



1. Artificial Intelligence Overview

Prof. Jesús Alfonso López Sotelo
jalopez@ua.edu.co

UAO - Universidad Autónoma de Occidente, Cali,
Colombia www.ua.edu.co

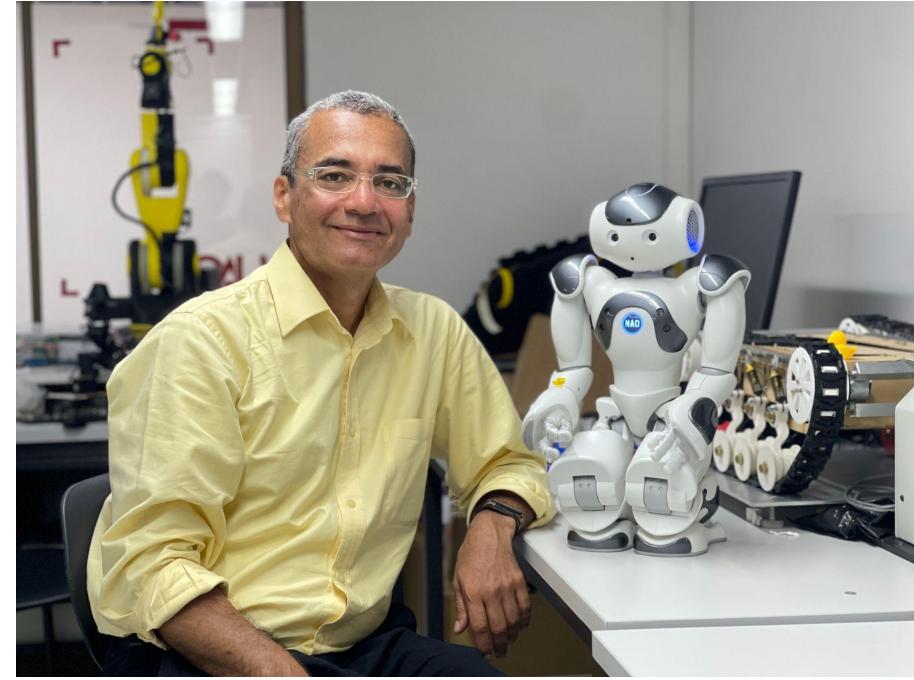


Jesús Alfonso López Sotelo

Nacido en Cali, Colombia. Es Ingeniero Electricista, Magíster en Automática y Doctor en Ingeniería.

Tiene más de 25 años de experiencia en docencia y en desarrollo de proyectos relacionados con Inteligencia Artificial. Sus áreas de interés son las redes neuronales artificiales y el aprendizaje profundo (Deep Learning), la Inteligencia Artificial en dispositivos de borde, los sistemas difusos, la computación evolutiva, la enseñanza de la inteligencia artificial y el impacto que esta tecnología pueda tener en nuestra sociedad.

Es investigador Asociado del sistema nacional de ciencia tecnología e innovación en Colombia de MinCiencias. Es miembro profesional de la IEEE donde pertenece al capítulo nacional de la sociedad de Inteligencia Computacional. Actualmente está vinculado a la Universidad Autónoma de Occidente en Cali y pertenece al Grupo de Investigación en Energías, GIEN. Ha publicado diversos artículos, capítulos de libro y libros en las temáticas de Redes Neuronales Artificiales, Deep Learning y otras técnicas de inteligencia artificial.



Perfil Linkedin

<https://www.linkedin.com/in/jesus-alfonso-lopez-sotelo-76100718/>

Universidad Autónoma de Occidente <https://www.uao.edu.co/>

Cali Colombia

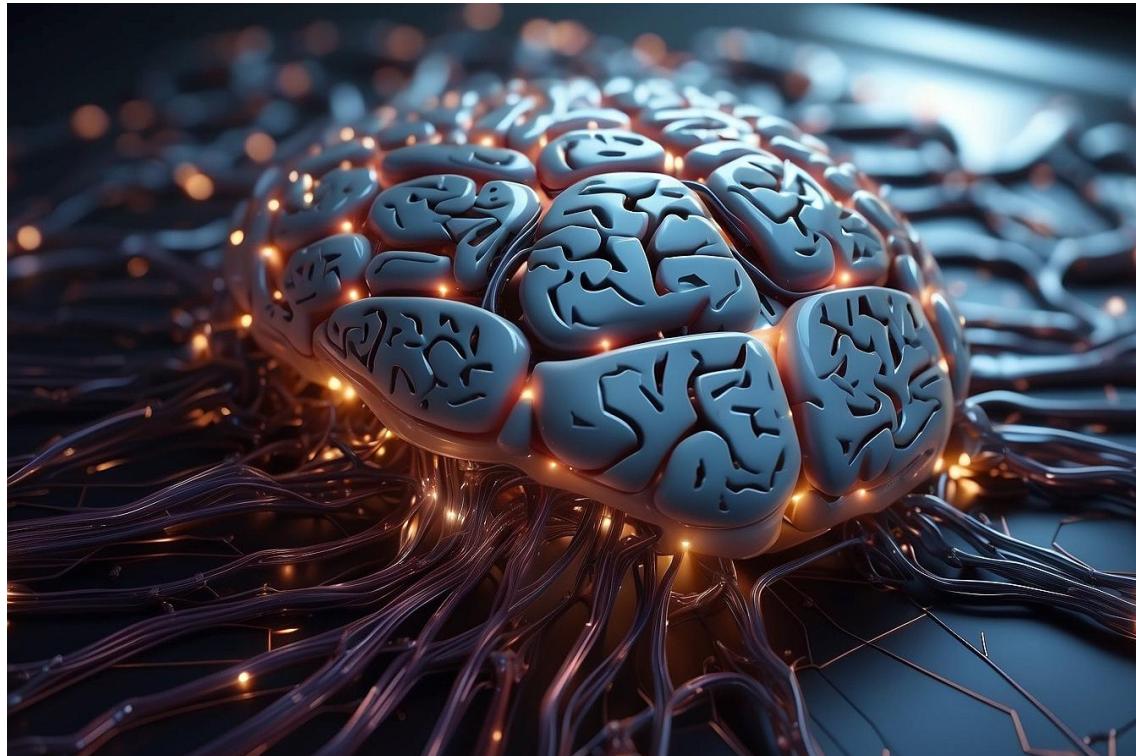


Uno de los campus universitarios
más sostenible del país

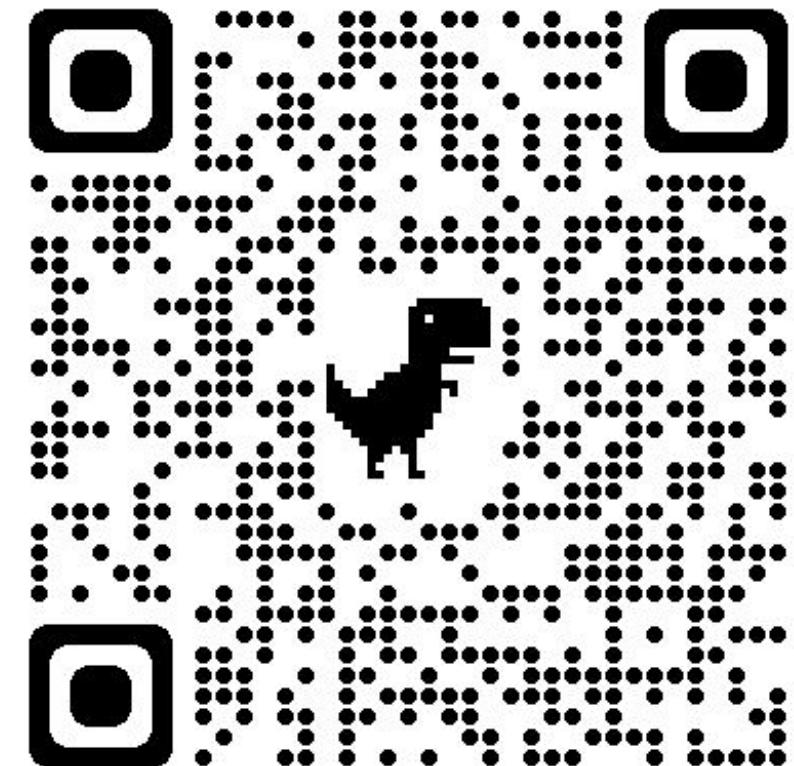
Programas Académicos Activos	72
Tecnológico	3
Profesional	26
Especialización	35
Maestrías	6
Doctorado	2

Especialización en Inteligencia Artificial. Modalidad Combinada

La especialización en Inteligencia Artificial de la UAO te proporciona los conocimientos y herramientas necesarias para el diseño e implementación de proyectos que involucren tecnologías como aprendizaje automático y aprendizaje profundo, considerando el impacto social de los mismos.



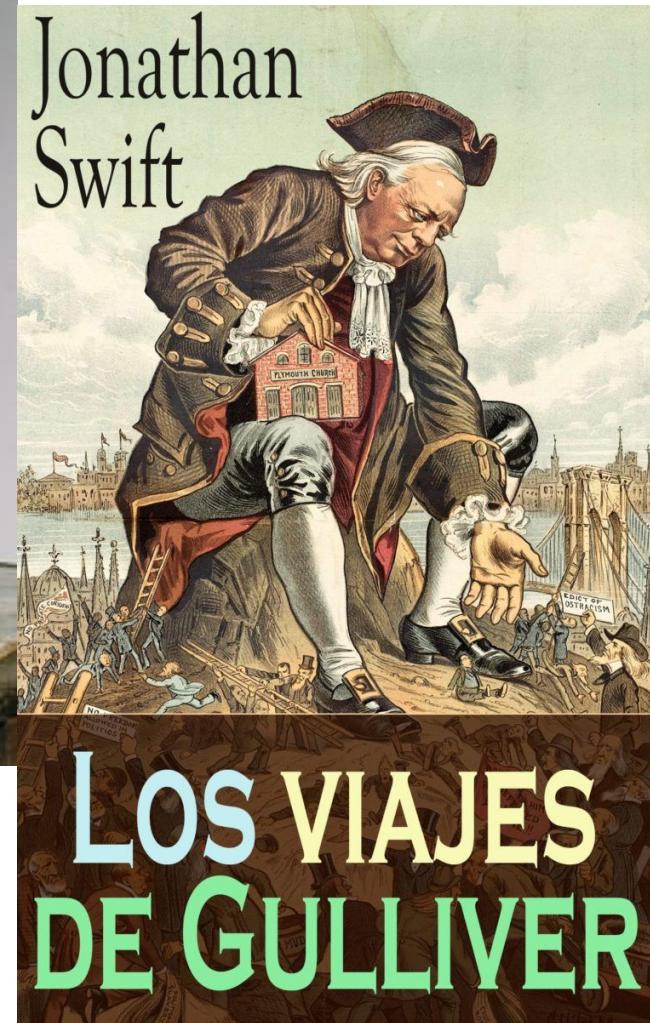
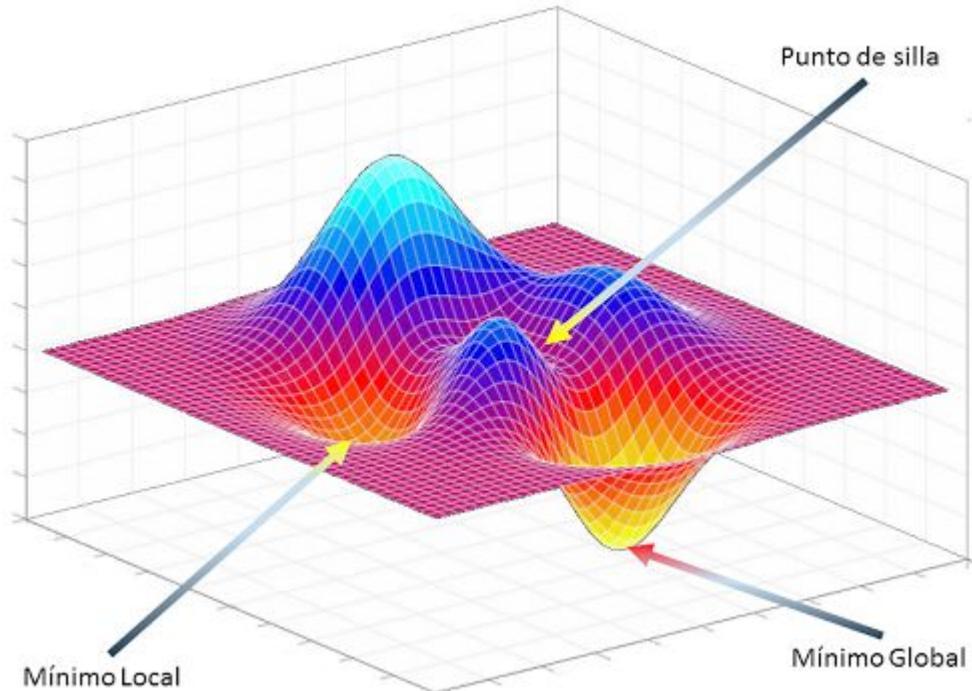
<https://www.uao.edu.co/programa/especializacion-en-inteligencia-artificial/>



¿Qué es la Inteligencia Artificial
(IA) y el Aprendizaje Automático?

¿Qué es Inteligencia Artificial?

Los antecedentes de la IA se pueden rastrear en la filosofía, en la técnica, en la literatura y en la matemáticas.

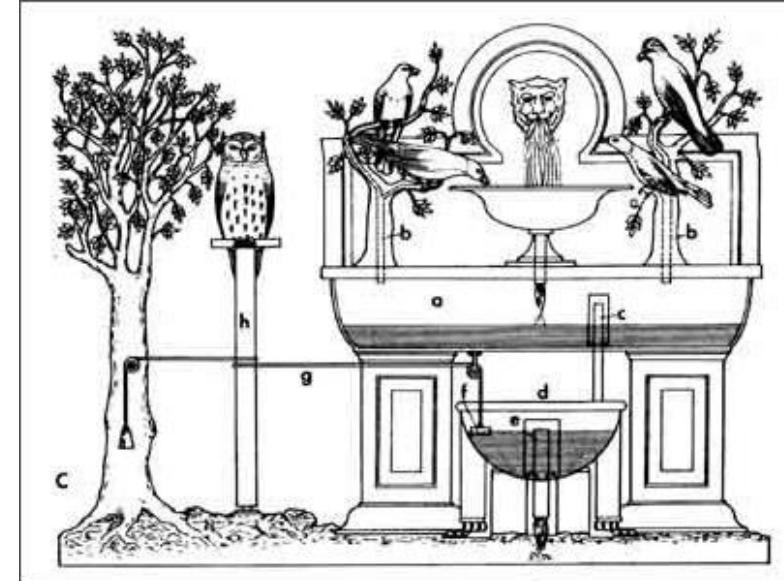


¿Qué es Inteligencia Artificial?

Mito de
Talos



Autómata de Herón



<https://mitologia.guru/personajes-mitologicos/talos/>

http://automata.cps.unizar.es/Historia/Webs/automatas_en_la_historia.htm

¿Qué es Inteligencia Artificial?

Autómatas de Pierre Jacques-Droz



<https://www.xatakaciencia.com/robotica/los-primeros-automatas-de-la-historia>

https://www.youtube.com/watch?v=vr0e_WsjkvY

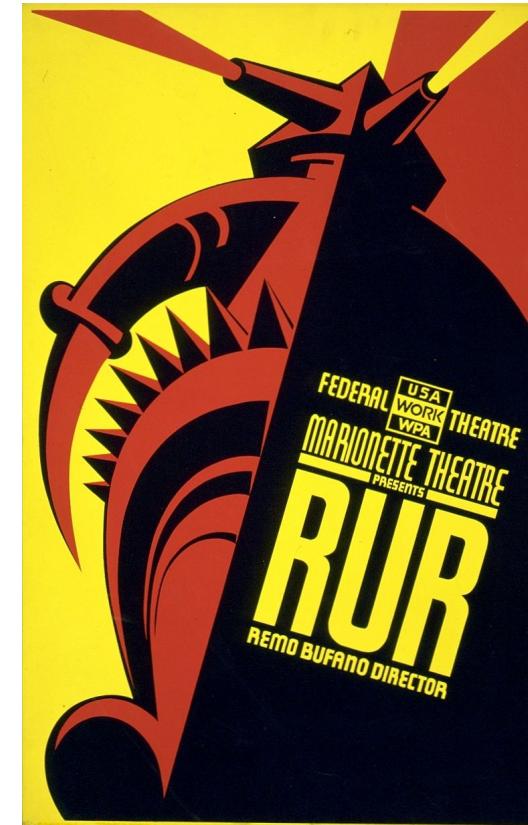
¿Qué es Inteligencia Artificial?

Frankenstein



<https://www.elperiodico.com/es/ciencia/20180309/frankenstein-ciencia-responsabilidad-y-estereotipos-de-genero-6679506>

R.U.R. (Robots Universales Rossum)



[https://es.wikipedia.org/wiki/R.U.R._\(Robots_Universales_Rossum\)](https://es.wikipedia.org/wiki/R.U.R._(Robots_Universales_Rossum))

¿Qué es Inteligencia Artificial?

Tres Leyes de la
Robótica de Asimov



https://historia.nationalgeographic.com.es/a/isaac-asimov-maestro-ciencia-ficción_15035

LEYES DE LA ROBÓTICA
POR ISAAC ASIMOV



1 Un robot no puede dañar a un ser humano ni, por inacción, permitir que un ser humano sufra daño

2 Un robot debe cumplir las órdenes de los seres humanos, excepto si dichas órdenes entran en conflicto con la Primera Ley

3 Un robot debe proteger su propia existencia en la medida en que ello no entre en conflicto con la Primera o la Segunda Ley

IEEE
UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ

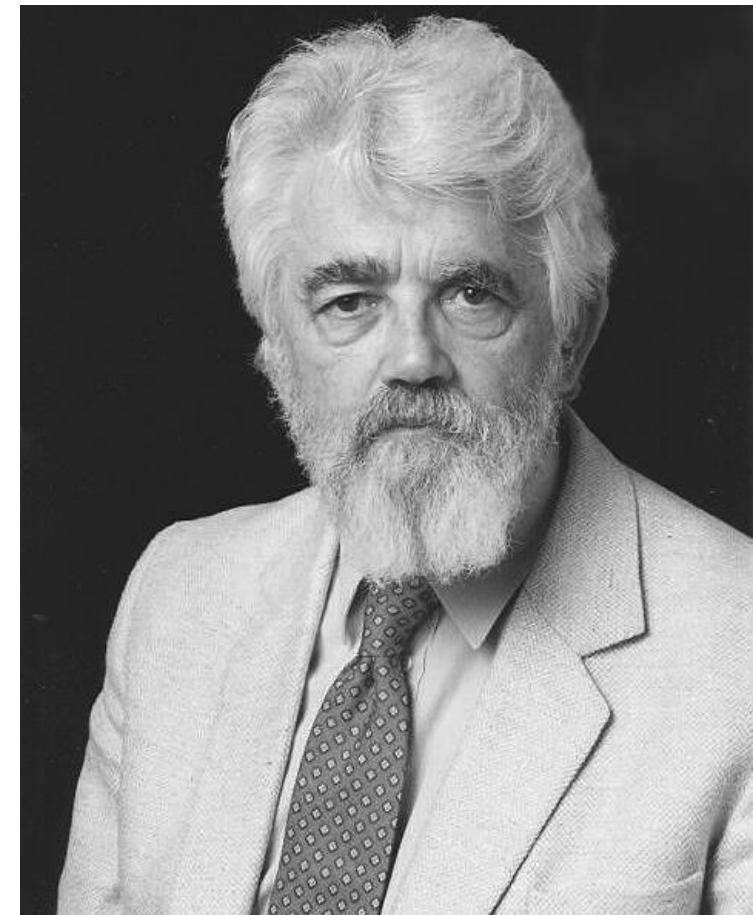


Ceimtuh-RAS

¿Qué es Inteligencia Artificial?

La Escuela de Verano en Dartmouth sobre Inteligencia Artificial (1956) se considera un evento importante en la historia de la IA y donde surgió el término inteligencia artificial, seleccionado por el informático John McCarthy.

La inteligencia artificial (IA) se puede definir como el campo de estudio y desarrollo de sistemas informáticos que pueden realizar tareas que normalmente requieren inteligencia humana. Estas tareas incluyen el aprendizaje, la percepción, el razonamiento, la resolución de problemas y la comprensión del lenguaje natural.



¿Qué es Inteligencia Artificial?



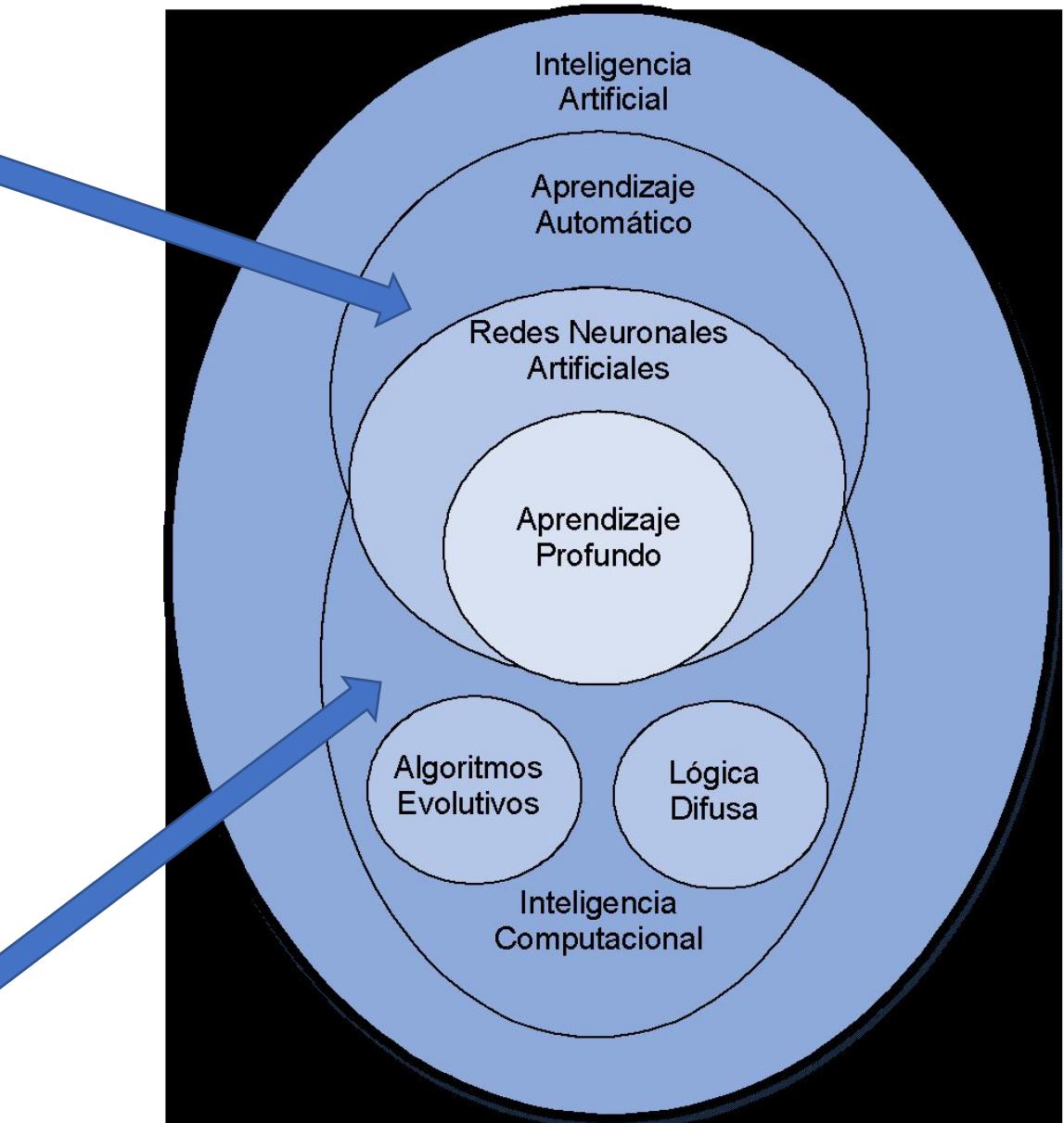
Timeline of AI Development

- 1950s-1960s: First AI boom - the age of reasoning, prototype AI developed
- 1970s: AI winter I
- 1980s-1990s: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- 1990s: AI winter II
- 1997: Deep Blue beats Gary Kasparov
- 2006: University of Toronto develops Deep Learning
- 2011: IBM's Watson won Jeopardy
- 2016: Go software based on Deep Learning beats world's champions

¿Qué es Inteligencia Artificial?

Machine Learning
(Aprendizaje Automático)

Algoritmos Bio-inspirados



Aprendizaje Automático

Datos es lo que Hay!

El aprendizaje automático es un subconjunto de la inteligencia artificial que tiene la capacidad de "aprender" (es decir, mejorar progresivamente el rendimiento en una tarea específica) con datos, sin ser programado explícitamente

Esto sucede en Internet en un minuto

Estimación de una selección de actividades y datos generados online en un minuto en 2021



Fuente: Lori Lewis vía AllAccess



Aprendizaje Automático

**Cuando tu amo viene
con olor a gato**



- Falta mucho para el viernes?
- Wei, apenas es Martes!
- ...



ia stoy arto

Detector de Firulais y Michis

Cuando en casa comienzan a decir que el perro sólo se echa en la alfombra y hay que regarlarlo

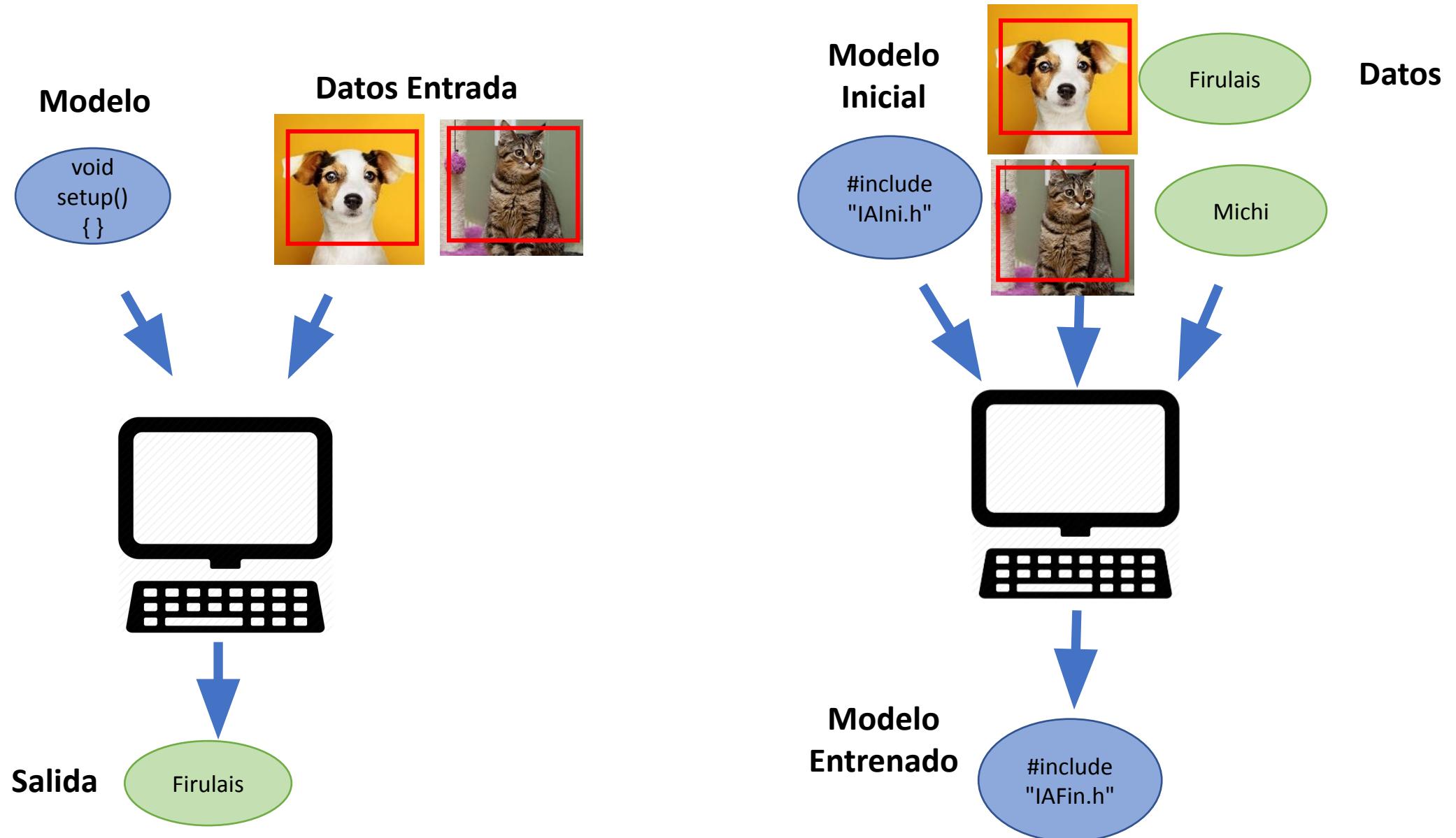


People with no idea about AI, telling me my AI will destroy the world



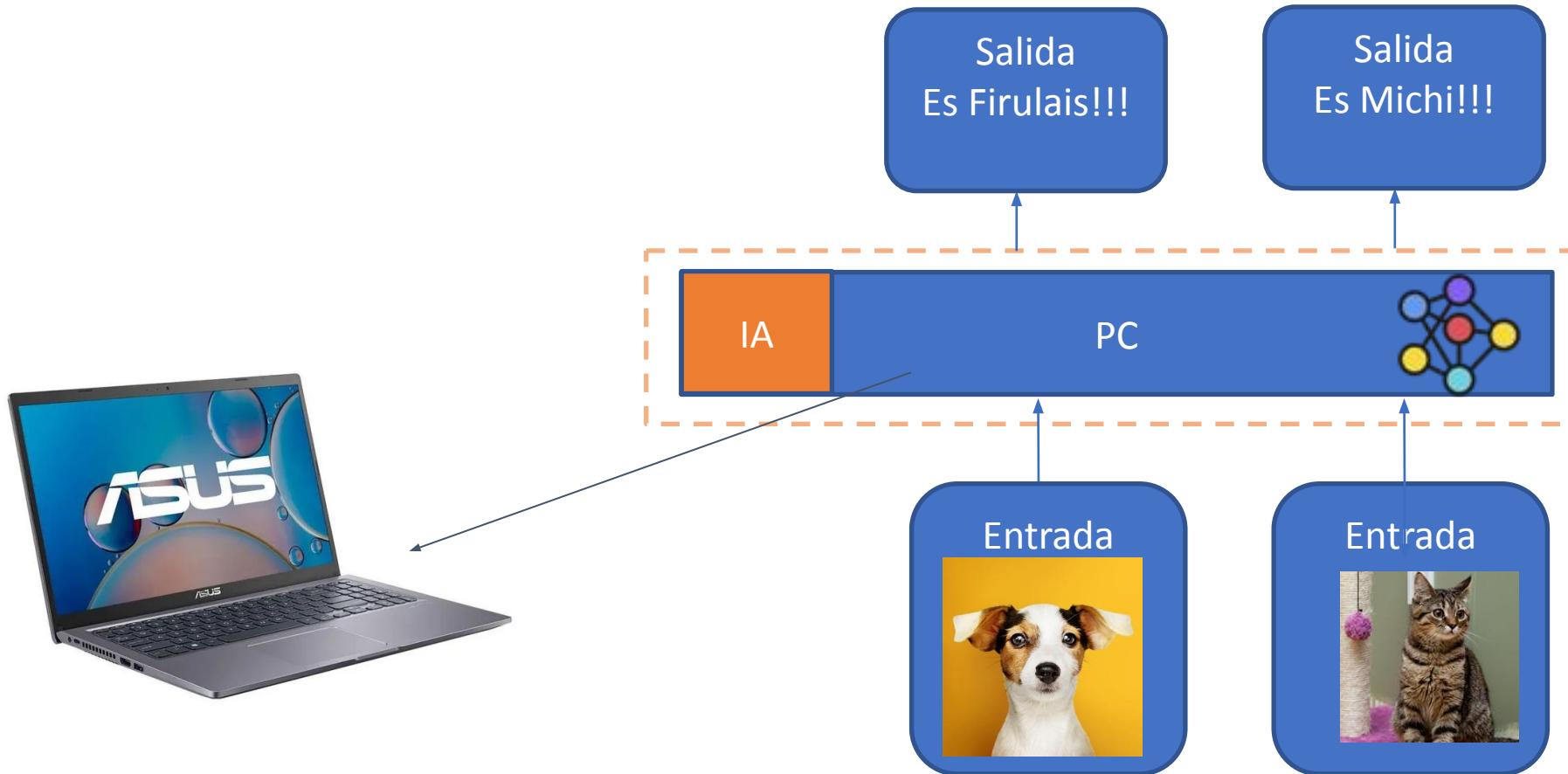
Me wondering why my neural network is classifying a cat as a dog..

Aprendizaje Automático



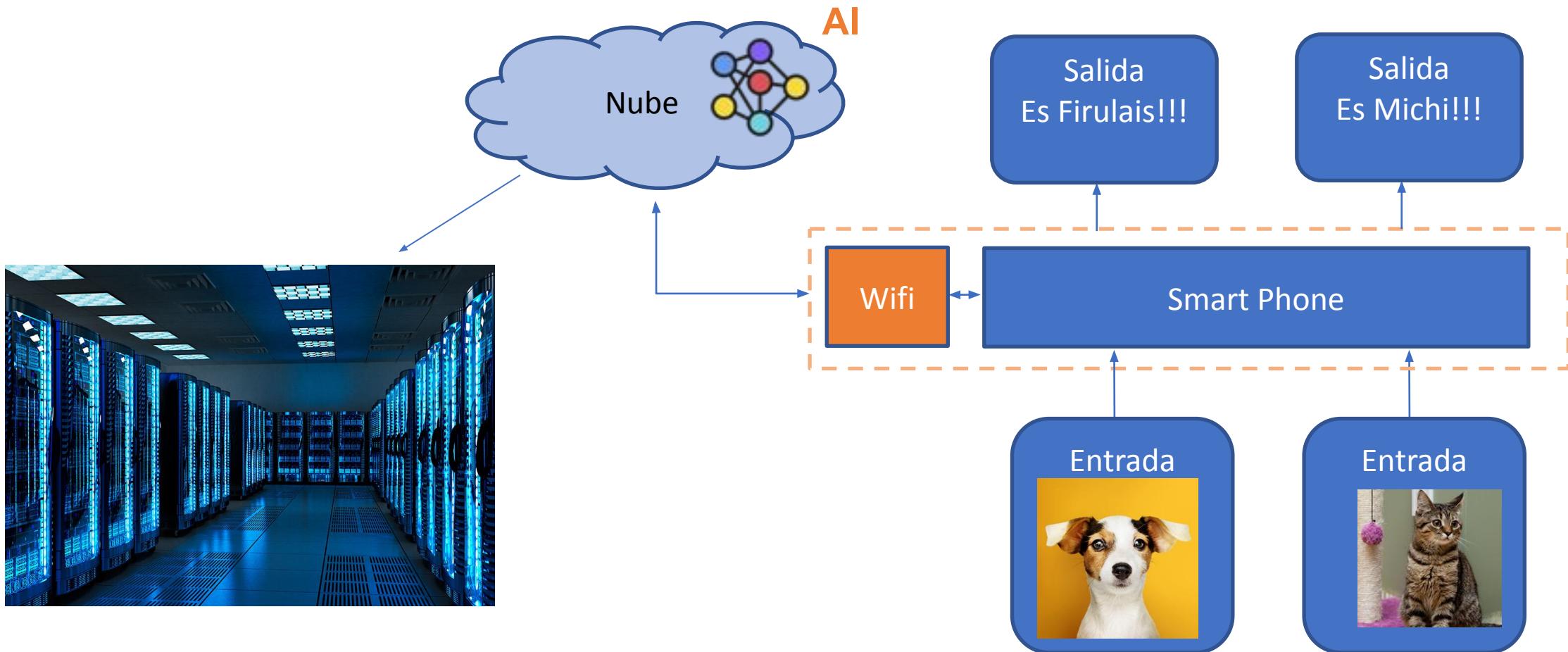
¿Cómo Utilizo el Modelo?

Despliegue Local (PC o Web Page)



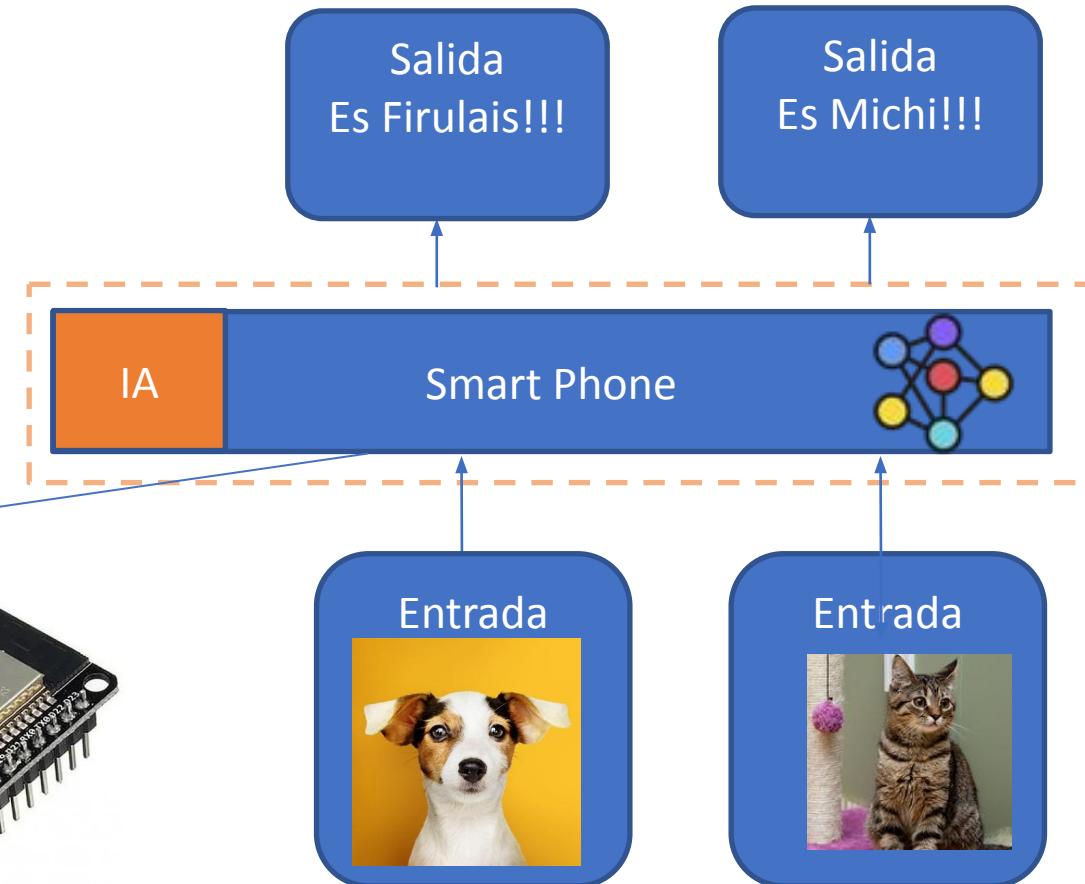
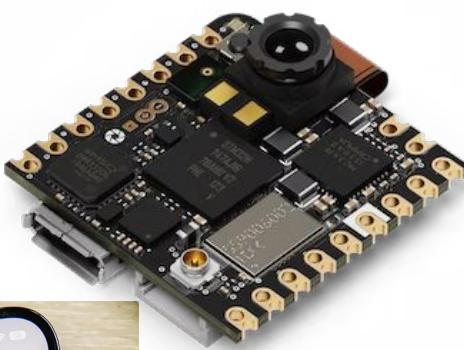
¿Cómo Utilizo el Modelo?

Inteligencia Artificial en la Nube



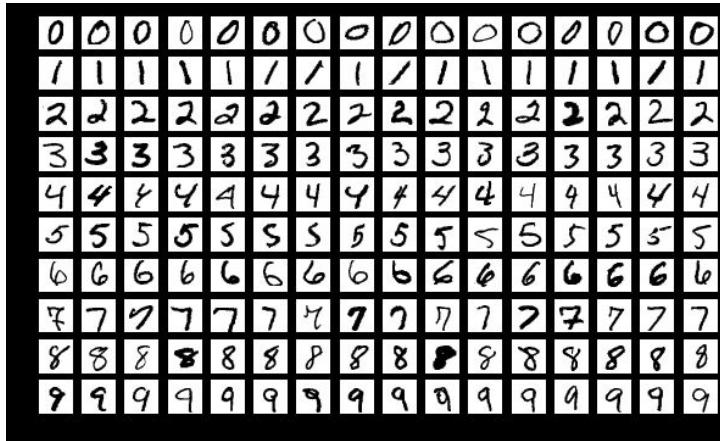
¿Cómo Utilizo el Modelo?

Edge AI



Facilitadores del Estado Actual de la Inteligencia Artificial

Accesibilidad a diversos data sets



By Josef Steppan - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=6481004>



(<http://image-net.org/>)



<https://www.facebook.com/kaggle/>

Dataset Search

Probar coronavirus covid-19 o global temperatures.

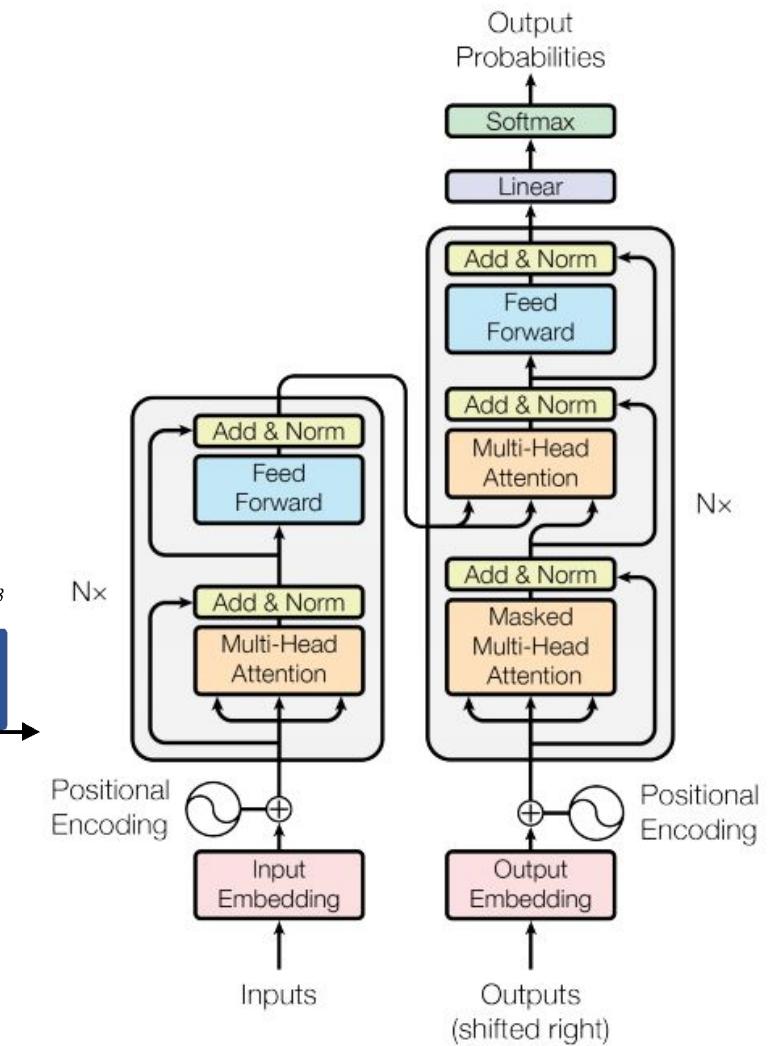
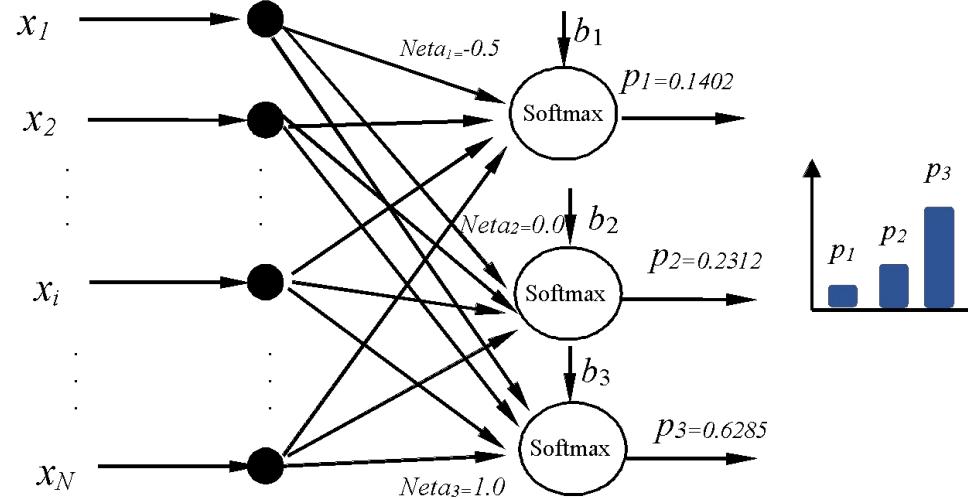
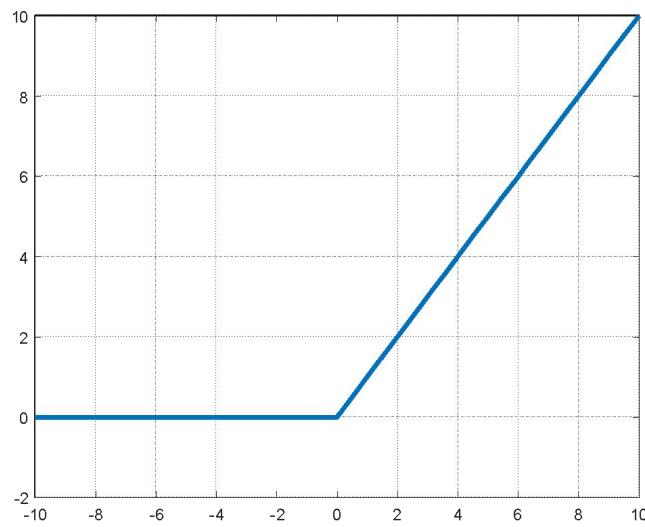
Obtén más información sobre cómo incluir tus conjuntos de datos en Búsqueda de Datasets.

<https://datasetsearch.research.google.com/>

Facilitadores del Estado Actual de la Inteligencia Artificial

Mejoras conceptuales

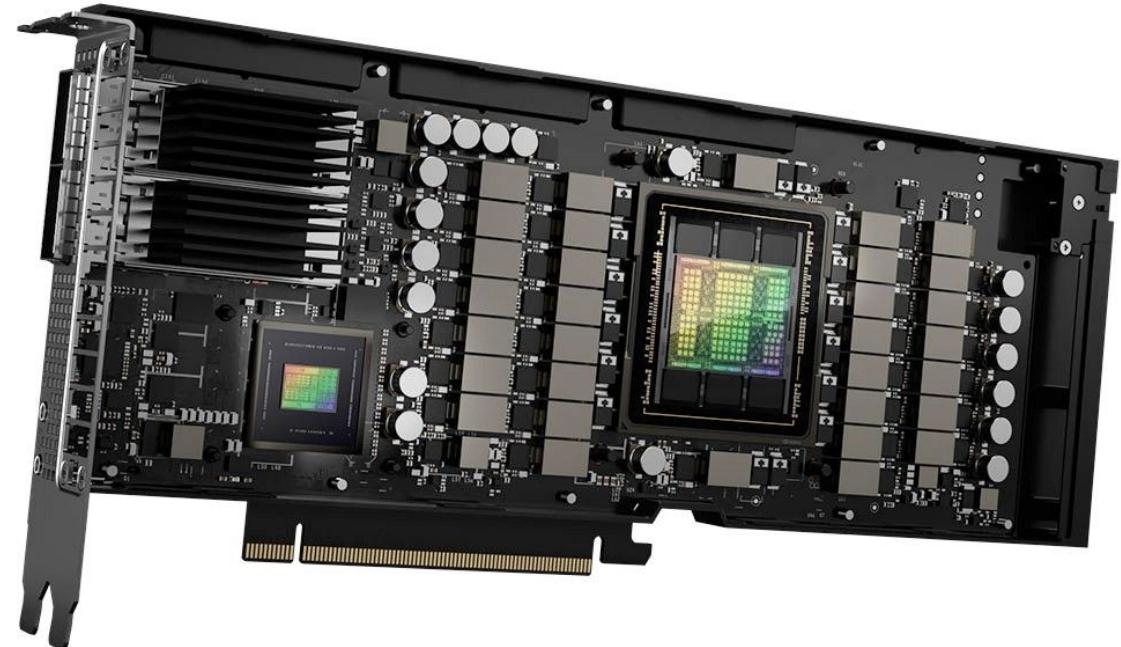
- Entrenamiento por lotes
- Mejoras en las arquitecturas (p.e. Nuevas funciones de activación)
- Nuevas arquitecturas



Facilitadores del Estado Actual de la Inteligencia Artificial

Se requiere una alta capacidad de cómputo

- Uso de GPU
- GPU especiales para data centers



<http://www.funkykit.com/news/pc-computers/nvidia-abandon-3-way-4-way-sli-configurations/>

<https://www.viperatech.com/product/nvidia-h100-tensor-core-gpu/>

Facilitadores del Estado Actual de la Inteligencia Artificial

Otros elementos

- Open source
- Software especializado (Diferenciación automática)



<https://arxiv.org/>



<https://github.com/>



Papers With Code

<https://paperswithcode.com/>



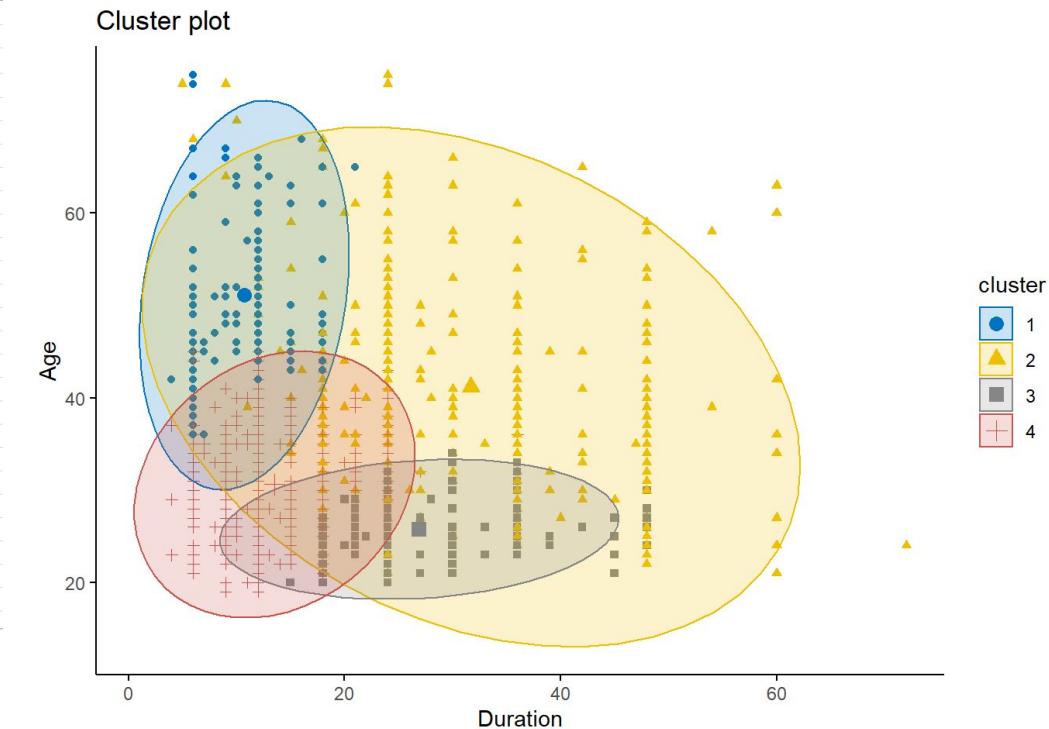
Hugging Face

<https://huggingface.co/>

Datos Para Modelos de IA

Datos Tabulados o Estructurados

Credit Risk Data											
Loan Purpose	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
Small Appliance	\$0	\$739	13	12	M	Single	23	Own	3	Unskilled	Low
Furniture	\$0	\$1,230	25	0	M	Divorced	32	Own	1	Skilled	High
New Car	\$0	\$389	19	119	M	Single	38	Own	4	Management	High
Furniture	\$638	\$347	13	14	M	Single	36	Own	2	Unskilled	High
Education	\$963	\$4,754	40	45	M	Single	31	Rent	3	Skilled	Low
Furniture	\$2,827	\$0	11	13	M	Married	25	Own	1	Skilled	Low
New Car	\$0	\$229	13	16	M	Married	26	Own	3	Unskilled	Low
Business	\$0	\$533	14	2	M	Single	27	Own	1	Unskilled	Low
Small Appliance	\$6,509	\$493	37	9	M	Single	25	Own	2	Skilled	High
Small Appliance	\$966	\$0	25	4	F	Divorced	43	Own	1	Skilled	High
Business	\$0	\$989	49	0	M	Single	32	Rent	2	Management	High
New Car	\$0	\$3,305	11	15	M	Single	34	Rent	2	Unskilled	Low
Business	\$322	\$578	10	14	M	Married	26	Own	1	Skilled	Low
New Car	\$0	\$821	25	63	M	Single	44	Own	1	Skilled	High
New Car	\$396	\$228	13	26	M	Single	46	Own	3	Unskilled	Low
Used Car	\$0	\$129	31	8	M	Divorced	39	Own	4	Management	Low
Furniture	\$652	\$732	49	4	F	Divorced	25	Own	2	Skilled	High
New Car	\$708	\$683	13	33	M	Single	31	Own	2	Skilled	Low
Repairs	\$207	\$0	28	116	M	Single	47	Own	4	Skilled	Low
Education	\$287	\$12,348	7	2	F	Divorced	23	Rent	2	Skilled	High
Furniture	\$0	\$17,545	34	16	F	Divorced	22	Own	4	Skilled	High
Furniture	\$101	\$3,871	13	5	F	Divorced	26	Rent	4	Skilled	High
Furniture	\$0	\$0	25	23	M	Married	19	Own	4	Skilled	High
Furniture	\$0	\$485	37	23	F	Divorced	27	Own	2	Management	High

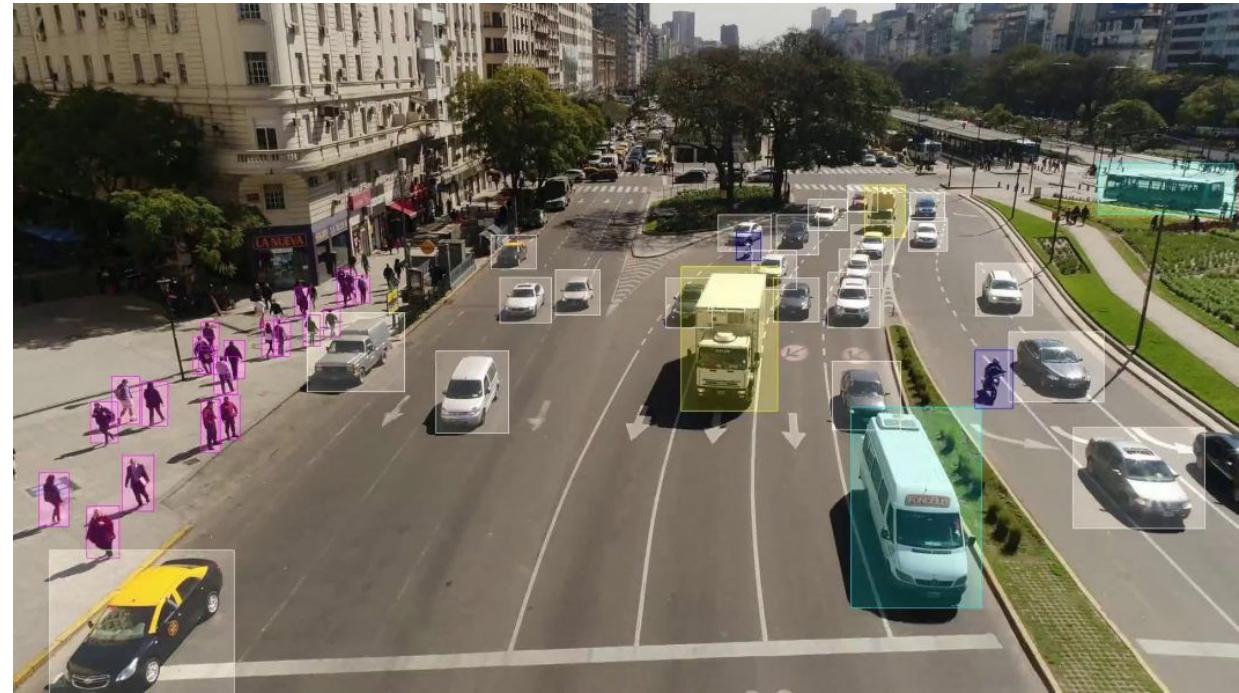


<https://media.cheggcdn.com/media/d52/d52c60c8-60d4-4e55-882f-3ed24306f8cb/phpR8NHxM>

<https://rpubs.com/sid9715/580607>

Datos Para Modelos de IA

Datos en Imágenes



[https://www.youtube.com/watch?
v=KS_4xjXNTxg&](https://www.youtube.com/watch?v=KS_4xjXNTxg&)

<https://viso.ai/applications/computer-vision-applications/>

Datos Para Modelos de IA

Datos de Lenguaje (hablado y escrito)



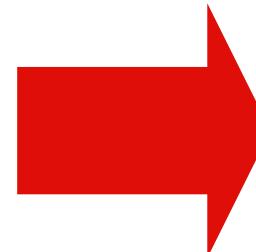
<https://www.grupoftp.com/noticias/el-futuro-de-los-chatbots/>



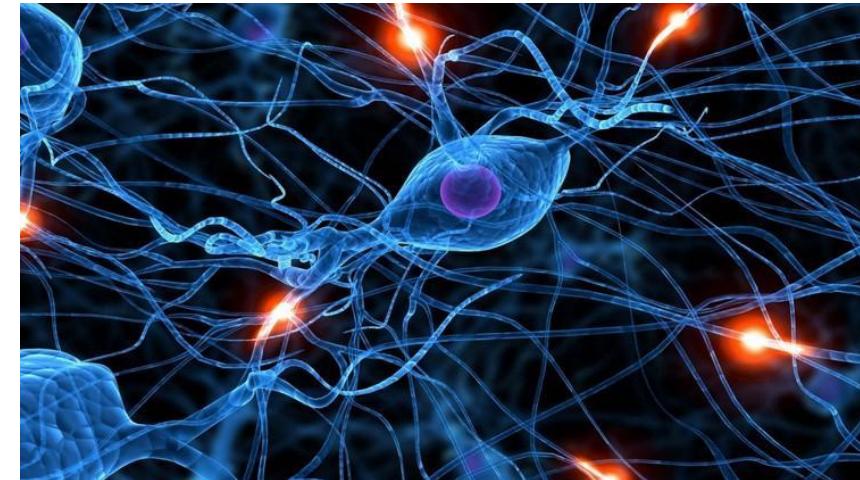
<https://analyticsindiamag.com/google-translate-machine-learning/>

Redes Neuronales Artificiales Alias Aprendizaje Profundo (Deep Learning)

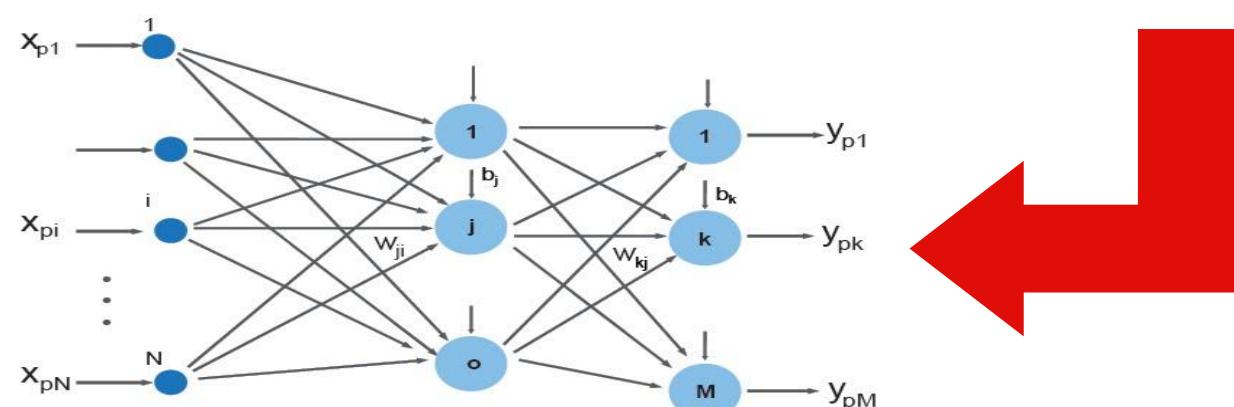
Redes Neuronales Artificiales (Alias Deep Learning)



<https://medium.com/espanol/avances-en-redes-neuronales-705c2efe53d2>



<https://medicine.wustl.edu/news/slow-steady-waves-keep-brain-humming/>

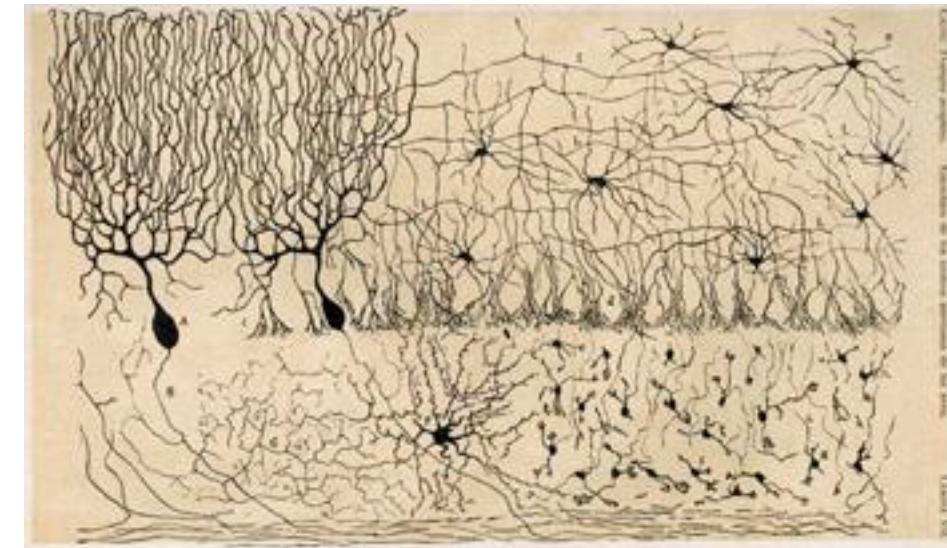
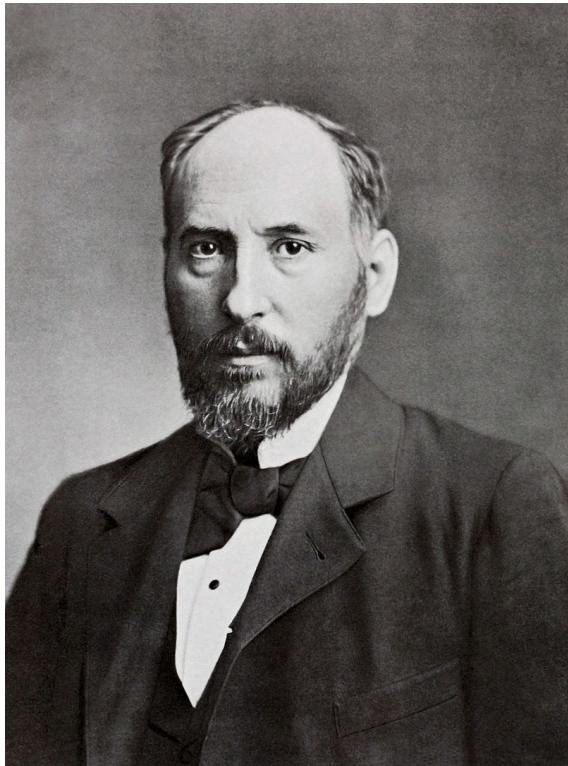


Fuente: Deep Learning. Teoría y Aplicaciones. Jesus Alfonso López. Alpha Editorial 2021

Los Bloques Básicos del Deep Learning

Capas Densas

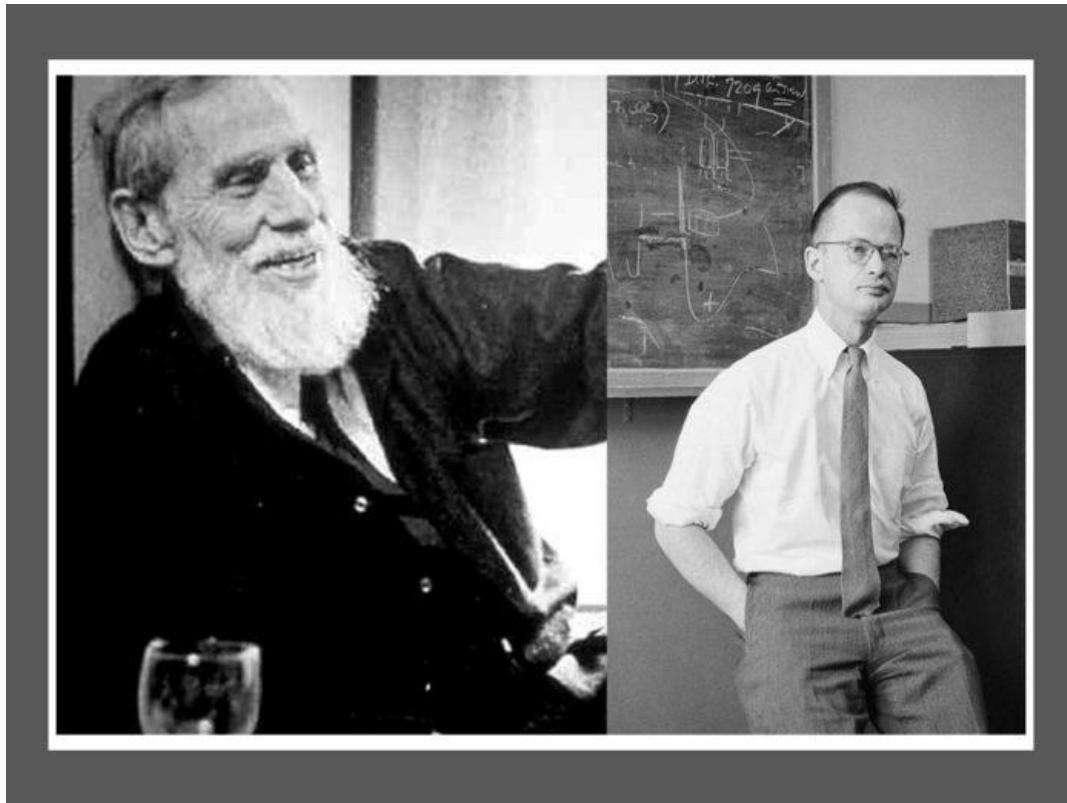
La Doctrina de la Neurona
Santiago Ramón y Cajal



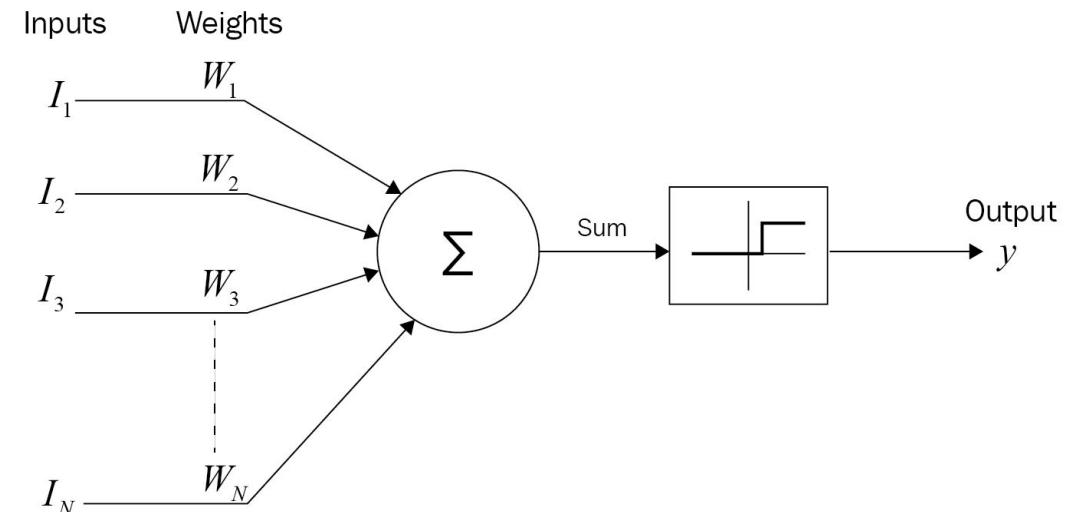
https://es.wikipedia.org/wiki/Doctrina_de_la_neurona

Capas Densas

McCulloch and Pitts



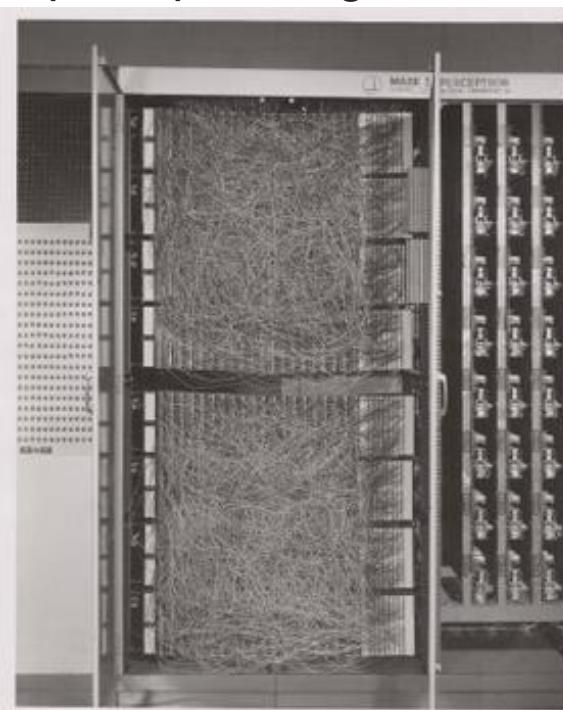
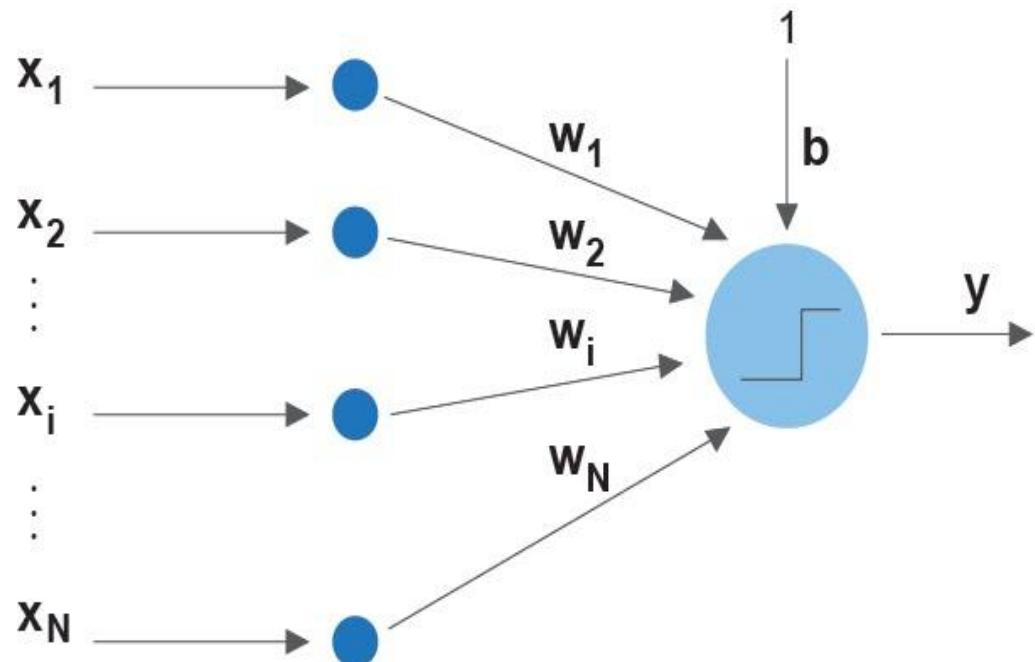
Modelo de Neurona Artificial



<https://machinelearningknowledge.ai/timeline/mcculloch-pitts-neuron-the-beginning/>

Capas Densas

Perceptron



The Mark I Perceptron machine was the first implementation of the perceptron algorithm

Frank Rosenblatt

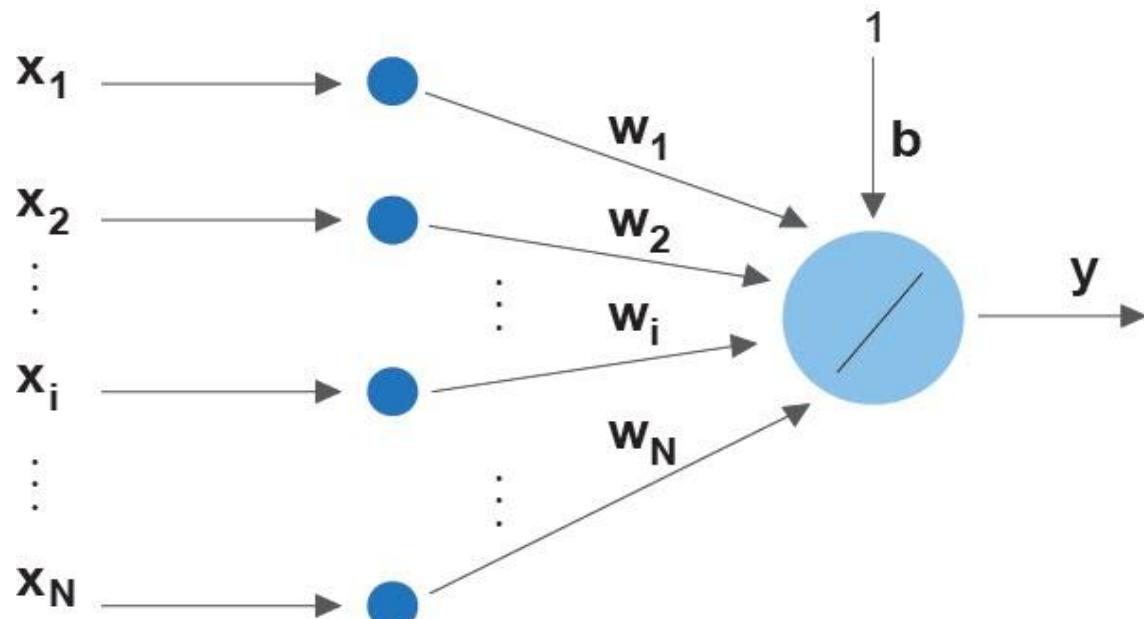


<https://en.wikipedia.org/wiki/Perceptron>

<https://blogs.umass.edu/comphon/2017/06/15/did-frank-rosenblatt-invent-de-learning-in-1962/>

Capas Densas

Adaline. Gradiente Descendente



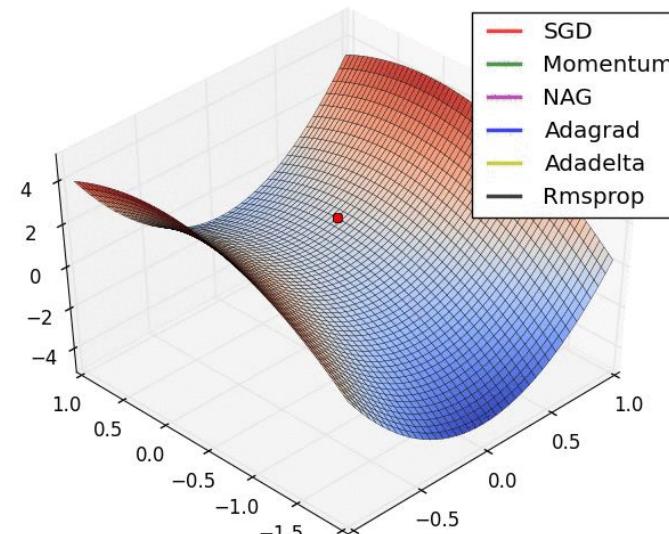
Regla Delta

$$w_i(t + 1) = w_i(t) + \Delta w_i(t)$$

$$w_i(t + 1) = w_i(t) + \alpha(d_i - y_i) x_i$$



<https://medium.com/invisible-illness/the-mountaineer-897387f8a902>



<http://dsdeepdive.blogspot.com/2016/03/optimizations-of-gradient-descent.html>

Capas Densas

Adaline. Gradiente Descendente

Augustin-Louis Cauchy

1847



https://en.wikipedia.org/wiki/Augustin-Louis_Cauchy

Haskell Brooks Curry

1944

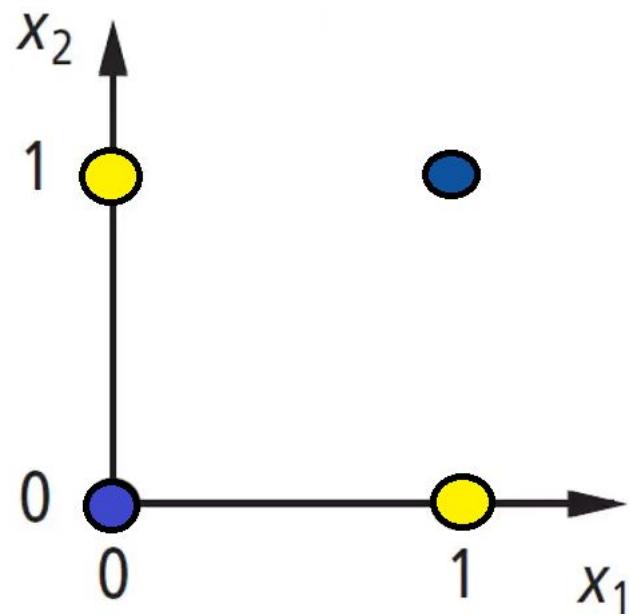


https://en.wikipedia.org/wiki/Haskell_Curry

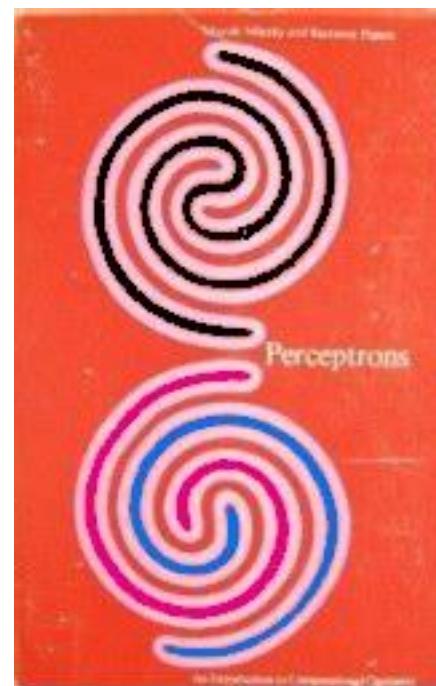
Capas Densas

Problemas de las Redes Monocapa

Problema de la XOR:
Separatividad No Lineal



Marvin Minsky and Seymour Papert

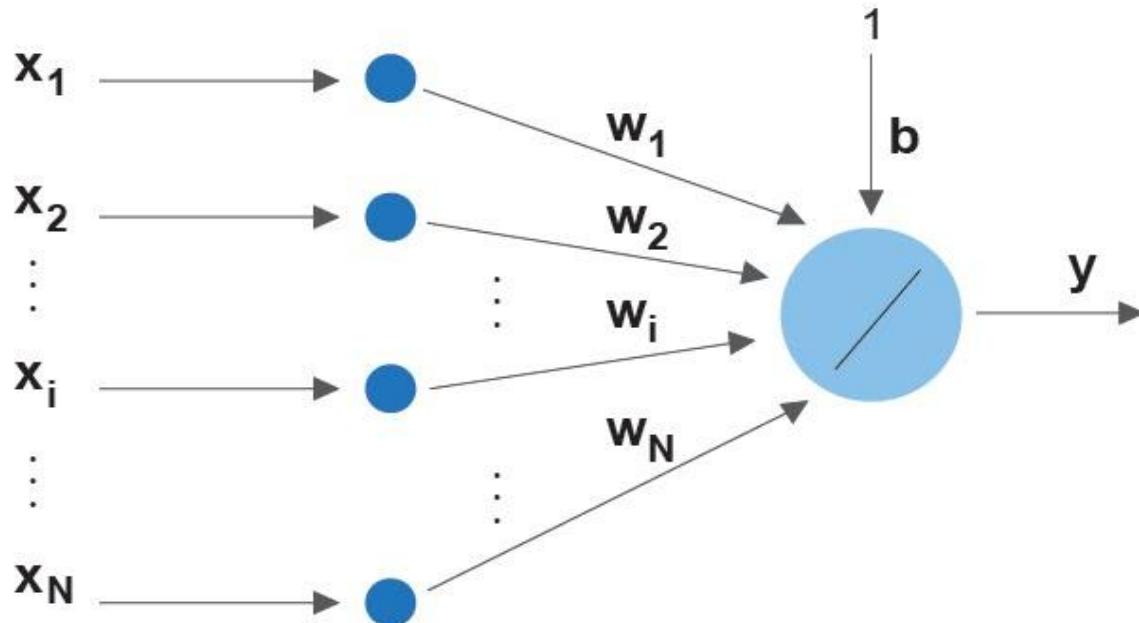


Winter!!!



Capas Densas

Backpropagation



David Rumelhart

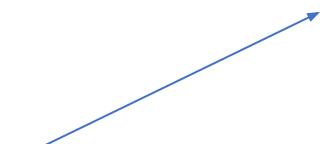


Geoffrey Hinton

Regla Delta Generalizada

$$w_i(t + 1) = w_i(t) + \Delta w_i(t)$$

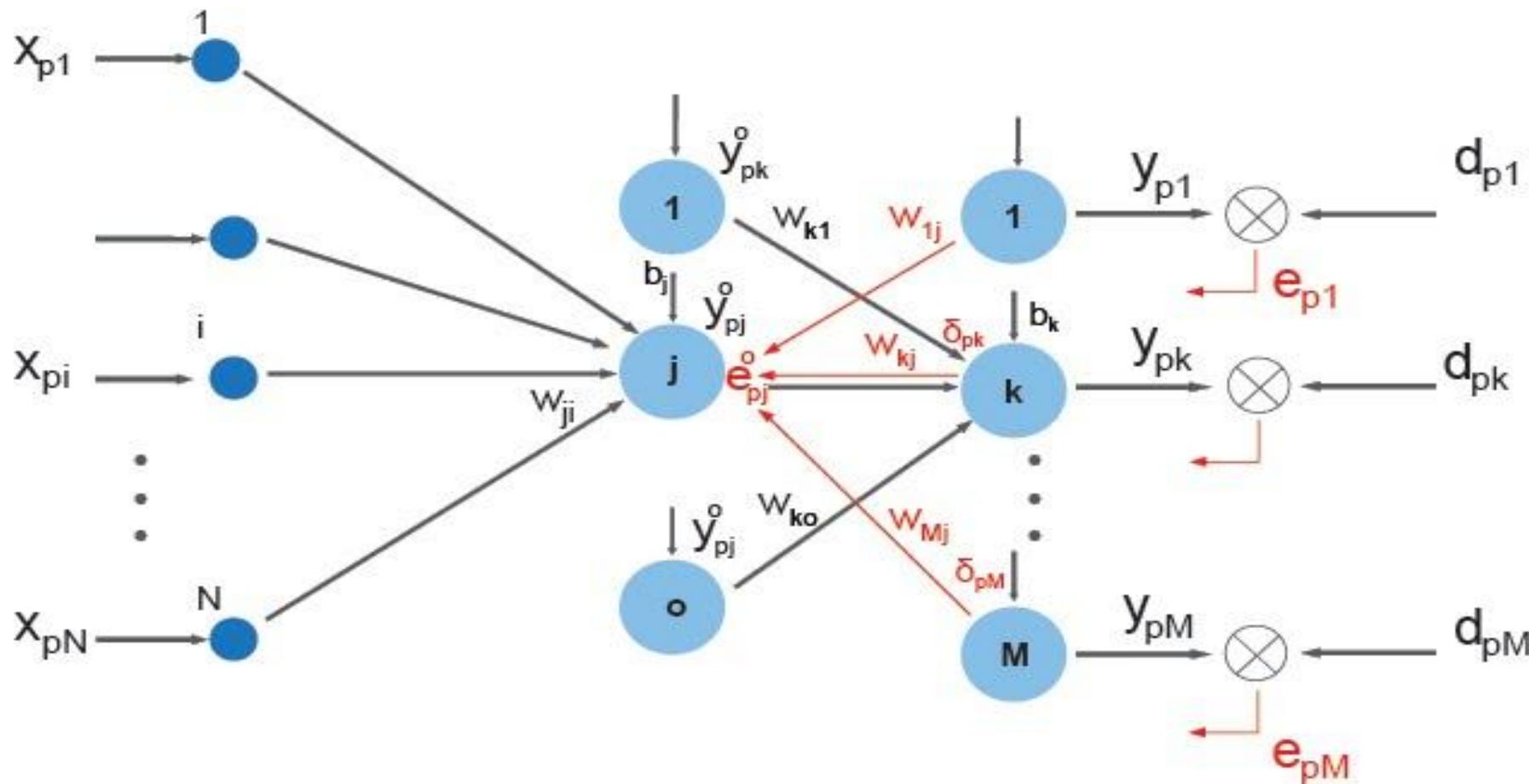
$$w_i(t + 1) = w_i(t) + \alpha(d_i - y_i) \text{Fact}'(\text{neta})x_i$$



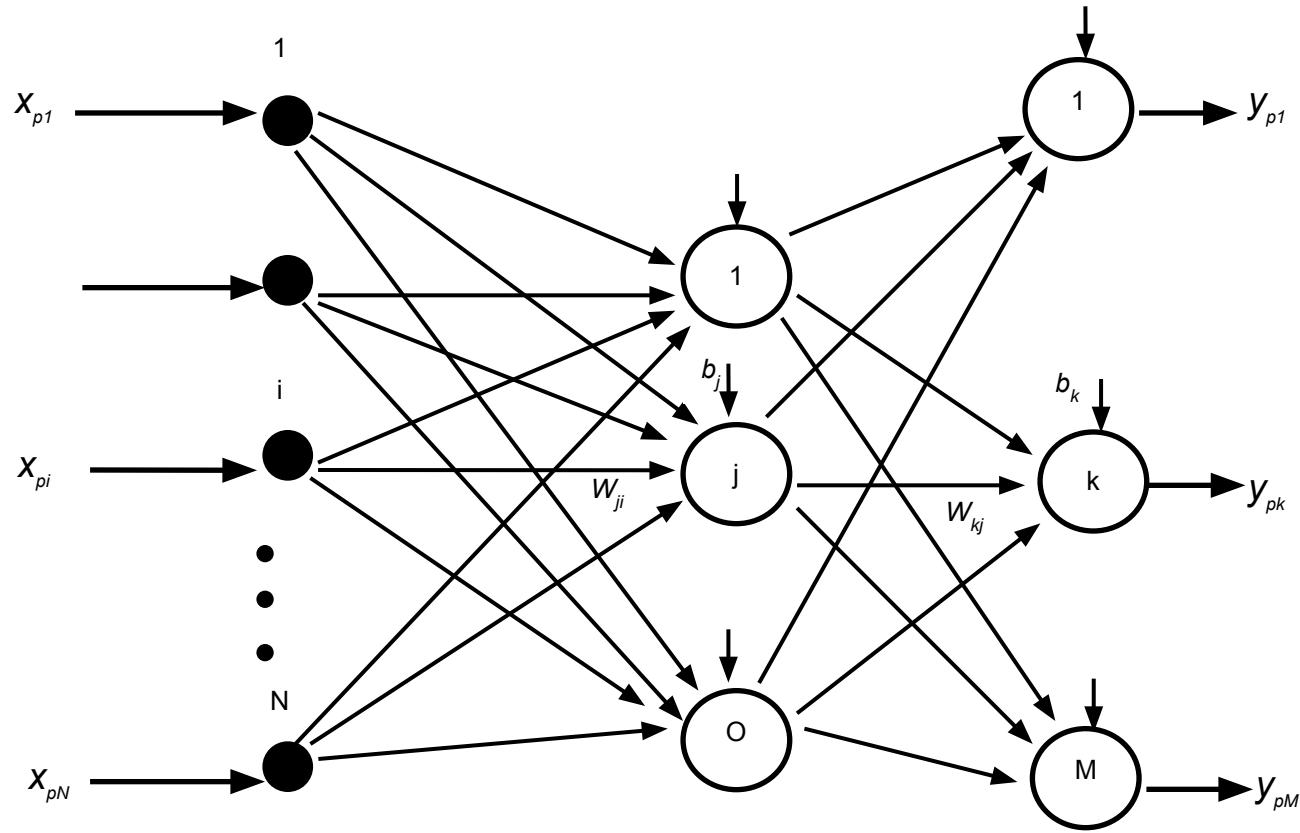
Es evidente en la capa de Salida pero
¿Cómo se calcula en las capas ocultas?

Capas Densas

Backpropagation



Capas Densas



Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

Los Bloques Básicos del Deep Learning

Capas Convolucionales

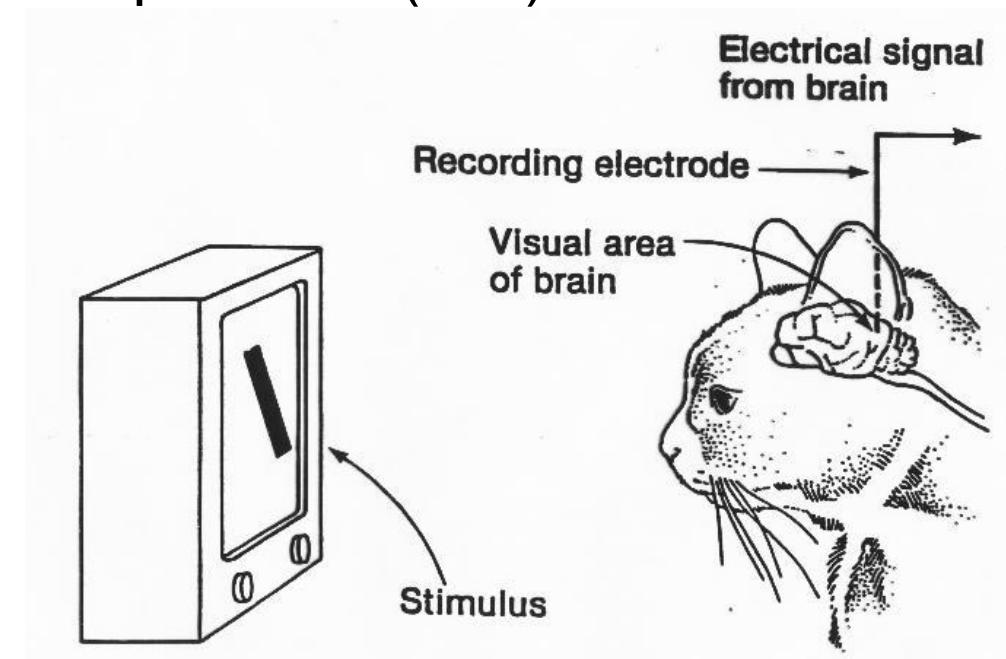
Hubel and Wiesel Cat Experiment (60s)

Patrones sencillos = luz y oscuridad, células simples

Patrones intermedios = Bordes o movimientos en determinadas direcciones, células complejas

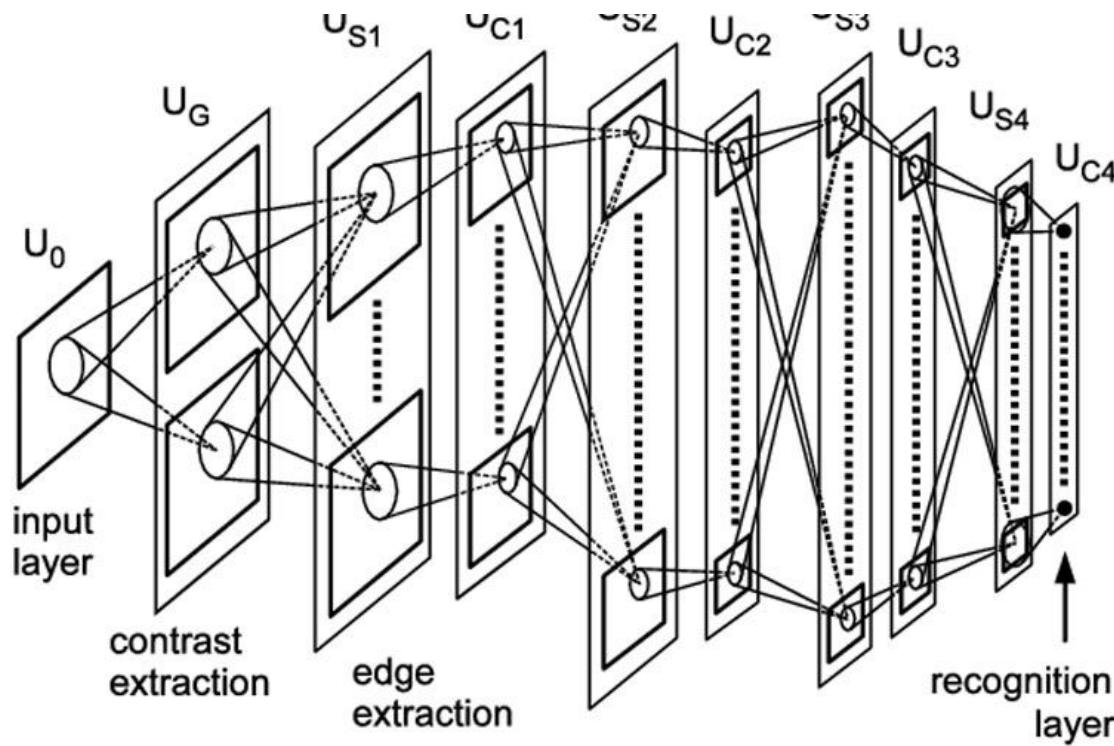
Patrones complejos = Dos bordes formando un ángulo recto , células hipercomplejas

Neuronas especializadas en detectar bordes, movimientos, profundidad estereoscópica o color.



Capas Convolucionales

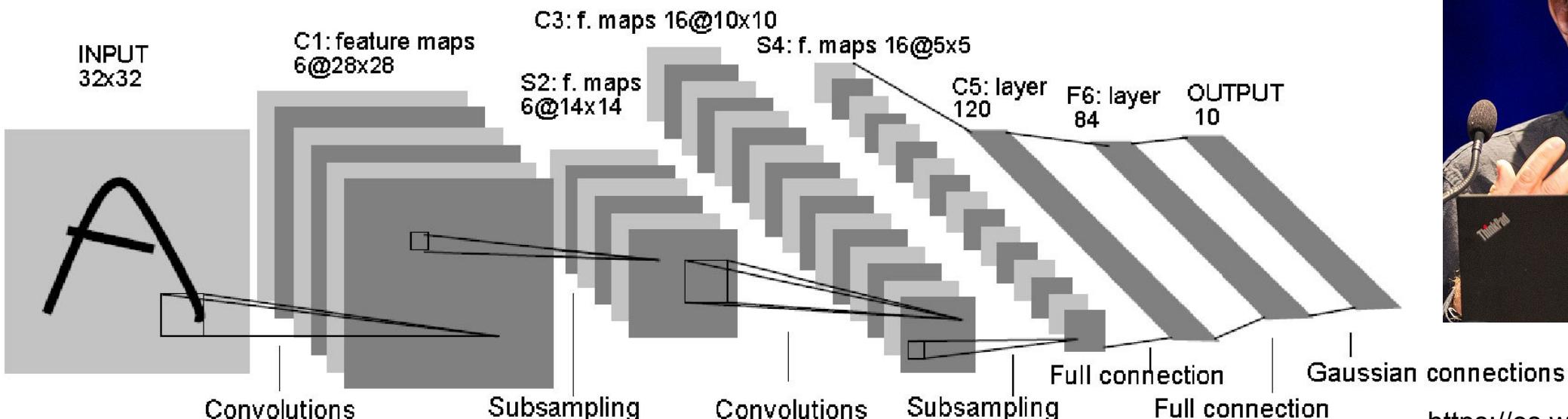
Kunihiko Fukushima and the architecture of the Neocognitron (1979)



Capas Convolucionales

Yann LeCun

LeNet-5 (1989)



https://es.wikipedia.org/wiki/Yann_LeCun

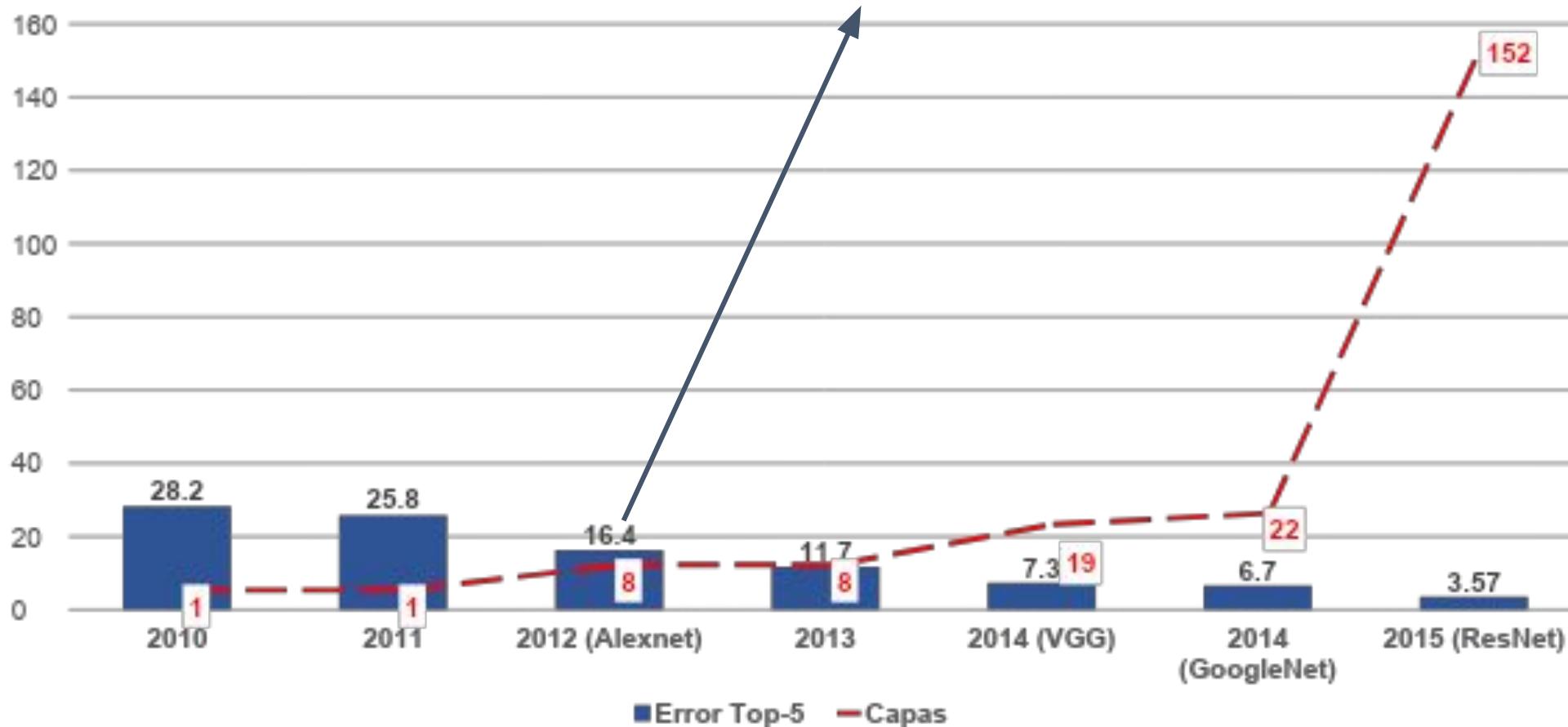
<http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf>

LeNet-1 Demo Video

https://www.youtube.com/watch?v=FwFduRA_L6Q

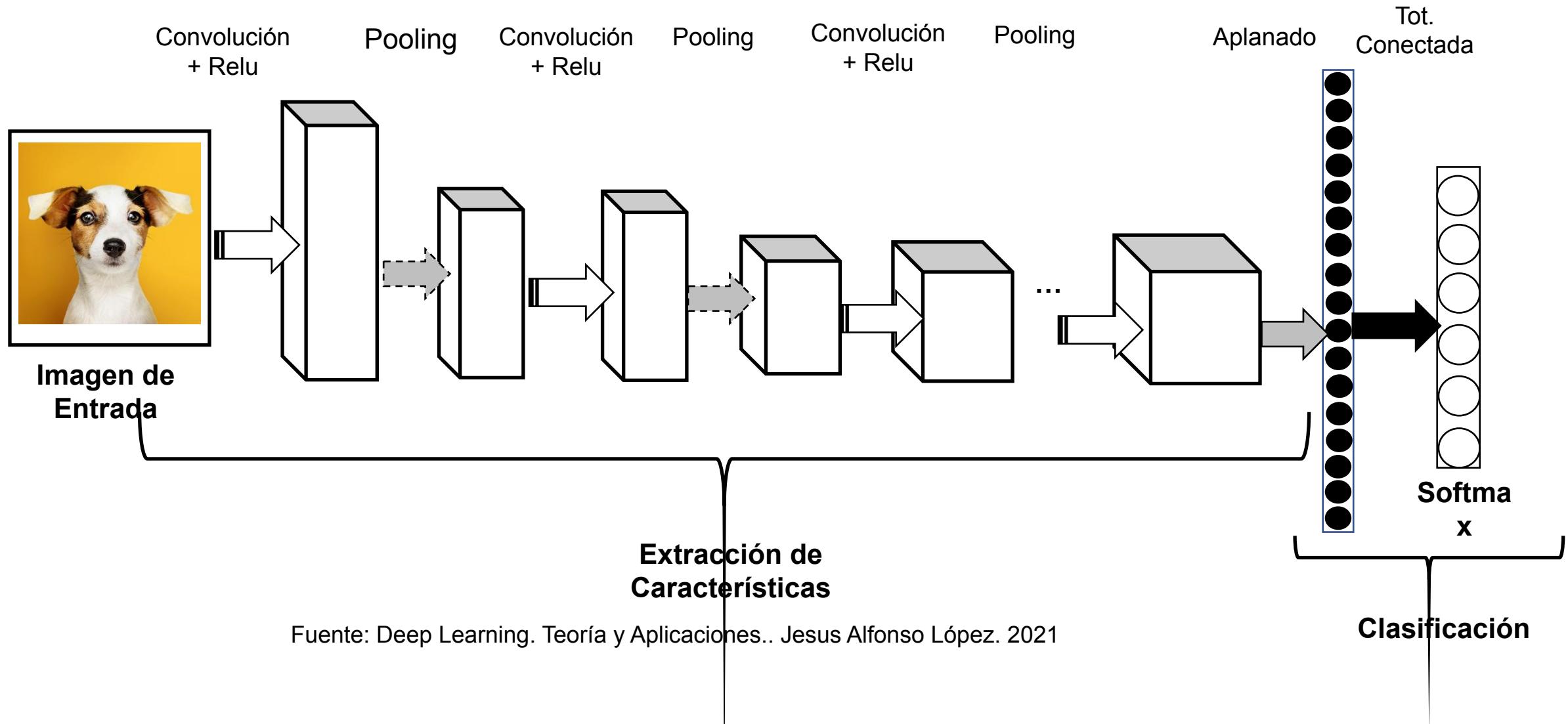
Capas Convolucionales

Imagenet Moment



Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

Capas Convolucionales



Capas Convolucionales

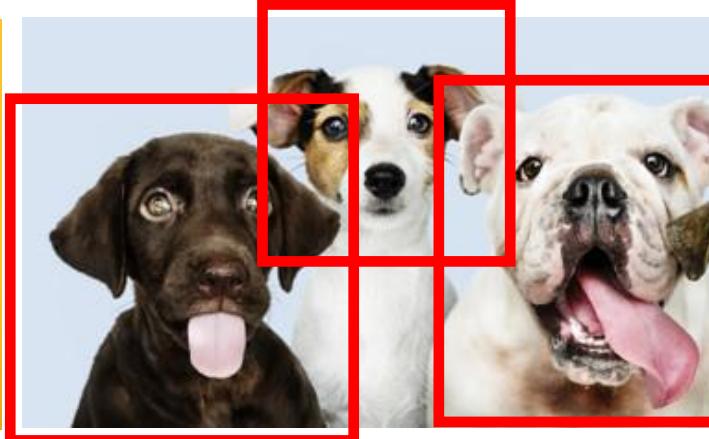
Clasificación



Clasificación y
Localización



Detección



Segmentación Semántica



Un Objeto

Varios Objetos

Capas Convolucionales



<https://securitytoday.com/articles/2019/03/01/the-flaws-and-dangers-of-facial-recognition.aspx>



<https://www.v7labs.com/blog/human-position-estimation-guide>

Los Bloques Básicos del Deep Learning

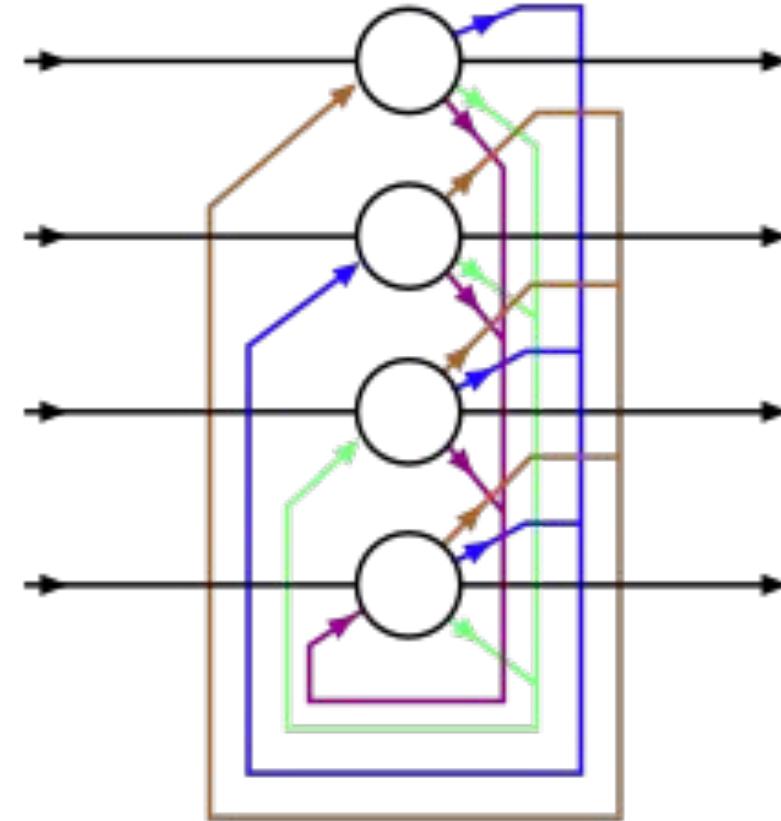
Capas Recurrentes

Red neuronal de Hopfield (1982)

John Joseph Hopfield



https://www.swarthmore.edu/bulletin/archive/wp/october-2009_john-hopfield-54.html



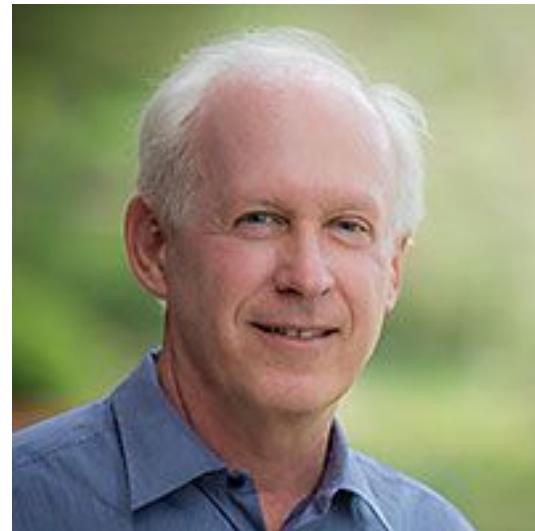
https://en.wikipedia.org/wiki/Hopfield_network

Los Bloques Básicos del Deep Learning

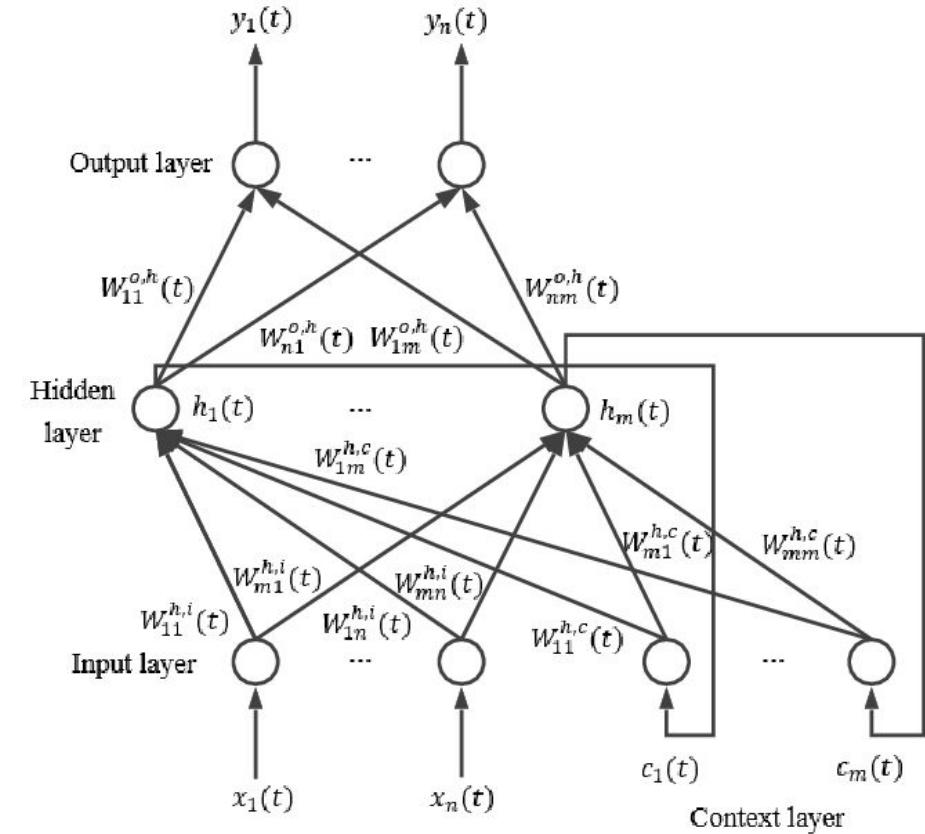
Capas Recurrentes

Red neuronal de Elman (1990)

Jeffrey Locke Elman



<https://crl.ucsd.edu/elman/obituary/>



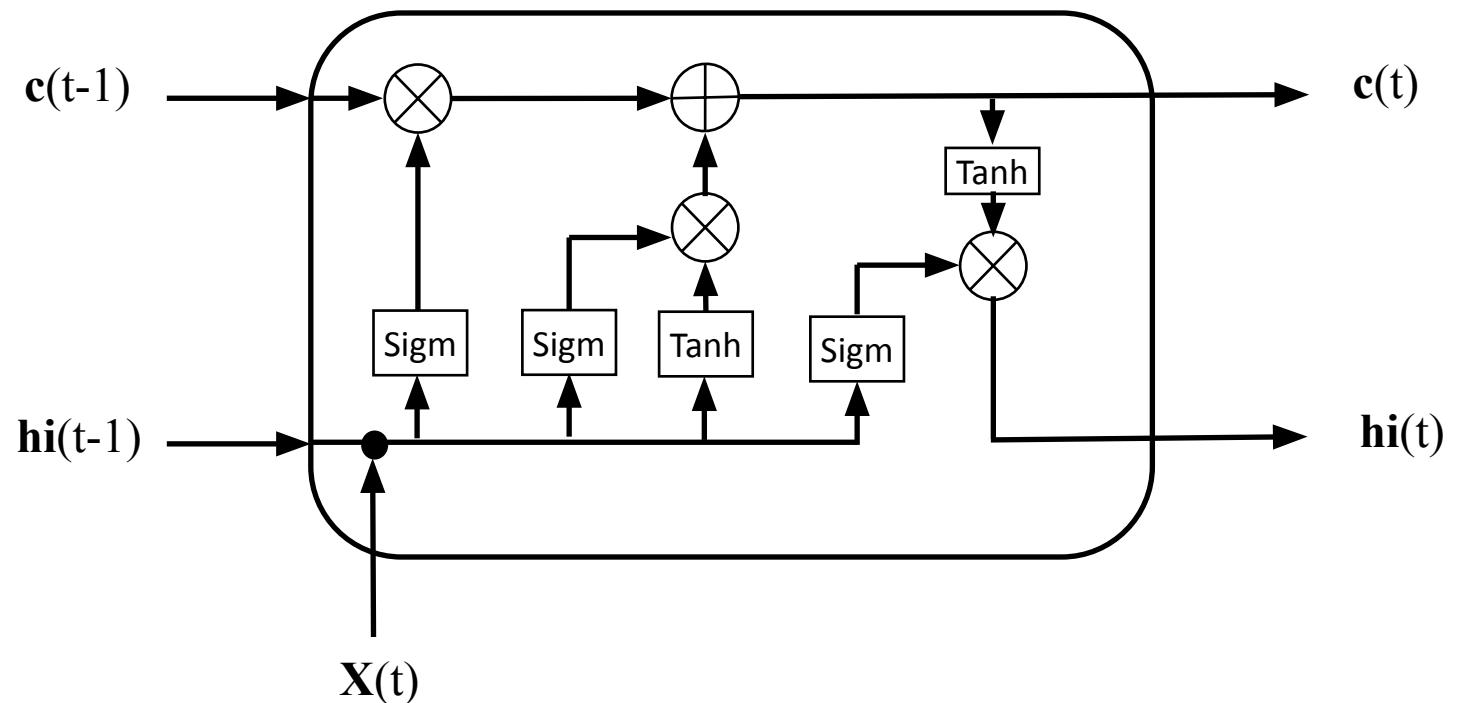
https://www.researchgate.net/figure/The-Structure-of-Elman-Neural-Network_fig1_322843098

Capas Recurrentes

Red neuronal de LSTM (1997)

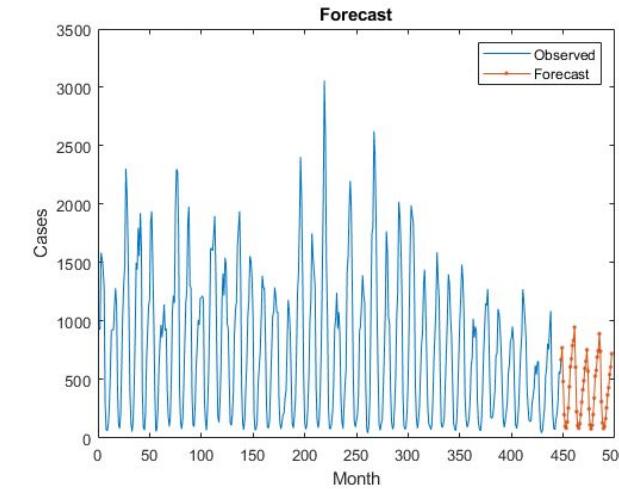


Jürgen Schmidhuber

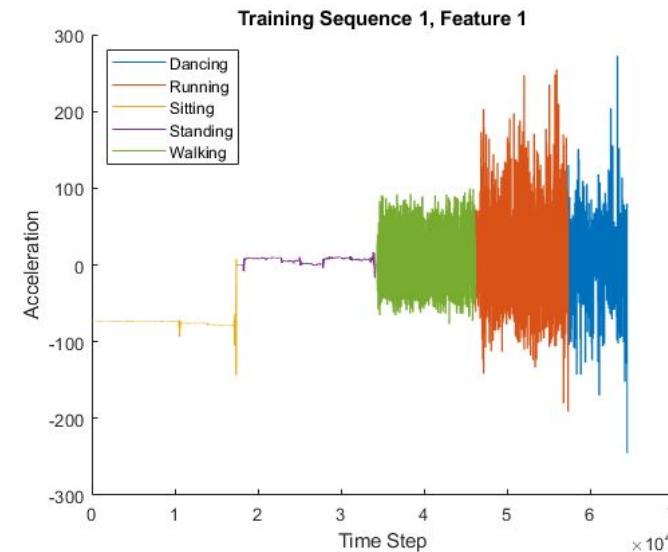


Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

Capas Recurrentes



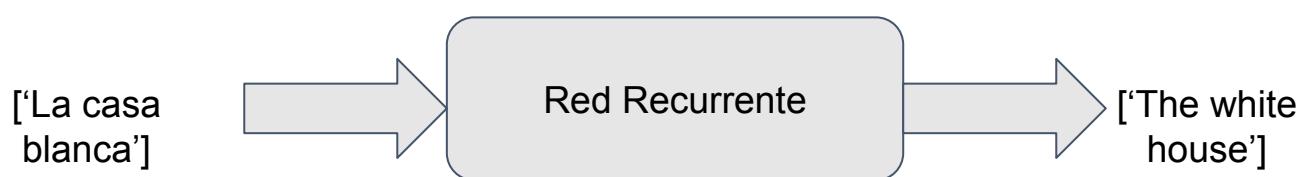
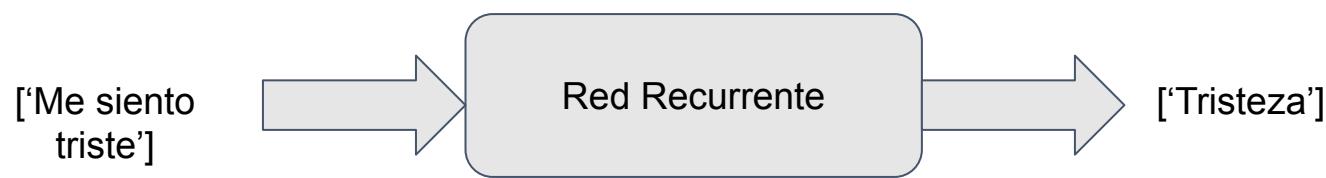
<https://la.mathworks.com/help/deeplearning/examples/time-series-forecasting-using-deep-learning.html>



<https://la.mathworks.com/help/deeplearning/examples/sequence-to-sequence-classification-using-deep-learning.html>

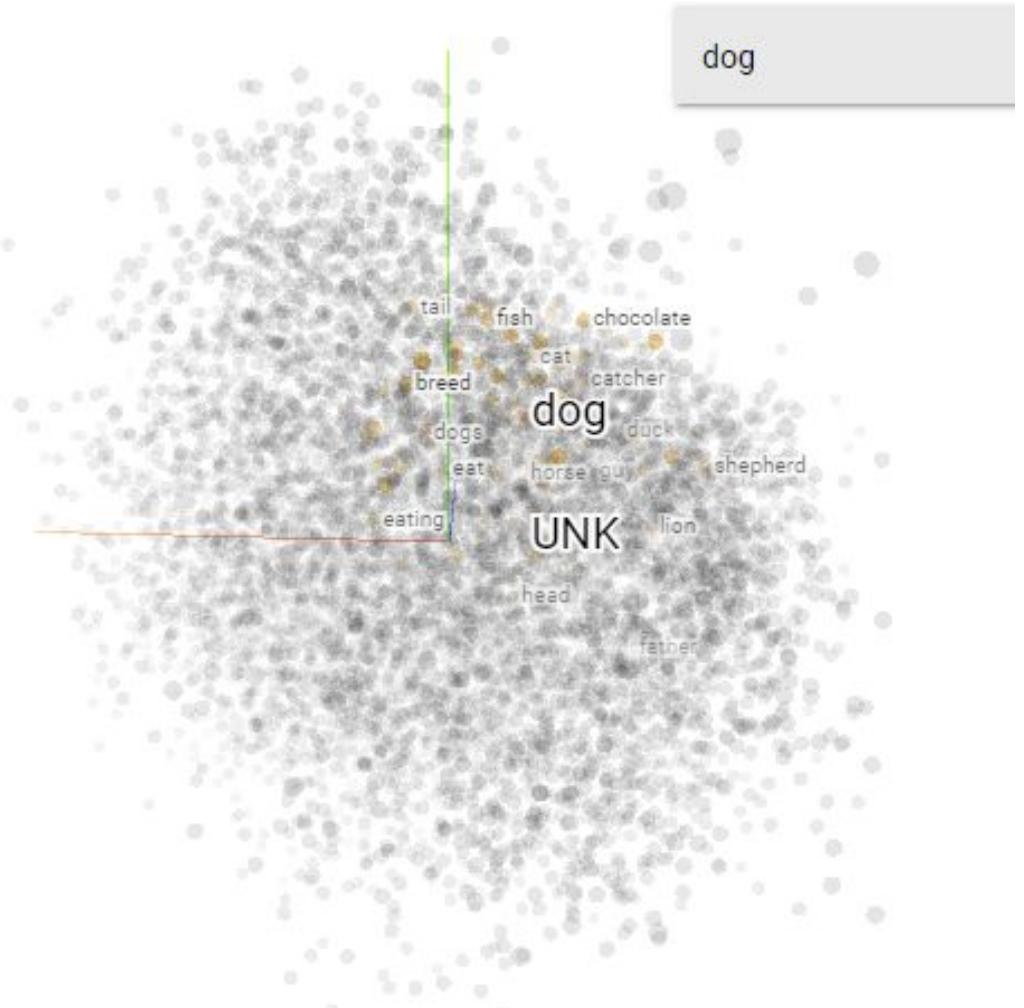
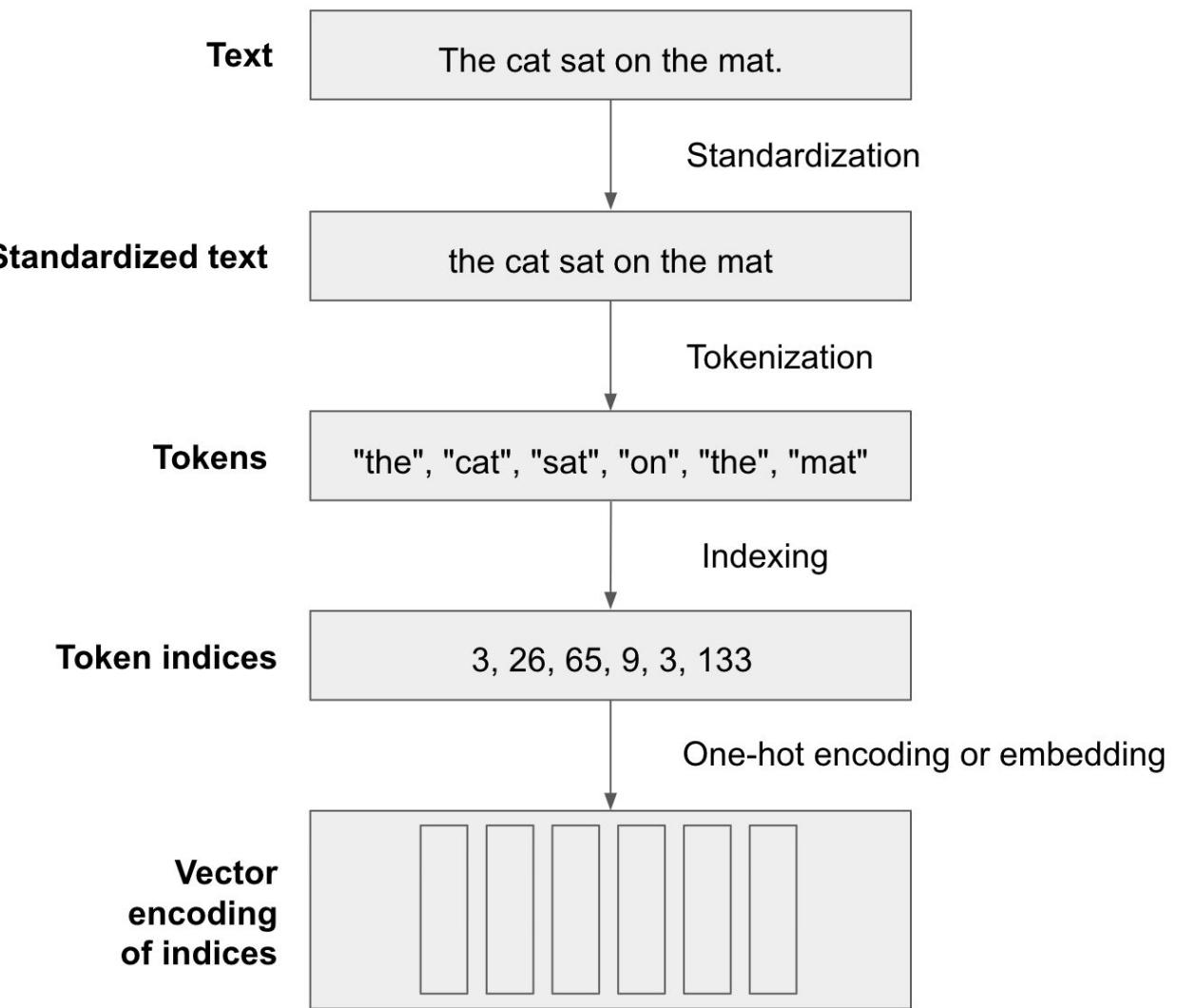
Capas Recurrentes

PLN (Procesamiento de Lenguaje Natural)



Modelos de Lenguaje

Embedding

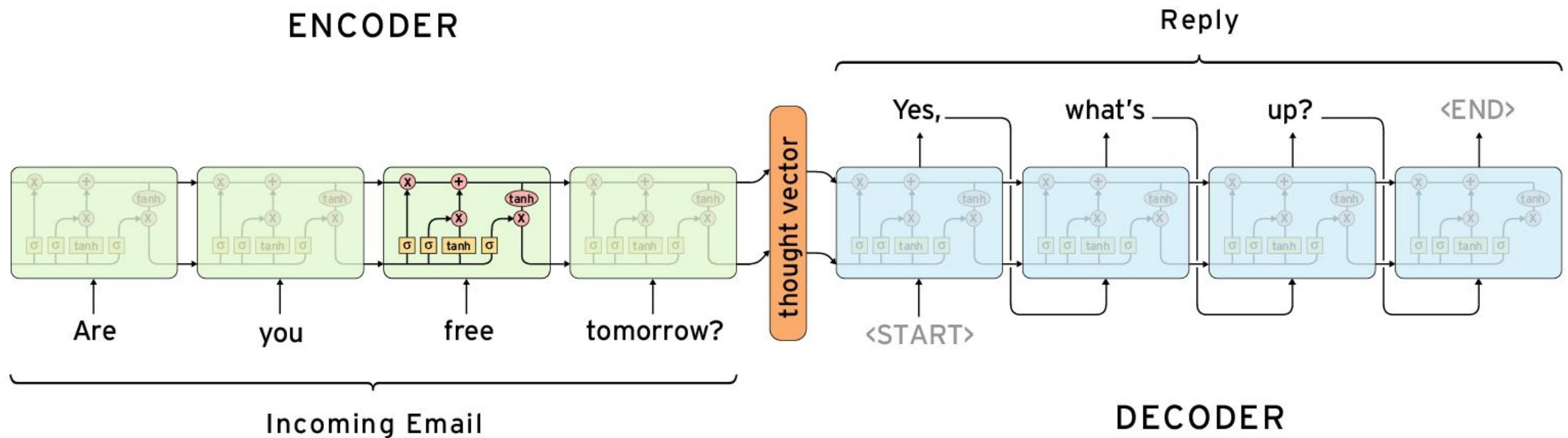
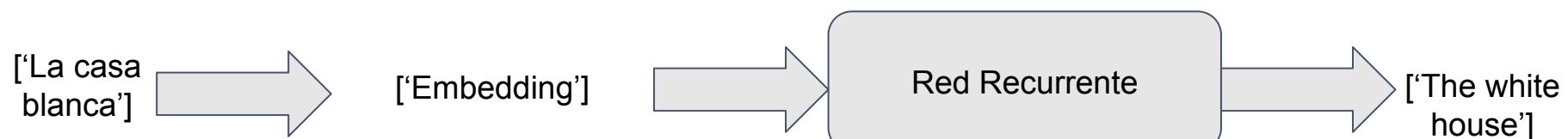


<https://projector.tensorflow.org/>

Modelos de Lenguaje

Secuencia a secuencia

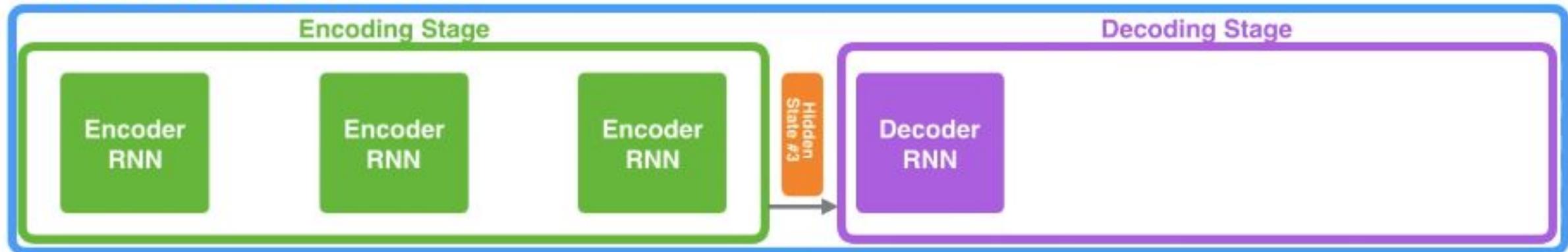
Aplicación tipo seq2seq. Por ejemplo traducción, sistemas de recomendación, chatbots
Tienen un Codificador y un Decodificador



Modelos de Lenguaje

Con entrada al decodificador se usan el último estado del codificador

Neural Machine Translation SEQUENCE TO SEQUENCE MODEL



<https://jamalmar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/>

Modelos de Lenguaje

Uso de la Atención

Teniendo en cuenta la atención se usan todos los estados ocultos del codificador

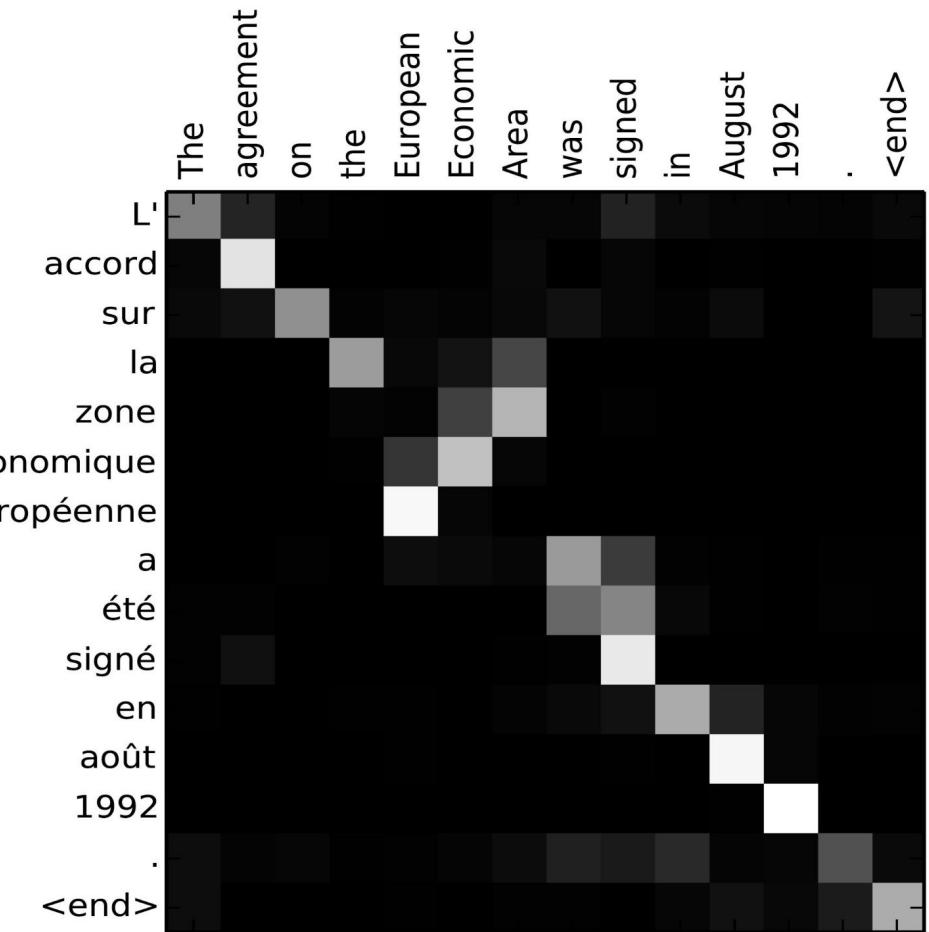
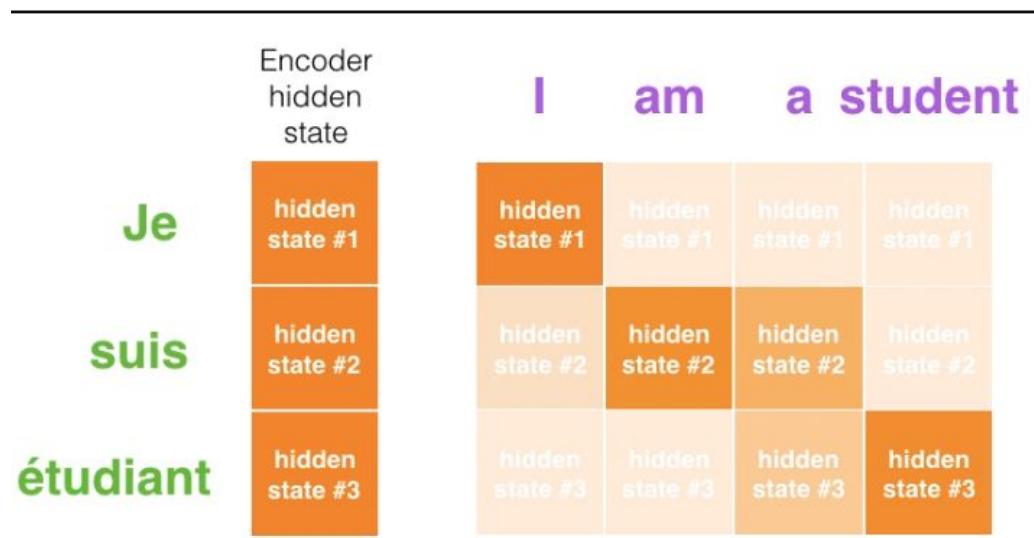
Neural Machine Translation SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



<https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/>

Atención

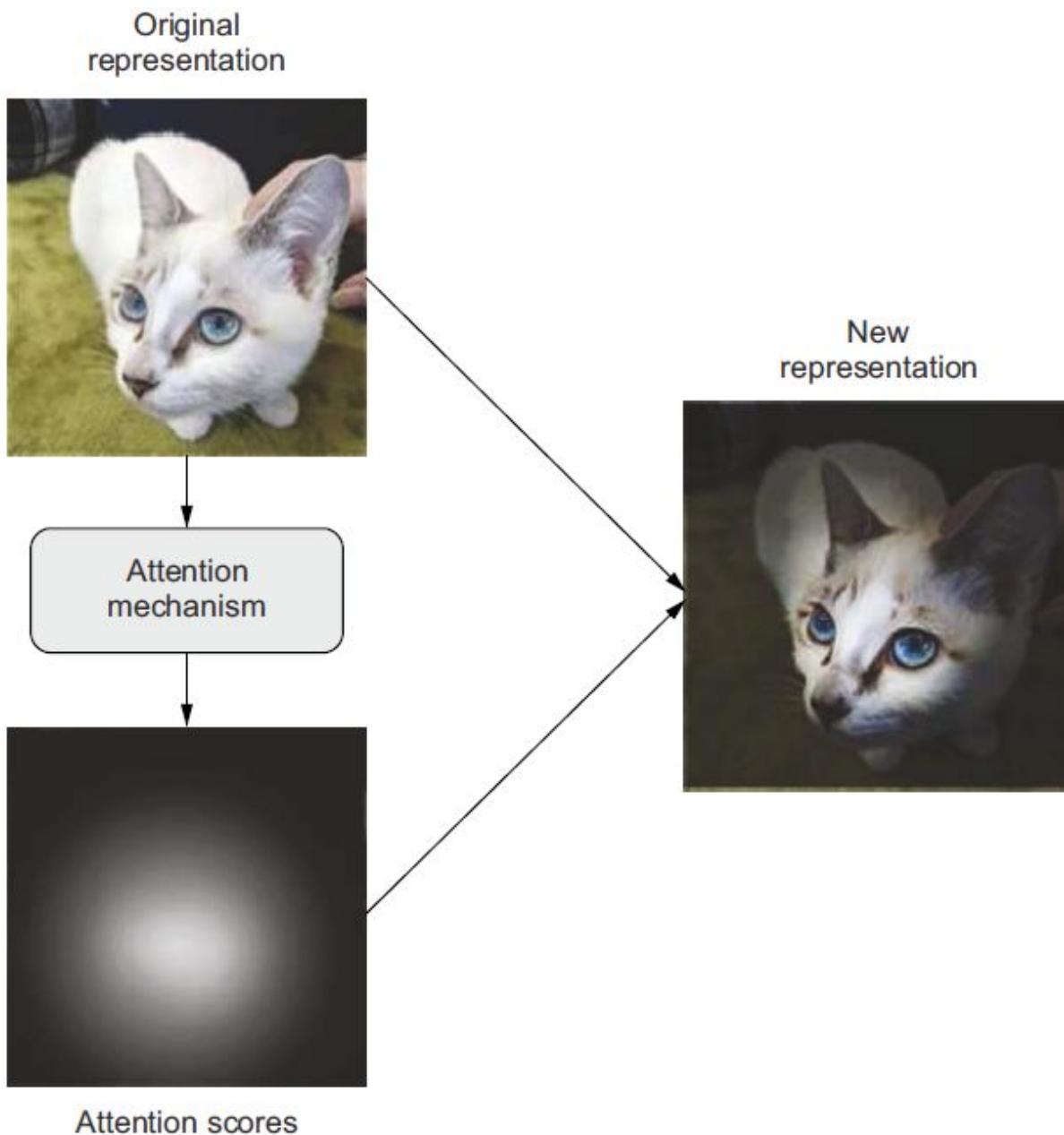
Esta es otra forma de ver a qué parte del texto de entrada prestamos atención en cada paso de decodificación



<https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/>

<https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>

Atención



Atención como mecanismo de representar el contexto de una información de entrada

Deep Learning with Python,
François Chollet
Manning; 2nd edición (21 Diciembre 2021)

Los Bloques Básicos del Deep Learning

Capas de Auto 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

Lukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Fórmula de la Self-Attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{keys}}}\right)\mathbf{V}$$

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 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

<https://arxiv.org/pdf/1706.03762.pdf>

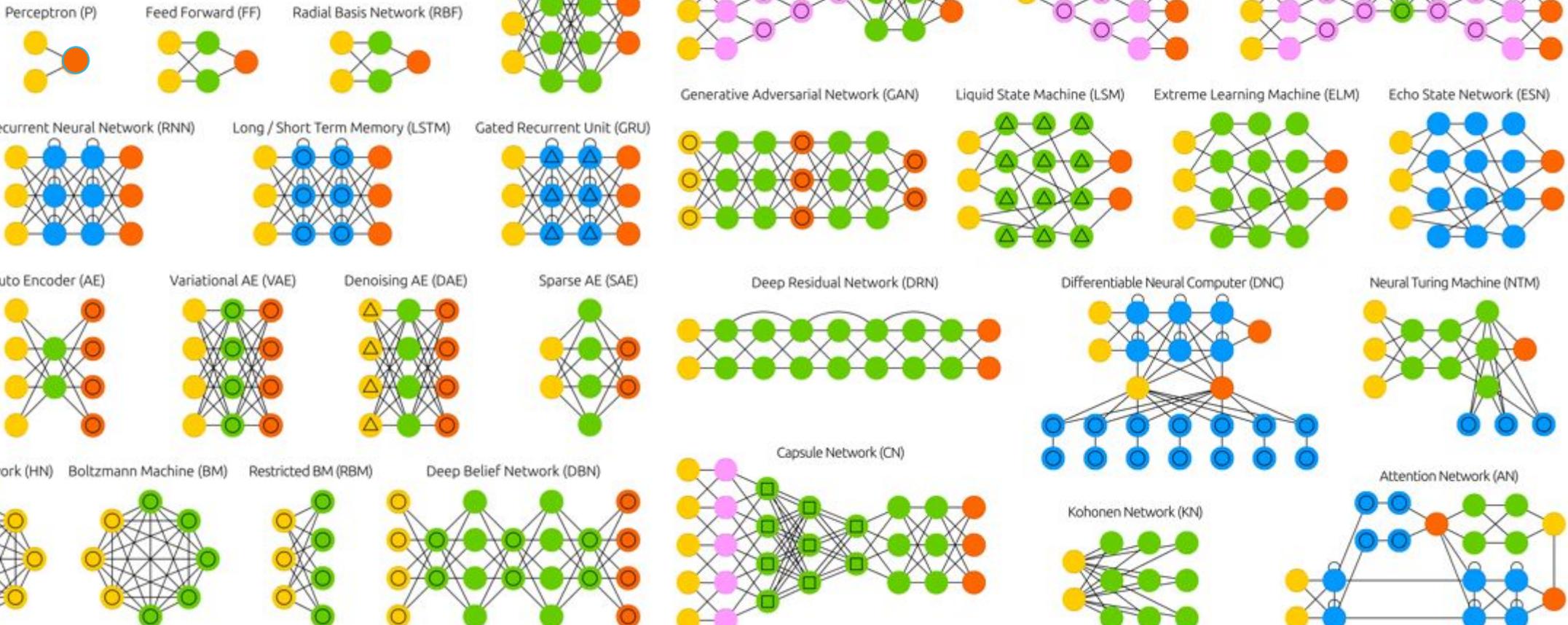
Deep Learning ZOO

- Input Cell
- Backfed Input Cell
- △ Noisy Input Cell
- Hidden Cell
- Probabilistic Hidden Cell
- △ Spiking Hidden Cell
- Capsule Cell
- Output Cell
- Match Input Output Cell

- Recurrent Cell
- Memory Cell
- △ Gated Memory Cell
- Kernel
- Convolution or Pool

A mostly complete chart of Neural Networks

©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org

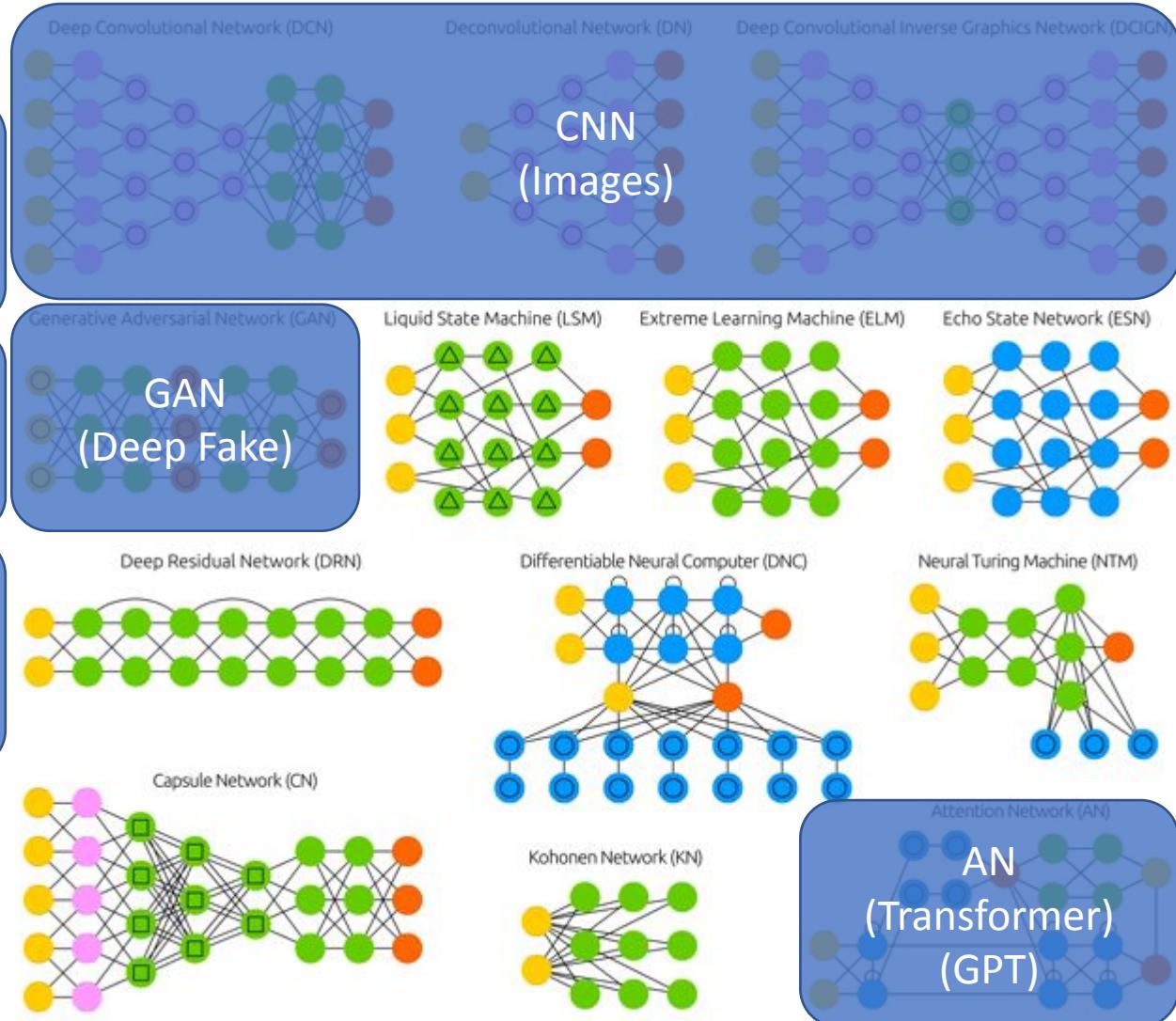
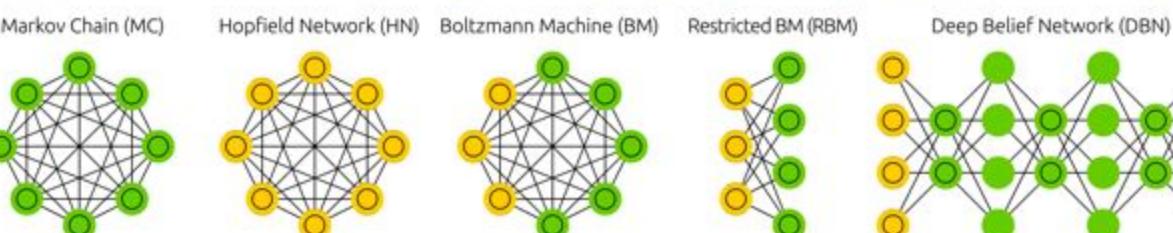
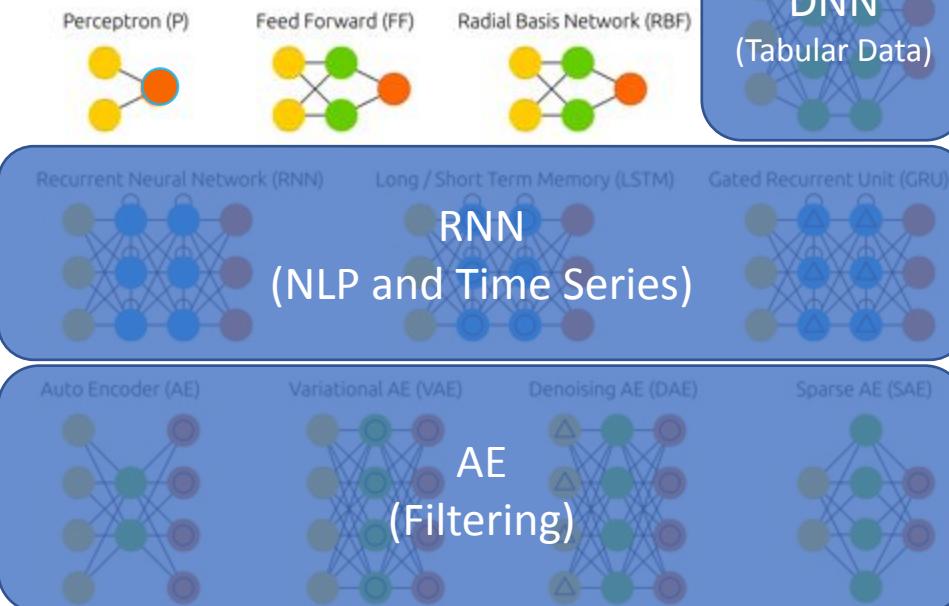


Deep Learning ZOO



A mostly complete chart of Neural Networks

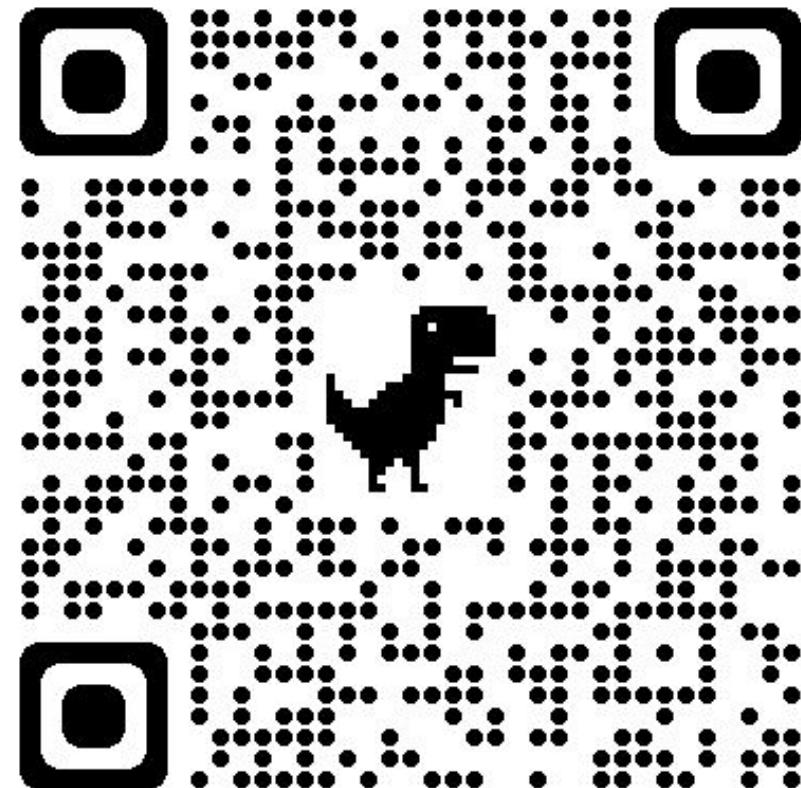
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org



Modelos de IA Basados en Transformers

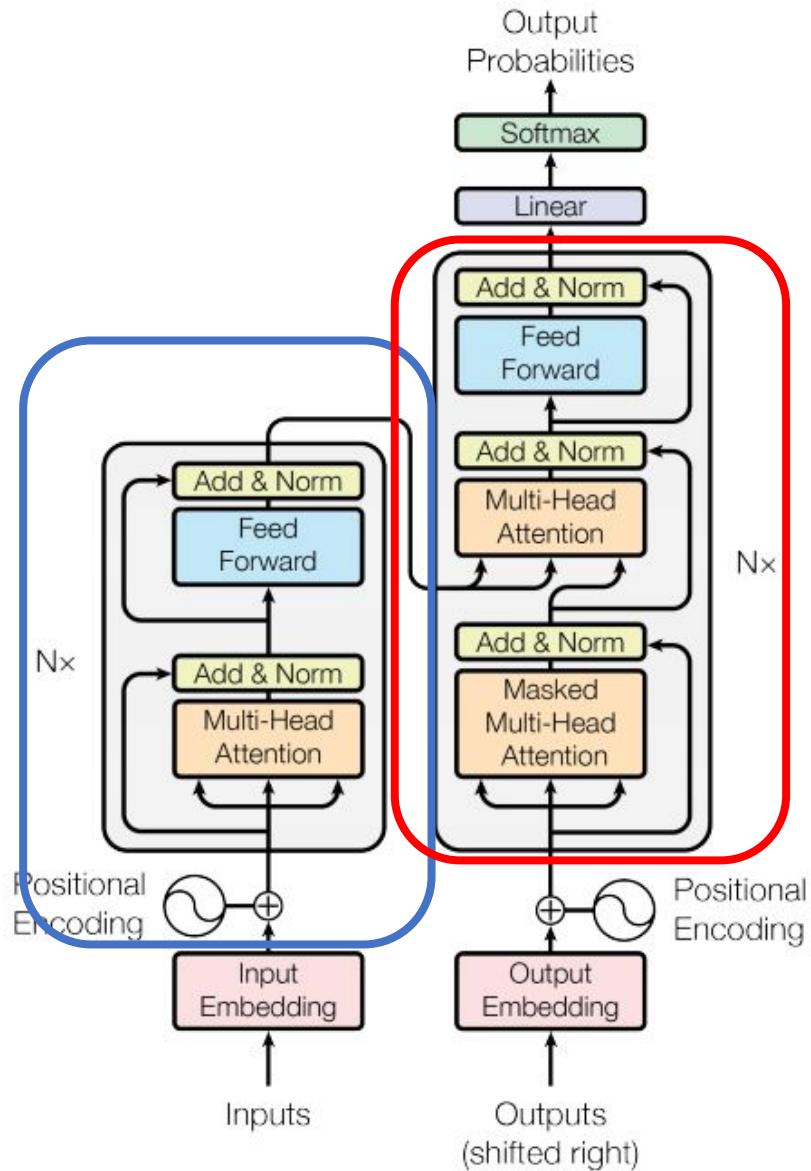
**¿Qué se te viene a la
mente cuando
escuchas la palabra
transformers?**

[https://app.wooclap.com/R
OJGUC?from=status-bar](https://app.wooclap.com/R
OJGUC?from=status-bar)



Transformers

Transformer



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

Lukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

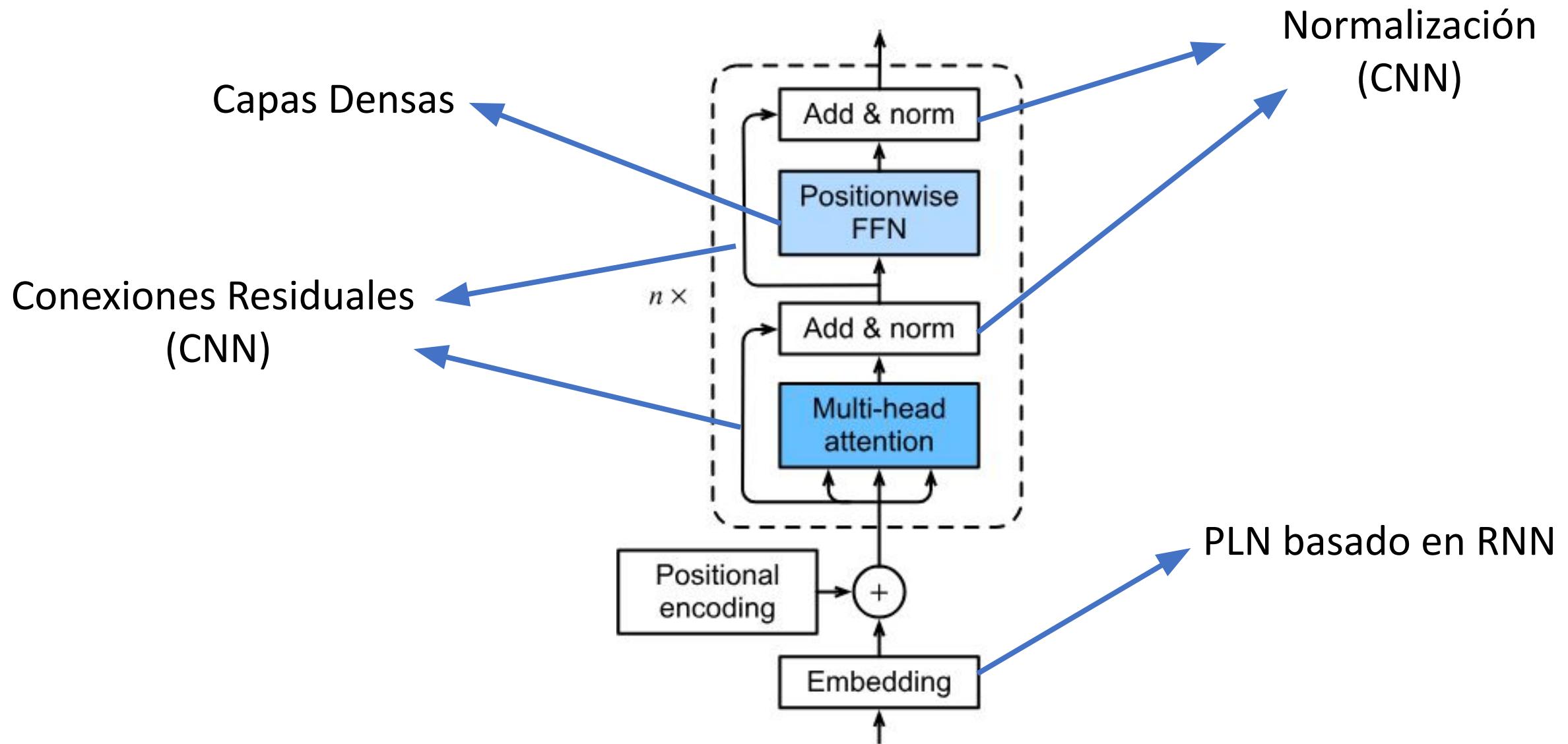
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

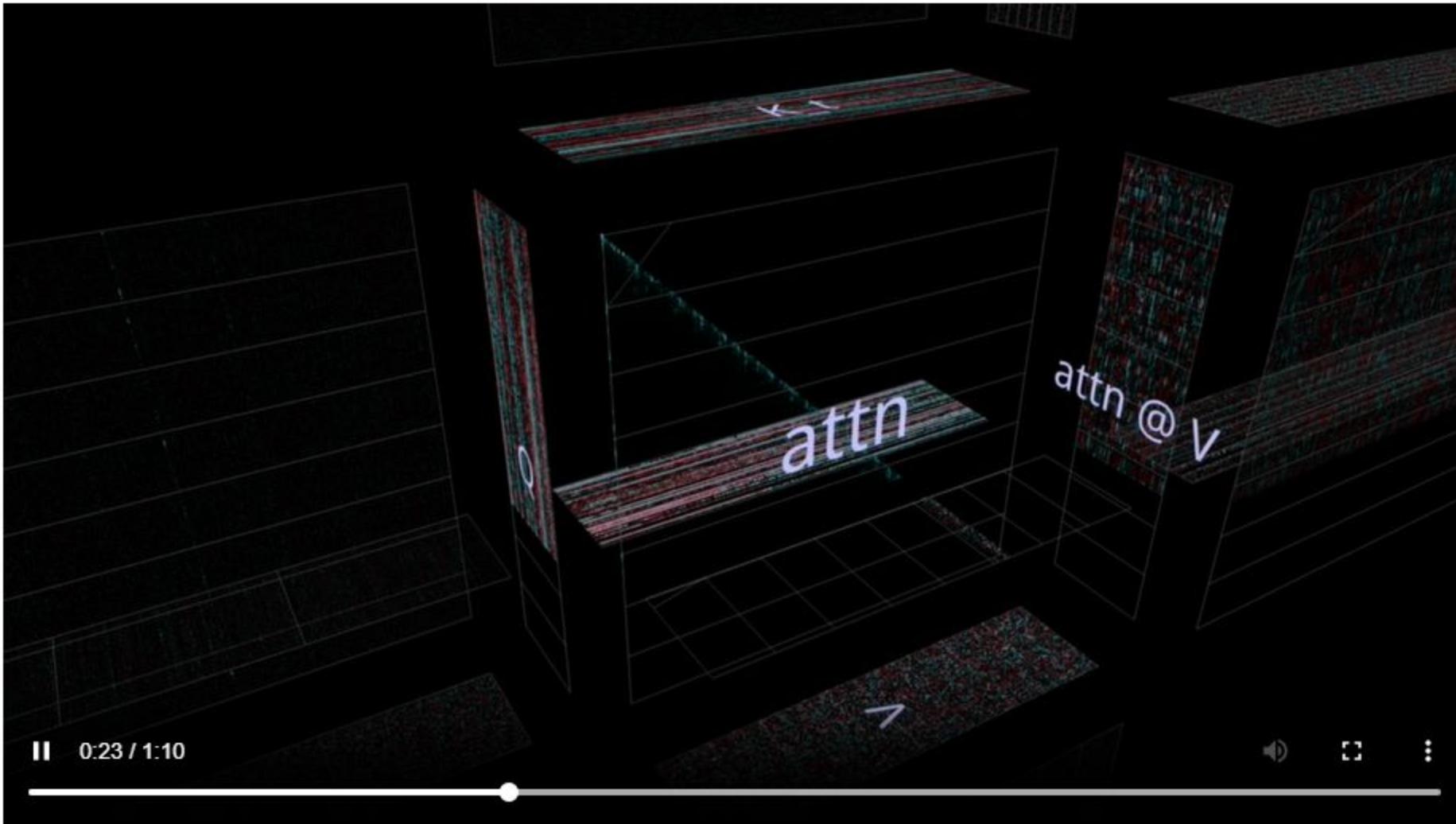
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

<https://arxiv.org/pdf/1706.03762.pdf>

Transformers

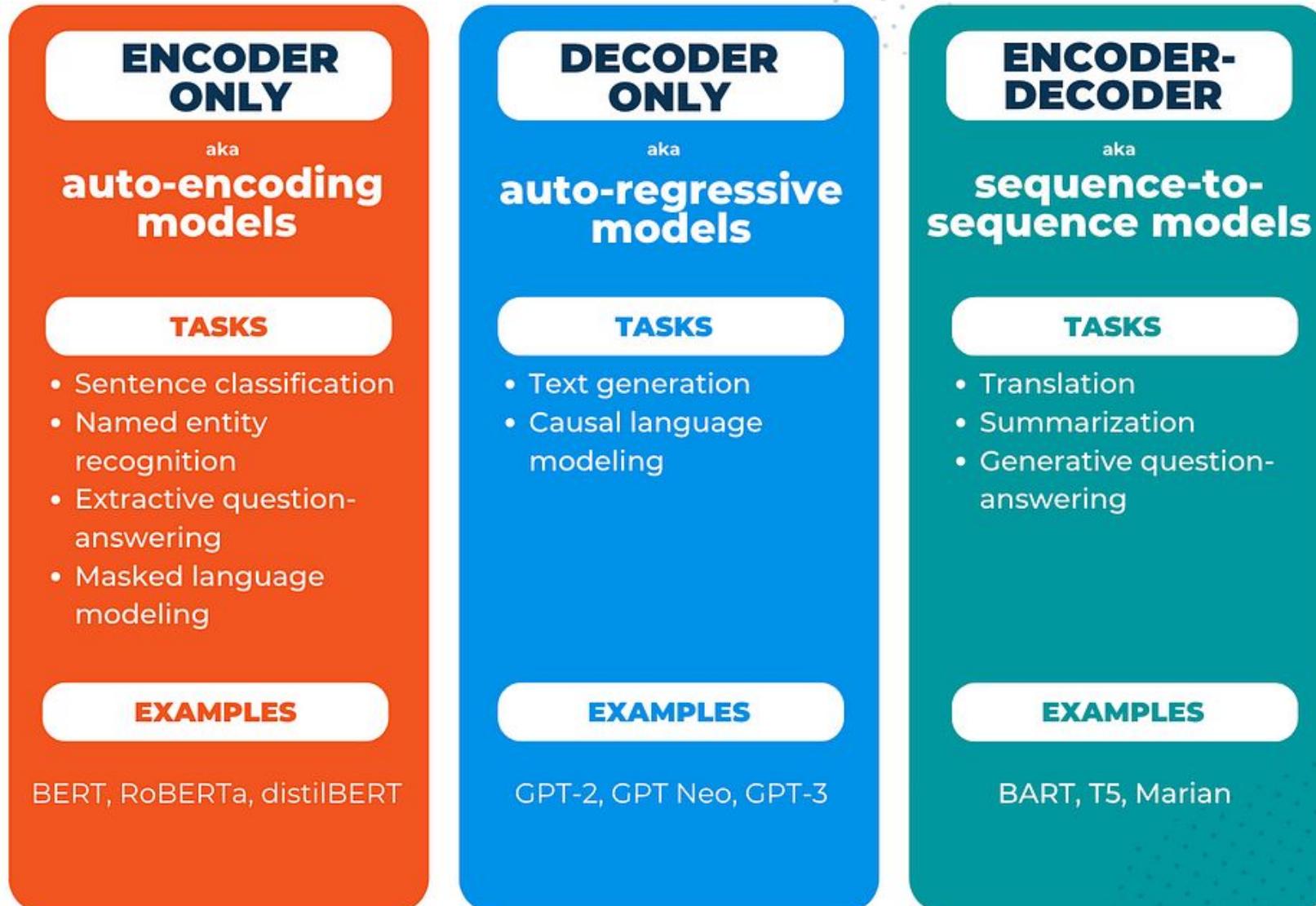


Transformers



<https://pytorch.org/blog/inside-the-matrix/>

Transformers y Modelos de Lenguaje



Transformers y Modelos de Lenguaje

BERT (Bidirectional Encoder Representations from Transformers)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

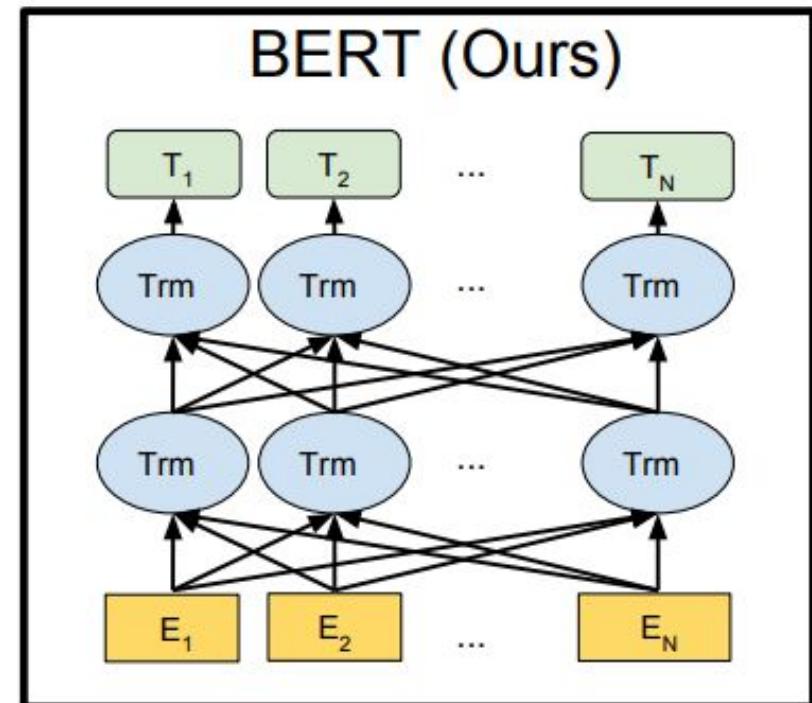
Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **Bidirectional Encoder Representations from Transformers**. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-

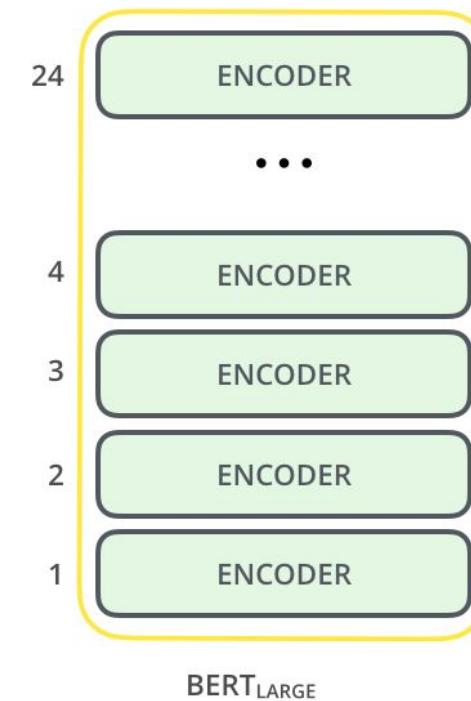
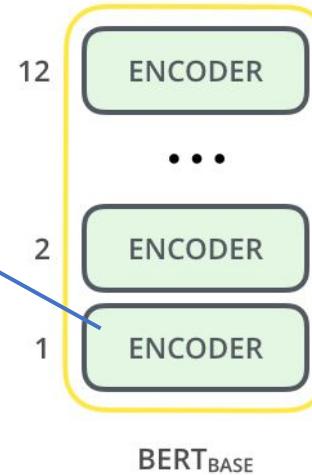
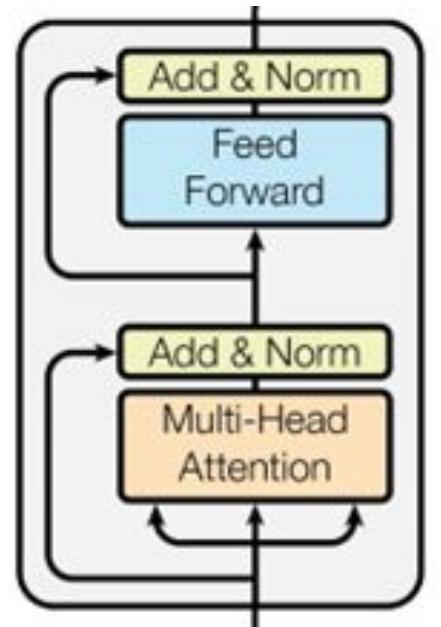


<https://arxiv.org/pdf/1810.04805.pdf>

Transformers y Modelos de Lenguaje

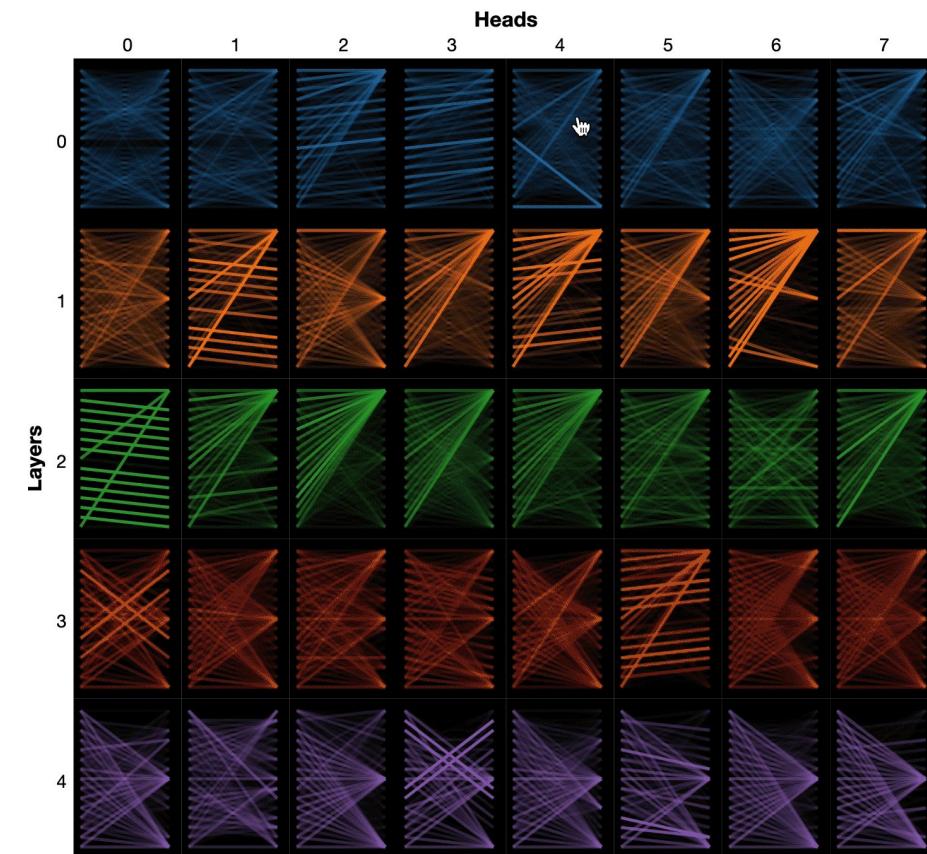
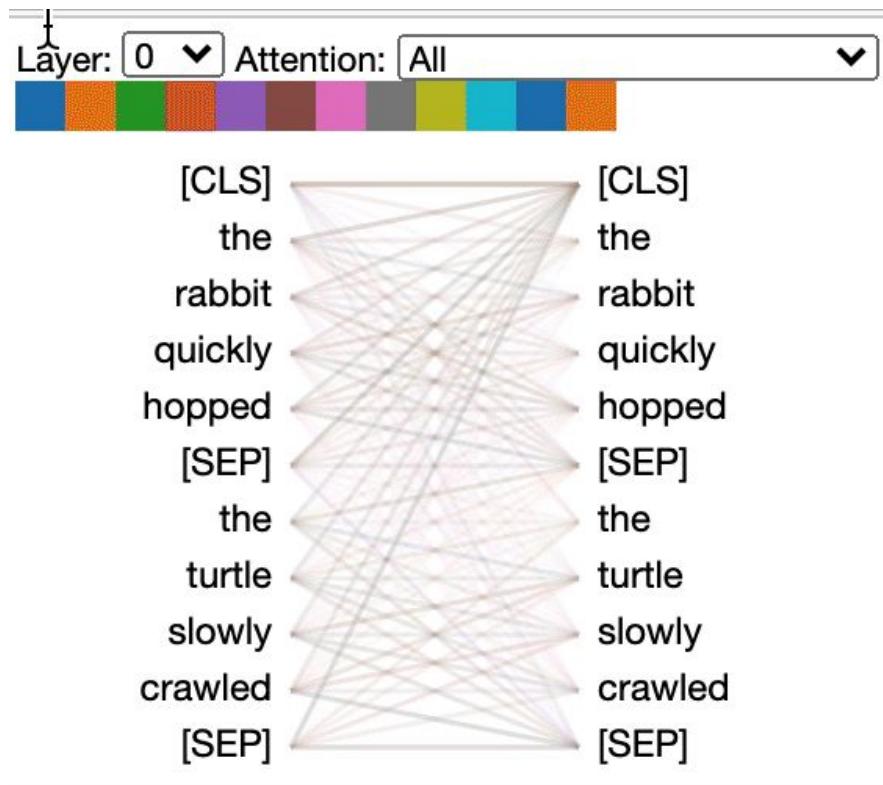
BERT (Bidirectional Encoder Representations from Transformers)

BERT-Base: 12-layers, 768-hidden, 12-attention-heads, 110M parameters
BERT-Large: 24-layers, 1024-hidden, 16-attention-heads, 340M parameters



<http://jalammar.github.io/illustrated-bert/>

Transformers y Modelos de Lenguaje



<https://github.com/jessevieg/bertviz>

<https://colab.research.google.com/drive/1hXIQ77A4TYS4y3UthWF-Ci7V7vVUoxmQ>

Transformers y Modelos de Lenguaje

GPT (Generative Pretrained Transformers)

12-layers, 768-hidden, 12-attention-heads, 117M parameters. Tamaño de secuencia 512

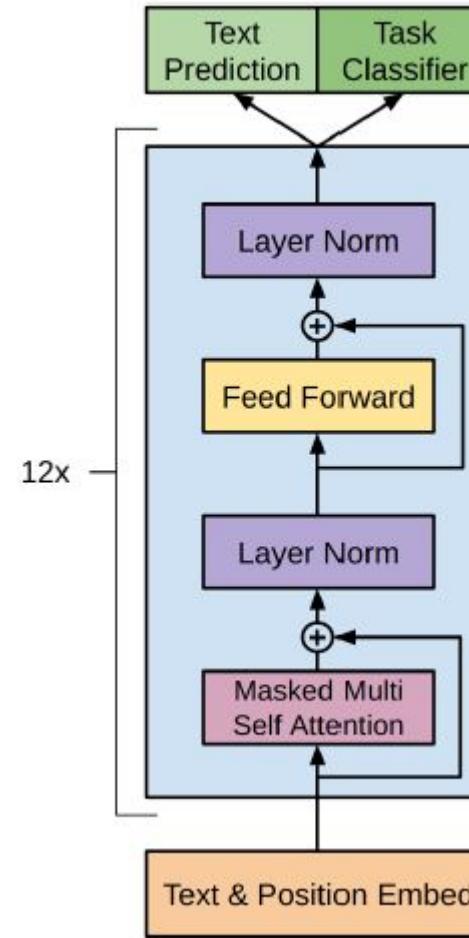
Improving Language Understanding by Generative Pre-Training

Alec Radford Karthik Narasimhan Tim Salimans Ilya Sutskever
OpenAI OpenAI OpenAI OpenAI
alec@openai.com karthikn@openai.com tim@openai.com ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

<https://openai.com/language-understanding-paper.pdf>



<https://openai.com/blog/language-unsupervised/>

Transformers y Modelos de Lenguaje

GPT-2

Language Models are Unsupervised Multitask Learners

Alec Radford *¹ Jeffrey Wu *¹ Rewon Child¹ David Luan¹ Dario Amodei **¹ Ilya Sutskever **¹

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and in-

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

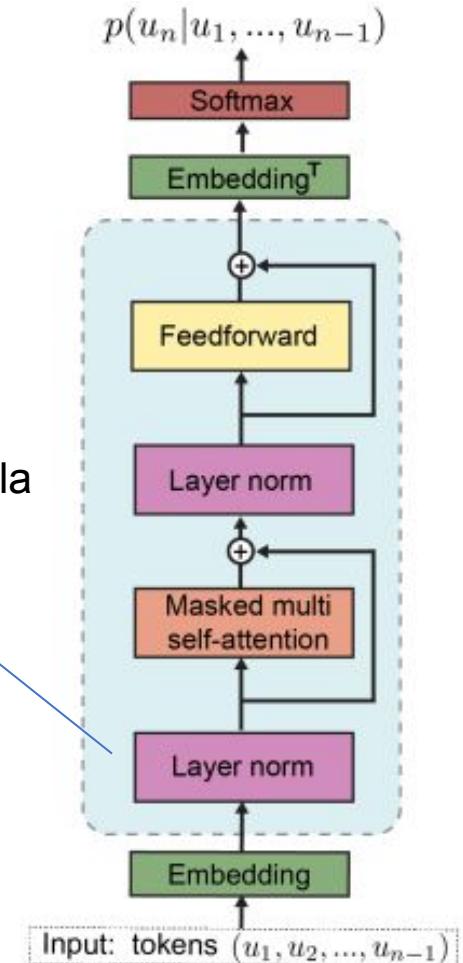
The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training

https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

https://www.researchgate.net/publication/335737829_Tracking_Naturalistic_Linguistic_Predictions_with_Deep_Neural_Language_Models

Se cambia de posición la capa de normalización



Transformers y Modelos de Lenguaje

GPT-2

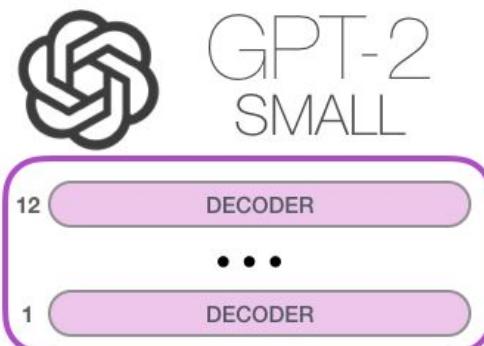
GPT-2: 12-layers, 768-hidden, 12-attention-heads, 117M parameters.

GPT-2: 24-layers, 1024-hidden, 12-attention-heads, 345M parameters.

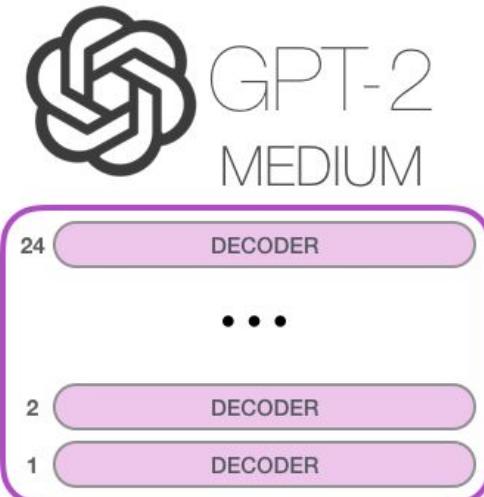
GPT-2: 36-layers, 1280-hidden, 12-attention-heads, 762M parameters.

GPT-2: 48-layers, 1600-hidden, 12-attention-heads, 1542M parameters.

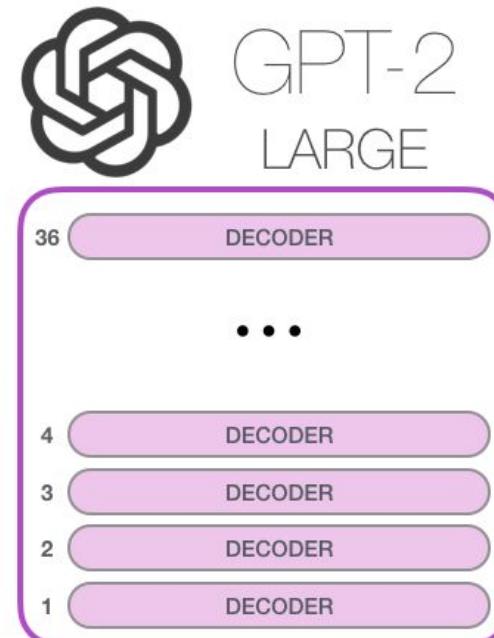
Tamaño de secuencia para todos los modelos 1024



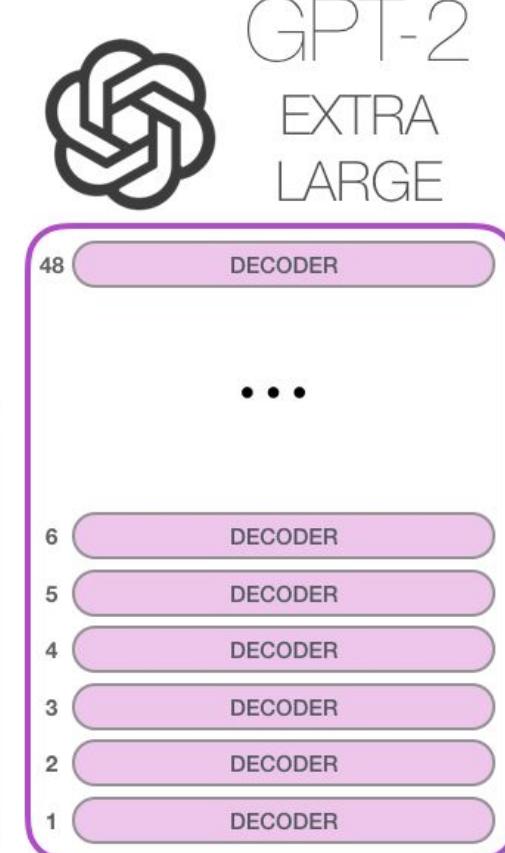
Model Dimensionality: 768



Model Dimensionality: 1024



Model Dimensionality: 1280



Model Dimensionality: 1600

Transformers y Modelos de Lenguaje

GPT-3

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3’s few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

Se explota la capacidad para Zero
Shot Learning
No se hace fine tuning
Es un modelo de lenguaje masivo

<https://arxiv.org/pdf/2005.14165.pdf>

Transformers y Modelos de Lenguaje

GPT-3

GPT-3: 96-layers, 12288-hidden, 96-attention-heads, 175×10^9 parameters.
Tamaño de secuencia 2048

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}

Transformers y Modelos de Lenguaje

Evaluating Large Language Models Trained on Code

Mark Chen ^{*1} Jerry Tworek ^{*1} Heewoo Jun ^{*1} Qiming Yuan ^{*1} Henrique Ponde de Oliveira Pinto ^{*1}
Jared Kaplan ^{*2} Harri Edwards ¹ Yuri Burda ¹ Nicholas Joseph ² Greg Brockman ¹ Alex Ray ¹ Raul Puri ¹
Gretchen Krueger ¹ Michael Petrov ¹ Heidiy Khlaaf ³ Girish Sastry ¹ Pamela Mishkin ¹ Brooke Chan ¹
Scott Gray ¹ Nick Ryder ¹ Mikhail Pavlov ¹ Alethea Power ¹ Lukasz Kaiser ¹ Mohammad Bavarian ¹
Clemens Winter ¹ Philippe Tillet ¹ Felipe Petroski Such ¹ Dave Cummings ¹ Matthias Plappert ¹
Fotios Chantzis ¹ Elizabeth Barnes ¹ Ariel Herbert-Voss ¹ William Hebgen Guss ¹ Alex Nichol ¹ Alex Paino ¹
Nikolas Tezak ¹ Jie Tang ¹ Igor Babuskin ¹ Suchir Balaji ¹ Shantanu Jain ¹ William Saunders ¹
Christopher Hesse ¹ Andrew N. Carr ¹ Jan Leike ¹ Josh Achiam ¹ Vedant Misra ¹ Evan Morikawa ¹
Alec Radford ¹ Matthew Knight ¹ Miles Brundage ¹ Mira Murati ¹ Katie Mayer ¹ Peter Welinder ¹
Bob McGrew ¹ Dario Amodei ² Sam McCandlish ² Ilya Sutskever ¹ Wojciech Zaremba ¹

Abstract

We introduce Codex, a GPT language model fine-tuned on publicly available code from GitHub, and study its Python code-writing capabilities. A distinct production version of Codex powers GitHub Copilot. On HumanEval, a new evaluation set we release to measure functional correctness for synthesizing programs from docstrings, our model solves 28.8% of the problems, while GPT-3 solves 0% and GPT-J solves 11.4%. Furthermore, we find that repeated sampling from the model is a surprisingly effective strategy for producing working solutions to difficult prompts. Using this method, we solve 70.2% of our problems with 100 samples per problem. Careful investigation of our model reveals its limitations, including difficulty with docstrings describing long chains of operations and with binding operations to variables. Finally, we discuss the potential broader impacts of deploying powerful code generation technologies, covering safety, security, and economics.

1. Introduction

Scalable sequence prediction models (Graves, 2014; Vaswani et al., 2017; Child et al., 2019) have become a general-purpose method for generation and representation learning in many domains, including natural language processing (Mikolov et al., 2013; Sutskever et al., 2014; Dai & Le, 2015; Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018), computer vision (Van Oord et al., 2016; Menick & Kalchbrenner, 2018; Chen et al., 2020; Bao et al., 2021), audio and speech processing (Oord et al., 2016, 2018; Dhariwal et al., 2020; Baevski et al., 2020), biology (Alley et al., 2019; Rives et al., 2021), and even across multiple modalities (Das et al., 2017; Lu et al., 2019; Ramesh et al., 2021; Zellers et al., 2021). More recently, language models have also fueled progress towards the longstanding challenge of program synthesis (Simon, 1963; Manna & Waldinger, 1971), spurred by the presence of code in large datasets (Husain et al., 2019; Gao et al., 2020) and the resulting programming capabilities of language models trained on these datasets (Wang & Komatsu, 2021). Popular language modeling objectives like masked language modeling (Devlin et al., 2018) and span prediction (Raffel et al., 2020) have also been adapted to train their programming counterparts CodeBERT (Feng et al., 2020) and PyMT5 (Clement et al.,

Codex

<https://copilot.github.com/>



<https://arxiv.org/pdf/2107.03374.pdf>

Transformers y Modelos de Lenguaje

Palm OPT

Explaining a joke

Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

PaLM explains an original joke with two-shot prompts

<https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

<https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/>

Transformers y Modelos de Lenguaje

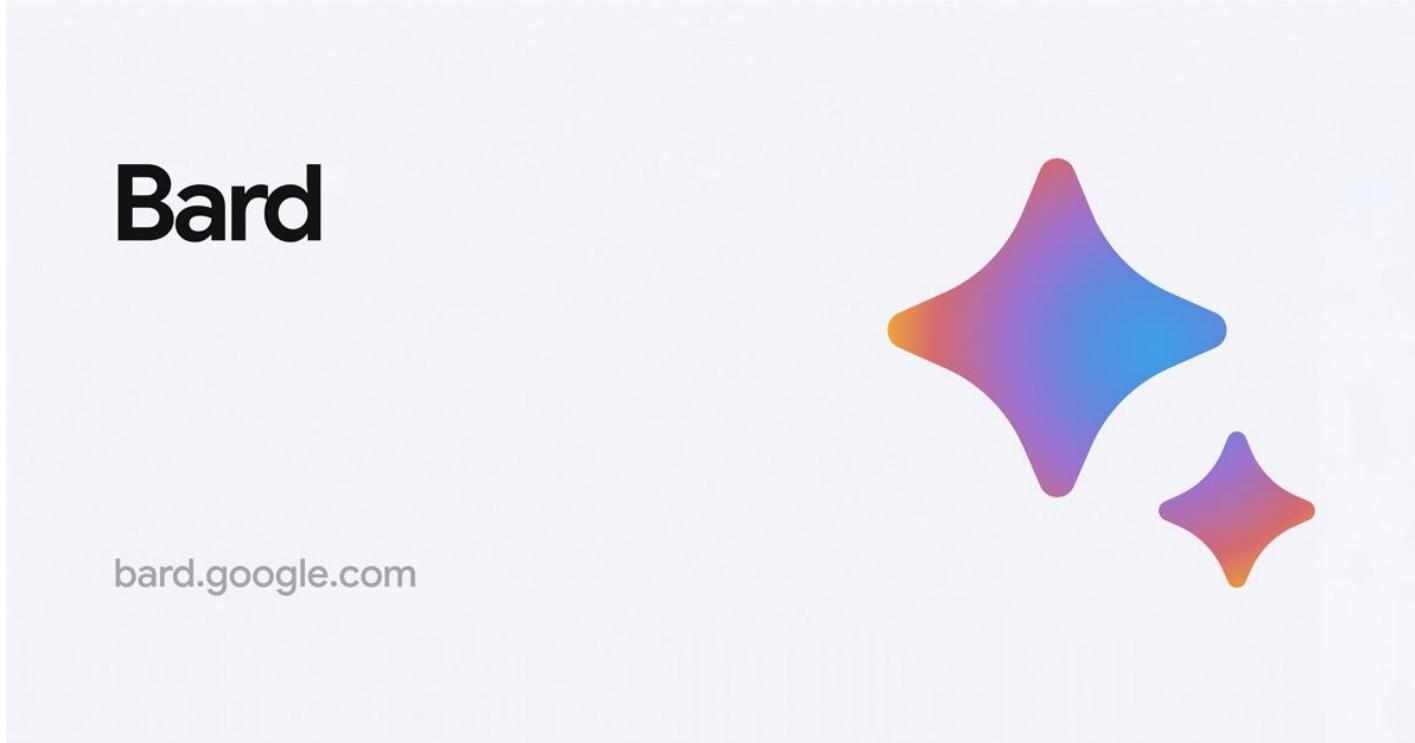


LaMDA

You would get to see a massive canyon, some frozen icebergs, geysers, and some craters.

<https://blog.google/technology/ai/lamda/>

Transformers y Modelos de Lenguaje



Bard

<https://bard.google.com/>

Transformers y Modelos de Lenguaje



ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

TRY CHATGPT ↗

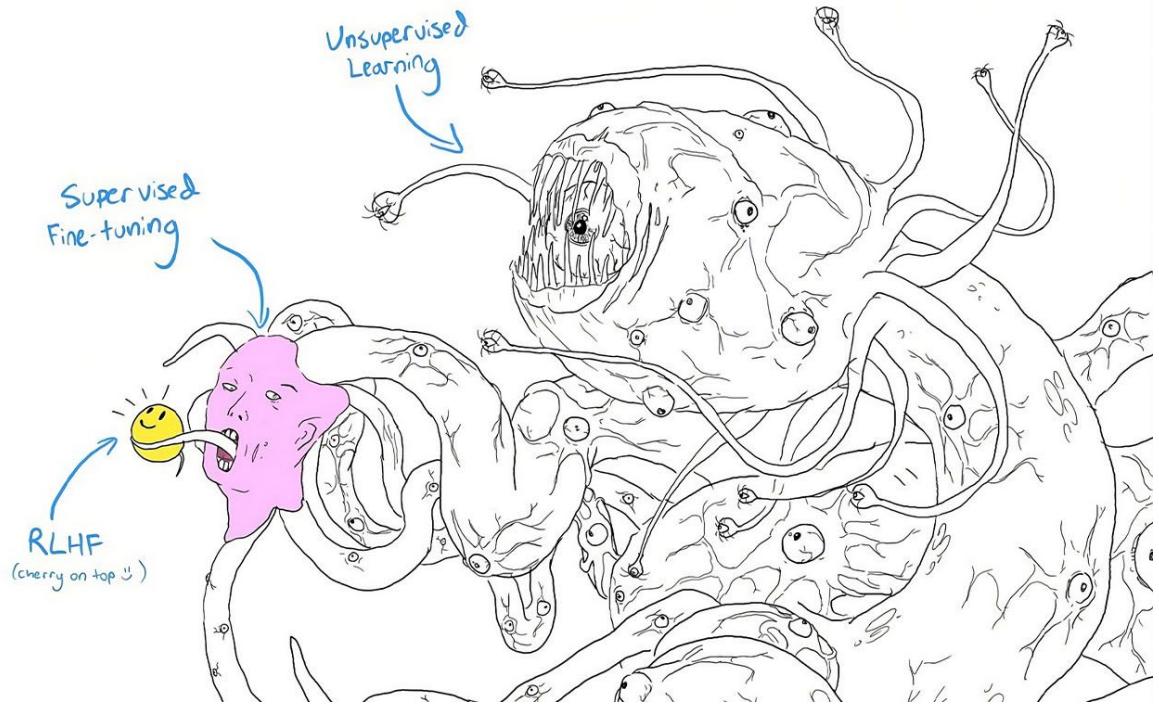
El caso ChatGPT

<https://openai.com/blog/chatgpt/>

Chat GPT

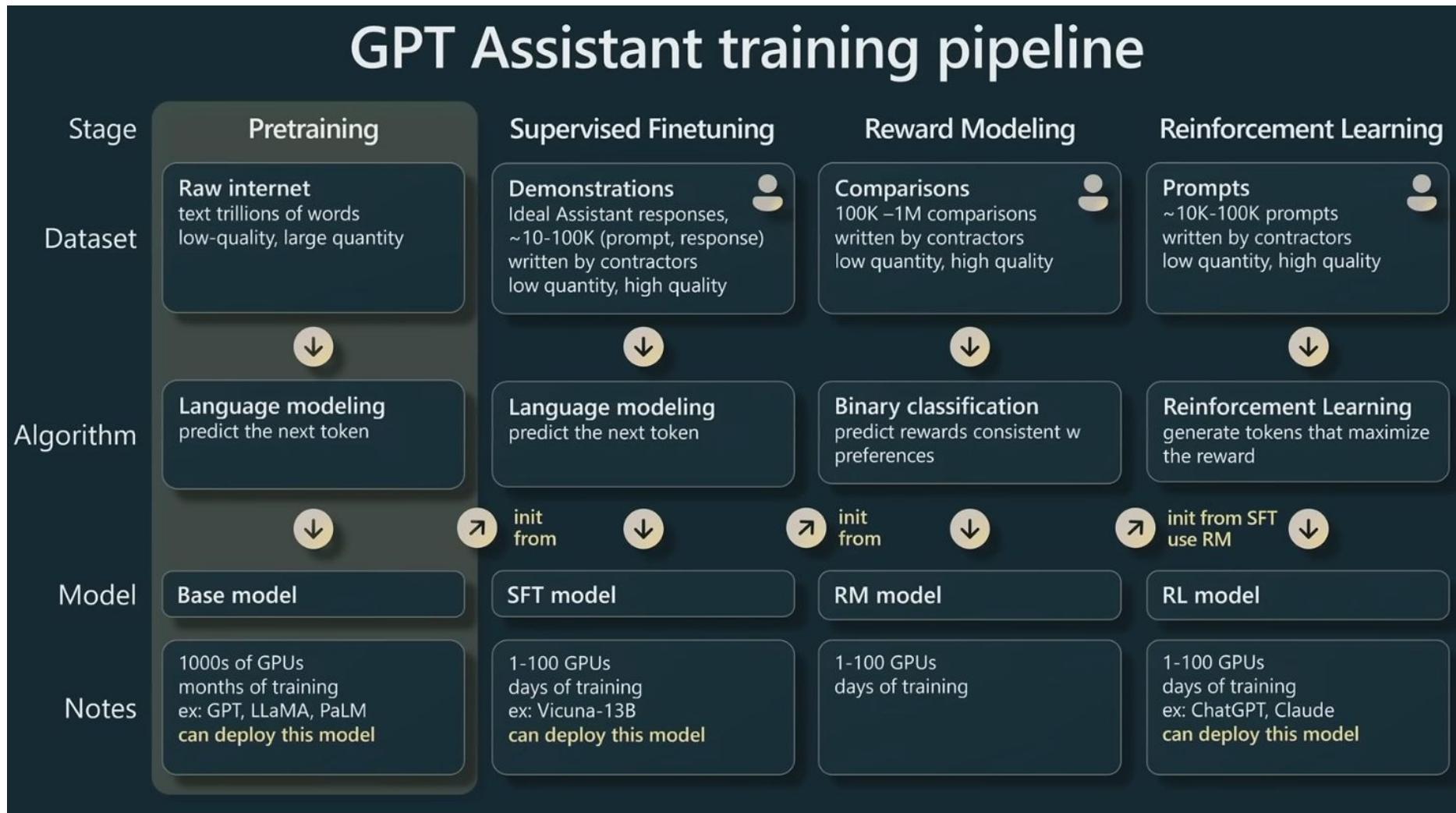
Reinforcement learning from Human Feedback

“Los modelos entrenados por RLHF pueden proporcionar respuestas que se alineen con los valores humanos, generar respuestas más detalladas y rechazar preguntas que sean inapropiadas o estén fuera del espacio de conocimiento del modelo” Por lo tanto pueden ser usados para disminuir el sesgo en la respuestas de los LLM



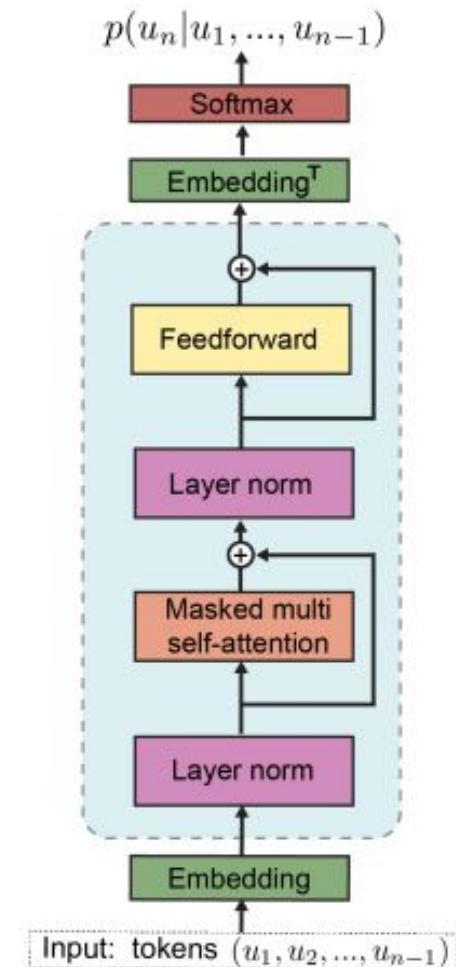
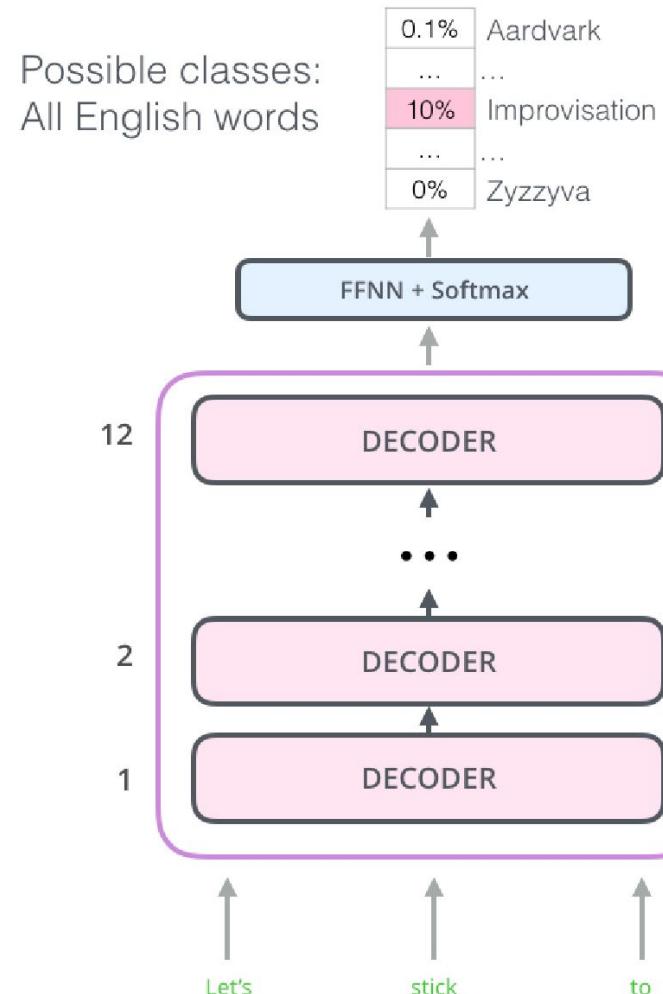
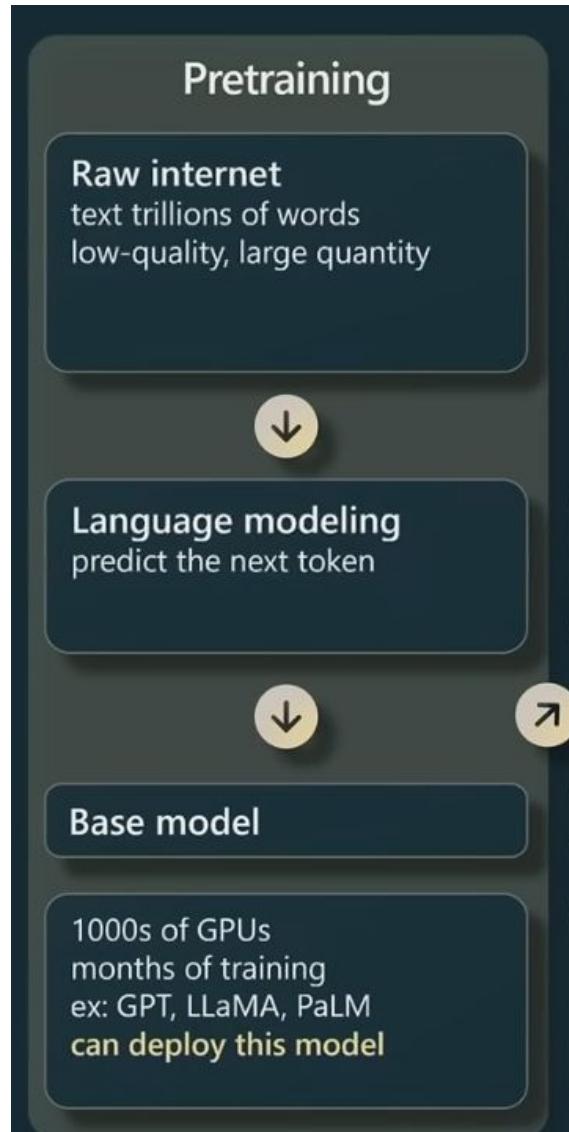
https://i.kym-cdn.com/entries/icons/original/000/044/025/shoggothhh_header.jpg

Chat GPT



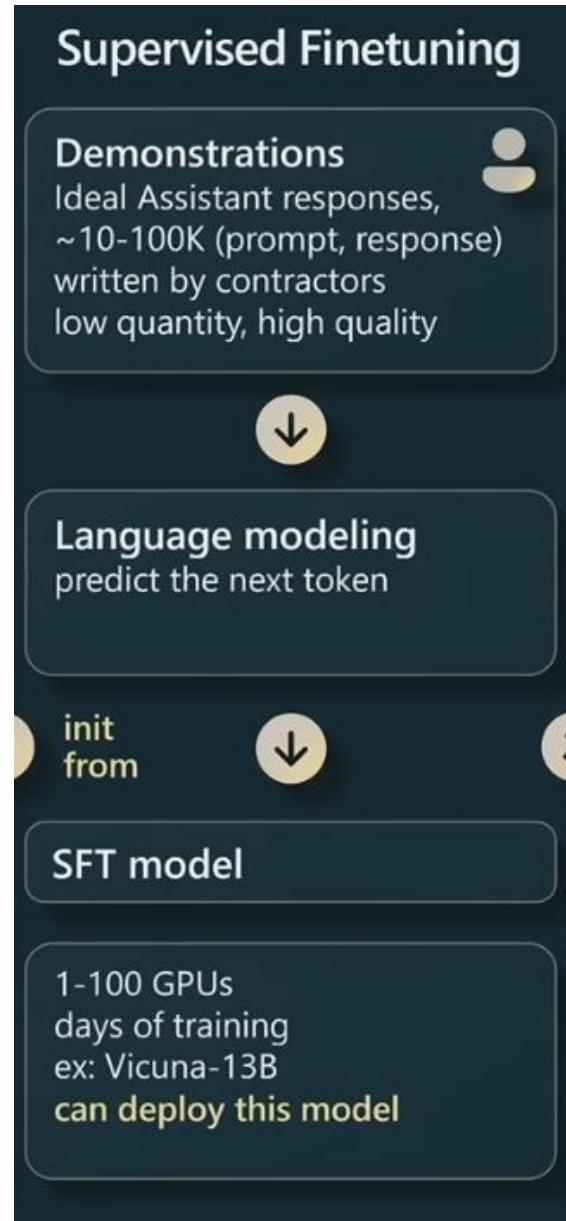
<https://www.youtube.com/watch?v=bZQun8Y4L2A&t=16s>

Chat GPT



<https://www.youtube.com/watch?v=bZQun8Y4L2A&t=16s>

Chat GPT



Finetune on many tasks (“instruction-tuning”)

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.

How would you accomplish this goal?

OPTIONS:

- Keep stack of pillow cases in fridge.
- Keep stack of pillow cases in oven.

Target

keep stack of pillow cases in fridge

Input (Translation)

Translate this sentence to Spanish:

The new office building was built in less than three months.

Target

El nuevo edificio de oficinas se construyó en tres meses.

Sentiment analysis tasks

Coreference resolution tasks

...



Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.

Hypothesis: It's not certain how many lessons you'll learn by your thirties.

Does the premise entail the hypothesis?

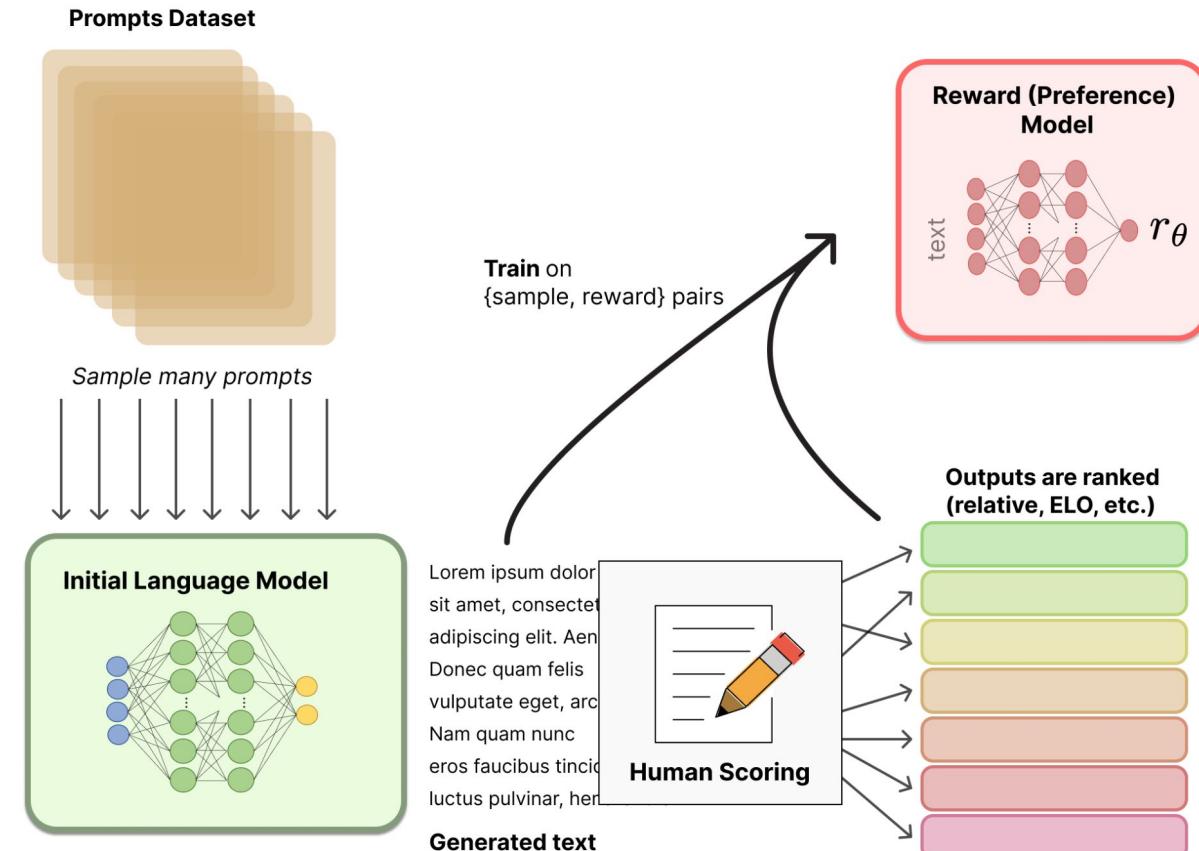
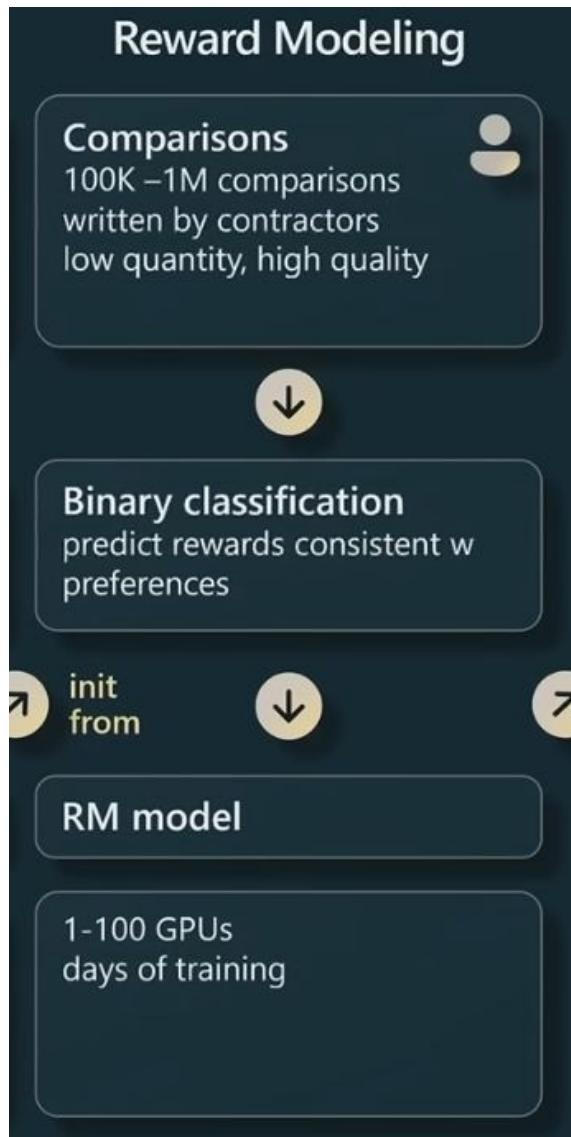
OPTIONS:

- yes
- it is not possible to tell
- no

FLAN Response

It is not possible to tell

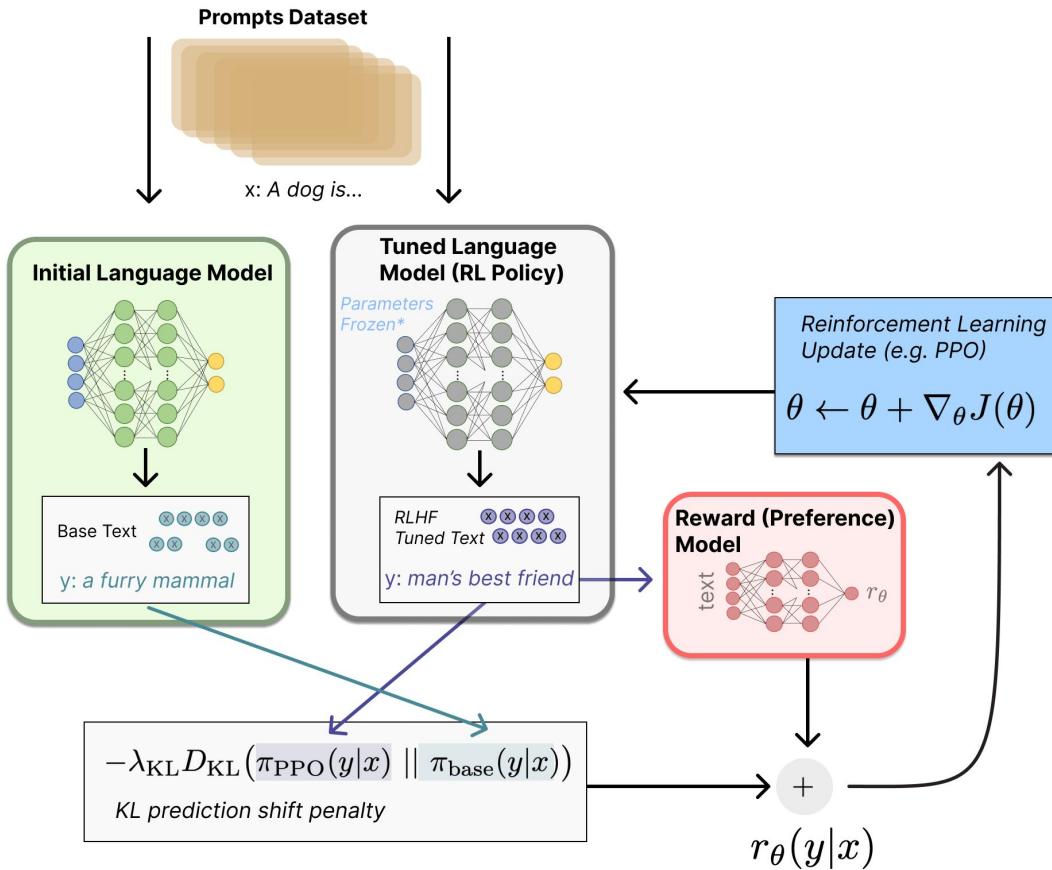
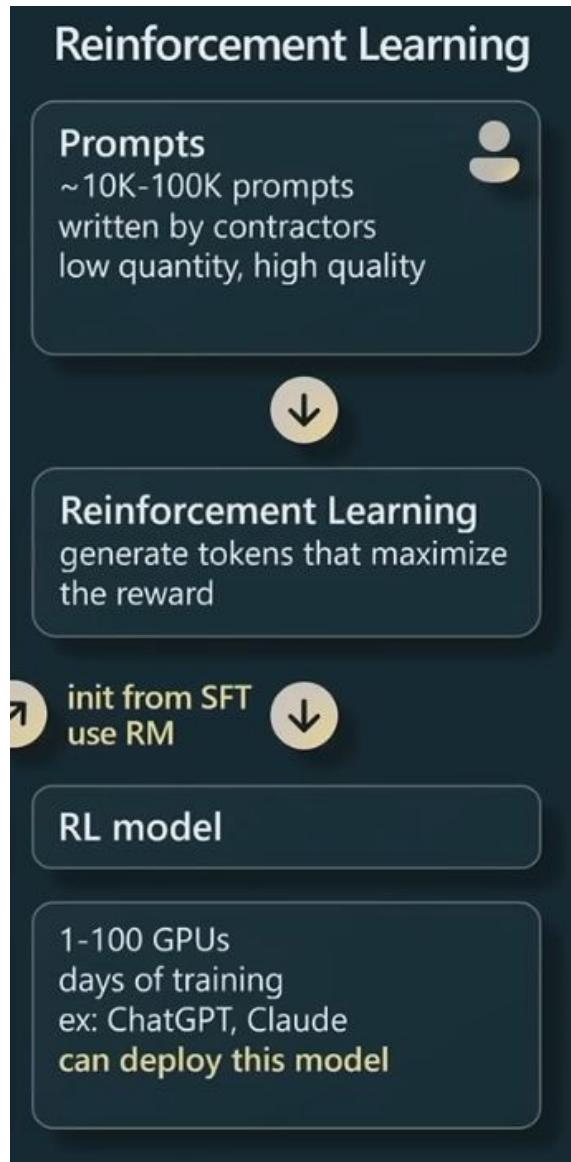
Chat GPT



<https://www.youtube.com/watch?v=bZQun8Y4L2A&t=16s>

<https://huggingface.co/blog/rhf>

Chat GPT



<https://www.youtube.com/watch?v=bZQun8Y4L2A&t=16s>

Transformers y Modelos de Imágenes

Vision Transformer

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*,
Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†

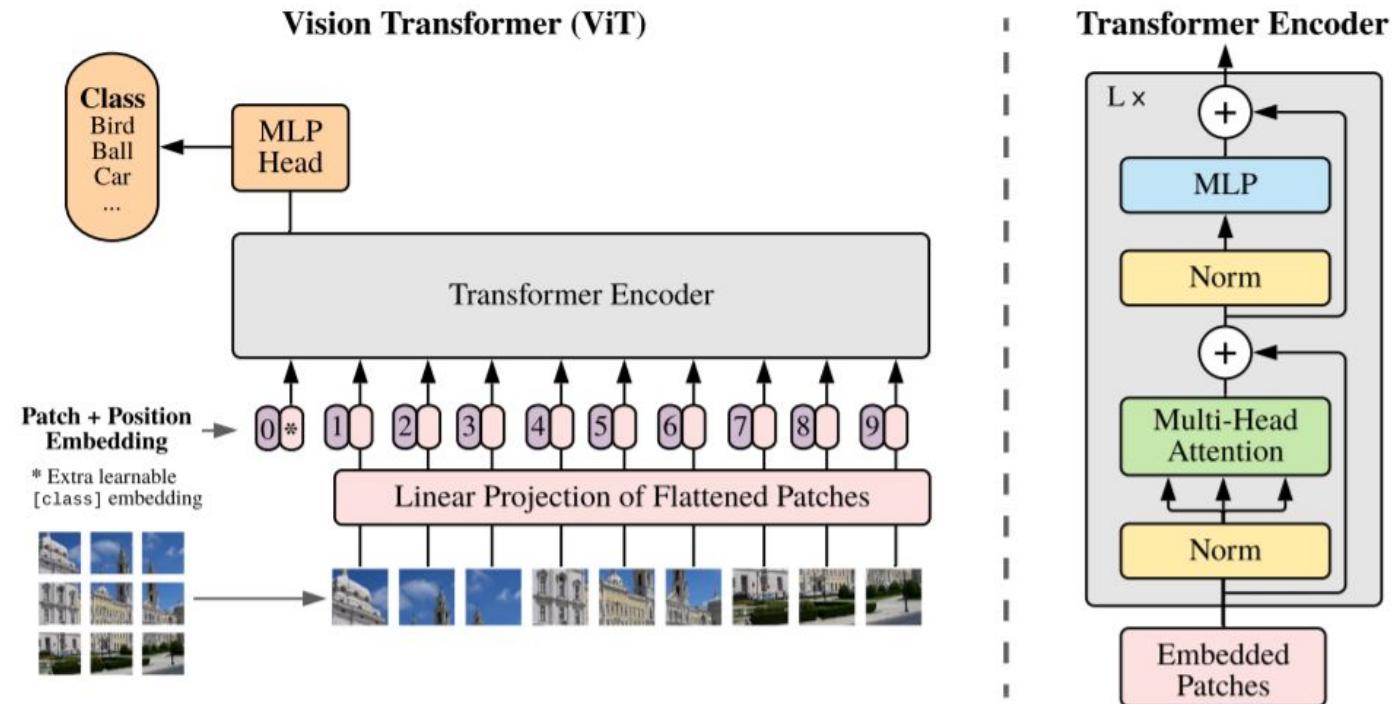
*equal technical contribution, †equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.¹



<https://arxiv.org/pdf/2010.11929.pdf>

Transformers y Modelos de Imágenes

Image GPT

Generative Pretraining from Pixels

Mark Chen¹ Alec Radford¹ Rewon Child¹ Jeff Wu¹ Heewoo Jun¹ Prafulla Dhariwal¹ David Luan¹
Ilya Sutskever¹

Abstract

Inspired by progress in unsupervised representation learning for natural language, we examine whether similar models can learn useful representations for images. We train a sequence Transformer to auto-regressively predict pixels, without incorporating knowledge of the 2D input structure. Despite training on low-resolution ImageNet without labels, we find that a GPT-2 scale model learns strong image representations as measured by linear probing, fine-tuning, and low-data classification. On CIFAR-10, we achieve 96.3% accuracy with a linear probe, outperforming a supervised Wide ResNet, and 99.0% accuracy with full fine-tuning, matching the top supervised pre-trained models. An even larger model trained on a mixture of ImageNet and web images is competitive with self-supervised benchmarks on ImageNet, achieving 72.0% top-1 accuracy on a linear probe.

ported strong results using a single layer of learned features (Coates et al., 2011), or even random features (Huang et al., 2014; May et al., 2017). The approach fell out of favor as the state of the art increasingly relied on directly encoding prior structure into the model and utilizing abundant supervised data to directly learn representations (Krizhevsky et al., 2012; Graves & Jaitly, 2014). Retrospective study of unsupervised pre-training demonstrated that it could even hurt performance in modern settings (Paine et al., 2014).

Instead, unsupervised pre-training flourished in a different domain. After initial strong results for word vectors (Mikolov et al., 2013), it has pushed the state of the art forward in Natural Language Processing on most tasks (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018). Interestingly, the training objective of a dominant approach like BERT, the prediction of corrupted inputs, closely resembles that of the Denoising Autoencoder, which was originally developed for images.

<https://openai.com/blog/gpt-2/>

https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf

Transformers y Modelos de Imágenes

Image GPT

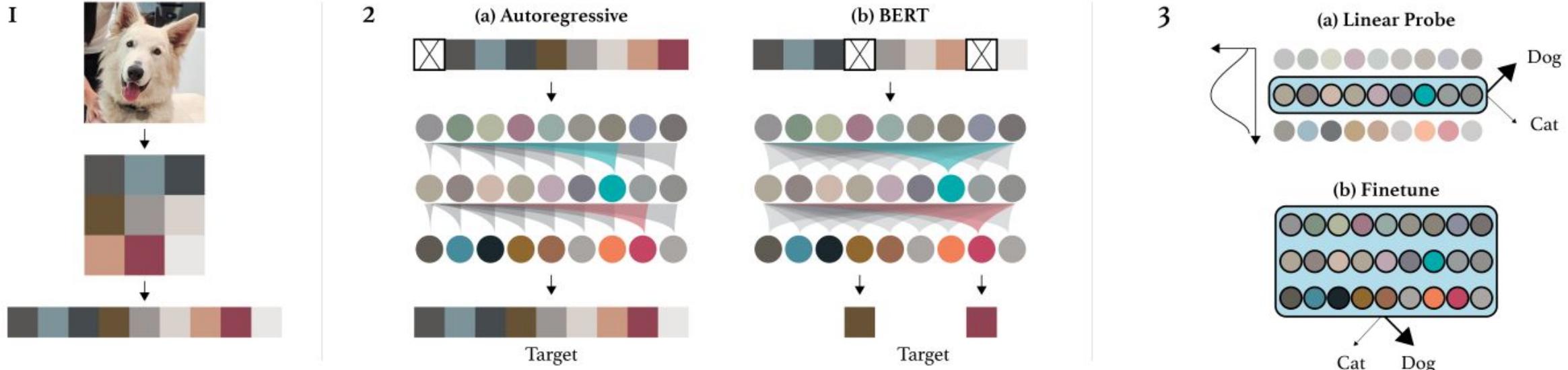


Figure 1. An overview of our approach. First, we pre-process raw images by resizing to a low resolution and reshaping into a 1D sequence. We then chose one of two pre-training objectives, auto-regressive next pixel prediction or masked pixel prediction. Finally, we evaluate the representations learned by these objectives with linear probes or fine-tuning.

<https://openai.com/blog/image-gpt/>

https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf

Transformers y Modelos de Imágenes

CLIP

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford ^{*1} Jong Wook Kim ^{*1} Chris Hallacy ¹ Aditya Ramesh ¹ Gabriel Goh ¹ Sandhini Agarwal ¹
Girish Sastry ¹ Amanda Askell ¹ Pamela Mishkin ¹ Jack Clark ¹ Gretchen Krueger ¹ Ilya Sutskever ¹

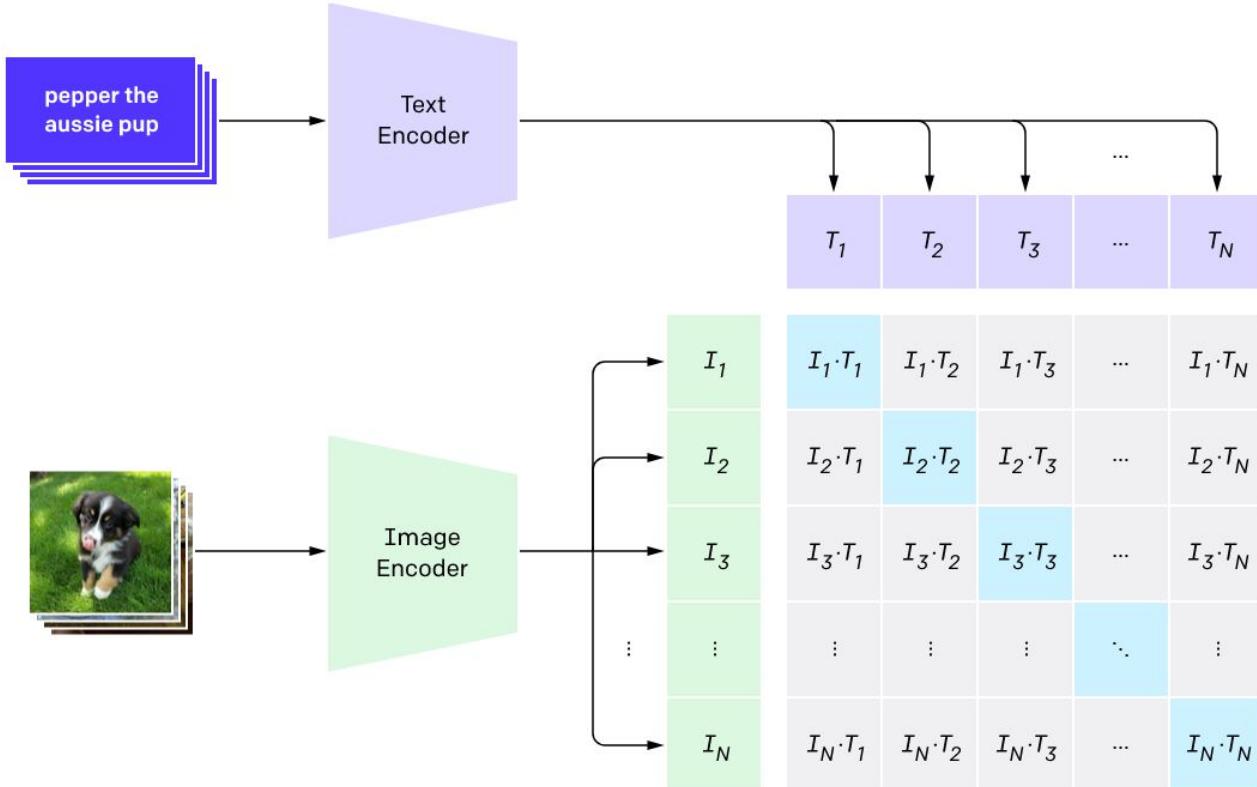
Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of “text-to-text” as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on

1. Contrastive pre-training



<https://openai.com/blog/clip/>

<https://arxiv.org/pdf/2010.11929.pdf>

Transformers y Modelos Generadores de Imágenes

Dalle-1

Zero-Shot Text-to-Image Generation

Aditya Ramesh¹ Mikhail Pavlov¹ Gabriel Goh¹ Scott Gray¹
Chelsea Voss¹ Alec Radford¹ Mark Chen¹ Ilya Sutskever¹

Abstract

Text-to-image generation has traditionally focused on finding better modeling assumptions for training on a fixed dataset. These assumptions might involve complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks supplied during training. We describe a simple approach for this task based on a transformer that autoregressively models the text and image tokens as a single stream of data. With sufficient data and scale, our approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.



<https://openai.com/blog/dalle-1/>

<https://arxiv.org/pdf/2102.12092.pdf>

Transformers y Modelos Generadores de Imágenes

Dalle-2

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh*
OpenAI
aramesh@openai.com

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Casey Chu*
OpenAI
casey@openai.com

Mark Chen
OpenAI
mark@openai.com

Abstract

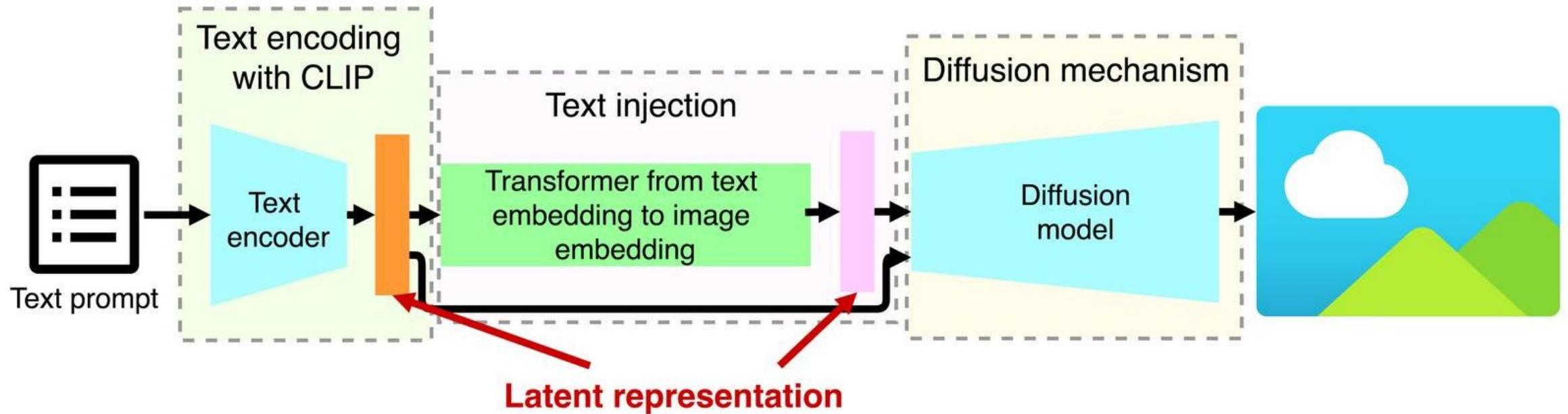
Contrastive models like CLIP have been shown to learn robust representations of images that capture both semantics and style. To leverage these representations for image generation, we propose a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. We show that explicitly generating image representations improves image diversity with minimal loss in photorealism and caption similarity. Our decoders conditioned on image representations can also produce variations of an image that preserve both its semantics and style, while varying the non-essential details absent from the image representation. Moreover, the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion. We use diffusion models for the decoder and experiment with both autoregressive and diffusion models for the prior, finding that the latter are computationally more efficient and produce higher-quality samples.

<https://openai.com/dall-e-2/>

<https://arxiv.org/pdf/2204.06125.pdf>

Transformers y Modelos Generadores de Imágenes

DALL-E 2



<https://newsletter.theaiedge.io/p/everything-you-needed-to-know-about>

Transformers y Modelos Generadores de Imágenes

Imagen

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†,
Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan,
S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans,
Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac, williamchan, mnorouzi}@google.com
{srbs, lala, jwhang, jonathanho, davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada

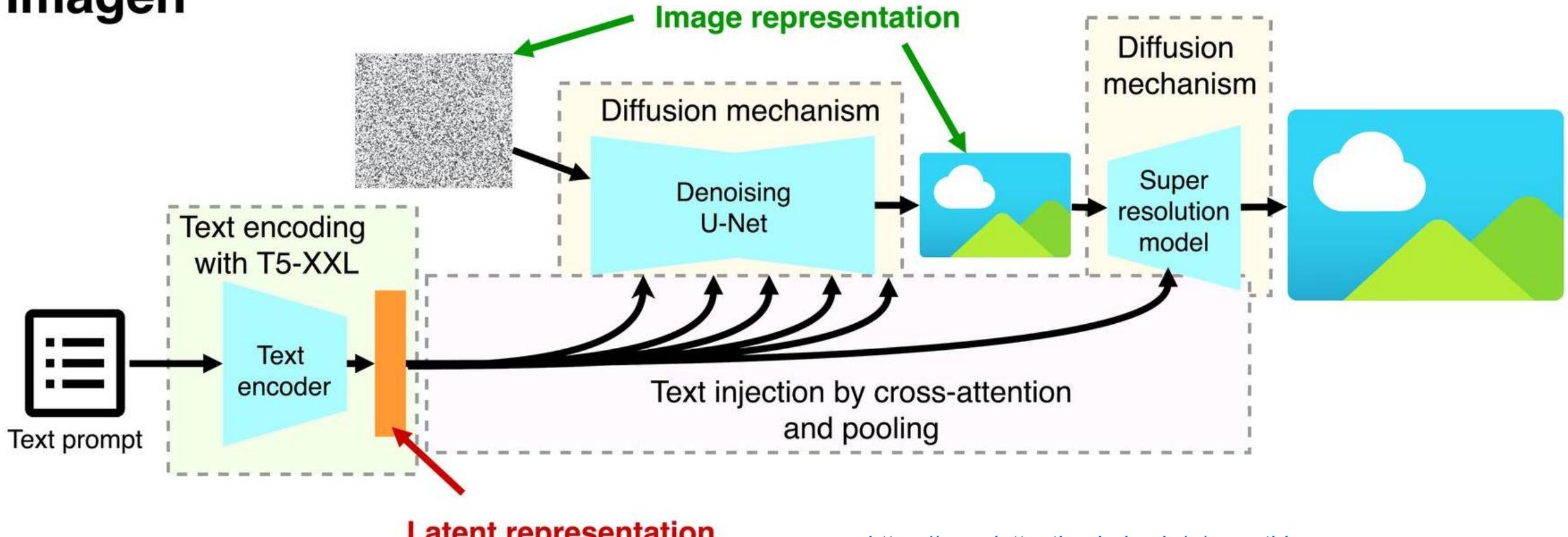
Abstract

We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g. T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, GLIDE and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment. See imagen.research.google for an overview of the results.

<https://arxiv.org/pdf/2205.11487.pdf>

Transformers y Modelos Generadores de Imágenes

Imagen



<https://newsletter.theaiedge.io/p/everything-you-needed-to-know-about>

Transformers y Modelos Generadores de Imágenes

Stable Diffusion

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach¹ * Andreas Blattmann¹ * Dominik Lorenz¹ Patrick Esser² Björn Ommer¹

¹Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany

²Runway ML

<https://github.com/CompVis/latent-diffusion>

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text



Figure 1. Boosting the upper bound on achievable quality with less aggressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f . Reconstruction FIDs [29] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

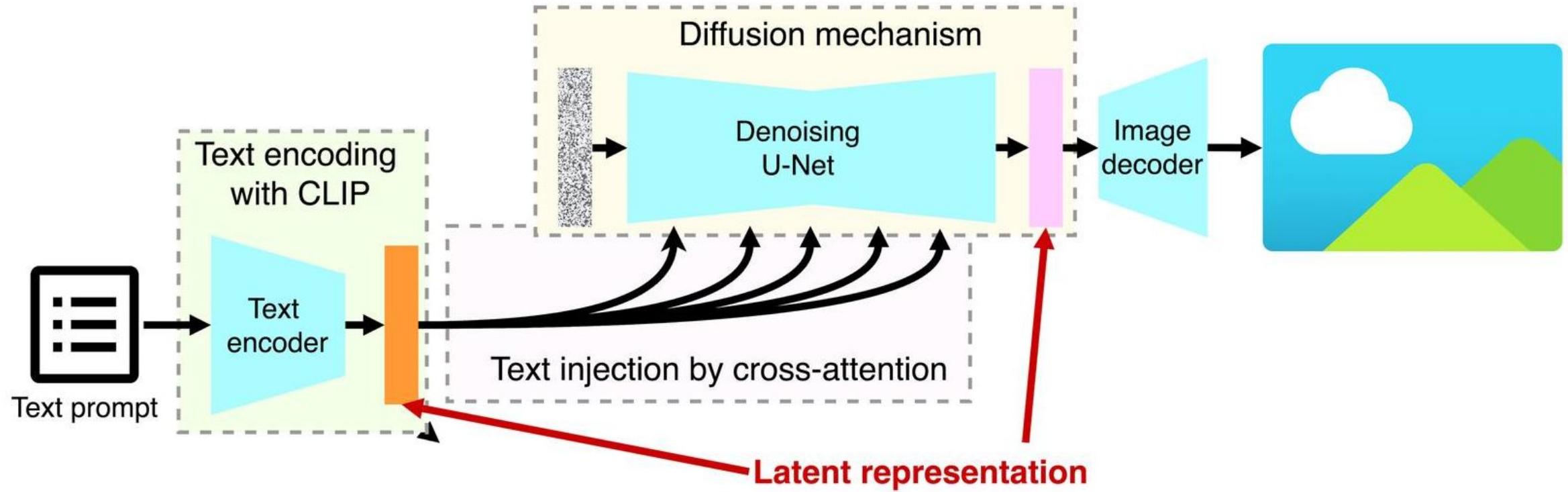
results in image synthesis F30.851 and beyond F7.45.48.571

<https://jalammar.github.io/illustrated-stable-diffusion/>

<https://arxiv.org/pdf/2112.10752.pdf>

Transformers y Modelos Generadores de Imágenes

Stable Diffusion



<https://newsletter.theaiedge.io/p/everything-you-needed-to-know-about>

Transformers y Modelos Multimodales

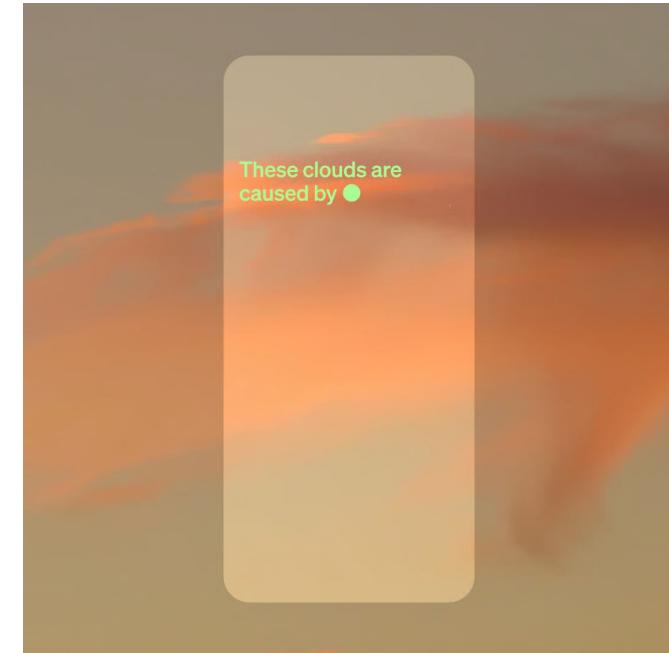
GPT-4



<https://openai.com/research/gpt-4>

GPT-4V

Recibe Imágenes y Texto



[https://openai.com/blog/chatgpt-can
-now-see-hear-and-speak?s=03](https://openai.com/blog/chatgpt-can-now-see-hear-and-speak?s=03)

Transformers y Modelos Multimodales

Inteligencia Artificial Multimodal o Polivalente

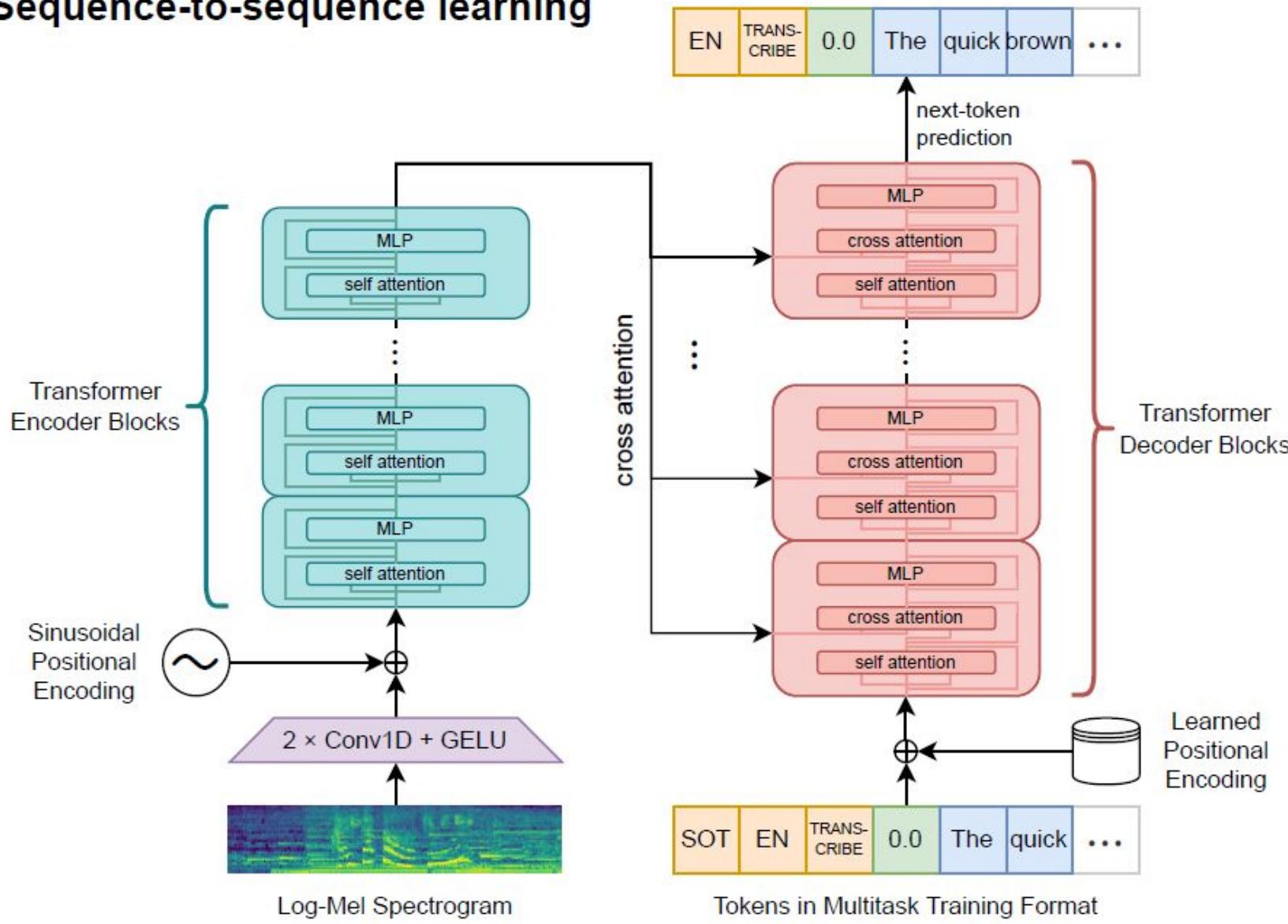


GEMINI

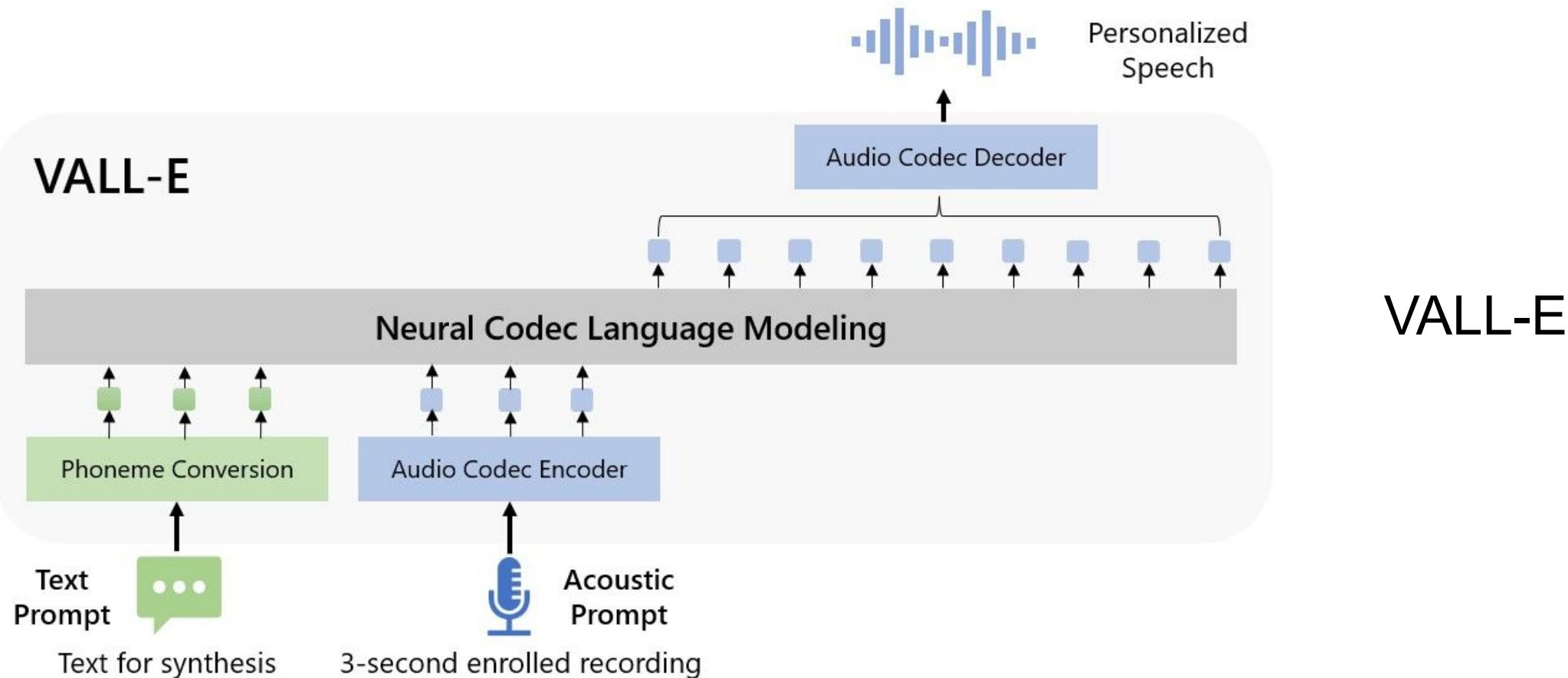
Sistema multimodal de Google
Actualmente en entrenamiento y/o
pruebas privadas

Transformers y Modelos de Audio

Sequence-to-sequence learning



Transformers y Modelos Generadores de Audio



<https://valle-demo.github.io/>

Transformers y Modelos Generadores de Audio

Music LM

MusicLM: Generating Music From Text

Andrea Agostinelli ^{*1} Timo I. Denk ^{*1}
Zalán Borsos ¹ Jesse Engel ¹ Mauro Verzetti ¹ Antoine Caillon ² Qingqing Huang ¹ Aren Jansen ¹
Adam Roberts ¹ Marco Tagliasacchi ¹ Matt Sharifi ¹ Neil Zeghidour ¹ Christian Frank ¹

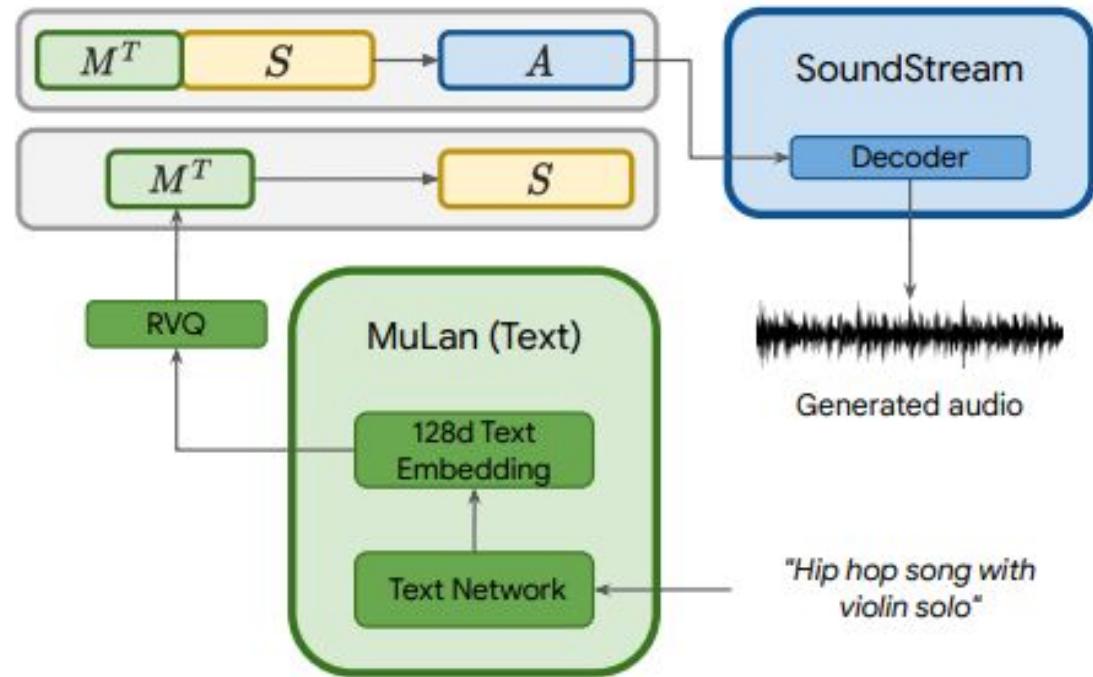
Abstract

We introduce MusicLM, a model for generating high-fidelity music from text descriptions such as “*a calming violin melody backed by a distorted guitar riff*”. MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text descriptions. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts. google-research.github.io/seanet/musiclm/examples

period of seconds. Hence, turning a single text caption into a rich audio sequence with long-term structure and many stems, such as a music clip, remains an open challenge.

AudioLM (Borsos et al., 2022) has recently been proposed as a framework for audio generation. Casting audio synthesis as a language modeling task in a discrete representation space, and leveraging a hierarchy of coarse-to-fine audio discrete units (or *tokens*), AudioLM achieves both high-fidelity and long-term coherence over dozens of seconds. Moreover, by making no assumptions about the content of the audio signal, AudioLM learns to generate realistic audio from audio-only corpora, be it speech or piano music, without any annotation. The ability to model diverse signals suggests that such a system could generate richer outputs if trained on the appropriate data.

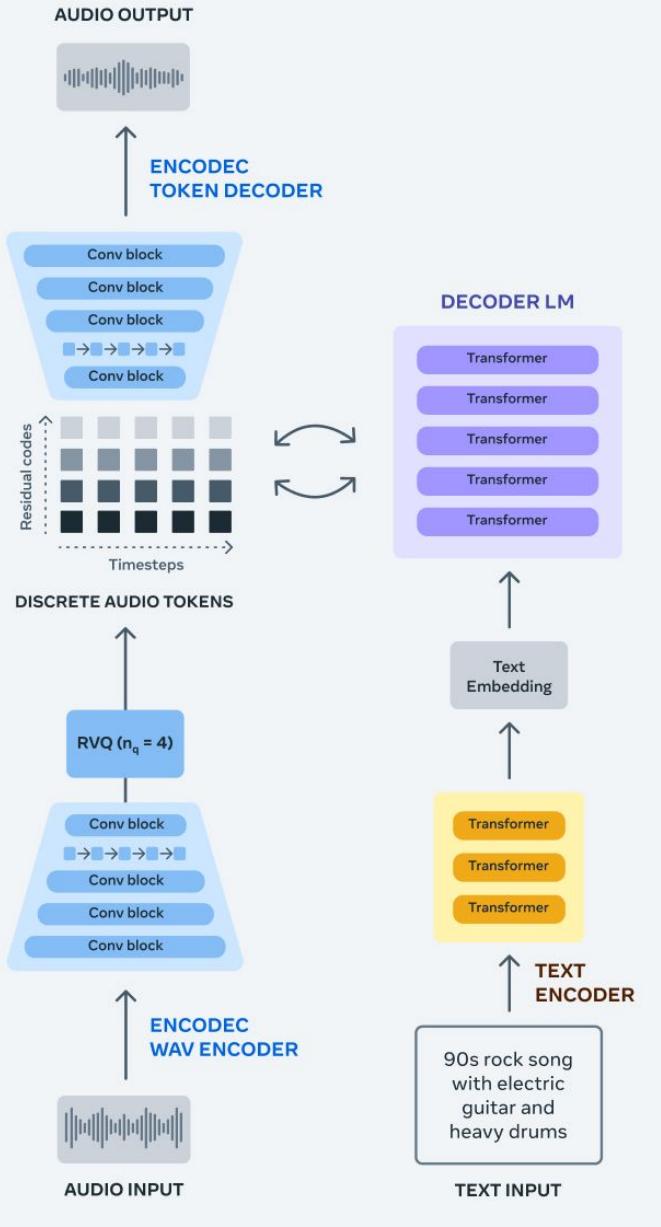
Besides the inherent difficulty of synthesizing high-quality and coherent audio, another impeding factor is the scarcity of paired audio-text data. This is in stark contrast with the image domain, where the availability of massive datasets



<https://arxiv.org/pdf/2301.11325.pdf>

Transformers y Modelos Generadores de Audio

Audiocraft



<https://audiocraft.metamodelab.com/>

Transformers y Modelos Generadores de Video

Texto a Video



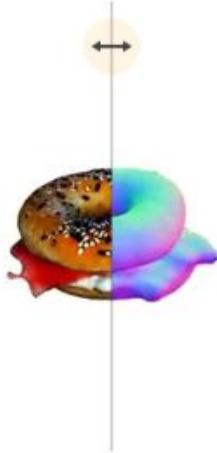
<https://makeavideo.studio/>



<https://www.nvidia.com/en-us/gpu-cloud/picasso/>

Transformers y Modelos Generadores de 3D

Texto a 3D



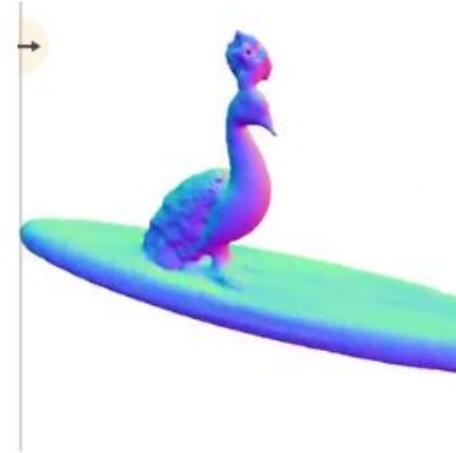
Reveal 3D mesh!

[...] a bagel filled with cream cheese and lox.



Reveal 3D mesh!

[...] an ice cream sundae.



Reveal 3D mesh!

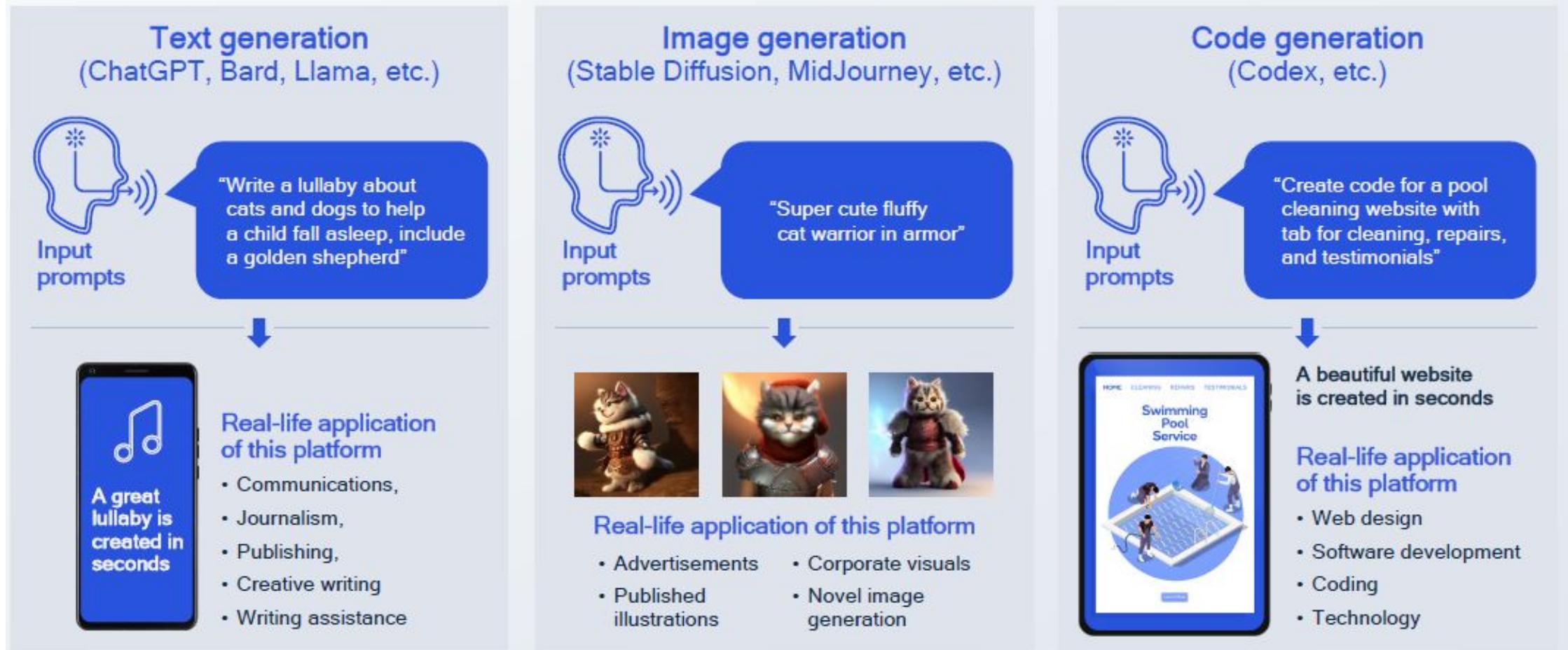
[...] a peacock on a surfboard.



<https://research.nvidia.com/labs/dir/magic3d/>

<https://www.nvidia.com/en-us/gpu-cloud/picasso/>

Inteligencia Artificial Generativa



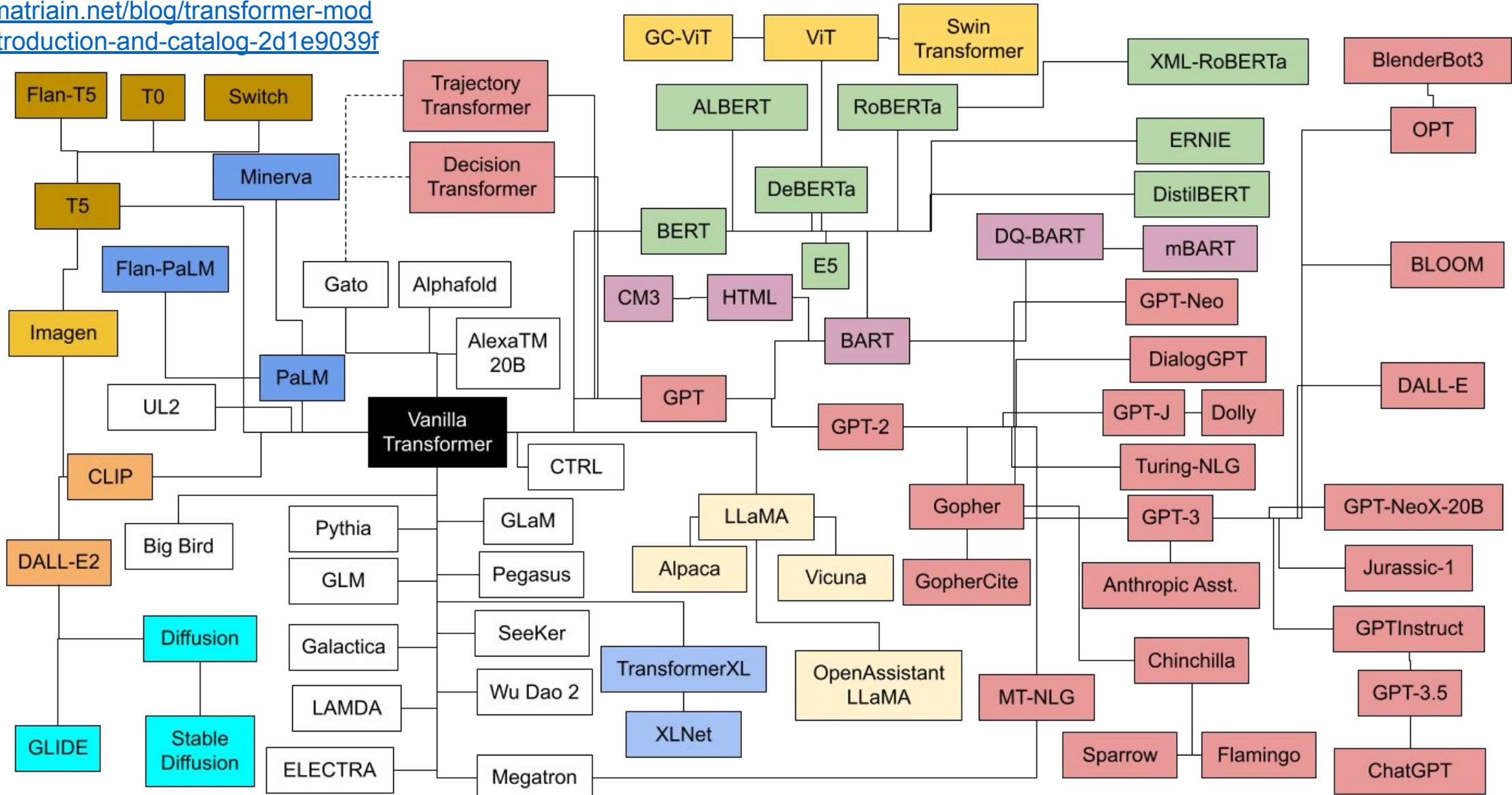
What is
generative AI?

AI models that create new and original content like text, images, video, audio, or other data

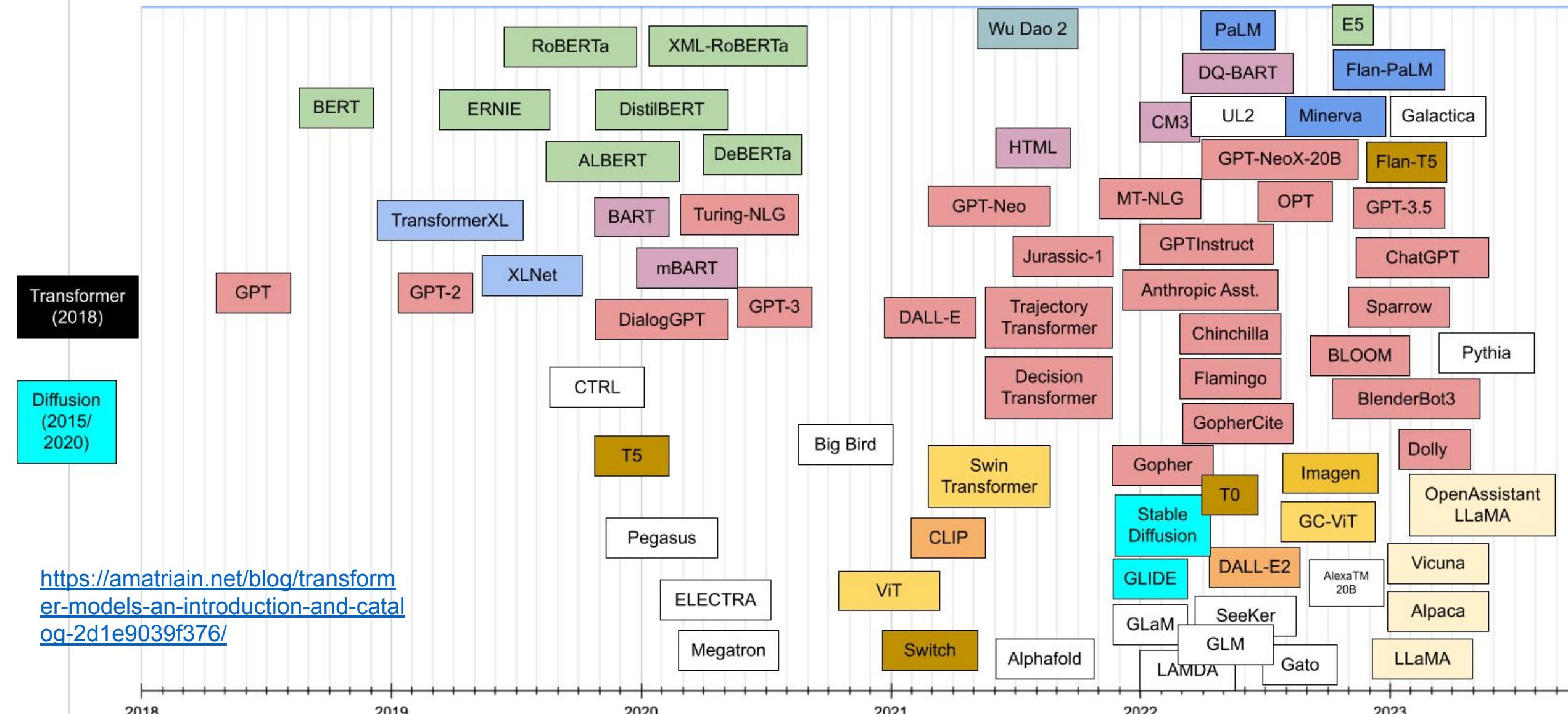
Generative AI, foundational models, and large language models are sometimes used interchangeably

Transformers ZOO

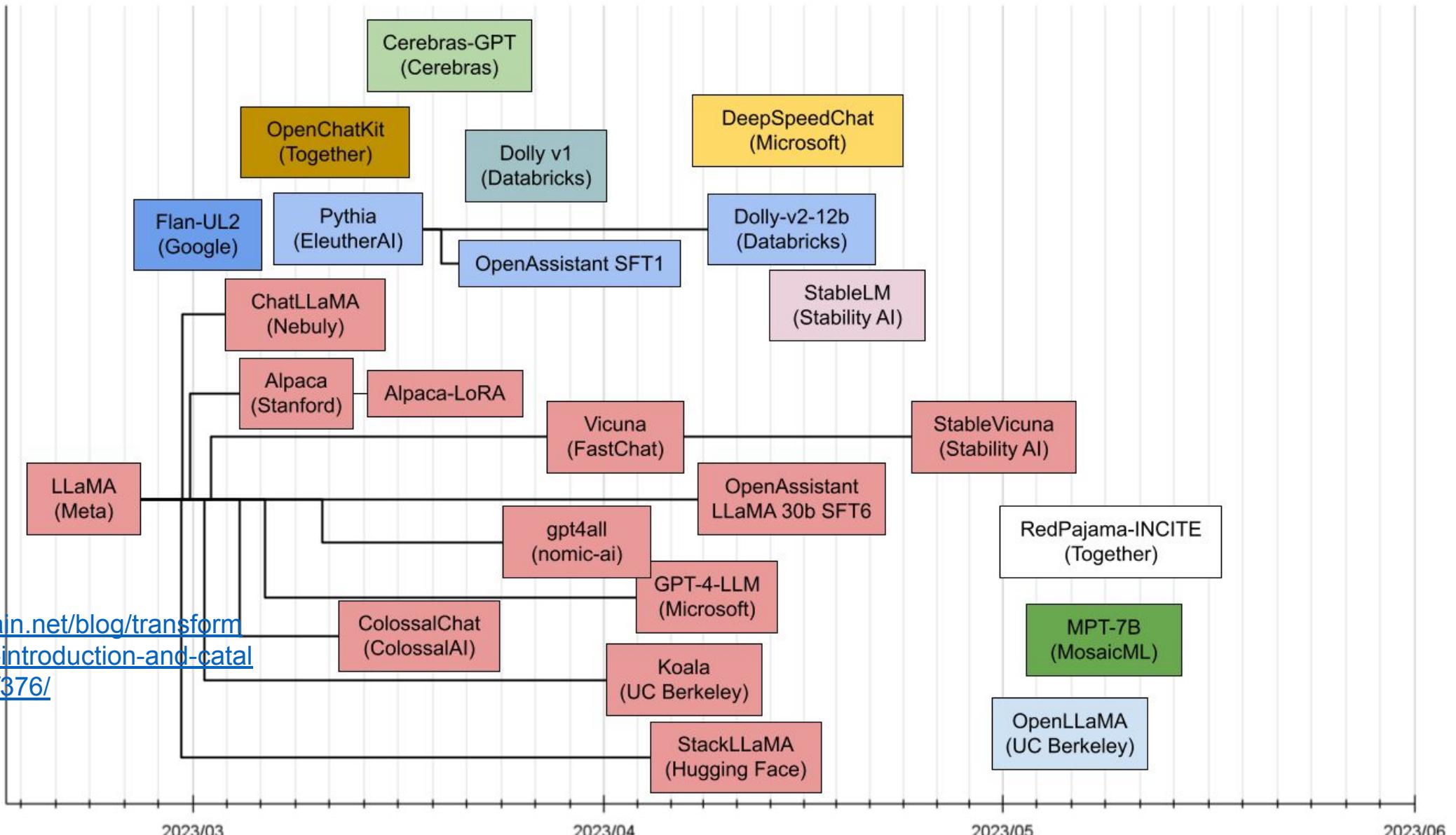
<https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e903f376/>



Transformers ZOO

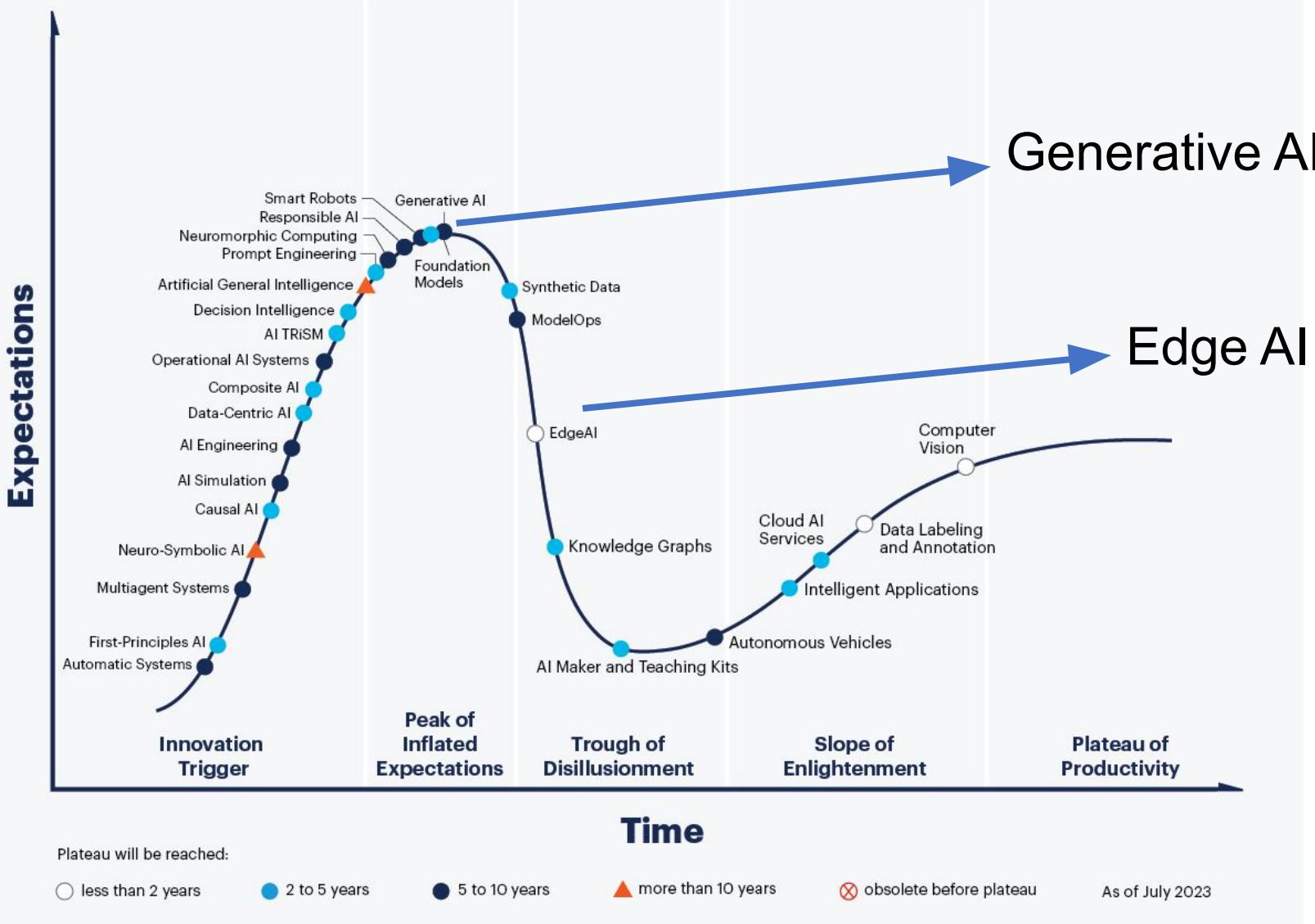


Transformers ZOO (Modelos Libres)



Los Extremos se Encuentran

Hype Cycle for Artificial Intelligence, 2023



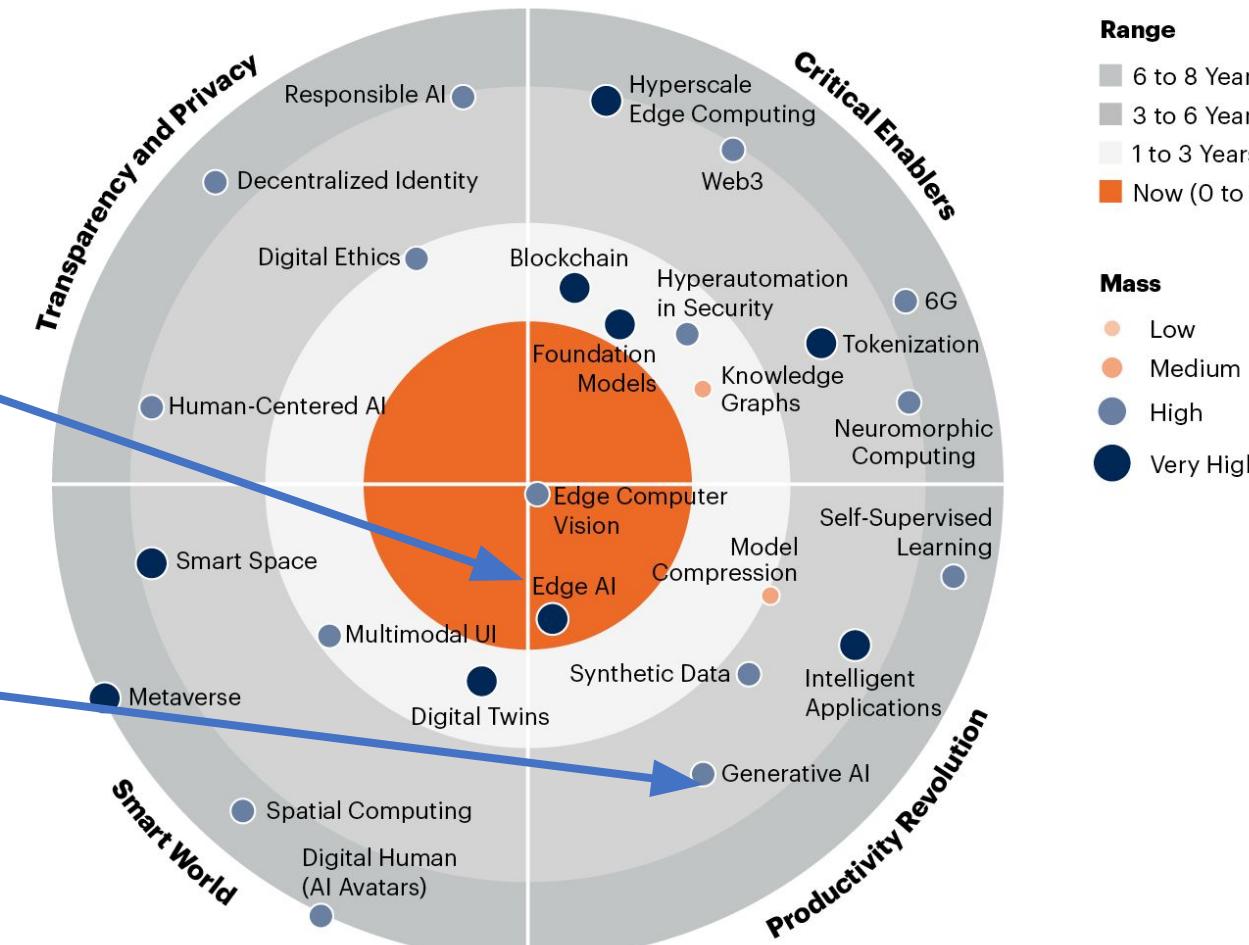
<https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle>

Edge AI y Generative AI

2023 Gartner Emerging Technologies and Trends Impact Radar

Edge AI
Generative AI

<https://www.gartner.com/en/articles/4-emerging-technologies-you-need-to-know-about>



Edge AI y Generative AI

Costo Económico y Ambiental

	GPU Type	GPU Power consumption	GPU-hours	Total power consumption	Carbon emitted (tCO ₂ eq)
OPT-175B	A100-80GB	400W	809,472	356 MWh	137
BLOOM-175B	A100-80GB	400W	1,082,880	475 MWh	183
LLaMA-7B	A100-80GB	400W	82,432	36 MWh	14
LLaMA-13B	A100-80GB	400W	135,168	59 MWh	23
LLaMA-33B	A100-80GB	400W	530,432	233 MWh	90
LLaMA-65B	A100-80GB	400W	1,022,362	449 MWh	173

Table 15: **Carbon footprint of training different models in the same data center.** We follow the formula from Wu et al. (2022) to compute carbon emission of train OPT, BLOOM and our models in the same data center. For the power consumption of a A100-80GB, we take the thermal design power (TDP) for NVLink systems, that is 400W. We take a PUE of 1.1 and a carbon intensity factor set at the national US average of 0.385 kg CO₂e per KWh.

<https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/>

... estimamos que utilizamos 2048 A100-80GB durante un período de aproximadamente 5 meses para desarrollar nuestros modelos

Esto muestra que el modelo más pequeño, LLaMA-7B, fue entrenado con 82.432 horas de GPU A100-80GB, con un costo de 36MWh y generando 14 toneladas de CO₂.
(Eso son aproximadamente 28 personas que vuelan de Londres a Nueva York).

Se estima que el costo del entrenamiento fue \$7,372,800

<https://simonwillison.net/2023/Mar/17/beat-chatgpt-in-a-browser/>

Edge AI y Generative AI

Costo Económico y Ambiental

HOME > TECH

ChatGPT could cost over \$700,000 per day to operate. Microsoft is reportedly trying to make it cheaper.

CLEAN ENERGY

Microsoft is hiring a nuclear energy expert to help power its AI and cloud data centers

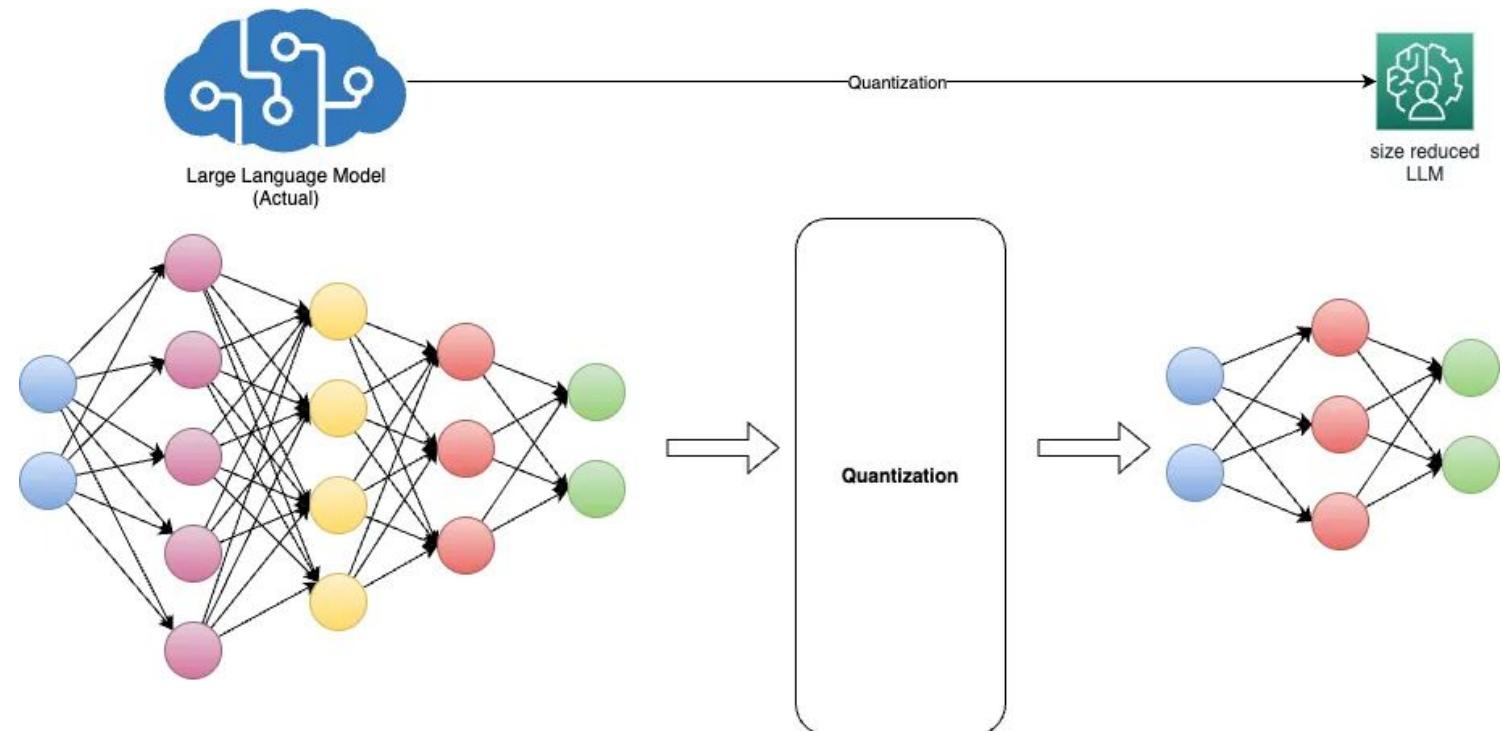
<https://www.cnbc.com/2023/09/25/microsoft-is-hiring-a-nuclear-energy-expert-to-help-power-data-centers.html>

La inteligencia artificial requiere mucha potencia informática y Microsoft está elaborando una hoja de ruta para impulsar ese proceso con pequeños reactores nucleares

<https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4>

Edge AI y Generative AI

- Optimización de los modelos de IA
- Cuantización
- Pruning
- Knowledge distillation

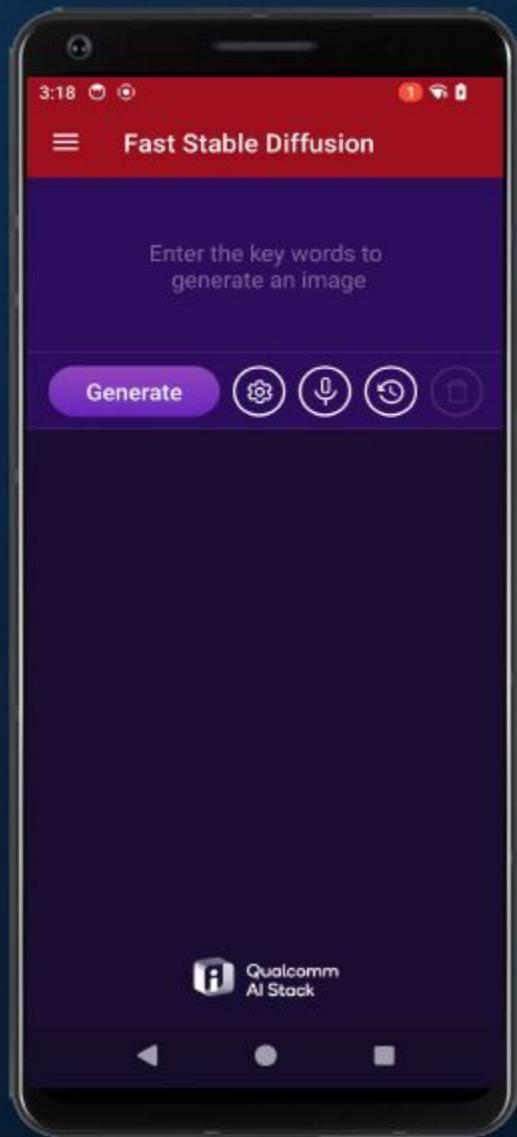


<https://int8.io/local-large-language-models-beginners-guide/>

<https://www.linkedin.com/pulse/quantization-what-you-should-understand-want-run-langs-pavan-mantha>

Edge AI y Generative AI

World's fastest AI
text-to-image
generative AI
on a phone



Takes less than 0.6 seconds for generating 512x512 images from text prompts

Efficient UNet architecture, guidance conditioning, and step distillation

Full-stack AI optimization to achieve this improvement

Edge AI y Generative AI

The image shows two smartphones side-by-side against a dark background. Both phones have a black screen with a purple gradient at the top. The left phone displays the "AI Assistant" app interface, showing a message "What is the most popular cookie?" and a response "The most popular cookie is chocolate chip." Below the screen is a purple input bar with a microphone icon and a placeholder "Enter your prompt here". The right phone displays the "Trip Planner" app interface, showing a message "I would like to go to San Diego from Toronto on December 10th and return on December 20th." followed by a travel plan summary: "Here is the travel plan for your destination", "Trip: YTO to SAN", "Date and time: Depart December 10, 2023; Return December 20, 2023", "Passengers: 1 adults, 0 children", and "Flight details: Round Trip". Below the screen is a blue input bar with a microphone icon and a placeholder "Enter your prompt here".

At
Snapdragon
Summit
2023

World's fastest
Llama 2-7B
on a phone

Up to 20 tokens per second

Demonstrating both chat and
application interaction on
device

World's first demonstration of
speculative decoding running
on a phone

Gracias

Prof. Jesús Alfonso López Sotelo
jalopez@uao.edu.co

UAO - Universidad Autónoma de Occidente, Cali,
Colombia www.uao.edu.co

