Pontificia Universidad Católica de Chile
Departamento de Ciencias de la Computación
IIC2523 - Sistemas Distribuidos

# GPU en Deep Learning

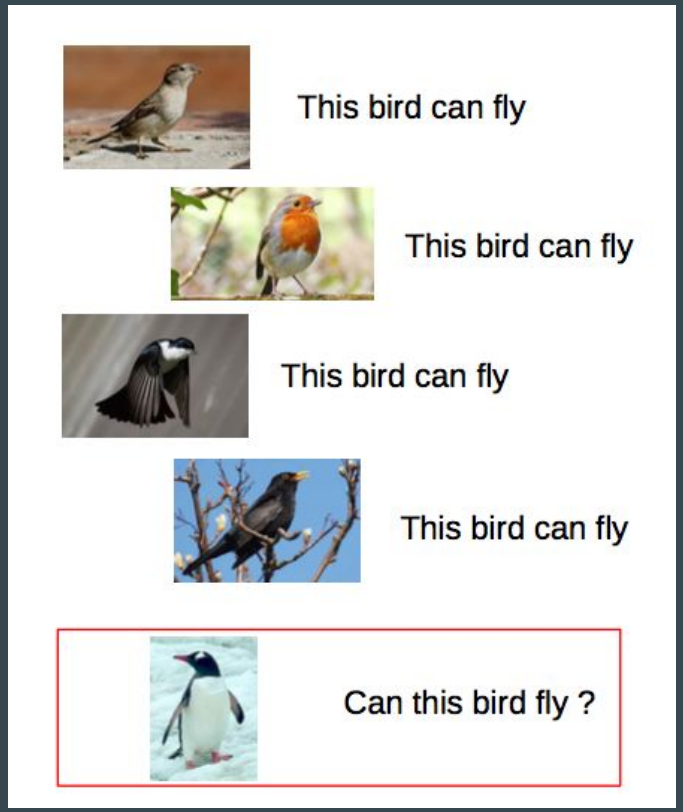
...

Patricio Cerda - Felipe Gómez

# Introducción: Graphics Processing Units (GPU)

- Nvidia “inventó" la GPU en el año 1999
- Nació para procesar gráficos de videojuegos o renderizado de imágenes en 3D
- Inicialmente se ignoraba su potencial en otras áreas

# Aprendizaje inductivo




 : Esto es un 5

 : Esto es un 3

 : Esto es un 2

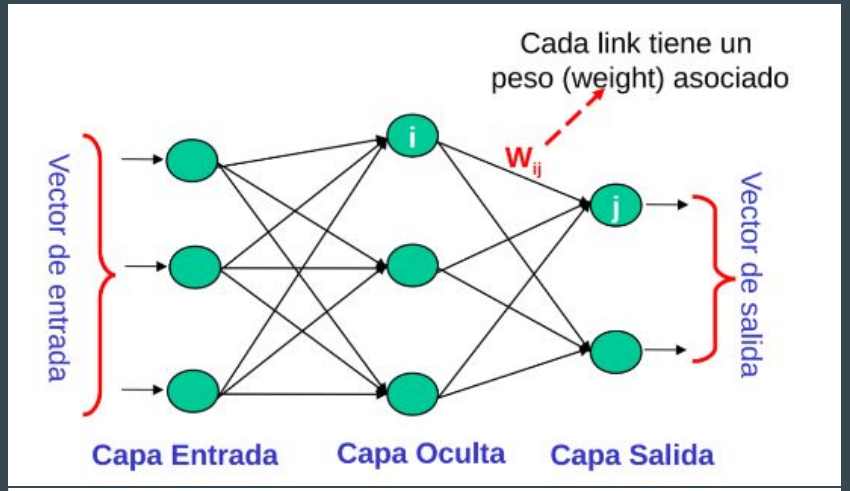
 : Esto es un 3

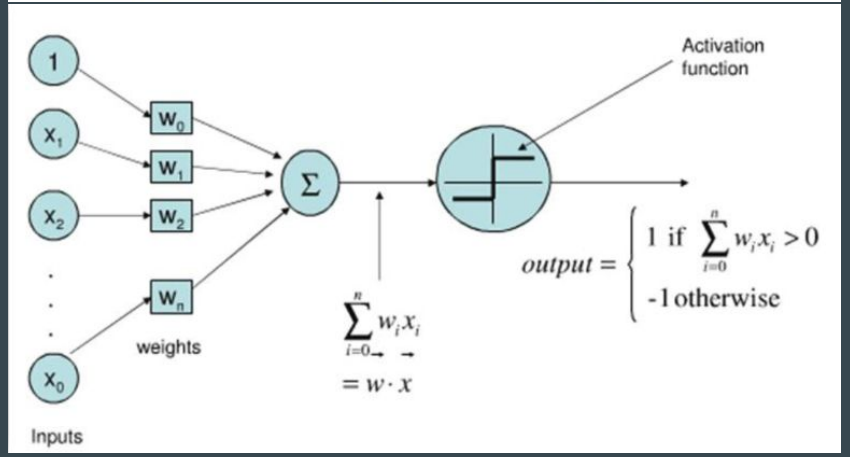
---

 : ¿Qué es esto?

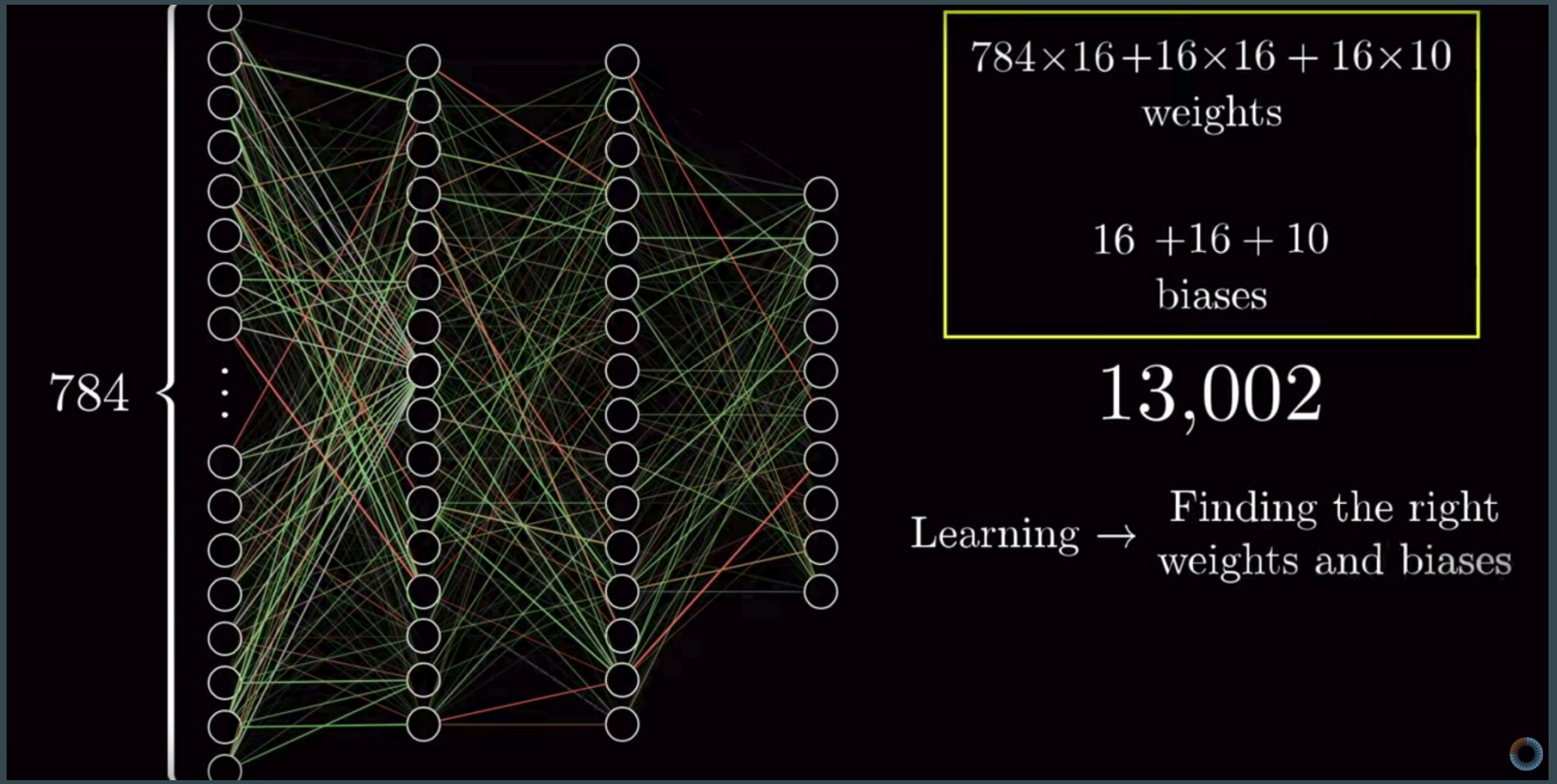
# Redes neuronales

- Sistemas de aprendizaje inductivo
- Estructura: capas de neuronas
- Cada neurona se activa en función de sus entradas



# Redes neuronales




Fuente: https://youtu.be/IHZwWFHWa-w

# *Deep Learning*

- Redes neuronales que aprenden representaciones jerárquicas composicionales

- "*Deep*": utilizan más de una capa oculta en su arquitectura

- La jerarquía permite un proceso de aprendizaje eficiente, pues capas de bajo nivel son compartidas por las de alto nivel

# Problema

LENTO

# *Problema*

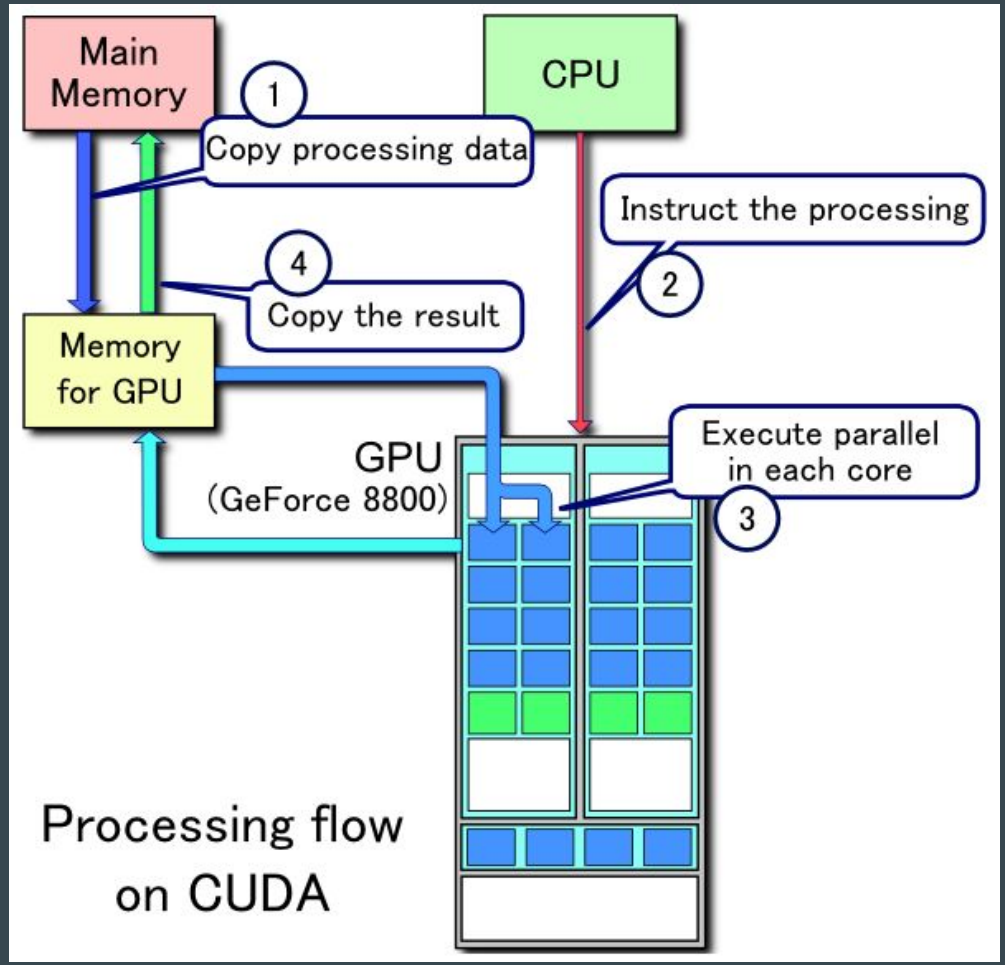
- Para el ejemplo de una imagen de solo 28 x 28 con dos capas de 16 neuronas cada una, se debe encontrar el valor óptimo de aproximadamente 13.000 parámetros
- La mayoría de los problemas complejos utilizan muchas más capas y más neuronas en cada una
- El problema se hacía infactible de solucionar en la práctica (problemas simples tardaban meses en solucionarse)

# Masificación de DL

2007: nVidia publica CUDA (*Compute Unified Device Architecture*). Facilita lograr flujo tipo SIMD sobre GPUs.

2009: primeros papers sobre sorprendentes ventajas de GPU

2012: primeros algoritmos de ML que utilizan GPU



# *Raina, Madhavan and Y. Ng. Large-scale Deep Unsupervised Learning using Graphics Processors* (2009)

| Package | Architecture | 576x1024 | 1024x4096 | 2304x16000 | 4096x11008 |
|---|---|---|---|---|---|
| Goto BLAS | Single CPU | 563s | 3638s | 172803s | 223741s |
| Goto BLAS | Dual-core CPU | 497s | 2987s | 93586s | 125381s |
| GPU | | 38.6s | 184s | 1376s | 1726s |
| GPU Speedup | | **12.9x** | **16.2x** | **68.0x** | **72.6x** |

| Package | Arch. | 20736x49152 | 36864x92928 |
|---|---|---|---|
| Goto | Single CPU | 38455s | 77246s |
| Goto | Dual-core | 32236s | 65235s |
| GPU | | 3415s | 6435s |
| GPU Speedup | | **9.4x** | **10.1x** |

| Method | Sparsity≈3% | 6% | 10% |
|---|---|---|---|
| Single CPU | 215s | 403s | 908s |
| Dual-core | 191s | 375s | 854s |
| GPU | 37.0s | 41.5s | 55.8s |
| Speedup | **5.2x** | **9.0x** | **15.3x** |

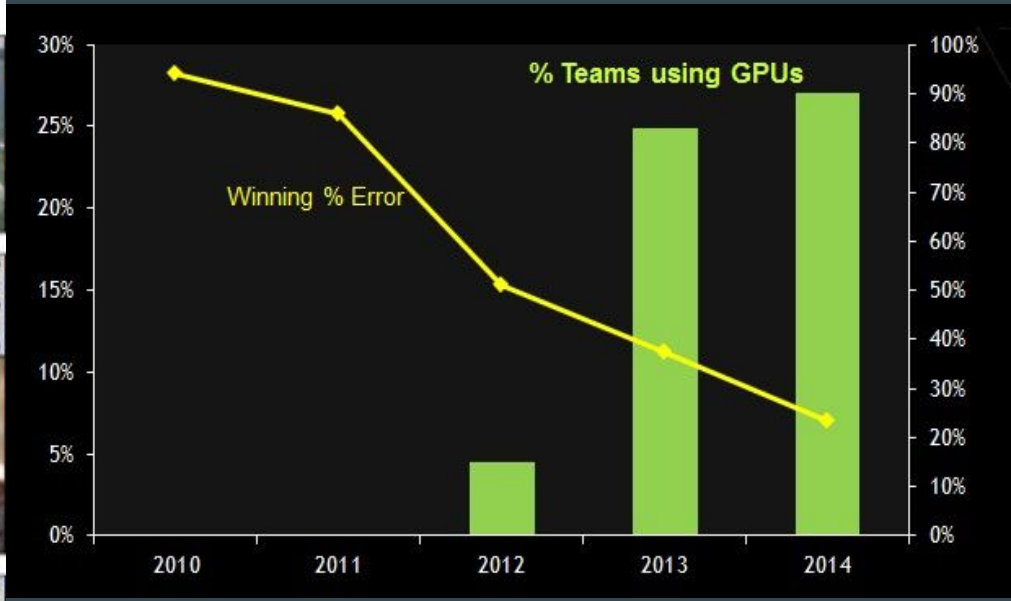
**GPGPU**

# Masificación de DL

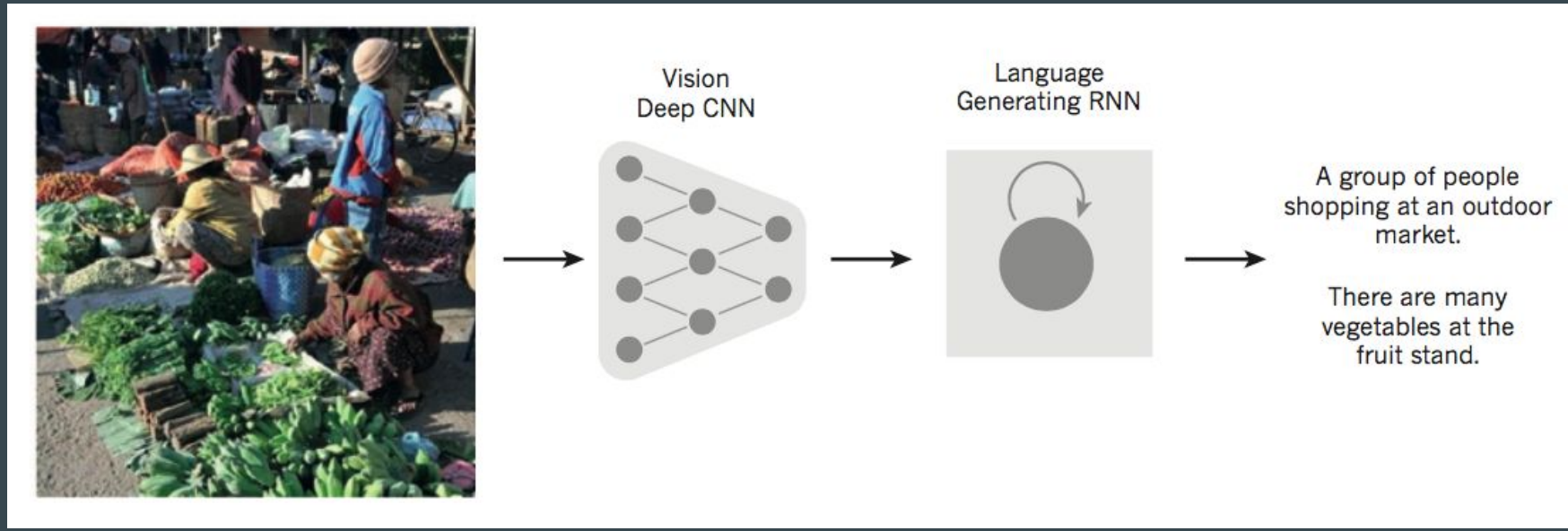
## ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

# Ejemplos



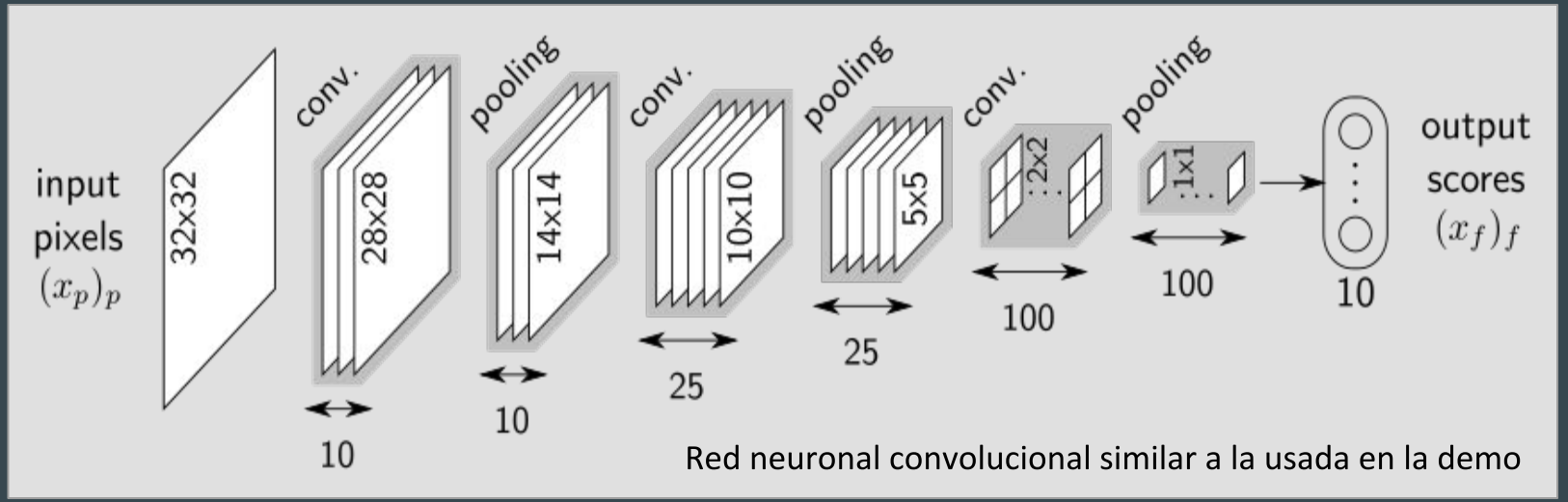

Automatic image captioning

# Ejemplos

Visual style transfer via Generative Adversarial Networks

# Demo: CPU vs GPU

Entrenaremos un modelo convolucional simple para resolver el problema de clasificación de dígitos (base de datos MNIST)



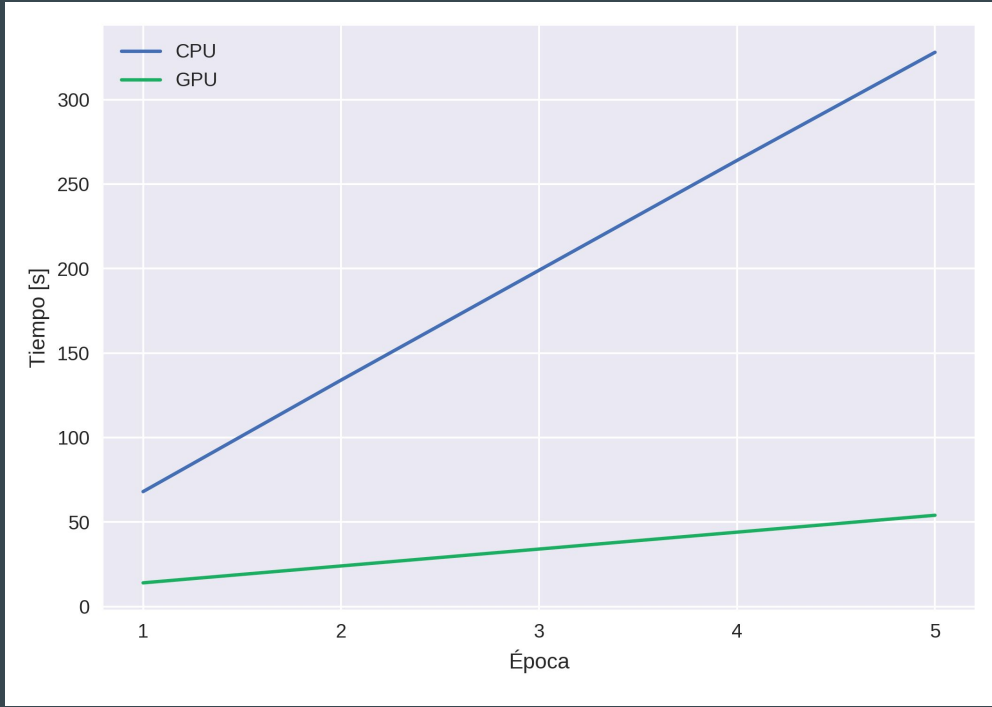

Red neuronal convolucional similar a la usada en la demo

# Demo: resultados

- **6 veces más rápido con *GPU* !**

- Nota: Para replicar, seleccionar el ambiente de ejecución respectivo en *Colaboratory*

- En modelos más sofisticados, el speedup puede ser aún mayor



Pontificia Universidad Católica de Chile
Departamento de Ciencias de la Computación
IIC2523 - Sistemas Distribuidos

# ¿Preguntas?

...

# Referencias bibliográficas


1. Francois Chollet - Deep Learning with Python, Manning (2017)
2. LeCun, Bengio, Hinton - Deep Learning, Nature (2015)
3. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. “Show and tell: a neural image caption generator” (2014)
4. Zhu J., Park T., Isola P., Efros, A. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Network” (2018)