Segovia
TECH TRAIN

ATAQUES ADVERSARIOS

JOSÉ MANUEL SIMÓN RAMOS

ÍNDICE

¿QUÉ ES UN ATAQUE ADVERSARIO?

¿QUÉ ES UN ATAQUE ADVERSARIO?

- Un ataque a
(aparentem
ha agregado
preparado)
imagen gen
a la red hac
como una c
la clase de

=

"gibbon"
99.3% confidence

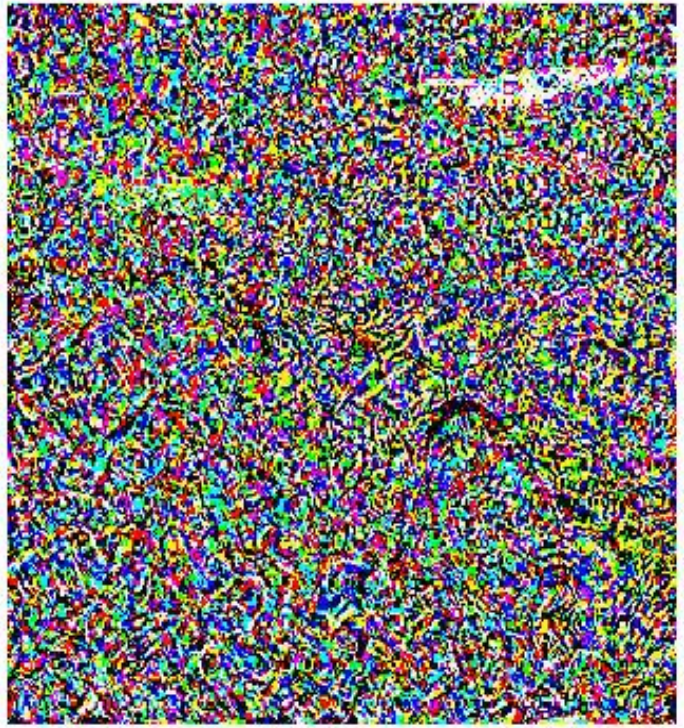
CARACTERÍSTICAS

- Se generan añadiendo ruido de manera premeditada a la imagen original.
- Generalmente las modificaciones que se realizan son imperceptibles para el ojo humano.
- No es necesario tener disponible el modelo a atacar.
- Altera tanto la salida de la red que hace que las demás clases no se tengan en cuenta.
- Es propagable: Existe la posibilidad de que un ataque que haya funcionado con un modelo funcione también para otro.

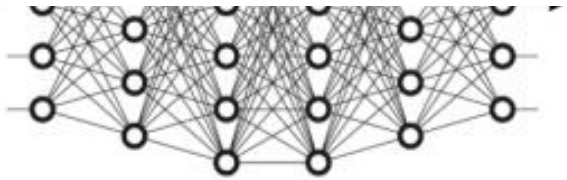
EJEMPLOS

Original image: sports car

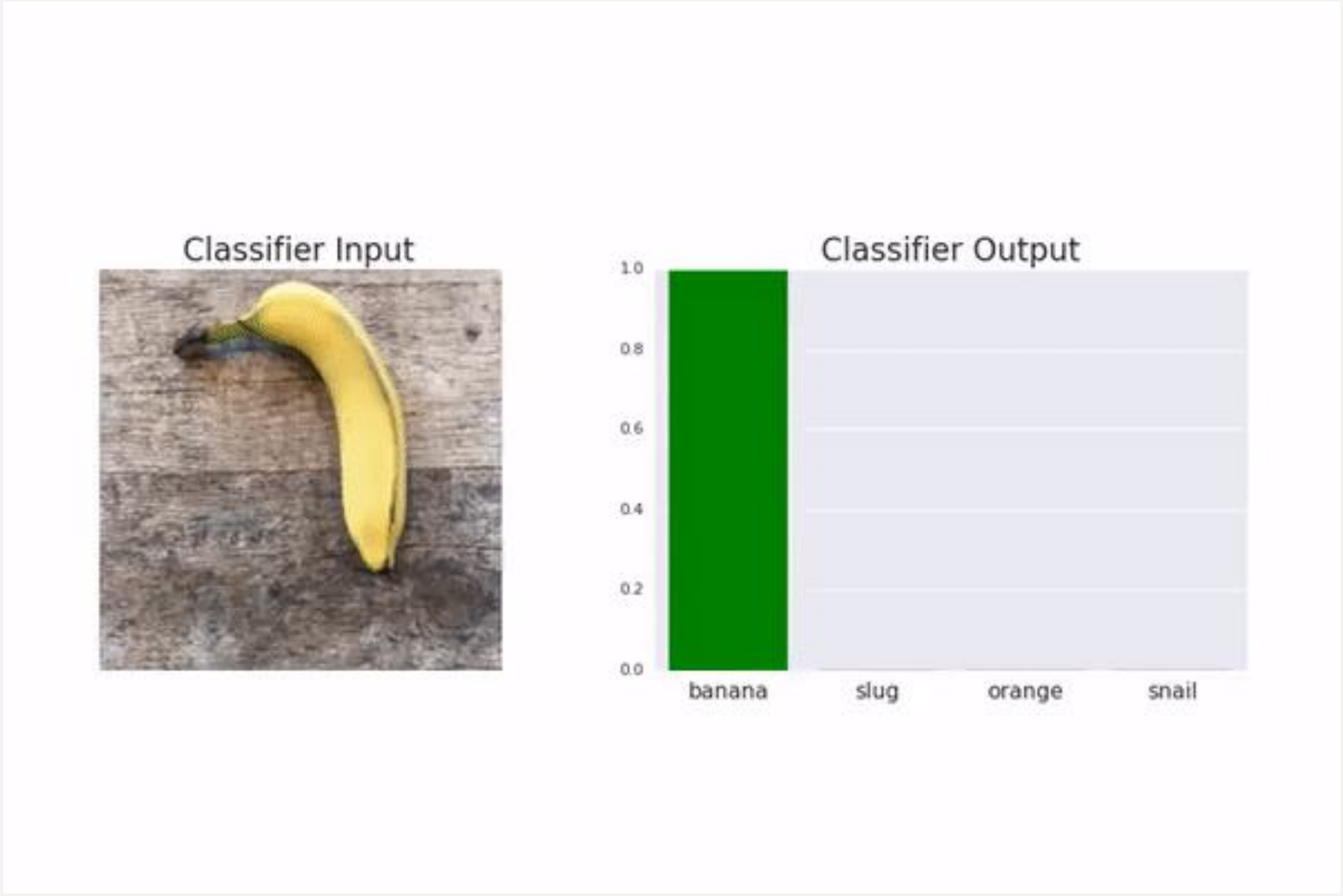
Attacking noise



Adversarial example: toaster



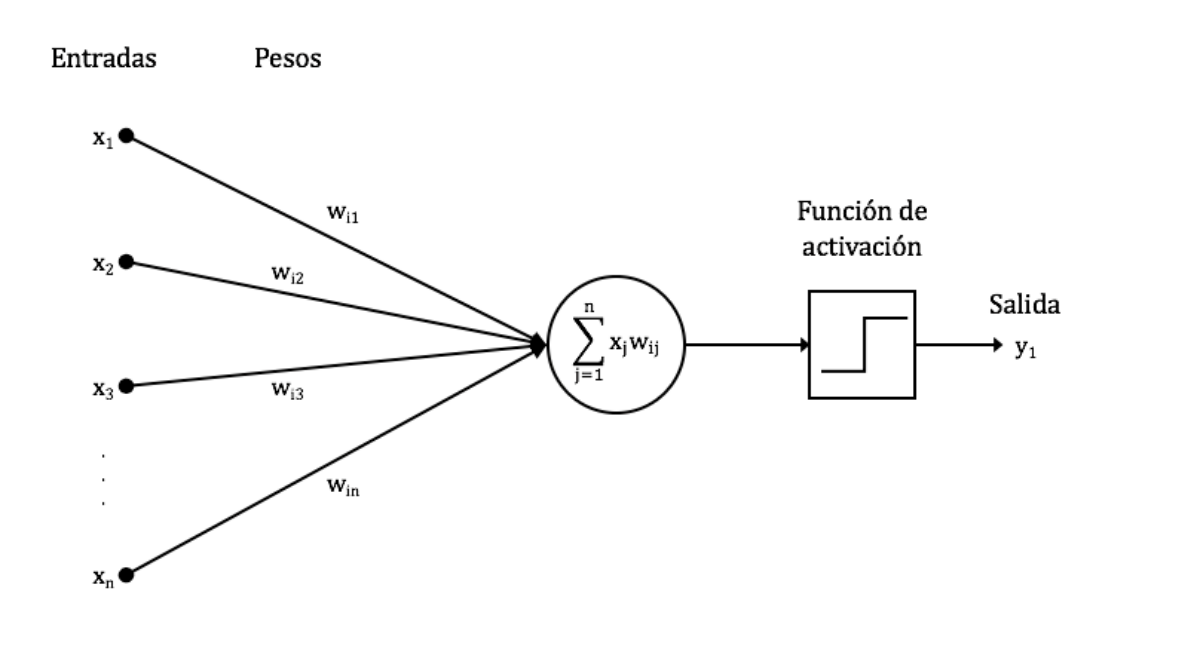
EJEMPLOS

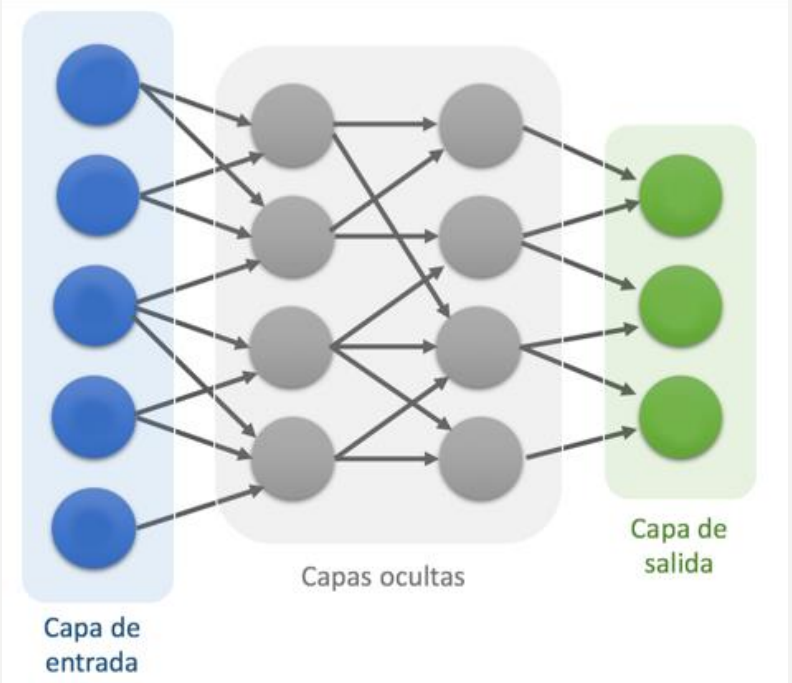

Classifier Input
Classifier Output
1.0
0.8
0.6
0.4
0.2
0.0
banana
slug
orange
snail

REDES NEURONALES

¿QUÉ ES UNA RED NEURONAL?

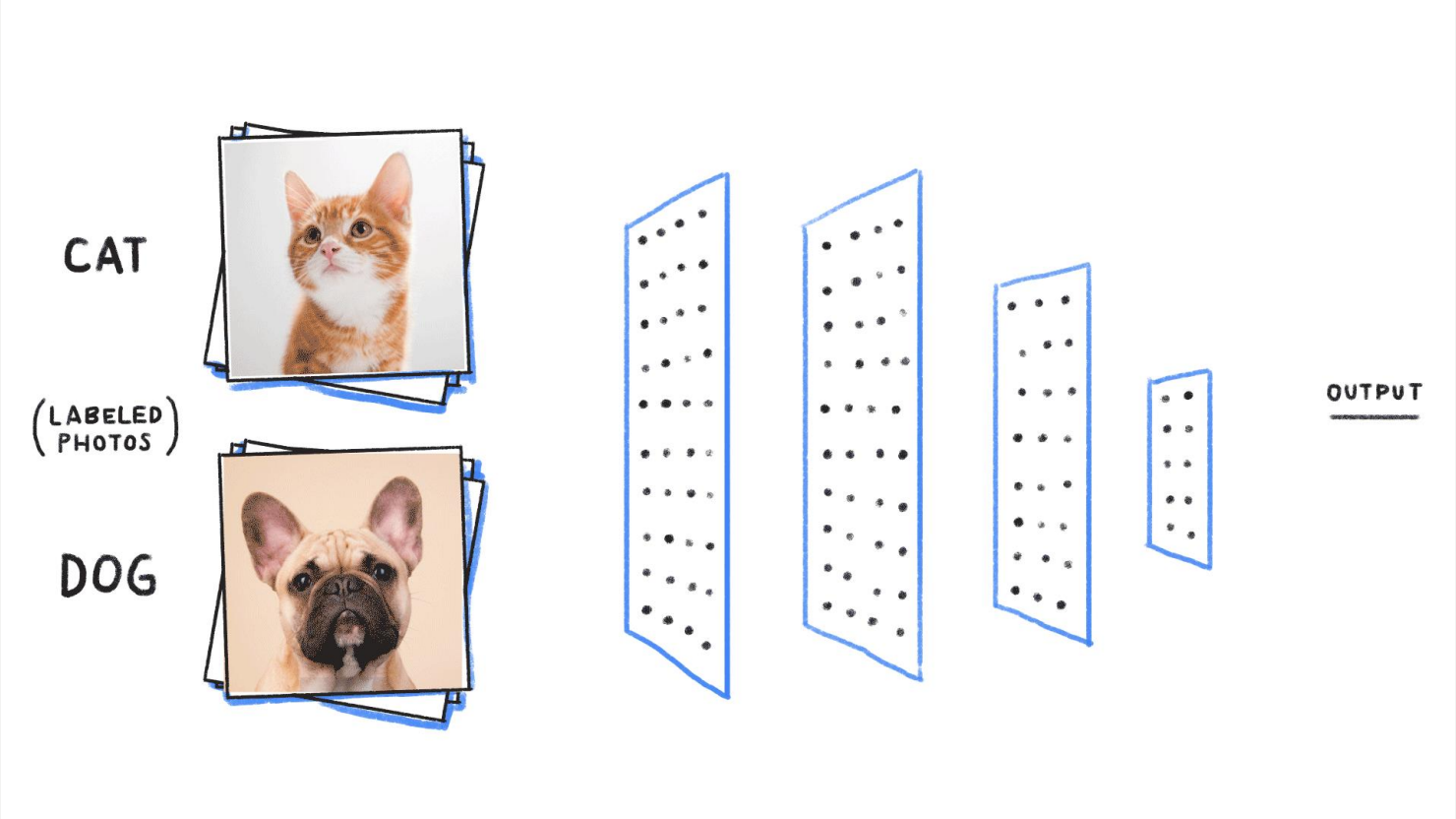
- Las redes neuronales son un modelo computacional inspirado en el comportamiento de las neuronas biológicas.
- Están constituidas por un conjunto de unidades (neuronas) que se encuentran conectadas entre sí formando capas.



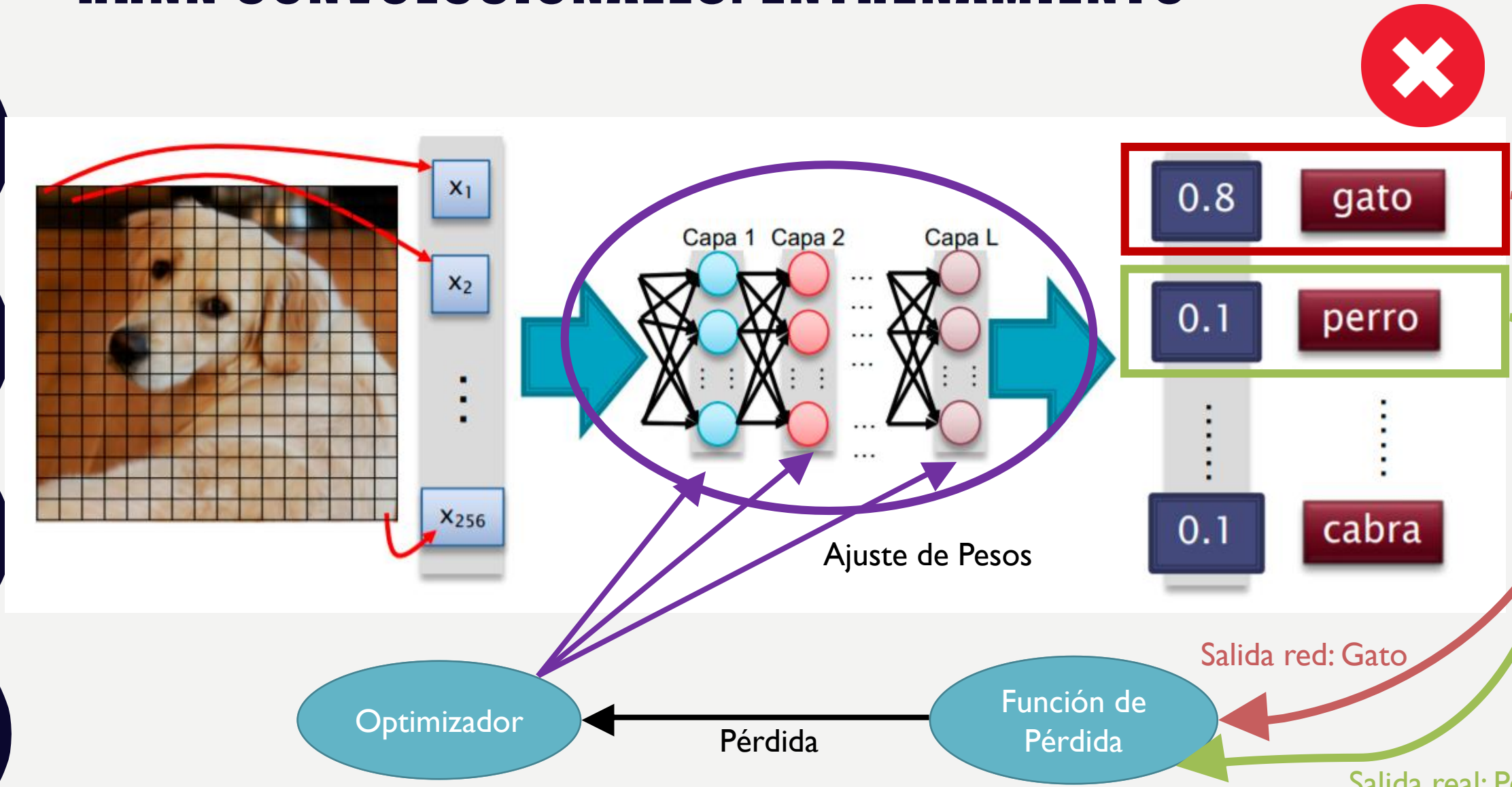


REDES NEURONALES CONVOLUCIONALES

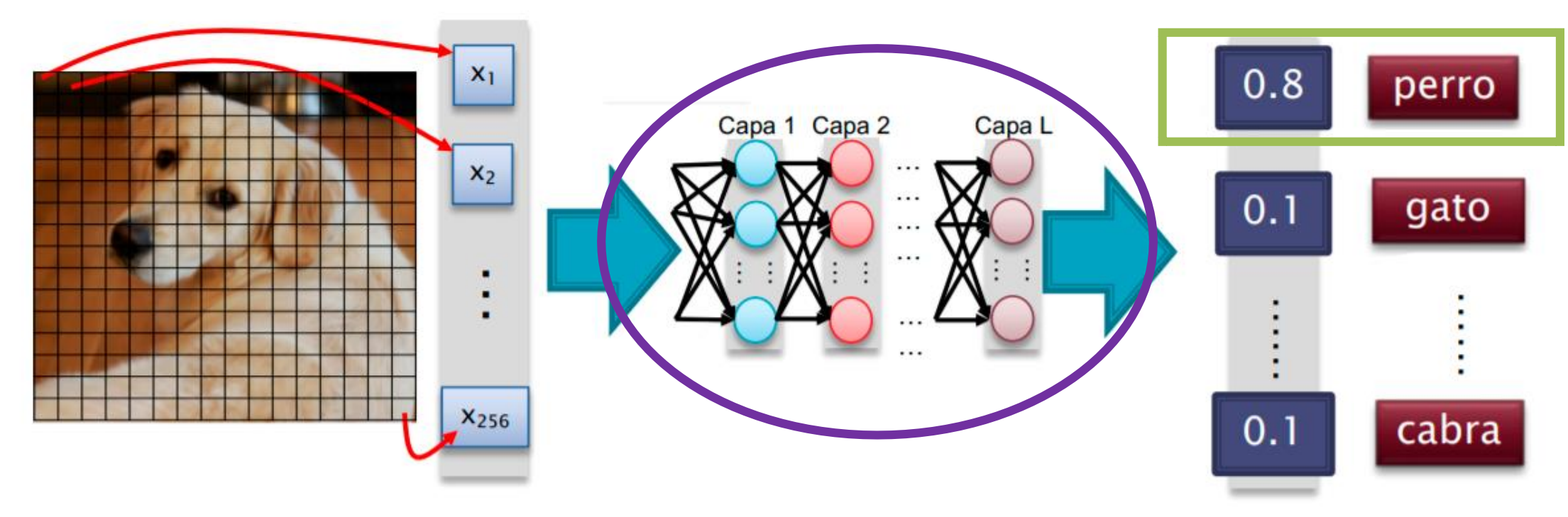
- Red Neuronal que procesa sus capas imitando al córtex visual del ojo humano para identificar distintas características y proporcionar a la red la capacidad de "ver".
- Uso principal → Tareas de visión artificial: Reconocimiento de imágenes



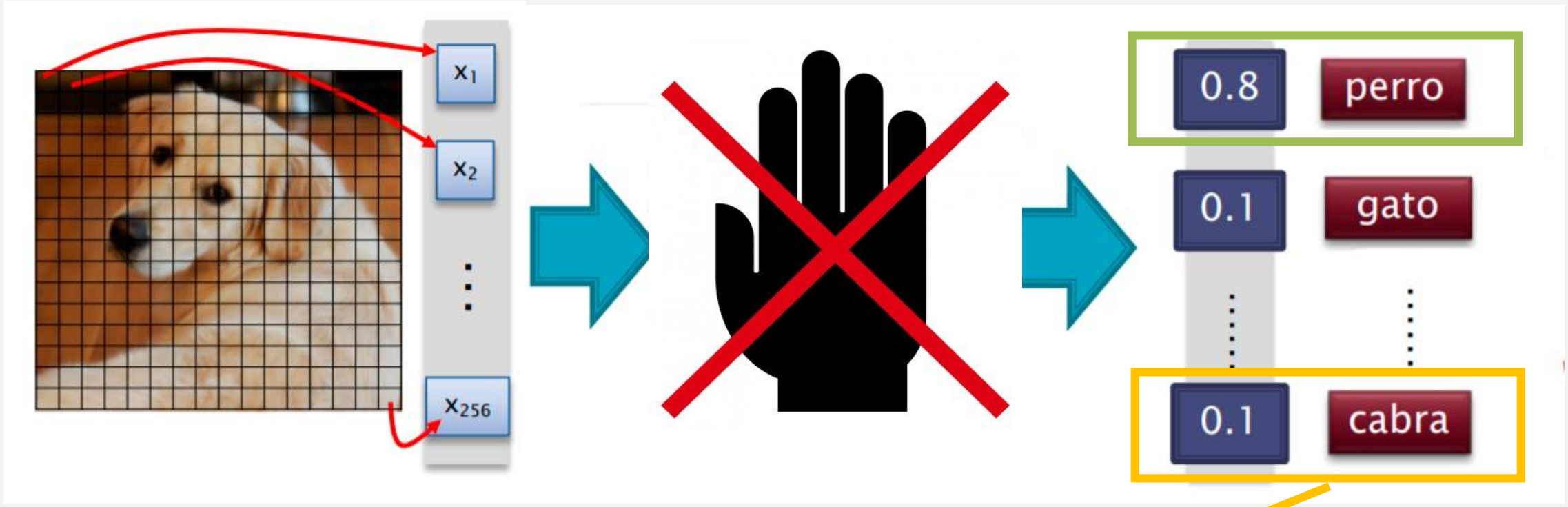
RRNN CONVOLUCIONALES: ENTRENAMIENTO



RRNN CONVOLUCIONALES: ENTRENAMIENTO



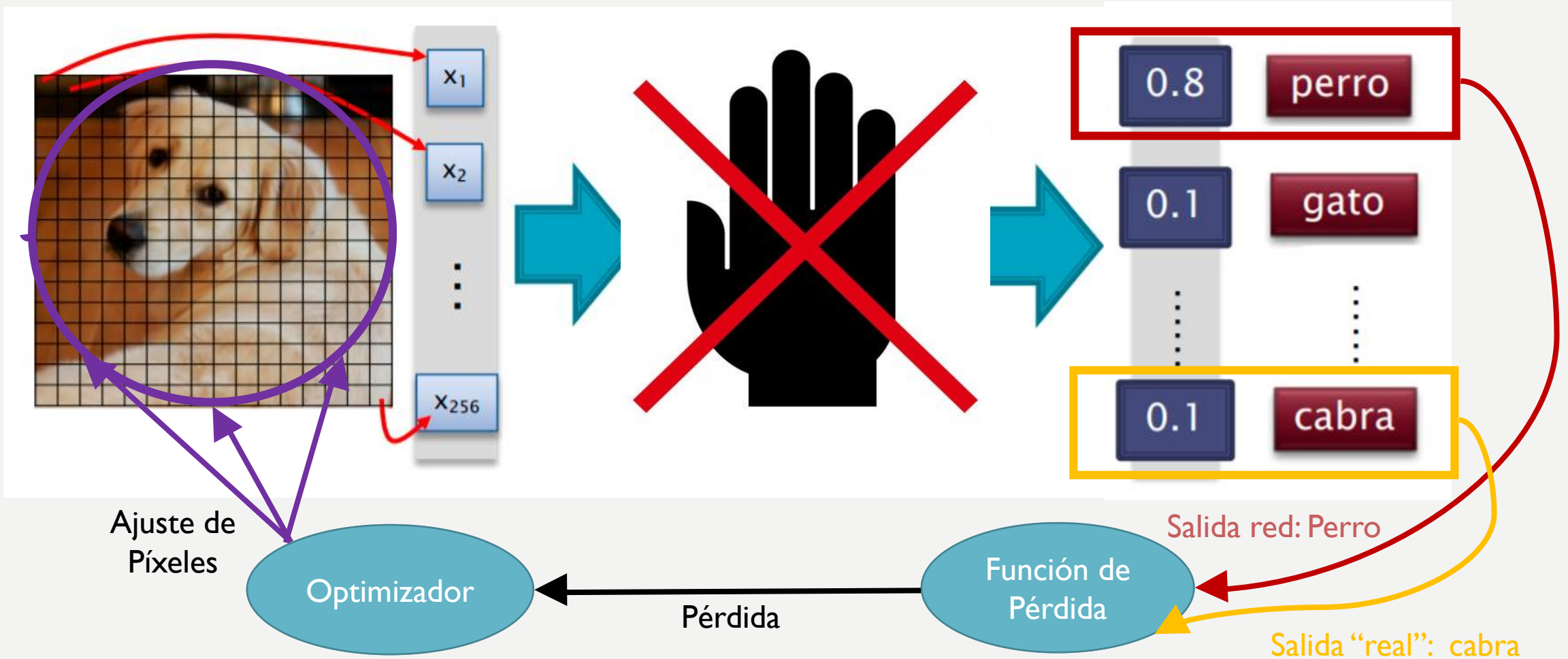
RRNN CONVOLUCIONALES: TEST



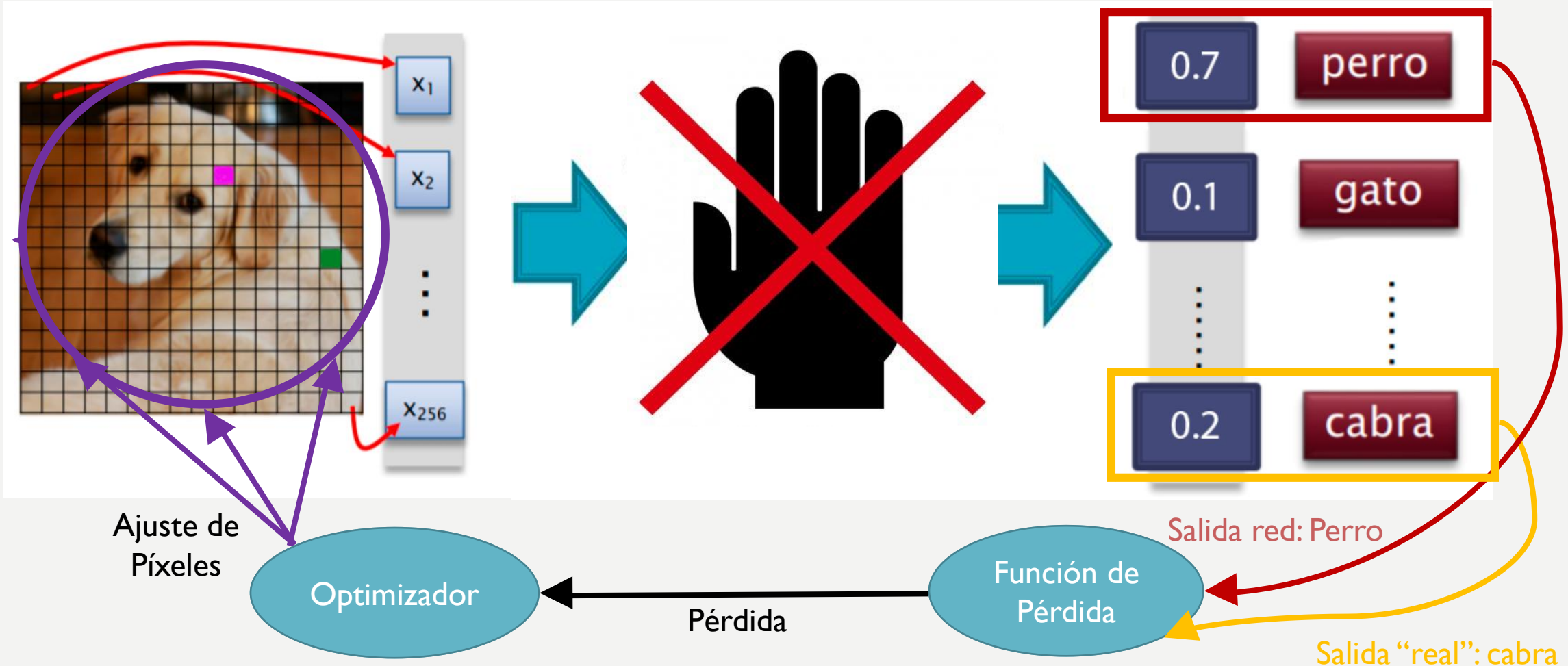
¿Y si quisiera que la red me devolviera "cabra"?

REALIZANDO UN ATAQUE ADVERSARIO

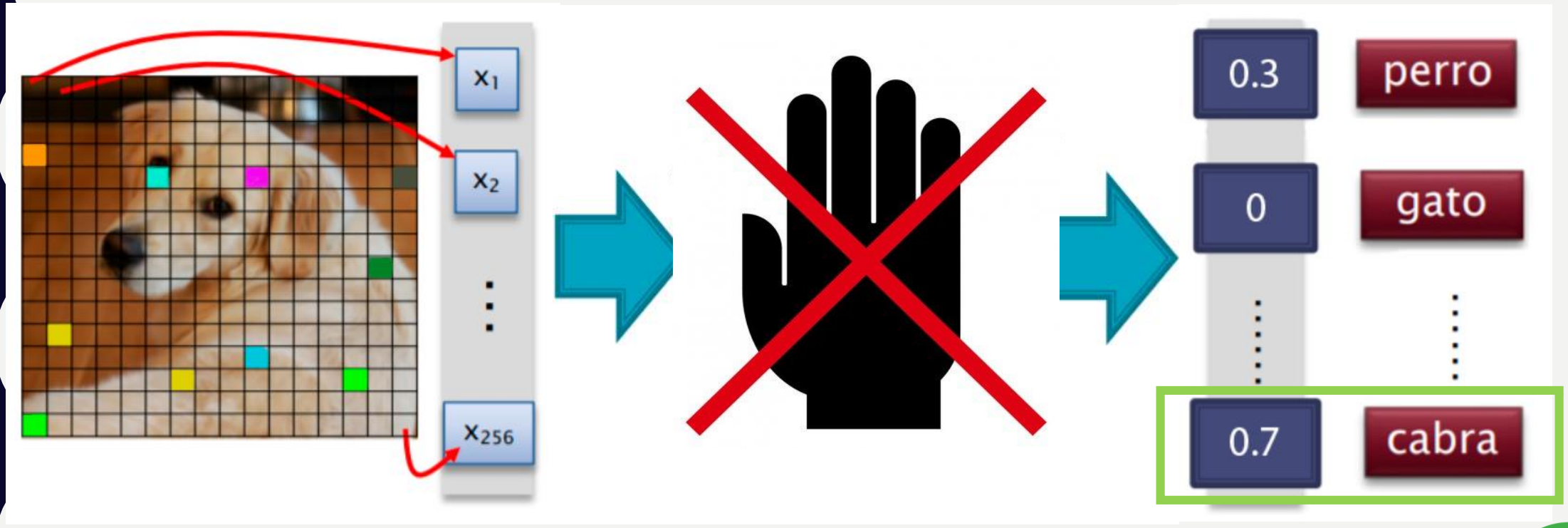
ATACANDO REDES NEURONALES



ATACANDO REDES NEURONALES



ATACANDO REDES NEURONALES



Conseguido!!

EJEMPLO PRÁCTICO

¿PREGUNTAS?

¡Muchas gracias por vuestra atención!

José Manuel Simón Ramos
jmsimonramos@gmail.com
linkedin.com/in/jose-manuel-simon-ramos/