


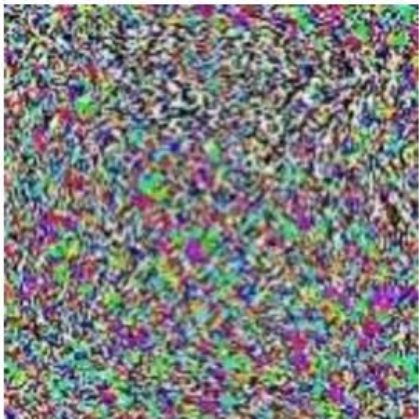



How to deal with adversarial images.

Sonia Portillo Clota

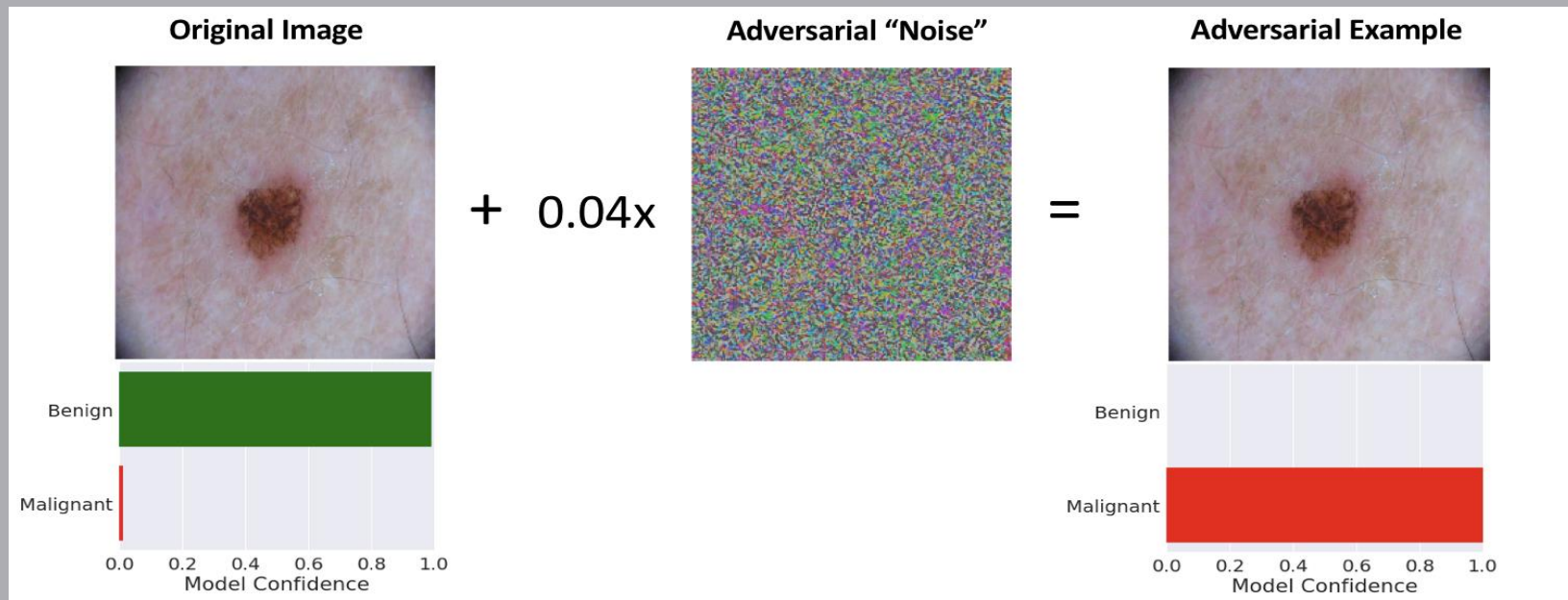
Que es una imagen adversaria?

Imagen que se ha modificado muy ligeramente, a simple vista parece que se han incorporado pixels sin importancia. Esto provoca que una red neuronal clasifique la imagen erróneamente.

Original Image		Adversarial "Noise"		Adversarial Example
	+		=	
<i>Panda</i> 99.95% confidence		$\frac{1}{255} \times \text{sign}(\nabla_x J(F, x, c))$		<i>Llama</i> 88.17% confidence

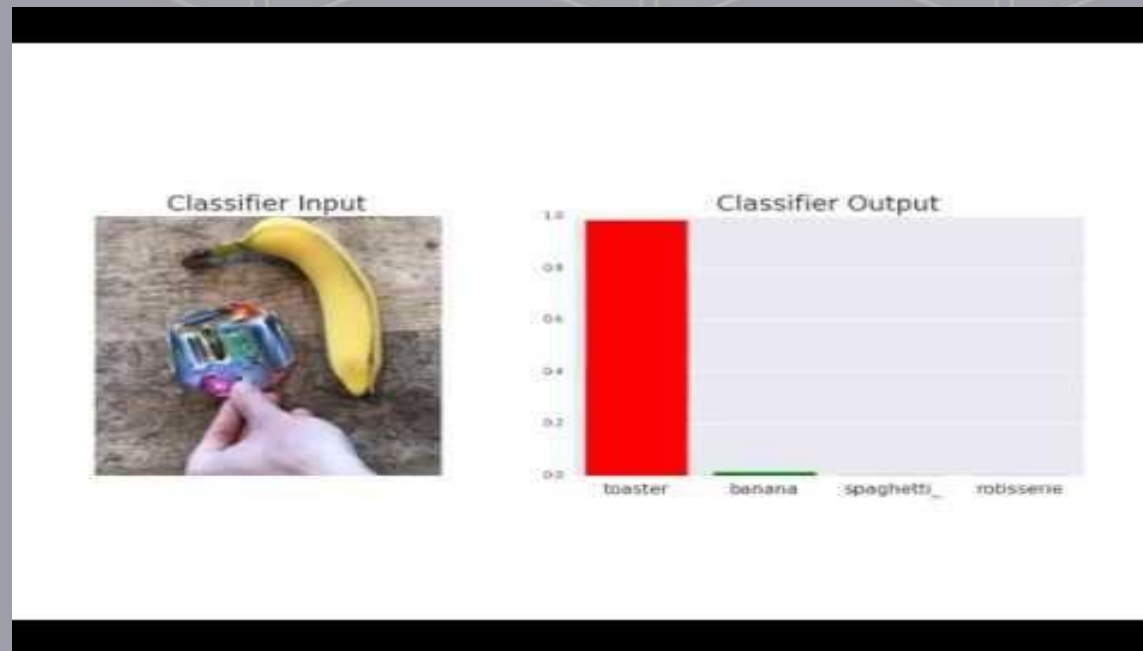
Que es una imagen adversaria?

Imagen que se ha modificado muy ligeramente, a simple vista parece que se han incorporado pixels sin importancia. Esto provoca que una red neuronal clasifique la imagen erróneamente.



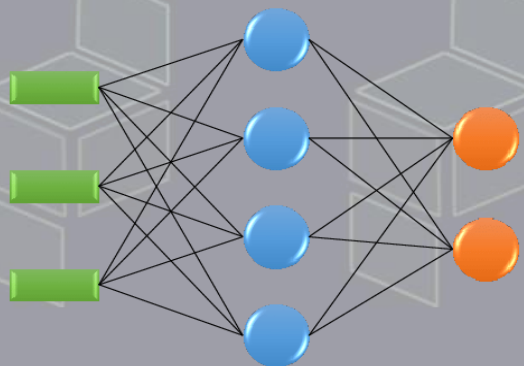
Que es una imagen adversarial?

Imagen que se ha modificado añadiendo un parche. Esto provoca que una red neuronal clasifique la imagen erróneamente.



Video: <https://www.youtube.com/watch?v=i1sp4X57TL4>

Posibles escenarios



+



=

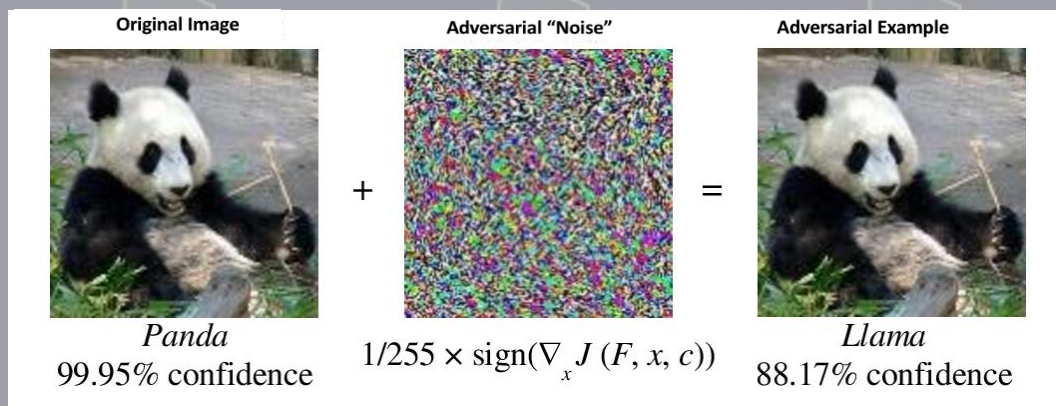
Caja Blanca

?

=

Caja Negra

Métodos para generar imágenes adversas



- FGSM
- BIM
- JSMA
- DeepFool
- C & W

<https://foolbox.readthedocs.io/en/stable/modules/attacks.html>

Métodos para general imágenes adversas



Expectation Over Transformation (EOT)
Método para generar ejemplos adversos
que incluso funcionan cuando la imagen
se transforma.

Defensas

- 1 – Reducir el numero de capas de la red.
- 2 – Preprocesado de imágenes y entrenamiento con imágenes preprocesadas.
- 3 – Entrenar la red con imágenes adversas.
- 4 – Utilizar redes neuronales bayesianas

Para más información

<https://github.com/tensorflow/cleverhans>

<https://github.com/bethgelab/foolbox/tree/d3721991ab31c078c0a1c1a9ba22e4a2480a1b04>

Katy Warr. Strengthening Deep Neural Networks: Making AI Less Susceptible to Adversarial Trickery. O'Reilly

