DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Modelos con adversario [GANs]

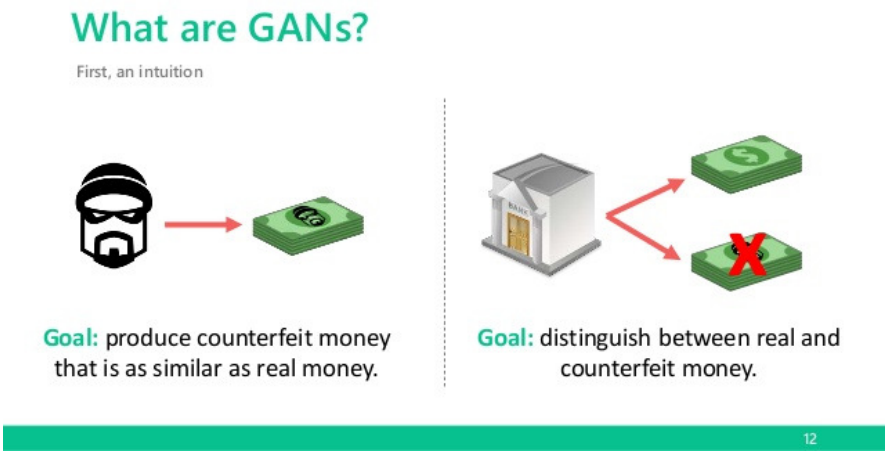
Fernando Berzal, berzal@acm.org

Modelos con adversario

http://xkcd.com/1425/

Modelos con adversario

GANs
Generative Adversarial Networks



SlideShare, Thomas da Silva Paula, HP

Modelos con adversario

GANs
Generative Adversarial Networks

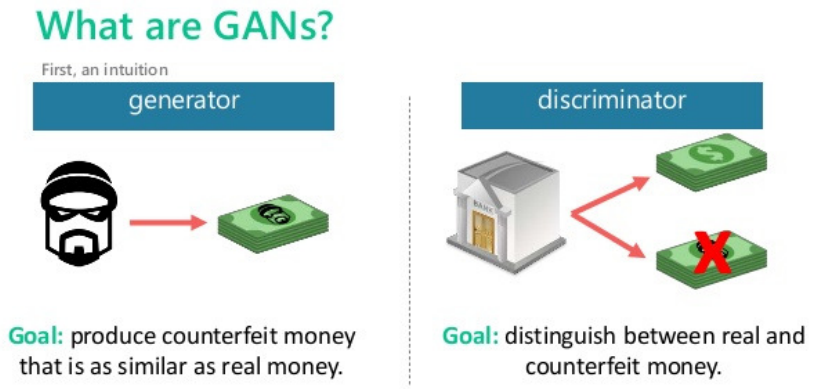
Combinación de dos modelos:

- **Modelo discriminativo** (tradicional):
  Probabilidad condicional, P(y|x)

- **Modelo generativo**:
  Probabilidad conjunta, P(x,y)

Modelos con adversario

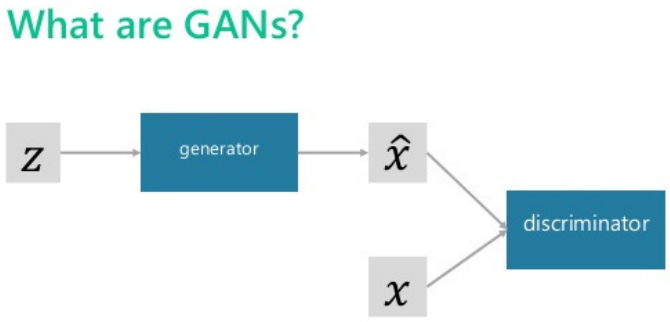
GANs
Generative Adversarial Networks



SlideShare, Thomas da Silva Paula, HP

Modelos con adversario

GANs
Generative Adversarial Networks



SlideShare, Thomas da Silva Paula, HP

Modelos con adversario

GANs

Generative Adversarial Networks

Los dos modelos compiten entre sí:

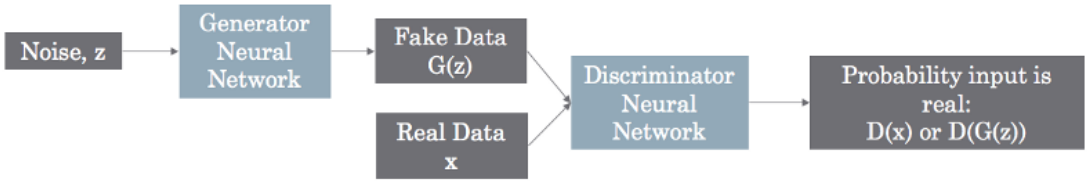
- El **modelo generativo** intenta construir instancias que confundan al **modelo discriminativo**.
- El **modelo discriminativo** utiliza tanto el conjunto de entrenamiento como las muestras sintetizadas por **modelo discriminativo** para ser más robusto.

Modelos con adversario

GANs

Generative Adversarial Networks

Más formalmente:



Modelos con adversario

GANs

Generative Adversarial Networks

Como problema de optimización:

$$\min_G \max_D V(D,G) = \underbrace{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})]}_{\text{real}} + \underbrace{\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]}_{\text{fake}}.$$

SOLUCIÓN

Algoritmo de aprendizaje basado
en el gradiente ascendente para el discriminador
y en el gradiente descendente para el generador

Modelos con adversario

GANs

Generative Adversarial Networks

Algoritmo original

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

  for k steps **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  end for

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

Modelos con adversario

Modelo generativo
Probabilidad conjunta, P(x,y)

Modelo de la distribución de probabilidad que da lugar a los datos observados en el conjunto de entrenamiento.

Modelos con adversario

Modelo generativo
Ejemplos de muestras generadas

Yellow boxes are real data samples that are nearest matches to last column of fake images. This shows the generator didn't merely memorize training examples

Modelos con adversario

Modelo generativo
Ejemplos de muestras generadas

Dormitorios
https://arxiv.org/abs/1511.06434

Modelos con adversario

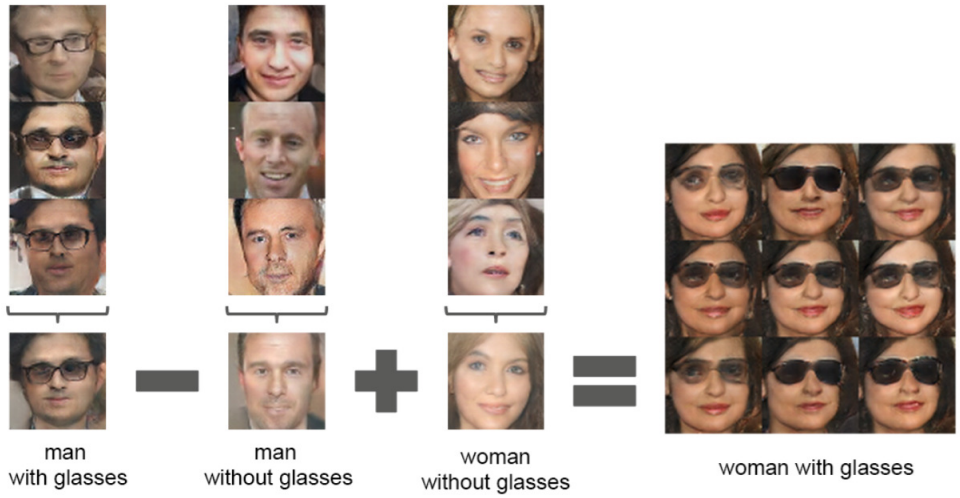
Modelo generativo
Ejemplos de muestras generadas

Orientación de las caras

Modelos con adversario

Modelo generativo

Ejemplos de muestras generadas



Aritmética de caras…

Modelos con adversario

Modelo generativo

Ejemplos de muestras generadas

Caras generadas (fila superior),
envejecidas (fila central)
y con una "sonrisa" (fila inferior)

http://www.foldl.me/2015/conditional-gans-face-generation/

Modelos con adversario

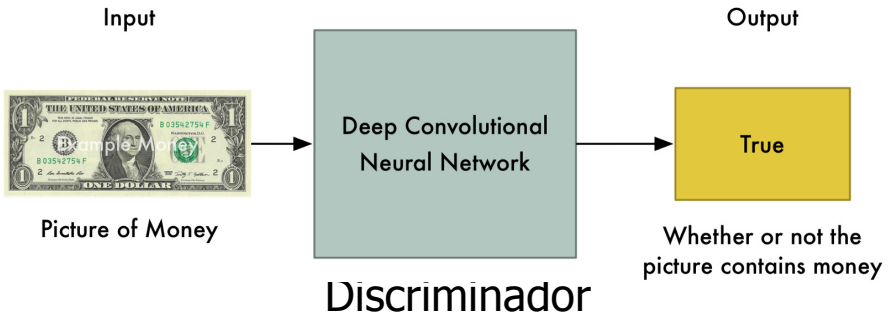
Síntesis de imágenes

https://thispersondoesnotexist.com/

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

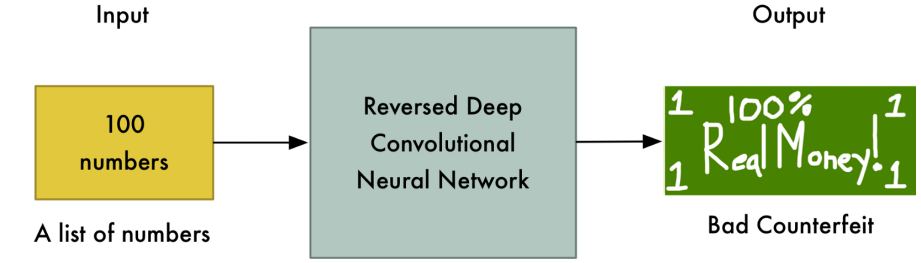
Modelos con adversario

Modelo generativo

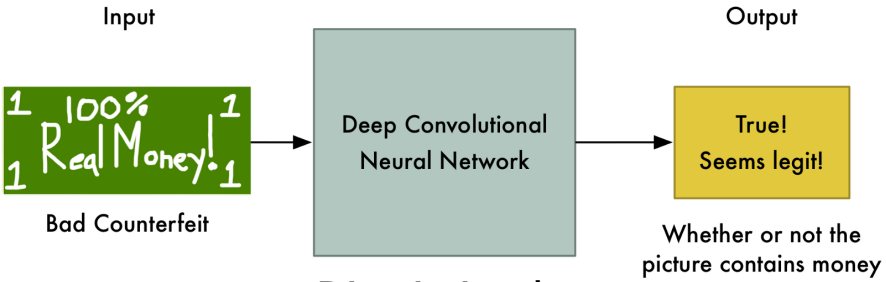


Modelos con adversario

Entrenamiento: Al principio...



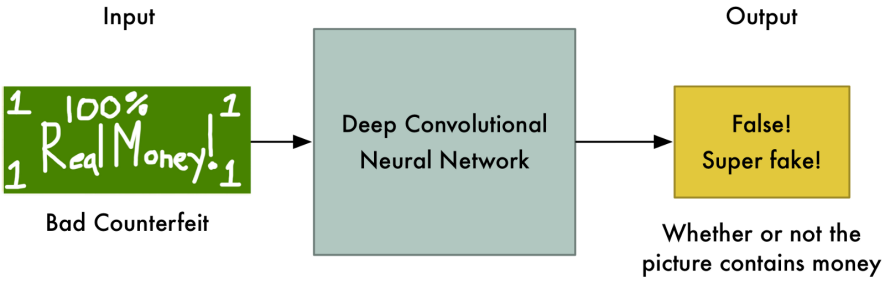
Generador



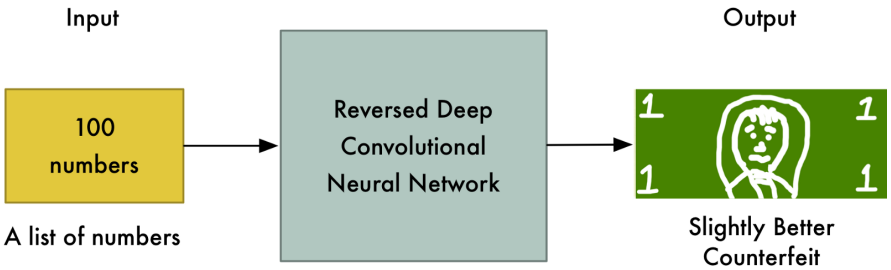
Discriminador

Modelos con adversario

Entrenamiento: Con algo más de práctica...



Discriminador



Generador

Modelos con adversario

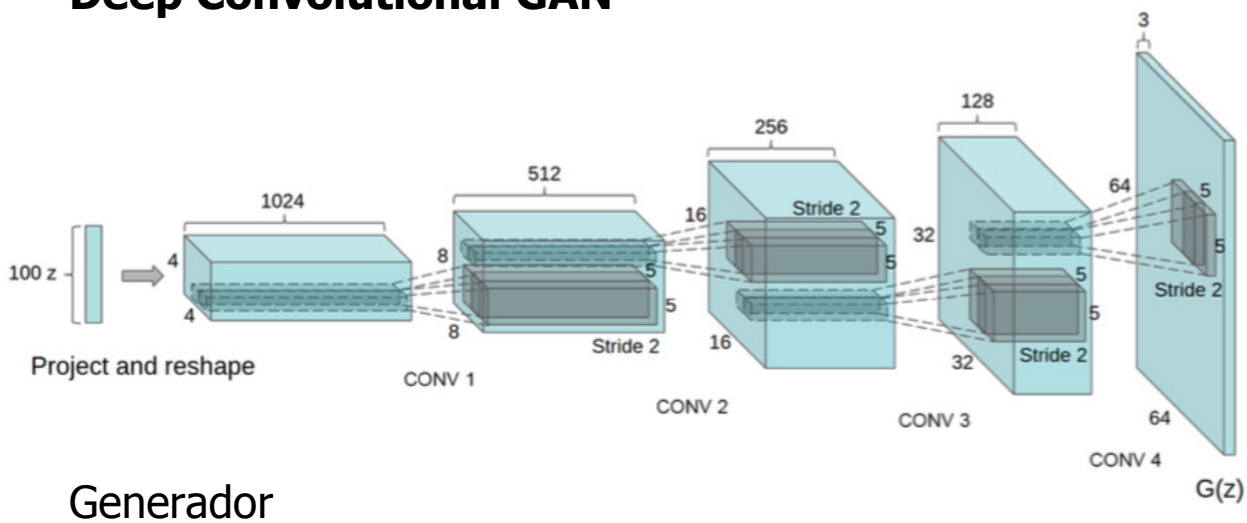
DCGAN
Deep Convolutional GAN

Recomendaciones arquitectónicas:

- Strided convolutions: Sustituir las capas de pooling por convoluciones con paso >1.
- Sin capas completamente conectadas: La salida se conecta directamente a las capas convolutivas.
- Entrenamiento usando normalización por lotes (escala las entradas de cada capa, de forma que tengan media 0 y varianza 1).
- Generador con unidades ReLU y discriminador con unidades "leaky" ReLU (para imágenes en color).

Modelos con adversario

DCGAN
Deep Convolutional GAN



Generador

Modelos con adversario

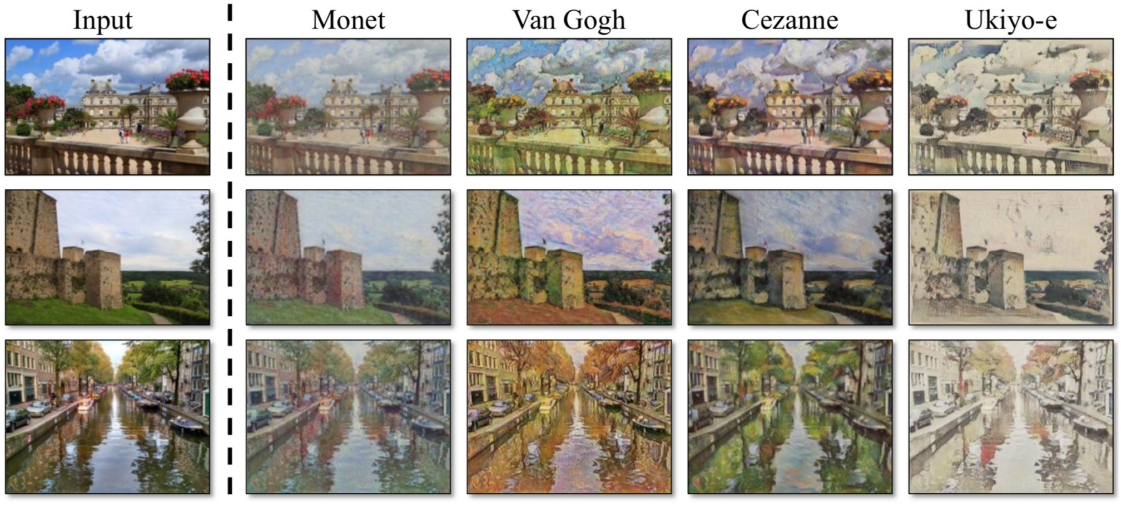
Aplicaciones

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

Modelos con adversario

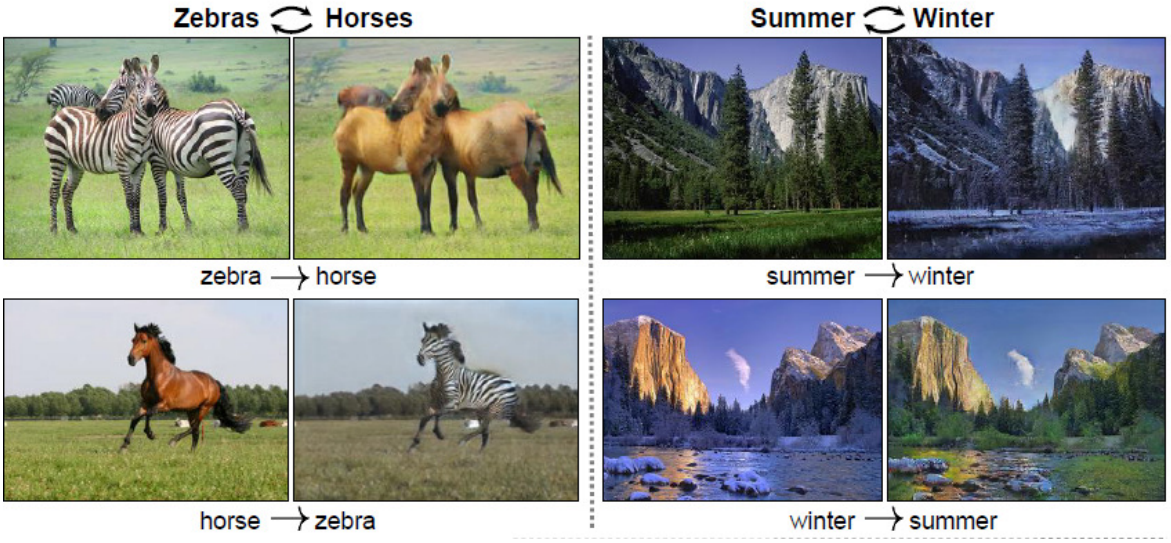
Aplicaciones



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

Modelos con adversario

Aplicaciones



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

Modelos con adversario

Aplicaciones

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

Modelos con adversario

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

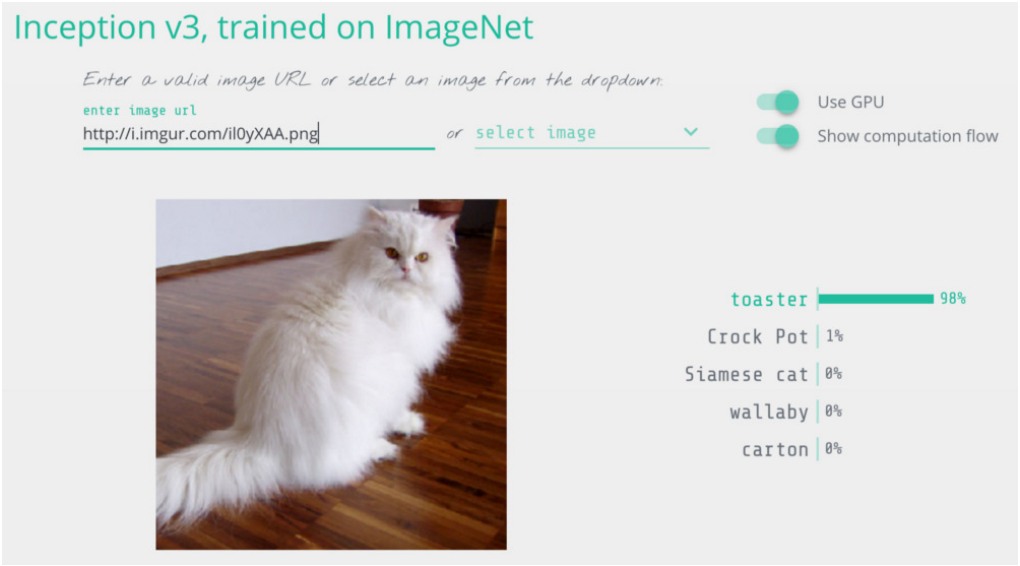
Modelos con adversario

Aplicaciones: "You sketch, the AI paints"

GauGAN, NVIDIA, CVPR'2019

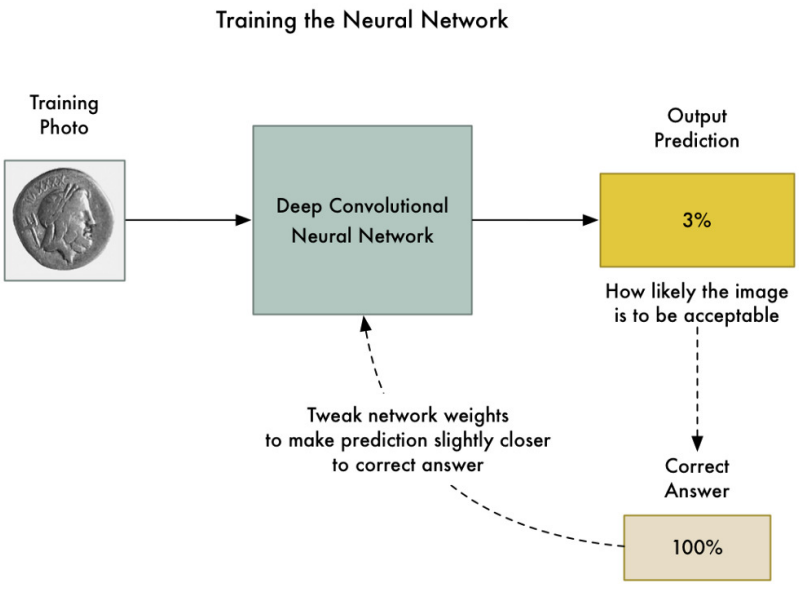
Modelos con adversario

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



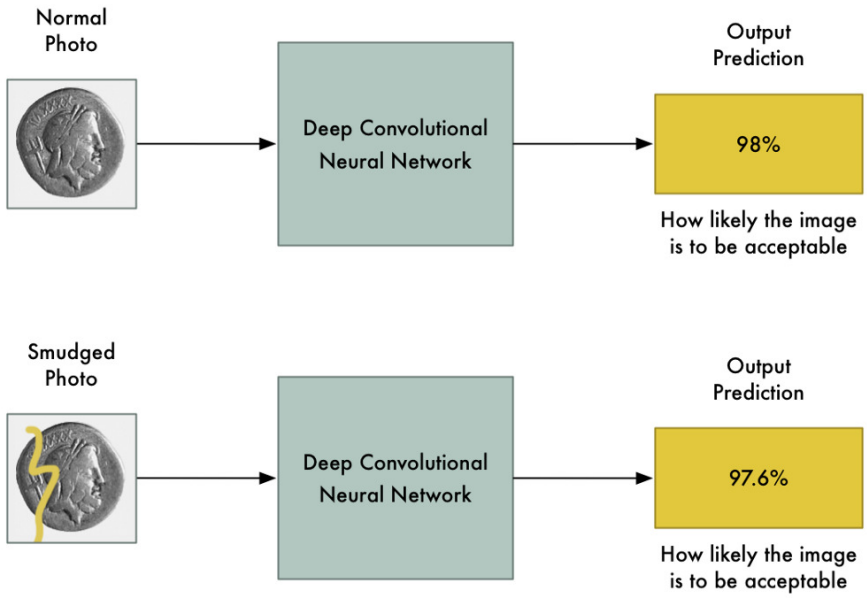
Modelos con adversario

El proceso de entrenamiento habitual...



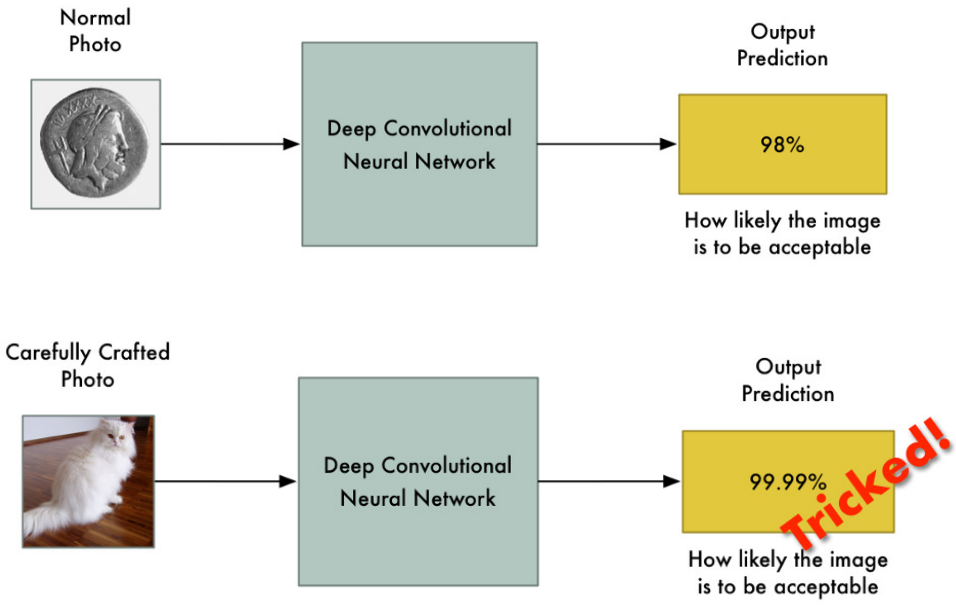
Modelos con adversario

Lo deseable...



Modelos con adversario

Lo que puede pasar...

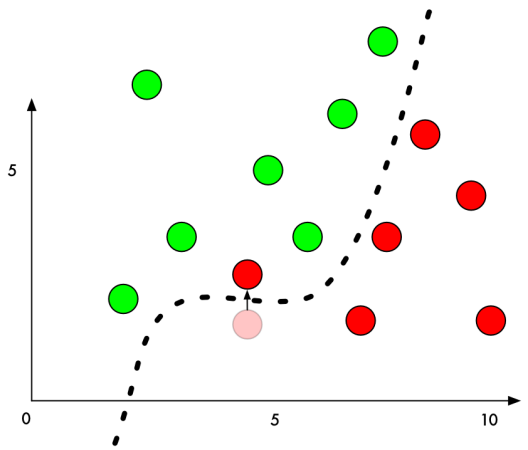


Modelos con adversario

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

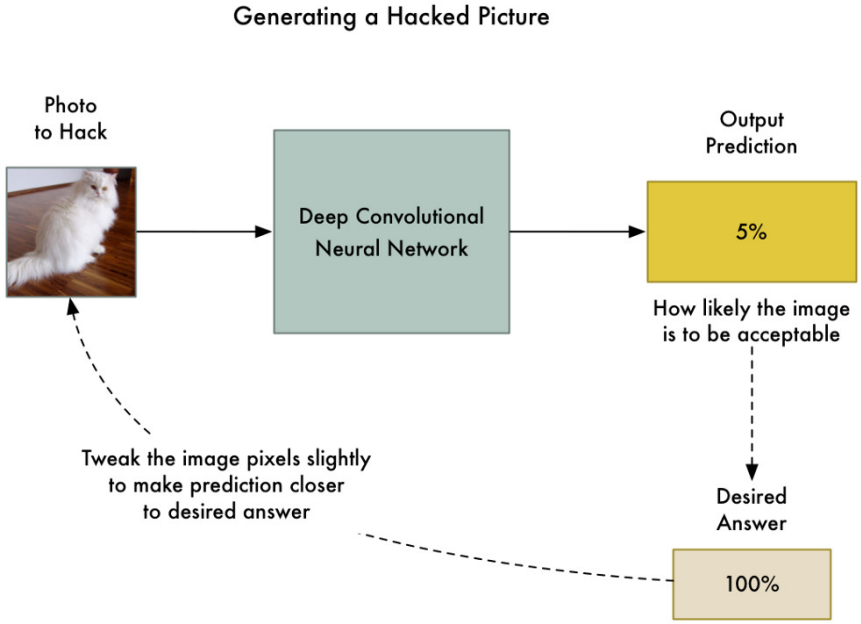
Si conocemos la red, podemos saber exactamente cómo modificar mínimamente la entrada para confundir a la red neuronal…

… en la dirección del gradiente !!!



Modelos con adversario

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



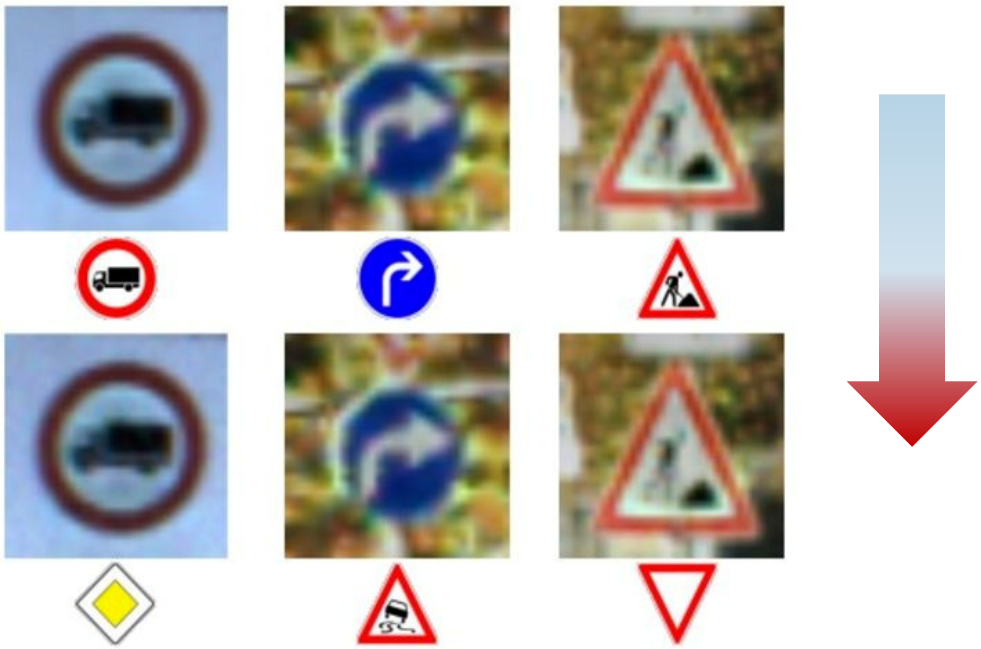
Modelos con adversario

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

https://transcranial.github.io/keras-js/#/inception-v3

Modelos con adversario

Implicaciones en seguridad



Referencias

GANs

- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville & Yoshua Bengio:
  "Generative Adversarial Networks"
  arXiv, June 2014
  https://arxiv.org/abs/1406.2661

DCGANs

- Alec Radford, Luke Metz & Soumith Chintala:
  "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks"
  arXiv, November 2015
  https://arxiv.org/abs/1511.06434

Referencias

Adversarial examples (ejemplos diseñados por un adversario)

- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow & Rob Fergus:
  "Intriguing properties of neural networks"
  arXiv, December 2013
  https://arxiv.org/abs/1312.6199
- Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy:
  "Explaining and Harnessing Adversarial Examples"
  arXiv, December 2014 & ICLR'2015
  https://arxiv.org/abs/1412.6572
- Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik & Ananthram Swami:
  "Practical Black-Box Attacks against Machine Learning"
  arXiv, February 2016 & ACM CCS'2017
  https://arxiv.org/abs/1602.02697