



REDES NEURONALES 2024

Clase 24 parte 2

Jueves 14 de noviembre 2024

FAMAF, UNIVERSIDAD NACIONAL DE CÓRDOBA

INSTITUTO DE FÍSICA ENRIQUE GAVIOLA (UNC-CONICET)

1 Redes Neuronales Convolucionales

- Conceptos básicos
- Capas convolucionales

Redes Neuronales Convolucionales

Conceptos básicos

Desde el origen de la inteligencia artificial, los científicos y científicas han buscado inspiración en el sentido de la vista. Hoy contamos un campo específico dentro del Aprendizaje Automático denominado **visión por computadora**. Este campo en particular está profundamente asociado a las mejores aplicaciones industriales, como es fácil imaginar.

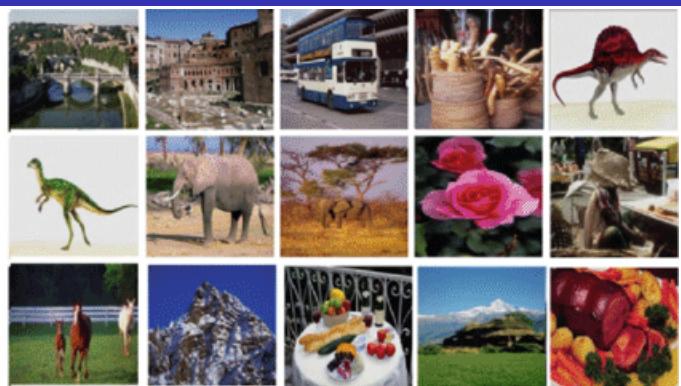
En el año 2006, un investigador llamado Fei-Fei Li comenzó a trabajar en el proyecto llama *ImageNet*, que consistió en la construcción de una base de imágenes de objetos. En pocos años consiguieron clasificar catorce millones de fotos de objetos, adecuadamente etiquetadas. Esta base de imágenes dio lugar a un competencia anual denominada ImageNet Large Scale Visual Recognition Challenge (ILSVRC) donde equipos de programadores de todo de todo el mundo compiten por clasificar las imágenes (un conjunto recortado) en aproximadamente mil categorías. El conjunto completo contienen veinte mil categorías.

En el año 2012 un sistema llamado AlexNet (creado por Alex Krizhevsky) ganó la competencia, rompiendo todas las marcas e usando por primera vez en este desafío una red neuronal convolucional y desde entonces su uso se ha difundido muchísimo, no solo para el tratamiento de imágenes, sino también para el procesamiento de series temporales y señales.

Las Redes Neuronales Convolucionales fueron introducidas en la década del 90 del siglo pasado, en el auge del paradigma conexionista de Inteligencia Artificial, para emular el sistema de procesamiento visual en mamíferos, que comienza en la retina. Esta arquitectura fue inicialmente publicada por el investigador japonés Kunihiko Fukushima, y dos años después fue modificada por Yann LeCun, hoy considerado el padre de la RNC. La versión de LeCun se llamó **ConvNets**.



(a) ImageNet Synset: One sample image from each category



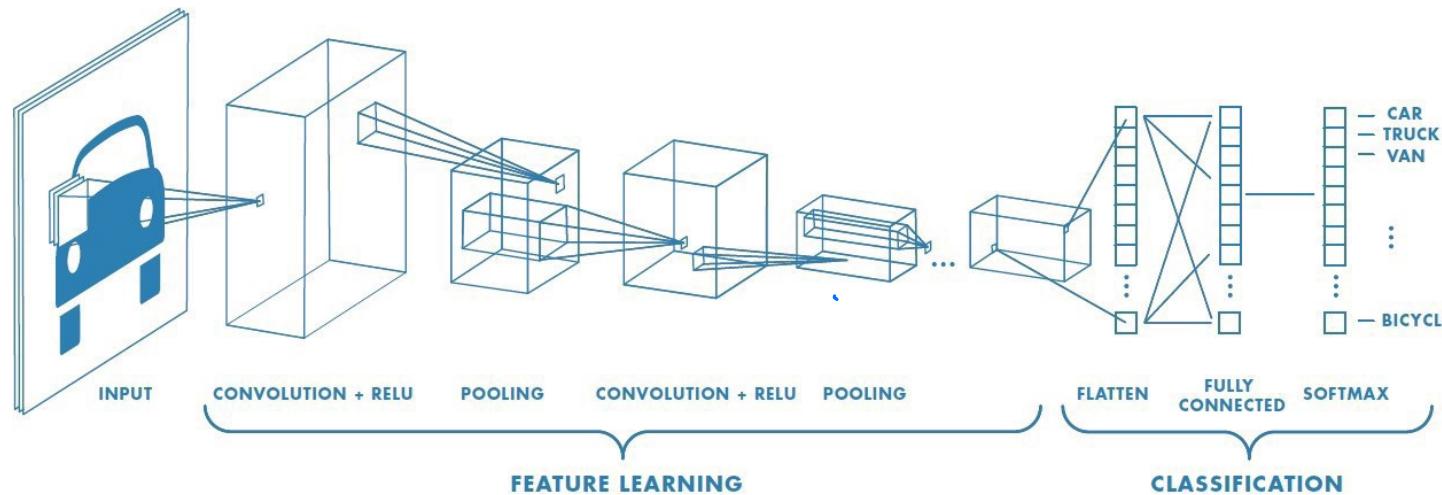
(b) Corel-1000 Dataset: Sample images from each category



(c) Caltech-256 Dataset: One sample image from each category



(d) Caltech-101 Dataset: One sample image from each category



Las RNC, a diferencia de las redes multi-capas feed-forward que vimos en la primera parte del curso, están diseñadas para conservar la relación espacial de los datos de entrada. Esto es muy importante en el caso de procesamiento de imágenes y de series temporales, donde la proximidad en los datos de entrada es un dato de vital importancia.

Las RNC están organizadas para recibir de entrada objetos tridimensionales, caracterizadas por el ancho W , el alto H y la profundidad D . La necesidad de tres dimensiones surge por ejemplo de la necesidad de considerar diferentes canales en las imágenes digitales, como por ejemplo la que usa en la convención RGB (red, green, blue).

Existen varias formas de definirlo, pero en términos generales diremos que son la generalización de matrices bidimensionales a objetos matemáticos multidimensionales.

Definición

Un **tensor** de orden d es una arreglo d -dimensional, donde llamamos *modos* a sus dimensiones. Denotamos un tensor como $T \in R^{n_1 \times n_2 \times \dots \times n_d}$, donde n_j es la cantidad de componentes presente en el j -ésimo modo, con $j = 1, 2, \dots, d$.

Un escalar t es un tensor de orden 0 con una única componente.

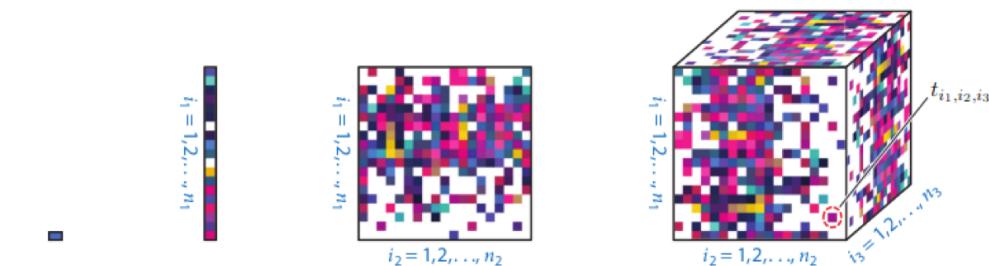
Un vector $\vec{t} \in R^n$ es un tensor de orden 1 con n componentes.

Una matriz $A \in R^{n \times m}$ es un tensor de orden 2, con n componentes en el primer modo y m componentes en el segundo modo.

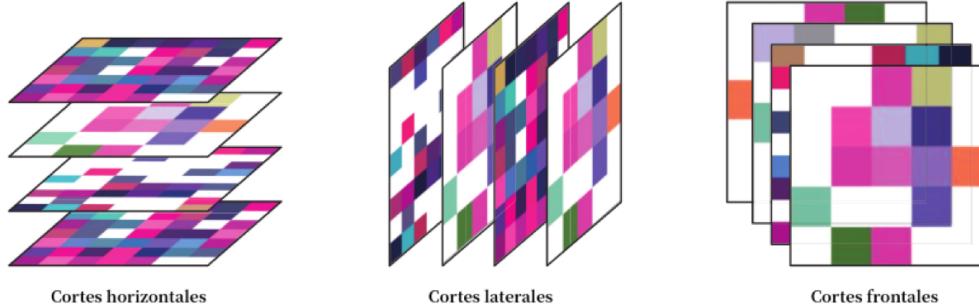
Para representar imágenes se usan tensores de tercer orden de tamaño $W \times H \times D$, con $W, H, D \in \mathbb{N}$, donde W y H son el número de pixeles en el ancho y el alto de la imagen, respectivamente y D representa la *profundidad*.

Si las imágenes son en blanco y negro o en tonos de grises, la profundidad será $D = 1$ y recuperamos una matriz bidimensional. Si las imágenes son en color y se usa la convención RGB, entonces $D = 3$.

Se suele usar un tensor de orden 4 para definir un mini-lote de imágenes (recuerden que ya definimos mini-lote o minibatch). En este caso tenemos $W \times H \times D \times B_s$ donde B_s es el tamaño del mini-lote.

Escalar: t Vector: \mathbf{t} Matriz: \mathbf{T} Tensor de tercer orden: \mathcal{T}

(a)



Cortes horizontales

Cortes laterales

Cortes frontales

(b)

Representación de tensores de distintos órdenes (a) y cortes de un tensor de tercer orden (b).

La arquitectura de las RNC se caracteriza por tener *Capas Ocultas Convolucionales*, que difieren de las capas densas (propias de las redes feed-forward), en el hecho de que cada neurona está conectada solamente a un región local de dimensión d (la dimensión de la entrada). El nombre de estas capas guarda relación con la operación de convolución matemática que denotamos, como siempre se hace, con el símbolo $*$.

Definición: convolución continua

Sean f y g dos funciones reales, continuas por tramos en \mathbb{R} .

Definimos la **convolución** de f y g como:

$$(f * g)(t) = (f(\tau) * g(\tau))(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau$$

Esto se puede extender al caso en que las funciones tengan dominio en los enteros \mathbb{Z} .

Definición: convolución discreta

Sean f y g dos funciones enteras, continuas por tramos en \mathbb{R} .

Definimos la **convolución** de f y g como:

$$(f * g)(m) = (f(n) * g(n))(m) = \sum_n f(n) g(m - n).$$

Definición: correlación cruzada continua

Sean f y g dos funciones reales, continuas por tramos en \mathbb{R} .

Definimos la **correlación cruzada** de f y g como:

$$(f * g)(t) = (f(\tau) * g(\tau))(t) = \int_{-\infty}^{\infty} f(\tau) g(t + \tau) d\tau$$

Definición: correlación cruzada discreta

Sean f y g dos funciones reales, continuas por tramos en \mathbb{R} .

Definimos la **correlación cruzada** de f y g como:

$$(f * g)(m) = (f(n) * g(n))(m) = \sum_n f(n) g(m + n).$$

Proposición

La correlación cruzada de dos funciones $f(\tau)$ y $g(\tau)$ es equivalente a la convolución de $f(-\tau)$ y $g(\tau)$, o sea,

$$(f(\tau) * g(\tau))(t) = (f(-\tau) * g(\tau))(t)$$

Las capas convolucionales de las redes convolucionales aplican la llamada correlación cruzada a un tensor de entrada. Aclaramos que una RNC suele tener muchas capas convolucionales.

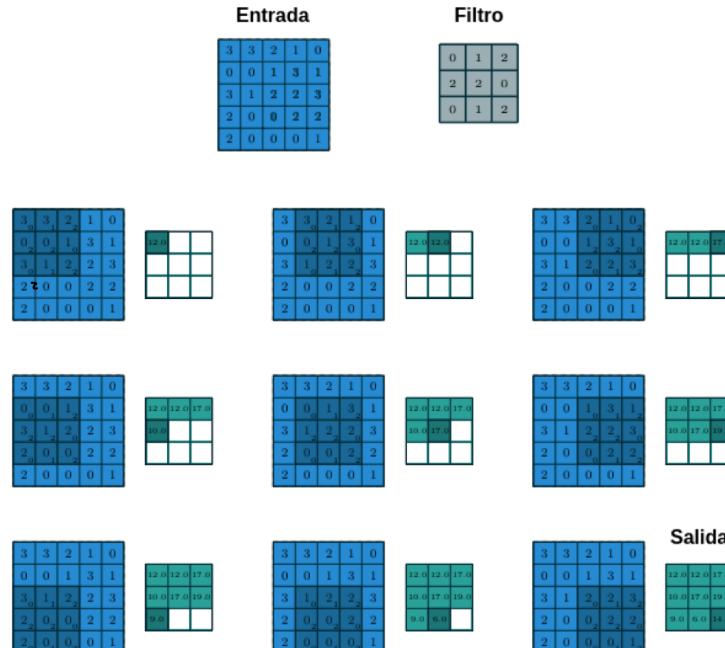
Definición de filtro de convolución

Un filtro de correlación es un tensor de tamaño $f_W \times f_H \times D$, donde f_W y f_H son el ancho y el alto del núcleo (kernel) y D ese la profundidad, que en general coincidará a lo largo de la RNC con la profundidad D de la entrada.

Las entradas de los filtros son los pesos de las conexiones que varían entre cortes frontales. Como las imágenes son representadas por tensores de orden 3 y cada corte frontal es de tamaño $W \times H$, donde cada entrada es el valor del pixel correspondiente (por ejemplo un número entero entre 0 y 255), se aplica la *correlación cruzada* discreta en dos dimensiones entre los pixeles seleccionados de la imagen y los pesos de un filtro. Esto es, se superponen el i -ésimo corte frontal del filtro sobre el i -ésimo corte frontal de la entrada de capa y se suman las multiplicaciones lugar a lugar:

$$(f * g)(m_1, m_2) = \sum_{n_1} \sum_{n_2} f(n_1, n_2) g(m_1 + n_1, m_2 + n_2).$$

Supongamos por simplicidad que procesamos una imágenes en tonos de gris, o sea, $D = 1$ y que $f_W = f_H = 3$. En la figura vemos un buen ejemplo de filtro de convolución.



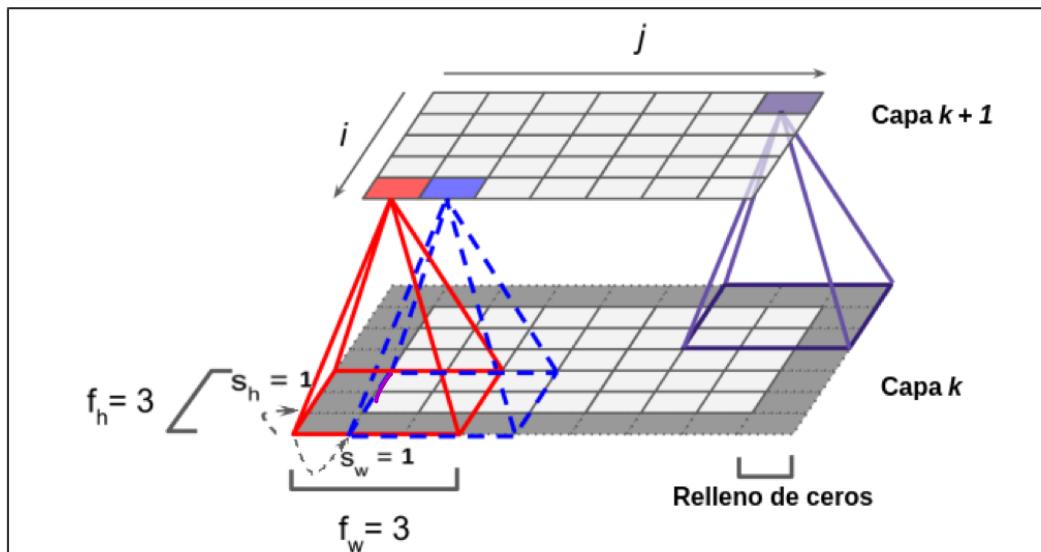
En el caso del un filtro cuadrado de ancho 3 y una imagen de entrada cuadrada de ancho 5 ($D = 1$), la salida de la capa de convolución es un red 3×3 . Esto es así porque la convolución no afecta a las filas y columnas de los extremos de las imágenes.

En el caso más general, supongamos una imagen con $D = 1$ (grises) cuadrada de ancho W y alto $H = W$. Si aplicamos un filtro bidimensional de ancho f_W (o sea, $f_H = f_W$ y $D = 1$), entonces perderemos $f_W - 1$ filas y $f_W - 1$ columnas, a las cuales no le aplicamos el filtro.

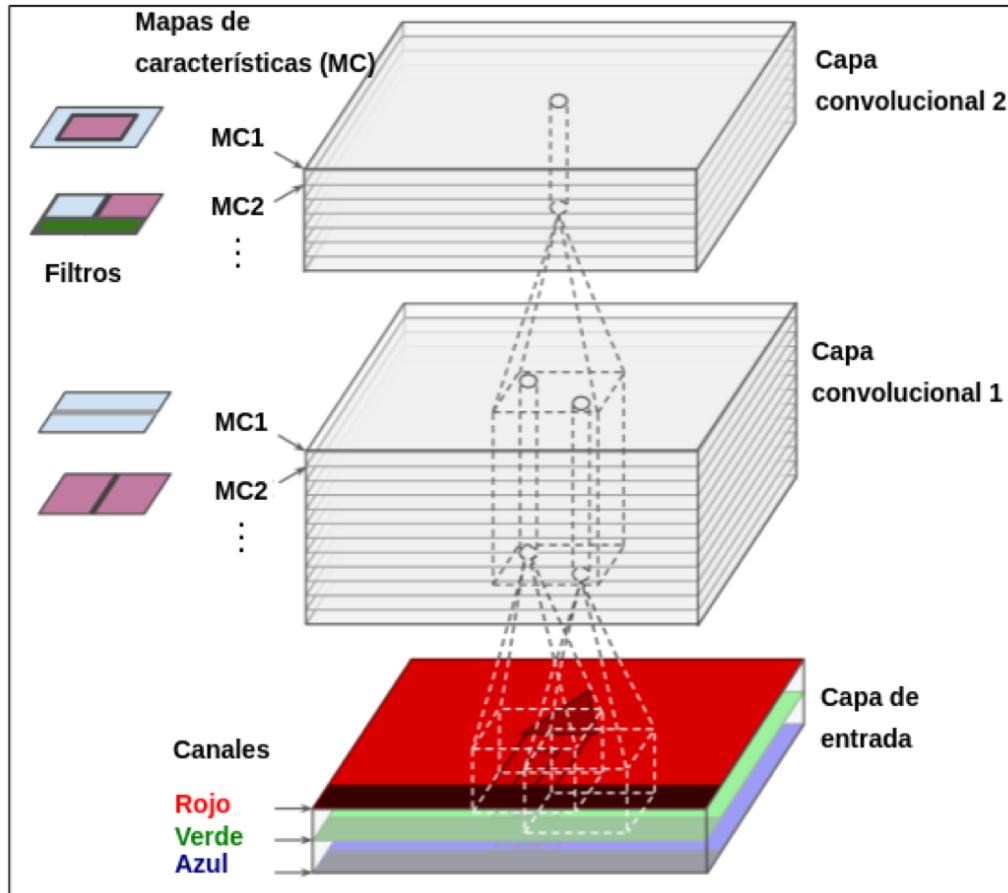
Definición de campo receptivo

El *Campo Receptivo* (RF por su nombre en inglés), es una región local, incluyendo la profundidad en la capa anterior a la que se conecta una neurona.

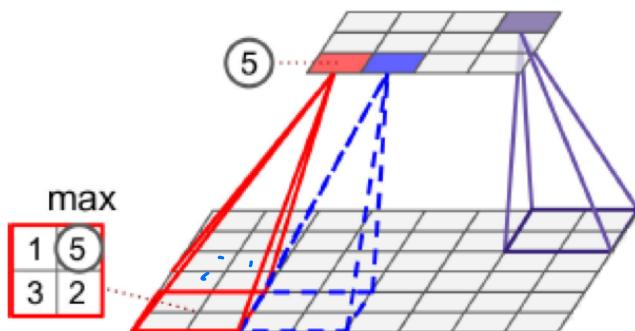
Las neuronas de la primera capa convolucional no están conectadas a cada uno de los pixeles de la imagen de entrada, sino solo a los pixeles de sus campos receptivos. Durante la propagación hacia adelante, se realiza la correlación cruzada de cada filtro a través del ancho y alto del volumen de entrada y se produce un mapa de características bidimensionales con los resultados de esta operación. No hay un único filtro, sino muchos, y entonces la salida de la red convolucional apila muchas salidas.

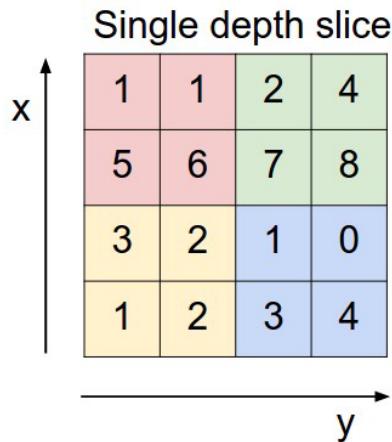


Al apilar estos mapas a lo largo de la profundidad, se obtiene el volumen de salida. Todas las neuronas dentro de un mapa de características (feature map) comparten los mismos pesos sinápticos (parámetros), lo que reduce significativamente el número de parámetros del modelo, pero además las neuronas de los diferentes mapas de características usan distintos parámetros. El tamaño del volumen de salida de cada capa convolucional depende del tamaño del núcleo, la profundidad D , el paso o *stride* (S_W y S_H) y el relleno de ceros o *zero padding* (P).



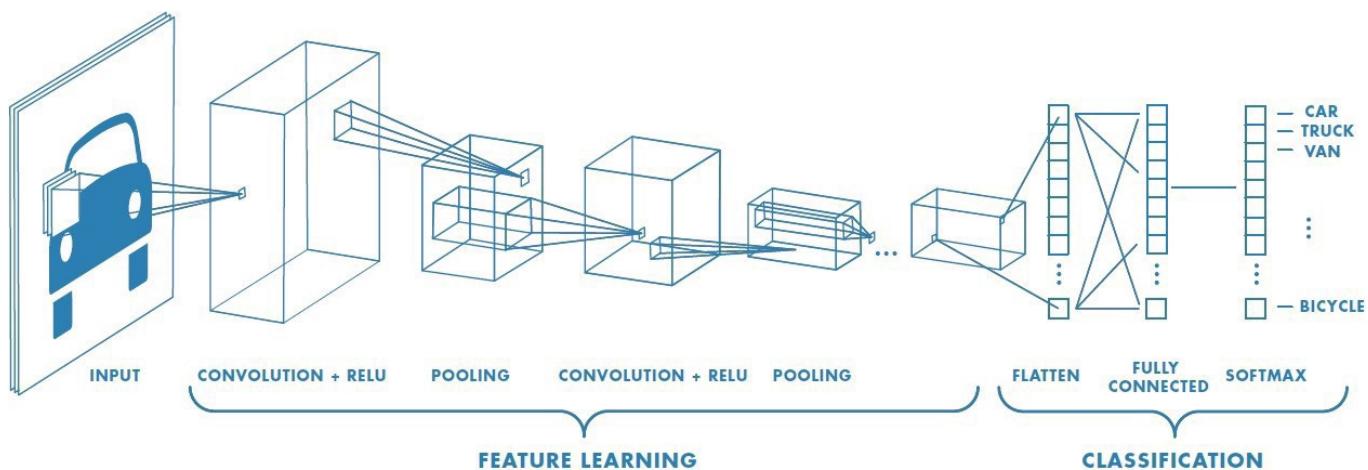
Capas de agrupamiento





max pool with 2×2 filters
and stride 2

6	8
3	4



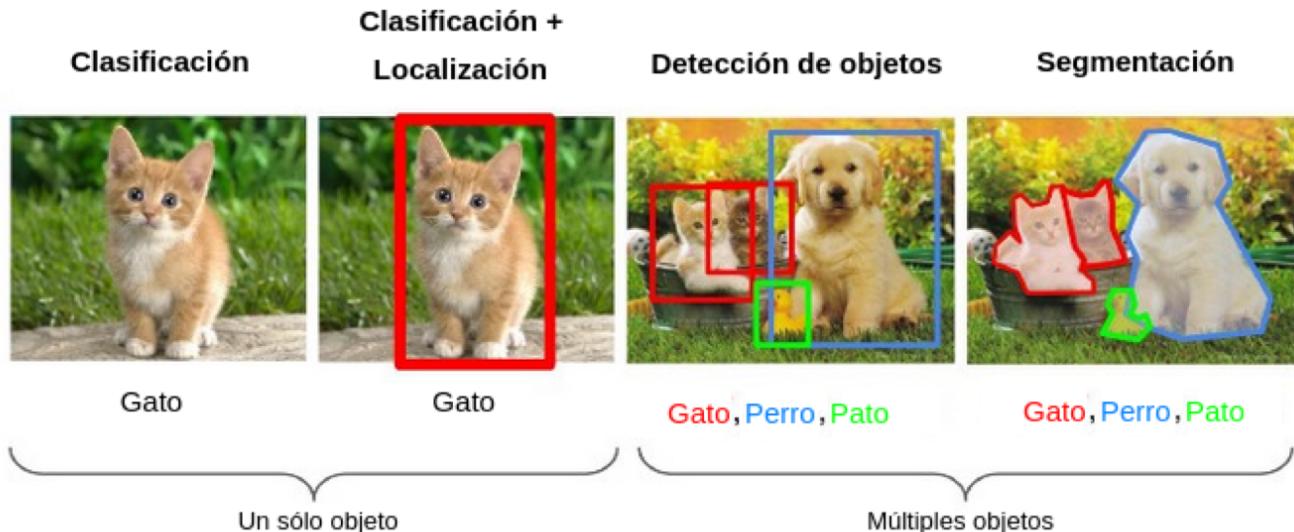


Figura 2.13. Comparación entre clasificación, detección y segmentación de objetos en imágenes.¹⁹

