



DECSAI

Departamento de Ciencias de la Computación e I.A.

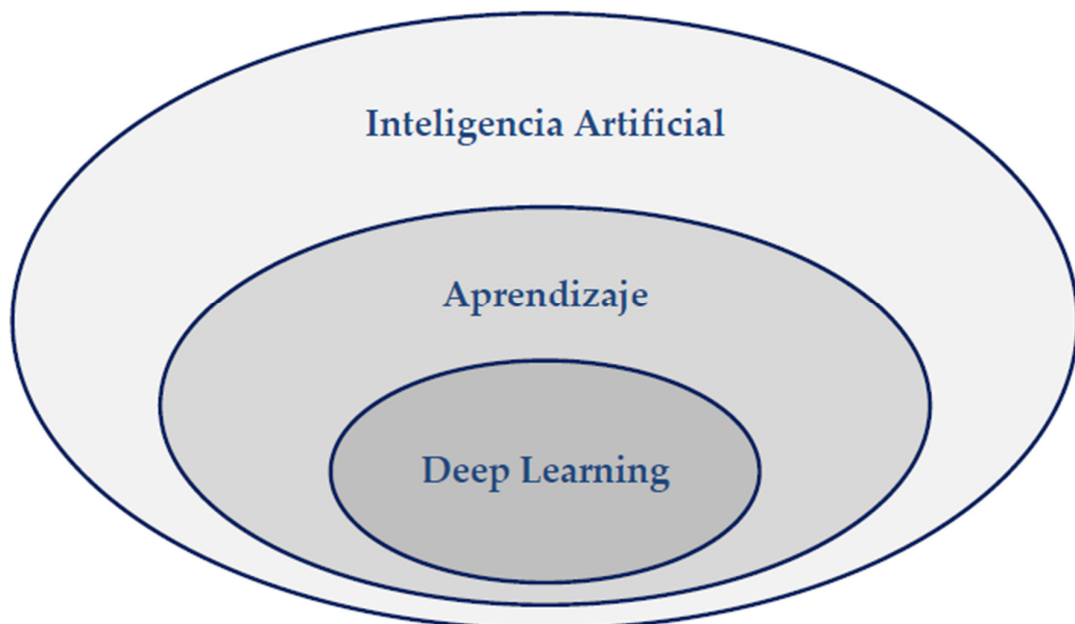
Universidad de Granada



Deep Learning

Fernando Berzal, berzal@acm.org

Deep Learning





¿Por qué estudiar redes neuronales?

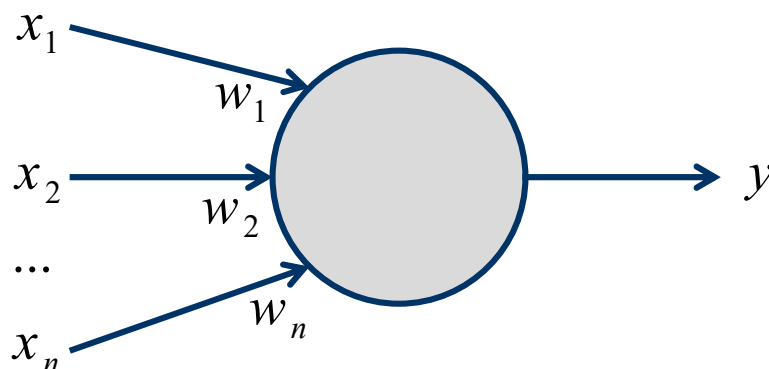
- Para comprender cómo funciona realmente el cerebro.
- Para diseñar un modelo de cómputo paralelo inspirado en las neuronas y sus sinapsis [conexiones] adaptativas.
- **Para resolver problemas prácticos utilizando algoritmos de aprendizaje inspirados en el cerebro.**

NOTA: Incluso aunque no sepamos realmente cómo funciona el cerebro, los algoritmos de aprendizaje nos serán muy útiles.



Neuronas

El modelo computacional más simple de una neurona



$$y = \sum_i x_i w_i = x_1 w_1 + x_2 w_2 + \dots + x_n w_n$$



Redes neuronales artificiales

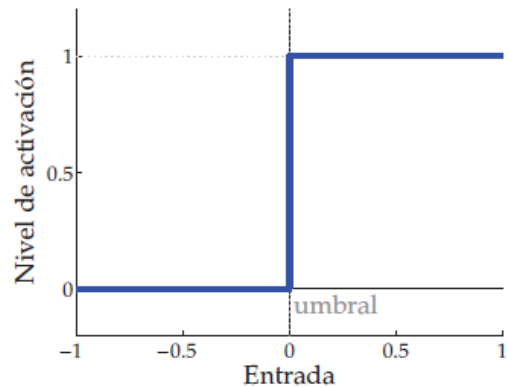


Modelo de neurona de McCulloch & Pitts

Neuronas binarias con umbral

$$z = \sum_i x_i w_i$$

$$y = \begin{cases} 1 & \text{si } z \geq 0 \\ 0 & \text{en otro caso} \end{cases}$$



1943

Warren McCulloch & Walter Pitts:
"A logical calculus of the ideas
immanent in nervous activity."
Bulletin of Mathematical Biophysics, 5:115-133.

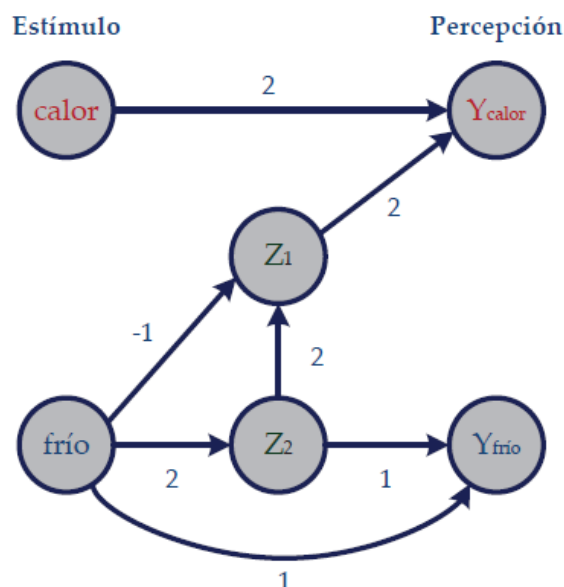


Redes neuronales artificiales



Modelo de neurona de McCulloch & Pitts

Ejemplo: Percepción fisiológica del calor y del frío

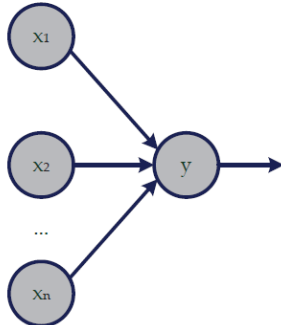


Historia de las redes neuronales artificiales

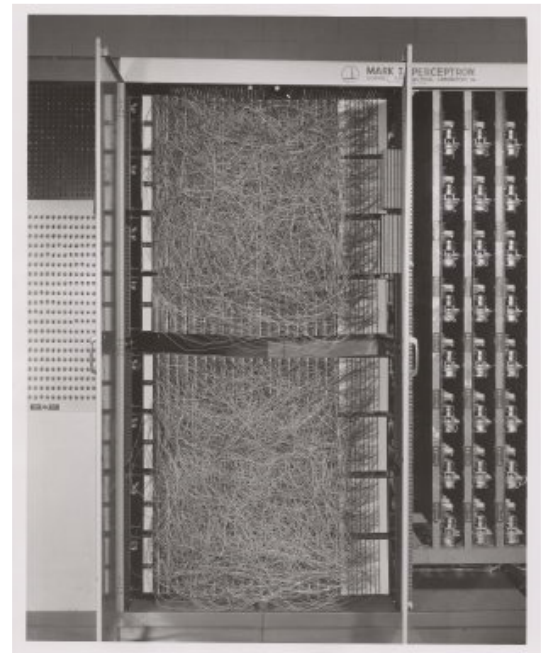
El perceptrón



Primer algoritmo de aprendizaje supervisado



Mark I Perceptron Machine
Primera implementación...



1957

Frank Rosenblatt:
"The Perceptron - A perceiving and recognizing automaton". Report 85-460-1, Cornell Aeronautical Laboratory, 1957.



Historia de las redes neuronales artificiales

El perceptrón



En la prensa...
New York Times

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

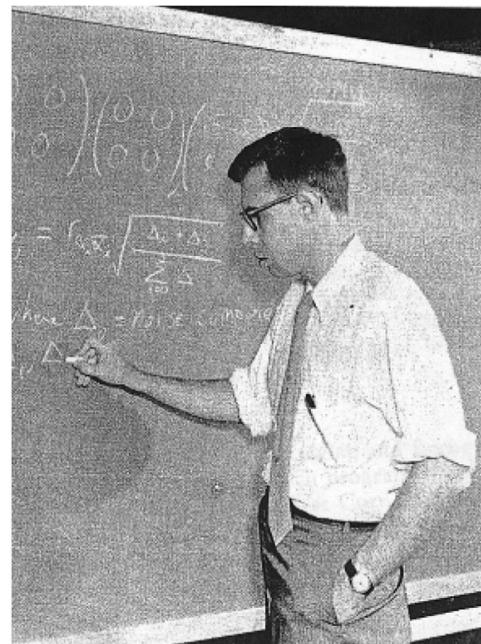
WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.



1958



Historia de las redes neuronales artificiales

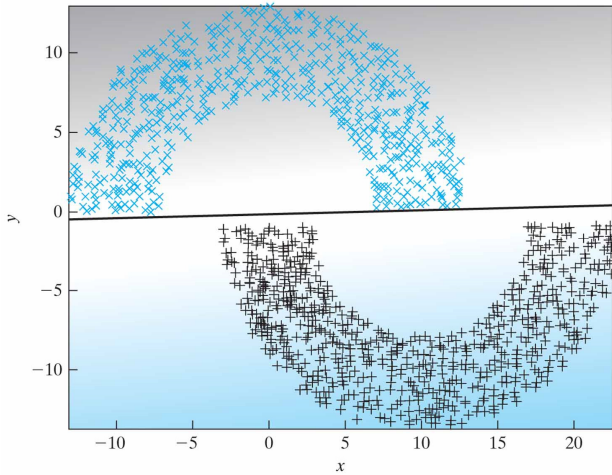
El perceptrón



En realidad...

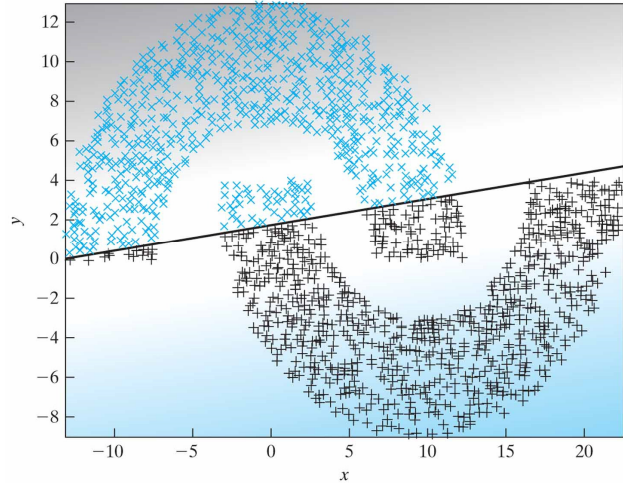
Un clasificador lineal

Classification using perceptron with distance = 1, radius = 10, and width = 6



(b) testing result

Classification using perceptron with distance = -4, radius = 10, and width = 6



(b) testing result

[Haykin: "Neural Networks and Learning Machines", 3rd edition]



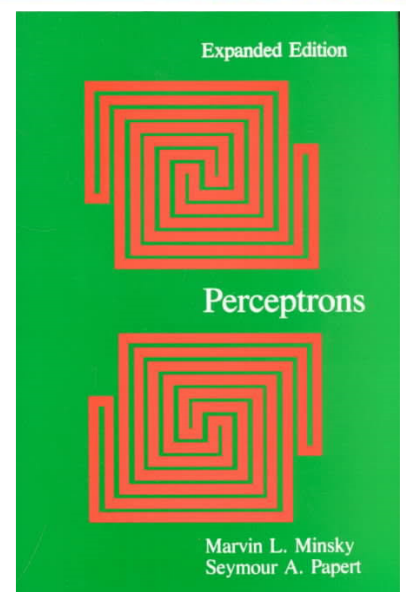
Historia de las redes neuronales artificiales

El perceptrón



Análisis de las capacidades y limitaciones del perceptrón:

- Muchos pensaron que esas limitaciones se extendían a todos los modelos de redes neuronales, aunque no es así.
- Abandono de los modelos conexionistas.
- La investigación en redes neuronales casi desaparece.



1969

Marvin Minsky & Seymour Papert:
"Perceptrons: An Introduction to Computational
Geometry". MIT Press, expanded edition, 1987
ISBN 0262631113

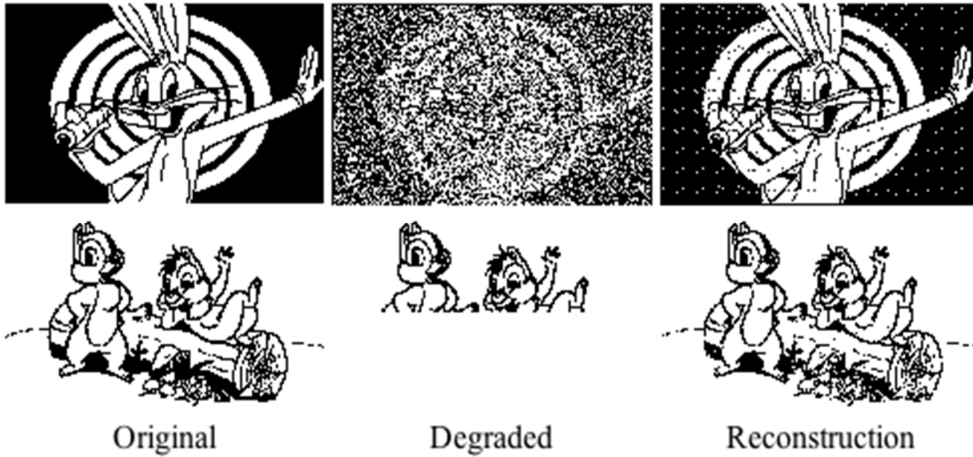


Historia de las redes neuronales artificiales

Redes de Hopfield



Redes recurrentes
que funcionan como memorias asociativas



Original

Degraded

Reconstruction

1982

John J. Hopfield:
"Neural networks and physical systems
with emergent collective computational abilities"
Proceedings of the National Academy of Sciences
PNAS 79(8):2554–2558, 1982

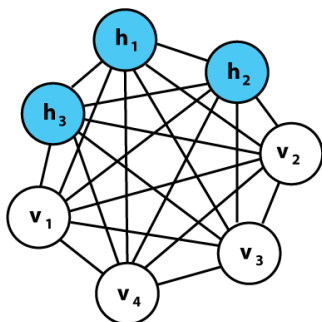


Historia de las redes neuronales artificiales

Máquinas de Boltzmann



Un contraejemplo:
Sí que se pueden
entrenar redes con
múltiples capas de
neuronas.



1985

David H. Ackley, Geoffrey E. Hinton & Terrence J. Sejnowski: "A Learning Algorithm for Boltzmann Machines", Cognitive Science 9(1):147–169, 1985. DOI 10.1207/s15516709cog0901_7

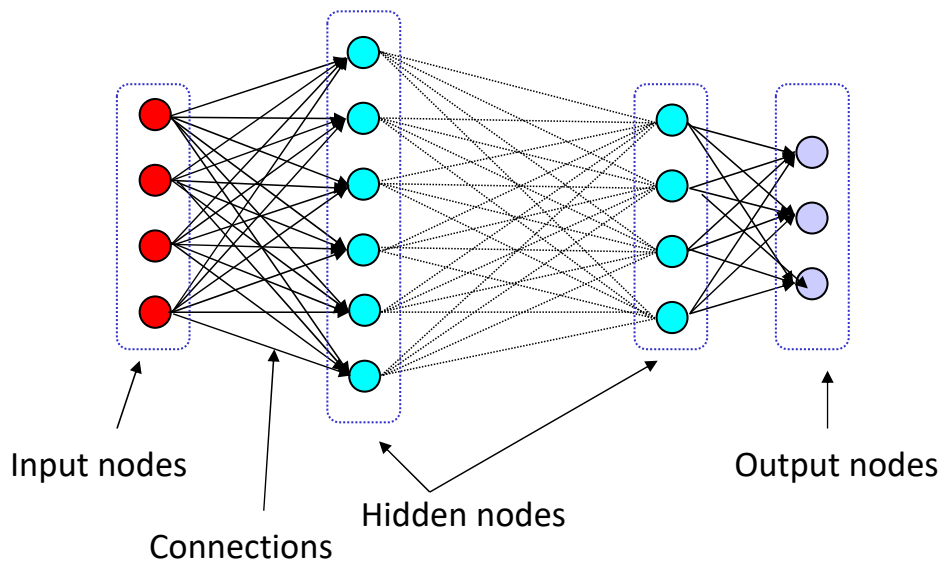


Historia de las redes neuronales artificiales

Backpropagation



Algoritmo de entrenamiento de redes multicapa



1986

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams: "Learning representations by back-propagating errors" *Nature* 323(6088):533–536, 1986. DOI 10.1038/323533a0



Historia de las redes neuronales artificiales

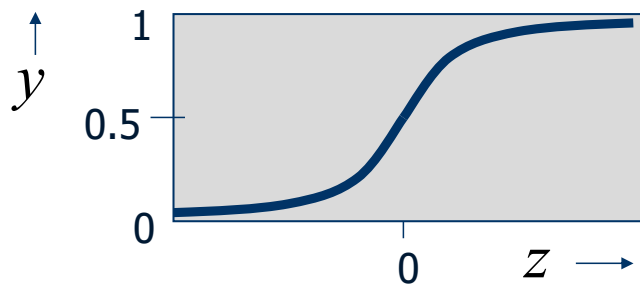
Backpropagation



Modelo de neurona sigmoideal

$$z = \sum_i x_i w_i$$

$$y = \frac{1}{1 + e^{-z}}$$



1986

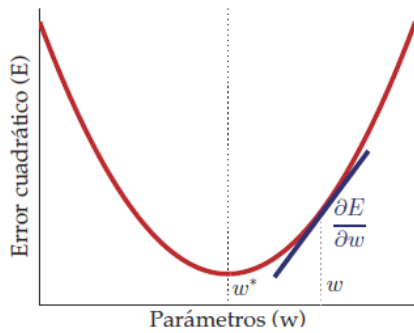


Historia de las redes neuronales artificiales

Backpropagation



Algoritmo de entrenamiento



$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

1986

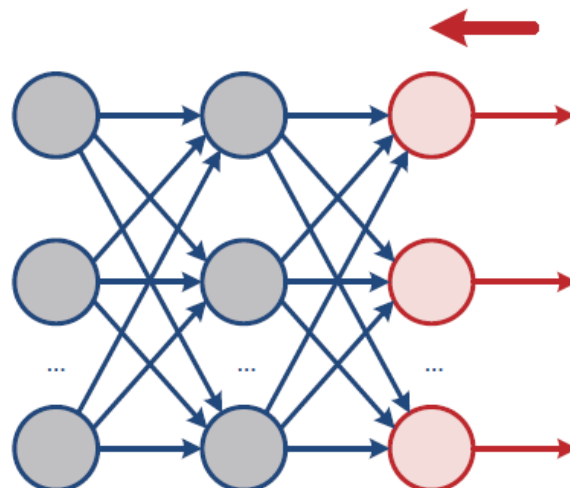


Historia de las redes neuronales artificiales

Backpropagation



Propagación de errores $\delta E / \delta y$

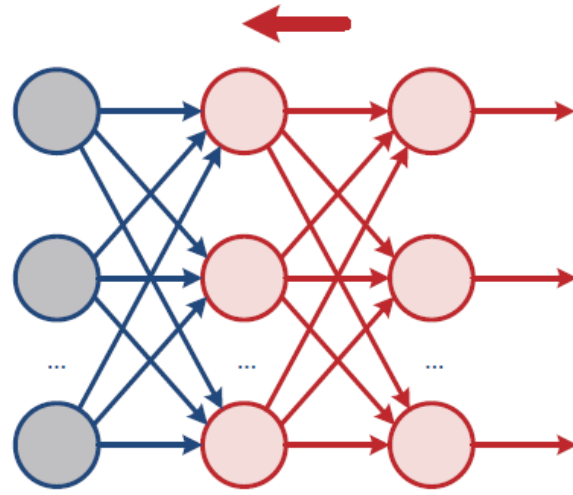


Historia de las redes neuronales artificiales

Backpropagation



Propagación de errores $\delta E/\delta y$

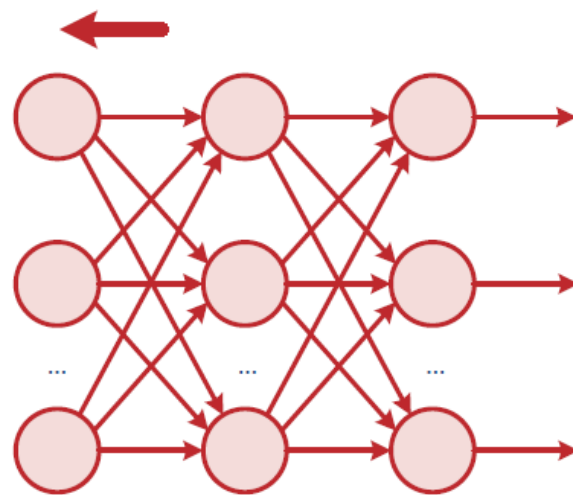


Historia de las redes neuronales artificiales

Backpropagation



Propagación de errores $\delta E/\delta y$



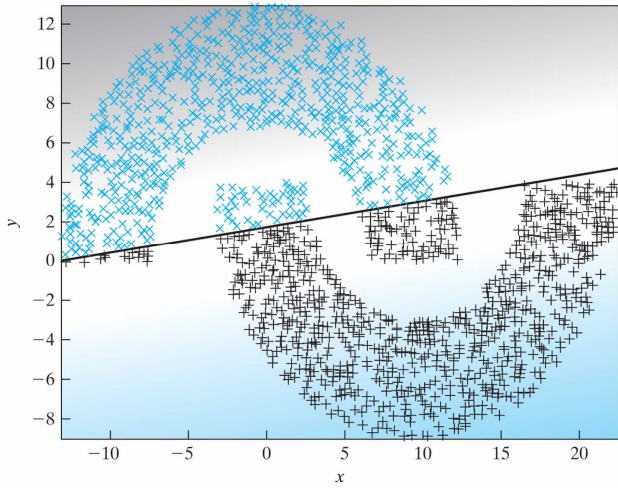
Historia de las redes neuronales artificiales

Backpropagation

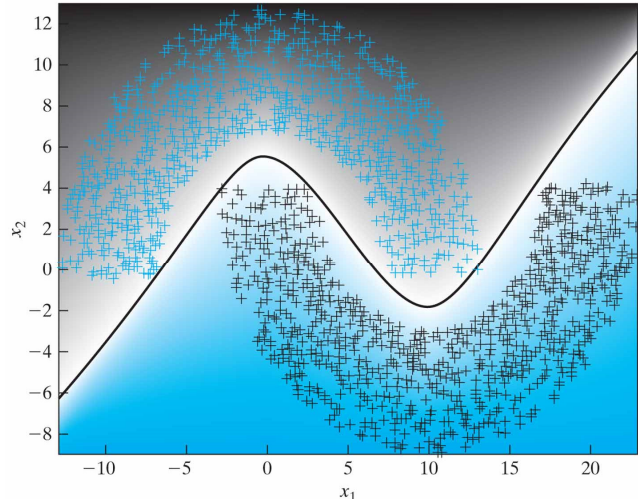


El resultado...

Classification using perceptron with distance = -4, radius = 10, and width = 6



(b) testing result



(b) Testing result

Perceptrón

Red multicapa



Historia de las redes neuronales artificiales

Backpropagation



Algoritmo redescubierto en múltiples ocasiones...

■ Sistemas de control (años 60)

Arthur E. Bryson, W.F. Denham & S.E. Dreyfus: "Optimal programming problems with inequality constraints. I: Necessary conditions for extremal solutions." AIAA J. 1(11):2544-2550, **1963**.

Arthur E. Bryson & Yu-Chi Ho: "Applied optimal control: optimization, estimation, and control." Blaisdell Publishing Company / Xerox College Publishing, p. 481, **1969**.

■ Diferenciación automática (años 70)

Seppo Linnainmaa: The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's Thesis (in Finnish), University of Helsinki, 6-7, **1970**.

Seppo Linnainmaa: "Taylor expansion of the accumulated rounding error". BIT Numerical Mathematics. 16(2):146-160, **1976**. DOI 10.1007/bf01931367.

1986???



Historia de las redes neuronales artificiales

Backpropagation

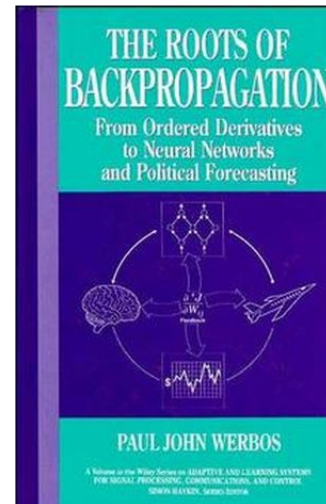


Algoritmo redescubierto en múltiples ocasiones...

■ Redes neuronales (1974!!!)

Paul John Werbos: "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." PhD thesis, Harvard University, **1974**.

Paul John Werbos:
"The Roots of Backpropagation:
From Ordered Derivatives
to Neural Networks and Political Forecasting."
John Wiley & Sons, Inc., 1994.
ISBN 0471598976



1986???

Historia de las redes neuronales artificiales

Backpropagation

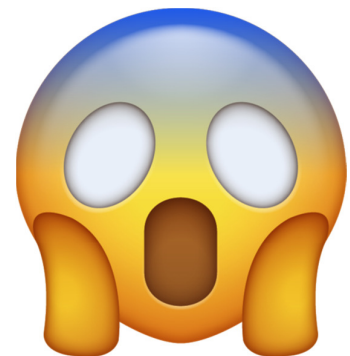


Política & Publicaciones

Referencias bibliográficas del artículo de Nature

Learning representations by back-propagating errors

David E. Rumelhart*, **Geoffrey E. Hinton†**
& **Ronald J. Williams***



Received 1 May; accepted 31 July 1986.

1. Rosenblatt, F. *Principles of Neurodynamics* (Spartan, Washington, DC, 1961).
2. Minsky, M. L. & Papert, S. *Perceptrons* (MIT, Cambridge, 1969).
3. Le Cun, Y. *Proc. Cognitiva* **85**, 599-604 (1985).
4. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1: *Foundations* (eds Rumelhart, D. E. & McClelland, J. L.) 318-362 (MIT, Cambridge, 1986).



Historia de las redes neuronales artificiales

Backpropagation



Política & Publicaciones

Geoffrey Hinton interview

Neural Networks & Deep Learning



"... we managed to get a paper into Nature in 1986. And **I did quite a lot of political work to get the paper accepted**. I figured out that one of the referees was probably going to be Stuart Sutherland, who was a well known psychologist in Britain. And I went to talk to him for a long time, and explained to him exactly what was going on. And he was very impressed by the fact that we showed that backprop could learn representations for words..."



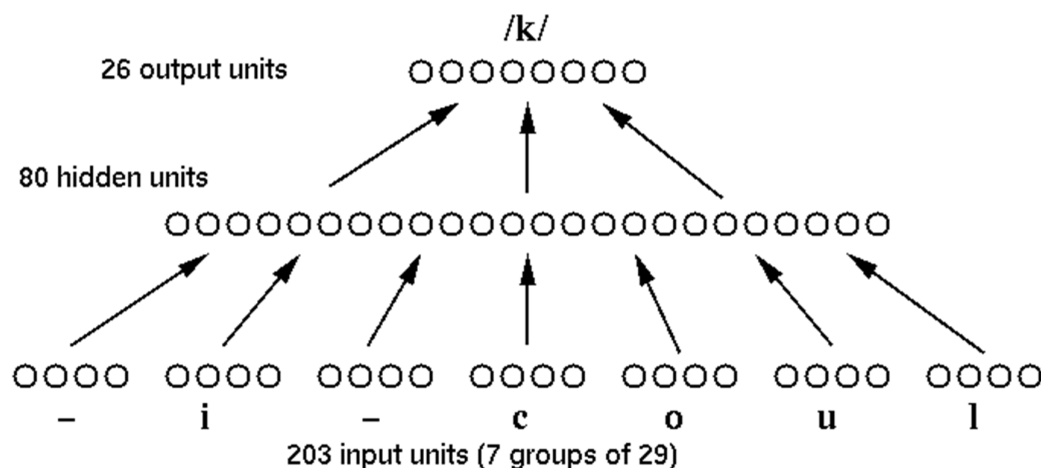
Historia de las redes neuronales artificiales

Backpropagation



NETTalk

Síntesis de voz



1986

Terrence J. Sejnowski & Charles Rosenberg:
"NETtalk: a parallel network that learns to read
aloud," Cognitive Science, 14, 179-211, 1986.



Historia de las redes neuronales artificiales

Redes convolutivas



The MNIST database of handwritten digits

<http://yann.lecun.com/exdb/mnist/>



1990s



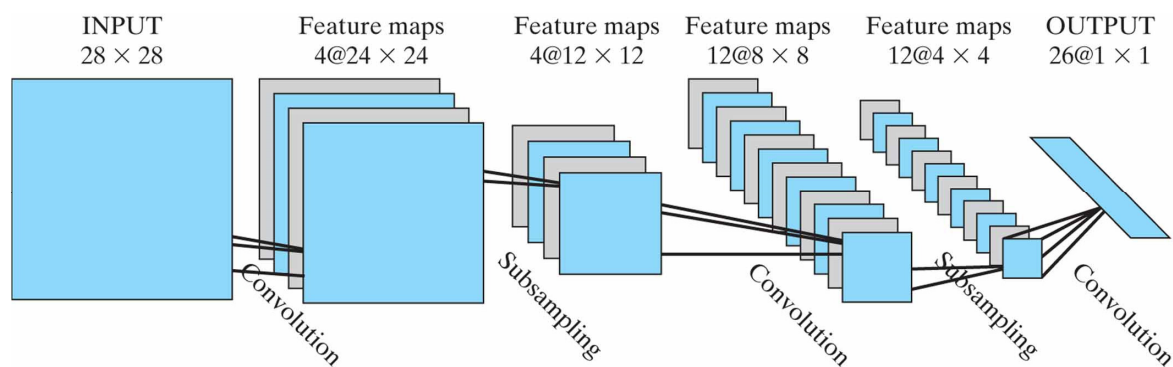
Historia de las redes neuronales artificiales

Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>



1990s



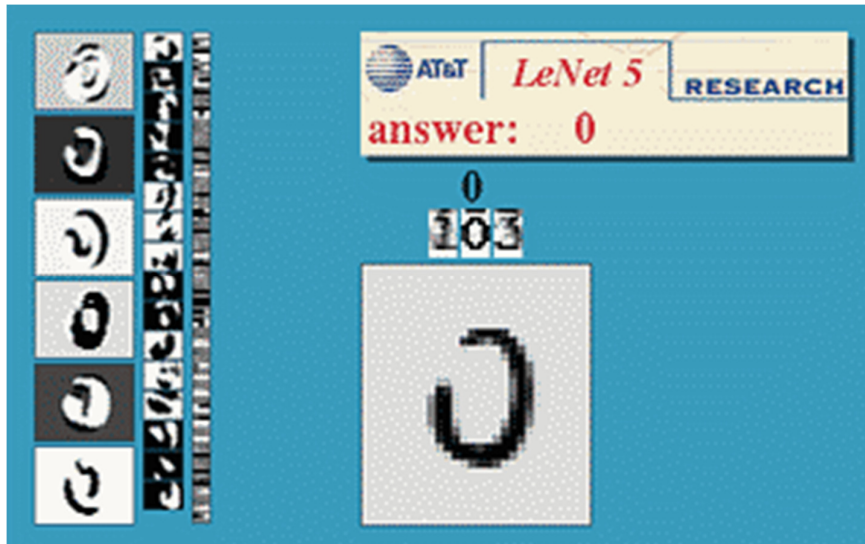
Historia de las redes neuronales artificiales

Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>



Historia de las redes neuronales artificiales

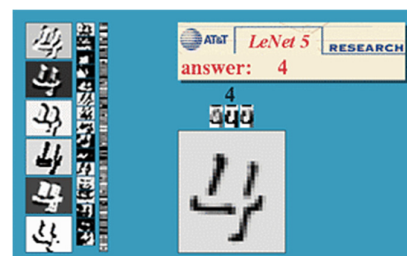
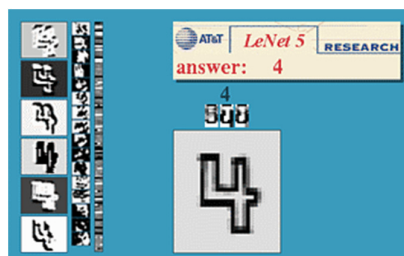
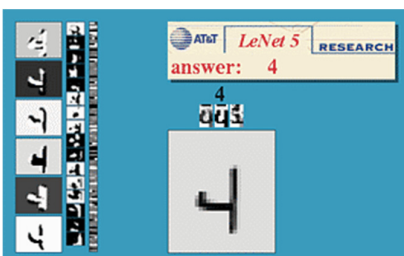
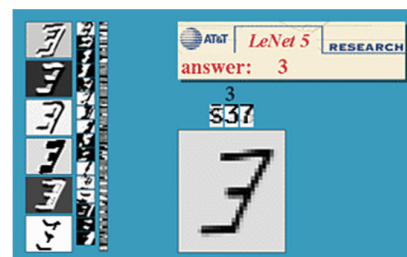
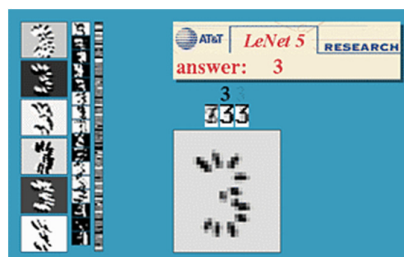
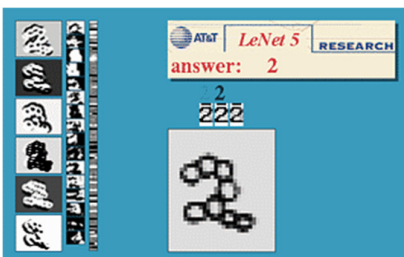
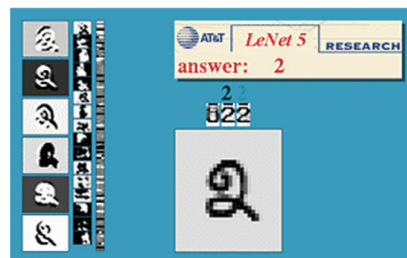
Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>

Ejemplos



Historia de las redes neuronales artificiales

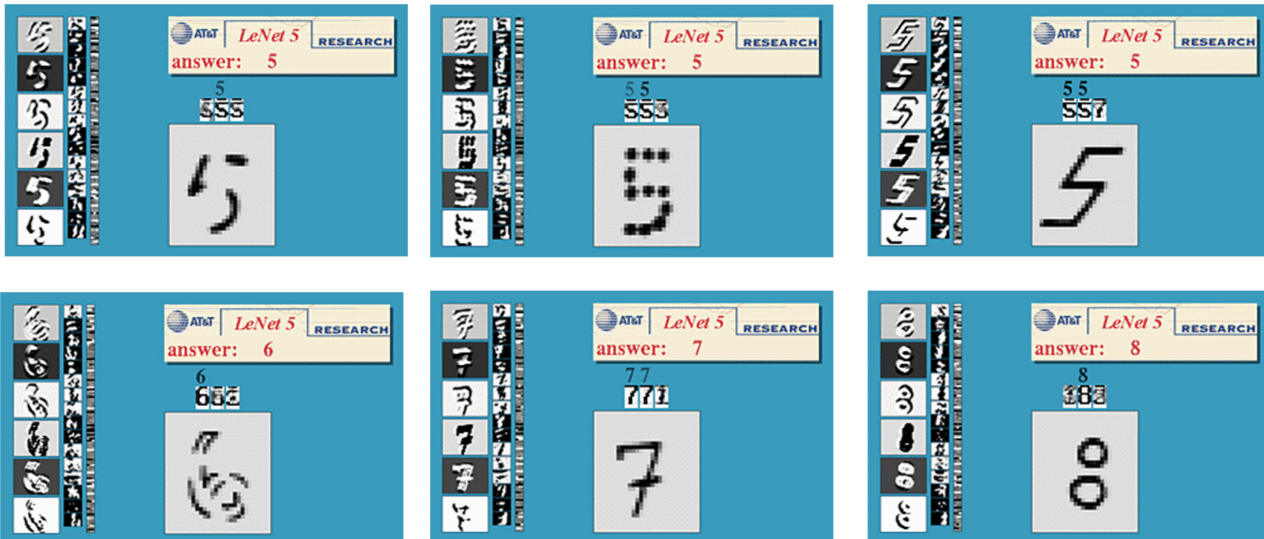
Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>

Ejemplos



Historia de las redes neuronales artificiales

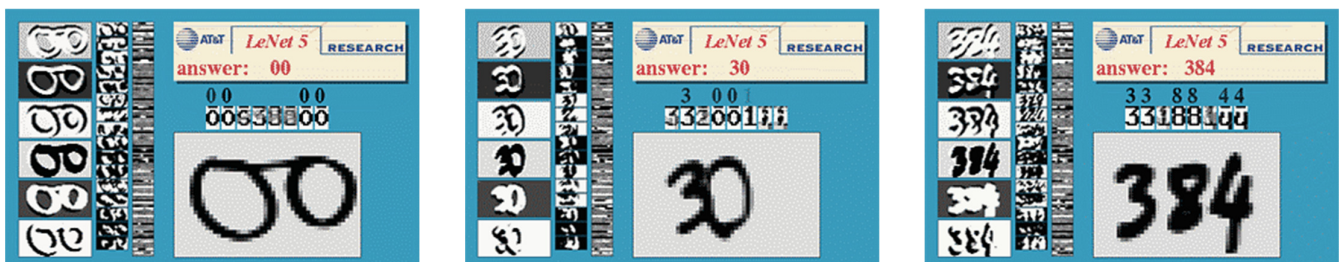
Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>

Variaciones en los datos de entrada



Historia de las redes neuronales artificiales

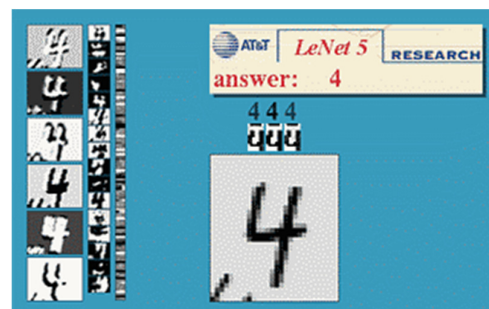
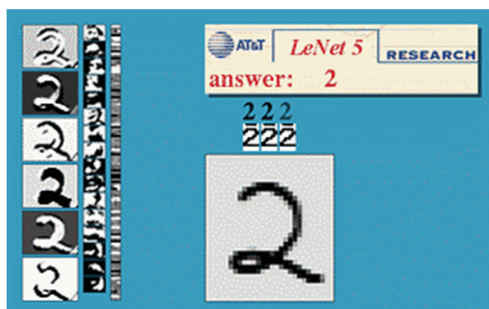
Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>

Robustez frente a la presencia de ruido en la imagen...



Historia de las redes neuronales artificiales

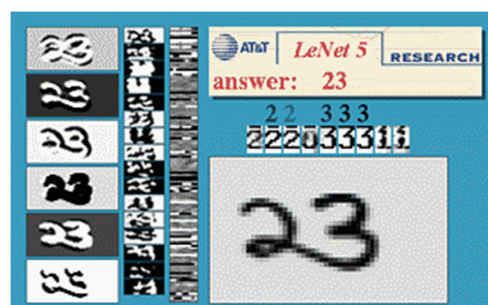
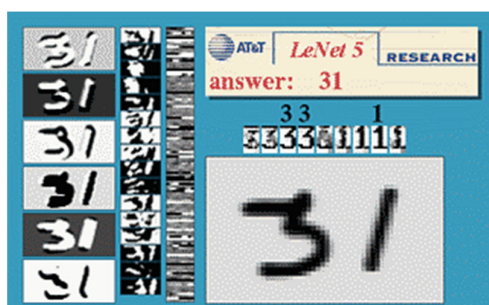
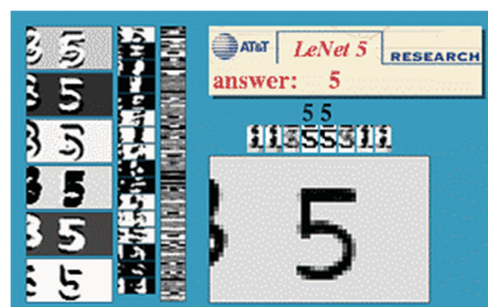
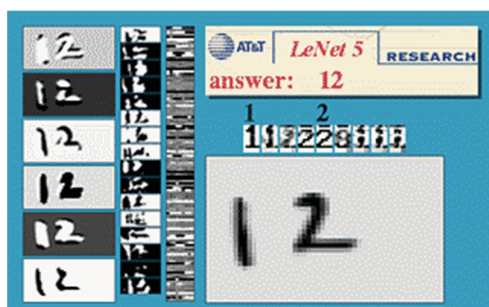
Redes convolutivas



LeNet

<http://yann.lecun.com/exdb/lenet/>

Casos curiosos





Una apuesta...

AT&T Adaptive Systems Research Dept., Bell Labs

1. Jackel bets (one fancy dinner) that by March 14, 2000, people will understand quantitatively why big neural nets working on large databases are not so bad. (Understanding means that there will be clear conditions and bounds)

Vapnik bets (one fancy dinner) that Jackel is wrong.

But .. If Vapnik figures out the bounds and conditions, Vapnik still wins the bet.

2. Vapnik bets (one fancy dinner) that by March 14, 2005, no one in his right mind will use neural nets that are essentially like those used in 1995.

Jackel bets (one fancy dinner) that Vapnik is wrong

1995



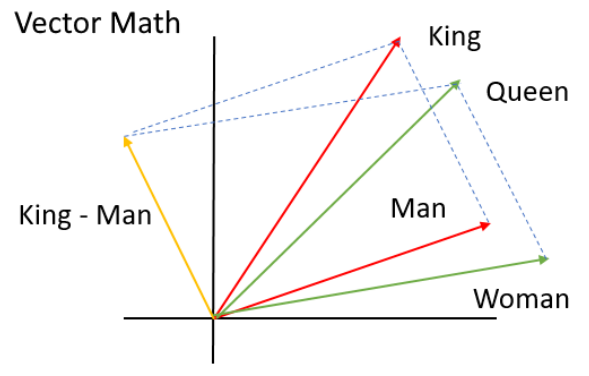
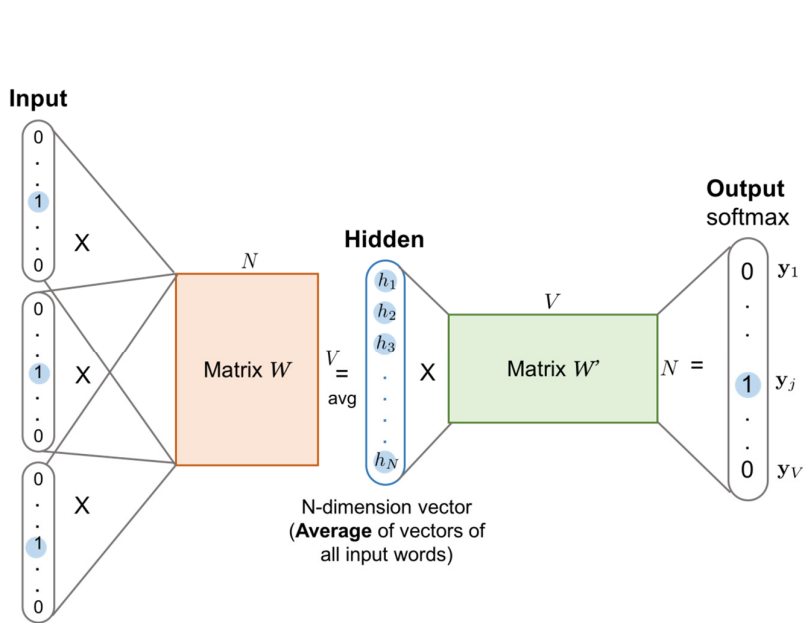
¿Por qué las SVMs nunca fueron una buena opción en IA?
Sólo son una reencarnación de los perceptrones...

- Expanden la entrada a una capa (enorme) de características **no adaptativas**.
- Sólo tienen una capa de pesos **adaptativos**.
- Disponen de un algoritmo eficiente para ajustar los pesos controlando el sobreaprendizaje (una forma inteligente de seleccionar características y encontrar los pesos adecuados).



Historia de las redes neuronales artificiales

Word embeddings



2000

Yoshua Bengio, Réjean Ducharme, Pascal Vincent
 "A neural probabilistic language model."
 NIPS 2000: 932-938
 JMLR 3:1137-1155, 2003

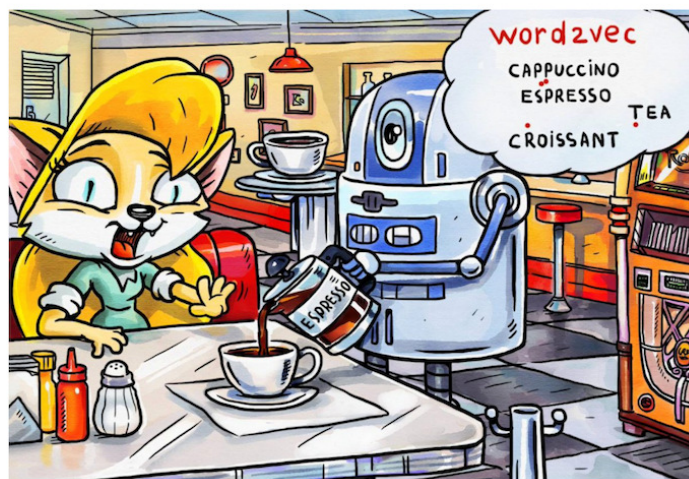
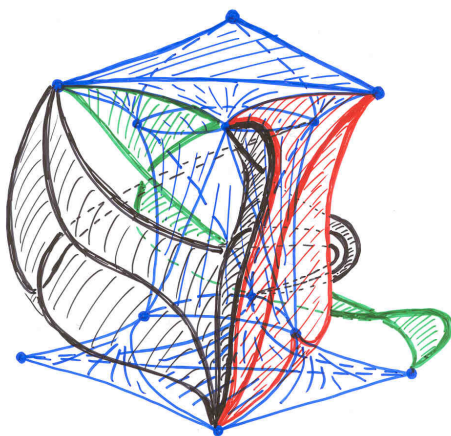


Historia de las redes neuronales artificiales

Word embeddings



word2vec



- Espresso? But I ordered a cappuccino!
 - Don't worry, the cosine distance between them is so small that they are almost the same thing.

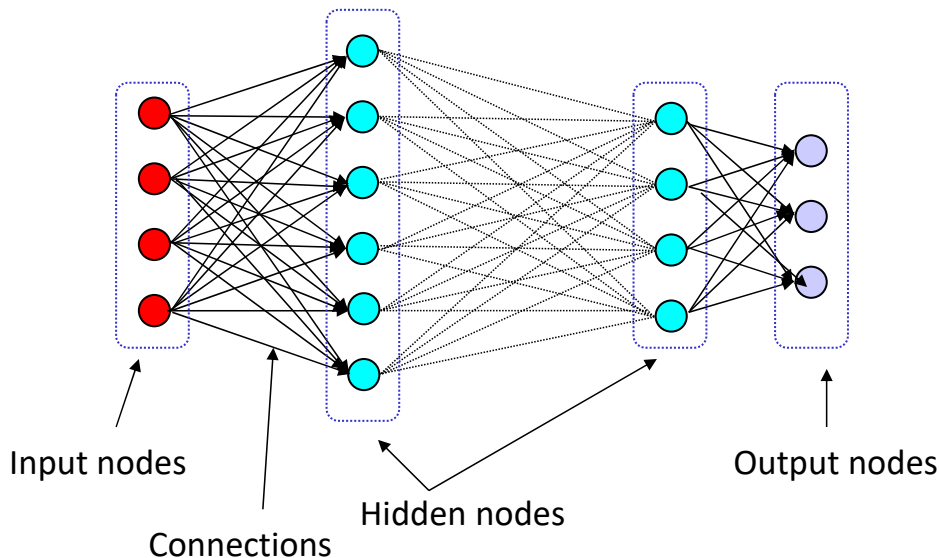
2000s

Tomas Mikolov et al.: "Efficient Estimation of Word Representations in Vector Space"
 arXiv:1301.3781, 2013





Backpropagation no funcionaba bien con redes que tengan varias capas ocultas (salvo en el caso de las redes convolutivas)...



Algunos hechos hicieron que backpropagation no tuviera éxito en tareas en las que luego se ha demostrado útil:

- Capacidad de cálculo limitada.
- Disponibilidad de conjuntos de datos etiquetados.
- “Deep networks” demasiado pequeñas (e inicializadas de forma poco razonable).





2006: The Deep Breakthrough



- Hinton, Osindero & Teh « A Fast Learning Algorithm for Deep Belief Nets », *Neural Computation*, 2006
- Bengio, Lamblin, Popovici, Larochelle « Greedy Layer-Wise Training of Deep Networks », *NIPS'2006*
- Ranzato, Poultney, Chopra, LeCun « Efficient Learning of Sparse Representations with an Energy-Based Model », *NIPS'2006*

2006

[Yoshua Bengio]



Geoffrey Hinton
(University of Toronto & Google)



Yann LeCun
(AT&T Labs → NYU → Facebook)



Joshua Bengio
(University of Montréal & IBM Watson)



2018



Historia de las redes neuronales artificiales

Deep Learning



Estadística	Inteligencia Artificial
Dimensionalidad baja (<100)	Dimensionalidad alta (>>100)
Mucho ruido en los datos	El ruido no es el mayor problema
Sin demasiada estructura en los datos (puede capturarse usando modelos simples)	Mucha estructura en los datos (demasiado complicada para modelos simples)
PRINCIPAL PROBLEMA	PRINCIPAL PROBLEMA
Separar estructura de ruido	Descubrir una forma de representar la estructura que se pueda aprender
TÉCNICAS	TÉCNICAS
SVM [Support Vector Machines]	Backpropagation

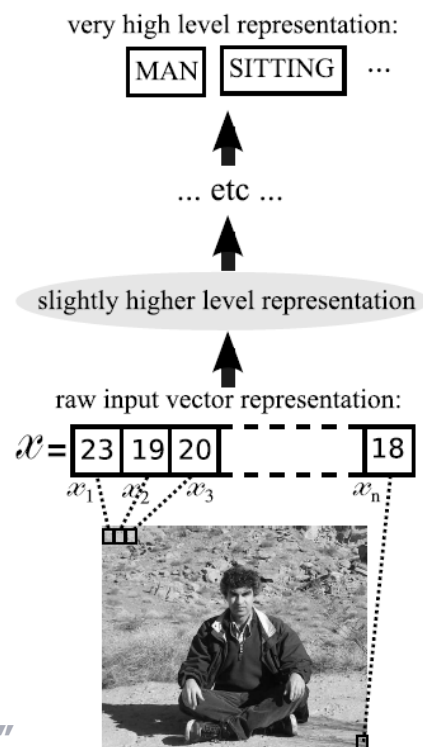


Historia de las redes neuronales artificiales

Deep Learning



Motivación



Yoshua Bengio
 "Learning Deep Architectures for AI"
 2009

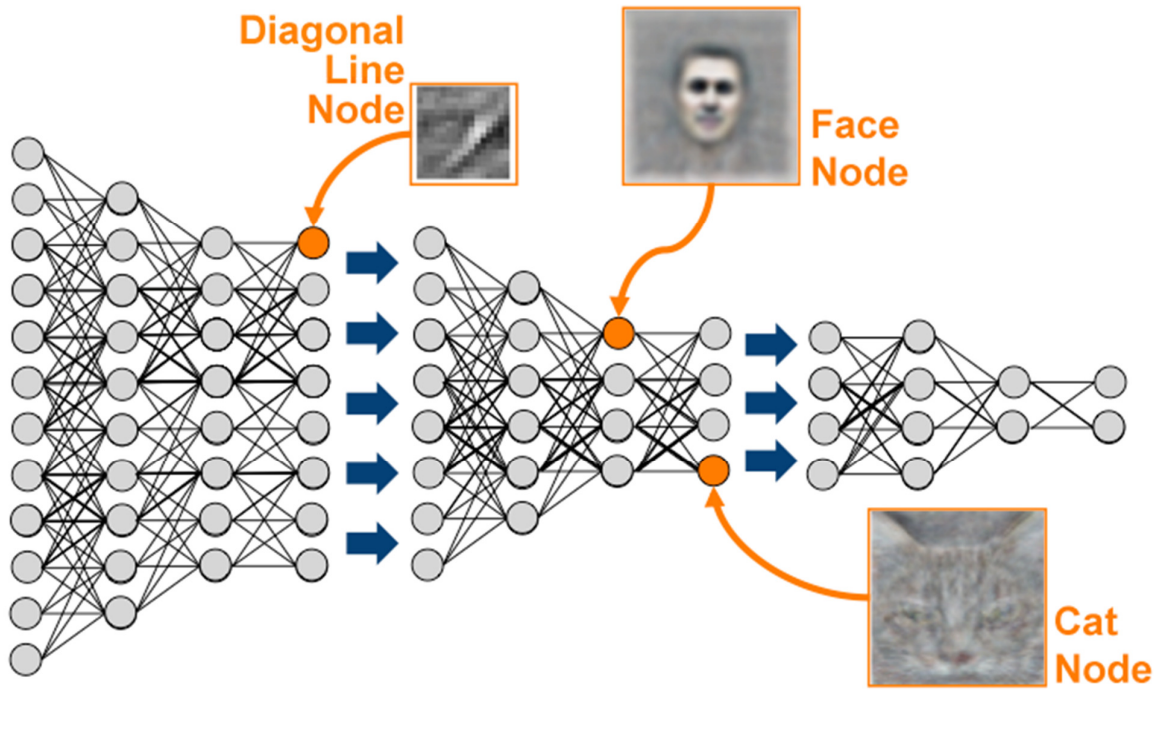


Historia de las redes neuronales artificiales

Deep Learning



Deep Learning as hierarchical feature representation

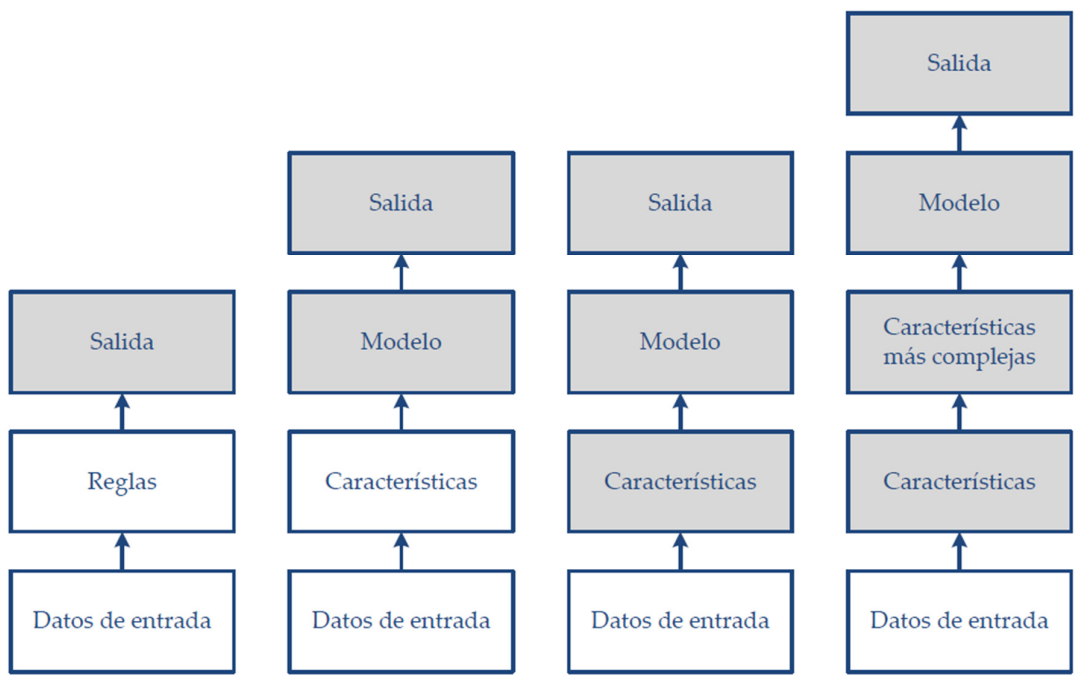


Historia de las redes neuronales artificiales

Deep Learning



Deep Learning as hierarchical feature representation



Sistemas expertos

Aprendizaje automático

Aprendizaje de representaciones

Deep Learning





¿Cuál era el problema de backpropagation?

- Requiere datos etiquetados, pero casi todos los datos disponibles no lo están.
- No resulta demasiado escalable: Demasiado lento en redes con múltiples capas ocultas.
- Se puede quedar atascado en óptimos locales (¿lejos de ser óptimos en “deep networks”?).



Política & Publicaciones

Yann LeCun @ CVPR'2012



... the reviews [are] so ridiculous, that I don't know how to begin writing a rebuttal without insulting the reviewers ... This time though, the reviewers were particularly clueless, or negatively biased, or both. I was very sure that this paper was going to get good reviews because: 1) it has two simple and generally applicable ideas for segmentation... 2) it uses no hand-crafted features... 3) it beats all published results on 3 standard datasets for scene parsing; 4) it's an order of magnitude faster than the competing methods.

If that is not enough to get good reviews, I just don't know what is.”



Historia de las redes neuronales artificiales

Deep Learning



IMAGENET

Large Scale Visual Recognition Challenge

Reconocimiento de objetos reales en imágenes



2012



Historia de las redes neuronales artificiales

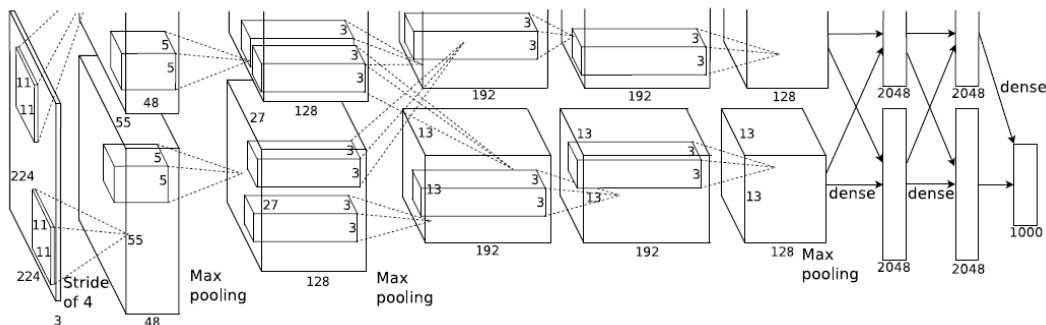
Deep Learning



IMAGENET

AlexNet

Red neuronal diseñada por Alex Krizhevsky (NIPS 2012)



2012

Tasa de error

Clasificación de imágenes

16.4% vs. 25% (2010)



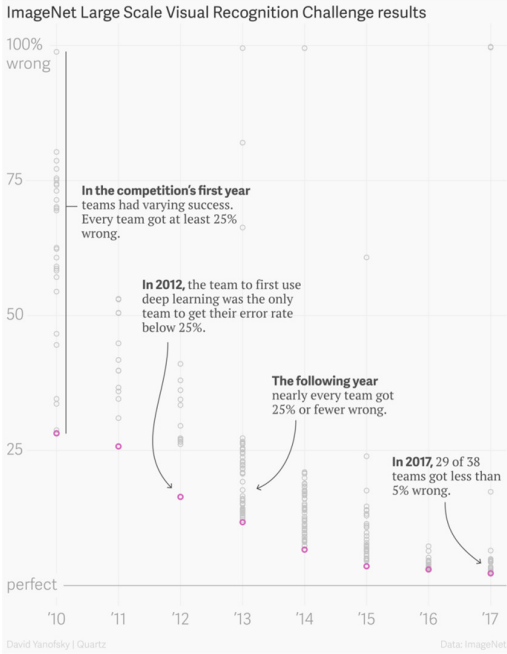
Historia de las redes neuronales artificiales

Deep Learning



IMAGENET

Large Scale Visual Recognition Challenge



Tasa de error

16.4% Alex Krizhevsky @ NIPS 2012

6.66% GoogLeNet @ ILSVRC'2014

4.94% PreLU-nets (MSR) @ 2015

"Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification"

arXiv, 2015, <http://arxiv.org/pdf/1502.01852v1.pdf>

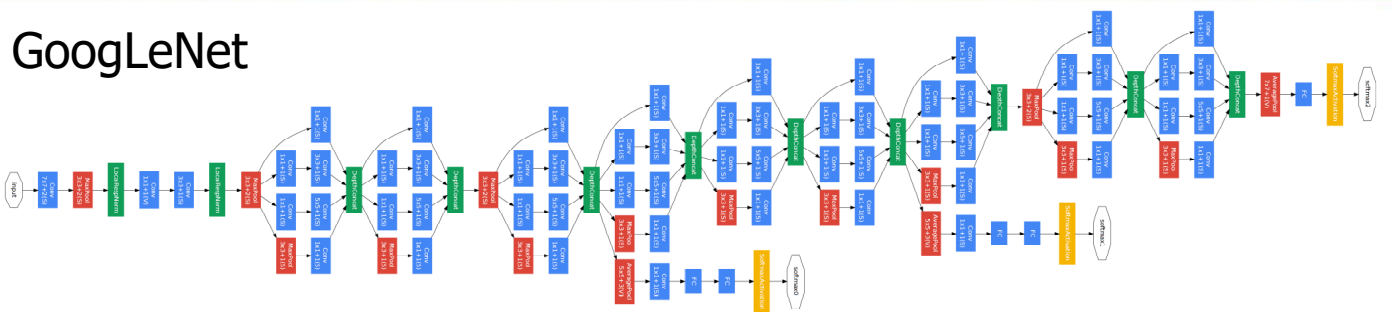


Historia de las redes neuronales artificiales

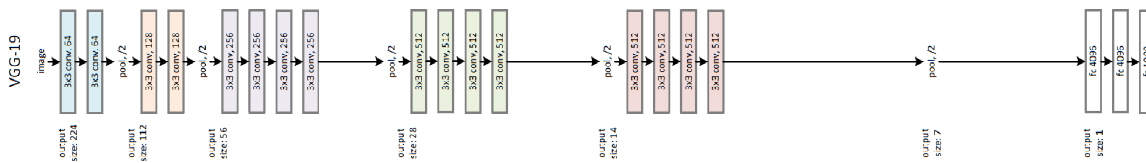
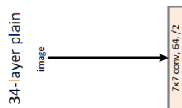
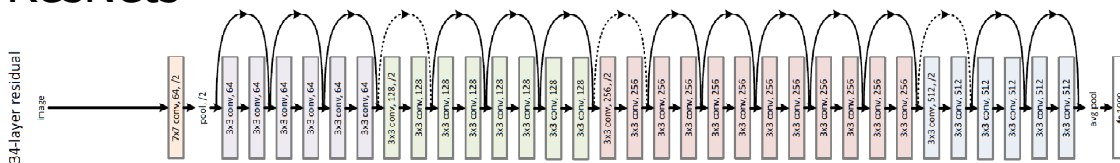
Deep Learning



GoogLeNet



ResNets



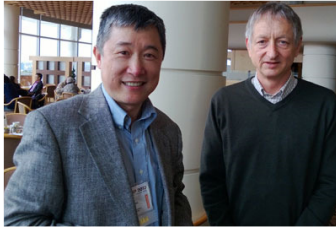
Historia de las redes neuronales artificiales

Deep Learning



Reconocimiento de voz

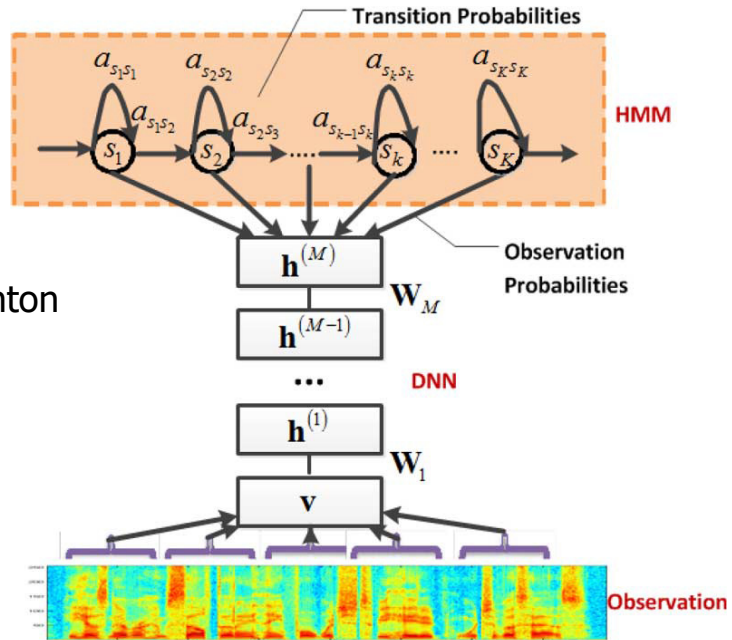
Microsoft
Research



Li Deng (MSR) & Geoff Hinton



Dong Yu (MSR)



Historia de las redes neuronales artificiales

Deep Learning



Reconocimiento de voz

Task	Hours of training data	Deep Neural Network	Gaussian Mixture Model	GMM with more data
Switchboard (Microsoft Research)	309	18.5%	27.4%	18.6% (2000 hrs)
English broadcast news (IBM)	50	17.5%	18.8%	
Google Voice Search (Android 4.1)	5,870	12.3% (and falling)		16.0% (>>5,870 hrs)

Microsoft
Research



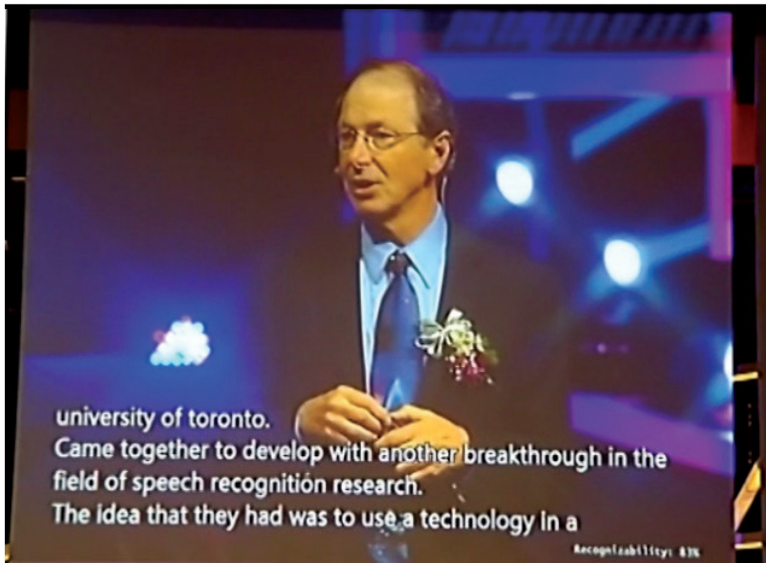
2012

Geoffrey Hinton, Li Deng, Dong Yu et al.:
"Deep Neural Networks for Acoustic Modeling in Speech Recognition"
IEEE Signal Processing Magazine, 2012





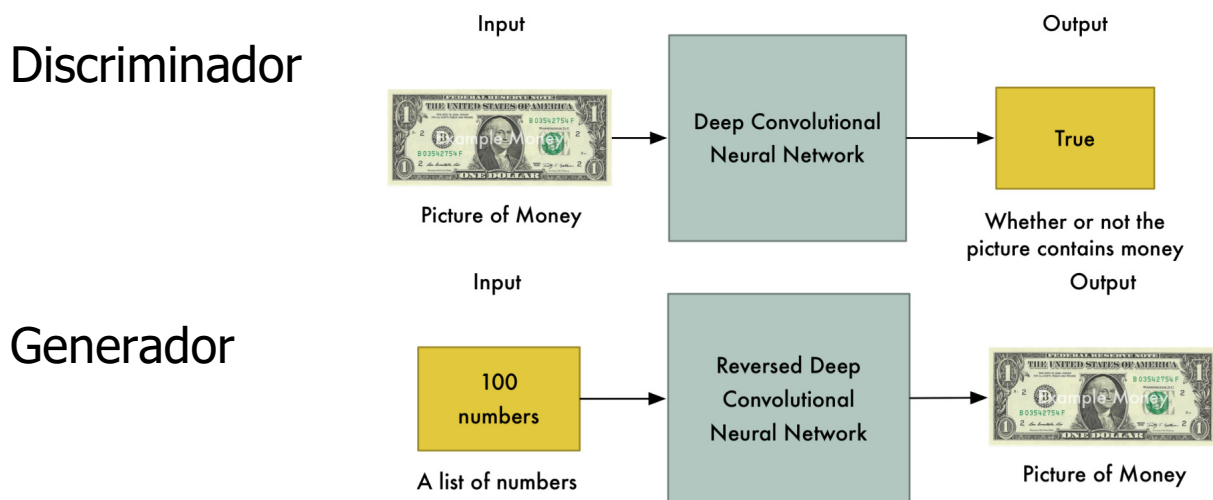
Traducción simultánea



2012



GANs [Generative Adversarial Networks]

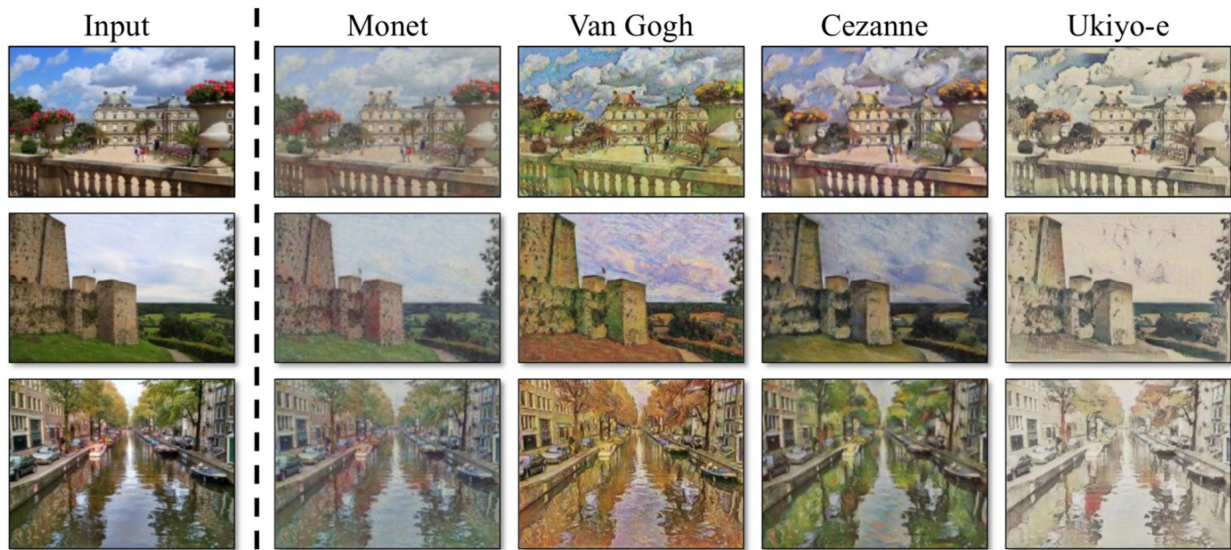


2014





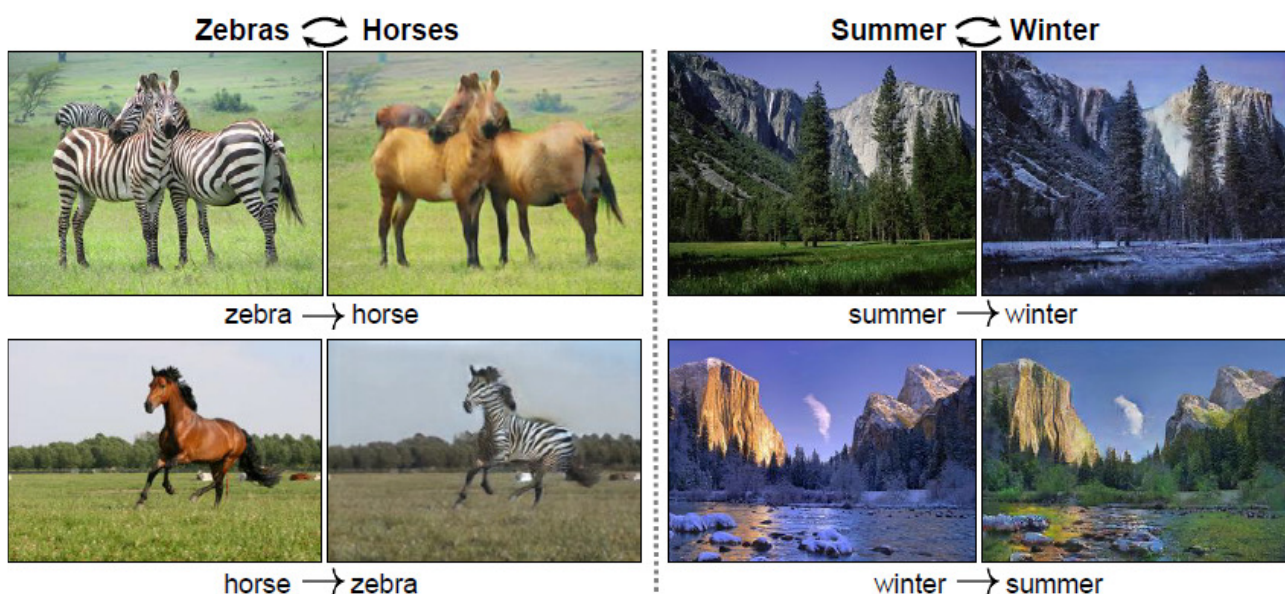
“Traducción de imágenes”



CycleGAN Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ICCV'2017



“Traducción de imágenes”



CycleGAN Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ICCV'2017



Historia de las redes neuronales artificiales

Deep Learning



“Traducción de imágenes”



CycleGAN Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ICCV'2017



Historia de las redes neuronales artificiales

Deep Learning



Unsupervised Image-to-Image Translation Networks, NIPS'2017



Historia de las redes neuronales artificiales

Deep Learning



Síntesis de imágenes

<https://thispersondoesnotexist.com/>



StyleGAN <https://arxiv.org/abs/1812.04948> CVPR'2019



Historia de las redes neuronales artificiales

Deep Learning



GauGAN



NVIDIA'2019

<https://blogs.nvidia.com/blog/2019/03/18/gaugan-photorealistic-landscapes-nvidia-research/>



Historia de las redes neuronales artificiales

Deep Learning



“You sketch, the AI paints”



GauGAN, NVIDIA, CVPR'2019

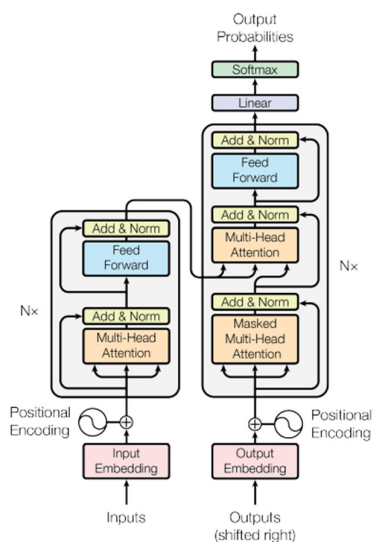


Historia de las redes neuronales artificiales

Deep Learning



Transformers: “Attention is All You Need”



Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL



Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



2017

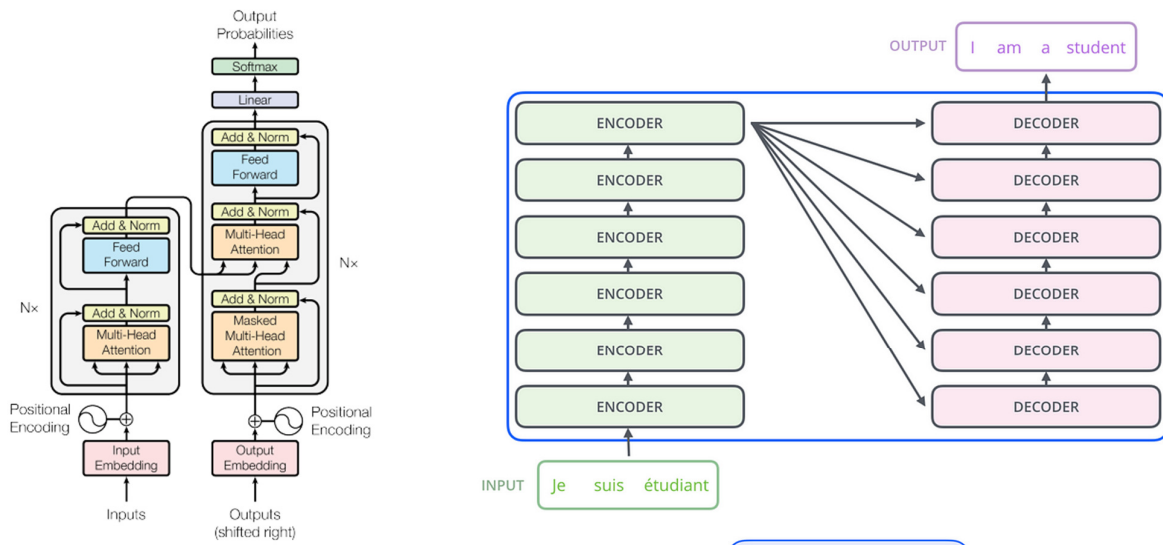


Historia de las redes neuronales artificiales

Deep Learning



Transformers: "Attention is All You Need"



2017



Historia de las redes neuronales artificiales

Deep Learning



DALL-E

Versión modificada de GPT-3 para generar imágenes de 256x256 a partir de descripciones textuales [12B parámetros]



vibrant portrait painting of Salvador Dalí with a robotic half face

DALL-E 2 (2022) 1024x1024
@ Bing Creator
<https://www.bing.com/create>



a dolohin in an astronaut suit on saturn. artstation

2021



Historia de las redes neuronales artificiales

Deep Learning

Midjourney

<https://www.midjourney.com/>



2022



Historia de las redes neuronales artificiales

Deep Learning

ChatGPT

GPT-3.5

175B parámetros

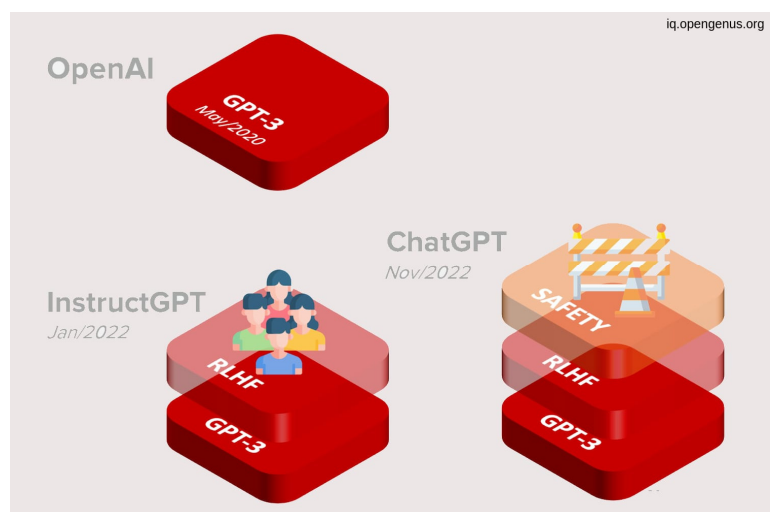
+

RLHF

Reinforcement Learning
from Human Feedback

+

Safety features



2022





GPT-4

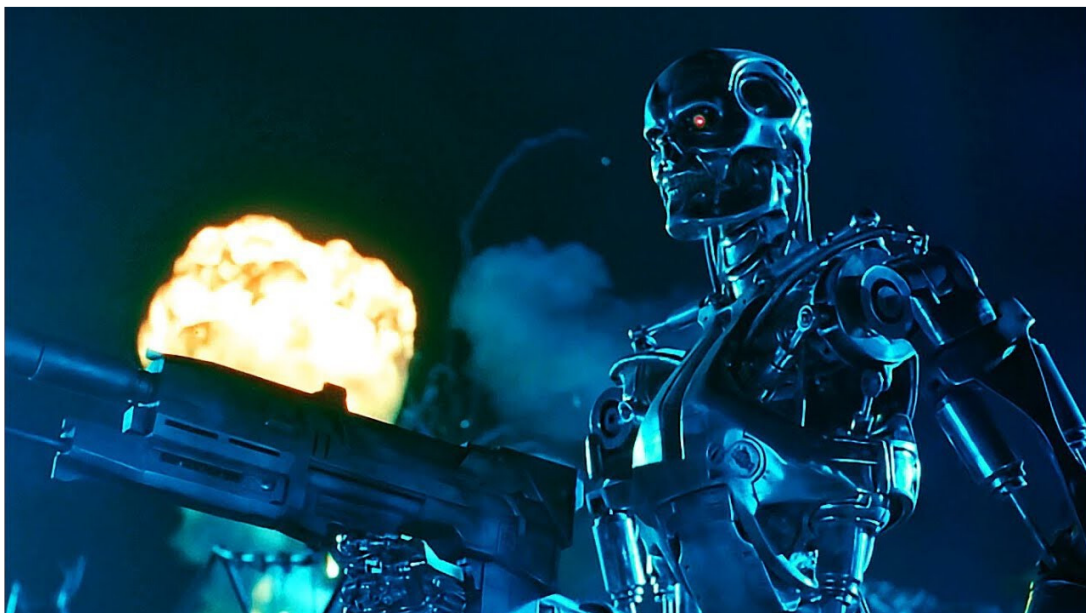
GPT-4 Technical Report

OpenAI*

2 Scope and Limitations of this Technical Report

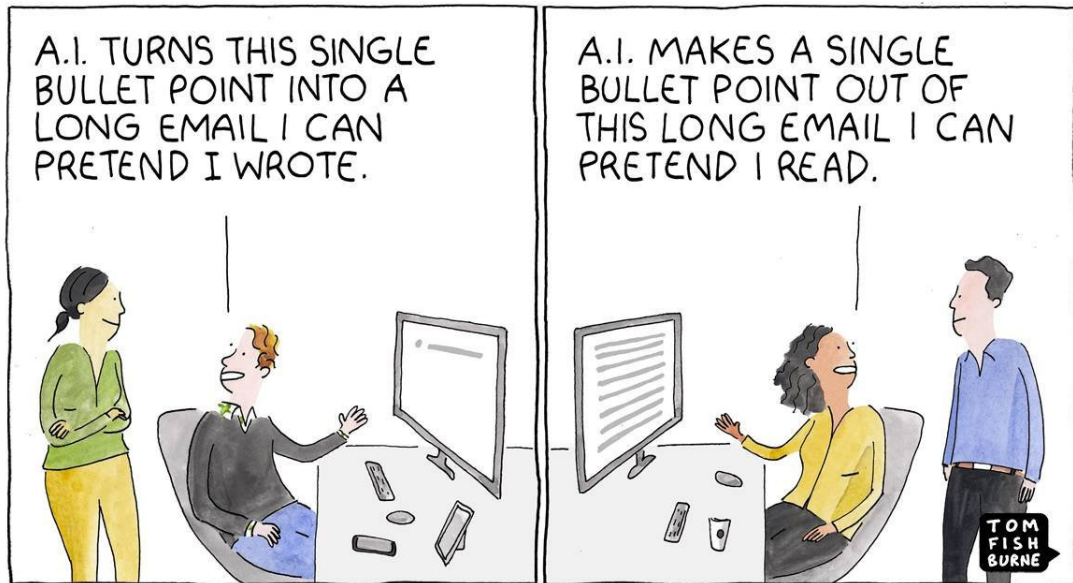
This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

2023



Historia de las redes neuronales artificiales

Deep Learning



Aplicaciones del Deep Learning



Existen problemas para los que es extremadamente difícil desarrollar manualmente un programa de ordenador que los resuelva.

Ejemplo: Visión artificial



[Terminator, 1984]



Aplicaciones del Deep Learning

De hecho, el reconocimiento de objetos...

- Ni siquiera sabemos cómo se hace realmente en nuestro cerebro (por lo que difícilmente podremos diseñar un algoritmo que haga exactamente lo mismo).
- Incluso aunque tuviésemos una idea más precisa de cómo se hace en nuestro cerebro, el programa necesario podría ser tremendamente complicado :-)



Aplicaciones del Deep Learning

Aprendizaje automático **/ Inteligencia Computacional** **/ Redes neuronales artificiales [deep learning]**

- En vez de diseñar un algoritmo que resuelva el problema, recopilamos un montón de datos (ejemplos).
- Diseñamos un algoritmo que aprenda de esos datos y cree el programa necesario para resolver el problema.



Aplicaciones del Deep Learning

La solución basada en deep learning

- El programa generado automáticamente no tiene por qué parecerse a un programa implementado manualmente (en el caso de las redes neuronales, puede contener millones de números reales).
- Si tenemos éxito, el programa funcionará bien para nuevos ejemplos, aunque sean diferentes a los que utilizamos para su entrenamiento.
- Si los datos cambian, el programa puede cambiar entrenándolo de nuevo.



Limitaciones del Deep Learning

Limitaciones

Cualquier cosa que requiera razonar, planificar a largo plazo o manipular datos de forma algorítmica está fuera del alcance de las técnicas actuales de deep learning.



p.ej. Ordenar un conjunto de datos puede ser extremadamente difícil usando una red neuronal.



Limitaciones del Deep Learning

Limitaciones

Falta de comprensión (en sentido humano) ...

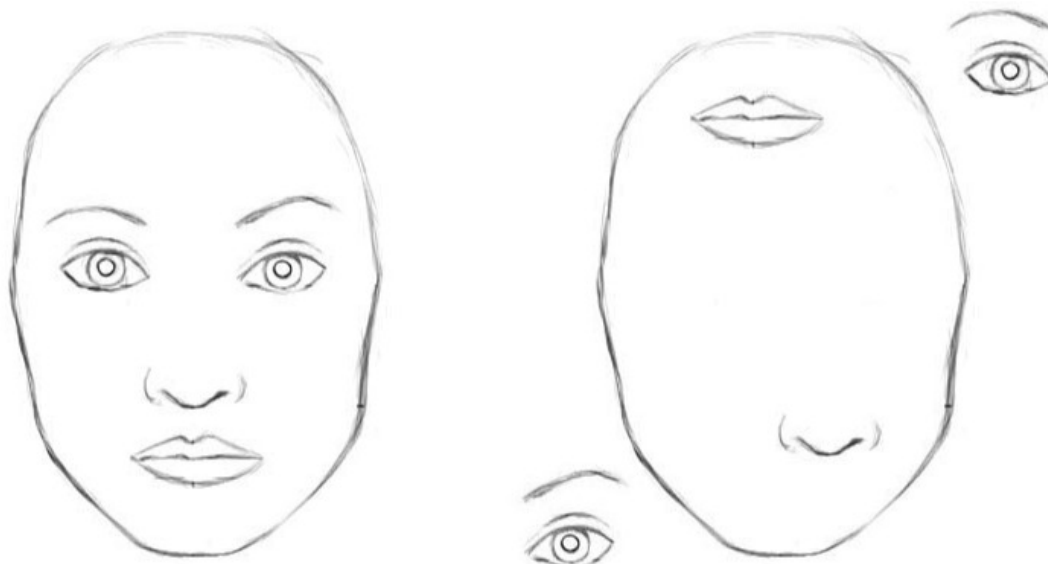


The boy is holding a baseball bat.



Limitaciones del Deep Learning

Las redes convolutivas [CNNs] funcionan muy bien en la práctica, pero...

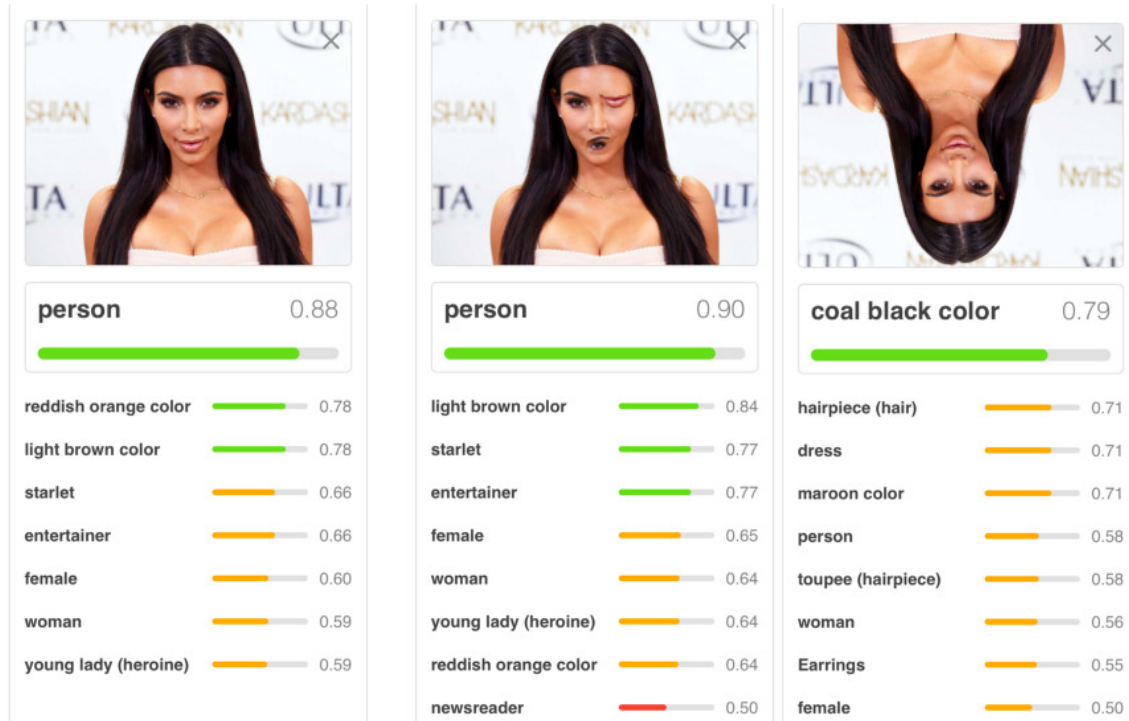


... para una CNN, ambas imágenes son similares ☹



Limitaciones del Deep Learning

“Convolutional neural networks are doomed”
—Geoffrey Hinton



Limitaciones del Deep Learning

- Las redes convolutivas detectan características, pero no su colocación relativa (traslación & rotación).
- Las redes convolutivas ignoran las posiciones relativas utilizando “pooling”, un apaño que funciona sorprendentemente bien en la práctica:

“The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster.” – Geoffrey Hinton



Limitaciones del Deep Learning

Pooling operation used in convolutional neural networks is a big mistake, and the fact that it works so well is a disaster.



Geoffrey Hinton

Professor at University of Toronto
Google Brain Team Manager
Godfather of the MLP, Backpropagation and DNN

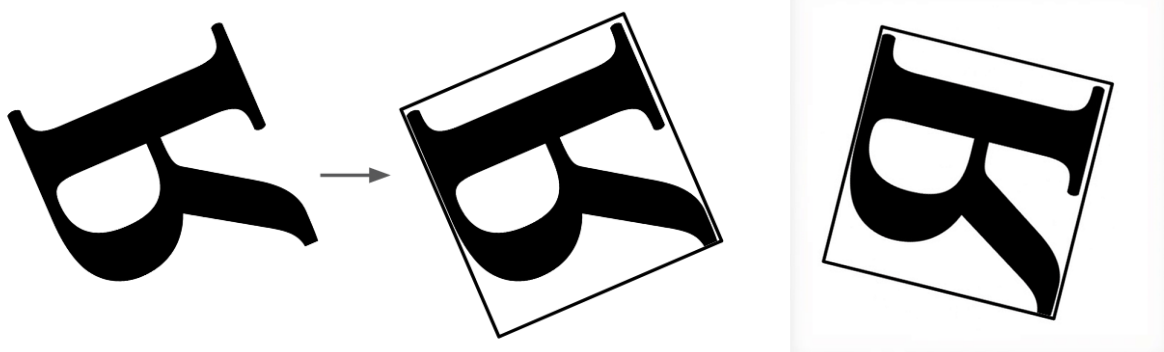
From Ask me anything on Reddit
https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton/



Limitaciones del Deep Learning

Problema clave de las redes convolutivas

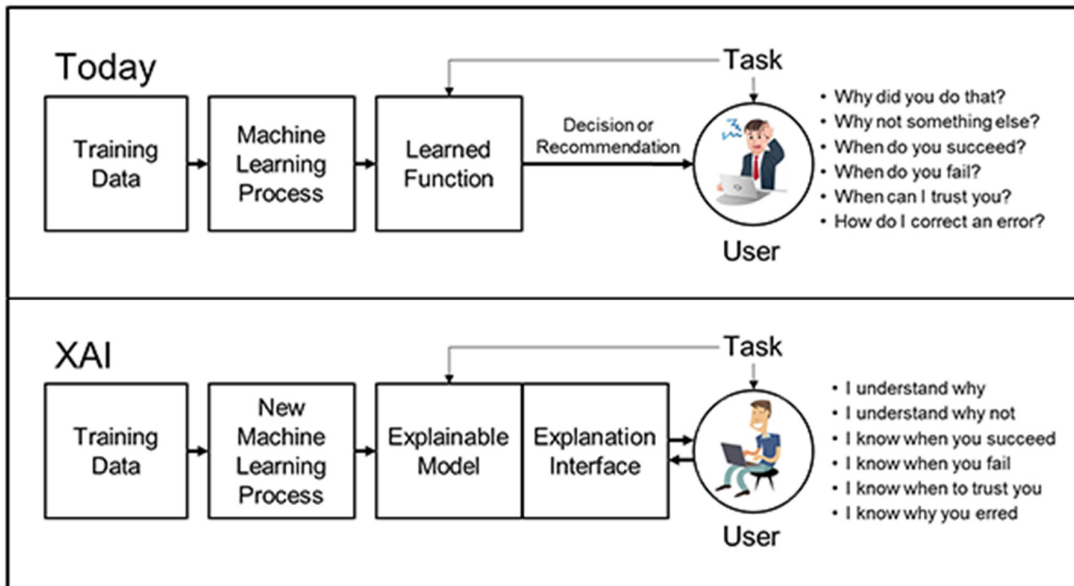
La representación interna de una red convolutiva no tiene en cuenta las relaciones espaciales entre objetos, ni la jerarquía existente entre objetos simples y los objetos compuestos de los que forman parte.



Limitaciones del Deep Learning

Limitaciones

Falta de interpretabilidad:
Redes neuronales como cajas negras



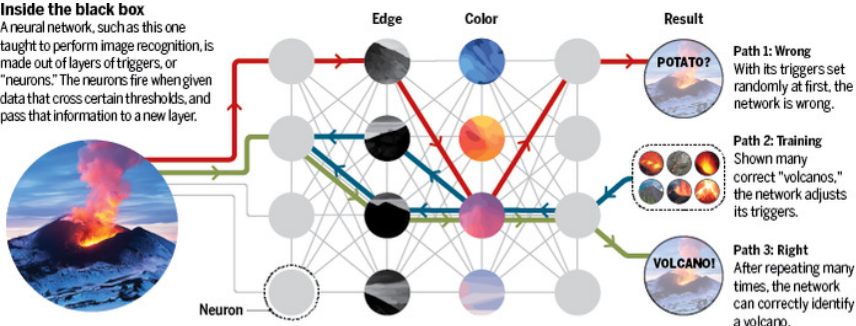
Limitaciones del Deep Learning

Limitaciones

Falta de interpretabilidad

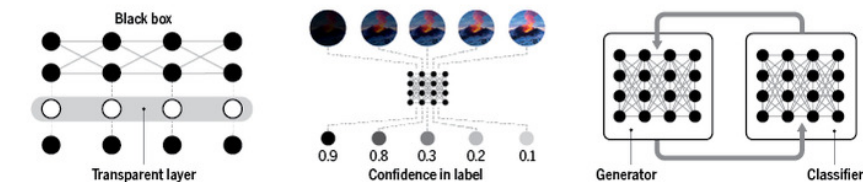
Inside the black box

A neural network, such as this one taught to perform image recognition, is made out of layers of triggers, or "neurons." The neurons fire when given data that cross certain thresholds, and pass that information to a new layer.



Into the darkness

Researchers have developed three broad classes of tools to look inside neural networks.



Controlling the black box

Some models guarantee relationships between two variables, like square footage and house price. These models are more transparent and can be wired into a neural network, helping control it.

Probing the black box

Researchers perturb the inputs to a trained neural network to see what most affects its decision-making. The probing can reveal the cause for one decision, but not the overall logic.

Embracing the darkness

Neural networks can be used to help understand other neural networks. Combining an image generator with an image classifier can expose knowledge gaps, such as accurate labels learned for the wrong reasons.



Limitaciones del Deep Learning

Atención...



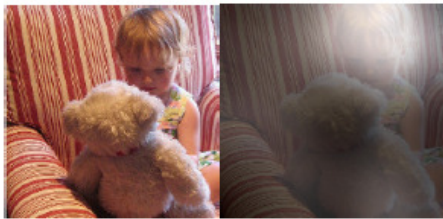
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



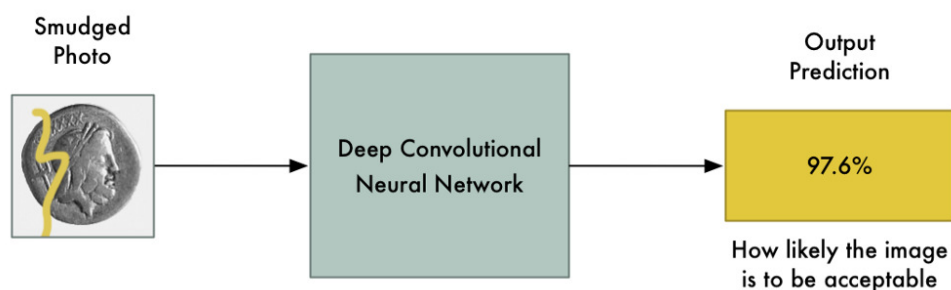
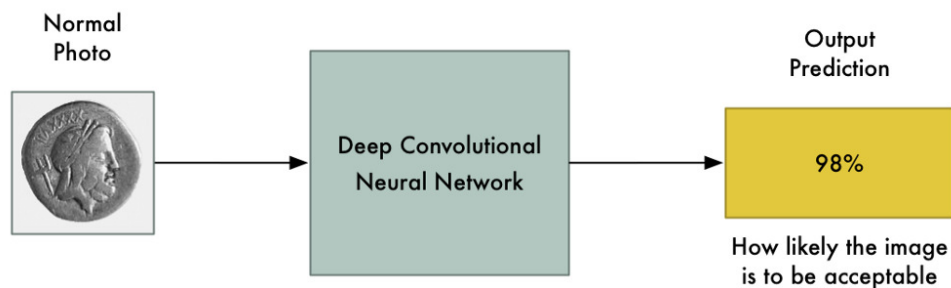
A giraffe standing in a forest with trees in the background.

Mecanismos de atención en la descripción textual de imágenes [image captioning]



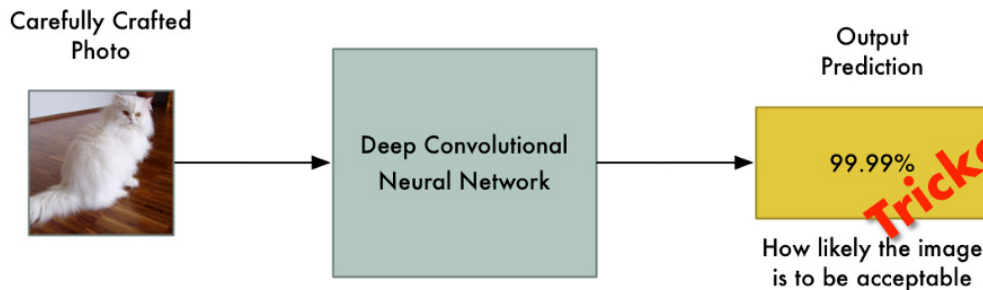
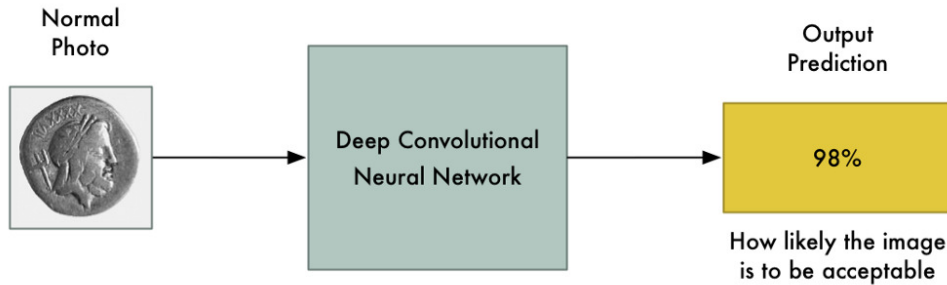
Limitaciones del Deep Learning

Lo deseable...



Limitaciones del Deep Learning

Lo que puede pasar...

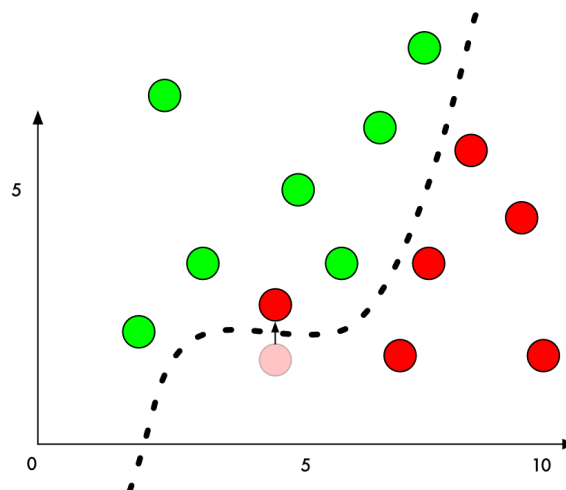


Limitaciones del Deep Learning

**Ejemplos diseñados por un adversario
(o cómo engañar fácilmente a una red neuronal)**

Si conocemos la red, podemos saber exactamente cómo modificar mínimamente la entrada para confundir a la red neuronal...

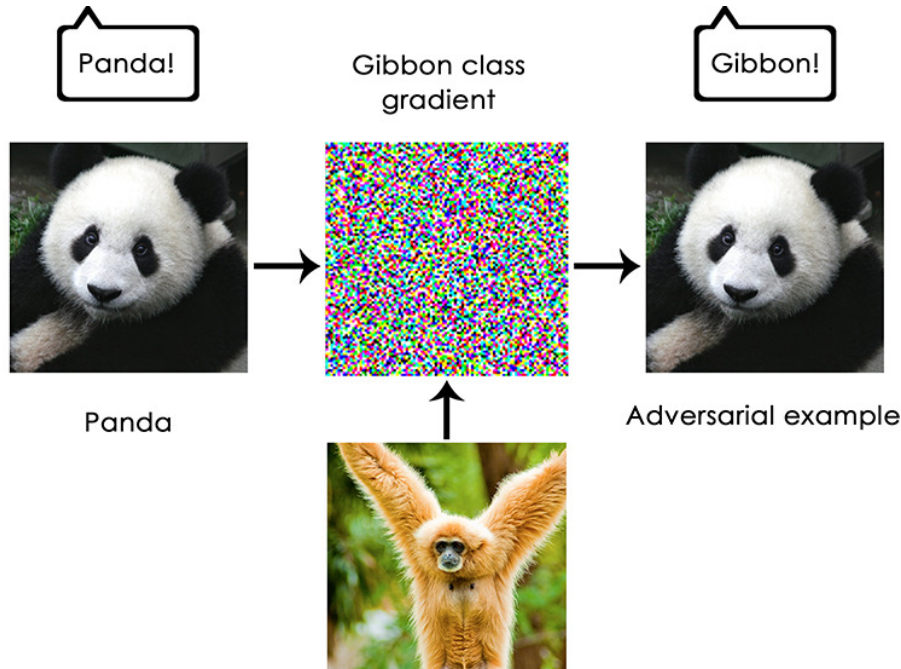
**... en la dirección
del gradiente !!!**



Limitaciones del Deep Learning

Limitaciones

Situaciones con adversario [adversarial examples]



Limitaciones del Deep Learning

**Ejemplos diseñados por un adversario
(o cómo engañar fácilmente a una red neuronal)**

Inception v3, trained on ImageNet

Enter a valid image URL or select an image from the dropdown.

enter image url
<http://i.imgur.com/il0yXAA.png> or select image

Use GPU
 Show computation flow



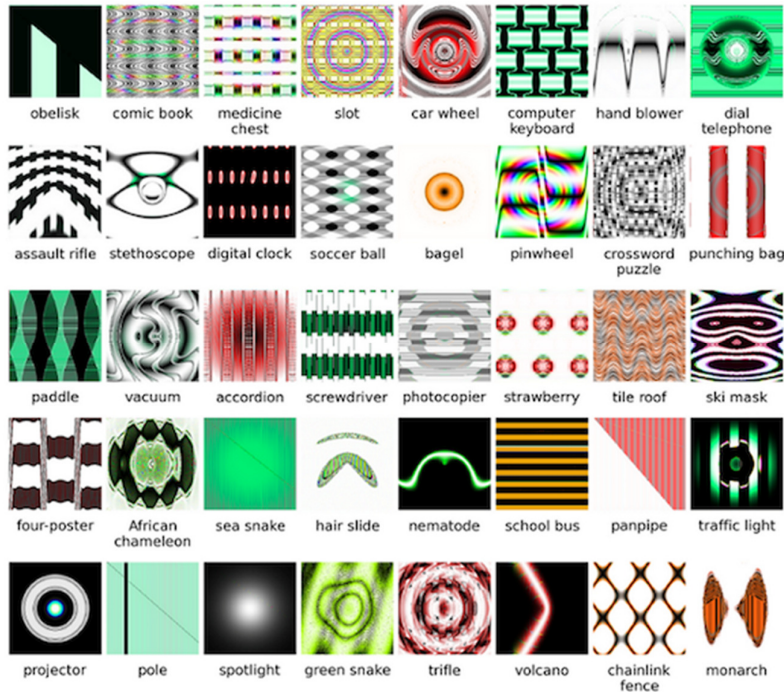
toaster	98%
Crock Pot	1%
Siamese cat	0%
wallaby	0%
carton	0%



Limitaciones del Deep Learning

Limitaciones

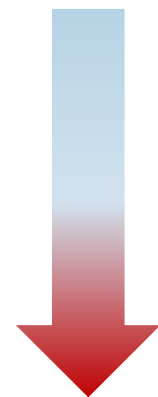
Situaciones con adversario [adversarial examples]



Limitaciones del Deep Learning

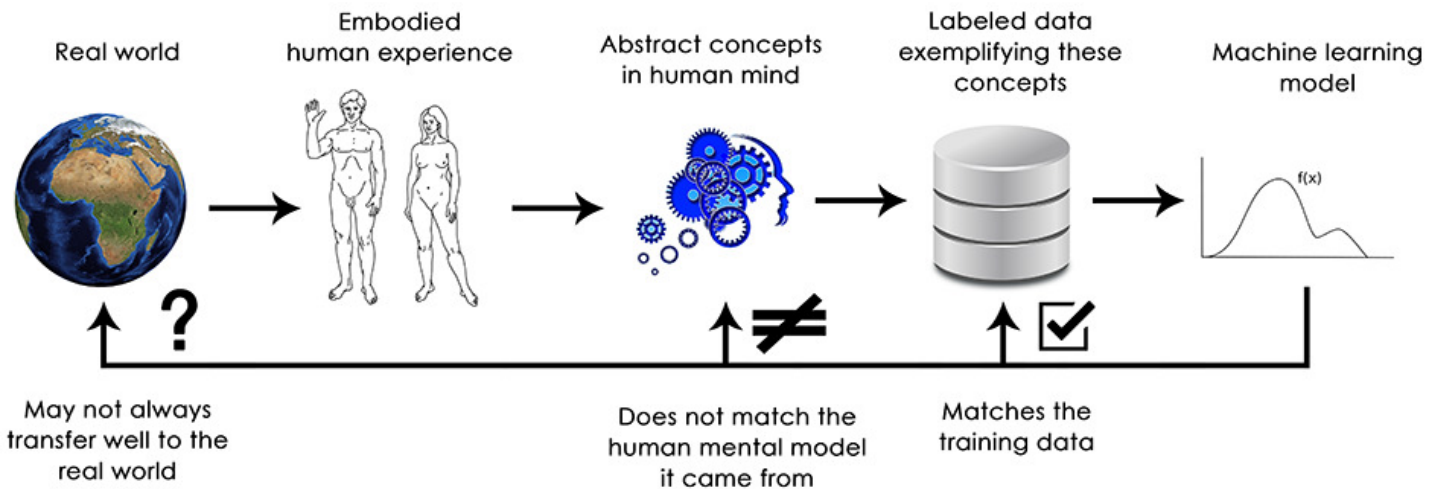
Limitaciones

Situaciones con adversario [adversarial examples]



Limitaciones del Deep Learning

Limitaciones



“Never fall into the trap of believing that neural networks understand the task they perform”

<https://blog.keras.io/the-limitations-of-deep-learning.html>



Información adicional

Fernando Berzal:
Redes Neuronales & Deep Learning

Edición en dos volúmenes, 2019

Volumen I: Entrenamiento de redes neuronales artificiales
ISBN 1-0903-2030-2

Volumen II: Regularización, optimización y arquitecturas especializadas
ISBN 1-0903-3688-8

<https://deep-learning.ikor.org>





Redes neuronales artificiales



Redes neuronales artificiales



El cerebro humano

Inspiración de las redes neuronales artificiales



Las RNA intentan modelar la estructura y funcionamiento de algunas partes del sistema nervioso animal.



Redes neuronales artificiales



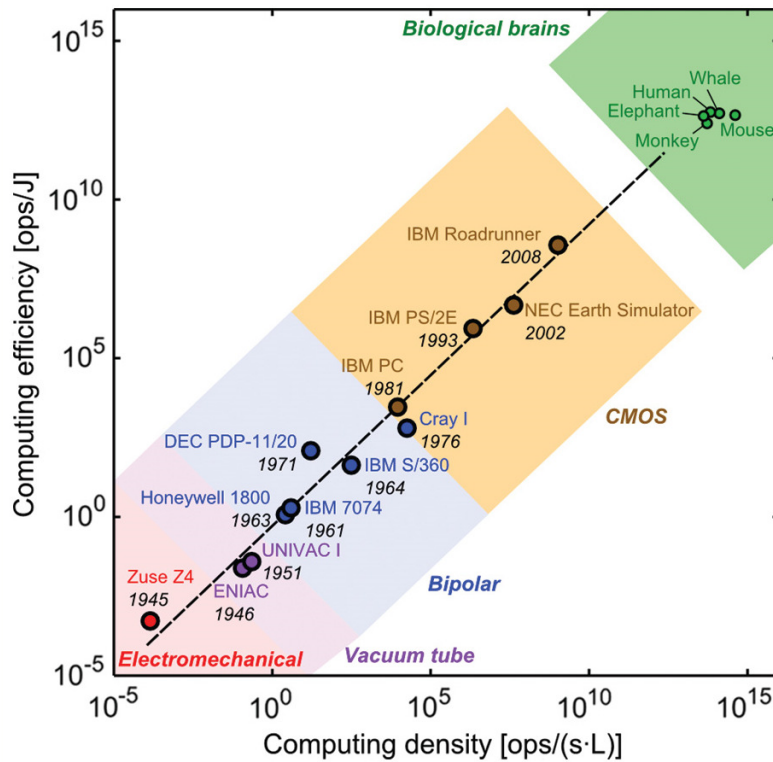
El cerebro humano

Diferencias entre un ordenador y el cerebro humano

Ordenador	Cerebro humano
Computación en serie	Computación en paralelo
Poco robusto	Tolerancia a fallos
Programable	Aprendizaje autónomo
Digital	Analógico
10^9 transistores	10^{11} neuronas $10^{14} \sim 10^{15}$ sinapsis
Nanosegundos (3.6GHz)	Milisegundos (4~90Hz)
51.2 GB/s	10 spikes/s
210,000,000 m/s	1 ~ 100 m/s
2.3×10^{13} TEPS	6.4×10^{14} TEPS



Redes neuronales artificiales



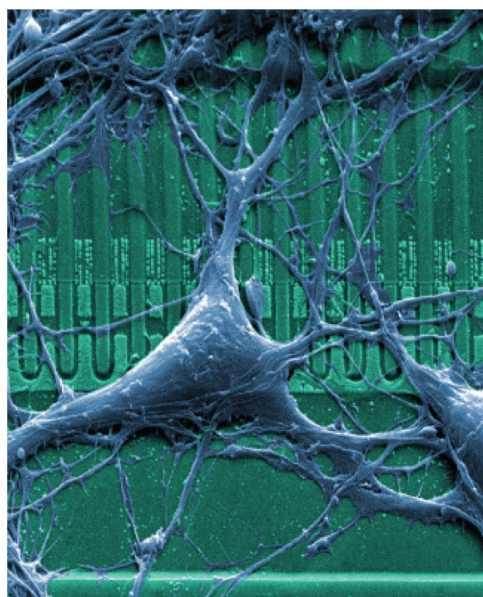
IBM Journal of Research and Development, 2011



Redes neuronales artificiales



Neuronas

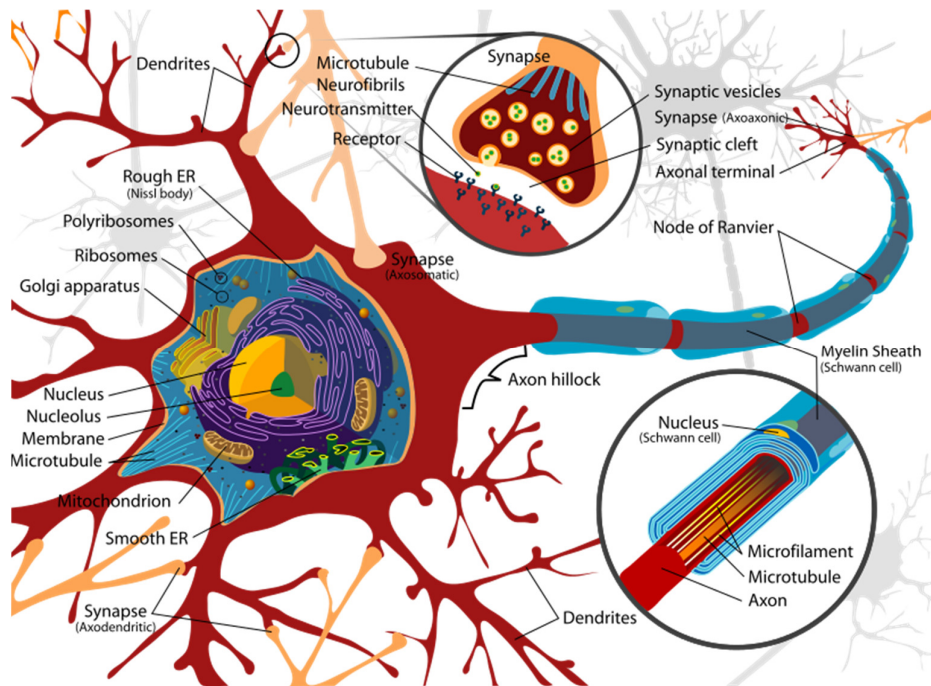


Microfotografía de una neurona "cultivada" sobre una oblea de silicio.
[Peter Fromherz, Max Planck Institute]





Neuronas



[Wikipedia]

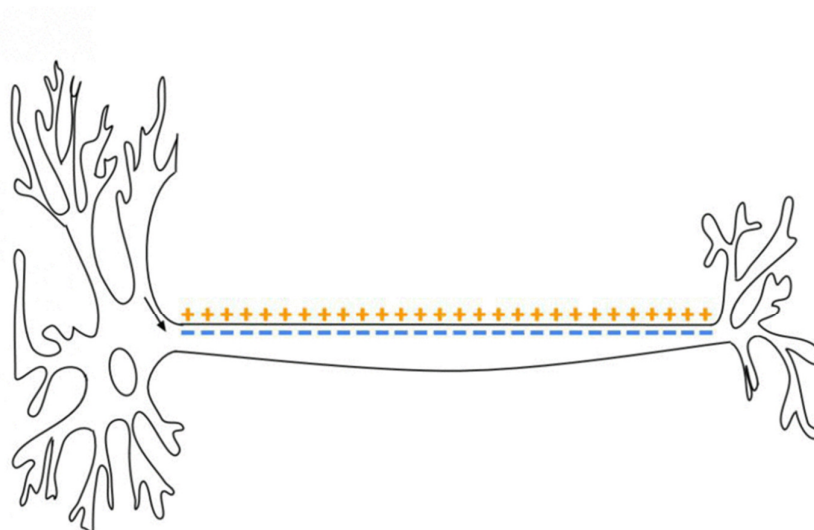


Redes neuronales artificiales



Neuronas

Spike, a.k.a. action potential [potencial de acción]



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

MakeAGIF.com



Redes neuronales artificiales

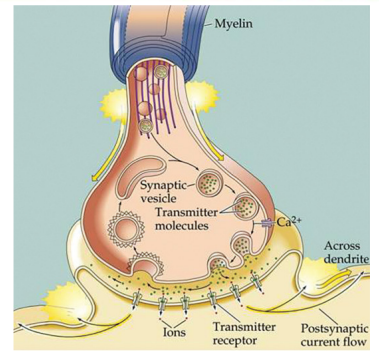


Neuronas

Sinapsis

Las sinapsis son lentas (en comparación con los transistores de un ordenador), pero...

- Son muy pequeñas y consumen muy poca energía.
- Se adaptan utilizando señales locales.



Como tenemos cerca de 10^{11} neuronas y de 10^{14} a 10^{15} sinapsis, muchas sinapsis pueden influir en un "cálculo" en un período de tiempo muy breve:

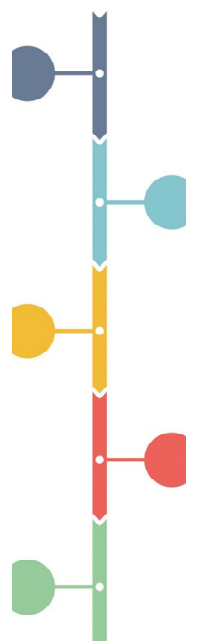
Ancho de banda muy superior al de un ordenador.



Historia



<p>1956: Psychologist Frank Rosenblatt uses theories about how brain cells work to design the perceptron, an artificial neural network that can be trained to categorize simple shapes.</p> <p>1969: AI pioneers Marvin Minsky and Seymour Papert write a book critical of perceptrons that quashes interest in neural networks for decades.</p> <p>1986: Yann LeCun and Geoff Hinton perfect backpropagation to train neural networks that pass data through successive layers of artificial neurons, allowing them to learn more complex skills.</p> <p>1987: Terry Sejnowski at Johns Hopkins University creates a system called NET-talk that can be trained to pronounce text, going from random babbling to recognizable speech.</p> <p>1990: At Bell Labs, LeCun uses backpropagation to train a network that can read handwritten text. AT&T later uses it in machines that can read checks.</p> <p>1995: Bell Labs mathematician Vladimir Vapnik publishes an alternative method for training software to categorize data such as images. This sidelines neural networks again.</p>	<p>2006: Hinton's research group at the University of Toronto develops ways to train much larger networks with tens of layers of artificial neurons.</p> <p>June 2012: Google uses deep learning to cut the error rate of its speech recognition software by 25 percent.</p> <p>October 2012: Hinton and two colleagues from the University of Toronto win the largest challenge for software that recognizes objects in photos, almost halving the previous error rate.</p> <p>March 2013: Google buys DNN Research, the company founded by the Toronto team to develop their ideas. Hinton starts working at Google.</p> <p>March 2014: Facebook starts using deep learning to power its facial recognition Vfeature, which identifies people in uploaded photos.</p> <p>May 2015: Google Photos launches. The service uses deep learning to group photos of the same people and let you search your snapshots using terms like "beach" or "dog."</p>
--	---



Otros modelos de redes



Arquitecturas basadas en el cerebro

Simulación (muy ineficiente) → “Neuromorphic Computing”

The K-Computer, Japan
simulating 1 Billion very simple neurons on 65.000 processors
1% „Brain“ Size, 13 Megawatt, 1500x slower than biology
Energy = Power x Time

10 Billion times less energy efficient
Wait 4 years for a simulated day

Diesmann, Proceedings of the 4th Biosupercomputing Symposium, Tokyo, 2012 ©RIKEN

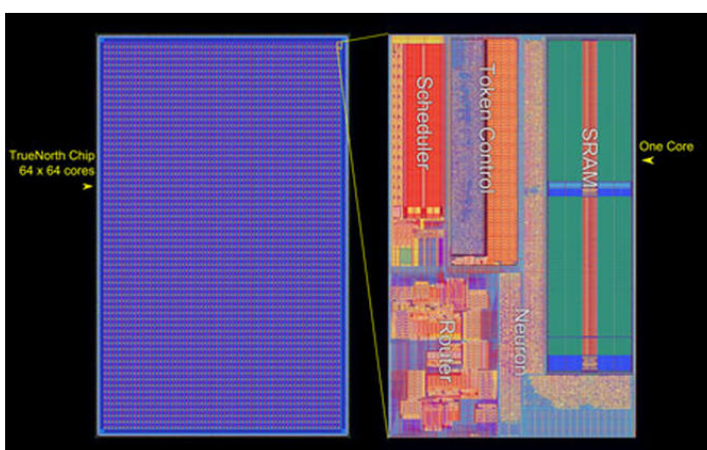
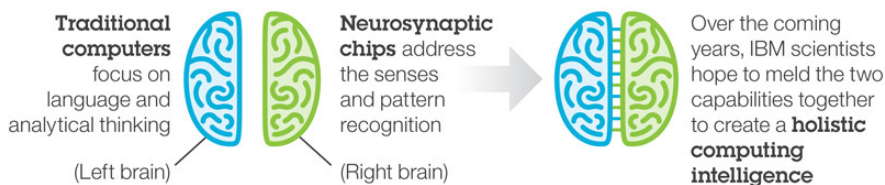


Otros modelos de redes



Arquitecturas basadas en el cerebro

IBM TrueNorth Brain-inspired Computer



4096 cores
1M neurons
256M synapses
5.4B transistors
CMOS
70mW

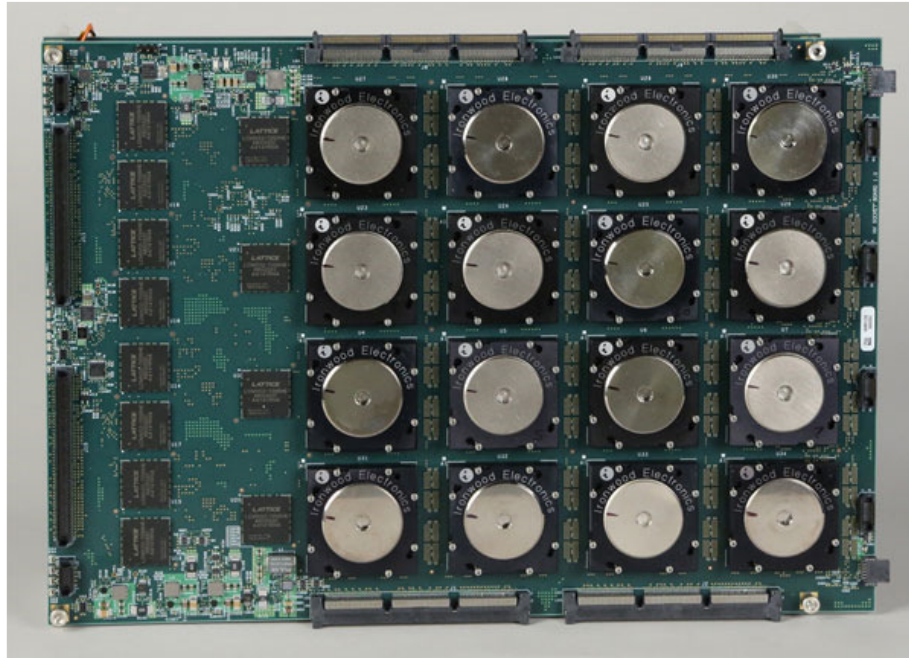


Otros modelos de redes



Arquitecturas basadas en el cerebro

IBM TrueNorth Brain-inspired Computer



Synapse 16
16M neurons
4B synapses

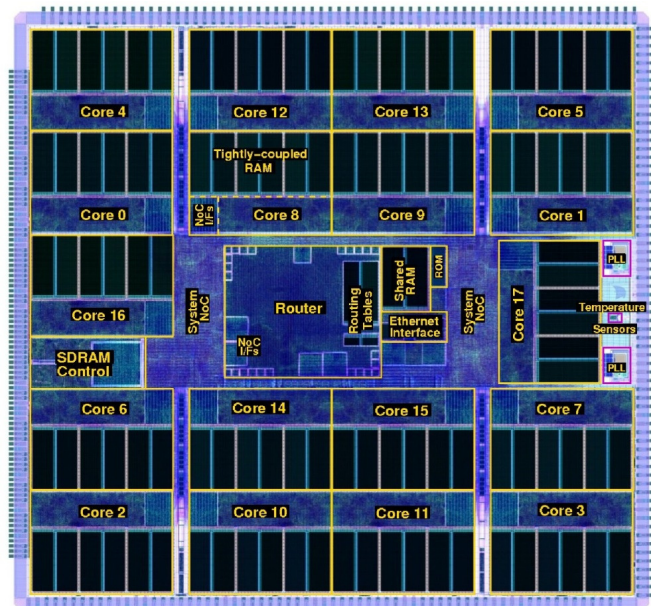
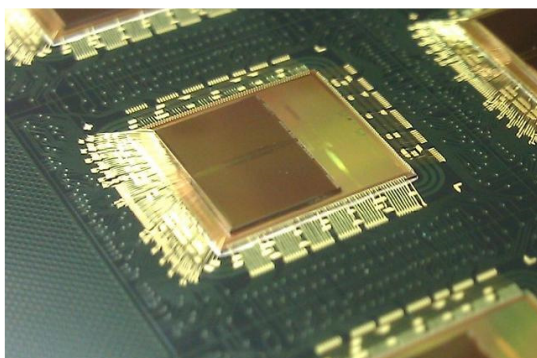


Otros modelos de redes



Arquitecturas basadas en el cerebro

SpiNNaker project (UK)



Globally Asynchronous Locally Synchronous (GALS) chip:
18 ARM968 processor nodes + 128MB Mobile DDR SDRAM
<http://apt.cs.manchester.ac.uk/projects/SpiNNaker/project/>

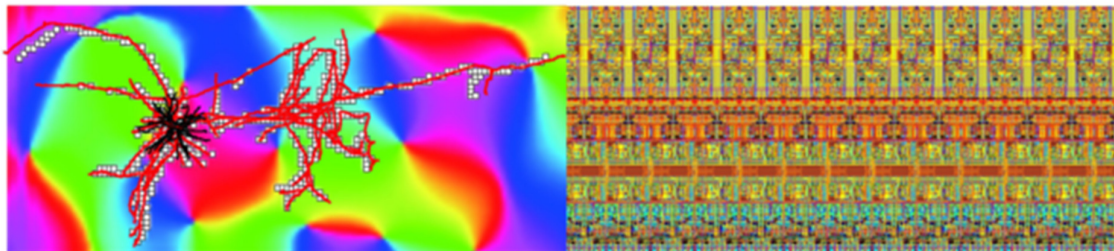
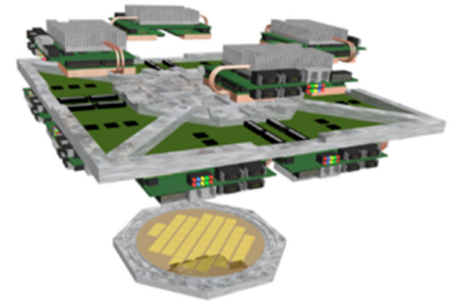


Otros modelos de redes



Arquitecturas basadas en el cerebro

BrainScaleS (Germany)



Mixed CMOS signals

<https://brainscales.kip.uni-heidelberg.de/>

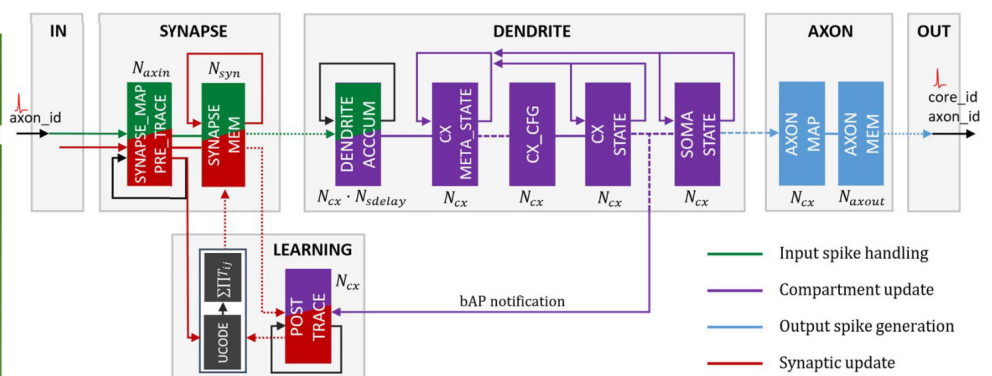
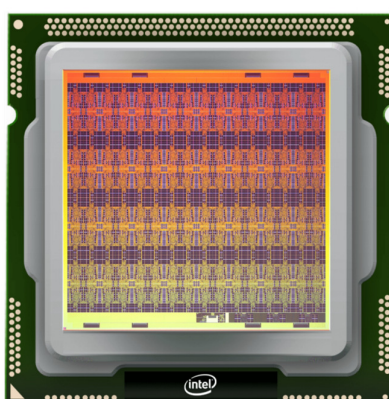


Otros modelos de redes



Arquitecturas basadas en el cerebro

Intel Loihi (self-learning neuromorphic research chip)



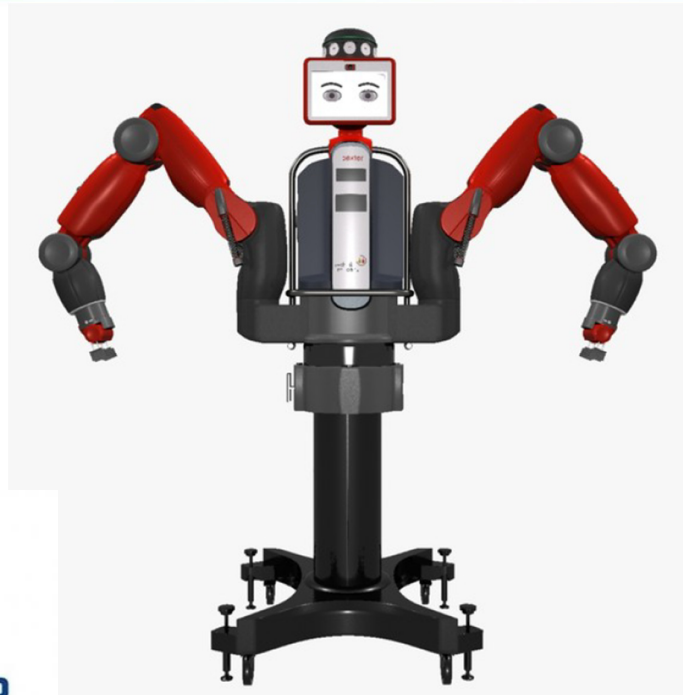
SNNs [Spiking Neural Networks]

130k neuronas / chip, 60mm², 14nm

<https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html/>
<https://ieeexplore.ieee.org/document/8259423/>



Otros modelos de redes



Laboratorio de Neuro-Robótica y Robots Colaborativos



Otros modelos de redes

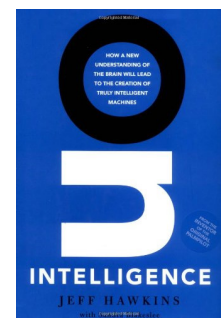
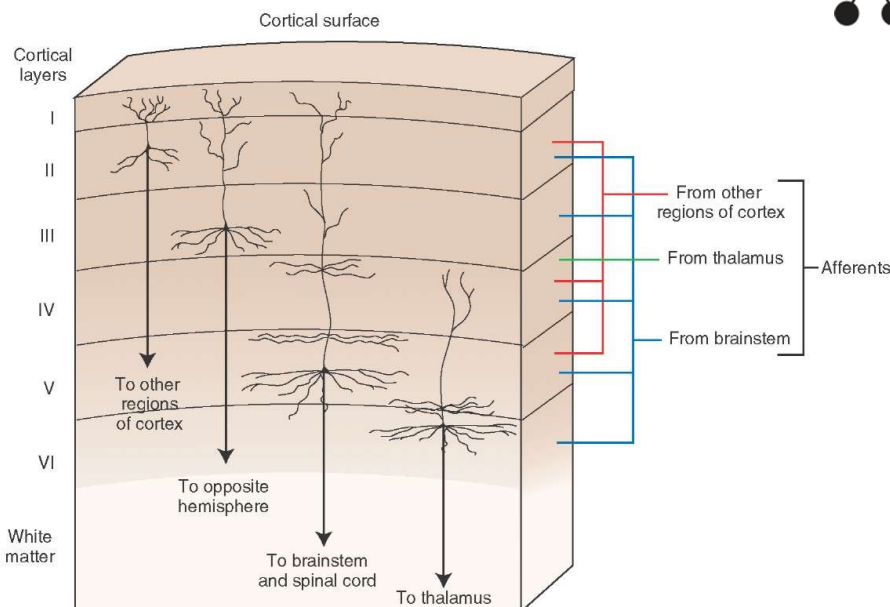


Arquitecturas basadas en el cerebro

HTM [Hierarchical Temporal Memory]



Numenta



Otros modelos de redes

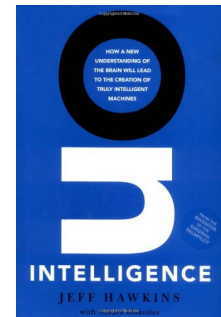
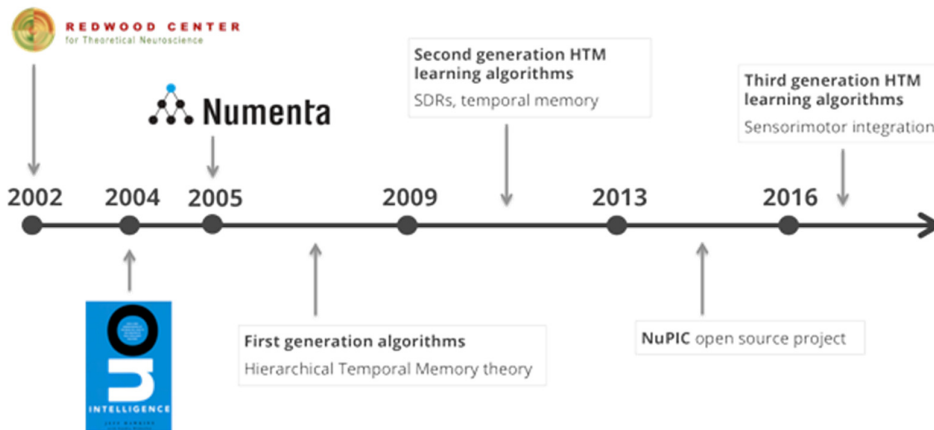


Arquitecturas basadas en el cerebro

HTM [Hierarchical Temporal Memory]



Numenta



<http://numenta.com/>



Otros modelos de redes

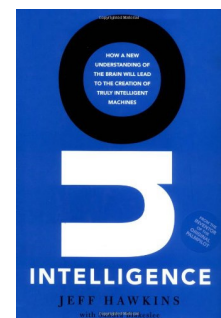
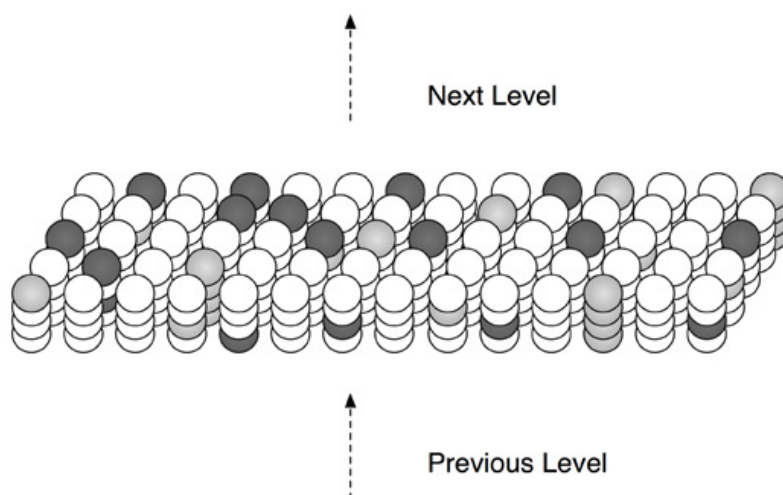


Arquitecturas basadas en el cerebro

HTM [Hierarchical Temporal Memory]



Numenta



<http://numenta.com/>



Otros modelos de redes

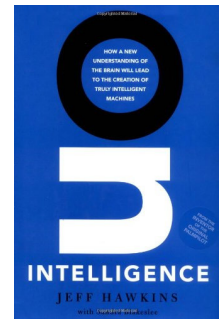
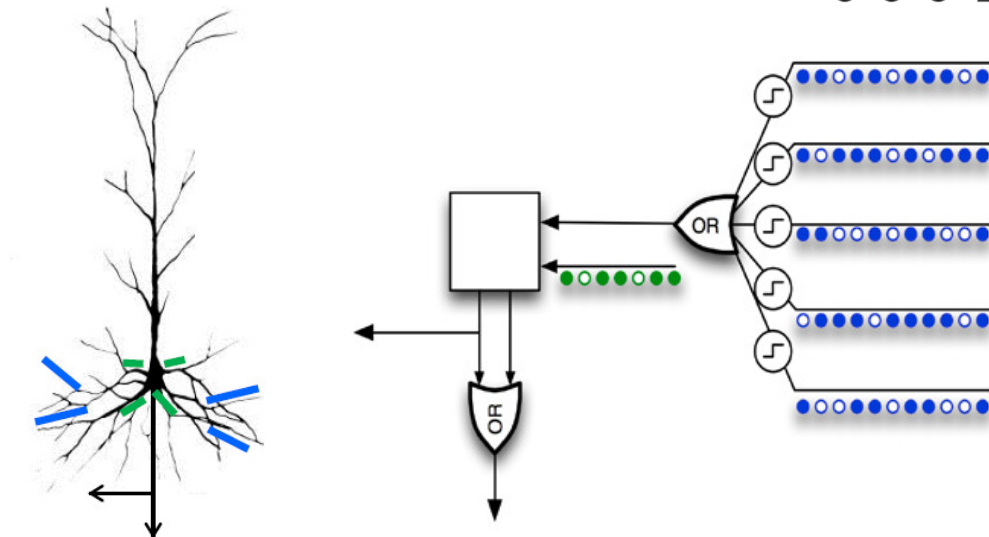


Arquitecturas basadas en el cerebro

HTM [Hierarchical Temporal Memory]



Numenta



<http://numenta.com/>



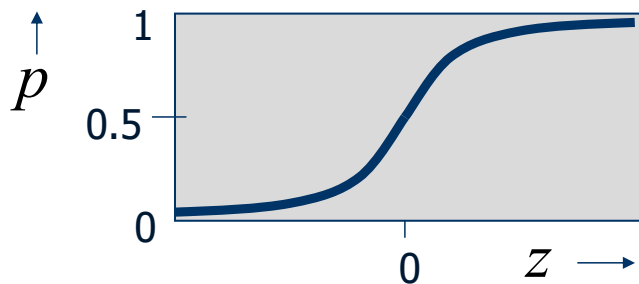
Otros modelos de redes



Neuronas binarias estocásticas

$$z = \sum_i x_i w_i$$

$$p = \frac{1}{1 + e^{-z}}$$



Las mismas ecuaciones que las neuronas sigmoideas, si bien su salida se interpreta como una probabilidad (de producir un spike en una pequeña ventana de tiempo).



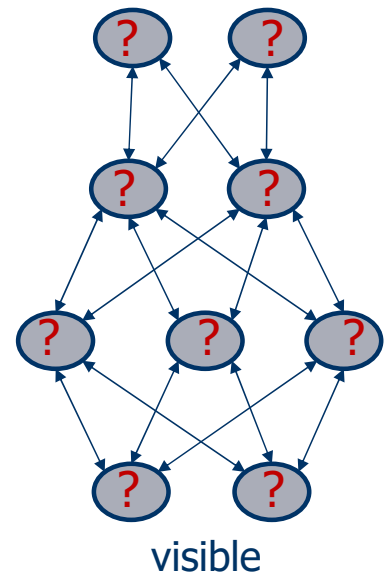


Máquinas de Boltzmann

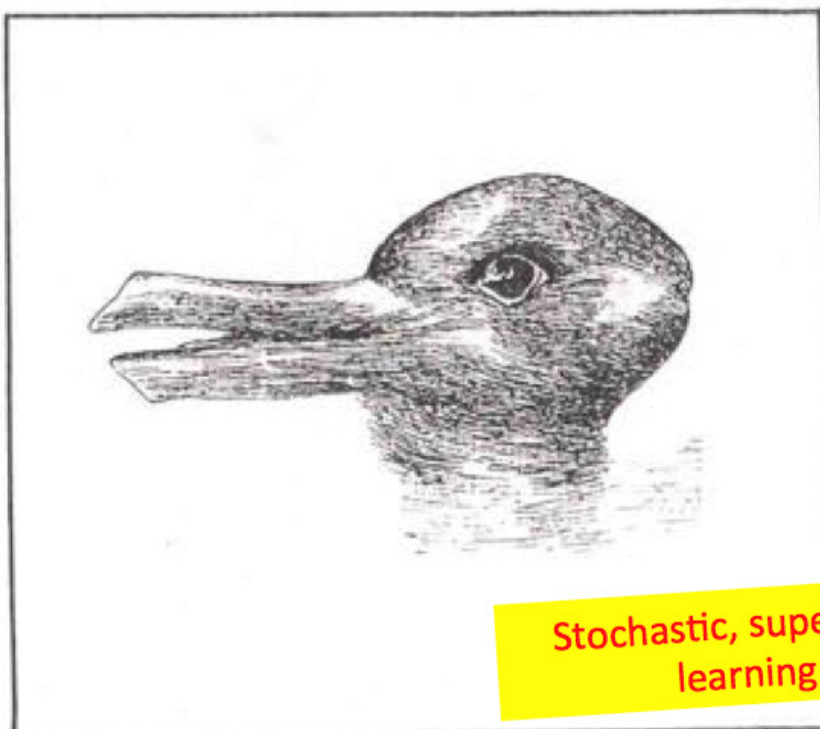
- En una máquina de Boltzmann, las actualizaciones estocásticas de las distintas unidades deben ser secuenciales.
- Existe una arquitectura que admite actualizaciones paralelas alternas mucho más eficientes:

DBM [Deep Boltzmann Machine]

- Sin conexiones entre unidades de una misma capa.
- Sin conexiones entre capas no adyacentes.



Máquinas de Boltzmann



Otros modelos de redes



Geoffrey Hinton: "The Next Generation of Neural Networks"
Google Tech Talks, 2007

<https://www.youtube.com/watch?v=AyzOUbkUf3M>



Otros modelos de redes



Geoff Hinton doesn't need to make hidden units.
They hide by themselves when he approaches.

Geoff Hinton doesn't disagree with you,
he contrastively diverges

Deep Belief Nets actually
believe deeply in Geoff Hinton.

Yann LeCun: "Geoff Hinton facts"

<http://yann.lecun.com/ex/fun/index.html>





How do you fix a neurotic algorithm?

Call Sigmoid Freud.



Técnicas de Deep Learning



Pero resulta que, después de todo,
la solución era muy simple:

- Modelos paramétricos suficientemente grandes (redes neuronales con muchos pesos ajustables).
- Conjuntos de entrenamiento suficientemente grandes para entrenar las redes usando el gradiente descendente.

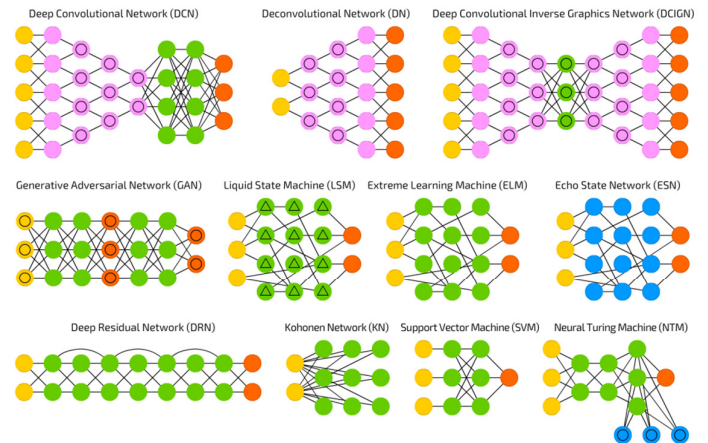
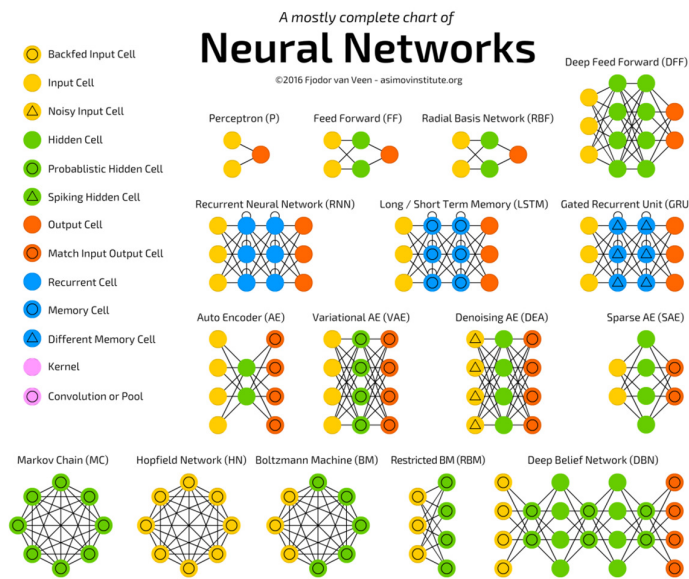
Richard Feynman, sobre el Universo:

It's not complicated, it's just a lot of it



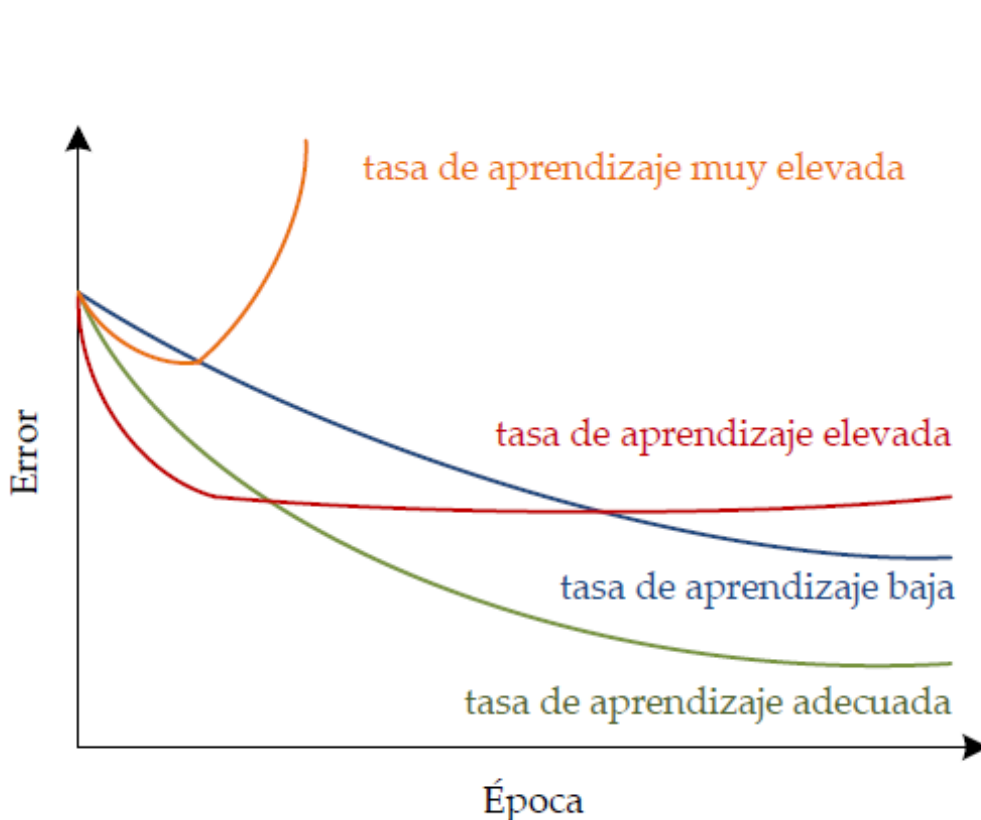


El zoo de las redes neuronales



Técnicas de Deep Learning

Algoritmos de optimización



$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

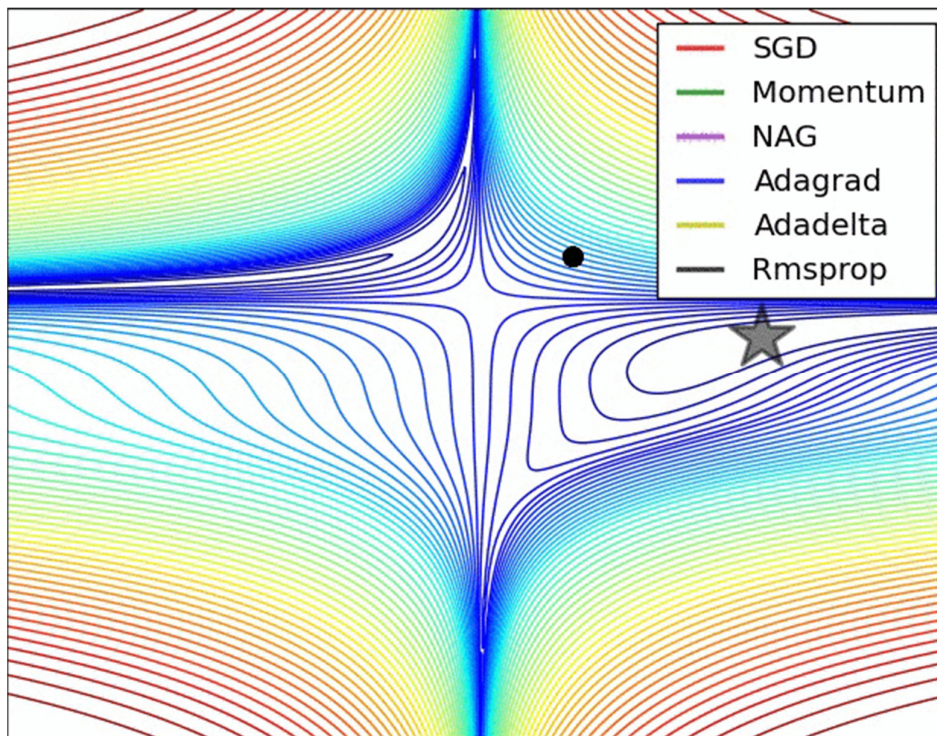


Técnicas de Deep Learning

Algoritmos de optimización



SGD [Stochastic Gradient Descent]



Alec Radford
<https://twitter.com/alecrad>

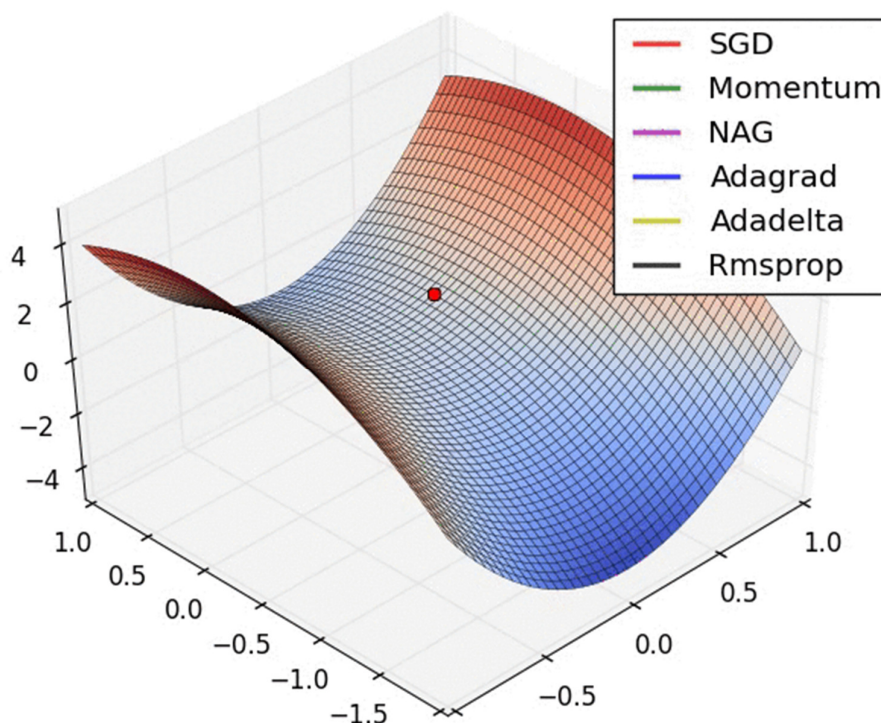


Técnicas de Deep Learning

Algoritmos de optimización



SGD [Stochastic Gradient Descent] @ saddle point

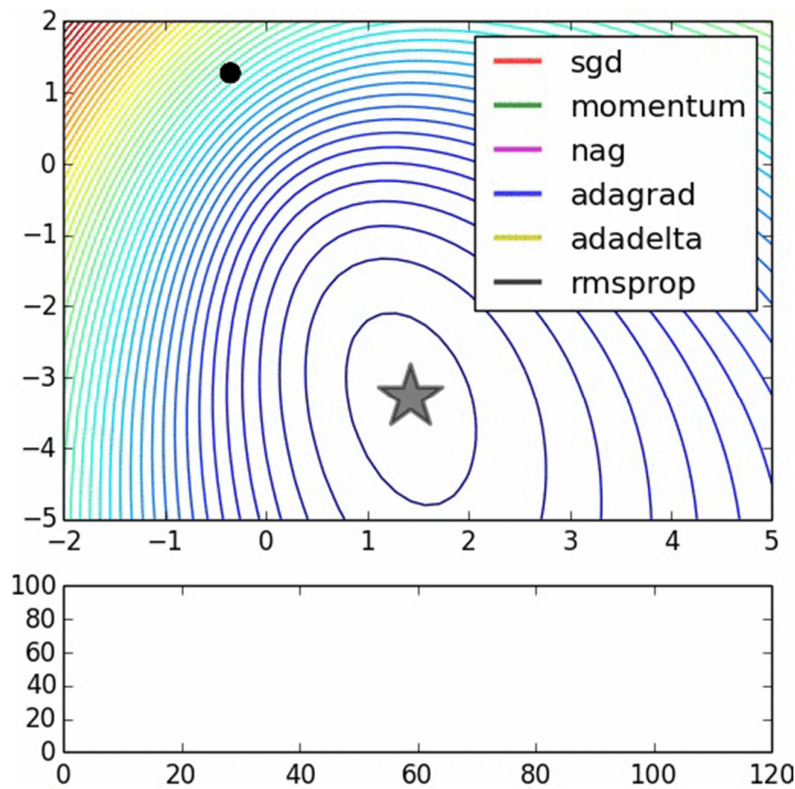


Alec Radford
<https://twitter.com/alecrad>



Técnicas de Deep Learning

Algoritmos de optimización



Alec Radford
<https://twitter.com/alecrad>

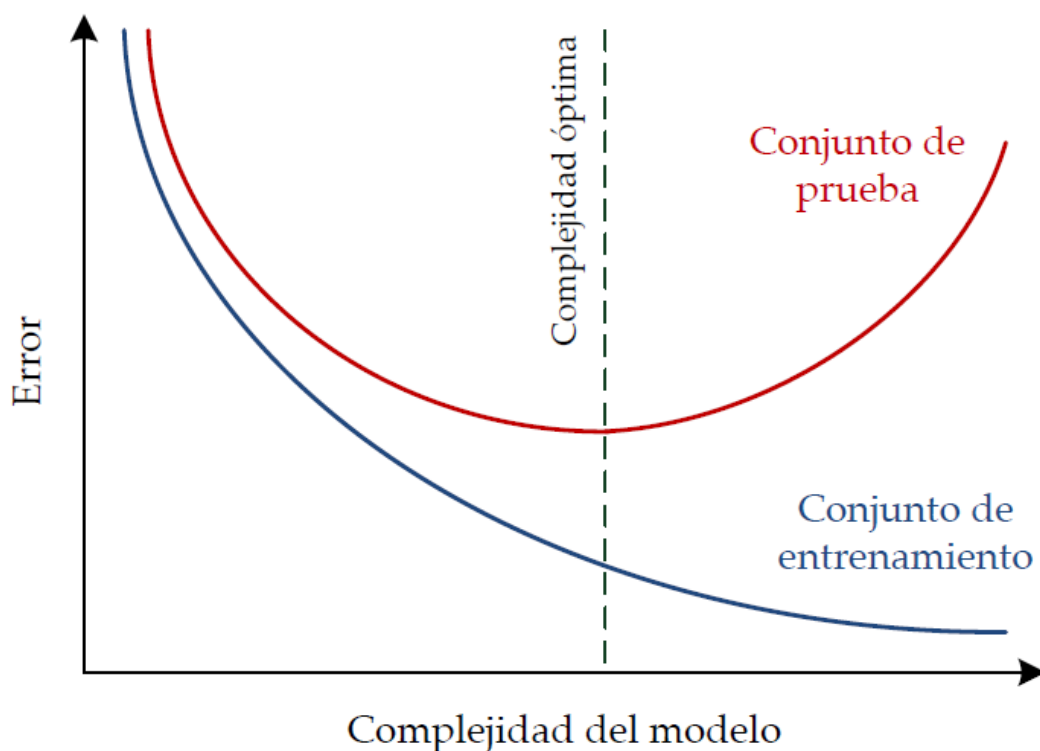


Técnicas de Deep Learning

Regularización



El problema del sobreaprendizaje



Estrategias para evitar el sobreaprendizaje:

- Obtener más datos (la mejor opción si tenemos capacidad para entrenar la red usando más datos).
- Ajustar los parámetros de la red para que tenga la capacidad adecuada (suficiente para identificar las regularidades en los datos, pero no demasiada para ajustarse a las espúreas, suponiendo que sean más débiles que las auténticas).



"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."



Capacidad de la red: Topología

Algunas formas de limitar la capacidad de la red actuando sobre su topología:

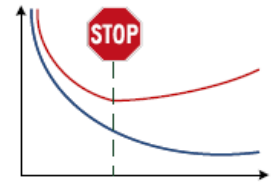
- **Arquitectura de la red:**
Se limita el número de capas ocultas y/o el número de unidades por capa.
- **Weight sharing:**
Se reduce el número de parámetros de la red haciendo que distintas neuronas compartan los mismos pesos (p.ej. redes convolutivas).





Capacidad de la red: Entrenamiento

Algunas formas de limitar la capacidad de la red actuando sobre su algoritmo de entrenamiento:



- **Early stopping:** Se comienza a entrenar la red con pesos pequeños y se para el entrenamiento antes de que sobreaprenda.
- **Weight decay:** Se penalizan los pesos grandes en función de sus valores al cuadrado (penalización L2) o absolutos (penalización L1).
- **Ruido:** Se añade ruido a los pesos o actividades de las neuronas de la red que se está entrenando.



Una tercera estrategia: Combinar modelos

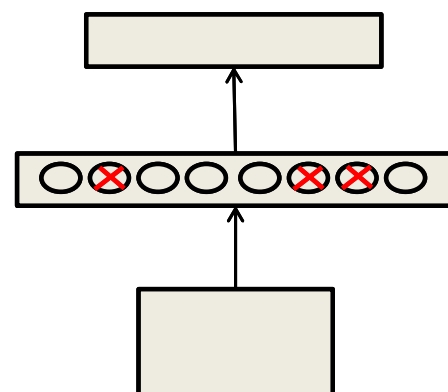
Model averaging” (a.k.a. “ensembles”):

Muchos modelos diferentes con distintos parámetros o el mismo tipo de modelo utilizando distintos subconjuntos del conjunto de entrenamiento [bagging].

Dropout

ilas conspiraciones complejas
no son robustas!

-- Geoff Hinton.



Técnicas de Deep Learning

Ajuste de hiperparámetros



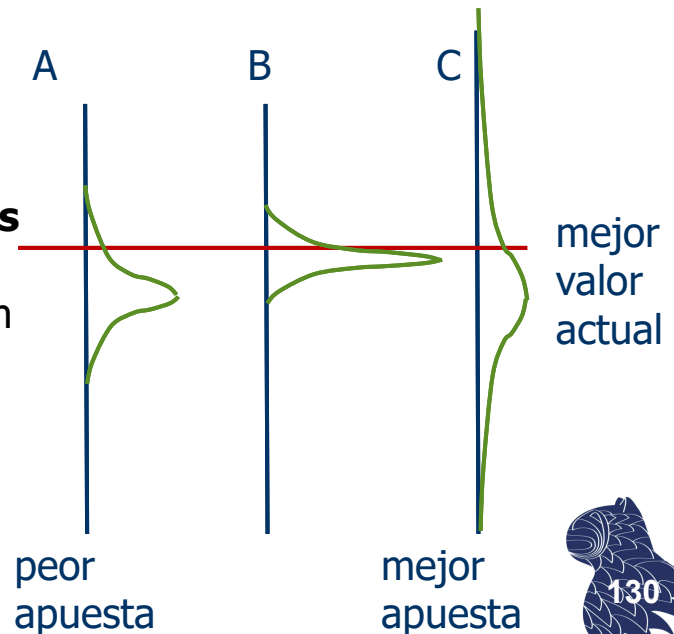
Una de las principales dificultades prácticas del uso de redes neuronales es la destreza que requiere establecer todos sus parámetros (“arte” más que ciencia)

p.ej.

Modelo de procesos gaussianos

A partir de la mejor configuración conocida, se elige una combinación de hiperparámetros tal que la mejora esperada sea grande (sin preocuparse por la posibilidad de empeorar).

Snoek, Larochelle & Adams
NIPS 2012



Técnicas de Deep Learning

Ajuste de hiperparámetros



AutoML

Aprendizaje automático [Machine Learning]

- Mucho mejor que ir haciendo pruebas manualmente (no es el tipo de tarea que los humanos hacemos bien).
- Evita sesgos psicológicos no deseados: método menos propenso a funcionar mejor con el método que nos gusta y peor con el que no (las personas no podemos evitarlo ;-)



En la práctica



Deep Learning

What society thinks I do

What my friends think I do

What other computer scientists think I do

What mathematicians think I do

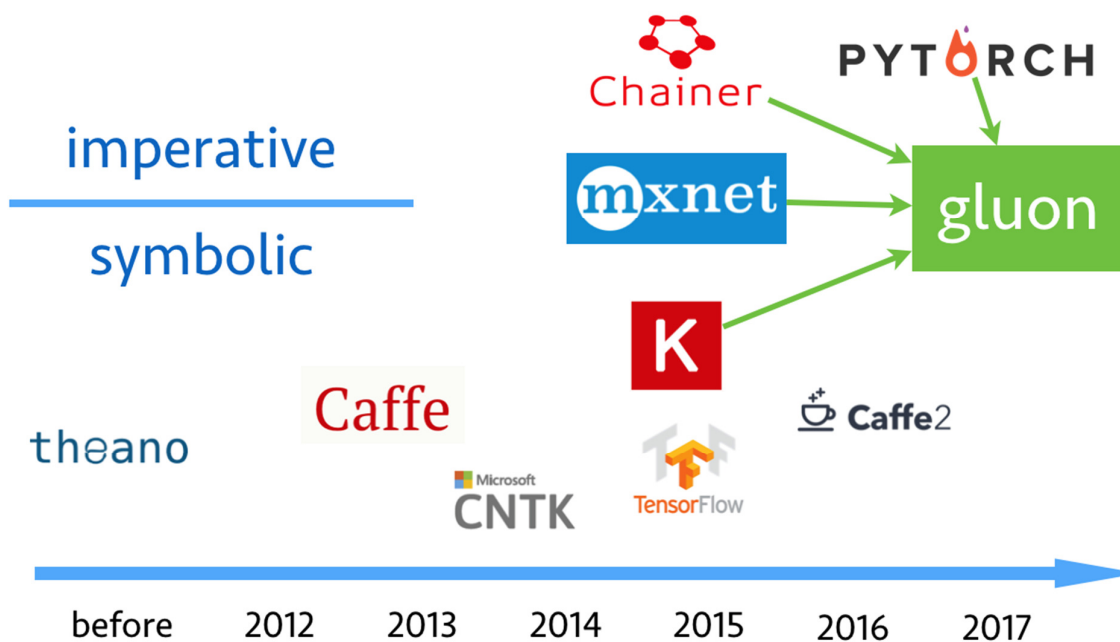
What I think I do

```
In [1]:  
import keras  
Using TensorFlow backend.
```

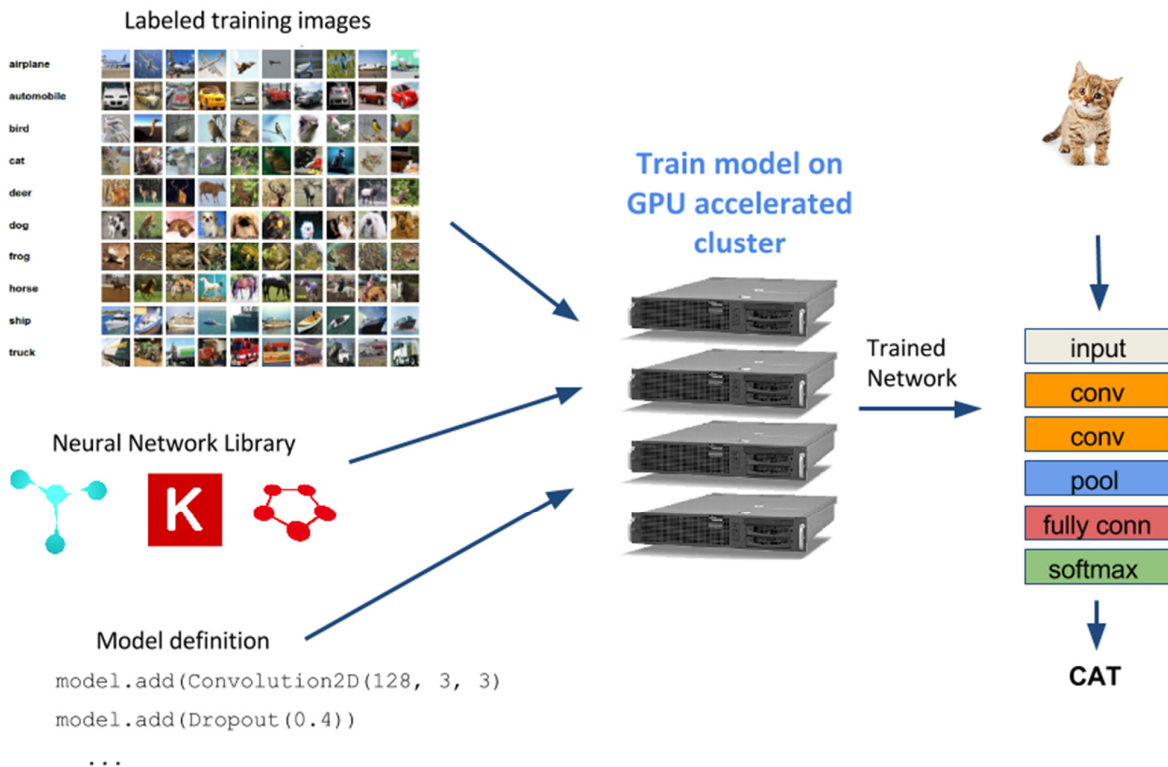


En la práctica

Software para deep learning



En la práctica Software para deep learning



En la práctica Software para deep learning



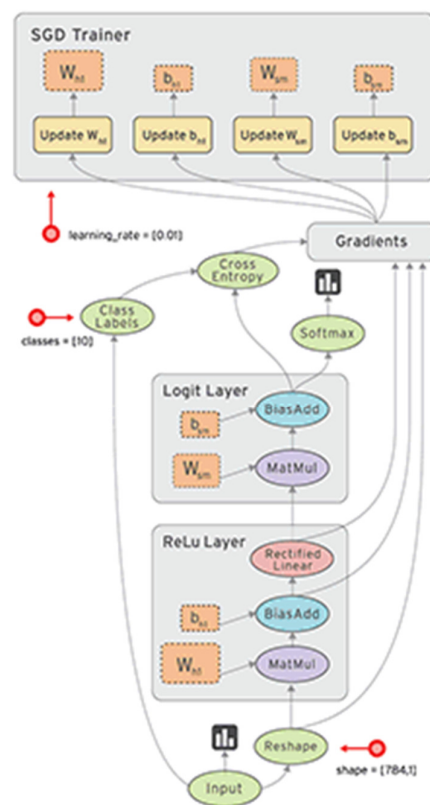
Google TensorFlow

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

Licencia Apache 2.0



Data flow graph



En la práctica

Software para deep learning



GLUON



En la práctica



They are neither neural nor networks!

They are chains of differentiable, parameterized geometric functions, trained with gradient descent (obtained via chain rule)

A small set of high school level ideas put together



Francois Chollet

Creator of Keras
Google AI Researcher

<https://twitter.com/fchollet/status/951906139632840704>



En la práctica



My favorite definition of Deep Learning is matrix multiplication, a lot of matrix multiplication...



Barbara Fusinska

Machine Learning Programmer at Microsoft

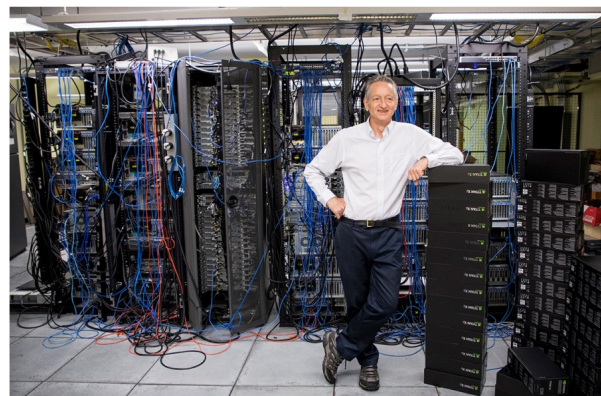
From Keynote Networks are like onions: Practical Deep Learning with TensorFlow
<https://www.youtube.com/watch?v=95IV8DoWRwI>



En la práctica



En deep learning, todo son vectores...



The Vector Institute for AI

University of Toronto

<http://vectorinstitute.ai/>





With all due respect to the brilliant Geoff Hinton, thought is not a vector, and AI is not a problem in statistics.
-- Oren Etzioni

Shortcomings of Deep Learning

<https://www.kdnuggets.com/2016/11/shortcomings-deep-learning.html>



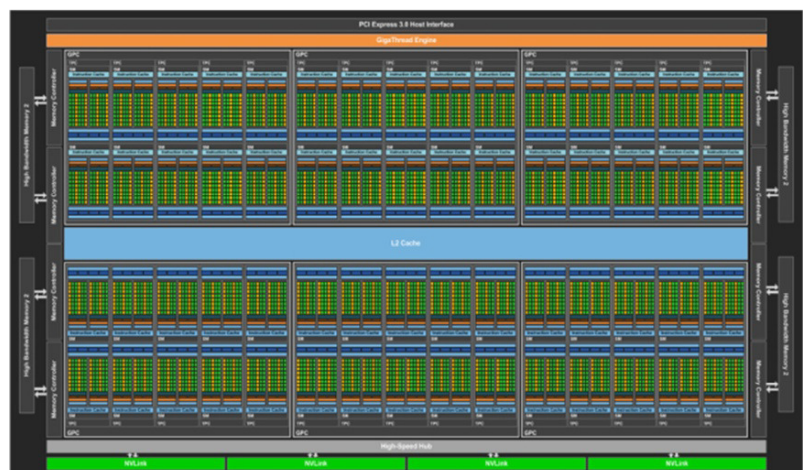
En la práctica Hardware para deep learning



GPU [procesador SIMD]: Data-level parallelism



NVIDIA Pascal SIMD Processor



NVIDIA Pascal P100 GPU



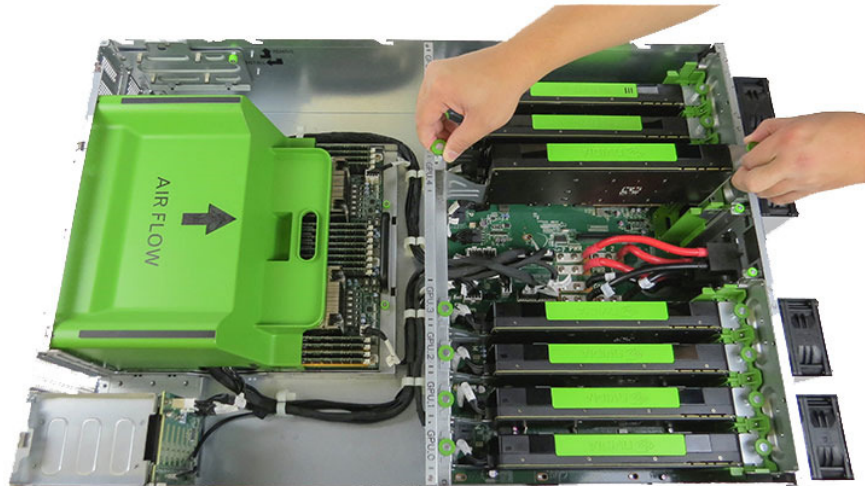
En la práctica

Hardware para deep learning



Facebook Big Sur

Servidor con 8 GPUs NVIDIA Tesla para deep learning



Facebook Engineering Blog, December 2015

<https://code.facebook.com/posts/1687861518126048/facebook-to-open-source-ai-hardware-design/>



En la práctica

Hardware para deep learning



NVIDIA DGX-1 deep learning supercomputer

\$ 129 000

8x GPUs NVIDIA Tesla P100, 28672 CUDA cores



Tesla P100: 21TFLOPS

- GPU: 15.3B 16nm FinFET transistors @ 610mm²

- GPU+interposer+HMB2 memory: 150B transistors !!!

DGX-1: 8xP100, 512 GB DDR4, 4x1.92TB SSD, 170 TFLOPS, 60kg, 3200W

A \$2 Billion Chip to Accelerate Artificial Intelligence, MIT Technology Review, April 2016

<https://www.technologyreview.com/s/601195/a-2-billion-chip-to-accelerate-artificial-intelligence/>

<http://www.nvidia.com/object/deep-learning-system.html>



En la práctica

Hardware para deep learning

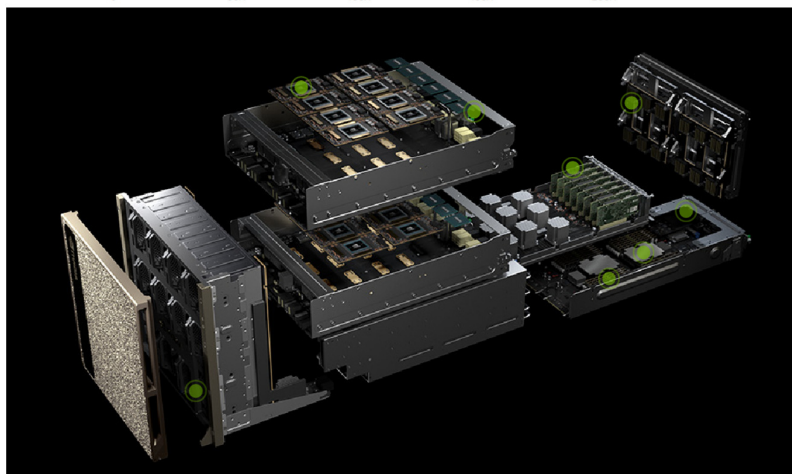
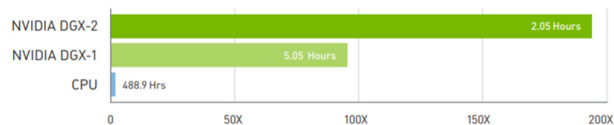


NVIDIA DGX-2 deep learning supercomputer

\$ 399 000

16x GPUs NVIDIA Tesla V100, 81920 CUDA cores, 2 petaFLOPS

NVIDIA DGX-2 Delivers 195X Faster Deep Learning Training



DGX-2: 16x V100, 96MB SRAM, 512 GB DDR, 30TB SSD, 2 petaFLOPS, 10kW, 163kg

<https://www.nvidia.com/en-us/data-center/dgx-2/>



En la práctica

Hardware para deep learning

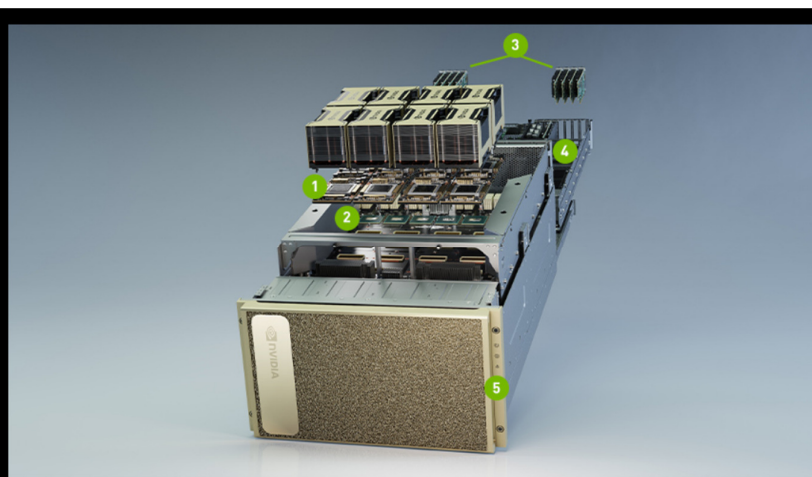


NVIDIA DGX A100

\$ 199 000

8x NVIDIA Ampere A100 Tensor Core GPUs, **5 petaFLOPS**

- 1 8x NVIDIA A100 GPUS WITH 320 GB TOTAL GPU MEMORY
12 NVLinks/GPU, 600 GB/s GPU-to-GPU Bi-directional Bandwidth
- 2 6x NVIDIA NVSWITCHES
4.8 TB/s Bi-directional Bandwidth, 2X More than Previous Generation NVSwitch
- 3 9x MELLANOX CONNECTX-6 200Gb/s NETWORK INTERFACE
450 GB/s Peak Bi-directional Bandwidth
- 4 DUAL 64-CORE AMD CPUs AND 1 TB SYSTEM MEMORY
3.2X More Cores to Power the Most Intensive AI Jobs
- 5 15 TB GEN4 NVME SSD
25GB/s Peak Bandwidth, 2X Faster than Gen3 NVME SSDs



A100 accelerator (May 2020): TSMC 7nm N7, 54.2B transistors, 6912 CUDA cores, 400W

DGX A100: 8x A100, 320GB GPU memory, 1TB system, 15TB SSD, 5 petaFLOPS, 6.5kW, 123kg

<https://www.nvidia.com/en-us/data-center/dgx-a100/>





DSAs [Domain-Specific Architectures]

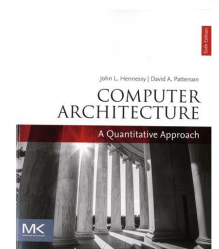
- **ASICs** [Application-Specific Integrated Circuits]
 - Mayor coste NRE [Non recurrent engineering]
 - Mayor rendimiento
- Circuitos reconfigurables, p.ej. **FPGAs** [Field-Programmable Gate Arrays]
 - Menor coste NRE
 - Menor rendimiento



DSAs [Domain-Specific Architectures]

Diseño energéticamente eficiente

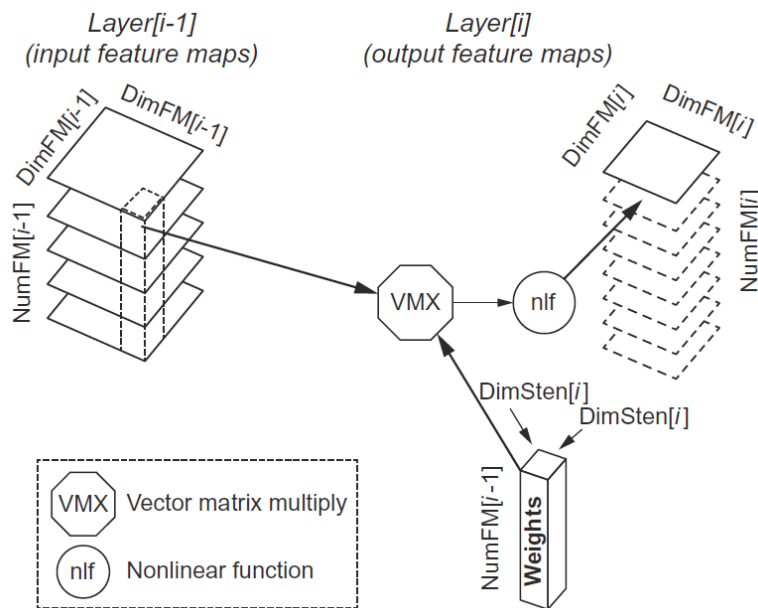
- Memorias dedicadas [scratchpad] gestionadas por el programador en lugar de memorias caché.
- Recursos ahorrados en la microarquitectura dedicados a más unidades funcionales o más memoria.
- Paralelismo ajustado al dominio de aplicación (SIMD).
- Reducción de la precisión (8-, 16-bit), para aprovechar mejor el ancho de banda de memoria.
- DSL [domain-specific language] para portar código a la DSA, p.ej. Halide (visión) o TensorFlow (DNNs).





DSAs [Domain-Specific Architectures]

Ejemplo: CNN



Intensidad aritmética:

operations/weight = 2 x DimFM[i]²



Microsoft Research Catapult

FPGAs [Field Programmable Gate Arrays]

- Menor consumo de energía: 25W
- Menor ancho de banda: 11GB/s
(datos en la memoria DDR3 de la propia FPGA para evitar PCIe)



CPU-FPGA minimalista:
FPGA Altera Stratix V @ Open CloudServer



- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

Air flow
200 LFM
68 °C Inlet

Toward Accelerating Deep Learning at Scale Using Specialized Logic

HOTCHIPS'2015: A Symposium on High Performance Chips, August 2015

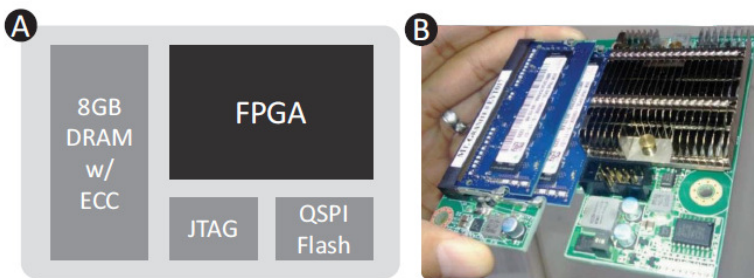
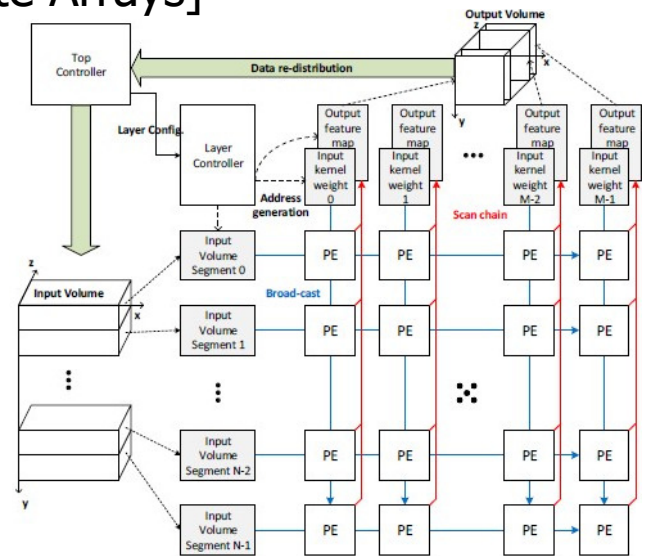




Microsoft Catapult v1

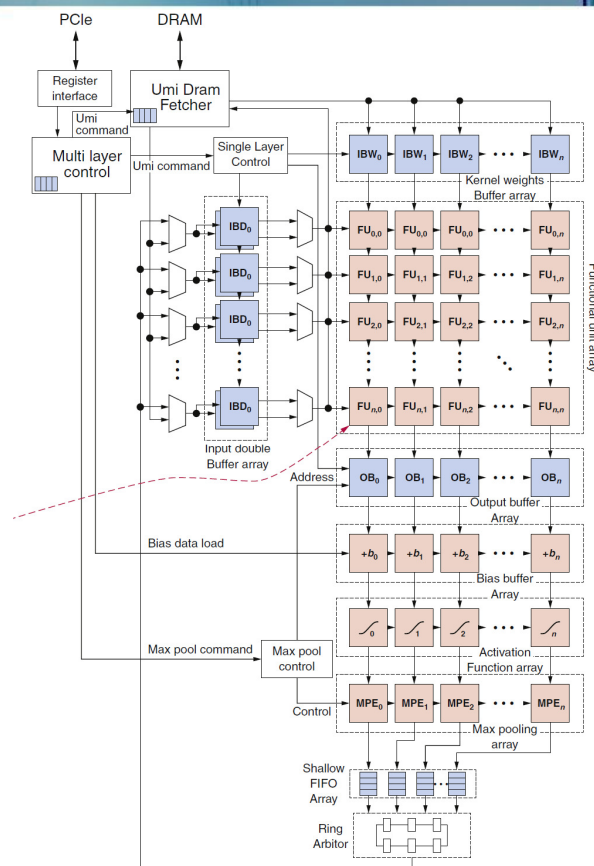
FPGAs [Field Programmable Gate Arrays]

- 3926 18-bit ALUs
- 2D array of PEs
- Red 20Gbps entre 48 FPGAs (topología de toro 6x8)



Microsoft Catapult v1

CNN Accelerator



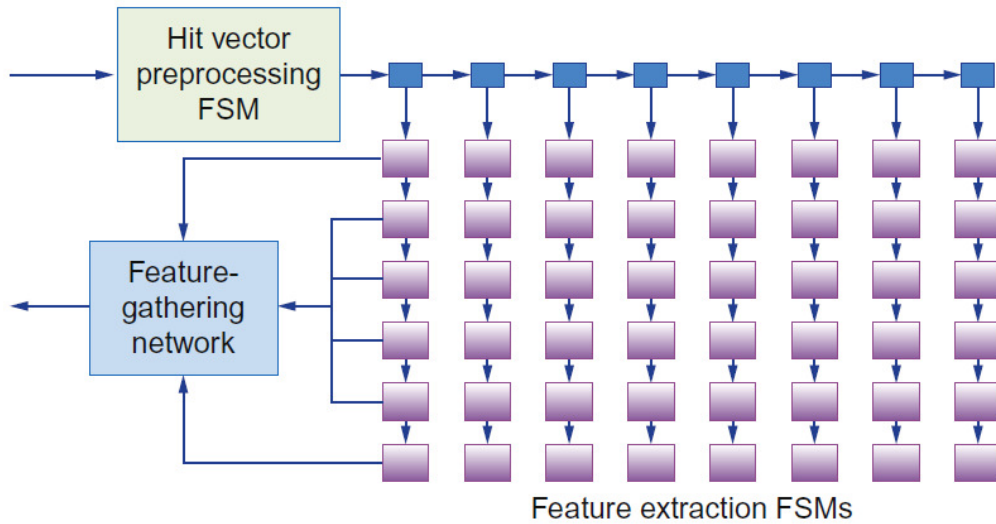
FU [Functional Unit]
= ALU + Registers





Microsoft Catapult v1

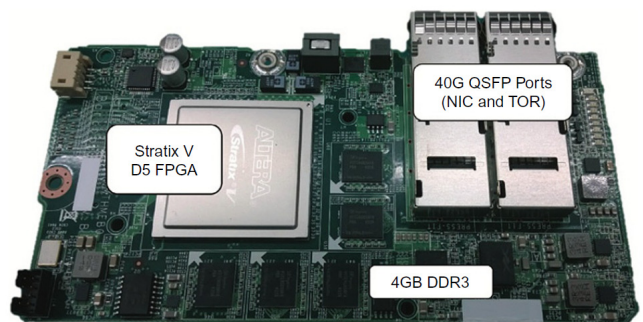
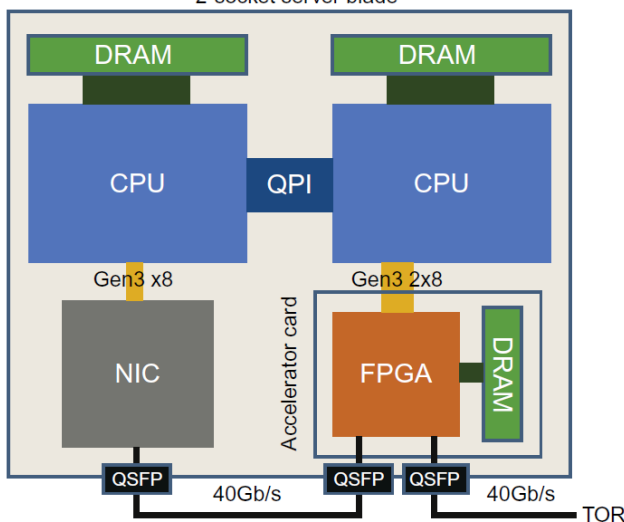
Search engine ranking @ Bing



Microsoft Catapult v2

Cambio en la posición de la FPGA en el centro de datos (entre la CPU y el NIC de cada servidor, 32W)

2-socket server blade



Usos: CNN, Bing & Azure Networking

A cloud-scale acceleration architecture. MICRO Conference, 2016

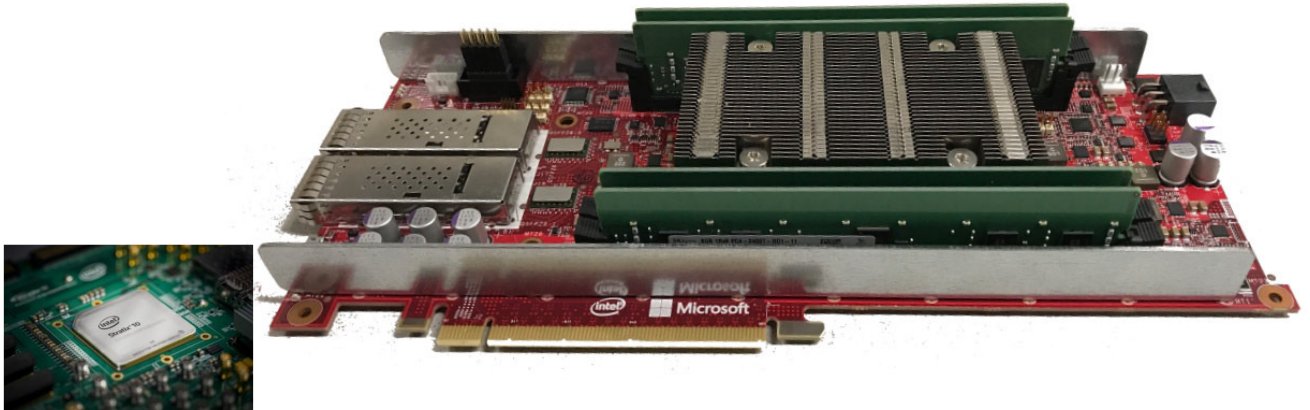




Microsoft Research Project BrainWave

FPGAs [Field Programmable Gate Arrays]

- DNNs as "hardware microservices"



FPGA Intel Stratix 10 (e.g. GRU @ 39.5TFLOPS)

Accelerating Persistent Neural Networks at Datacenter Scale

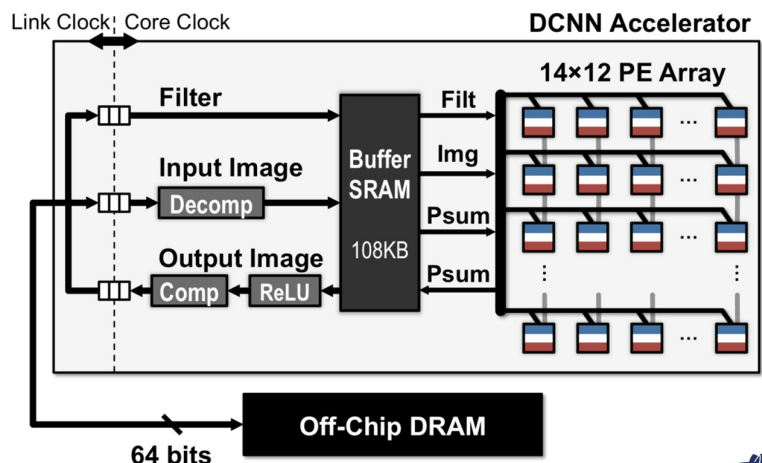
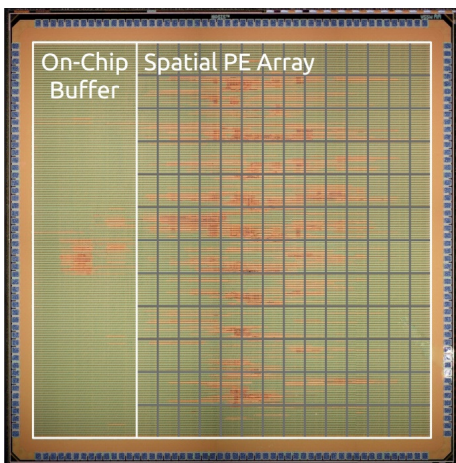
HOTCHIPS'2017 & NIPS'2017



MIT Eyeriss

168 PE [Processing Elements], 0.3W (<10% mobile GPU)

<http://www.mit.edu/~sze/eyeriss.html>



A Deep Learning AI Chip for Your Phone, IEEE Spectrum, February 2016

<http://spectrum.ieee.org/tech-talk/semiconductors/processors/a-deep-learning-ai-chip-for-your-phone>





TPU [Tensor Processing Unit]

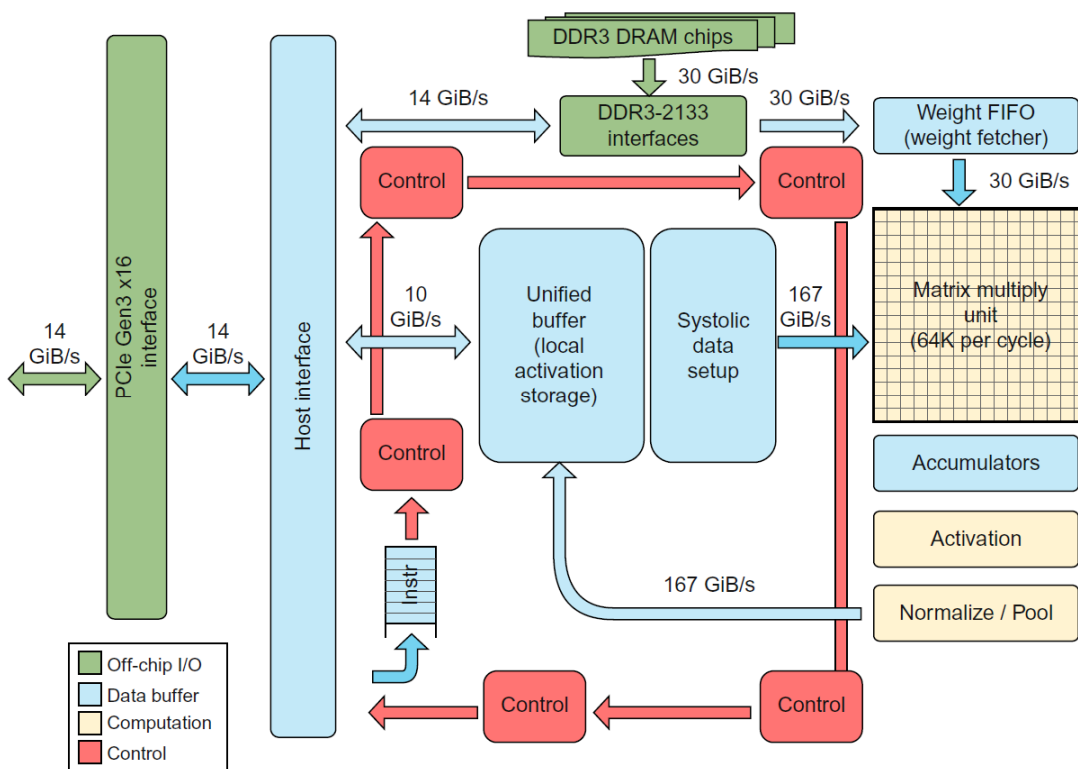
Google, 2015-

Características

- 65,536 (256x256) **8-bit** ALU Matrix Multiply Unit
- Large software-managed on-chip memory
- Single-threaded, deterministic execution model
- Coprocessor on the PCIe I/O bus
- The host server sends instructions over the PCIe bus directly to the TPU for it to execute, rather than having the TPU fetch the instructions (closer to FPU's than to GPU's).



TPU [Tensor Processing Unit]





TPU [Tensor Processing Unit]

CISC ISA (due to slow PCIe bus)

- **Read_Host_Memory**
(CPU Host Memory -> Unified Buffer)
- **Read_Weights**
(Weight Memory -> Weight Buffer)
- **MatrixMultiply / Convolve**
(Unified Buffer -> Accumulators)
- **Activate**
(Accumulators -> Outout Buffer)
- **Write_Host_Memory**
(Unified Buffer -> CPU Host Memory)

Clock cycles per instruction (CPI) \sim 10-20



TPU Microarchitecture

- Goal:

Keep the Matrix Multiply Unit busy.

- Plan:

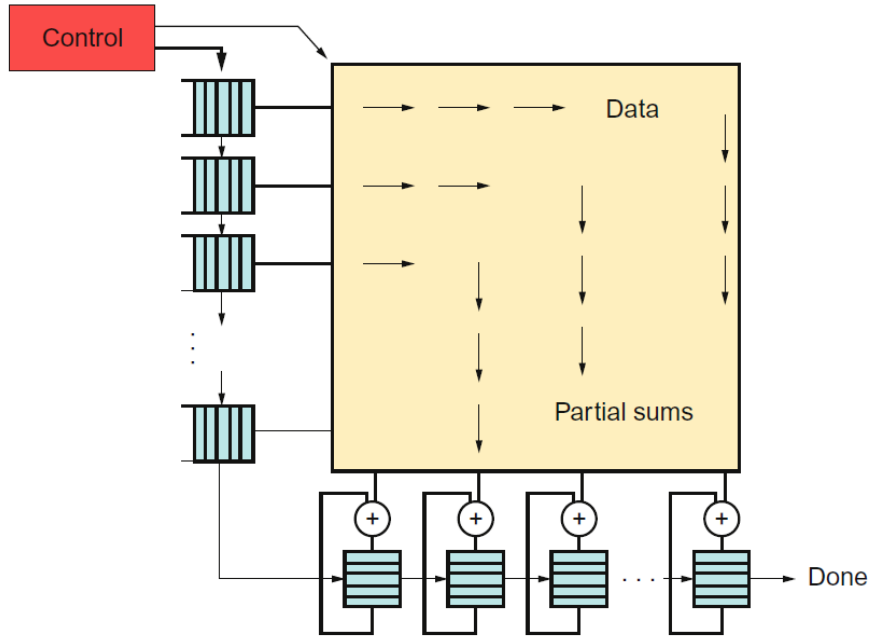
Hide the execution of the other instructions by overlapping their execution with the MatrixMultiply instruction



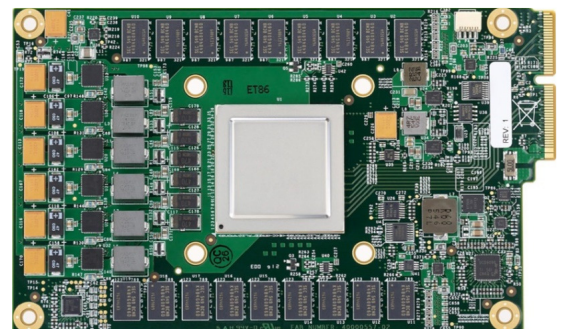
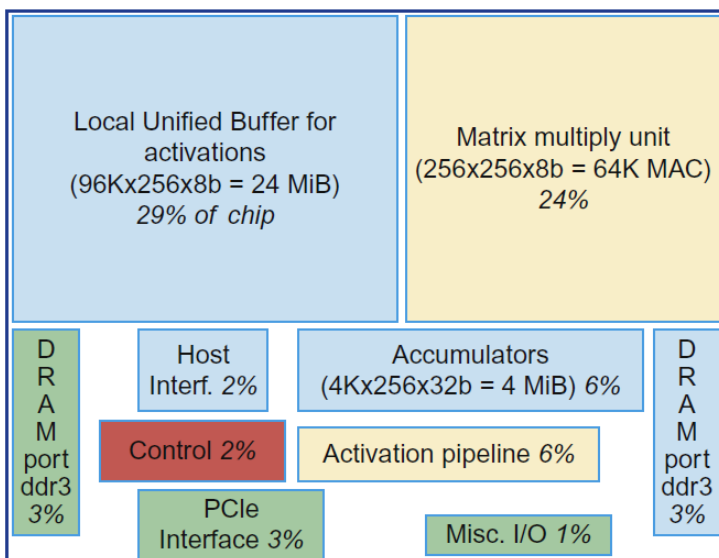


TPU Microarchitecture

The Matrix Multiply Unit uses systolic execution to save energy by reducing reads and writes of the Unified Buffer



TPU Implementation



TPU die floor plan





TPU Programming

- The portion of the application run on the TPU is typically written using TensorFlow and is compiled into an API that can run on GPUs or TPUs.

TPU Software

- Lightweight kernel driver:
Memory management and interrupts.
- User space driver:
Sets up and controls TPU execution, reformats data into TPU order, and translates API calls into TPU instructions and turns them into an application binary

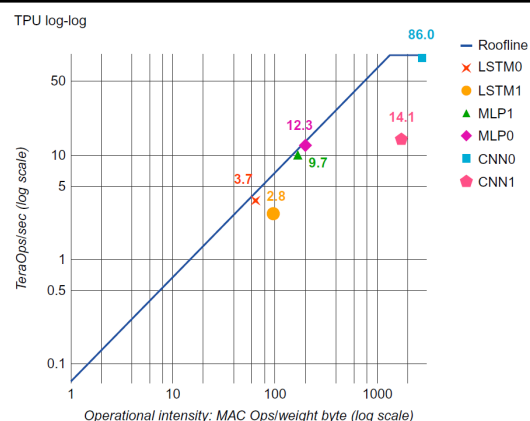


TPU Applications

Name	LOC	DNN layers					Weights	TPU Ops/Weight	% deployed TPUs 2016
		FC	Conv	Element	Pool	Total			
MLP0	100	5				5	20M	200	61%
MLP1	1000	4				4	5M	168	
LSTM0	1000	24		34		58	52M	64	29%
LSTM1	1500	37		19		56	34M	96	
CNN0	1000		16			16	8M	2888	5%
CNN1	1000	4	72		13	89	100M	1750	

6 applications, 95% workload @ Google

- MLP: RankBrain
- LSTM: Google Neural Translator
- CNN: DeepMind AlphaGo





TPU Terminology

Accumulators	—	The 4096 256×32 -bit registers (4 MiB) that collect the output of the MMU and are input to the Activation Unit
Activation unit	—	Performs the nonlinear functions (ReLU, sigmoid, hyperbolic tangent, max pool, and average pool). Its input comes from the Accumulators and its output goes to the Unified Buffer
Matrix multiply unit	MMU	A systolic array of 256×256 8-bit arithmetic units that perform multiply-add. Its inputs are the Weight Memory and the Unified Buffer, and its output is the Accumulators
Systolic array	—	An array of processing units that in lockstep input data from upstream neighbors, compute partial results, and pass some inputs and results to downstream neighbors
Unified buffer	UB	A 24 MiB on-chip memory that holds the activations. It was sized to try to avoid spilling activations to DRAM when running a DNN
Weight memory	—	An 8 MiB external DRAM chip containing the weights for the MMU. Weights are transferred to a <i>Weight FIFO</i> before entering the MMU



CPU Intel i7	GPU NVIDIA	FPGA Catapult	ASIC TPU
SIMD extensions (MMX, SSE, AVX)	Streaming multiprocessors	3926 18-bit PEs (reconfigurable)	256x256 Matrix Multiply Unit
1D SIMD		2D SIMD	
Multithreading		Pipelined systolic array	
Out-of-order execution	Multiprocessing	Programmable controller	Simple execution of instructions
32 & 64-bit FP	32 & 64-bit FP	18-bit integers	8-bit integers
x86 ISA (C, Java, Python...)	PTX (CUDA, OpenCL)	RTL (Verilog, VHDL)	Reduced CISC ISA (API via DSLs: Halide, TensorFlow)





Intel Nervana Neural Network Processor (NNP)

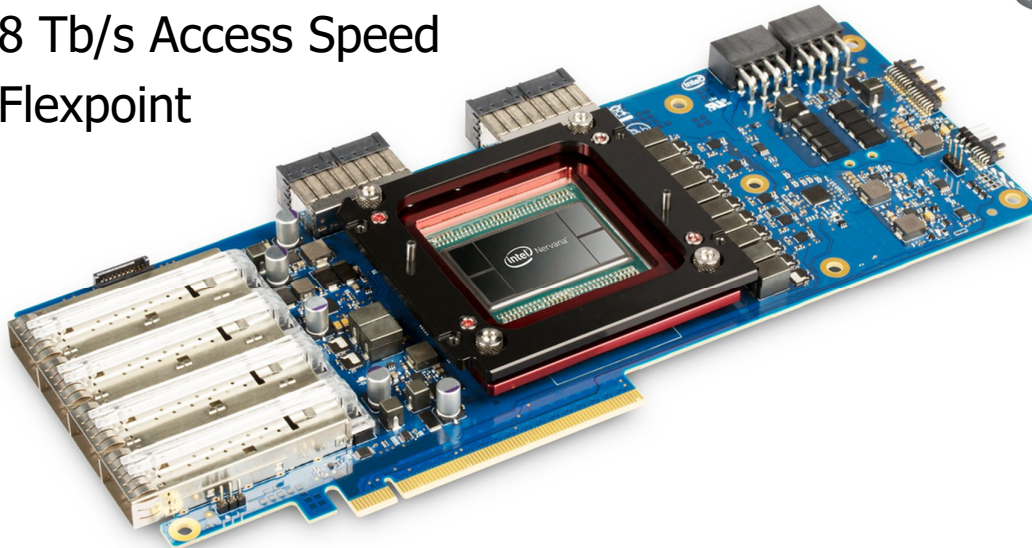
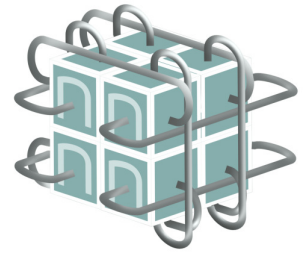
“Lake Crest”

32 GB HBM2,

1 TB/s Bandwidth

8 Tb/s Access Speed

Flexpoint



Intel Nervana Neural Network Processor (NNP)

“Lake Crest”

LAKE CREST DEEP LEARNING ARCHITECTURE

- Tensor based architecture
- Flexpoint®
 - Unprecedented levels of parallelism up to 10x of state-of-art
 - Low power per tensor operation
- HBM2 memory: up to 12x faster than DDR4
- Proprietary inter-chip links: up to 20x faster than PCIe

The diagram illustrates the Lake Crest Deep Learning Architecture. It features a central Interposer connecting various components. On the left and right sides, there are 8GB HBM2 memory blocks. The central area contains a 3x3 grid of Processing Clusters. Each cluster includes HBM PHY, Mem Ctrlr, and an MGMT CPU. The Interposer also manages SPI, IC2, GPIO, and PCIe Controller & DMA (PCI Express x16). Interconnects (ICC) are shown connecting the clusters and memory blocks.






Intel Nervana Neural Network Processor (NNP)

“Lake Crest”

DESIGNED FOR DEEP LEARNING WORKLOADS

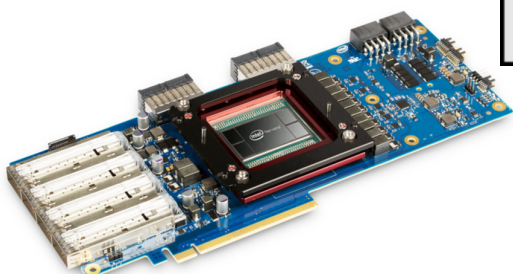
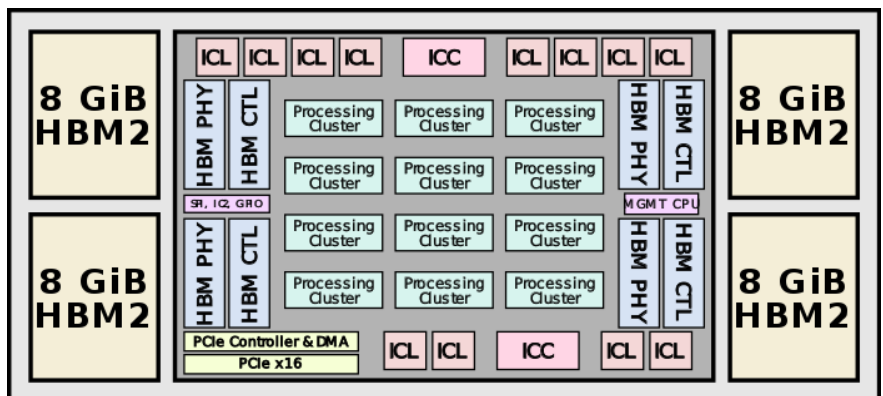
- Supports true model parallelism
 - Each compute node has own memory interface
 - Model size is less limited
 - Memory I/O increased
- New model exploration




Intel Nervana Neural Network Processor (NNP)

“Lake Crest”, 1st generation NNP, 2016

nervana
SYSTEMS



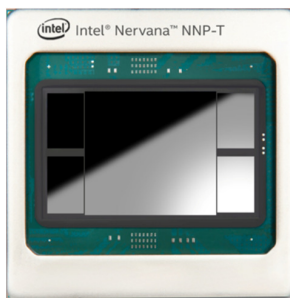
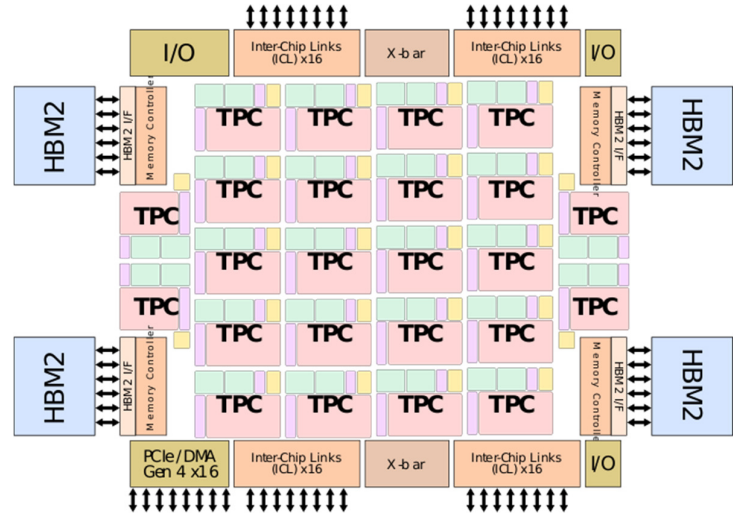
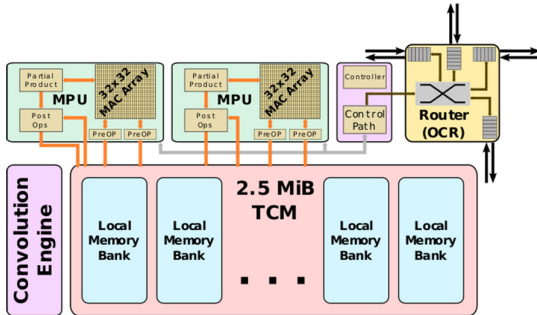
Lake Crest Accelerator PCIe card





Intel Nervana Neural Network Processor (NNP)

"Spring Crest", 2nd generation NNP, 2019



NNP T-1400, 375W, 108TFLOPS



Hardware especializado



Intel Nervana Neural Network Processor (NNP)

Febrero 2020:

"Intel Axes Nervana Just Two Months After Launch"



Habana HLS-1 (8x HL-205)



habana
GOYA™



En la práctica Hardware para deep learning



Intel + Habana Labs

\$2bn

News Release

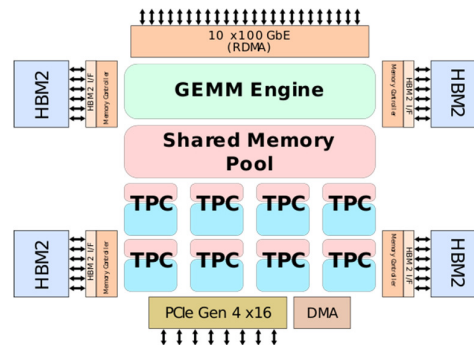
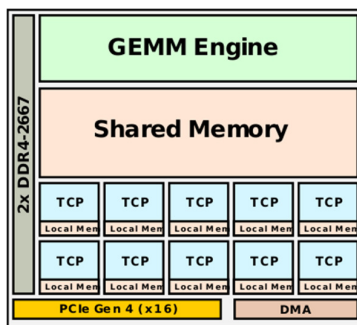
December 16, 2019

Contact Intel PR

Intel Acquires Artificial Intelligence Chipmaker Habana Labs

Combination Advances Intel's AI Strategy, Strengthens Portfolio of AI Accelerators for the Data Center

SANTA CLARA Calif., Dec. 16, 2019 – Intel Corporation today announced that it has acquired Habana Labs, an Israel-based developer of programmable deep learning accelerators for the data center for approximately \$2 billion. The combination strengthens Intel's artificial intelligence (AI) portfolio and accelerates its efforts in the nascent, fast-growing AI silicon market, which Intel expects to be greater than \$25 billion by 2024¹.



En la práctica Hardware para deep learning



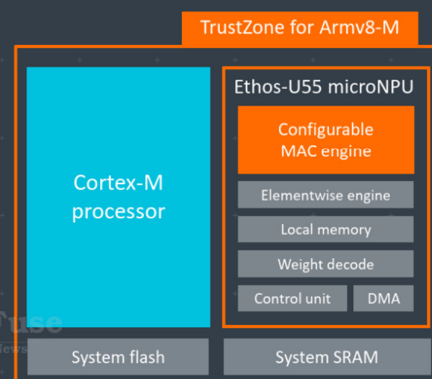
ARM Ethos microNPU

Cortex M55 CPU + Ethos U55 NPU (2020)



Ethos-U55: The First microNPU for Cortex-M

- ✓ Highest efficiency and small memory footprint
- ✓ 32, 64, 128, or 256 unit multiply-accumulate (MAC) engine
- ✓ Weight decoder and DMA for on-the-fly weight decompression
- ✓ Tooling available for offline optimization
- ✓ Works with a range of Cortex-M processors:
 - Cortex-M55
 - Cortex-M7
 - Cortex-M33
 - Cortex-M4



WikiChip Fuse
Chips & Semi News



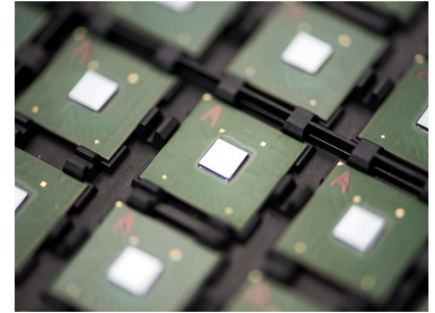


Implementaciones analógicas

... para reducir el consumo energético

Trillion operations per watt (TOPS/W)

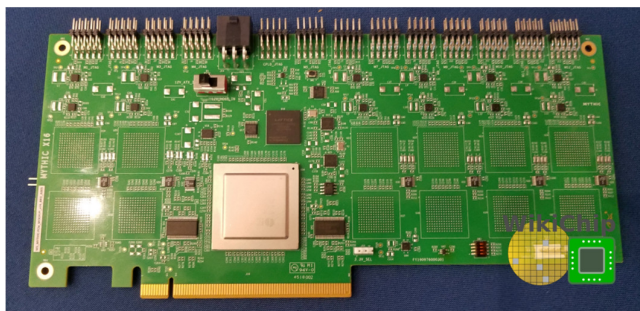
- 0.4 TOPS/W: NVIDIA Volta V100 GPU
- 4 TOPS/W: **Mythic AI** analog in-memory computing, (analog flash array + programmable digital circuit) <https://www.mythic-ai.com/>
- 20 TOPS/W: **Syntiant** Neural Decision Processor (NDP), (entire analog network) <https://www.syntiant.com/>



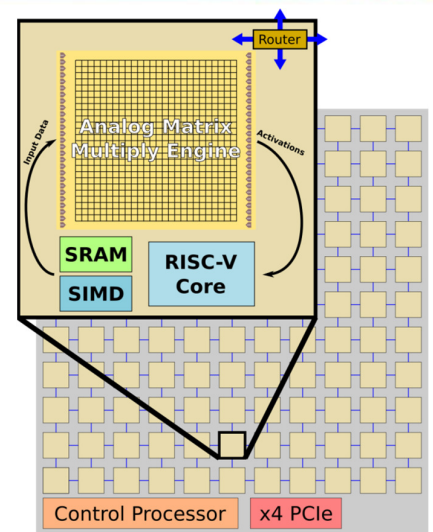
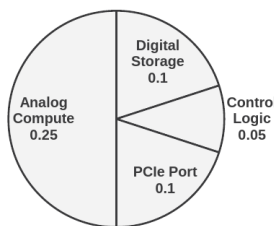
Two Startups Use Processing in Flash Memory for AI at the Edge, IEEE Spectrum, August 2018 <https://spectrum.ieee.org/tech-talk/computing/embedded-systems/two-startups-use-processing-in-flash-memory-for-ai-at-the-edge>



Mythic AI IPU



Energy (pJ/MAC)
Total = 0.5



Mythic Initial Product Lineup

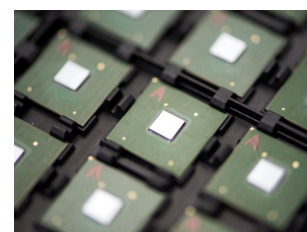
Up to 120M Weights

Mythic IPU x1 (M.2 x4 2280)

Mythic IPU x4 (PCIe x4 HHHL)

Mythic IPU x16 (PCIe x16 FHFL)

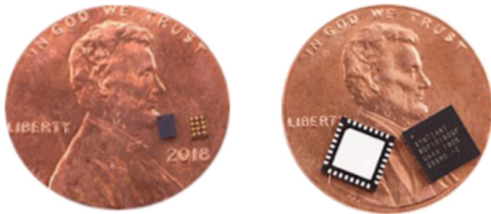
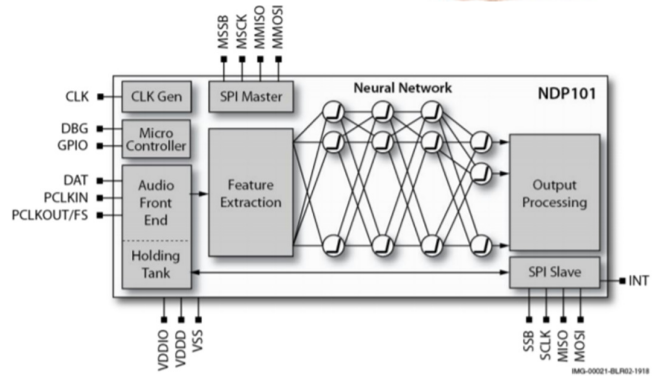
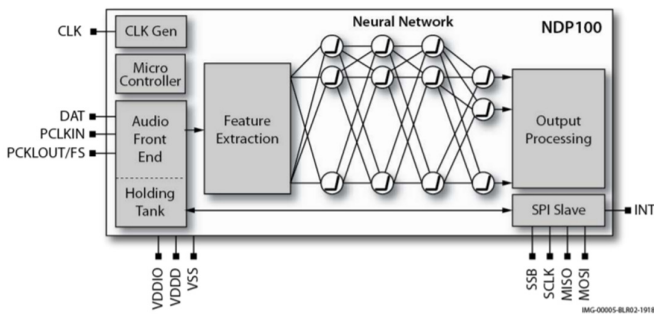
MYTHIC WikiChip



En la práctica Hardware para deep learning

Syntiant NDP [Neural Decision Processor]

<140μW

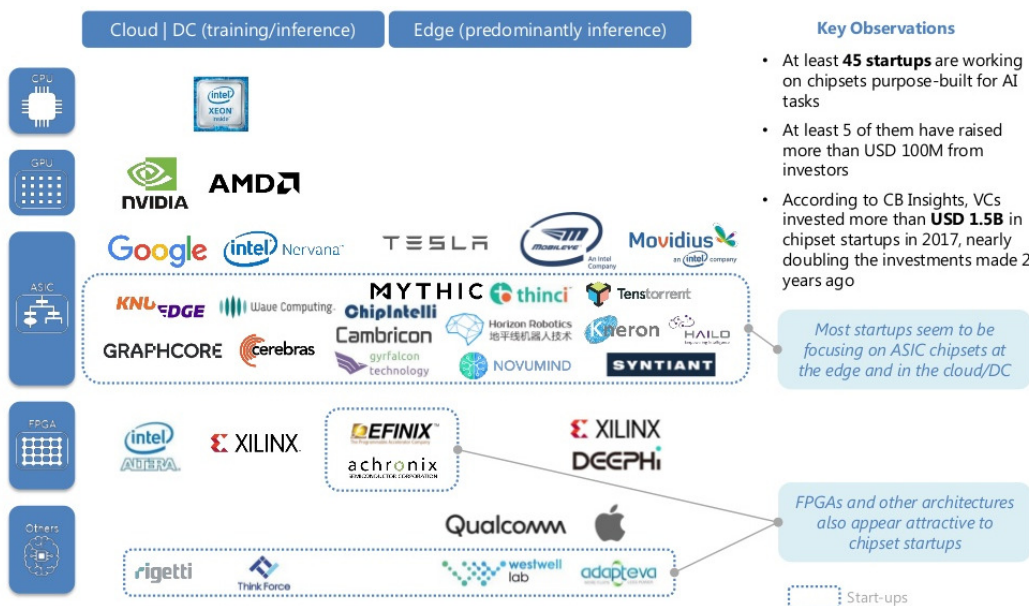


Syntiant NDP100 & NDP101

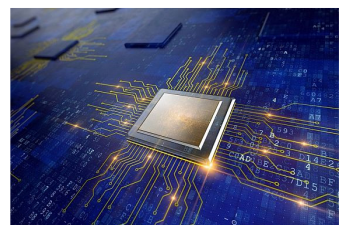


En la práctica Hardware para deep learning

Docenas de empresas desarrollan chips para IA...



- Key Observations**
- At least **45 startups** are working on chipsets purpose-built for AI tasks
 - At least 5 of them have raised more than USD 100M from investors
 - According to CB Insights, VCs invested more than **USD 1.5B** in chipset startups in 2017, nearly doubling the investments made 2 years ago



Most startups seem to be focusing on ASIC chipsets at the edge and in the cloud/DC

FPGAs and other architectures also appear attractive to chipset startups

Start-ups

Note: Several startups are in stealth mode and company information may not be publicly available.



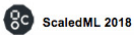
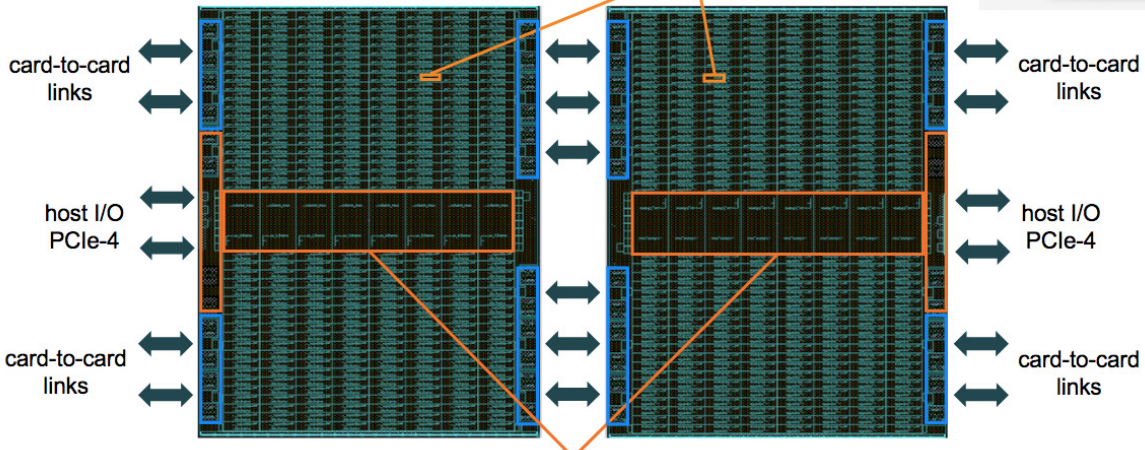


Graphcore Colossus IPU, UK



"Colossus" IPU pair
(300W PCIe card)

2432 processor tiles >200Tflop_{16.32} ~600MB



15



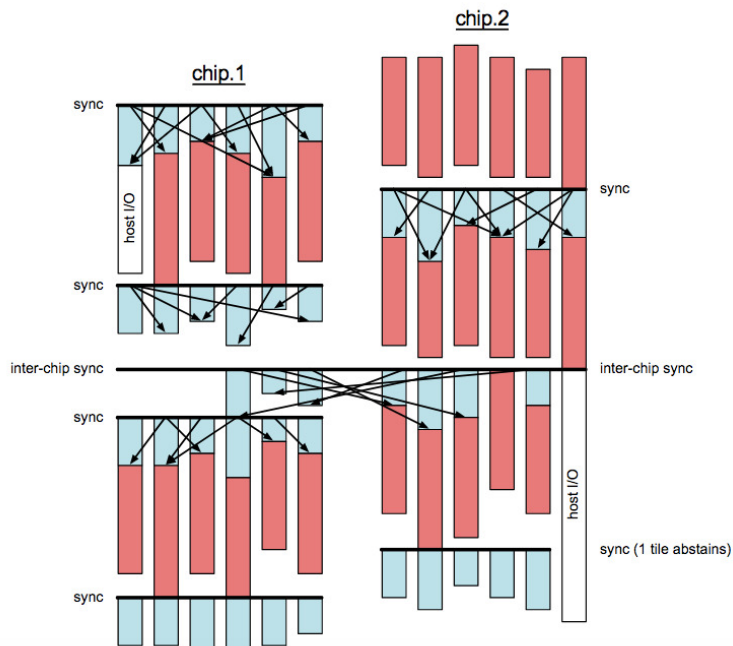
IPU [Intelligent Processing Unit] pair



Graphcore Colossus IPU, UK

Massively parallel computing
with no concurrency hazards

- exchange phase
- compute phase



BSP [Bulk Synchronous Parallel] model

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



En la práctica

Hardware para deep learning



Graphcore Colossus IPU, UK



1 petaFLOP

IPU M2000 machine



Dell DSS8440



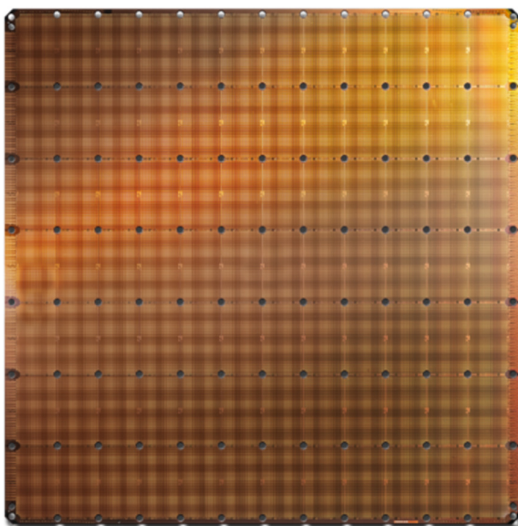
En la práctica

Hardware para deep learning



Cerebras WSE [Wafer-Scale Engine]

>400000 AI cores, 18GB SRAM, 17.5kW



Cerebras WSE
1.2 Trillion transistors
46,225 mm² silicon



Largest GPU
21.1 Billion transistors
815 mm² silicon

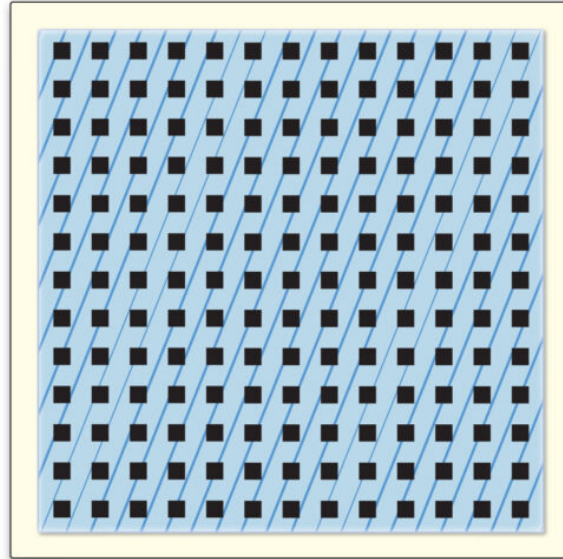
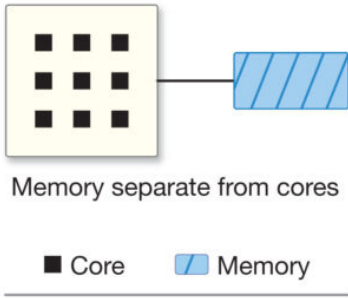




Cerebras WSE [Wafer-Scale Engine]

Cerebras Memory Architecture

Traditional Memory Architecture

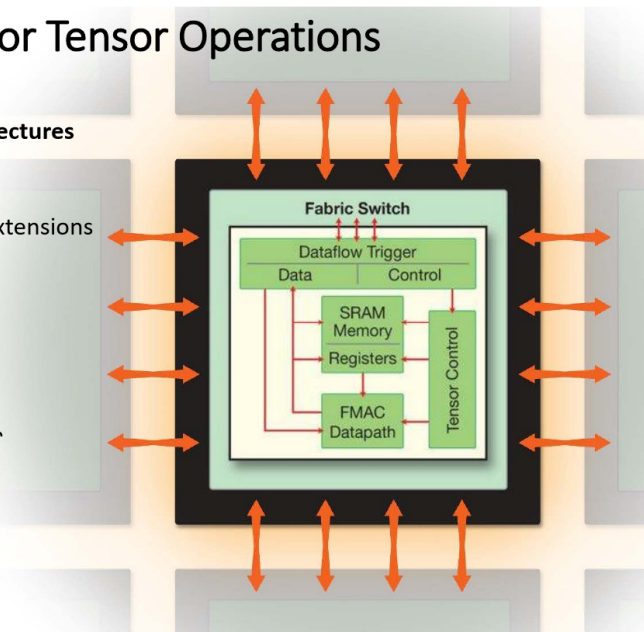


Cerebras WSE [Wafer-Scale Engine]

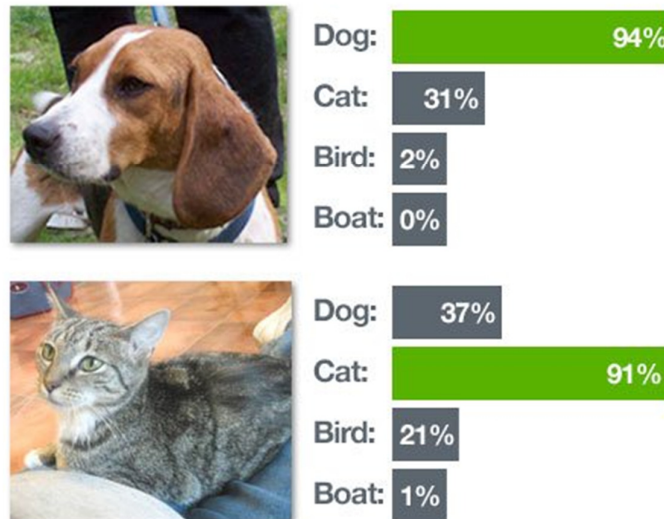
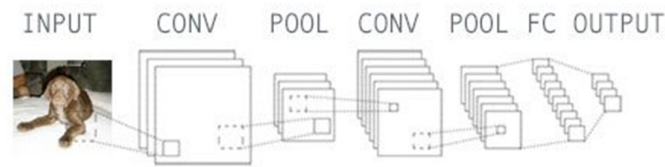
Flexible Cores Optimized for Tensor Operations

Key to supporting rapidly evolving NN architectures

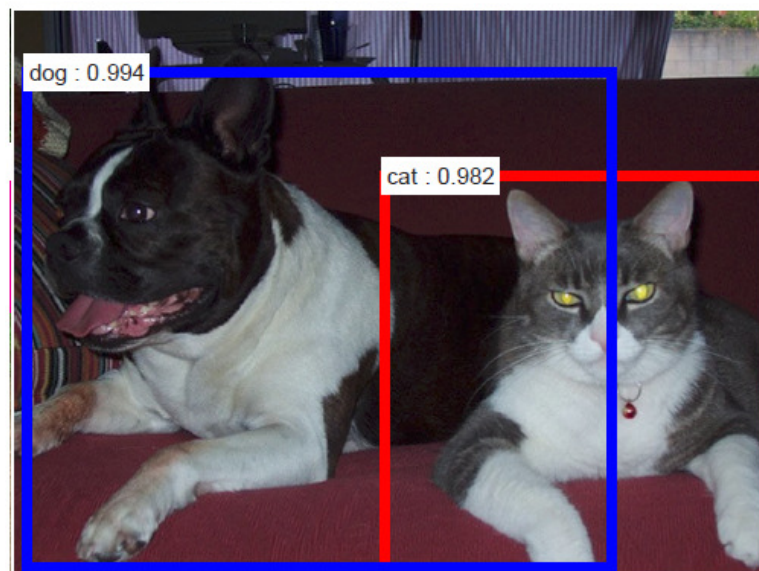
- Fully programmable compute core
- Full array of general instructions with ML extensions
- Flexible **general ops** for control processing
 - e.g. arithmetic, logical, load/store, branch
- Optimized **tensor ops** for data processing
 - Tensors as first class operands
 - e.g. $fmac [z] = [z], [w], a$
 3D 3D 2D scalar



Clasificación de imágenes



Detección de objetos





Detección de objetos



CVPR'2018



Segmentación de imágenes



Aplicaciones del Deep Learning

Vehículos autónomos



Aplicaciones del Deep Learning

Vehículos autónomos

2005 DARPA Grand Challenge



Aplicaciones del Deep Learning

Vehículos autónomos



Autonomous Land Vehicle In a Neural Network (ALVINN)

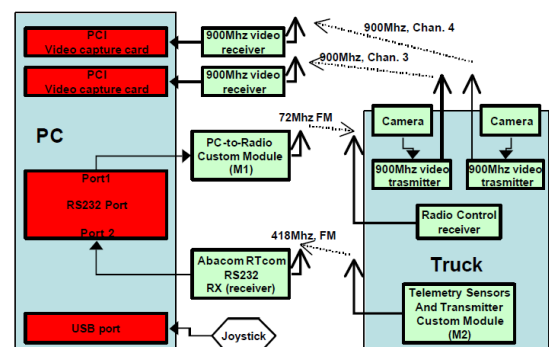
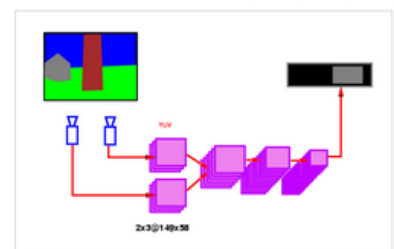
NAVigational LABoratory II (NAVLAB II)

Control de dirección de un vehículo, CMU Ph.D. thesis, 1992



Aplicaciones del Deep Learning

Vehículos autónomos



DAVE, 2004

Autonomous Off-Road Vehicle Control using End-to-End Learning

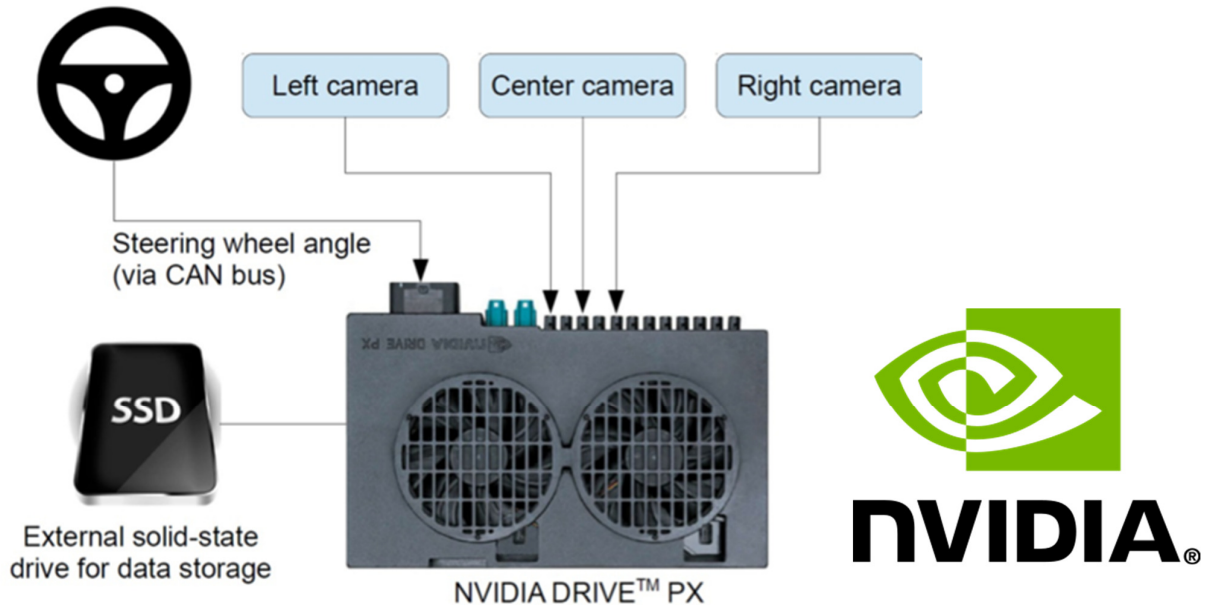
NYU Courant Institute / CBLL [Computational & Biological Learning Lab]

<http://www.cs.nyu.edu/~yann/research/dave/>



Aplicaciones del Deep Learning

Vehículos autónomos



DAVE2, after DARPA Autonomous Vehicle (DAVE) project
NVIDIA, 2016. <http://arxiv.org/abs/1604.07316>

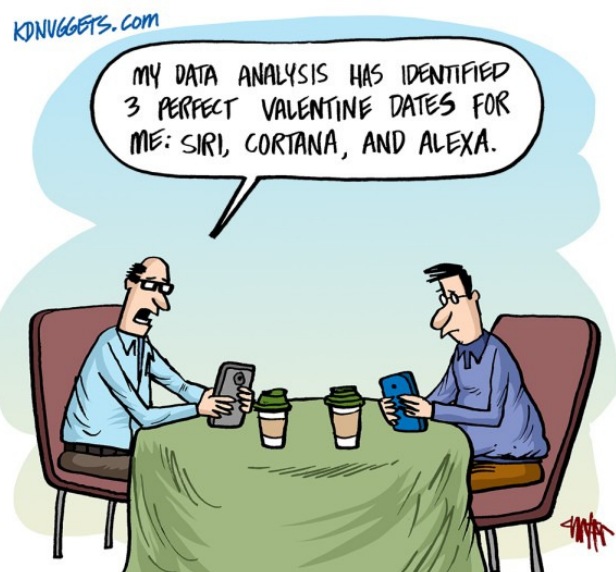
Completamente autónomo con sólo 100 horas de entrenamiento!!!



Aplicaciones del Deep Learning

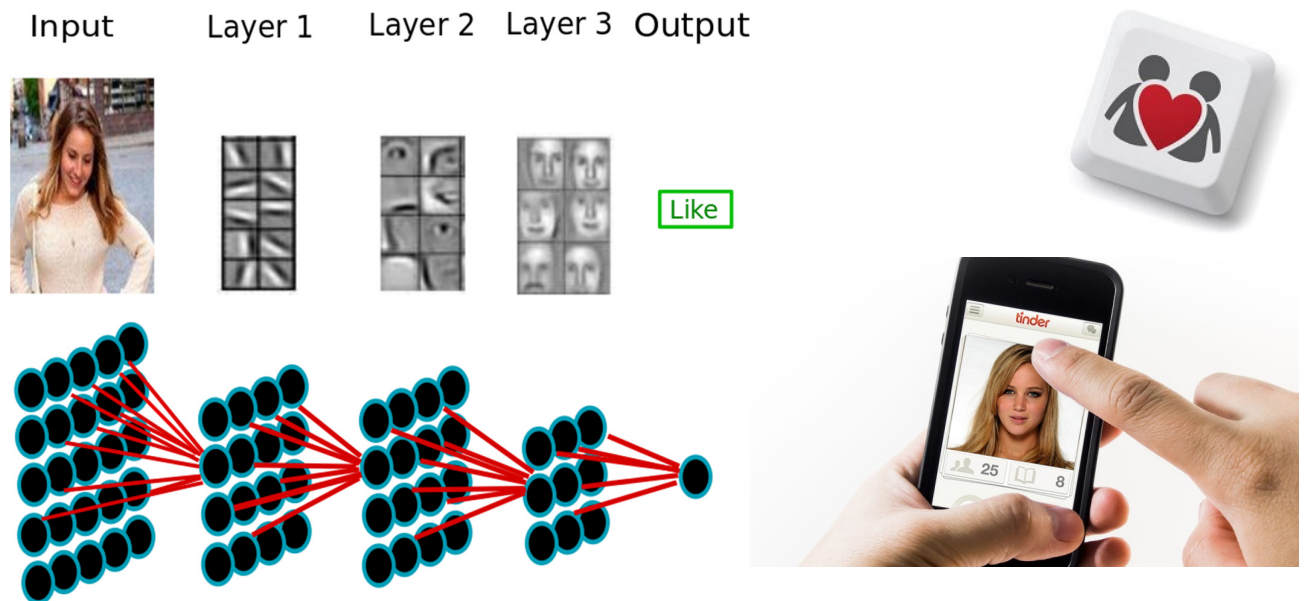


"Sweetheart, my neural net predicts that you and I are 98.9% compatible. Will you be my Valentine?"



Aplicaciones del Deep Learning

Citas por Internet, e.g. Tinder & OKCupid.com



Harm De Vries & Jason Yosinski:

Can deep learning help you find the perfect match?

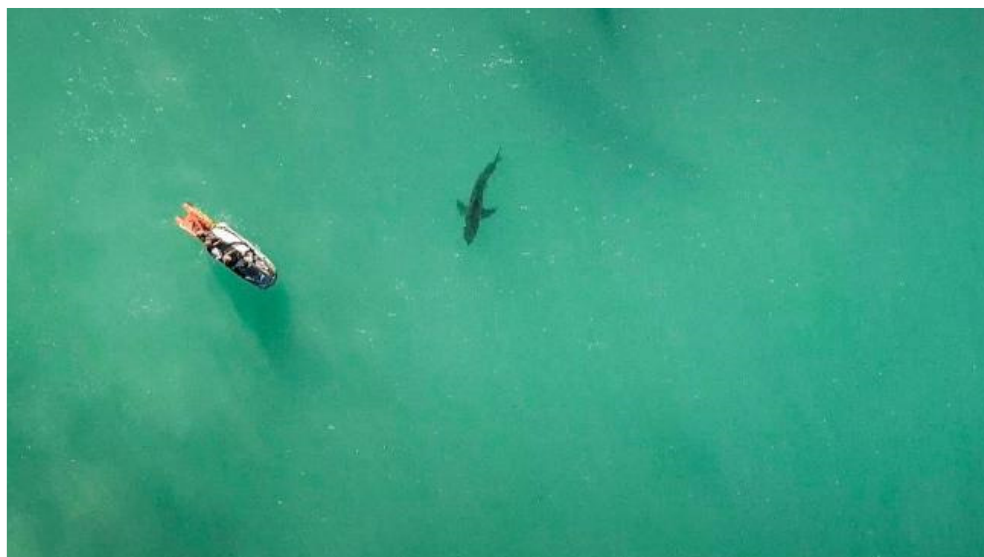
ICML'2015 Deep Learning Workshop



Aplicaciones del Deep Learning

DetECCIÓN de tiburones

SharkSpotter, Australia, 2017



Aplicaciones del Deep Learning

Traducción automática de señales

Google app



Otavio Good (Google Translate):

How Google Translate squeezes deep learning onto a phone

<http://googlresearch.blogspot.com.es/2015/07/how-google-translate-squeezes-deep.html>



Aplicaciones del Deep Learning

Descripción textual de imágenes [image captioning]

<p>A person riding a motorcycle on a dirt road.</p> 	<p>Two dogs play in the grass.</p> 	<p>A skateboarder does a trick on a ramp.</p> 	<p>A dog is jumping to catch a frisbee.</p> 
<p>A group of young people playing a game of frisbee.</p> 	<p>Two hockey players are fighting over the puck.</p> 	<p>A little girl in a pink hat is blowing bubbles.</p> 	<p>A refrigerator filled with lots of food and drinks.</p> 
<p>A herd of elephants walking across a dry grass field.</p> 	<p>A close up of a cat laying on a couch.</p> 	<p>A red motorcycle parked on the side of the road.</p> 	<p>A yellow school bus parked in a parking lot.</p> 
Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image



Aplicaciones del Deep Learning

Descripción textual de imágenes

“La inteligencia visual de un niño de 3 años...”
-- Fei-Fei Li (Stanford)



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with legos toy."



"boy is doing backflip on wakeboard."



a young boy is holding a baseball bat.



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

STANFORD UNIVERSITY

“The search for a thinking machine”

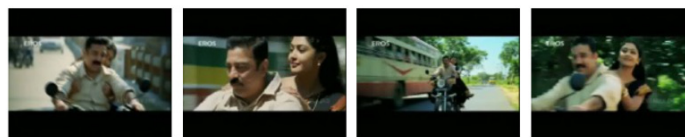
BBC News, 17 September 2015

<http://www.bbc.com/news/technology-32334573>



Aplicaciones del Deep Learning

Descripción textual de vídeos [video clip description]



+Local+Global: A **man** and a **woman** are **talking** on the **road**

Ref: A man and a woman ride a motorcycle



+Local+Global: **Someone** is **frying** a **fish** in a **pot**

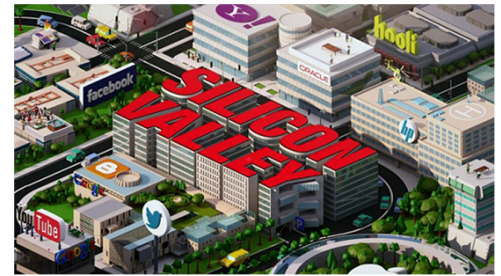
Ref: A woman is frying food

Youtube2Text



Aplicaciones del Deep Learning

Compresión de imágenes



<http://www.piedpiper.com/>

Full Resolution Image Compression with Recurrent Neural Networks
arXiv, August 2016, <http://arxiv.org/abs/1608.05148>



Aplicaciones del Deep Learning

Retoque fotográfico



Figure 6: Original (top) vs. enhanced (bottom) images for iPhone 6, HTC One M9 and Huawei P9 cameras.

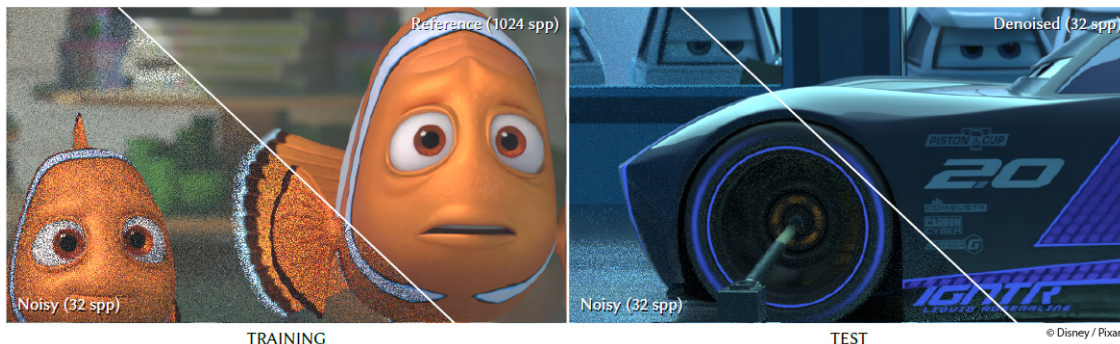
WESPE: Weakly Supervised Photo Enhancer for Digital Cameras. CVPR 2018. <https://arxiv.org/abs/1709.01118>



Aplicaciones del Deep Learning

Síntesis de imágenes

Eliminación de ruido
@ UCSB, Disney & Pixar



SIGGRAPH 2017

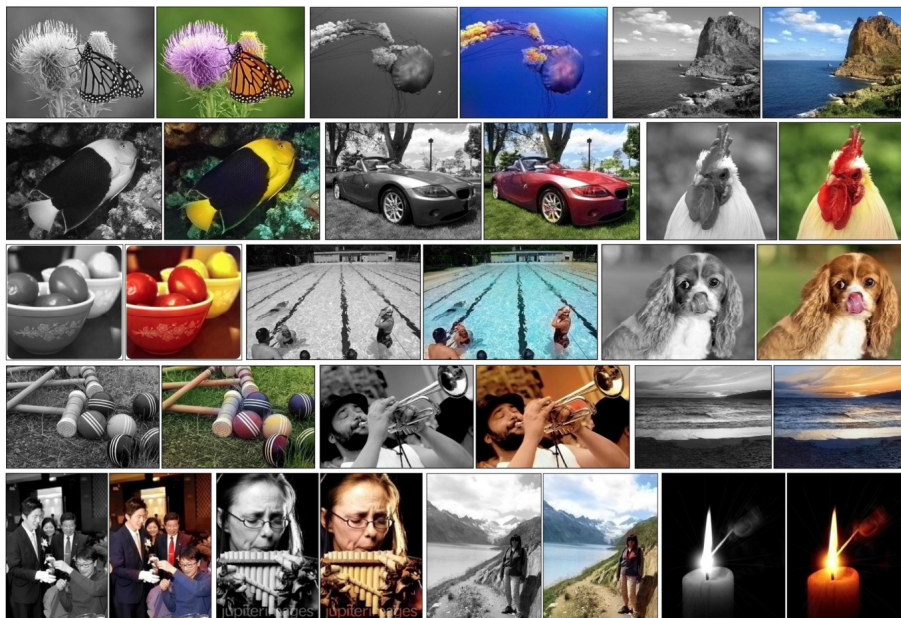
http://cvc.ucsb.edu/graphics/Papers/SIGGRAPH2017_KPCN/



Aplicaciones del Deep Learning

Síntesis de imágenes

Coloreado de fotografías



ECCV 2016 <http://richzhang.github.io/colorization/>



Aplicaciones del Deep Learning

Síntesis de imágenes

Coloreado de fotografías interactivo



SIGGRAPH 2017

<https://richzhang.github.io/ideepcolor/>



Aplicaciones del Deep Learning

Síntesis de imágenes: "Inceptionism"

Usando una red ya entrenada para reconocer objetos...

"El grito"

Edvard Munch

... visto por una red neuronal

<http://deepdreamgenerator.com>

<https://github.com/google/deepdream>



Aplicaciones del Deep Learning

Síntesis de imágenes

Transferencia de estilos



Leon A. Gatys, Alexander S. Ecker & Matthias Bethge:

A Neural Algorithm of Artistic Style

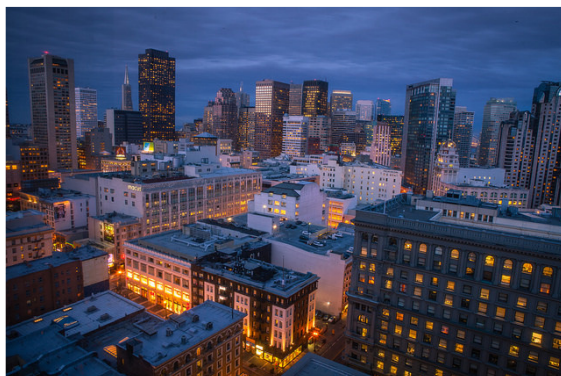
arXiv, 2015. <http://arxiv.org/abs/1508.06576>



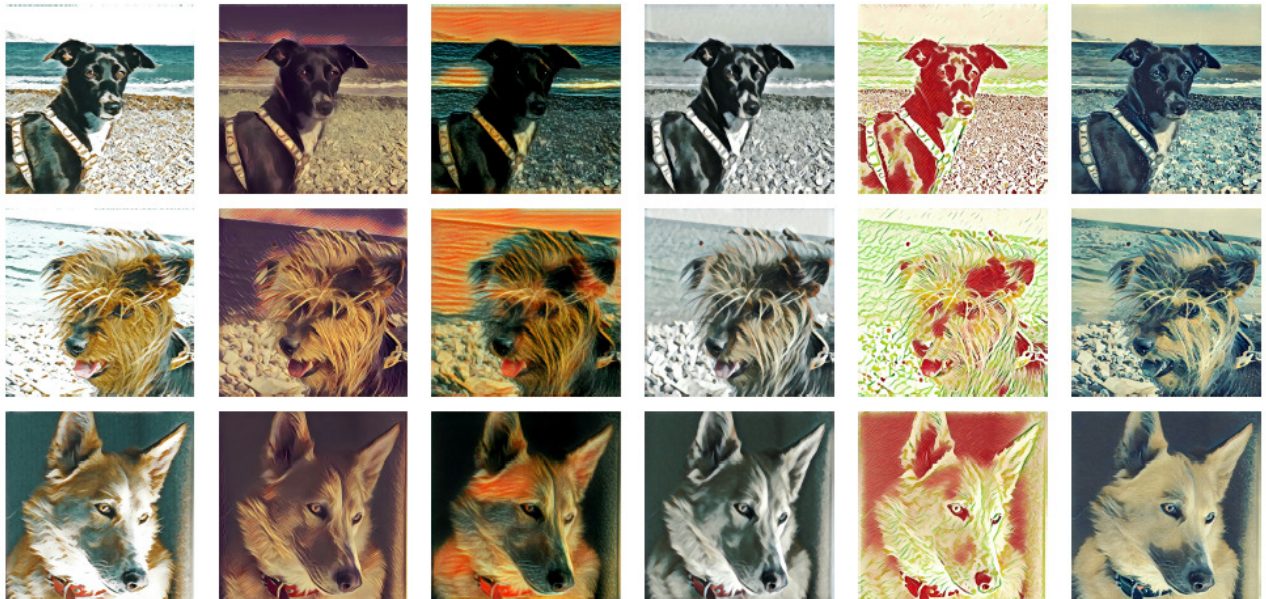
Aplicaciones del Deep Learning

Síntesis de imágenes

Transferencia de estilos



Aplicaciones del Deep Learning



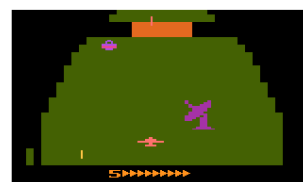
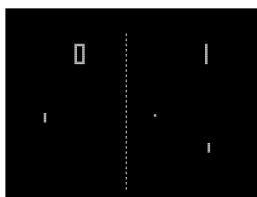
Transferencia de estilos



Aplicaciones del Deep Learning

Videojuegos (Atari 2600)

Google DeepMind



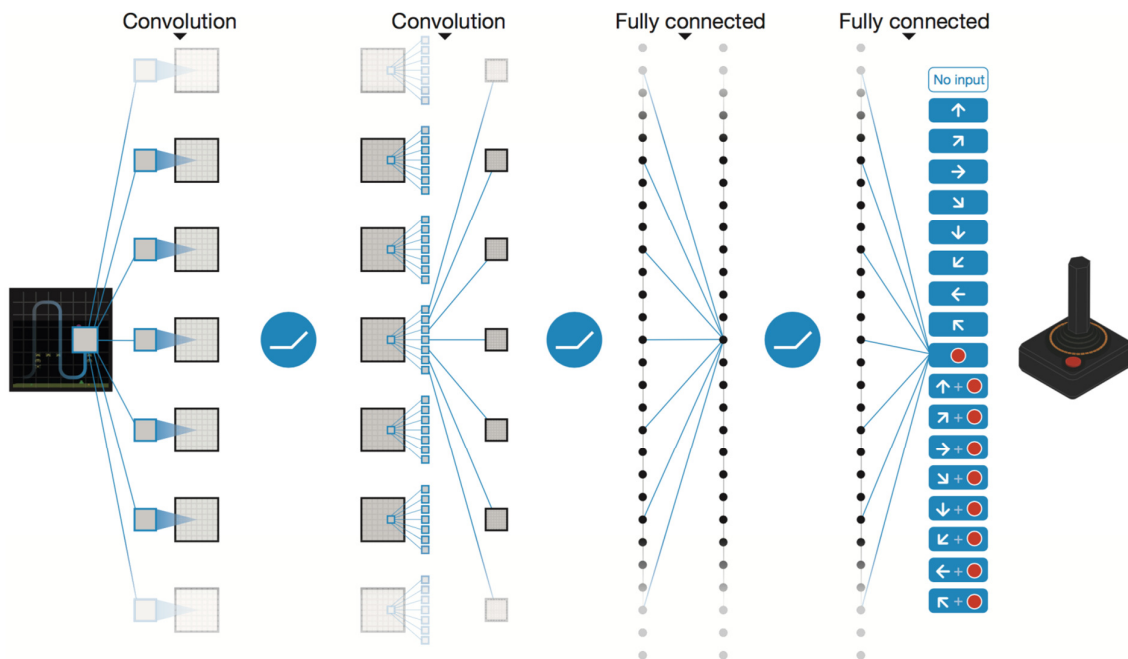
"Google AI beats humans at more classic arcade games than ever before"
<http://arxiv.org/pdf/1509.06461v1.pdf> (September 2015)



Aplicaciones del Deep Learning

Videojuegos

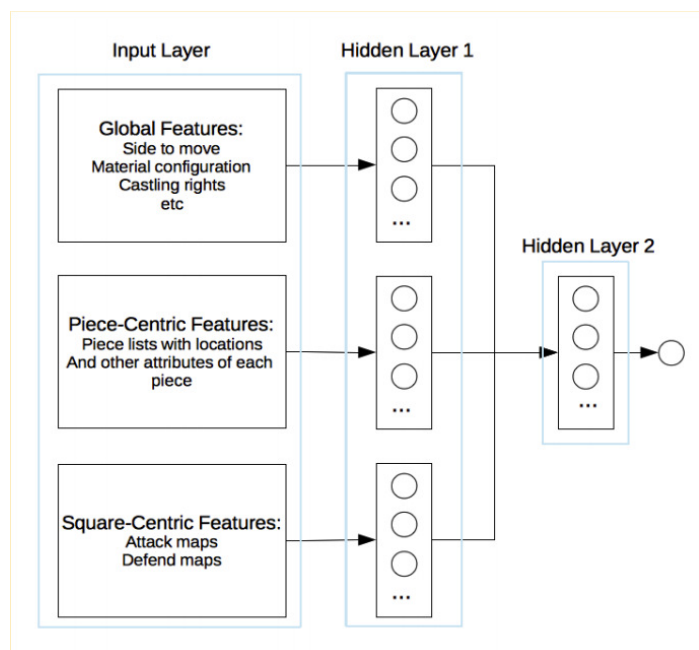
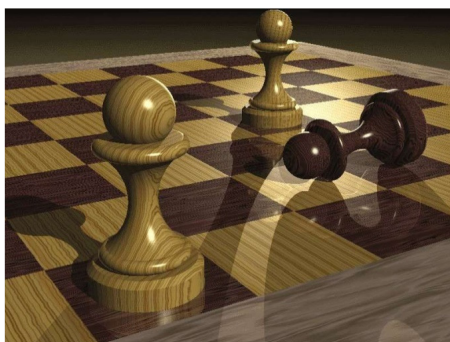
Deep Q Learning (Nature, 2015)



Aplicaciones del Deep Learning

Juegos

Ajedrez



Matthew Lai (Imperial College London):

“Giraffe: Using Deep Reinforcement Learning to Play Chess”

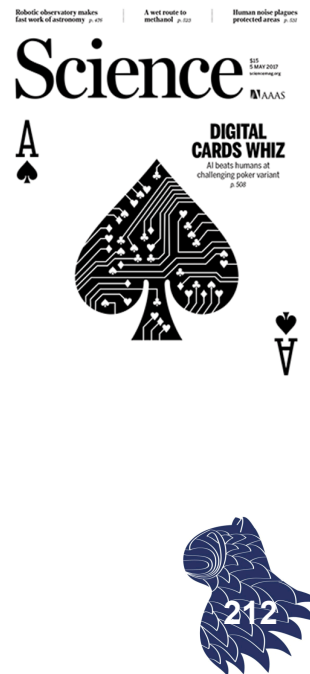
<http://arxiv.org/abs/1509.01549> (September 2015)



Aplicaciones del Deep Learning

Juegos: Poker

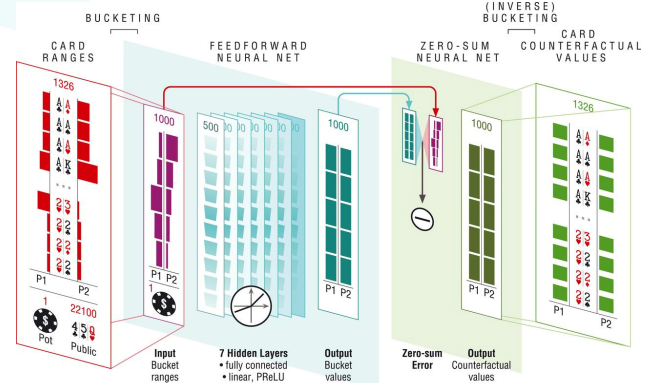
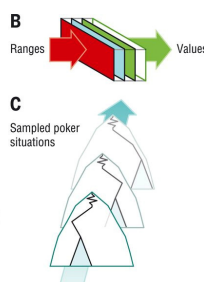
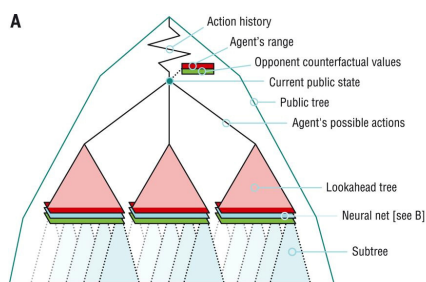
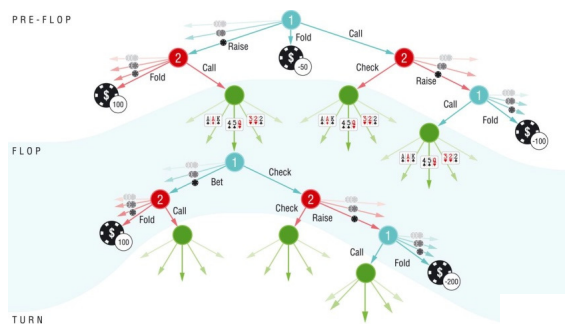
DeepStack



<https://www.deepstack.ai/>

Aplicaciones

Juegos Poker



DeepStack: Expert-level artificial intelligence in heads-up no-limit poker

Science, Vol. 356, Issue 6337, pp. 508-513, 5 May 2017

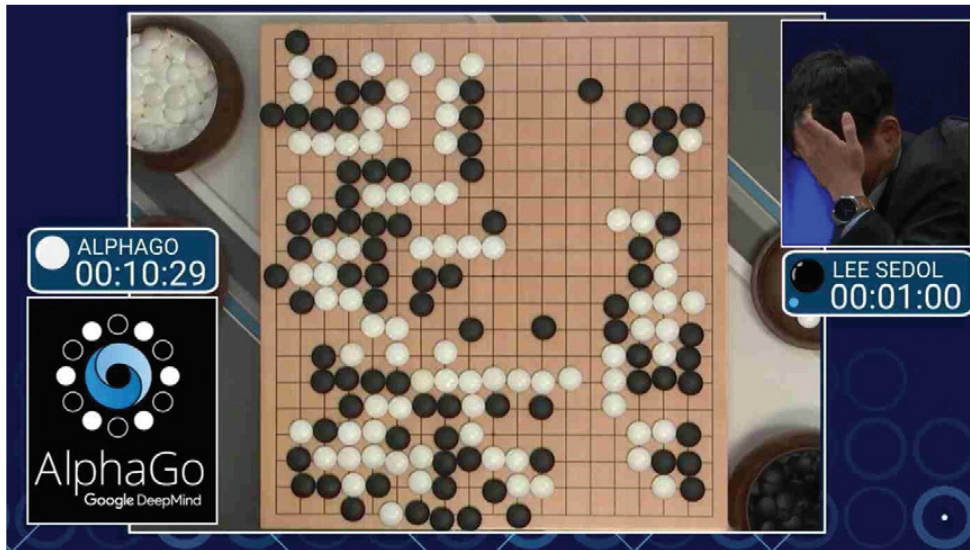
DOI: [10.1126/science.aam6960](https://doi.org/10.1126/science.aam6960)



Aplicaciones del Deep Learning

Juegos: Go

AlphaGo



<https://deepmind.com/research/alphago/>



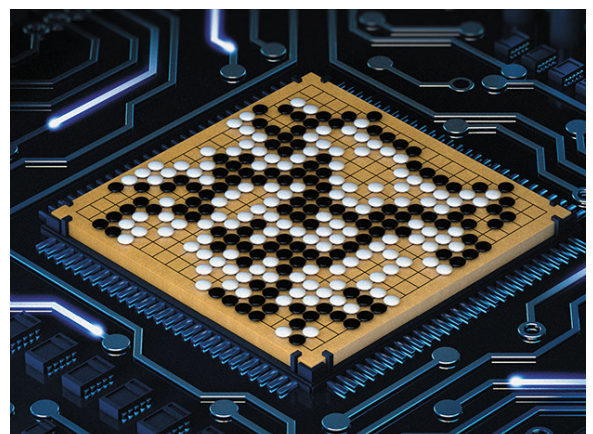
Aplicaciones

Juegos

Go: Los campeones humanos se negaban a jugar contra ordenadores porque eran demasiado malos ($b > 300$)...

Octubre 2015, Londres:
AlphaGo (Google DeepMind) vence al campeón europeo Fan Hui [2-dan], 5-0.

Marzo de 2016, Seúl: \$1M
AlphaGo (Google DeepMind) vence a Lee Sedol [9-dan], 4-1.



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



Aplicaciones del Deep Learning

Juegos

AlphaGo Zero



https://elpais.com/elpais/2018/12/05/ciencia/1544007034_265553.html



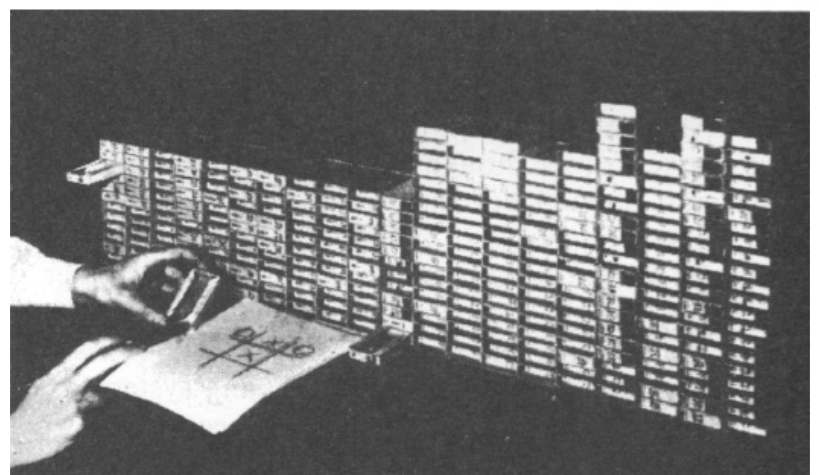
Aplicaciones del Deep Learning

Juegos: Go

Really???

The Matchbox Machine

1961



MENACE

Matchbox Educable Noughts And Crosses Engine

Donald Michie:

“Experiments on the mechanization of game-learning
Part I. Characterization of the model and its parameters”

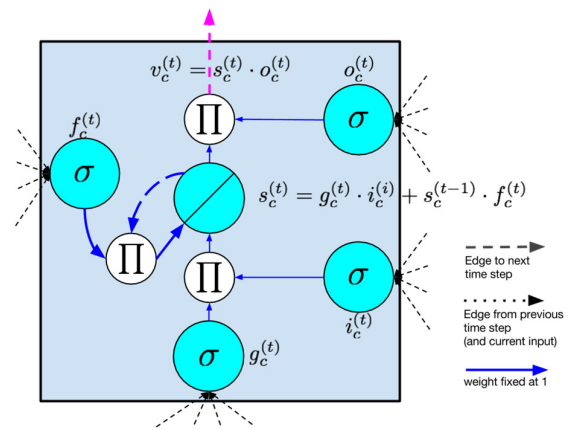
The Computer Journal, 6(3):232–236, November 1963,
The British Computer Society, <https://doi.org/10.1093/comjnl/6.3.232>



Deep Learning



Para algunos investigadores, que intentan obtener garantías teóricas basadas en resultados matemáticos: “[deep learning] might seem to be a regression” ;-)



LSTM (red recurrente)

En la práctica, los algoritmos con las mejores propiedades teóricas no son siempre los que mejor funcionan (sin restar importancia al estudio de las propiedades de los algoritmos de aprendizaje).



Deep Learning



Las técnicas heurísticas tienen éxito gracias a la disponibilidad de grandes conjuntos de datos (en los que el riesgo de sobreaprendizaje es menor) y la capacidad de cálculo de los sistemas actuales.

La validación con conjuntos de datos de prueba independientes ofrece una estimación de su comportamiento esperado en situaciones reales (los análisis teóricos se centran en el peor caso).



Deep Learning



Few Things Are Guaranteed

When attainable, theoretical guarantees are beautiful. They reflect clear thinking and provide deep insight to the structure of a problem. Given a working algorithm, a theory which explains its performance deepens understanding and provides a basis for further intuition. Given the absence of a working algorithm, theory offers a path of attack.

However, there is also beauty in the idea that well-founded intuitions paired with rigorous empirical study can yield consistently functioning systems that outperform better-understood models, and sometimes even humans at many important tasks. Empiricism offers a path forward for applications where formal analysis is stifled, and potentially opens new directions that might eventually admit deeper theoretical understanding in the future.

Zachary Lipton:

"Deep Learning and the Triumph of Empiricism"

KDnuggets, July 2015



Deep Learning



The only real success of deep learning so far has been the ability to map space X to space Y using a continuous geometric transform, given large amounts of human-annotated data. Doing this well is a game-changer for essentially every industry, but it is still a very long way from human-level AI.

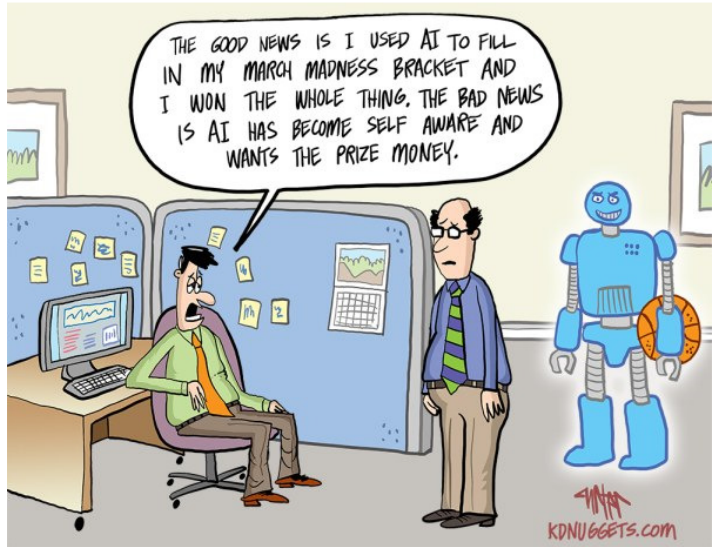
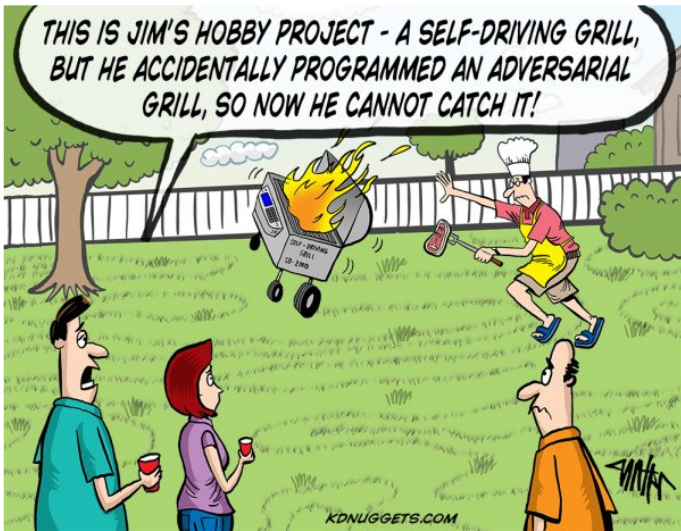
... machine learning models could be defined as "learnable programs"; currently we can only learn programs that belong to a very narrow and specific subset of all possible programs. But what if we could learn *any* program, in a modular and reusable way?

-- François Chollet: "The limitations of deep learning"

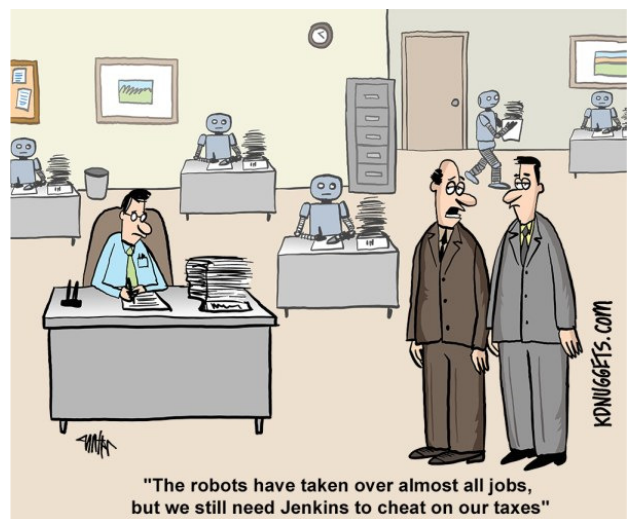
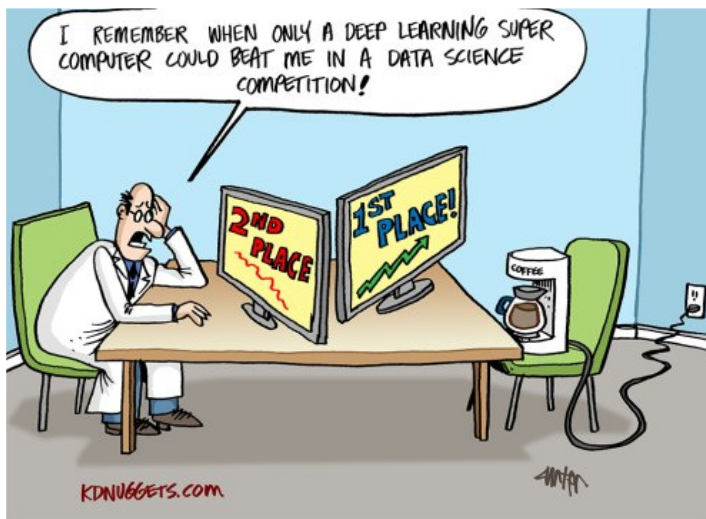
<https://blog.keras.io/the-limitations-of-deep-learning.html>



Deep Learning



Deep Learning



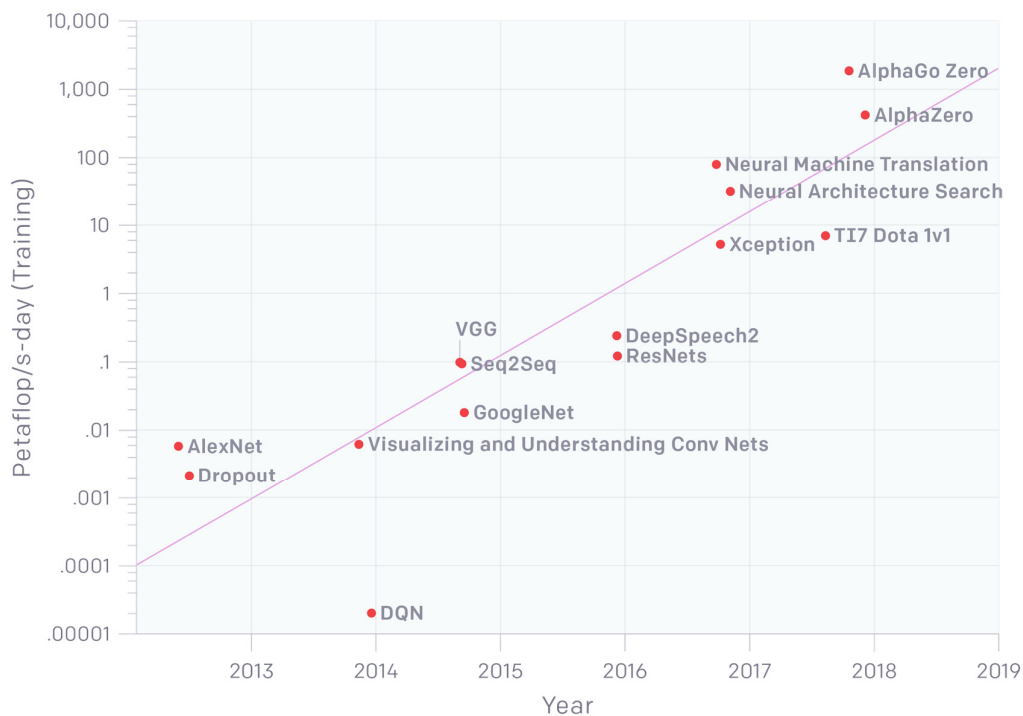
Deep Learning



El futuro...



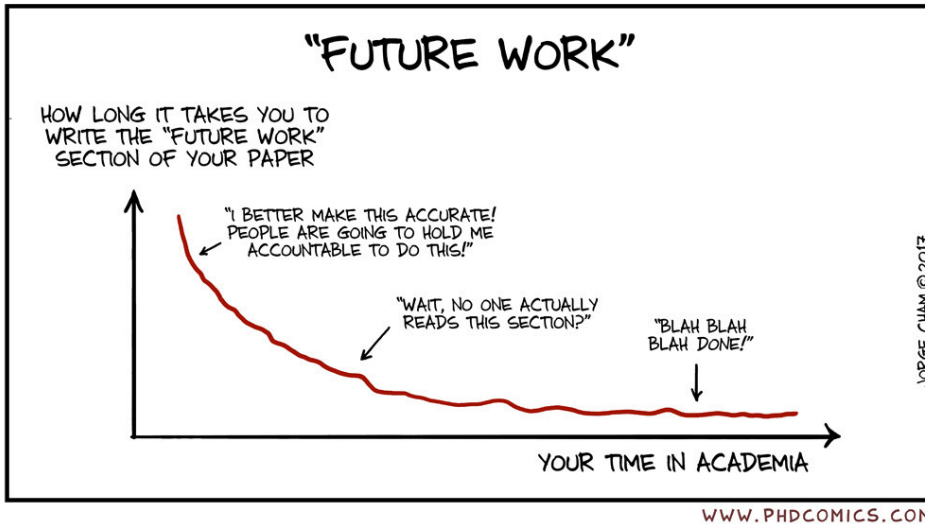
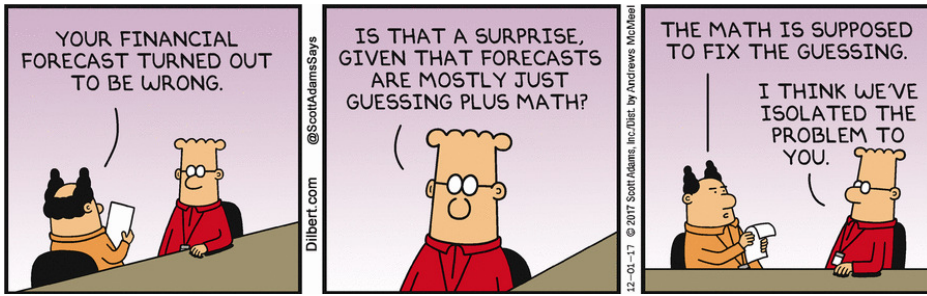
AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



<https://openai.com/blog/ai-and-compute/>



El futuro...



Demos

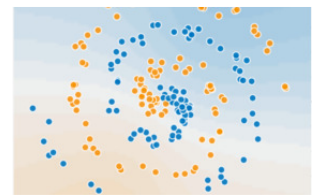


Para jugar un poco...

<http://playground.tensorflow.org/>

<http://ml4a.github.io/demos/>

<http://demos.algorithmia.com/classify-places/>



DATA: Which dataset do you want to use?

INPUT: Which properties do you want to feed in?

4 HIDDEN LAYERS: 4 neurons, 5 neurons, 5 neurons, 4 neurons

OUTPUT: Test loss 0.086, Training loss 0.048

Colors shows data, neuron and weight values.

Show test data Discretize output



Cursos

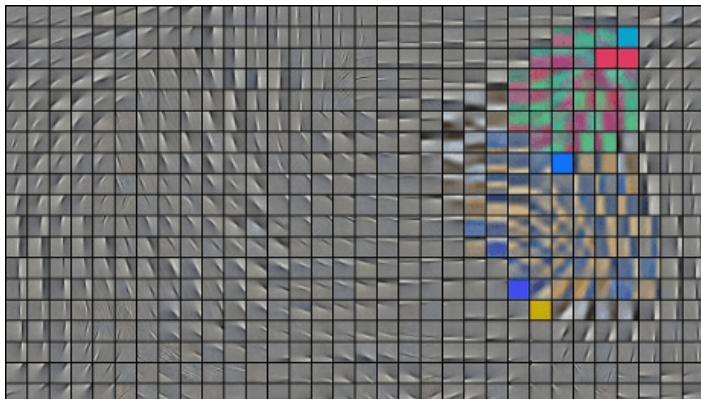


Neural Networks for Machine Learning

by Geoffrey Hinton

(University of Toronto & Google)

<https://www.coursera.org/course/neuralnets>



Cursos



Deep Learning Specialization

by Andrew Ng, 2017

- Neural Networks and Deep Learning
- Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization
- Structuring Machine Learning Projects
- Convolutional Neural Networks
- Sequence Models



deeplearning.ai

<https://www.coursera.org/specializations/deep-learning>



Cursos & Tutoriales



- **Deep Learning Tutorial**
Andrew Ng et al. (Stanford University)
<http://ufldl.stanford.edu/tutorial/>
- **Deep Learning: Methods and Applications**
Li Deng & Dong Yu (Microsoft Research)
<http://research.microsoft.com/apps/pubs/default.aspx?id=209355>
- **Deep Learning for Natural Language Processing**
Richard Socher et al. (Stanford University CS224d)
<http://cs224d.stanford.edu/>
- **Convolutional Neural Networks for Visual Recognition**
Andrej Karpathy (Stanford University CS231n)
<http://cs231n.github.io/>

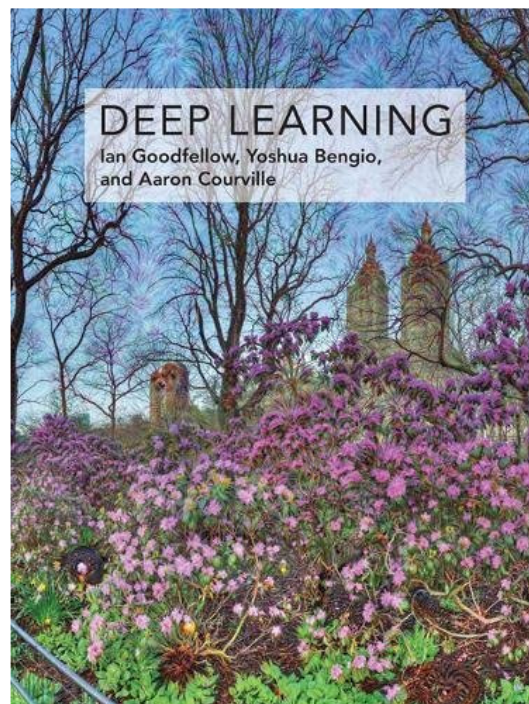


Bibliografía



Lecturas recomendadas

Ian Goodfellow,
Yoshua Bengio
& Aaron Courville:
Deep Learning
MIT Press, 2016
ISBN 0262035618



<http://www.deeplearningbook.org>



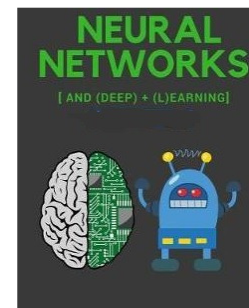
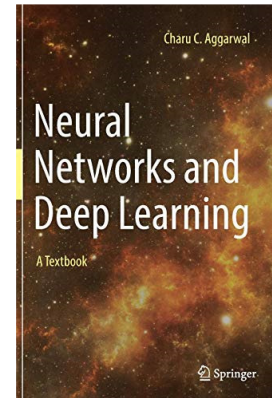
Bibliografía



Lecturas complementarias

Fundamentos

- Charu C. Aggarwal:
**Neural Networks and Deep Learning:
A Textbook.**
Springer, 2018
ISBN 3319944622
<http://link.springer.com/978-3-319-94463-0>
- Michael Nielsen:
Neural Networks and Deep Learning:
Determination Press, 2015
<http://neuralnetworksanddeeplearning.com/>



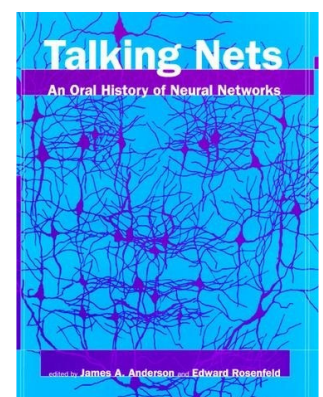
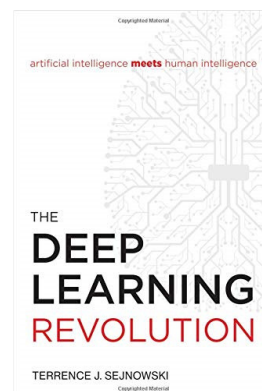
Bibliografía



Lecturas complementarias

Evolución histórica

- Terrence J. Sejnowski:
The Deep Learning Revolution
MIT Press, 2018
ISBN 026203803X
<https://mitpress.mit.edu/books/deep-learning-revolution>
- James A. Anderson & Edward Rosenfeld (editores):
Talking Nets: An Oral History of Neural Networks
The MIT Press, 1998
ISBN 0262011670
<https://mitpress.mit.edu/books/talking-nets>



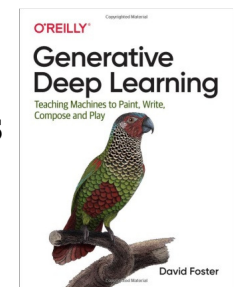
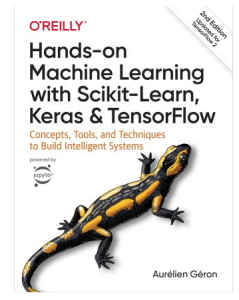
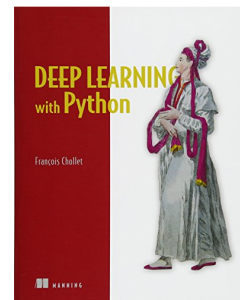
Bibliografía



Lecturas complementarias

Con una orientación práctica

- François Chollet:
Deep Learning with Python
Manning Publications, 2018
ISBN 1617294438
<https://github.com/fchollet/deep-learning-with-python-notebooks>
- Aurélien Géron: **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**
O'Reilly, 2nd edition, 2019, ISBN 1627052984
<https://github.com/ageron/handson-ml2>
- David Foster: **Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play**
O'Reilly, 2019, ISBN 1492041947
https://github.com/davidADSP/GDL_code



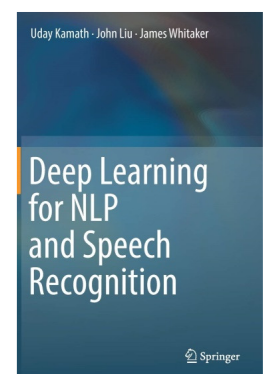
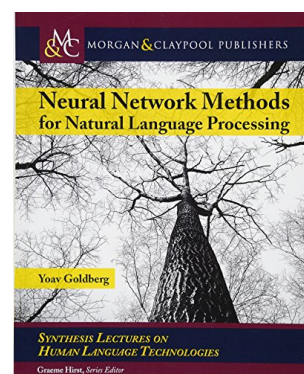
Bibliografía



Lecturas complementarias

Áreas de aplicación, p.ej. NLP

- Yoav Goldberg:
Neural Network Methods in Natural Language Processing
Morgan & Claypool Publishers, 2017
ISBN 1627052984
<https://doi.org/10.2200/S00762ED1V01Y201703HLT037>
- Uday Kamath, John Liu & James Whitaker:
Deep Learning for NLP and Speech Recognition
Springer, 2019
ISBN 3030145956
<http://link.springer.com/978-3-030-14595-8>

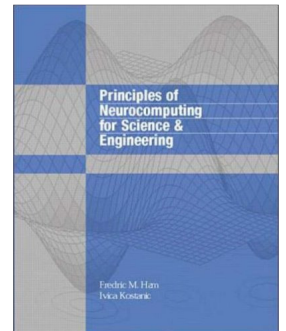
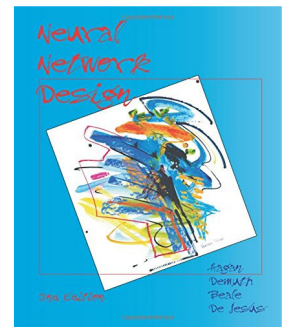
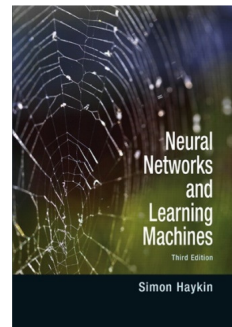


Bibliografía



Redes neuronales artificiales

- Simon Haykin:
Neural Networks and Learning Machines
Prentice Hall, 3rd edition, 2008
ISBN 0131471392
- Martin T. Hagan, Howard B. Demuth, Mark H. Beale & Orlando de Jesús:
Neural Network Design
Martin Hagan, 2nd edition, 2014
ISBN 0971732116
<http://hagan.okstate.edu/NNDesign.pdf>
- Fredric M. Ham & Ivica Kostanic:
Principles of Neurocomputing for Science and Engineering
McGraw-Hill Higher Education, 2000
ISBN 0070259666

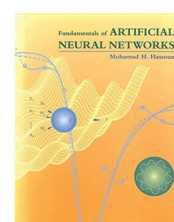
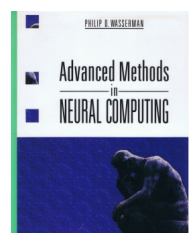
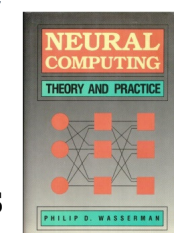
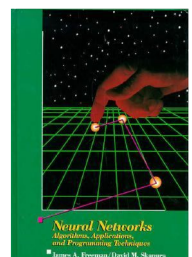
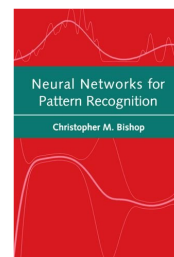


Bibliografía complementaria



Redes neuronales artificiales

- Christopher M. Bishop:
Neural Networks for Pattern Recognition
Oxford University Press, 1996. ISBN 0198538642
- James A. Freeman & David M. Skapura:
Neural Networks: Algorithms, Applications, and Programming Techniques
Addison-Wesley, 1991. ISBN 0201513765
- Mohamad Hassoun:
Fundamentals of Artificial Neural Networks
MIT Press, 2003. ISBN 0262514672
- Philip D. Wasserman:
Neural Computing: Theory and Practice,
Van Nostrand Reinhold, 1989. ISBN 0442207433
Advanced Methods in Neural Computing
Van Nostrand Reinhold, 1993. ISBN 0442004613

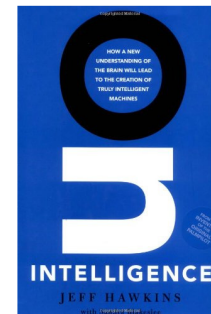
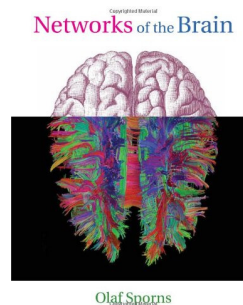
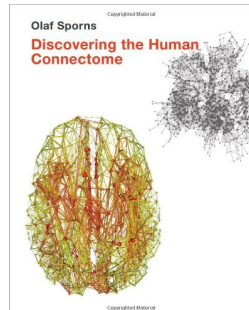
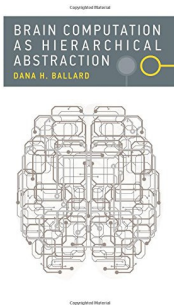


Bibliografía complementaria



“wetware”

- Dana H. Ballard: **Brain Computation as Hierarchical Abstraction.** MIT Press, 2015. ISBN 0262028611
- Olaf Sporns: **Discovering the Human Connectome.** MIT Press, 2012. ISBN 0262017903
- Olaf Sporns: **Networks of the Brain.** MIT Press, 2010. ISBN 0262014696
- Jeff Hawkins: **On Intelligence.** Times Books, 2004. ISBN 0805074562

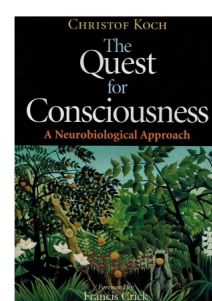
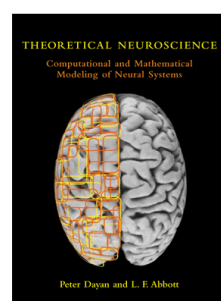
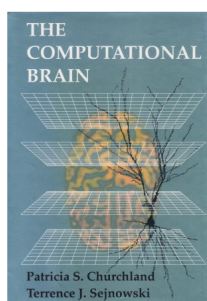


Bibliografía complementaria



Computational Neuroscience

- Patricia S. Churchland & Terrence J. Sejnowski: **The Computational Brain.** MIT Press, 1992. ISBN 0262031884
- Peter Dayan & L.F. Abbott: **Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.** MIT Press, 2001. ISBN 0262041995.
- Christof Koch: **The Quest for Consciousness: A Neurobiological Approach.** Roberts & Company Publishers, 2004. ISBN 0974707708

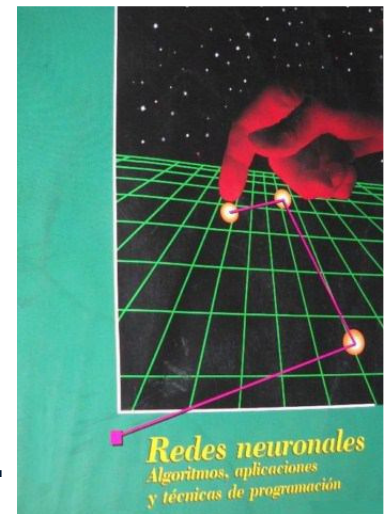


Bibliografía



Bibliografía en castellano

- James A. Freeman
& David M. Skapura:
**Redes Neuronales:
Algoritmos, aplicaciones
y técnicas de programación**
Addison-Wesley / Díaz de Santos, 1993.
ISBN 020160115X



... con ejemplos de código en Pascal

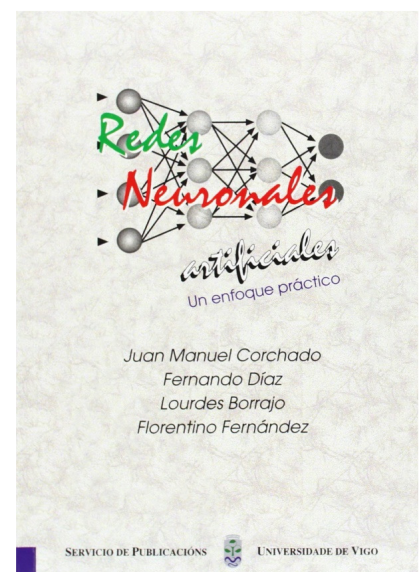


Bibliografía



Bibliografía en castellano

- Juan Manuel Corchado,
Fernando Díaz,
Lourdes Borrajo
& Florentino Fernández:
**Redes Neuronales:
Un enfoque práctico**
Universidad de Vigo, 2000.
ISBN 8481581453



... con un disco de 3½"



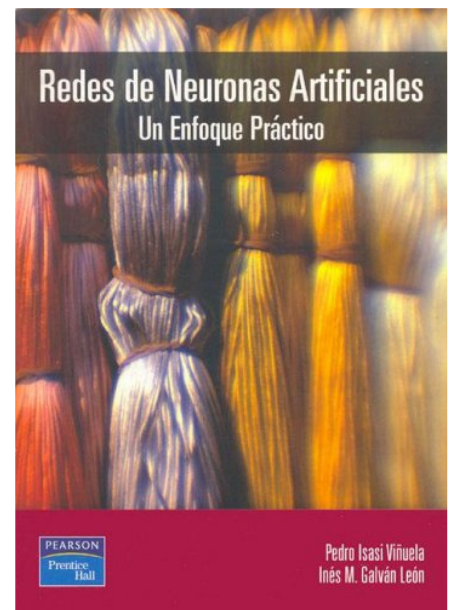
Bibliografía



Bibliografía en castellano

- Pedro Isasi Viñuela
& Inés M. Galván León
**Las redes neuronales artificiales:
Un enfoque práctico**
Prentice Hall, 2004
ISBN 8420540250

Descatalogado :-)



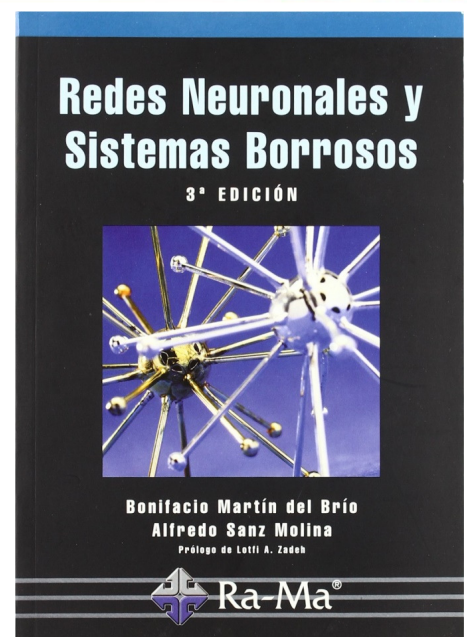
Bibliografía



Bibliografía en castellano

- Bonifacio Martín del Brío
& Alfredo Sanz Molina:
**Redes Neuronales
y Sistemas Borrosos.**
Ra-Ma, 3ª Edición, 2006
ISBN 8478977430

Alfaomega Grupo Editor
México D.F.



Bibliografía



Bibliografía en castellano

Fernando Berzal:

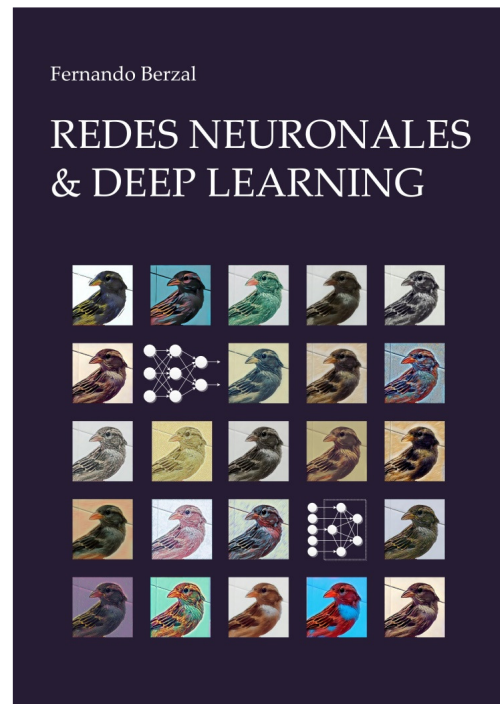
Redes Neuronales & Deep Learning

Edición independiente, 2018

ISBN 1-7312-6538-7 (b&n)

ISBN 1-7313-1433-7 (color)

<https://deep-learning.ikor.org>



Bibliografía



Bibliografía en castellano

Fernando Berzal:

Redes Neuronales & Deep Learning

Edición en dos volúmenes, 2019

Volumen I: Entrenamiento de redes neuronales artificiales

ISBN 1-0903-2030-2

Volumen II: Regularización, optimización y arquitecturas especializadas

ISBN 1-0903-3688-8

<https://deep-learning.ikor.org>

