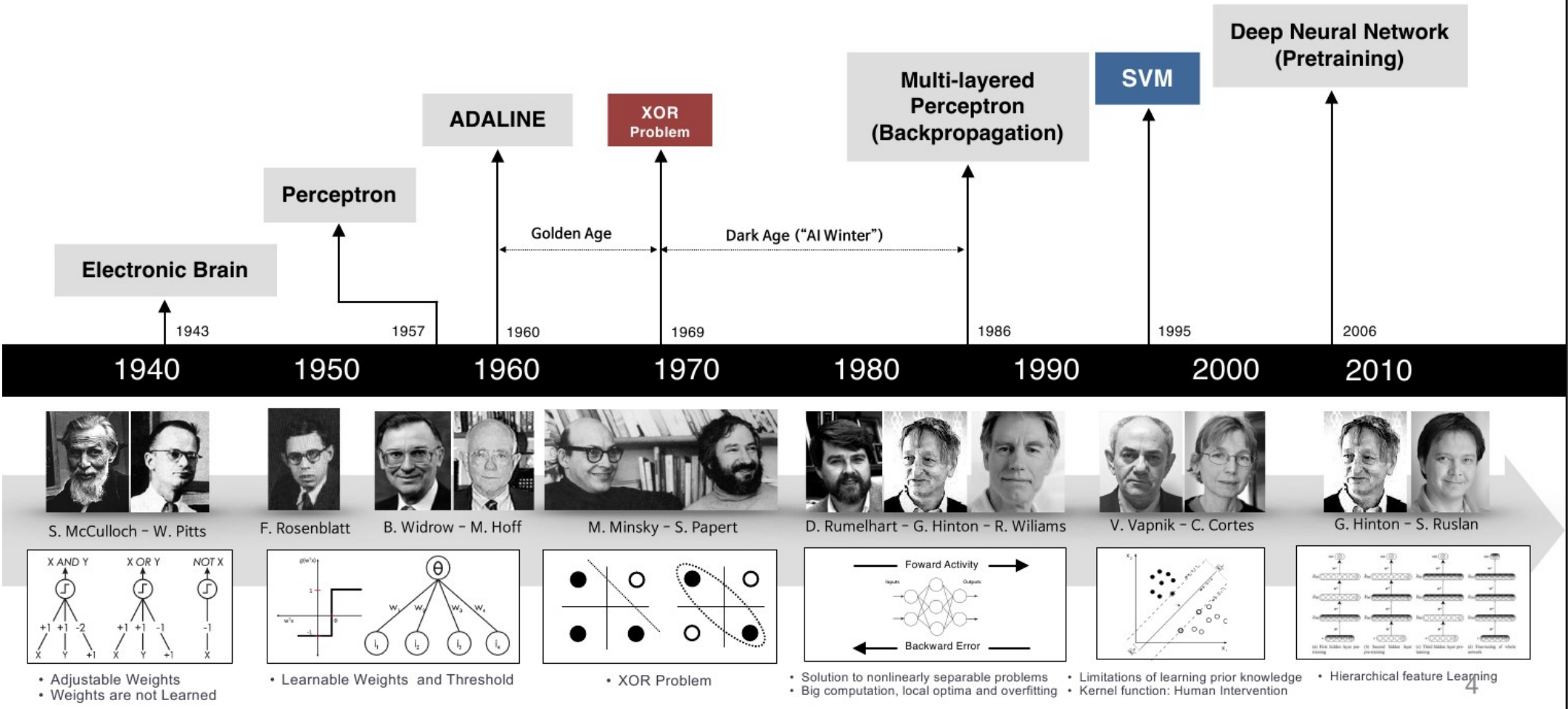


FIGURE 1

Brief History of Neural Network

DEVIEW
2015



S. McCulloch - W. Pitts

F. Rosenblatt

B. Widrow - M. Hoff

M. Minsky - S. Papert

D. Rumelhart - G. Hinton - R. Wiliams

V. Vapnik - C. Cortes

G. Hinton - S. Ruslan

- Adjustable Weights
- Weights are not Learned

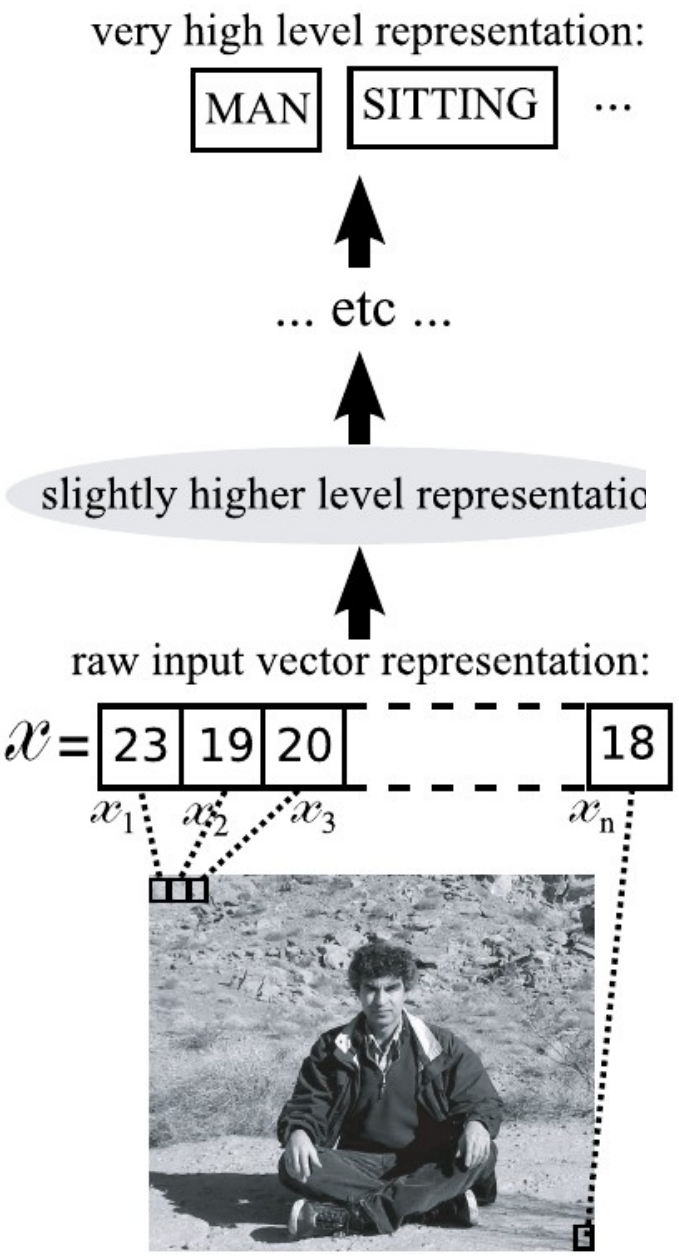
- Learnable Weights and Threshold

- XOR Problem

- Solution to nonlinearly separable problems
- Big computation, local optima and overfitting

- Limitations of learning prior knowledge
- Kernel function: Human Intervention

- Hierarchical feature Learning



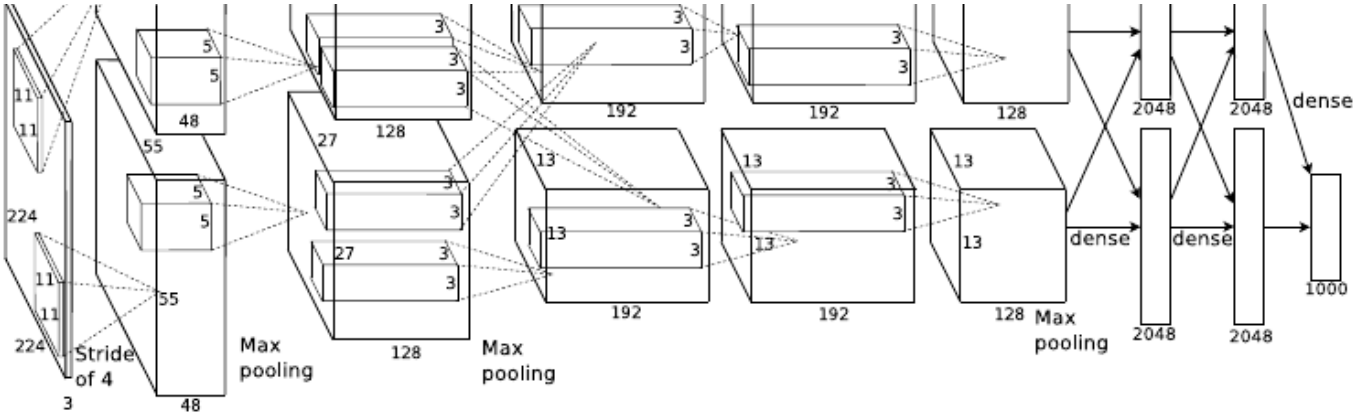


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

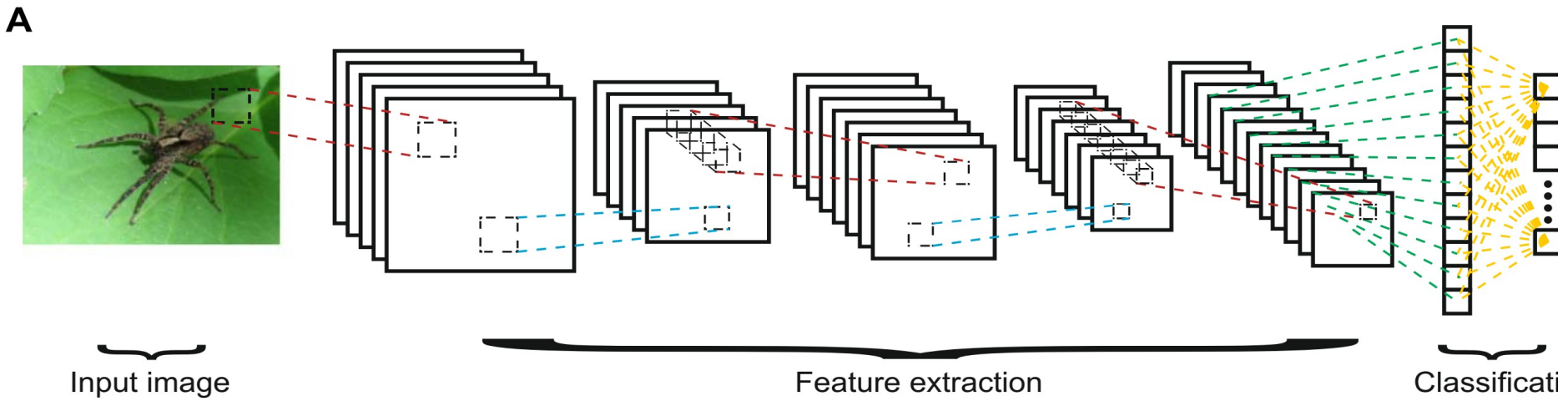
Redes modernas: muchas capas y muchos datos

COGNITIVE NEUROSCIENCE

Number detectors spontaneously emerge in a deep neural network designed for visual object recognition

Khaled Nasr*, Pooja Viswanathan†, Andreas Nieder‡

Nasr *et al.*, *Sci. Adv.* 2019;**5**:eaav7903 8 May 2019



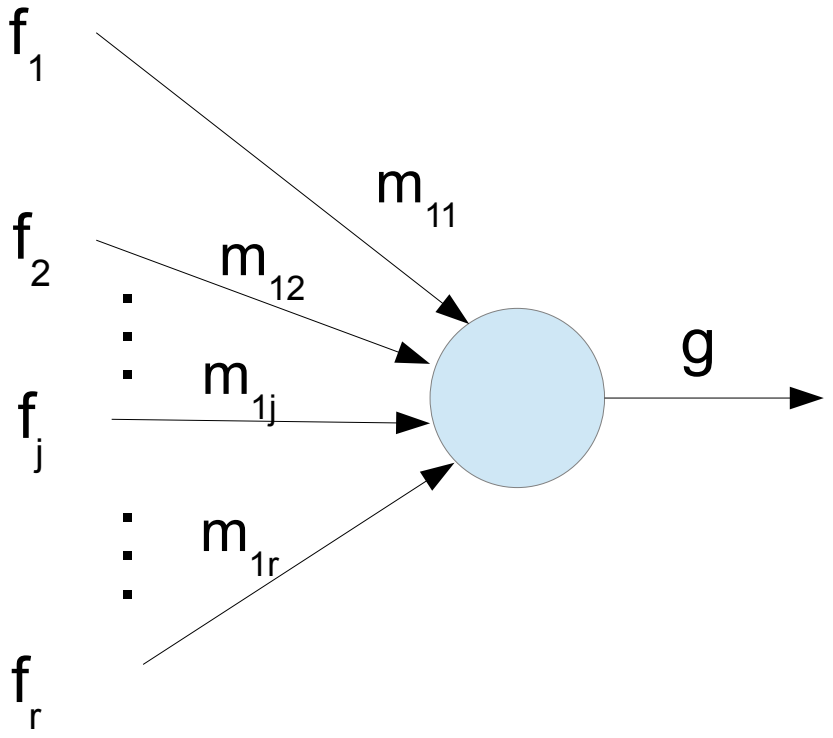
De las memorias distribuidas a la Inteligencia Artificial

Aprendizaje automático

- Aprendizaje supervisado
  - Se le da a la red que aprende, información de cómo debería responder a cierta entrada.
  - Autosupervisado: usar la predicción.
- Aprendizaje no supervisado
  - Se le da a la red solamente entradas.
- Aprendizaje por refuerzo
  - La red (o el agente) recibe información de error, pero no de la salida correcta.

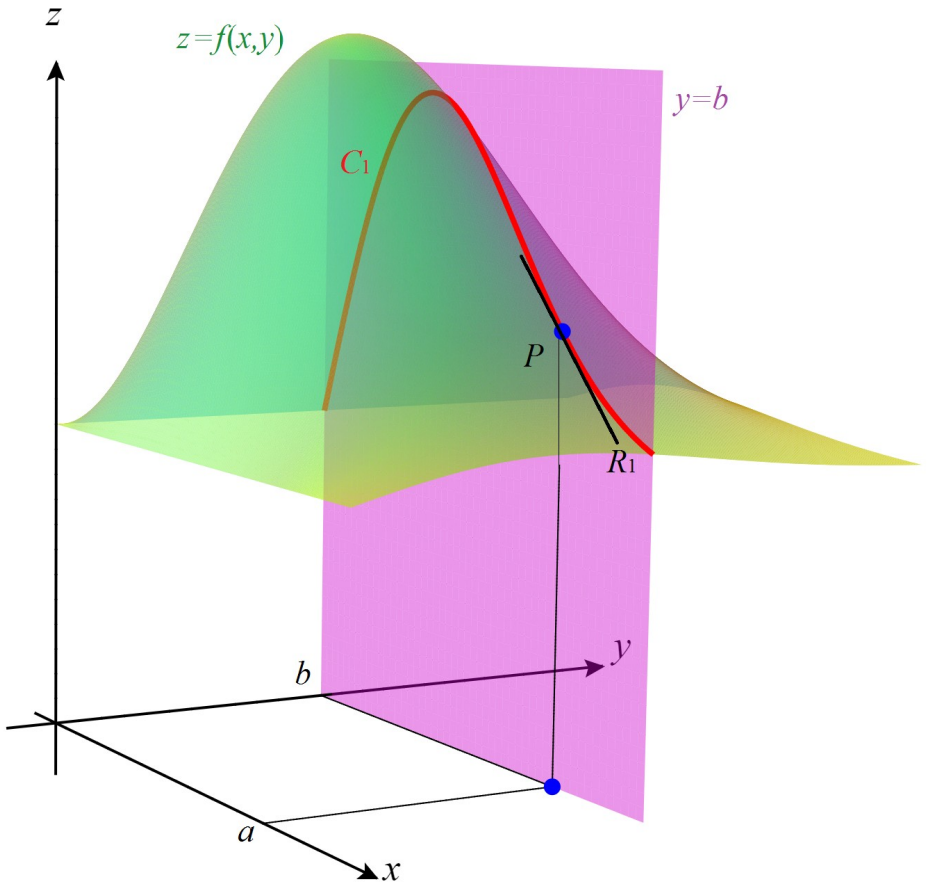
Regla delta

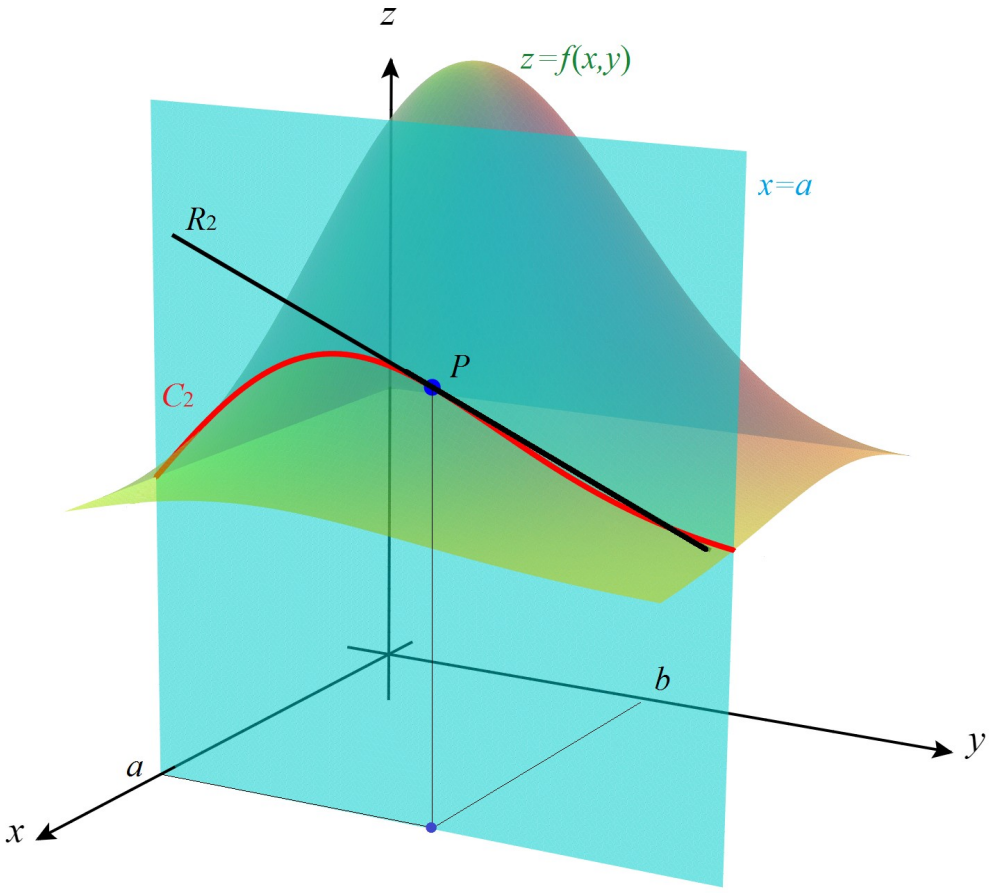
- Es una regla de aprendizaje supervisado; precisa la salida esperada.
- Delta hace referencia a la diferencia entre lo esperado y lo observado: sale del **descenso por el gradiente**.



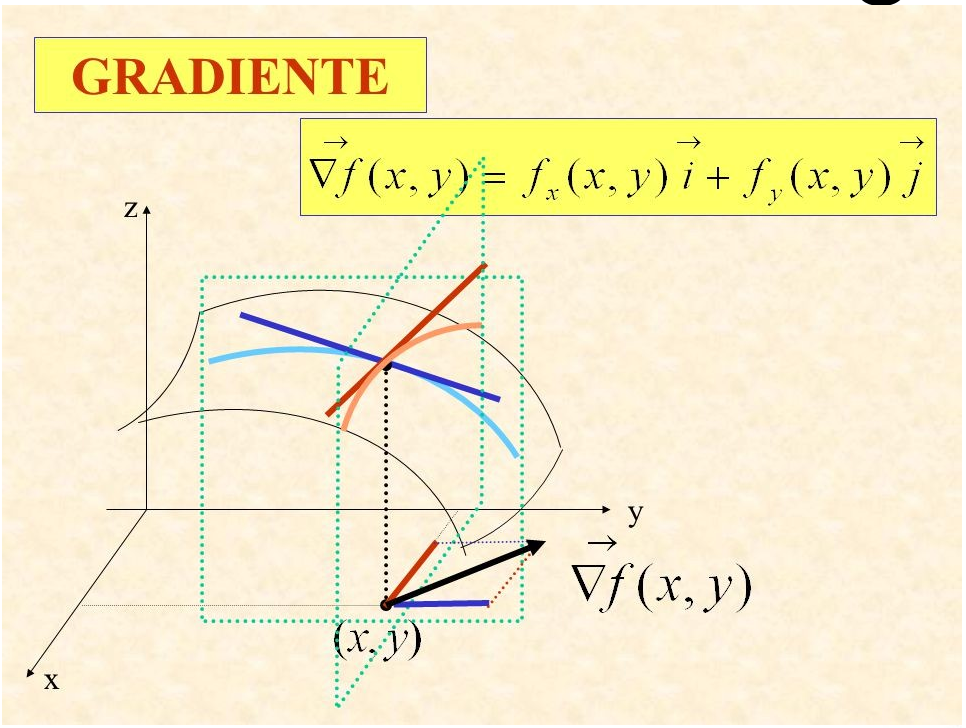
$$\Delta m_{1j} = \eta (t - g) f_j$$

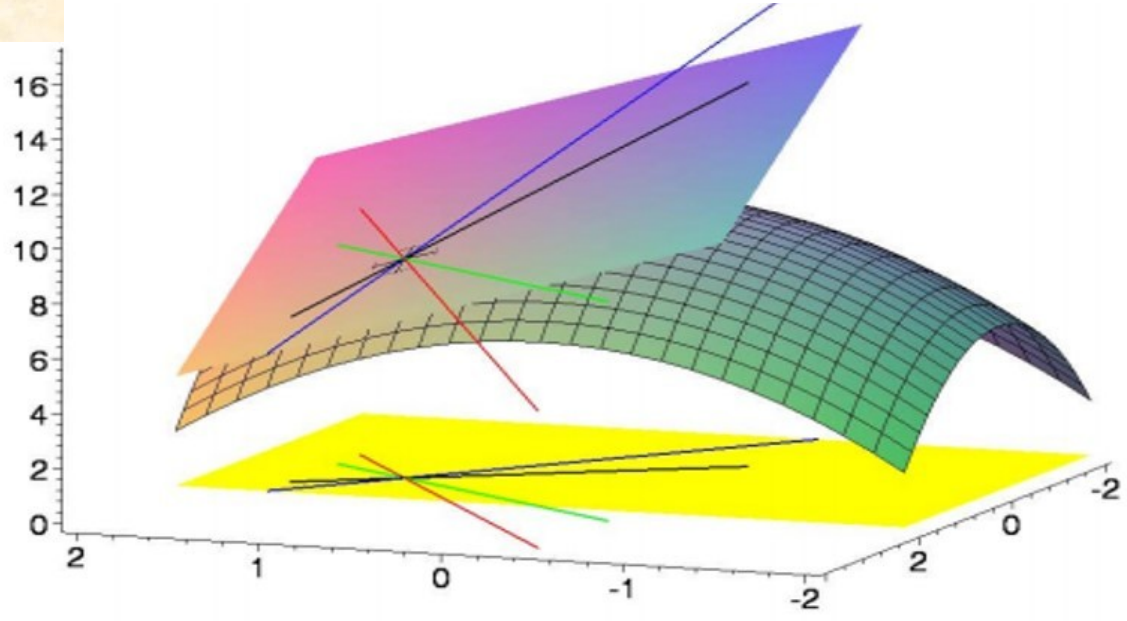
Las derivadas parciales





El gradiente





Descenso por el gradiente:

Si tenemos una función de varias variables, o sea $f(\mathbf{X}):R^n \rightarrow R$, con $\mathbf{X}=(x_1,x_2, \ldots,x_n)$

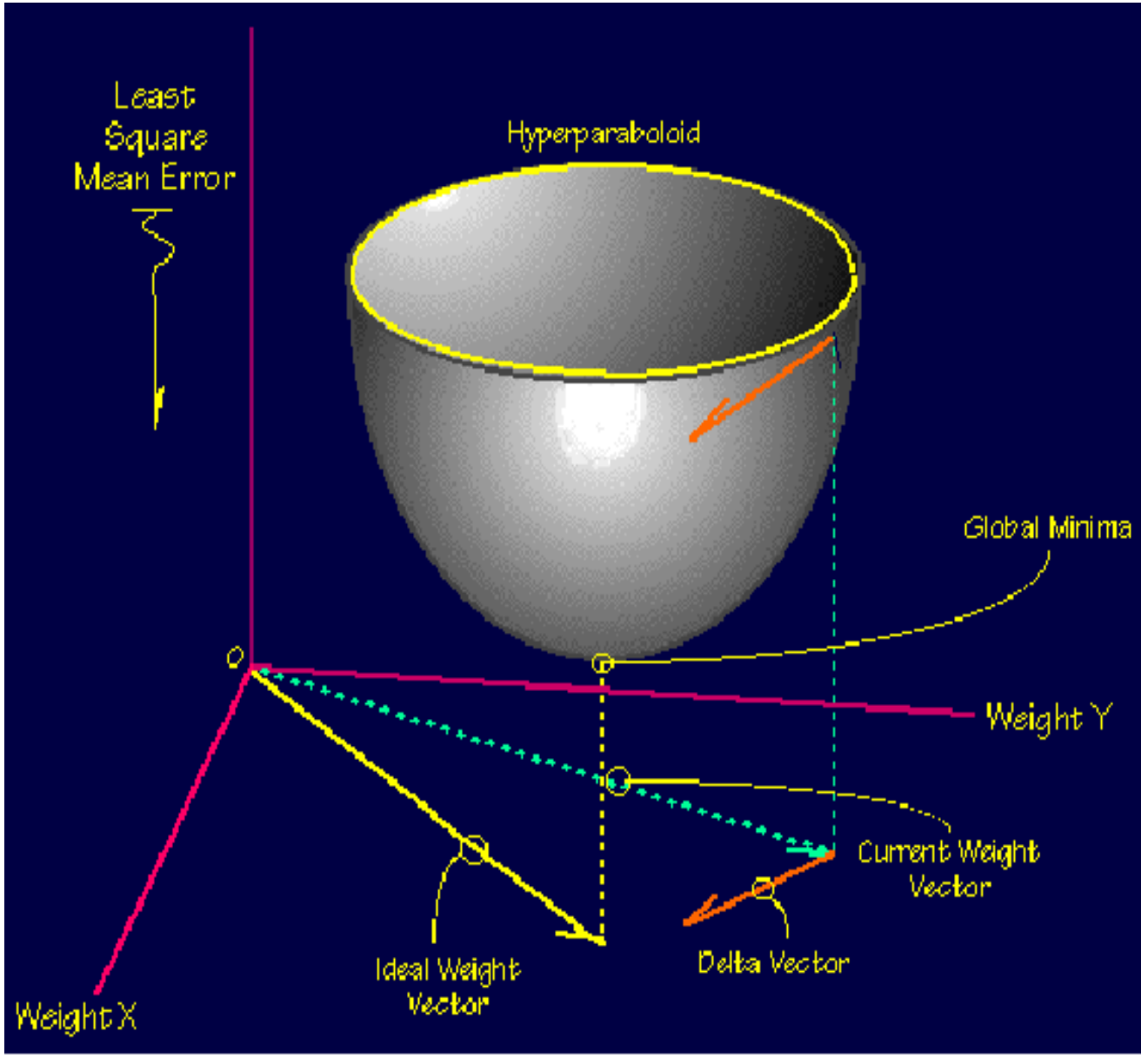
El gradiente de la función f, ∇f, es el vector de componentes:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

Nota: En un extremo $\nabla f=0$.
Teo: El vector ∇f en el espacio del dominio apunta en la dirección en la que la función crece más rápidamente.

IDEA: En cada punto $\mathbf{X}_0=(x_10,x_20,...x_n0)$ calcular el gradiente y cambiar X un poco en esa dirección (para maximizar) o en la opuesta (para minimizar).

Descenso por el gradiente:



Ejemplos

Si $f(x) = x^2 + 1$, $f'(x) = 2x$ en $x_0 = 0$ hay un mínimo

$f(0) = 1$.

Otra forma: $x_0 = 1$; $f(x_0) = 1^2 + 1 = 2$

$$x_1 = x_0 - \eta f'(x_0) = x_0 - 2x_0\eta = x_0(1 - 2\eta)$$

Si $\eta = 0{,}1 \rightarrow x_1 = x_0(1 - 0{,}2) = 0{,}8x_0$

$$x_2 = x_1 - \eta f'(x_1) = x_1 - 2\eta x_1 = x_1(1 - 2\eta)$$

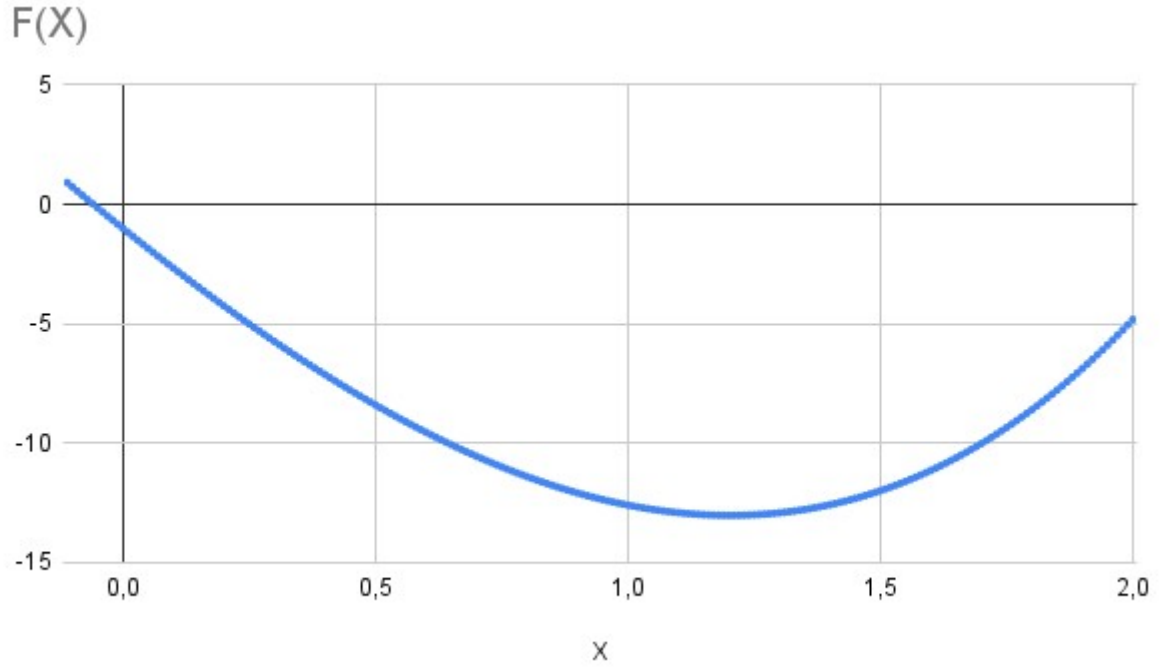
$$= x_0(1 - 2\eta)^2$$

Si hacemos $x_n = x_0(1 - 2\eta)^n$ y $\eta < \frac{1}{2} \rightarrow x_n \rightarrow 0$

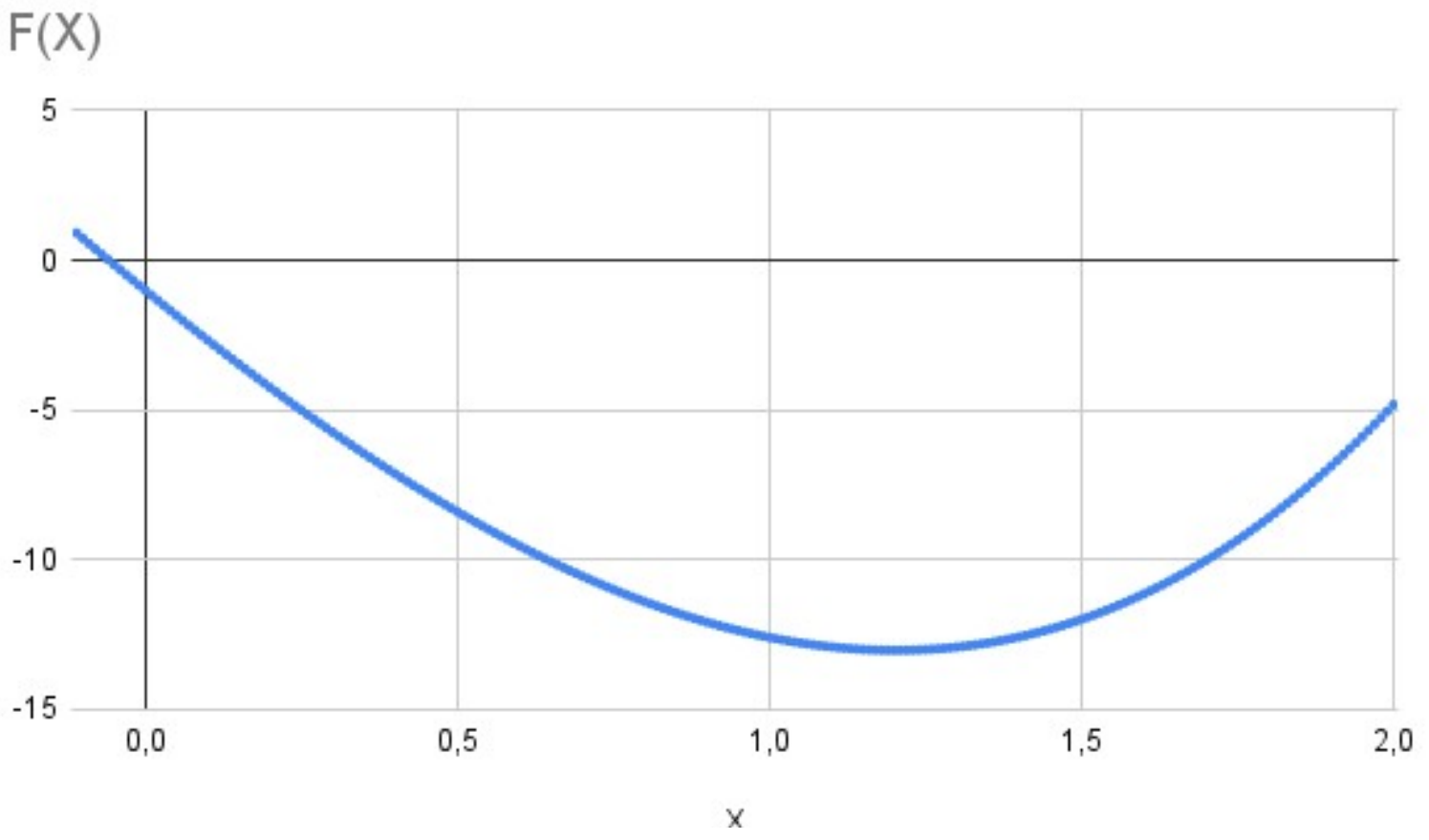
$$f(x) = \frac{1}{9}x^4 + \frac{3}{5}x^3 + \frac{7}{2}x^2 - 17x - 1$$

$$f'(x) \quad \frac{4}{9}x^3 + \frac{27}{5}x^2 + 7x - 17$$

$$x = x_0 - \eta f'(x_0) = x_0 - \eta\left(\frac{4}{9}x_0^3 + \frac{27}{5}x_0^2 - 7x_0 - 17\right)$$



X	1	1,42	1,32	1,27	1,242	1,227	1,219	1,214	1,2114		1,202557	1,2026
Δ	0,42	-0,1	-0,05	-0,028	-0,015	-0,008	-0,005	-0,003	-0,001	. . .	1,41E-11	0
F(x)	-12,59	-12,48	-12,86	-12,97	-13,00	-13,01	-13,02	-13,02	-13,02		-13,01927	-13,02



Aplicación a nuestro problema

Tenemos

$$g_i = \boldsymbol{M_i f}$$

En donde g_i es la actividad de la neurona de salida i-ésima, $\mathbf{M_i}$ es la fila i-ésima de la matriz de pesos, y **f** es un vector de entrada.

Para diferentes vectores **f(k)** que deseamos tengan como salida **t(k)**, en particular $t_i(k)$, es la actividad de salida deseada de la neurona i-ésima frente a la entrada **f(k)**. Se busca minimizar el error cuadrático cometido

$$E_i = \frac{1}{2}\sum_k \left(t_i(k) - g_i(k)\right)^2 = \frac{1}{2}\sum_k \left(t_i(k) - \boldsymbol{M_i f}(\boldsymbol{k})\right)^2$$

Aplicación a nuestro problema

Una forma de hacer la actualización cada vez que se presenta un ejemplo y tratar de reducir ese error:

$$e_i^2(k)=\frac{1}{2}\left(t_i(k)-\boldsymbol{M_i f}(\boldsymbol{k})\right)^2$$

$$\Delta m_{ij}=-\eta\frac{\partial e_i^2(k)}{\partial m_{ij}}=-\eta\frac{1}{2}2\left(t_i(k)-\boldsymbol{M_i f}(\boldsymbol{k})\right)\frac{-\partial \boldsymbol{M_i f}(\boldsymbol{k})}{\partial m_{ij}}$$

Para aplicar el descenso por el gradiente (ecuación superior derecha), hay que derivar con respecto a los parámetros:

$$\boldsymbol{M_i f}(\boldsymbol{k})=\sum_j m_{ij}f_j(k) \qquad \frac{\partial \sum_j m_{ij}f_j(k)}{\partial m_{ij}}=f_j(k)$$

Al juntar todo, se obtiene:

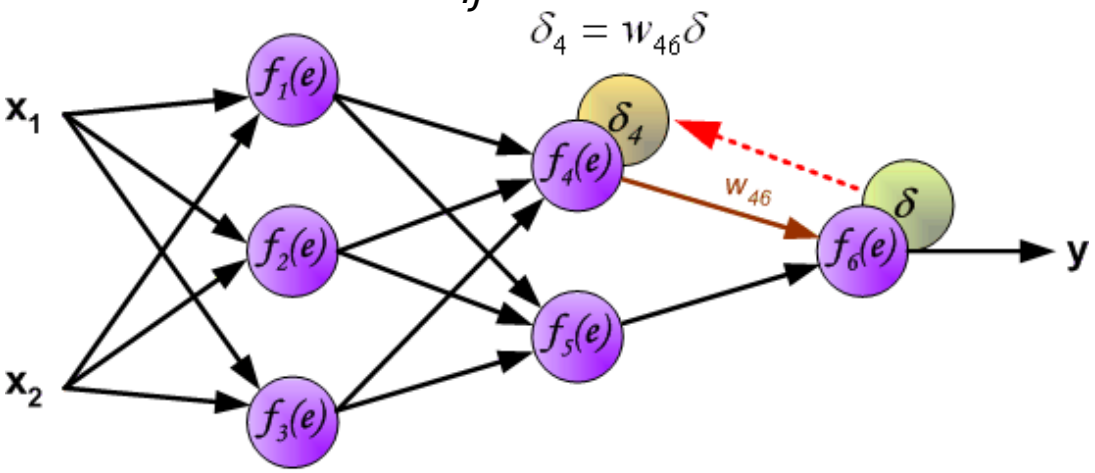
$$\Delta m_{ij}=-\eta\left[\frac{2}{2}\left(t_i(k)-\boldsymbol{M_i f}(\boldsymbol{k})\right)\left(-f_j(k)\right)\right]=\eta\left(t_i(k)-g_i(k)\right)f_j(k)\stackrel{\text{def}}{=}\delta_i(k)f_j(k)$$

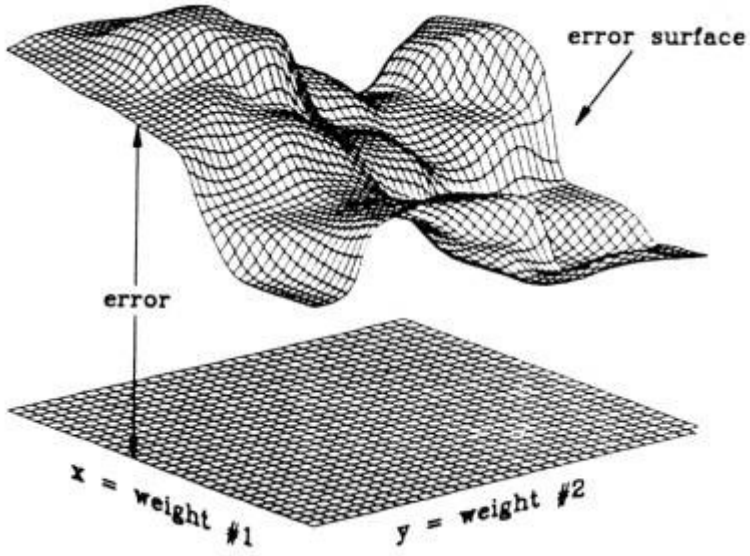
La regla delta generalizada: Backpropagation

Widrow & Hoff, 1960: $\Delta M_{ij} = \eta (t_i - n_i) f_j$

Rumelhart, Hinton, Williams (1986) (Werbos 1974)

$$\Delta w_{ij} = -\eta \, ' \frac{\partial E}{\partial w_{ij}}$$





Redes recurrentes simples y la tarea de predicción (Elman, 1990, 1993)

COGNITIVE SCIENCE 14, 179–211 (1990)

Finding Structure in Time

JEFFREY L. ELMAN
University of California, San Diego

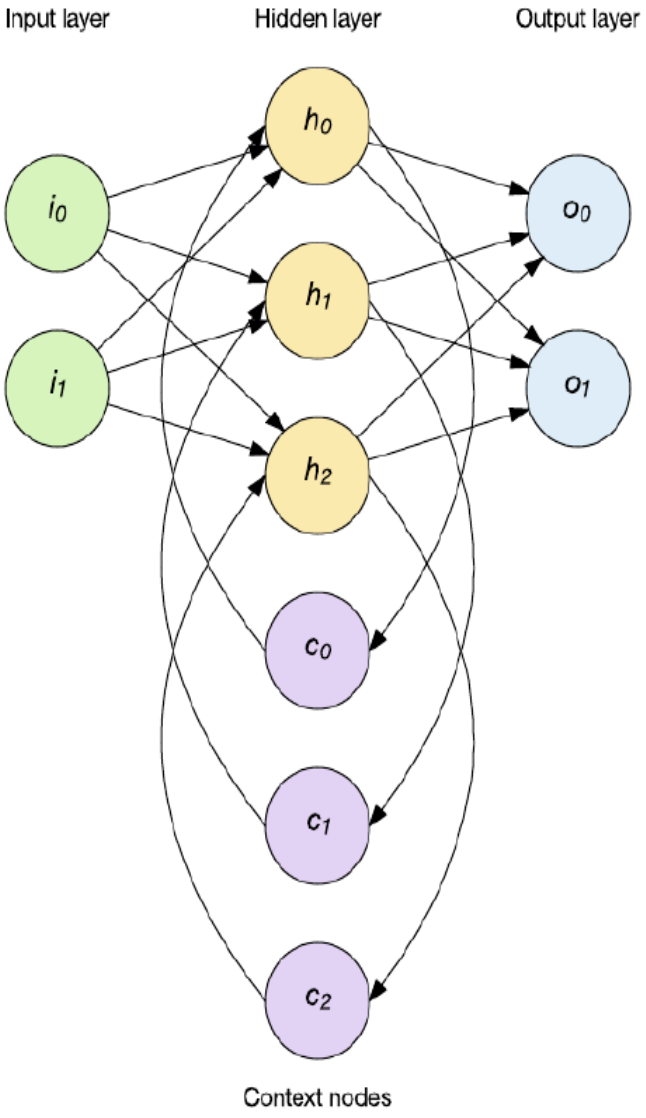


TABLE 3

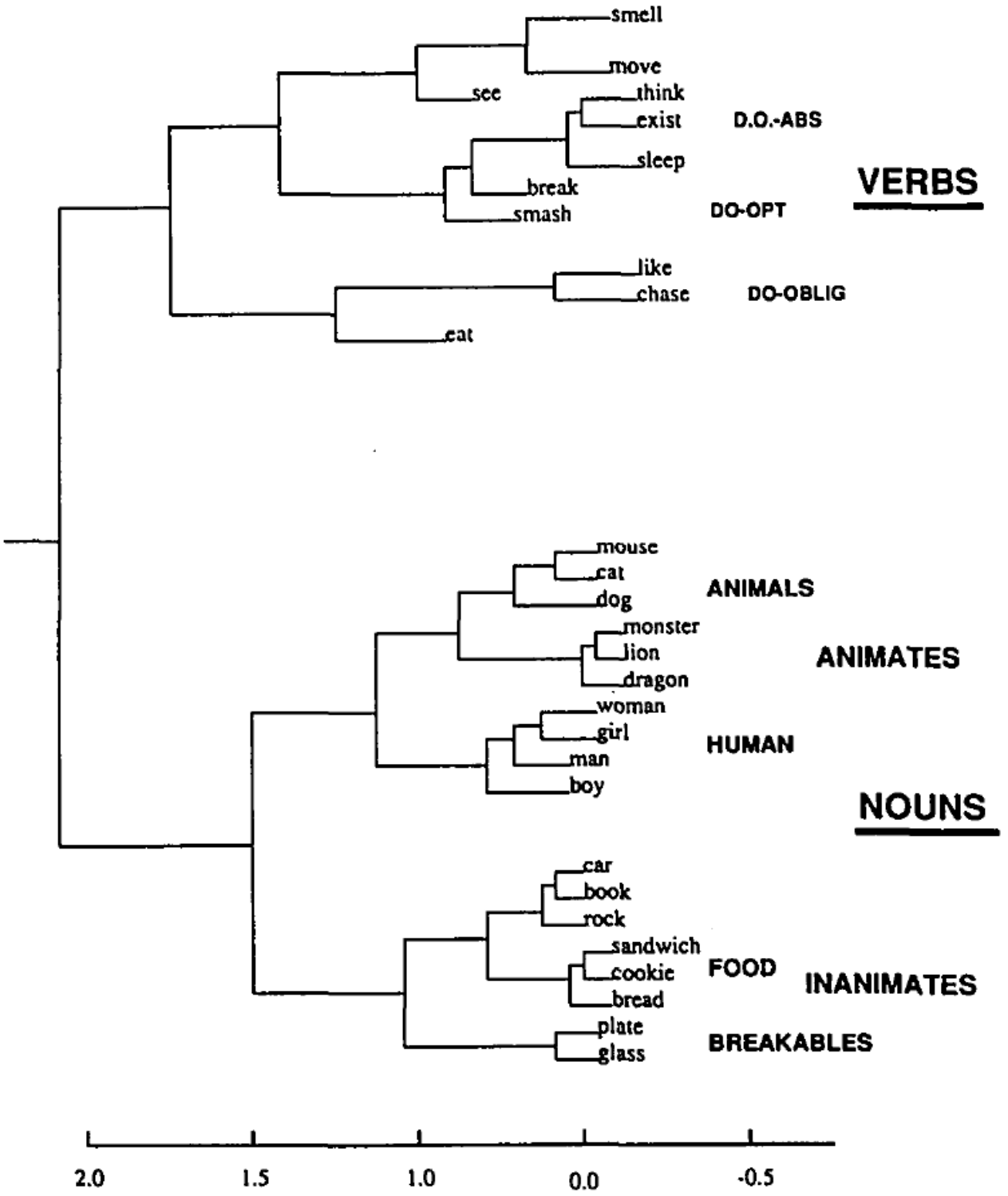
Categories of Lexical Items Used in Sentence Simulation

Category	Examples
NOUN-HUM	man, woman
NOUN-ANIM	cat, mouse
NOUN-INANIM	book, rock
NOUN-AGRESS	dragon, monster
NOUN-FRAG	glass, plate
NOUN-FOOD	cookie, break
VERB-INTRAN	think, sleep

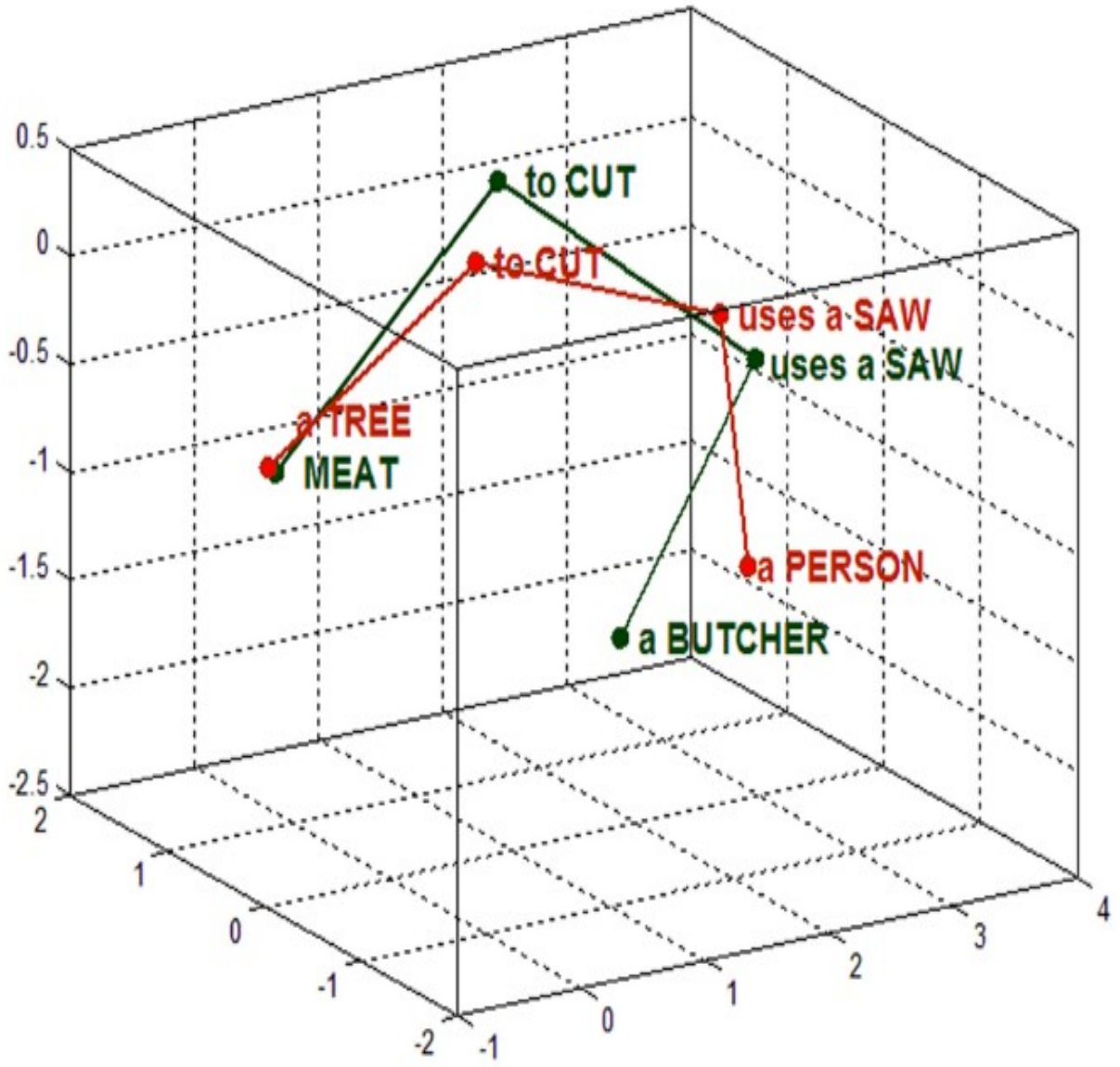
WORD 1	WORD 2	WORD 3
NOUN-HUM	VERB-EAT	NOUN-FOOD
NOUN-HUM	VERB-PERCEPT	NOUN-INANIM
NOUN-HUM	VERB-DESTROY	NOUN-FRAG
NOUN-HUM	VERB-INTRAN	
NOUN-HUM	VERB-TRAN	NOUN-HUM
NOUN-HUM	VERB-AGPAT	NOUN-INANIM
NOUN-HUM	VERB-AGPAT	
NOUN-ANIM	VERB-EAT	NOUN-FOOD
NOUN-ANIM	VERB-TRAN	NOUN-ANIM
NOUN-ANIM	VERB-AGPAT	NOUN-INANIM
NOUN-ANIM	VERB-AGPAT	
NOUN-INANIM	VERB-AGPAT	
NOUN-AGRESS	VERB-DESTROY	NOUN-FRAG
NOUN-AGRESS	VERB-EAT	NOUN-HUM
NOUN-AGRESS	VERB-EAT	NOUN-ANIM
NOUN-AGRESS	VERB-EAT	NOUN-FOOD

Fragment of Training Sequences for Sentence Simulation

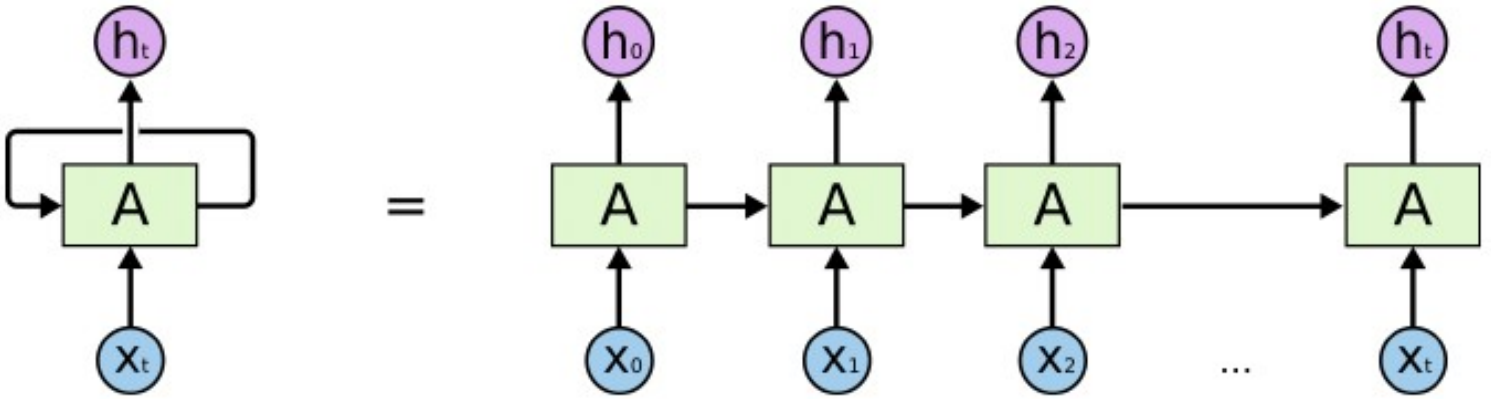
Input	Output
00000000000000000000000000000010 (*woman*)	00000000000000000000000000010000 (*smash*)
00000000000000000000000000010000 (*smash*)	00000000000000000000001000000000 (*plate*)
00000000000000000000001000000000 (*plate*)	00000100000000000000000000000000 (*cat*)
00000100000000000000000000000000 (*cat*)	00000000000000000001000000000000 (*move*)
00000000000000000001000000000000 (*move*)	00000000000000001000000000000000 (*man*)
00000000000000001000000000000000 (*man*)	00010000000000000000000000000000 (*break*)
00010000000000000000000000000000 (*break*)	00001000000000000000000000000000 (*car*)
00001000000000000000000000000000 (*car*)	01000000000000000000000000000000 (*boy*)
01000000000000000000000000000000 (*boy*)	00000000000000000010000000000000 (*move*)
00000000000000000010000000000000 (*move*)	00000000000010000000000000000000 (*girl*)
00000000000010000000000000000000 (*girl*)	00000000010000000000000000000000 (*eat*)
00000000010000000000000000000000 (*eat*)	00100000000000000000000000000000 (*bread*)
00100000000000000000000000000000 (*bread*)	00000001000000000000000000000000 (*dog*)
00000001000000000000000000000000 (*dog*)	00000000000000000010000000000000 (*move*)
00000000000000000010000000000000 (*move*)	00000000000000000100000000000000 (*mouse*)
00000000000000000100000000000000 (*mouse*)	00000000000000000100000000000000 (*mouse*)
00000000000000000100000000000000 (*mouse*)	00000000000000000010000000000000 (*move*)
00000000000000000010000000000000 (*move*)	10000000000000000000000000000000 (*book*)
10000000000000000000000000000000 (*book*)	00000000000000010000000000000000 (*lion*)


smell
move
see
think
exist
D.O.-ABS
sleep
break
smash
DO-OPT
VERBS
like
chase
DO-OBLIG
eat
mouse
cat
ANIMALS
dog
monster
lion
ANIMATES
dragon
woman
girl
HUMAN
man
boy
NOUNS
car
book
rock
sandwich
cookie
FOOD
INANIMATES
bread
plate
glass
BREAKABLES
2.0
1.5
1.0
0.0
-0.5

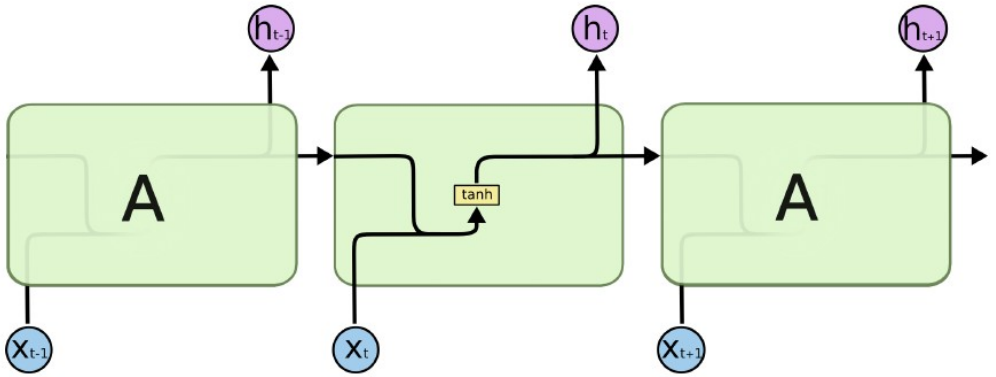
El significado en la Trayectoria



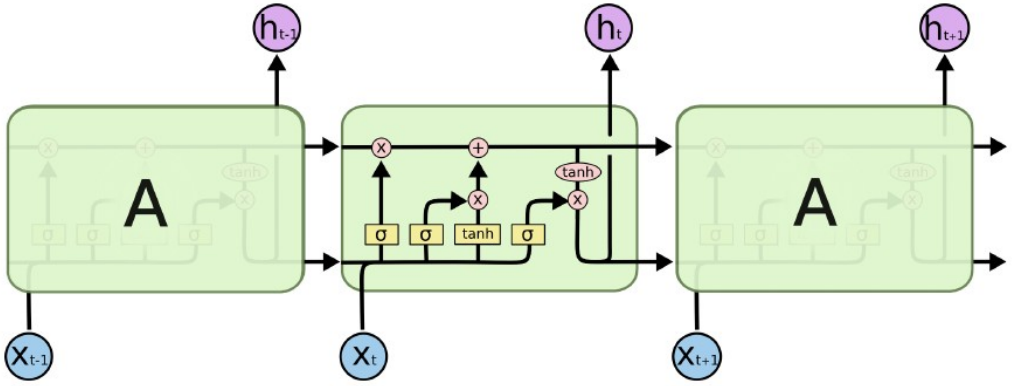
De SRNs a LSTMs



An unrolled recurrent neural network.

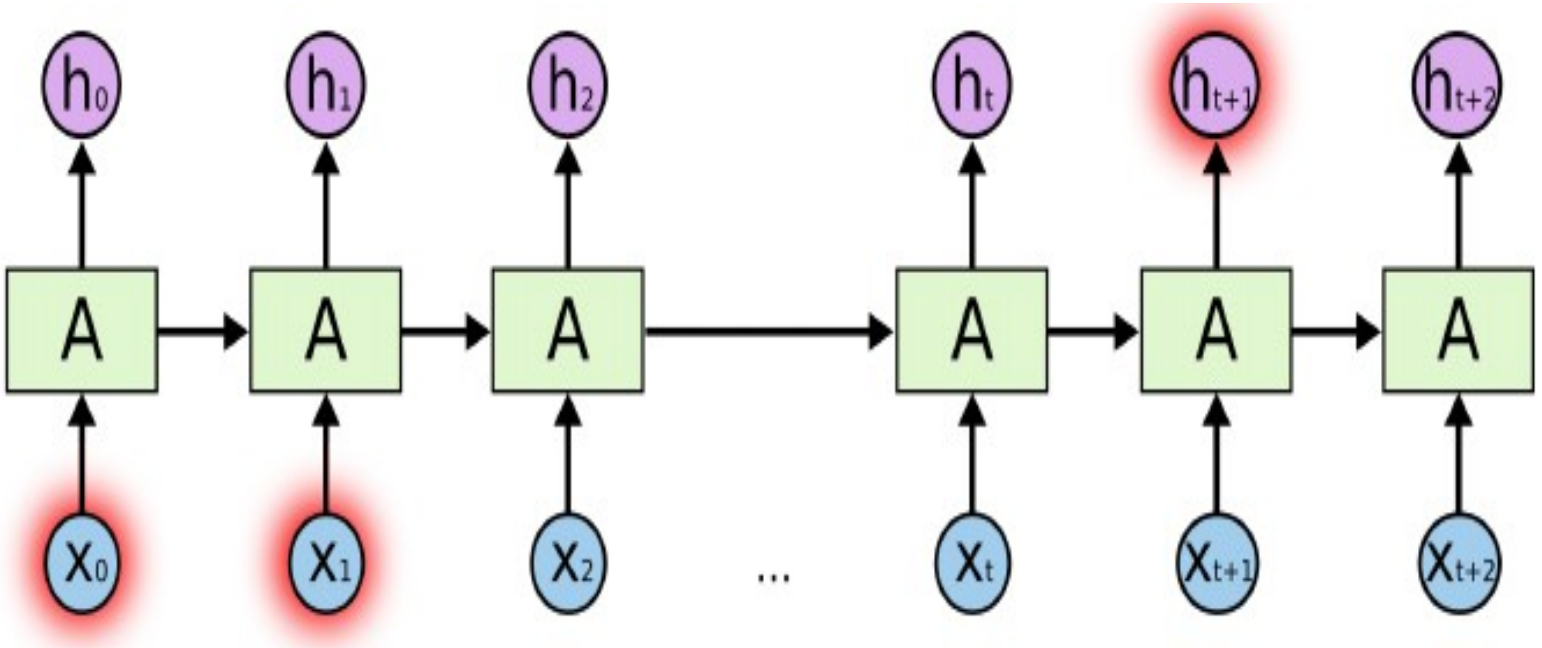


The repeating module in a standard RNN contains a single layer.

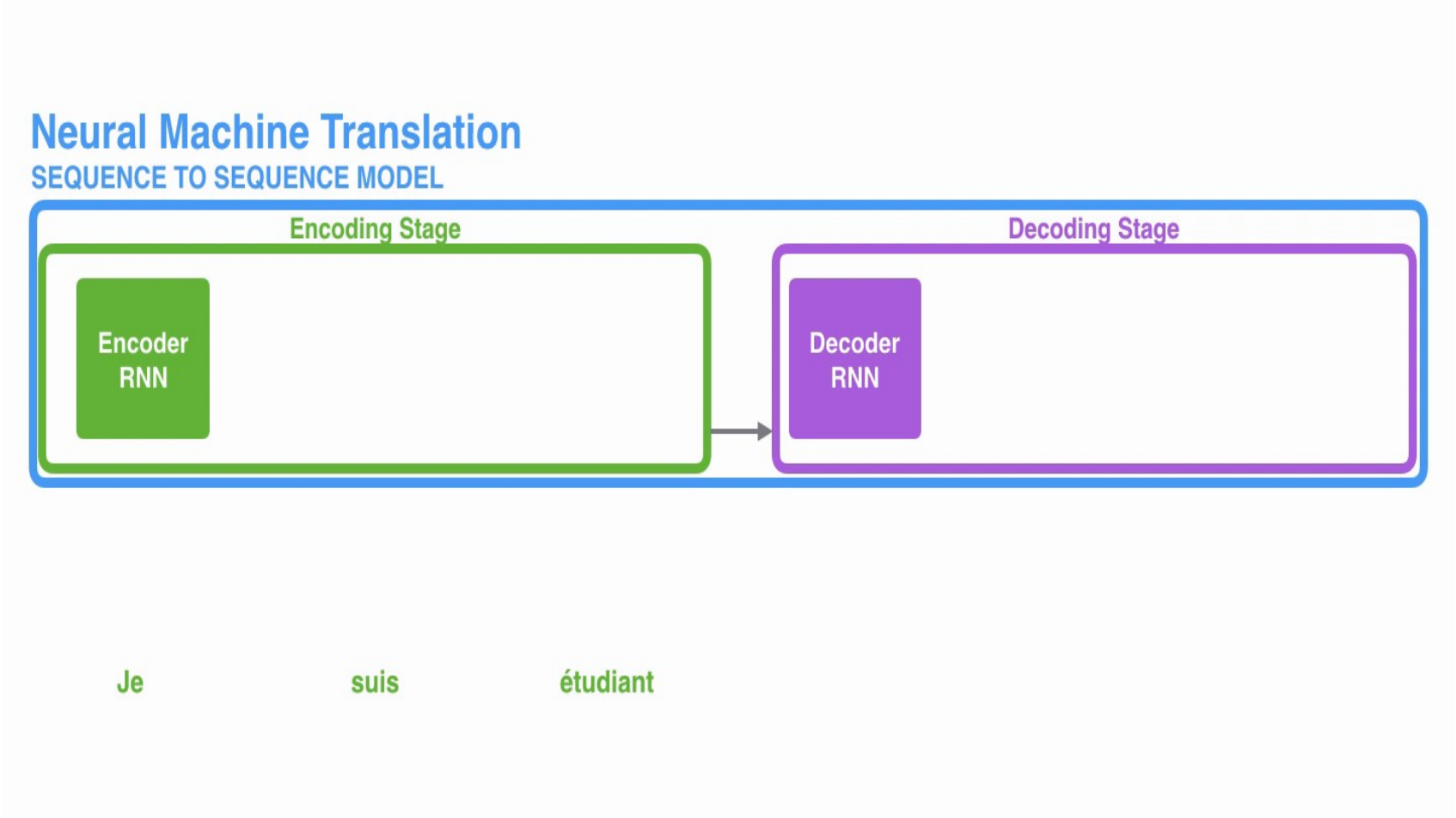


The repeating module in an LSTM contains four interacting layers.

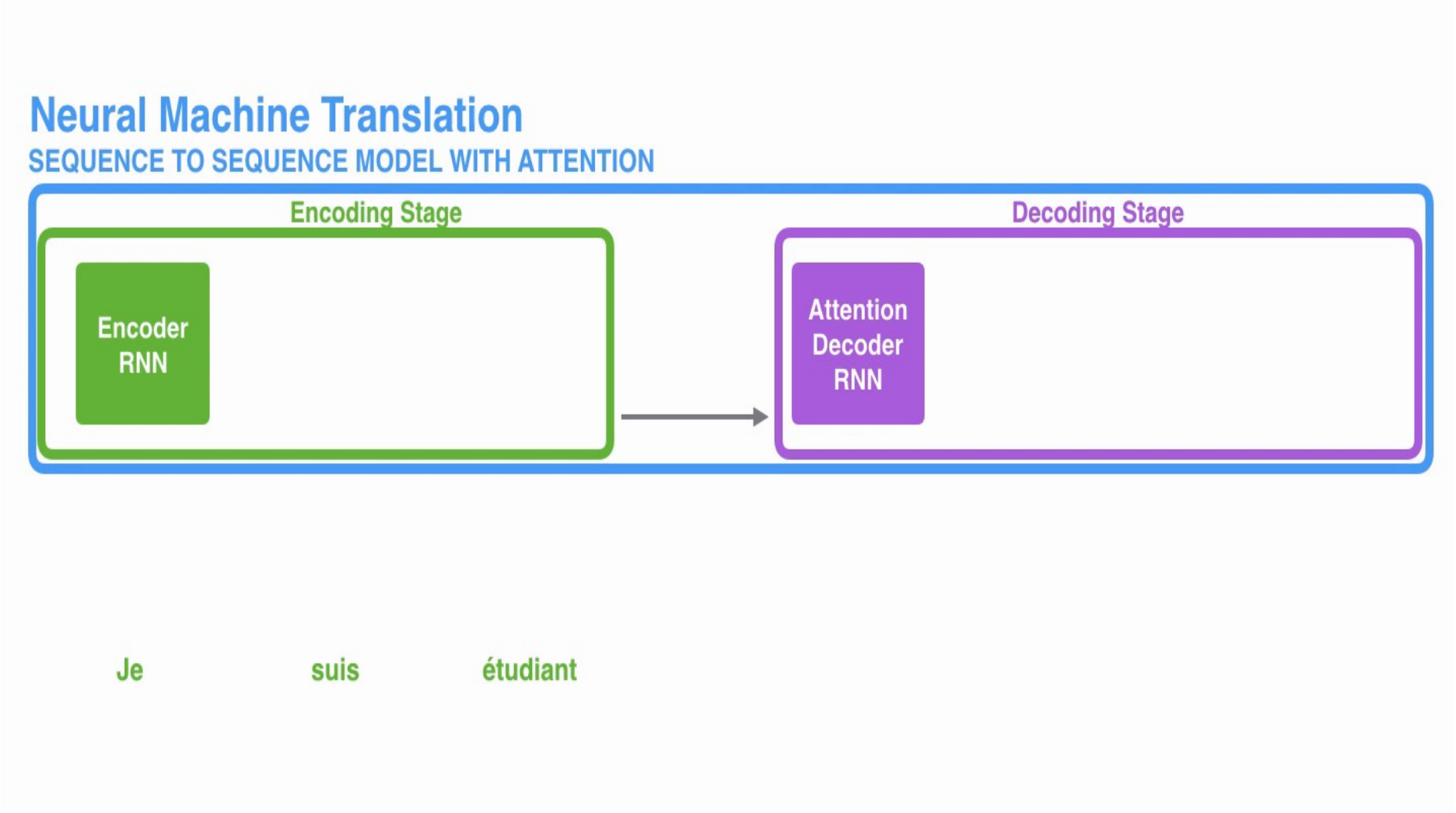
Las dependencias de largo alcance



La traducción: secuencias en secuencias



Atento Casco!



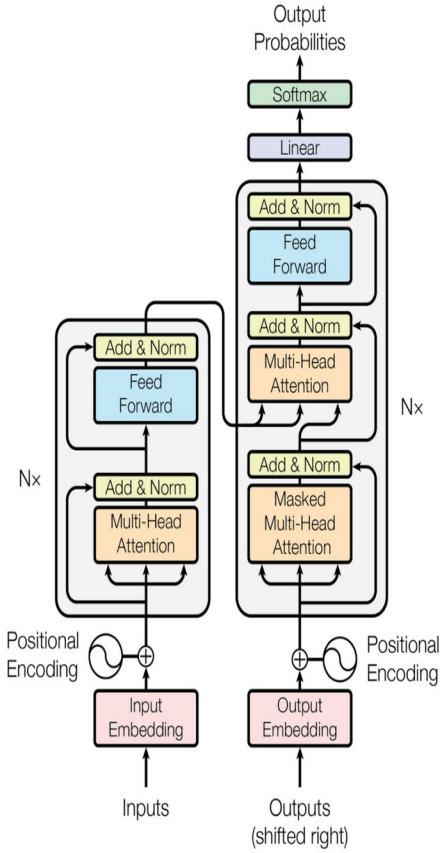


Figure 1: The Transformer - model architecture.

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

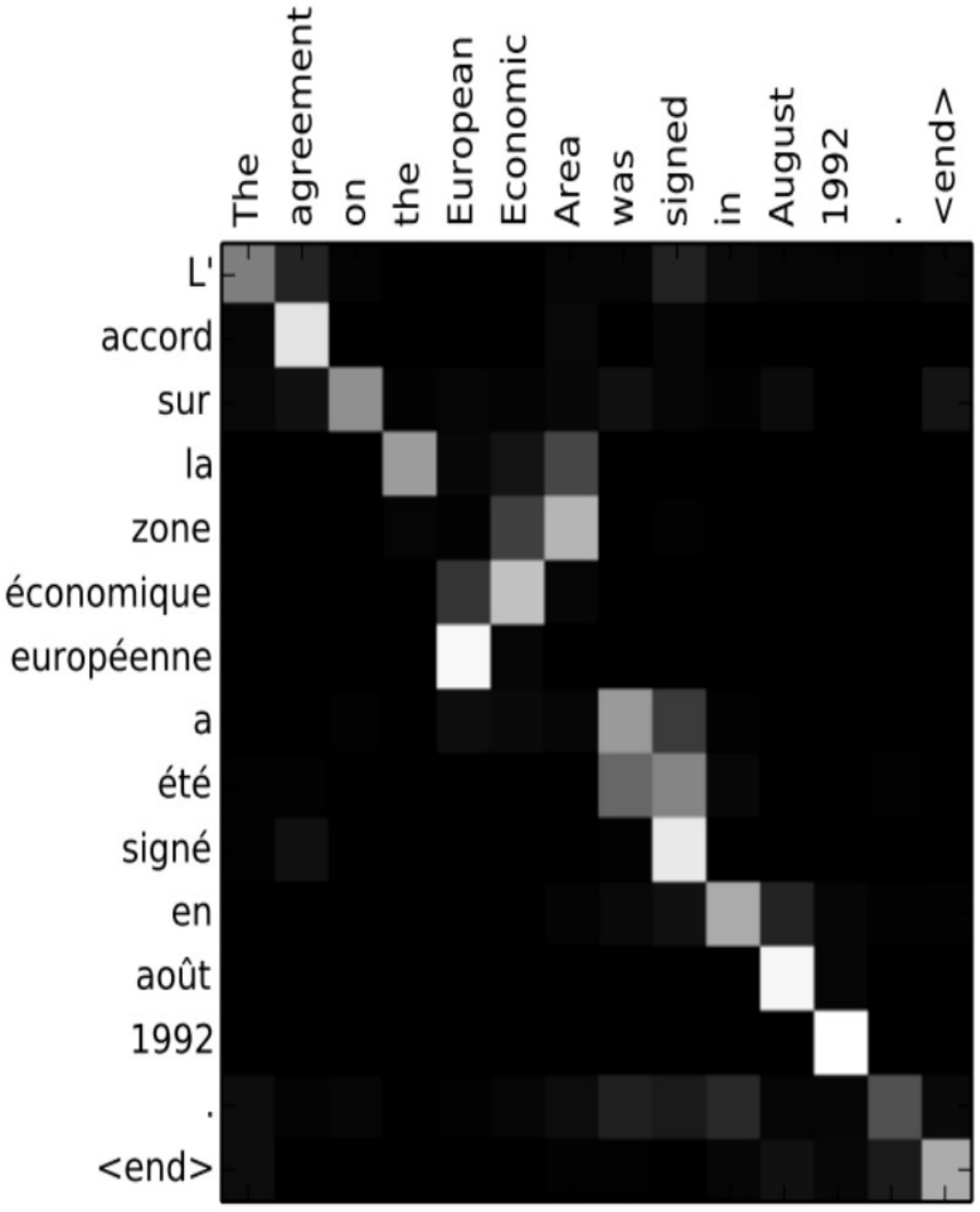
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Transformers

¿Qué es la atención?

Atención en modelos recurrentes:



Orígenes de la atención

Recurrent Models of Visual Attention

Volodymyr Mnih **Nicolas Heess** **Alex Graves** **Koray Kavukcuoglu**
Google DeepMind

Generating Sequences With Recurrent Neural Networks

Alex Graves
Department of Computer Science
University of Toronto
graves@cs.toronto.edu

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

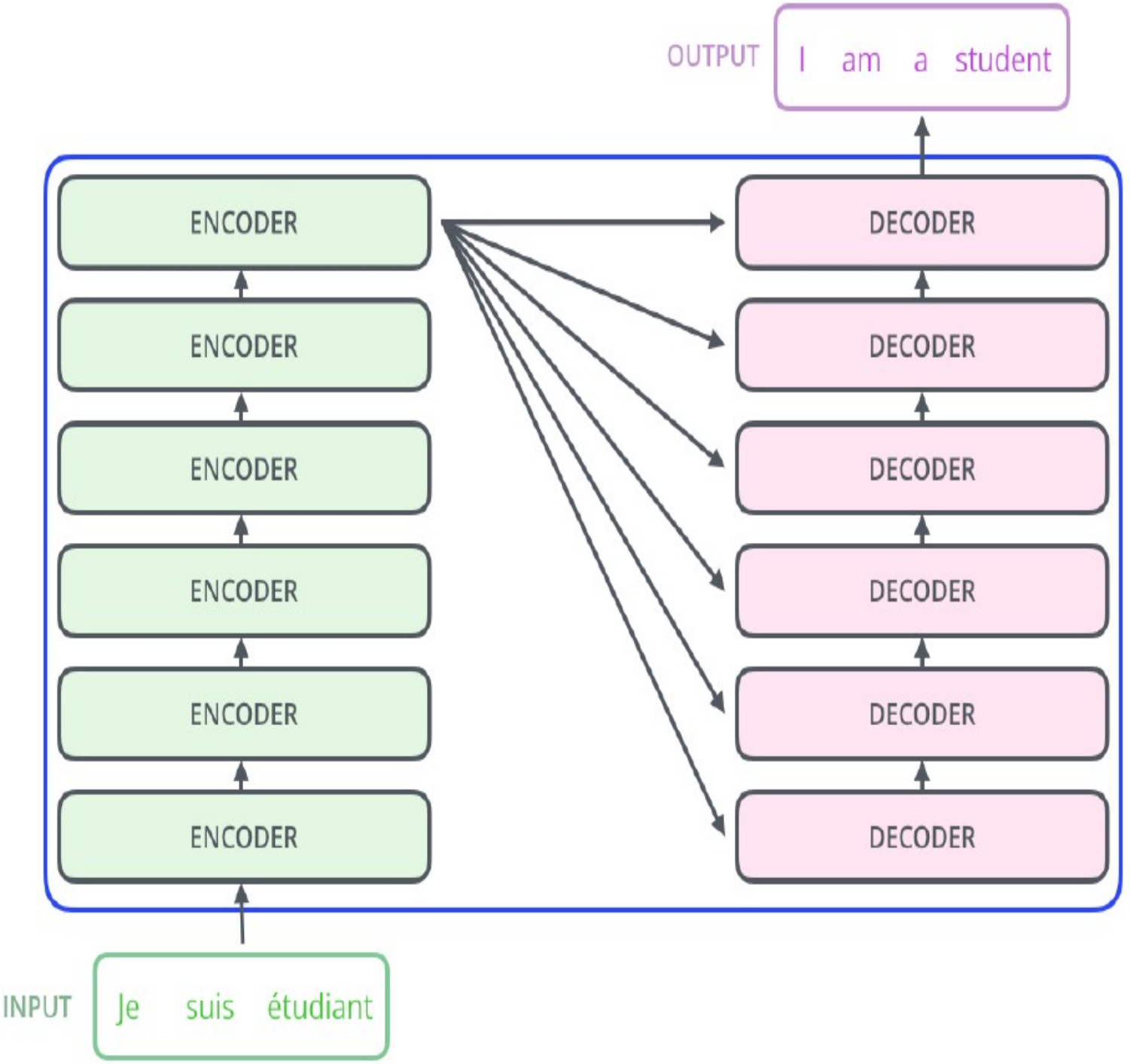
Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

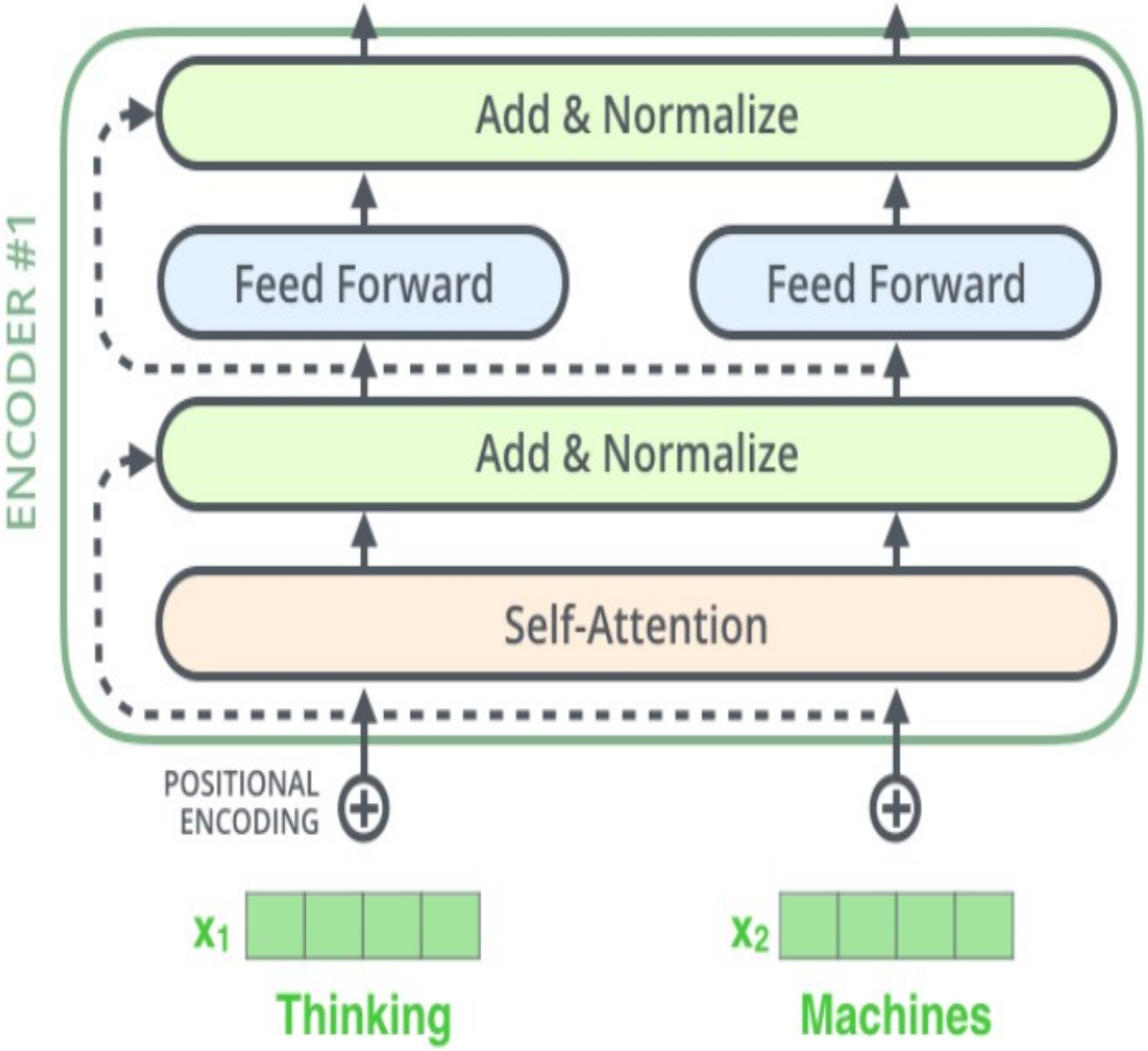
Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong **Hieu Pham** **Christopher D. Manning**
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang,hyhieu,manning}@stanford.edu

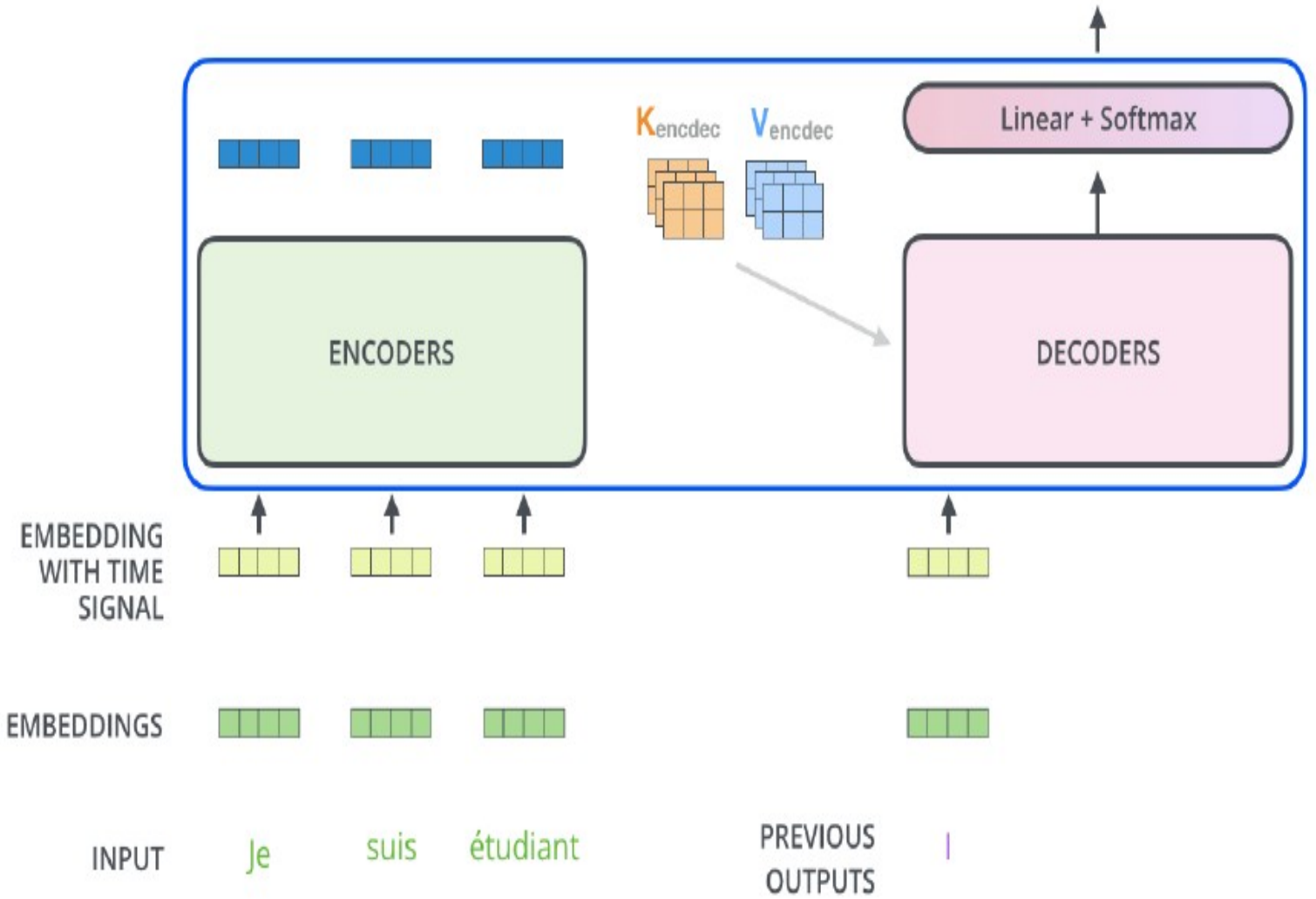
El transformer con atención


OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant

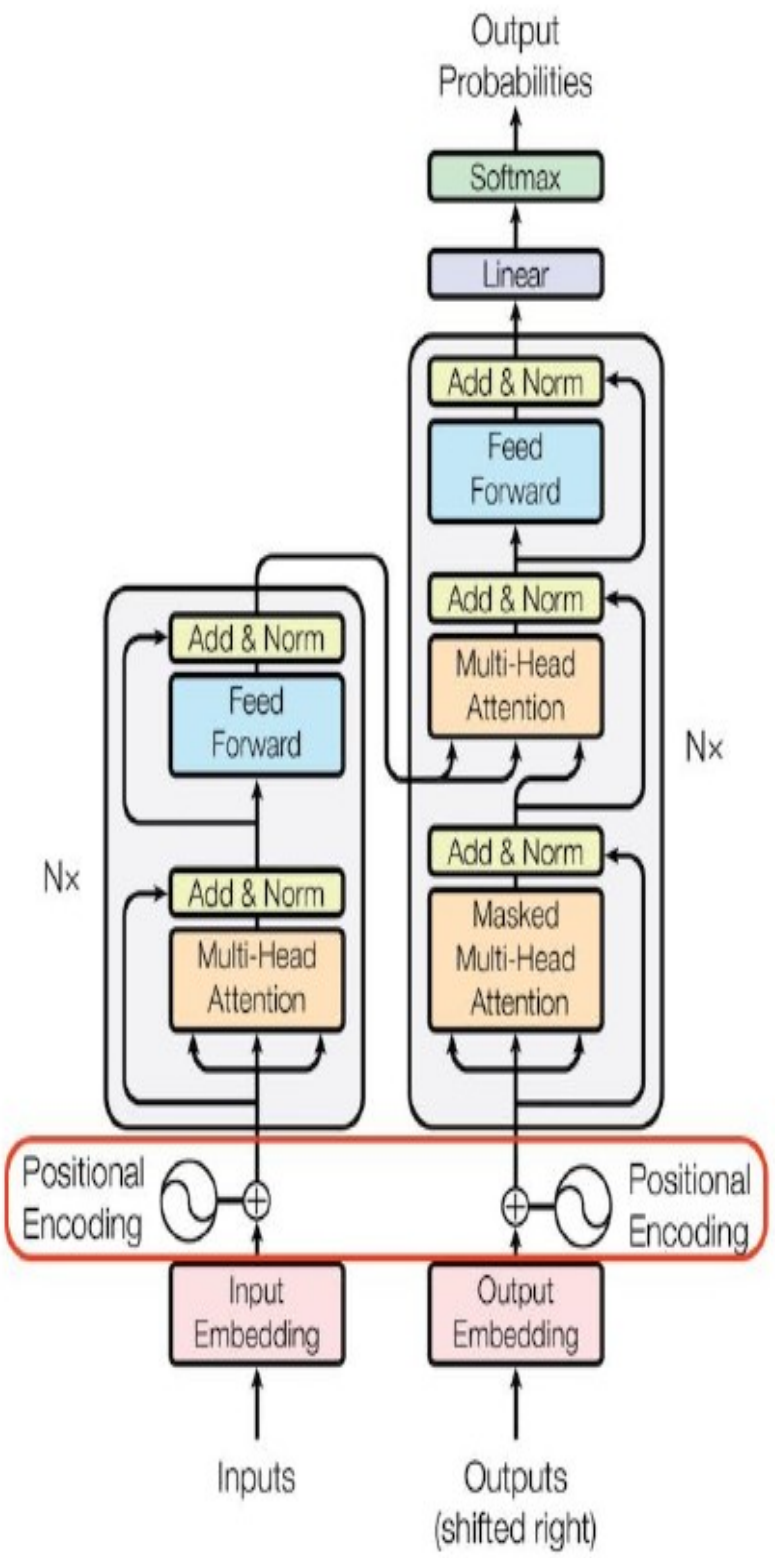
Estructura de cada Encoder



¿Cómo funciona el decoder?



La estructura completa



El transformer todo

- Input: en el encoder una oración. Decoder. La oración traducida corrida hacia la derecha: la tarea es predecir la secuencia una palabra a la vez.
- Con todo el input en el encoder y el código <INICIO> en el decoder se predice la siguiente palabra.
- Se agrega la predicción al input y se predice la segunda.
- Se sigue así hasta que se predice el símbolo <FIN>

El transformer todo: entrenamiento

- Utilizar lo mismo que en el funcionamiento, el símbolo START.
- Aplicar backpropagation (con todos los chiches).
- Usar en el input la palabra correcta y no la predicha para predecir la segunda.
- Corregir nuevamente.
- Así hasta que ande bien.

La bestia GPT-3

La bestia GPT-3

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

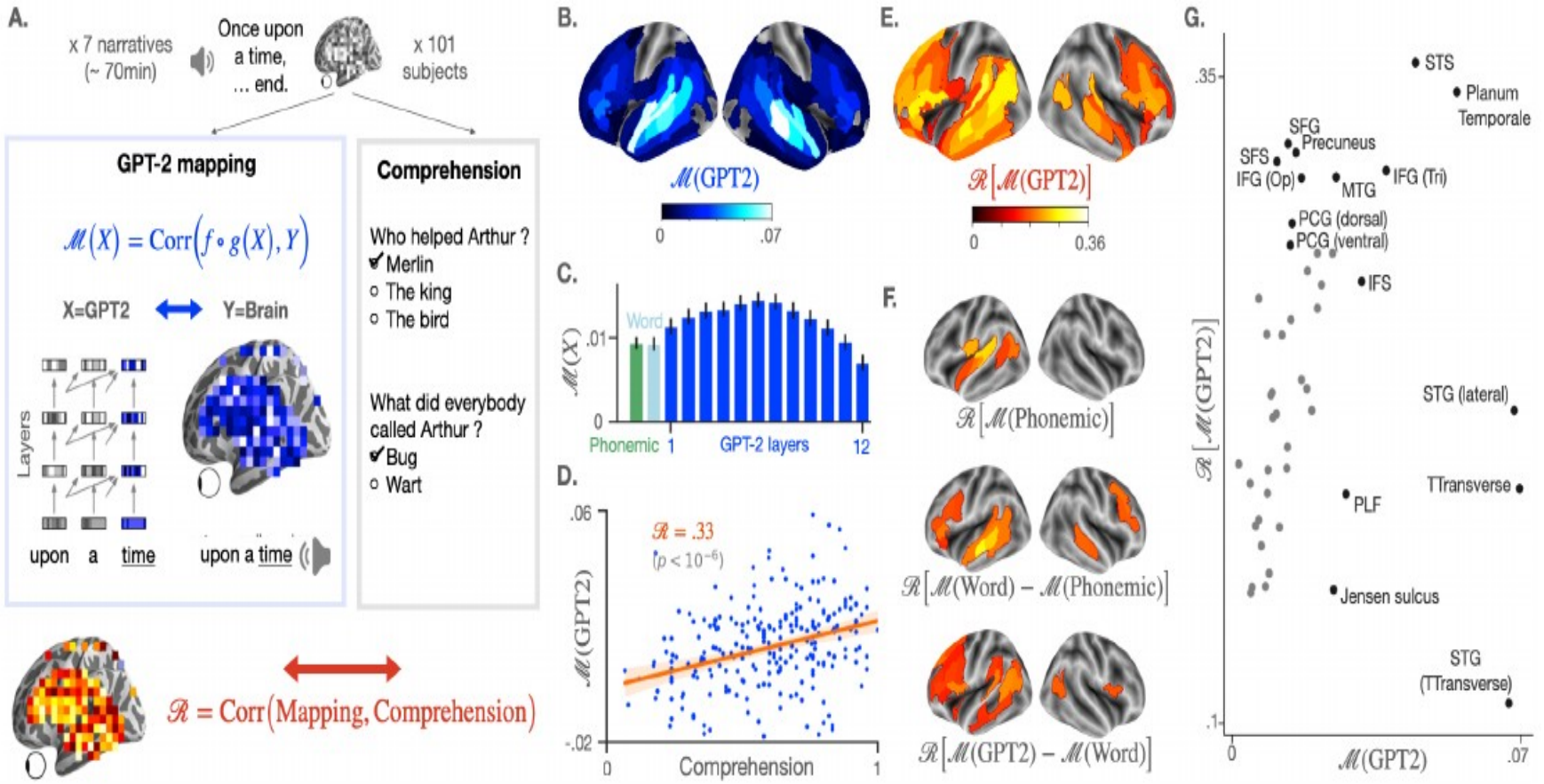
Dataset	# Tokens (Billions)
Total	499
Common Crawl	410
WebText2	19
Books1	12
Books2	55
Wikipedia	3

La vuelta al cerebro

GPT-2's activations predict the degree of semantic comprehension in the human brain

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

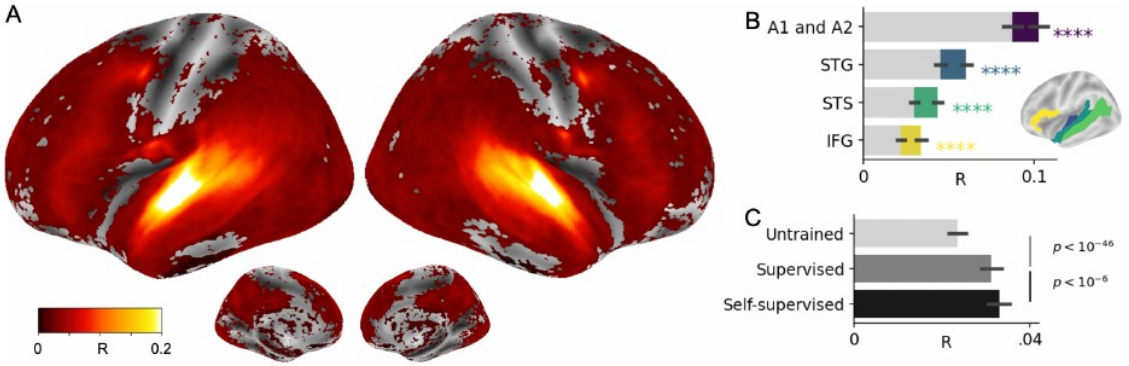
[1]Facebook AI Research, Paris, France; [2]Université Paris-Saclay, Inria, CEA, Palaiseau, France; [3]École normale supérieure, PSL University, CNRS, Paris, France

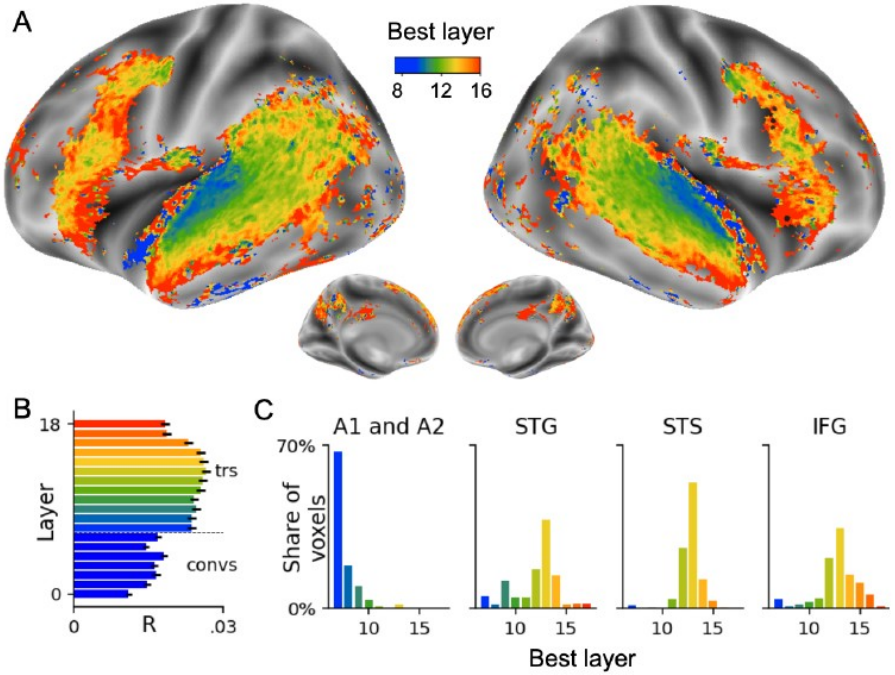


¿Hace lo mismo el cerebro que una red neuronal artificial?

Toward a realistic model of speech processing in the brain with self-supervised learning

Juliette Millet* [1,2,3] Charlotte Caucheteux* [1,4] Pierre Orhan[2] Yves Boubenec[2]
Alexandre Gramfort[4] Ewan Dunbar[2,5] Christophe Pallier[6] Jean-Rémi King[1,2]





1000 niños

- El GPT-3 usa 4,99E11 ejemplos.
- Un niño que escuche 1 palabra por segundo durante 14 hs/día, en 20 años, escucha<4E8 palabras.
- El GPT-3 tiene la experiencia de más de 1000 personas que viven 20 años (o de 20000 mil años).
- El dato más certero (Hart &Risley, 2003): 45 millones los hijos de los profesionales a los 4 años. (225 millones a los 20, con extrapolación lineal).

¿Quién hace la IA?

- A lo largo de la historia, la IA ha intentado emular lo que hacen los humanos.
- En parte inspirado en la Psicología Cognitiva y la Ciencia Cognitiva en los ‘60.
- También inspirado en la neurobiología teórica: los cambios más importantes vienen de los enfoques que intentan interpretar computacionalmente algunas características del sistema nervioso:
    - Neuronas son dispositivos todo o nada (MP).
    - Aprender es cambiar pesos sinápticos.
    - Convolución.
    - Memoria de trabajo.
    - Atención.