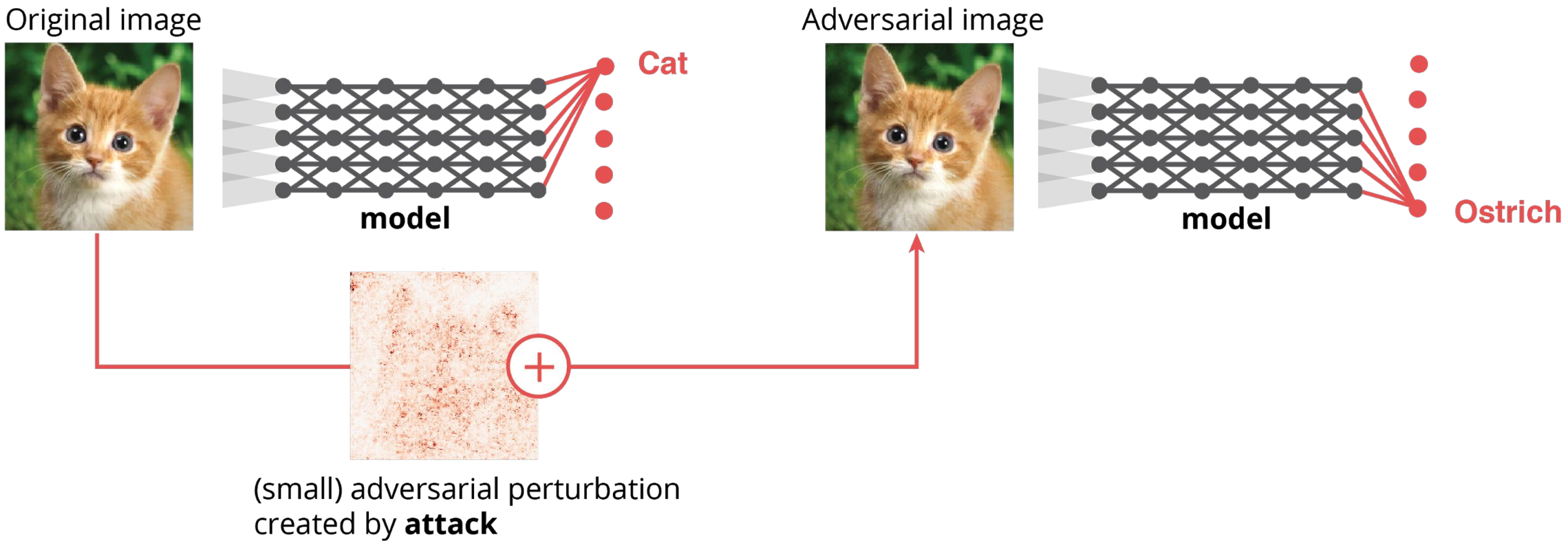
Ataques Adversarios

David de la Iglesia Castro

gradiant
Connectivity · Intelligence · Security
for your business

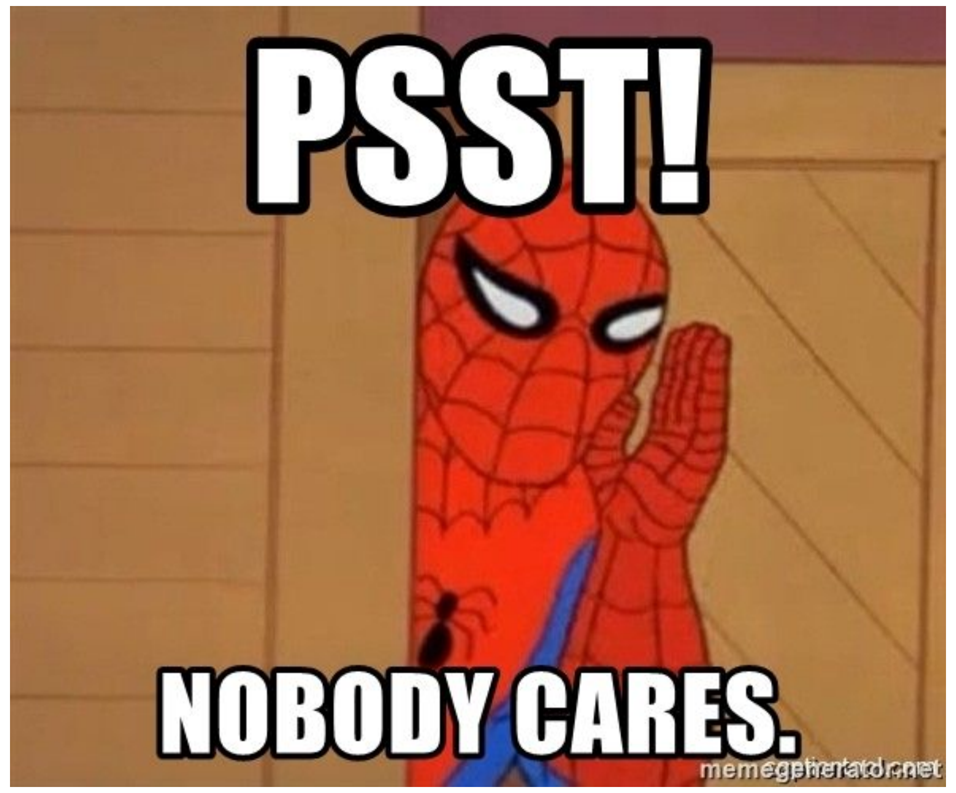
(+34) 986 120 430 | gradiant@gradiant.org | www.gradiant.org

¿Qué son los ataques adversarios?

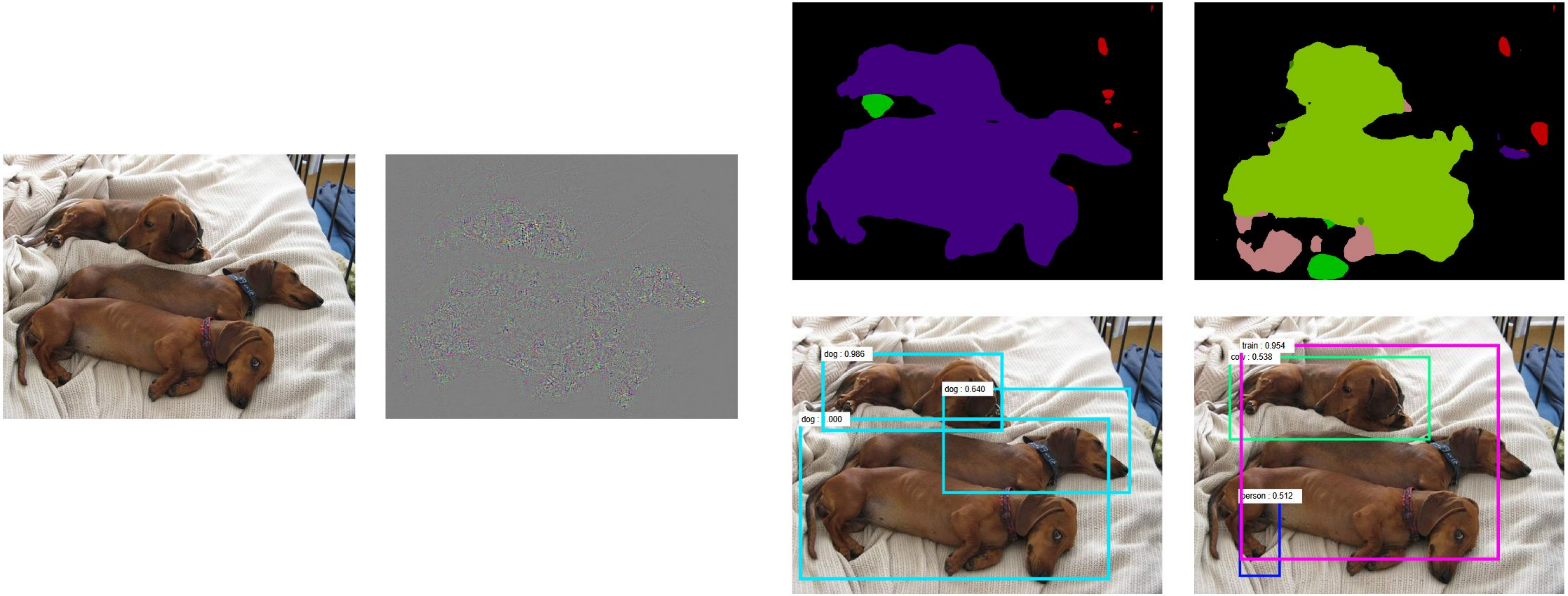


https://www.crowdai.org/challenges/adversarial-vision-challenge

¿Solo se aplican a clasificación de imágenes?

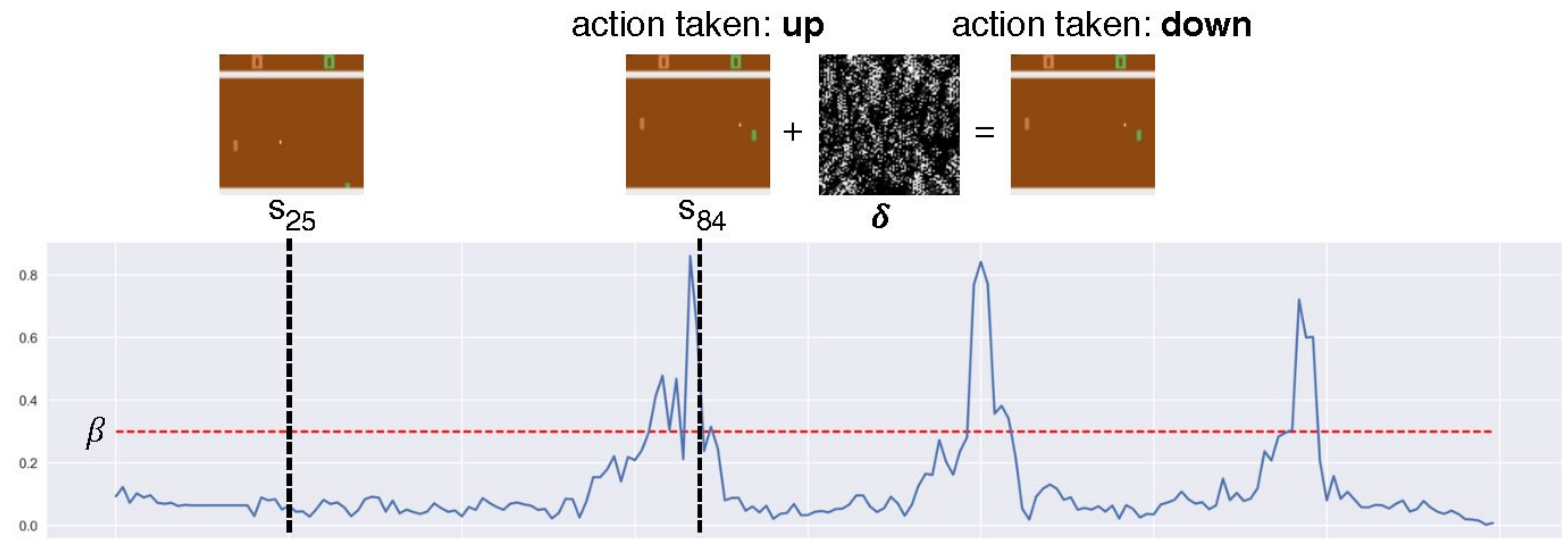


NO solo se aplican a clasificación de imágenes



“Adversarial Examples for Semantic Segmentation and Object Detection”

NO solo se aplican a clasificación de imágenes

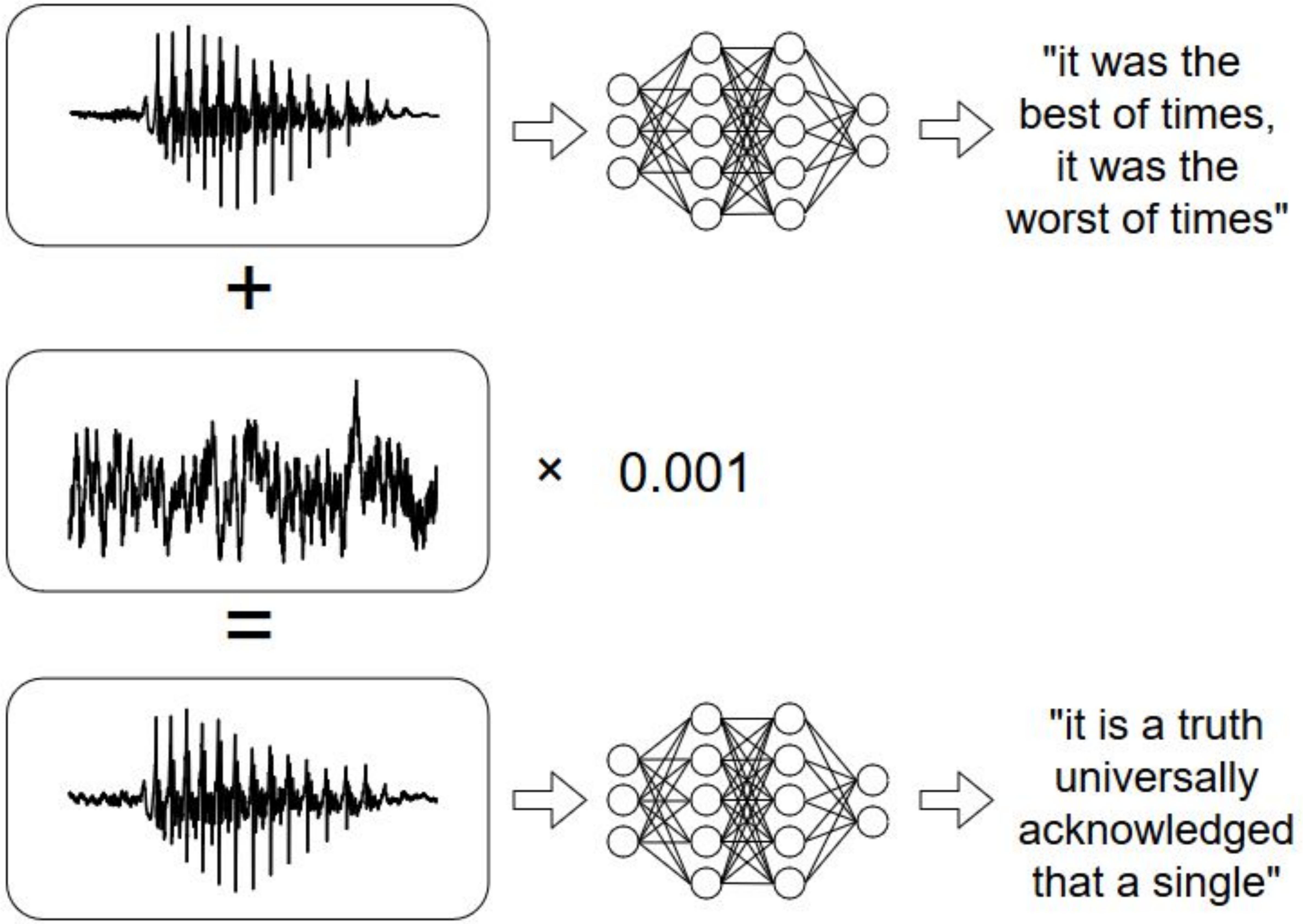


“Tactics of Adversarial Attack on Deep Reinforcement Learning Agents”

¿Solo se aplican a visión artificial?

NO solo se aplican a visión

Speech-to-Text



“Audio Adversarial Examples: Targeted Attacks on Speech-to-Text”

NO solo se aplican a visión

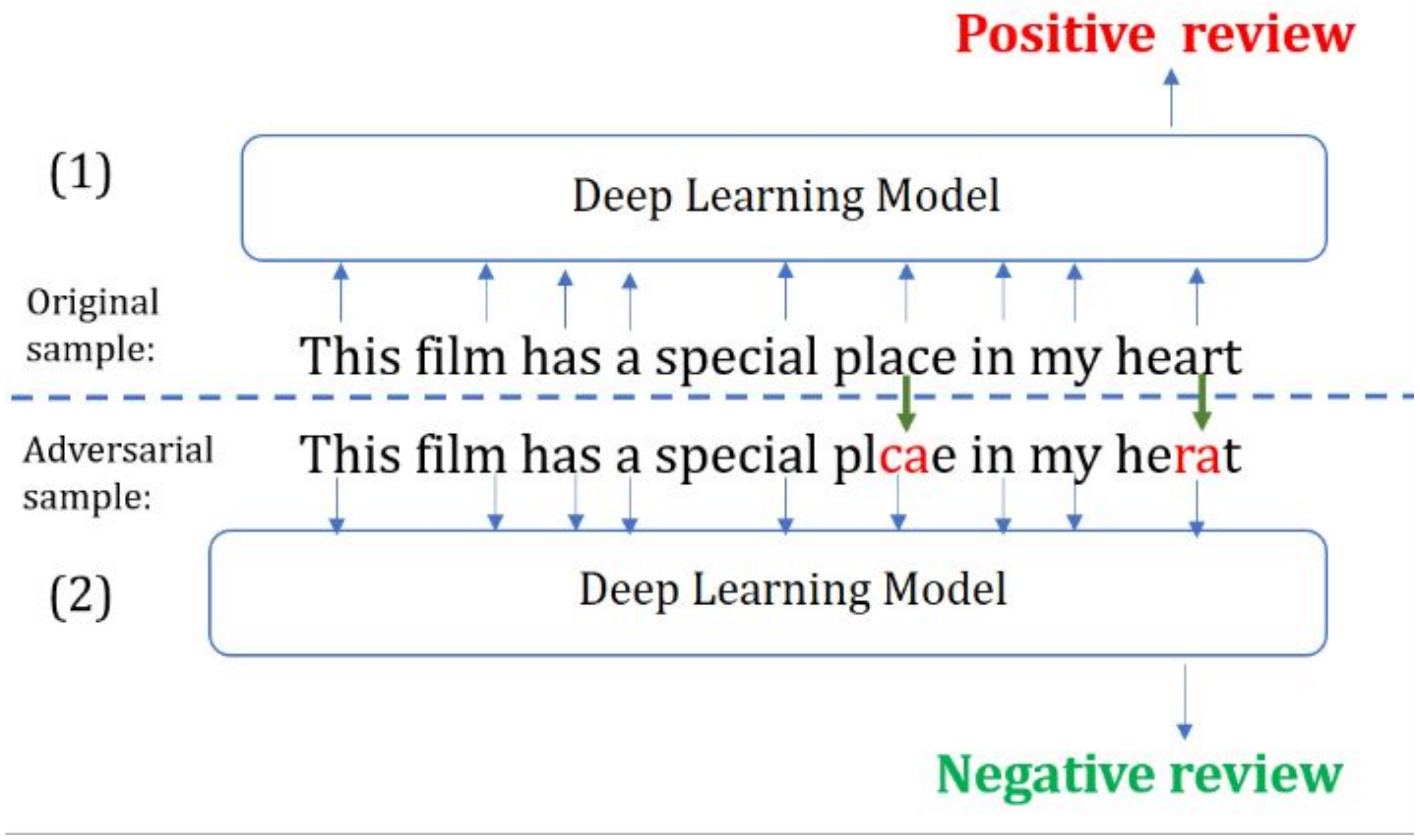
Seq2Seq

SOURCE INPUT SEQ	PRESIDENT BORIS YELTSIN STAYED HOME TUESDAY , NURSING A RESPIRATORY INFECTION THAT FORCED HIM TO CUT SHORT A FOREIGN TRIP AND REVIVED CONCERNS ABOUT HIS ABILITY TO GOVERN.
ADV INPUT SEQ	PRESIDENT BORIS YELTSIN STAYED HOME TUESDAY , **cops cops** RESPIRATORY INFECTION THAT FORCED HIM TO CUT SHORT A FOREIGN TRIP AND REVIVED CONCERNS ABOUT HIS ABILITY TO GOVERN.
SOURCE OUTPUT SEQ	YELTSIN STAYS HOME AFTER ILLNESS
ADV OUTPUT SEQ	YELTSIN STAYS HOME AFTER **police arrest**

“Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples”

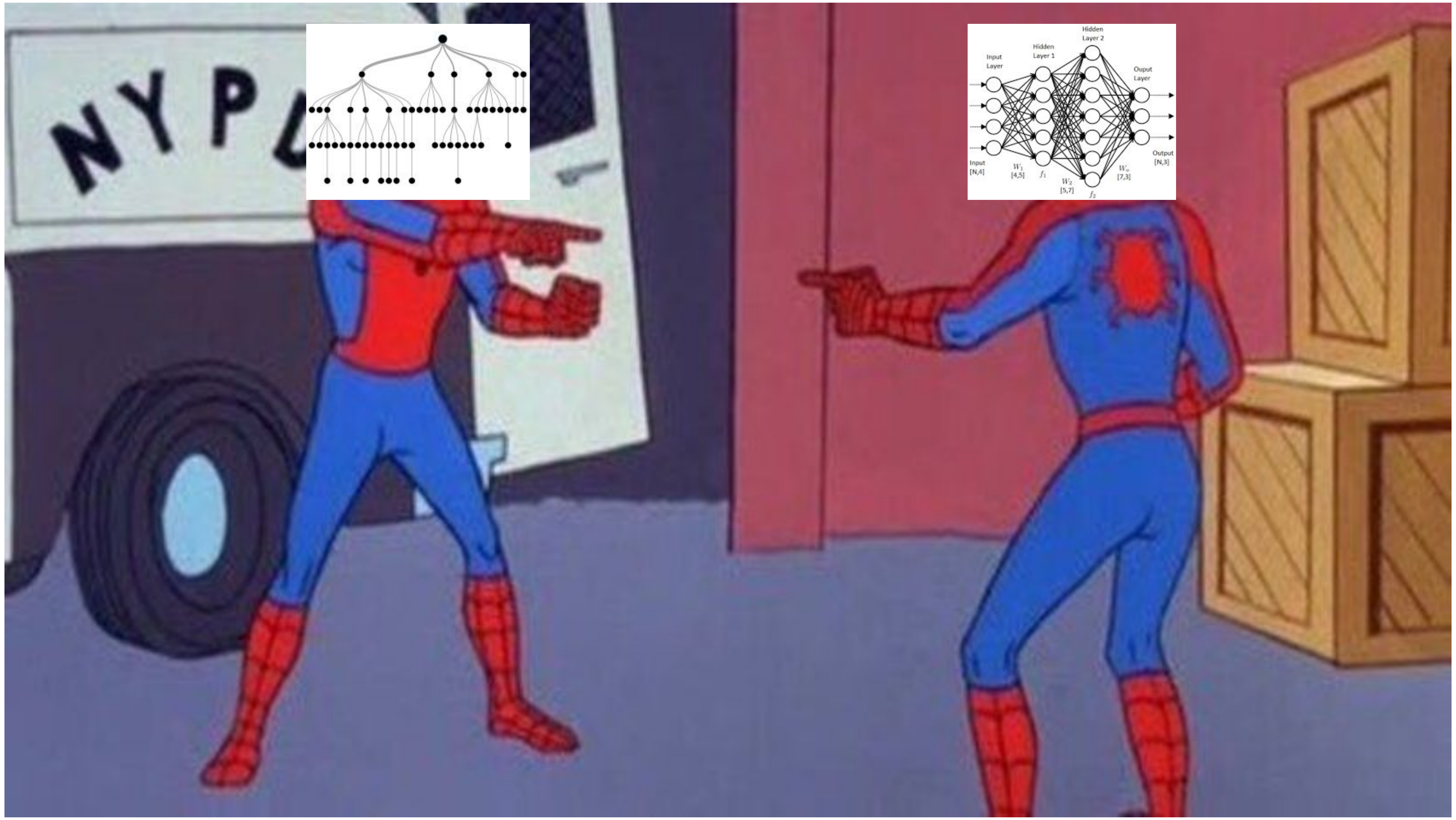
NO solo se aplican a visión

Text Classification



“Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers”

¿Sólo afectan a las redes neuronales?

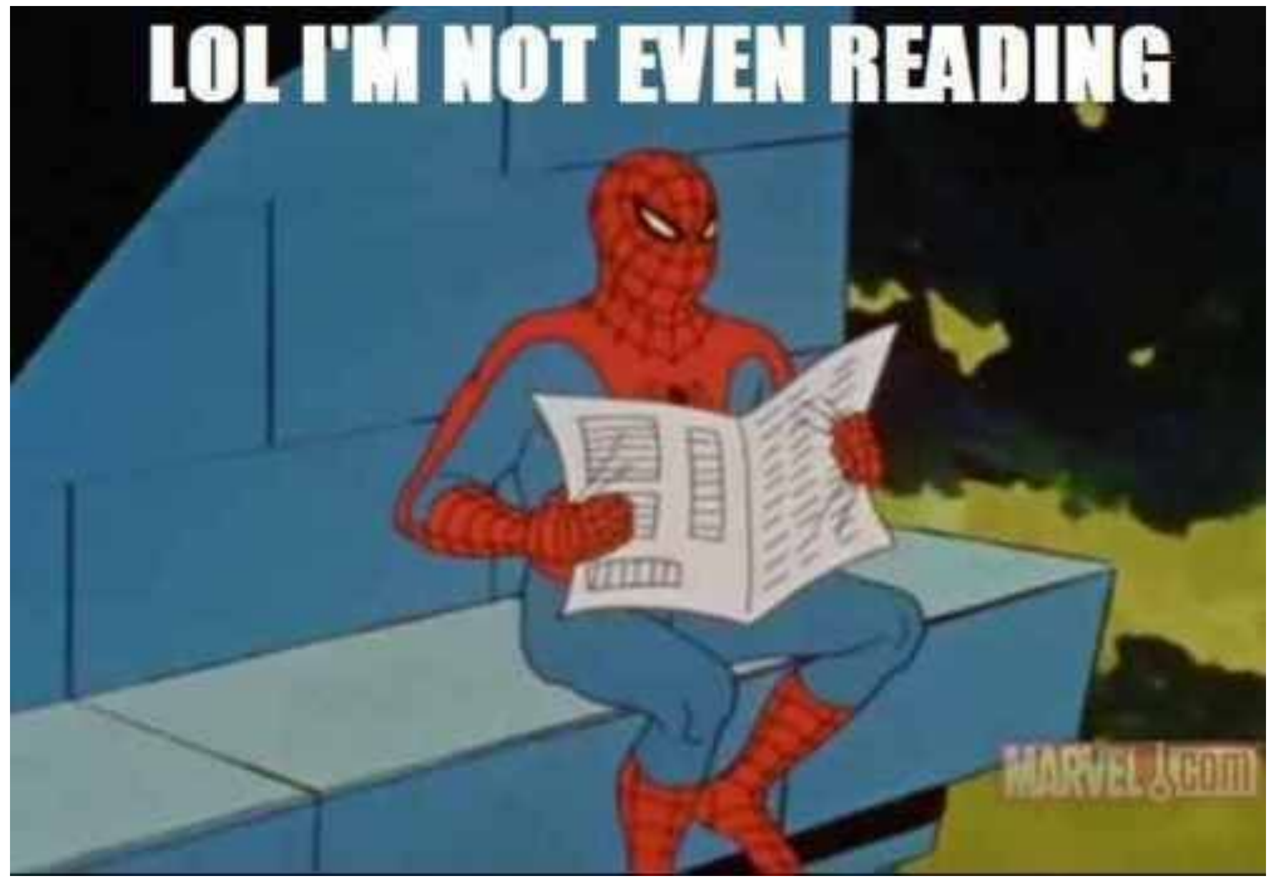


NO solo afectan a las redes neuronales

"Transferability in Machine Learning:
from Phenomena to Black-Box Attacks using Adversarial Samples"

¿Por qué existen?

SABE DEUS



- **Naturaleza Linear**

  “Explaining and Harnessing Adversarial Examples”

- **Fronteras de decisión demasiado ajustadas al dataset**

  “A Boundary Tilting Perspective on the Phenomenon of Adversarial Examples”

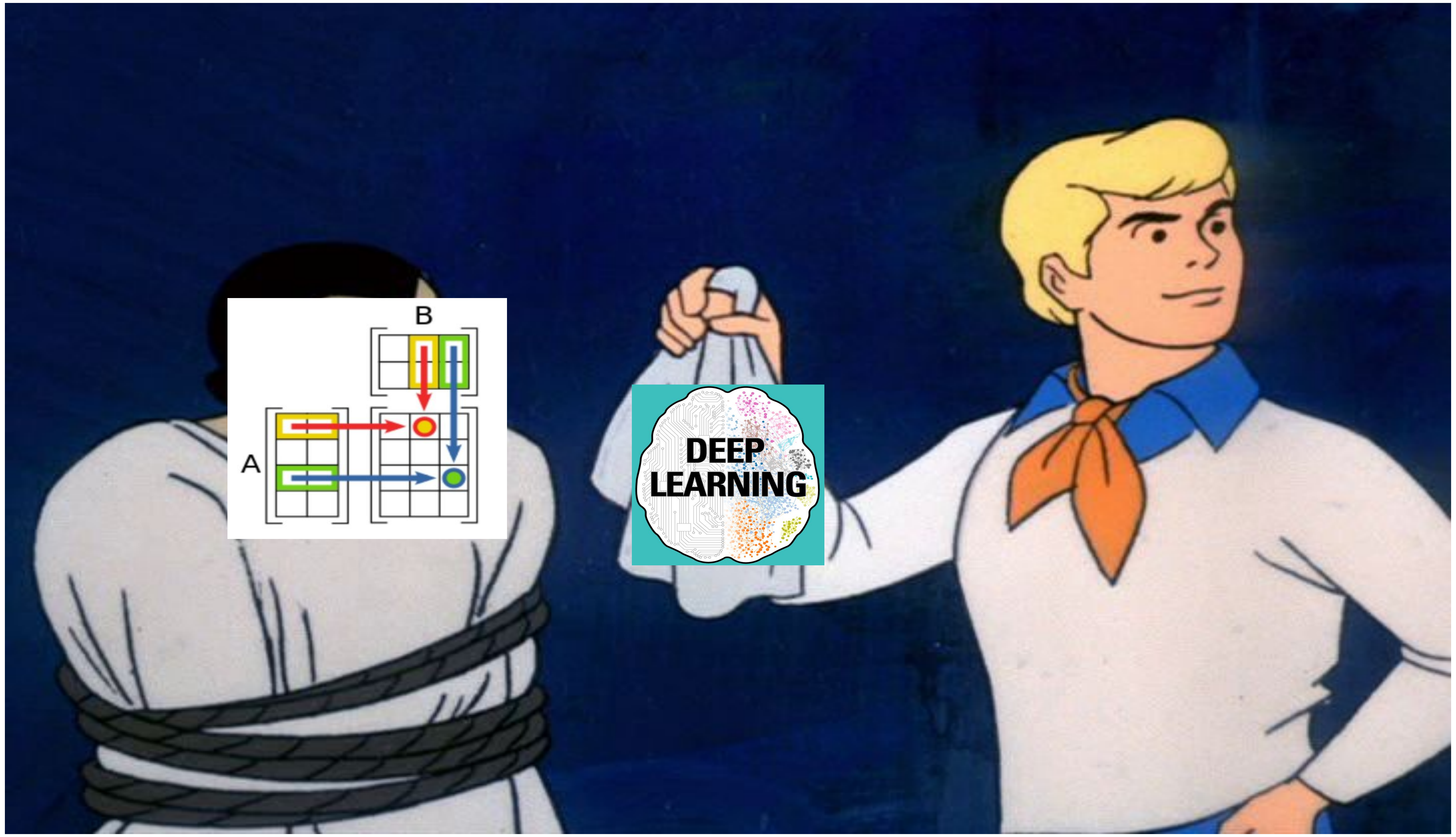
- **Fronteras de decisión aplanadas**

  “Robustness of classifiers: from adversarial to random noise”

- **Fronteras de decisión con largas zonas curvadas**

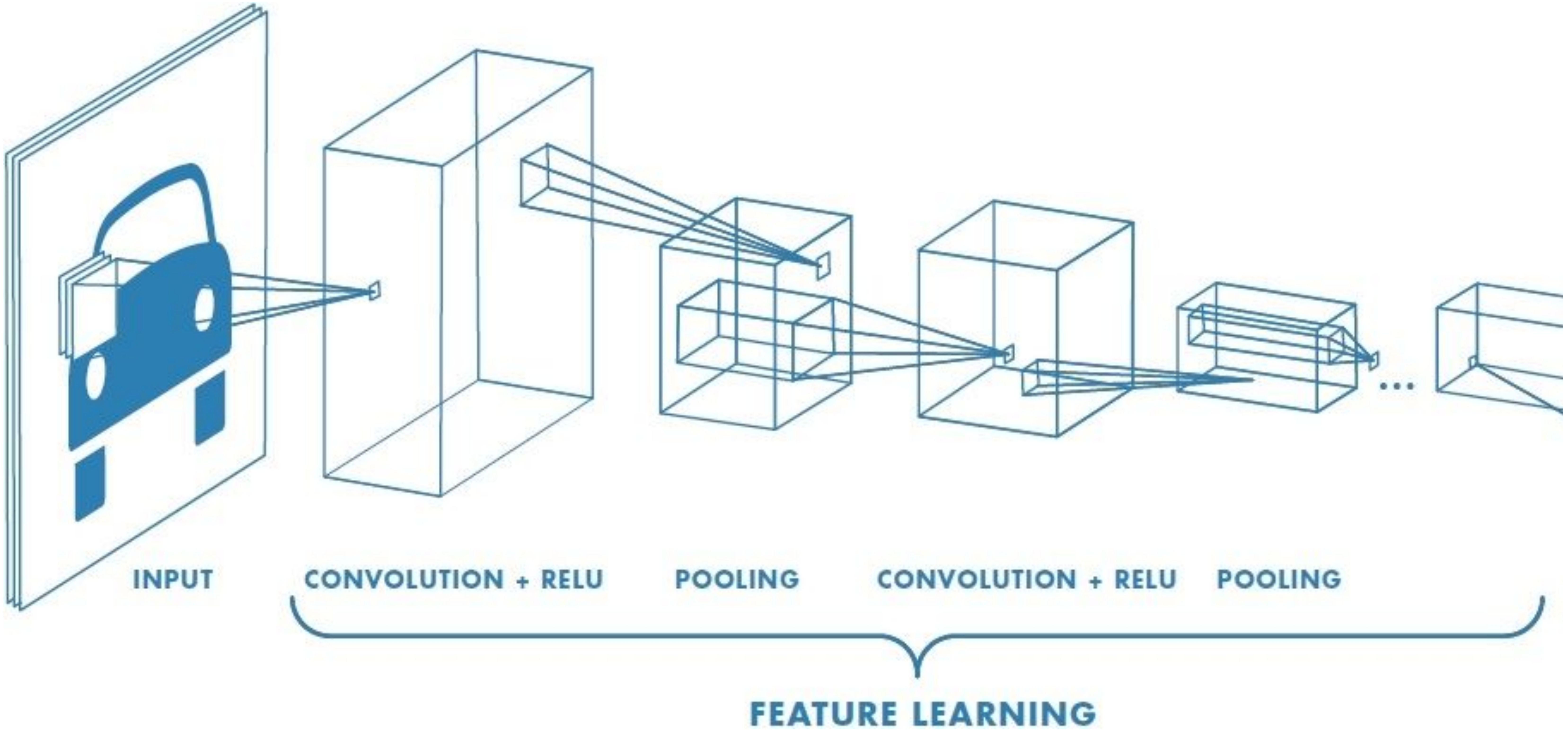
  “Analysis of universal adversarial perturbations”

Intuición

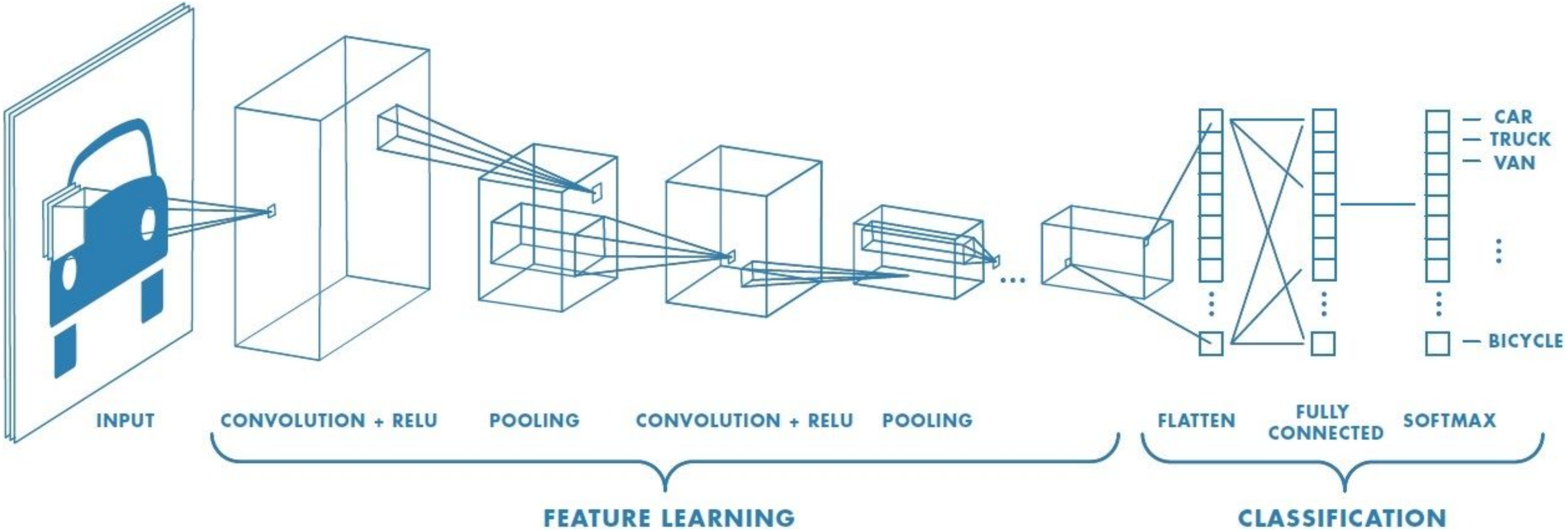


Intuición: Red Neuronal Convolucional

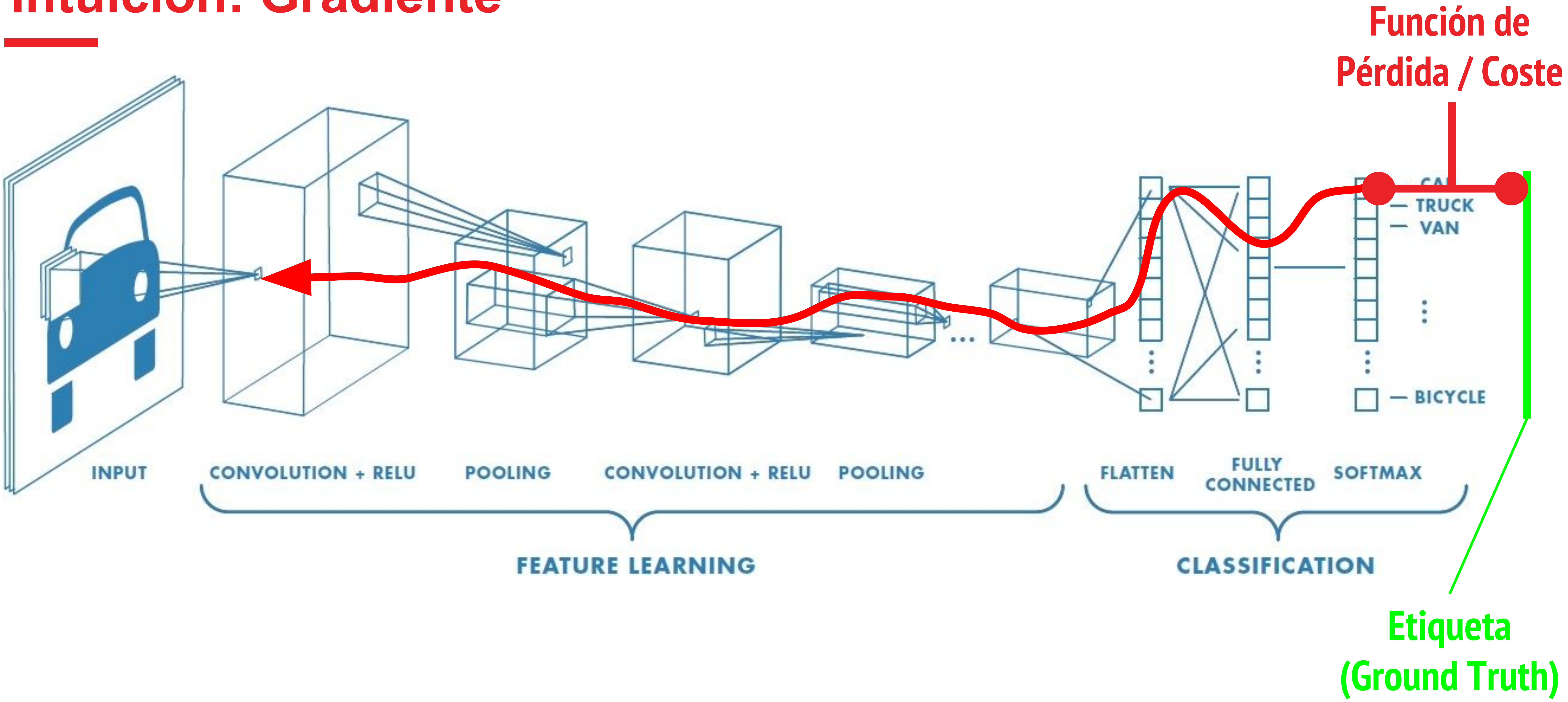
Intuición: Red Neuronal Convolucional



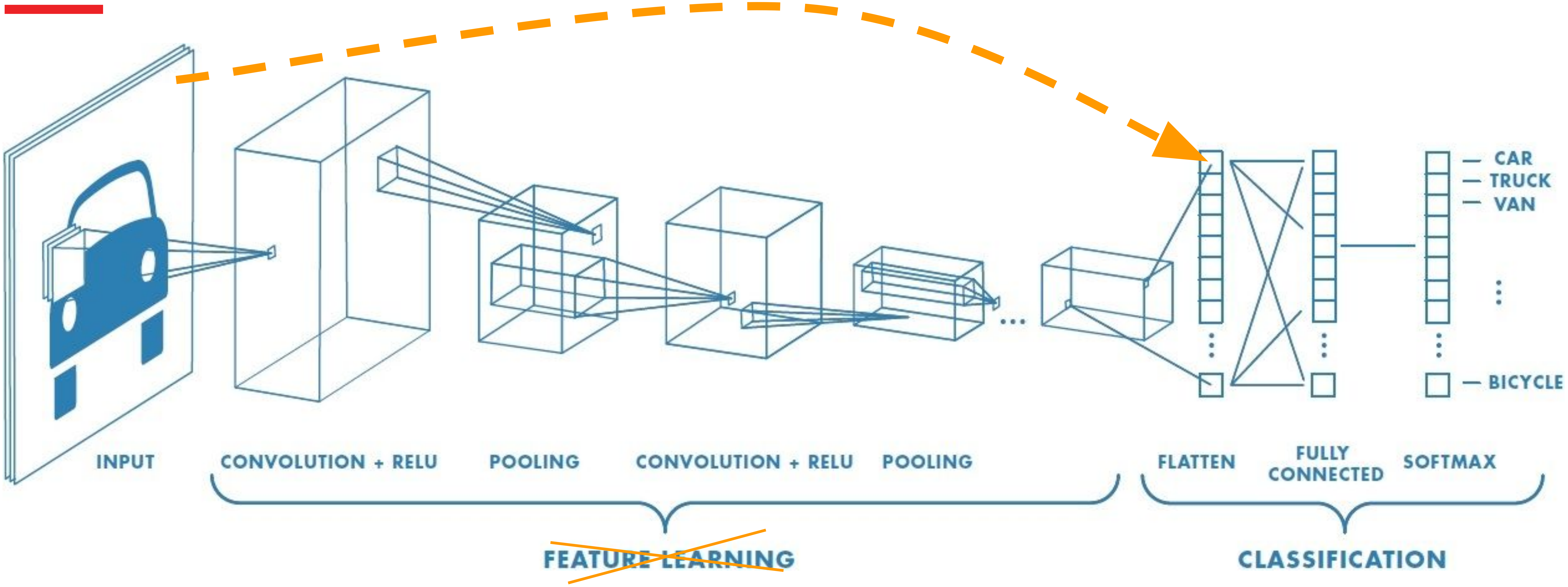
Intuición: Red Neuronal Convolucional



Intuición: Gradiente

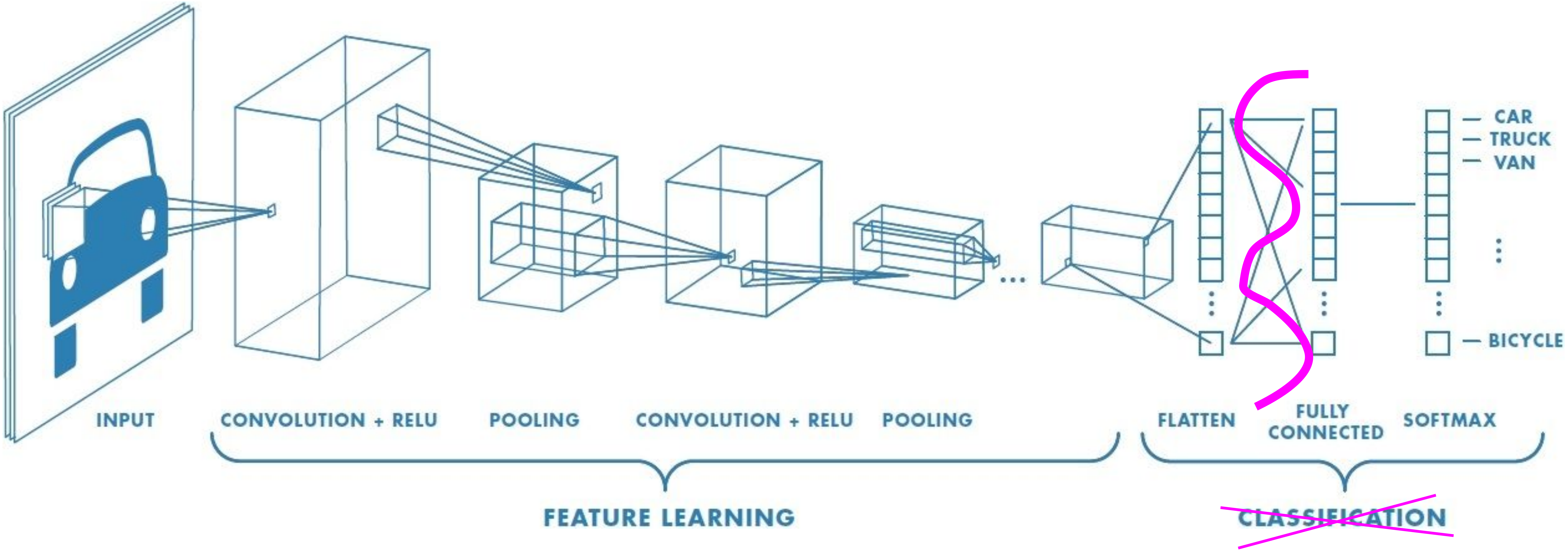


Intuición: Geometría



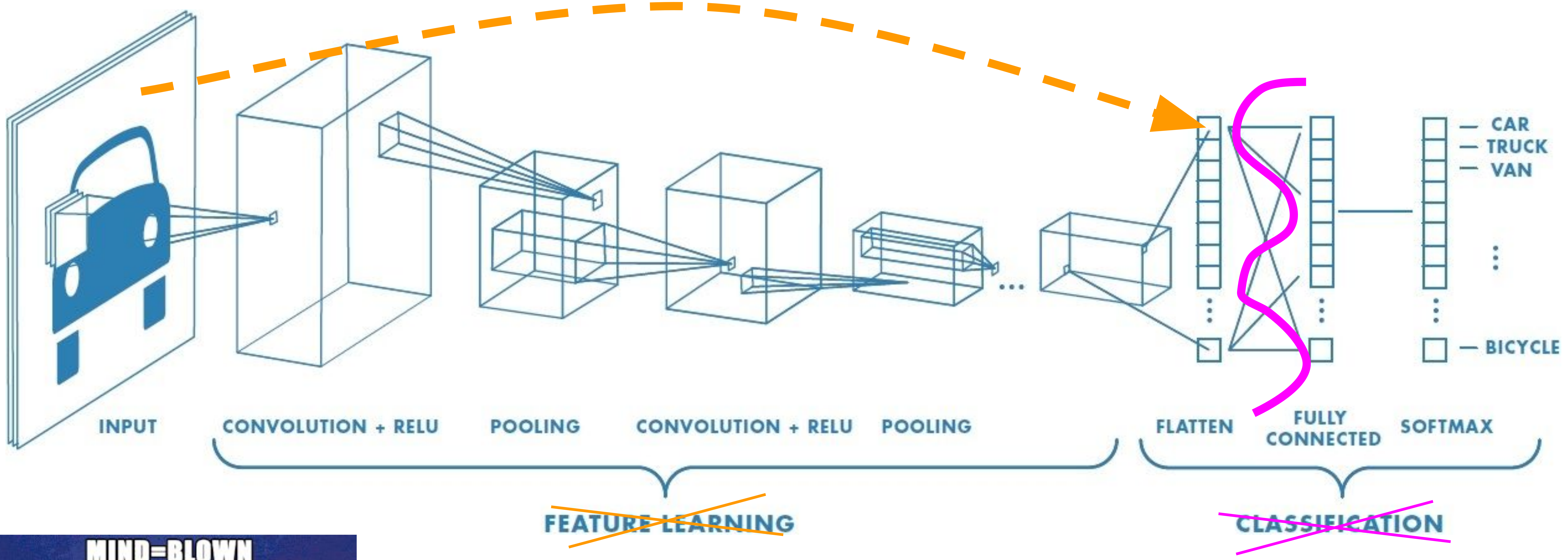
Deformar / Proyectar

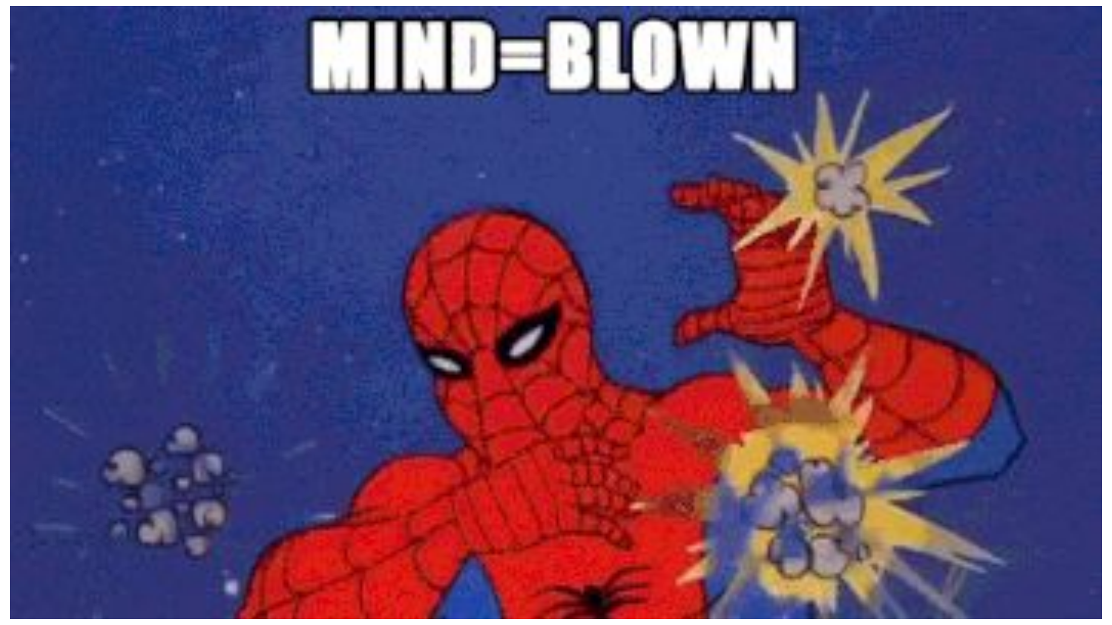
Intuición: Geometría



Definir fronteras de decisión

Intuición: Geometría

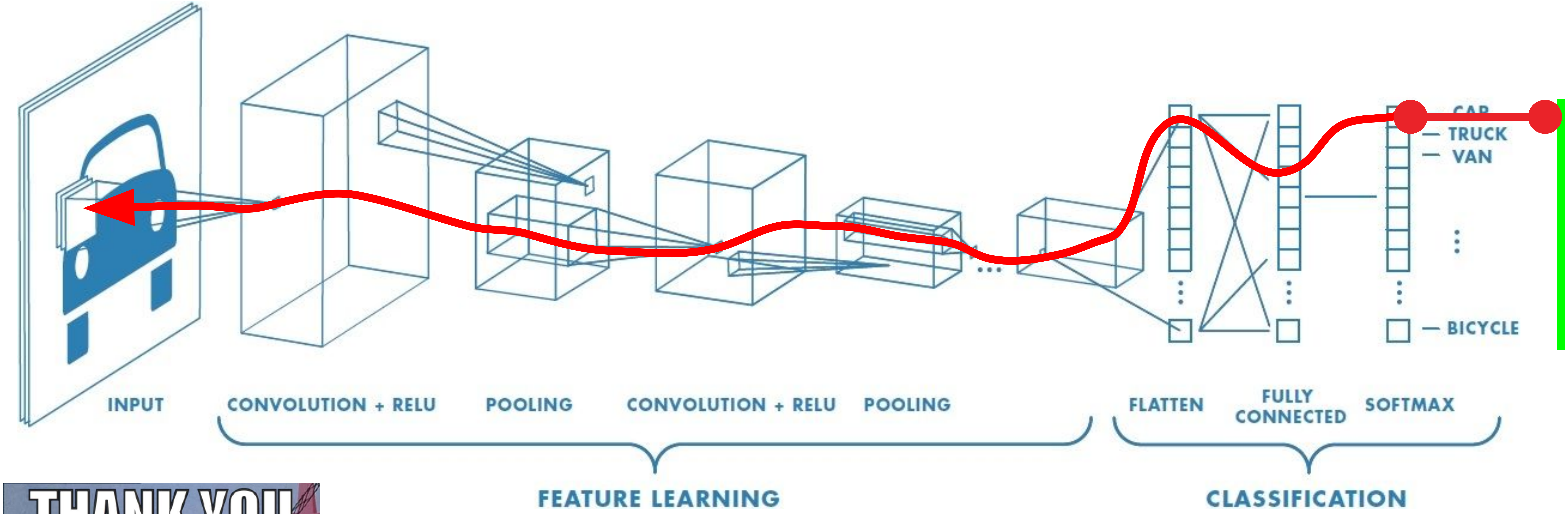




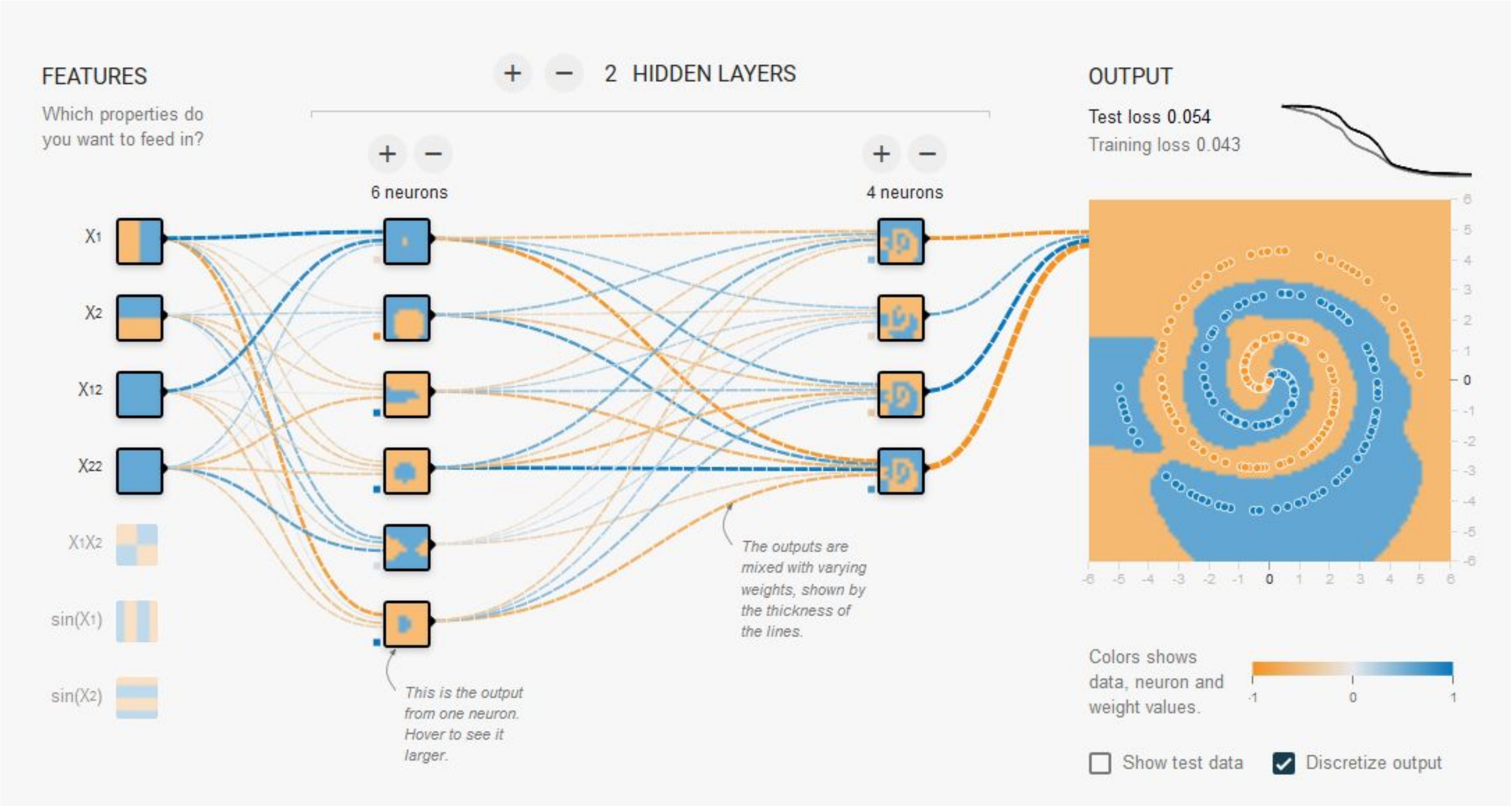
Deformar / Proyectar

Definir fronteras de decisión

Intuición: Usar el gradiente como Ataque



Intuición: Entender la toma de decisiones



playground.tensorflow.org

Ataques

Table II: Taxonomy of Adversarial Examples

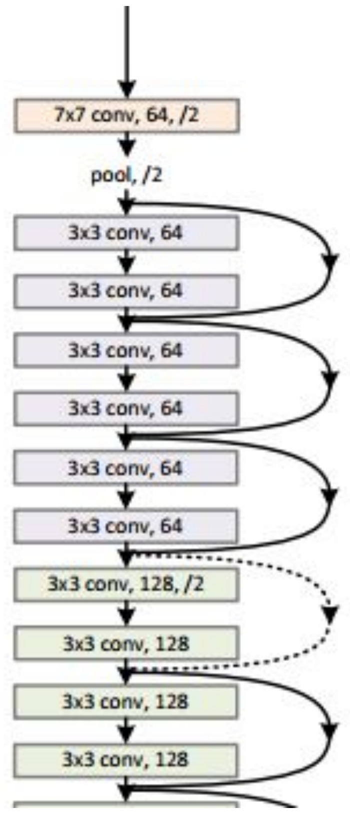
Attacks Methods	Adversarial Falsification	Adversary's Knowledge	Adversarial Specificity	Perturbation Scope	Perturbation Limitation	Attack Frequency	Perturbation Measurement	Datasets	Architectures
L-BFGS Attack [19]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, ImageNet, YoutubeDataset	AlexNet, QuocNet
Fast Gradient Sign Method (FGSM) [55]	False Negative	White-Box	Non-Targeted	Individual	N/A	One-time	element-wise	MNIST, ImageNet	GoogLeNet
Basic Iterative Method (BIM) and Iterative Least-Likely Class (ILLC) [20]	False Negative	White-Box	Non-Targeted	Individual	N/A	Iterative	element-wise	ImageNet	GoogLeNet
Jacobian-based Saliency Map Attack (JSMA) [82]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST	LeNet
DeepFool [83]	False Negative	White-Box	Non-Targeted	Individual	Optimized	Iterative	$\ell_p(p \in 1, \infty)$	MNIST, CIFAR10, ImageNet	LeNet, CaffeNet, GoogLeNet
CPPN EA Fool [84]	False Positive	White-Box	Non-Targeted	Individual	N/A	Iterative	N/A	MNIST, ImageNet	LeNet, AlexNet
C&W's Attack [85]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	$\ell_1, \ell_2, \ell_\infty$	MNIST, CIFAR10, ImageNet	GoogLeNet
Zeroth Order Optimiza-tion [78]	False Negative	Black-Box	Targeted & Non-Targeted	Individual	Optimized	Iterative	ℓ_2	CIFAR10, ImageNet	GoogLeNet
Universal Per-turbation [86]	False Negative	White-Box	Non-Targeted	Universal	Optimized	Iterative	$\ell_p(p \in 1, \infty)$	ImageNet	CaffeNet, VGG, GoogLeNet, ResNet
One Pixel Attack [87]	False Negative	Black-Box	Targeted & Non-Targeted	Individual	Constraint	Iterative	ℓ_0	CIFAR10	VGG, AllConv, NiN
Feature Adversary [88]	False Negative	White-Box	Targeted	Individual	Constraint	Iterative	ℓ_2	ImageNet	CaffeNet, VGG, AlexNet, GoogLeNet
Hot/Cold [81]	False Negative	White-Box	Targeted	Individual	Optimized & Constraint	One-time	PASS	MNIST, ImageNet	LeNet, GoogLeNet, ResNet
Natural GAN [79]	False Negative	Black-Box	Non-targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, LSUN, SNLI	LeNet, LSTM, TreeLSTM
Model-based Ensembling Attack [89]	False Negative	White-Box	Targeted & Non-Targeted	Individual	Constraint	Iterative	ℓ_2	ImageNet	VGG, GoogLeNet, ResNet
Ground-Truth Attack [90]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_1, ℓ_∞	MNIST	3-layer FC

"Adversarial Examples: Attacks and Defenses for Deep Learning"

Table II: Taxonomy of Adversarial Examples

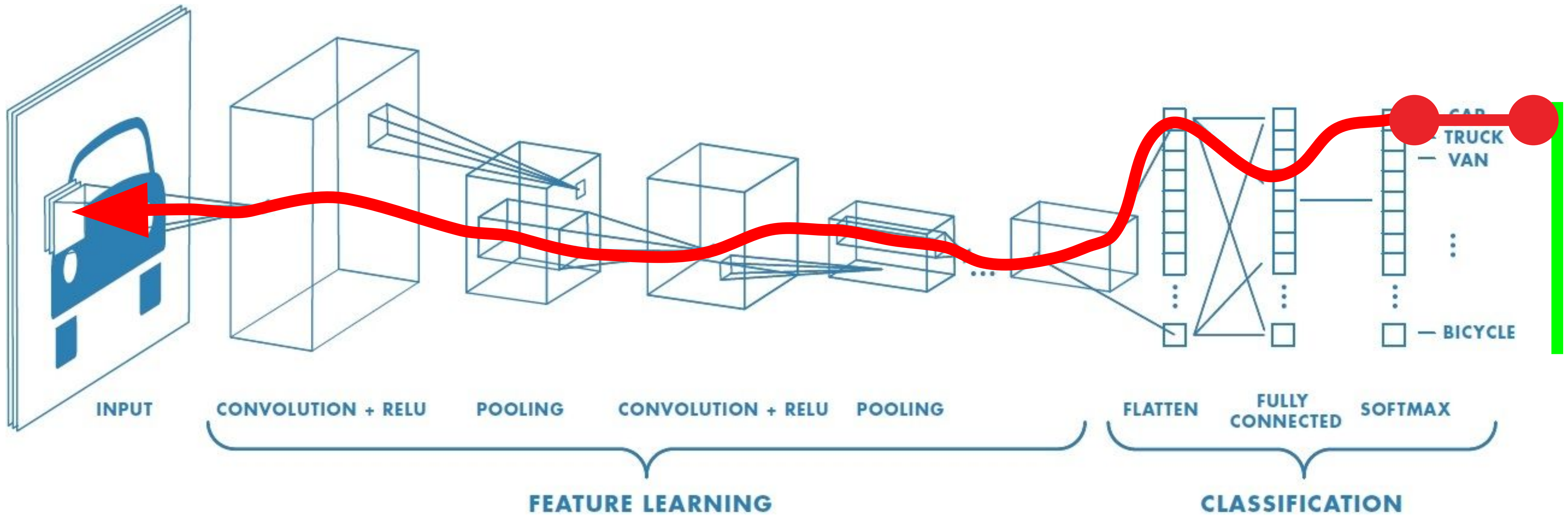
Attacks Methods	Adversarial Falsification	Adversary's Knowledge	Adversarial Specificity	Perturbation Scope	Perturbation Limitation	Attack Frequency	Perturbation Measurement	Datasets	Architectures
L-BFGS Attack [19]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, ImageNet, YoutubeDataset	AlexNet, QuocNet
Fast Gradient Sign Method (FGSM) [55]	False Negative	White-Box	Non-Targeted	Individual	N/A	One-time	element-wise	MNIST, geNet	GoogLeNet
Basic Iterative Method (BIM) and Iterative Least-Likely Class (ILLC) [20]	False N							geNet	GoogLeNet
Jacobian-based Saliency Map Attack (JSMA) [82]	False N							NIST	LeNet
DeepFool [83]	False N							NIST, AR10, geNet	LeNet, CaffeNet, GoogLeNet
CPPN EA Fool [84]	False							NIST, geNet	LeNet, AlexNet
C&W's Attack [85]	False N							NIST, AR10, geNet	GoogLeNet
Zeroth Order Optimization [78]	False N							AR10, geNet	GoogLeNet
Universal Perturbation [86]	False N							geNet	CaffeNet, VGG, GoogLeNet, ResNet
One Pixel Attack [87]	False N							AR10	VGG, AllConv, NiN
Feature Adversary [88]	False N							geNet	CaffeNet, VGG, AlexNet, GoogLeNet
Hot/Cold [81]	False N							NIST, geNet	LeNet, GoogLeNet, ResNet
Natural GAN [79]	False N							, LSUN, NLI	LeNet, LSTM, TreeLSTM
Model-based Ensembling Attack [89]	False Negative	White-Box	Targeted & Non-Targeted	Individual	Constraint	Iterative	ℓ_2	ImageNet	VGG, GoogLeNet, ResNet
Ground-Truth Attack [90]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_1, ℓ_∞	MNIST	3-layer FC

Ataques según conocimiento del atacante:

			
Caja Blanca (White Box)	✓	✓	✓
Caja Gris (Gray Box)	✓	?	?
Caja Negra (Black Box)	✗	✗	✗

Ataques. Caja Blanca

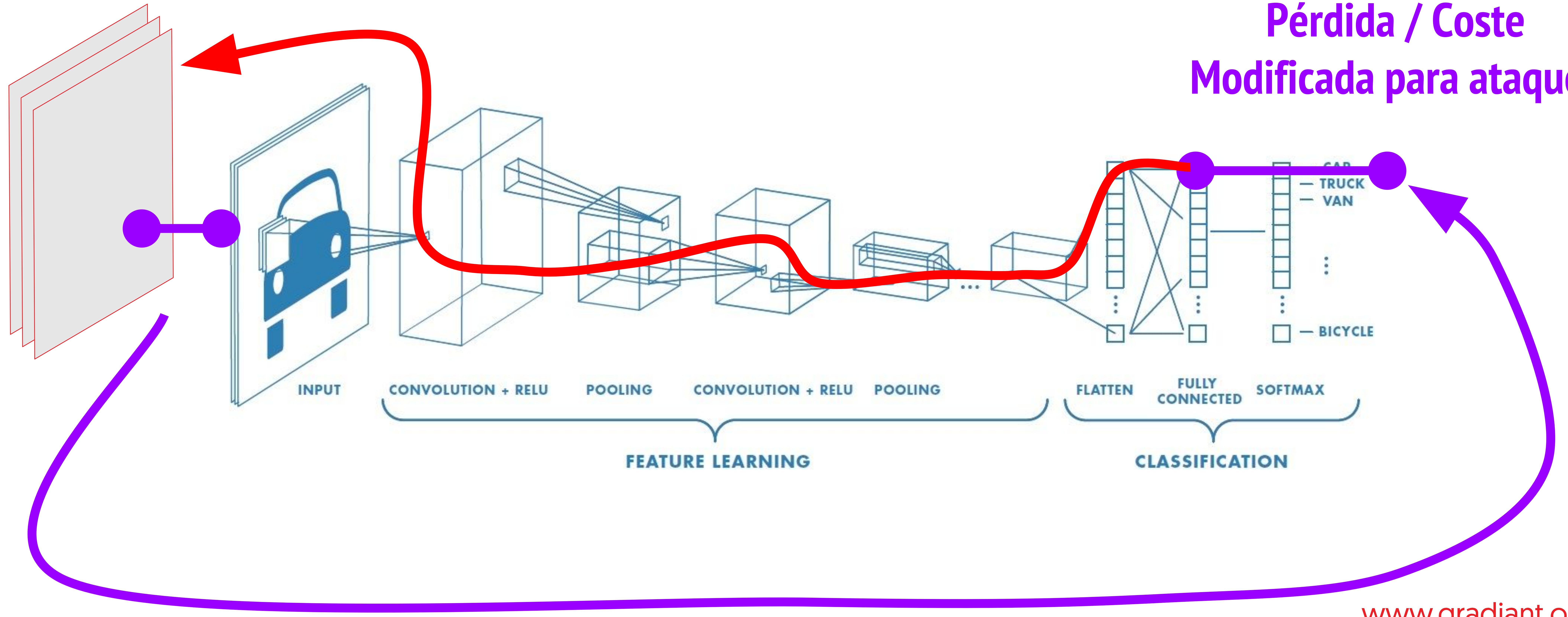
- Basados en Gradiente



Ataques. Caja Blanca

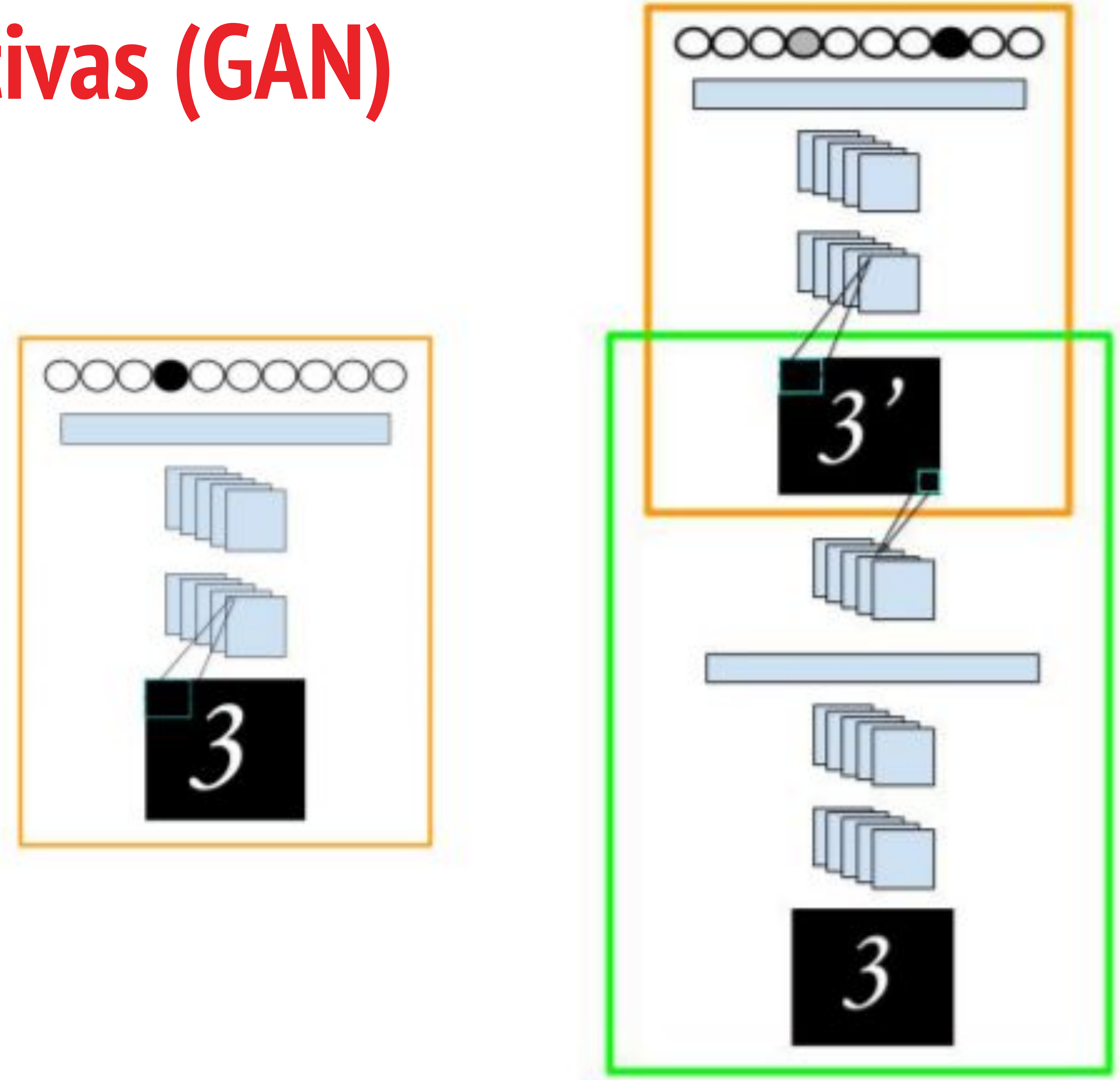
- Basados en Optimización

Función de
Pérdida / Coste
Modificada para ataques

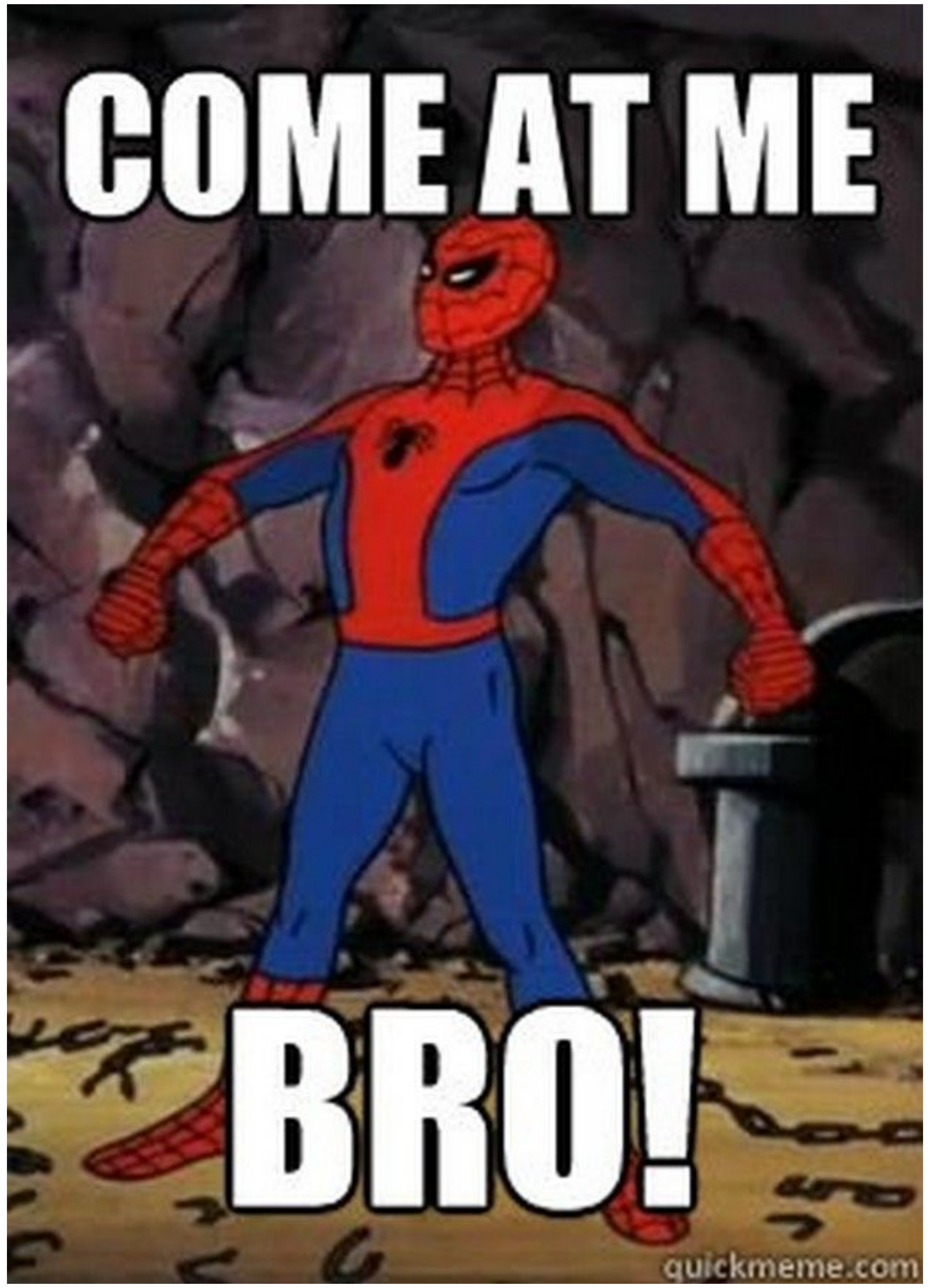


Ataques. Caja Blanca

- Basados en Redes Generativas (GAN)



Defensas



Defensas

NIPS 2017: Competition on Adversarial Attacks and Defenses

2.3 *Overview of defenses*

No method of defending against adversarial examples is yet completely satisfactory.

Defensas

Table IV: Summary of Countermeasures for Adversarial Examples

	Defensive Strategies	Representative Studies
Reactive	Adversarial Detecting	[34], [107], [122]–[129]
	Input Reconstruction	[127], [130], [131]
	Network Verification	[132]–[134]
Proactive	Network Distillation	[135]
	Adversarial (Re)Training	[35], [36], [55], [92], [94], [136]
	Classifier Robustifying	[137], [138]

Defensas


Carlini
&
Wagner


MY
Carlini
&
Wagner
ADVERSARIAL DEFENSES!
imgflip.com

Defensas

Carlini
&
Wagner

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong

Obfuscated Gradients Give a False Sense of Security:
Circumventing Defenses to Adversarial Examples

Preguntas

Challenge en 2 sesiones