

Jak wprowadzić w błąd sieć neuronową?

Adam Żychowski

Motywacja

- ▶ sieci neuronowe zaczęły podejmować decyzje wpływające na życie, zdrowie, sekrety (self-driving cars, skanery na lotniskach, diagnozowanie chorób, face recognition itd.)
- ▶ coraz więcej mówi się o bezpieczeństwie sieci neuronowych i coraz więcej zasobów przeznaczanych jest na konstruowanie ataków
- ▶ „uczenie maszynowe działa, ale łatwo je zepsuć”

[Szegedy et al. 2014]



“panda”
57.7% confidence

+ .007 ×



“nematode”
8.2% confidence

=

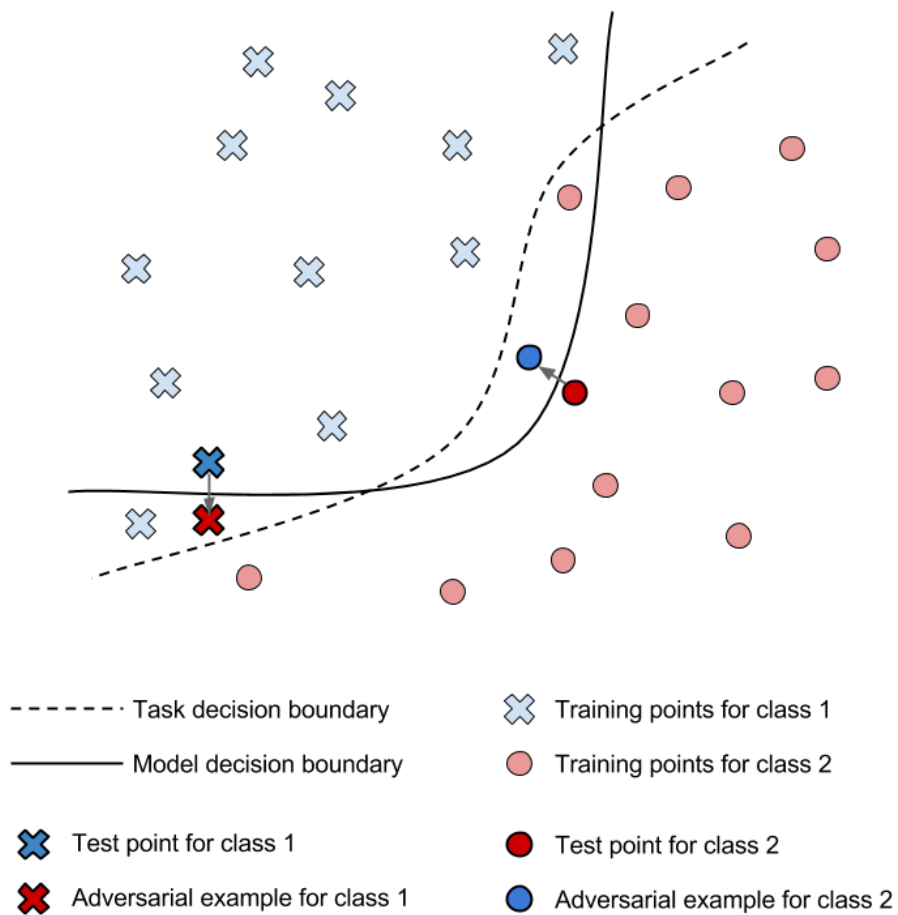


“gibbon”
99.3 % confidence

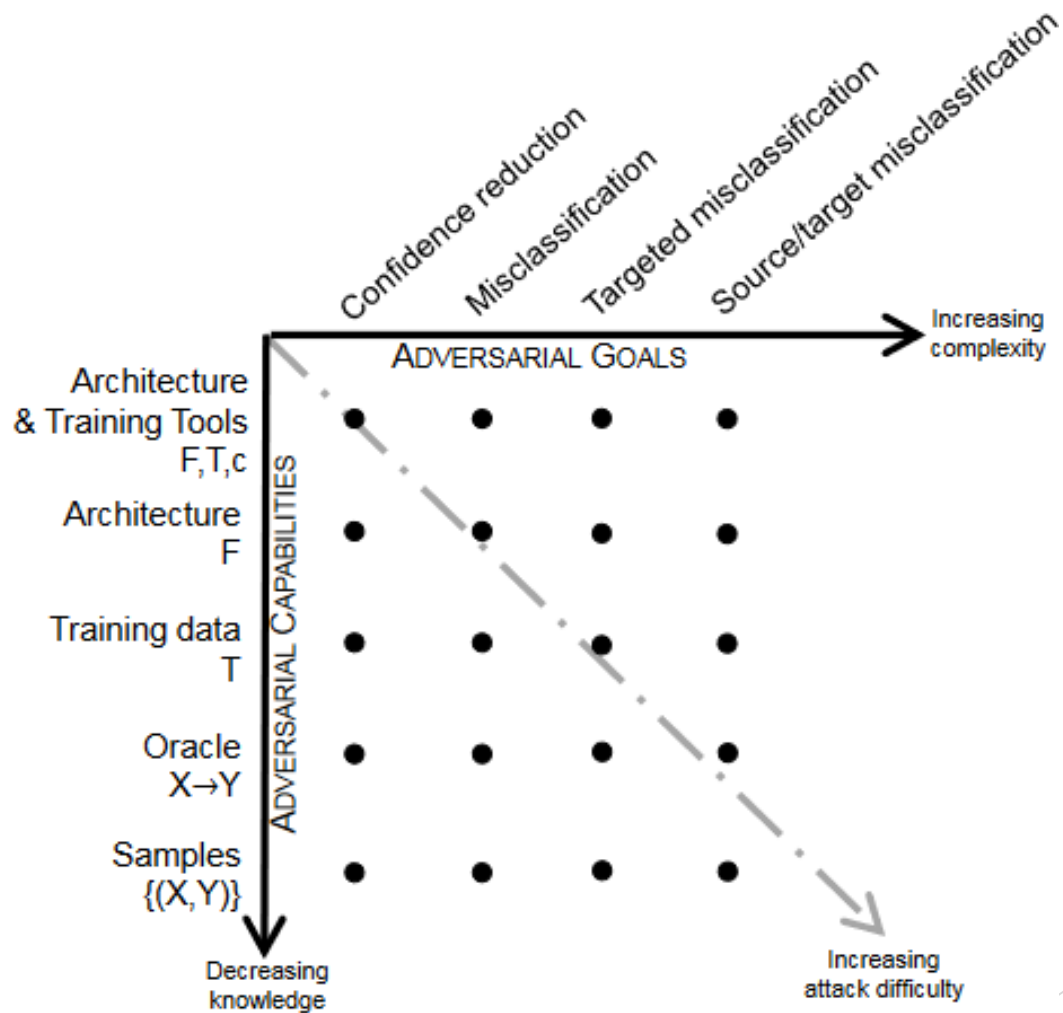
Minimize $\|r\|_2$ subject to:

1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$

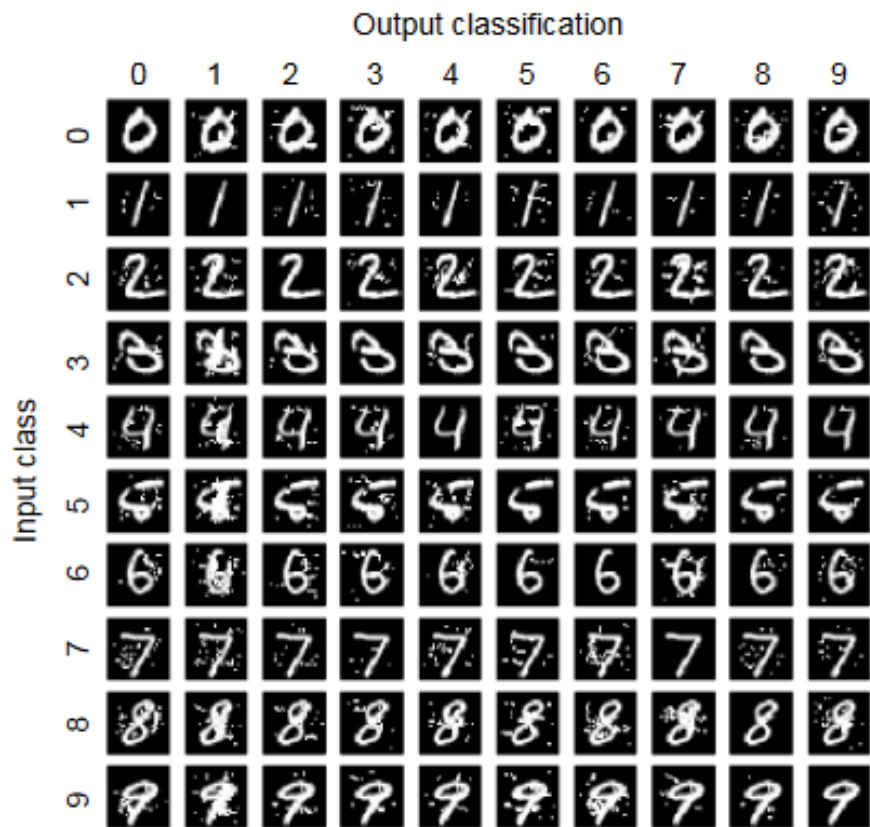
Dlaczego to działa?



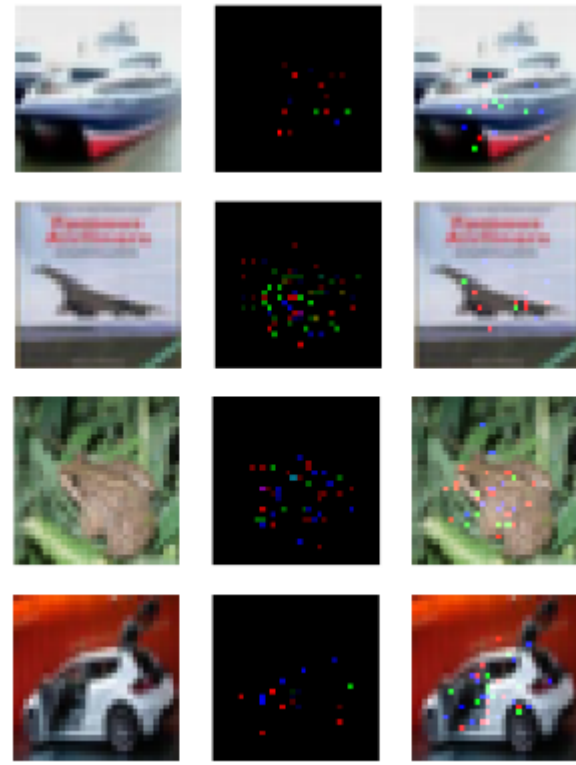
Podział ataków



[Papernot et al. 2016]



[2]



One-pixel attack

[Su et al. 2019]

	AllConv	NiN	VGG16
OriginAcc	85.6%	87.2%	83.3%
Targeted	19.82%	23.15%	16.48%
Non-targeted	68.71%	71.66%	63.53%

CIFAR-10 dataset

AllConv



SHIP
CAR(99.7%)



HORSE
DOG(70.7%)



CAR
AIRPLANE(82.4%)



DEER
AIRPLANE(49.8%)



HORSE
DOG(88.0%)

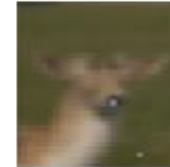
NiN



HORSE
FROG(99.9%)



DOG
CAT(75.5%)



DEER
DOG(86.4%)



BIRD
FROG(88.8%)



SHIP
AIRPLANE(62.7%)

VGG



DEER
AIRPLANE(85.3%)



BIRD
FROG(86.5%)



CAT
BIRD(66.2%)



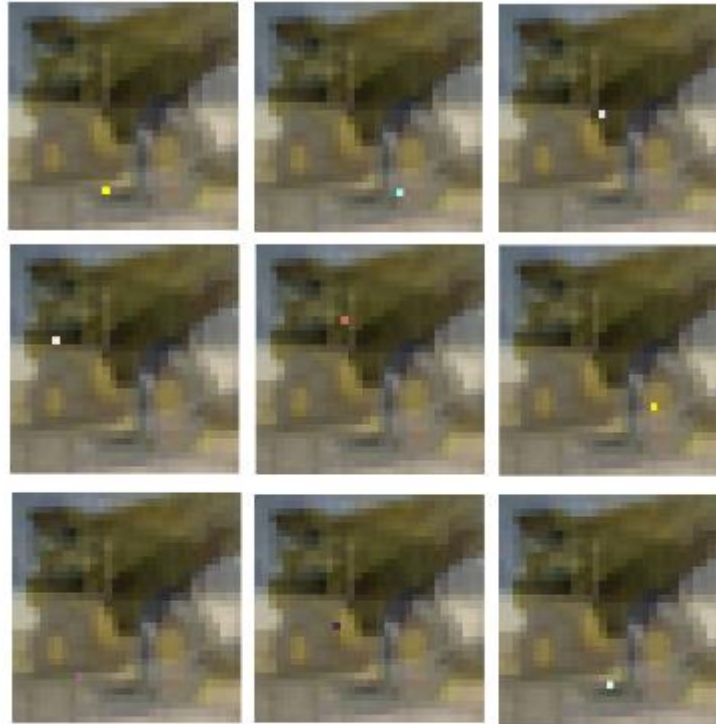
SHIP
AIRPLANE(88.2%)



CAT
DOG(78.2%)

One-pixel attack

[Su et al. 2019]



Original image (dog)

Airplane	Automobile	Bird
Cat	Deer	Frog
Horse	Ship	Truck

Target classes

One-pixel attack

[Su et al. 2019]

- ▶ ImageNet dataset
- ▶ AlexNet network
- ▶ 16% obrazów „podatnych” na one-pixel attack



Cup(16.48%)
Soup Bowl(16.74%)



Bassinet(16.59%)
Paper Towel(16.21%)



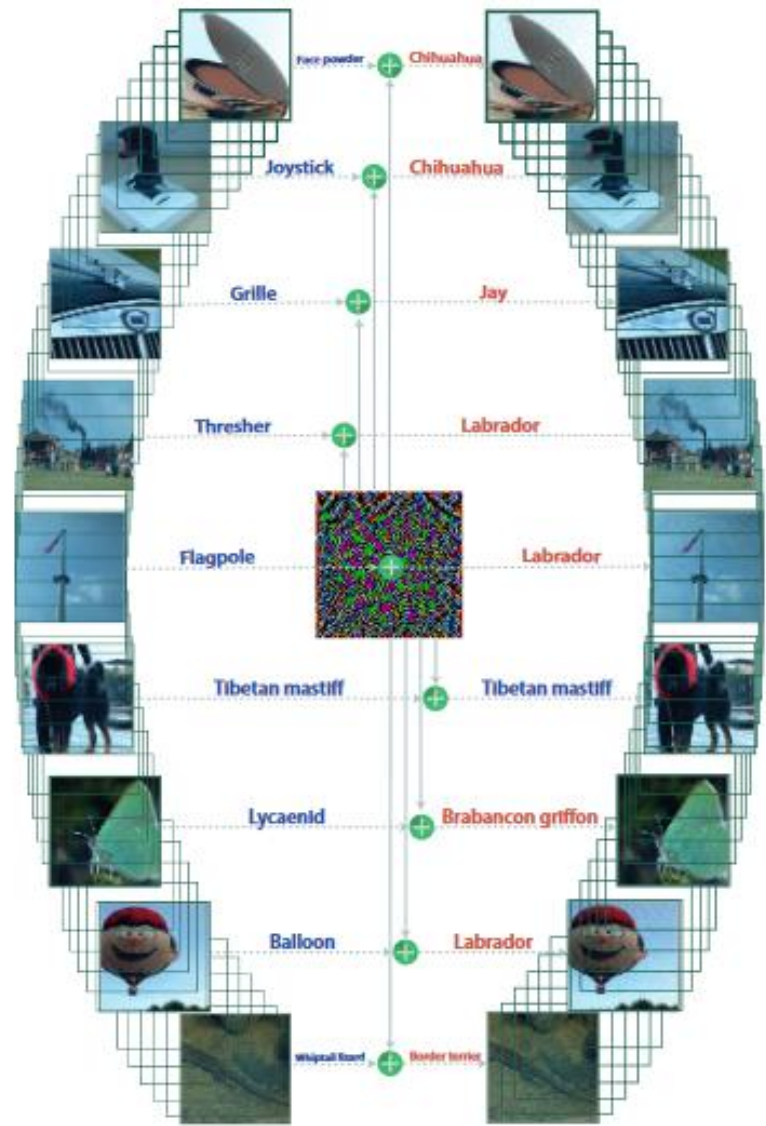
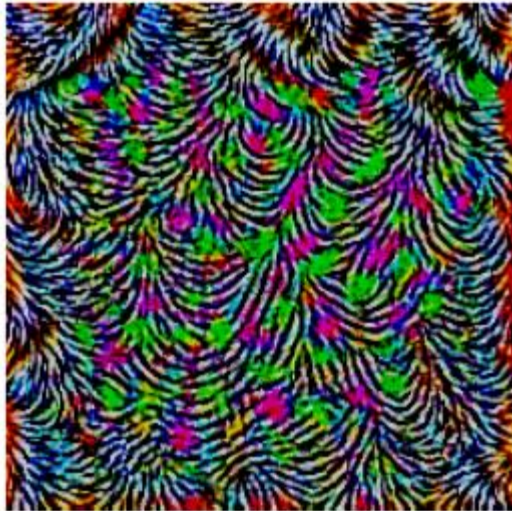
Teapot(24.99%)
Joystick(37.39%)



Hamster(35.79%)
Nipple(42.36%)

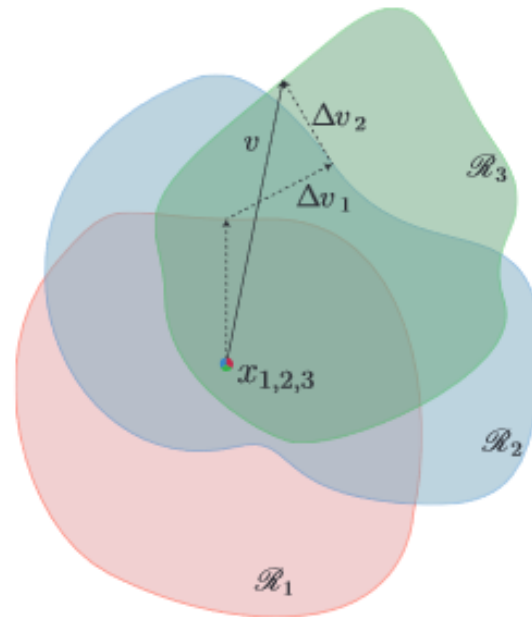
Universal adversarial perturbations

[Moosavi-Dezfooli et al. 2017]



Universal adversarial perturbations

[Moosavi-Dezfooli et al. 2017]



	CaffeNet [8]	VGG-F [2]	VGG-16 [17]	VGG-19 [17]	GoogLeNet [18]	ResNet-152 [6]
X	85.4%	85.9%	90.7%	86.9%	82.9%	89.7%
Val.	85.6	87.0%	90.3%	84.5%	82.0%	88.5%



wool



Indian elephant



Indian elephant



African grey



tabby



African grey



common newt



carousel



grey fox



macaw



three-toed sloth



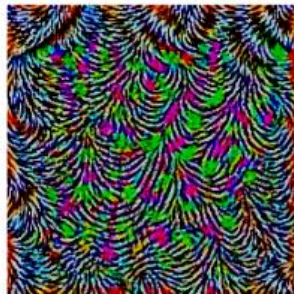
macaw



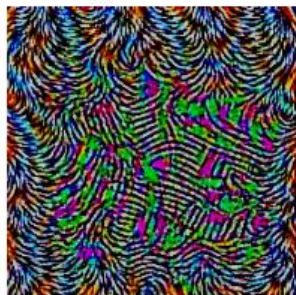
(a) CaffeNet



(b) VGG-F



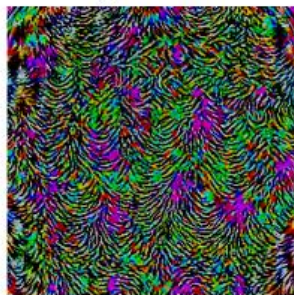
(c) VGG-16



(d) VGG-19



(e) GoogLeNet

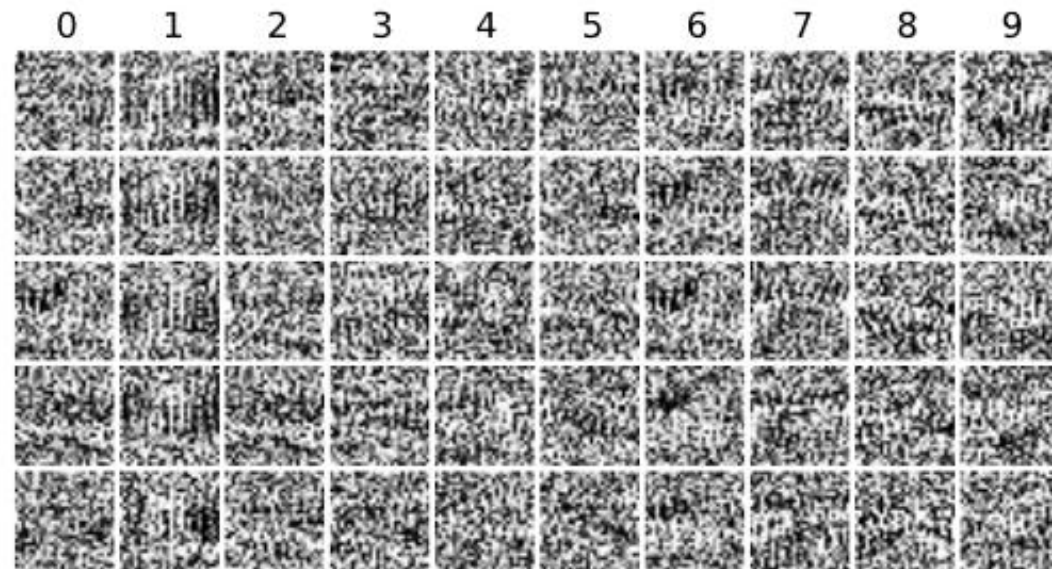


(f) ResNet-152

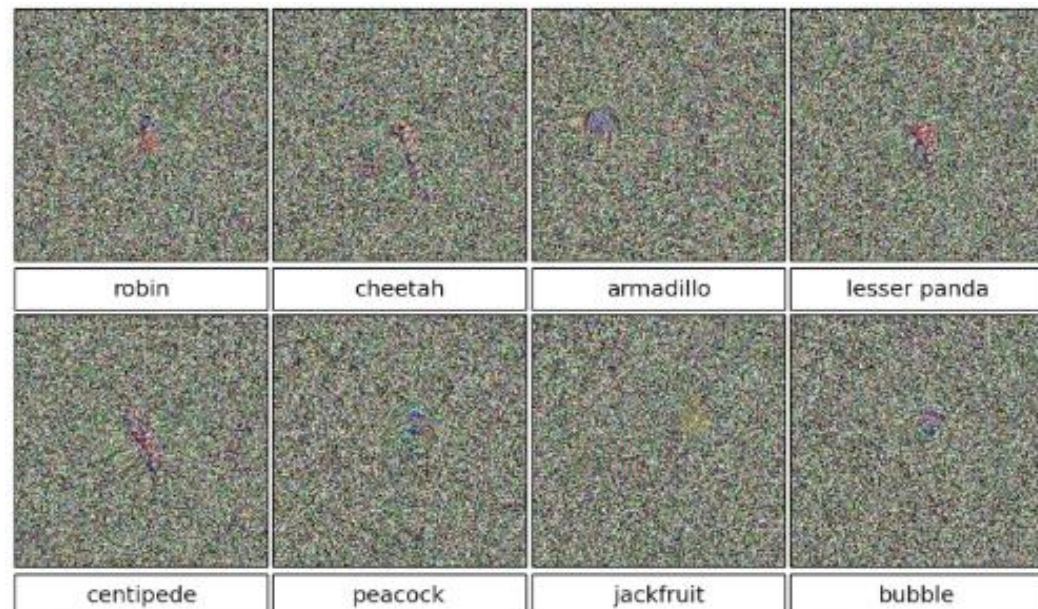
High confidence predictions for unrecognizable images

[Nguyen et al. 2015]

MNIST:



ImageNet:

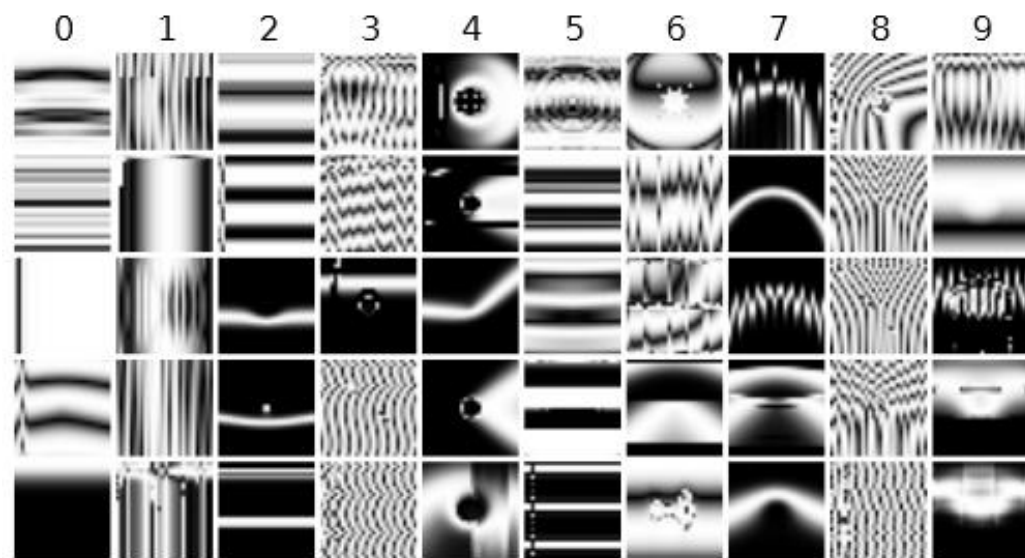


ponad 99% pewności

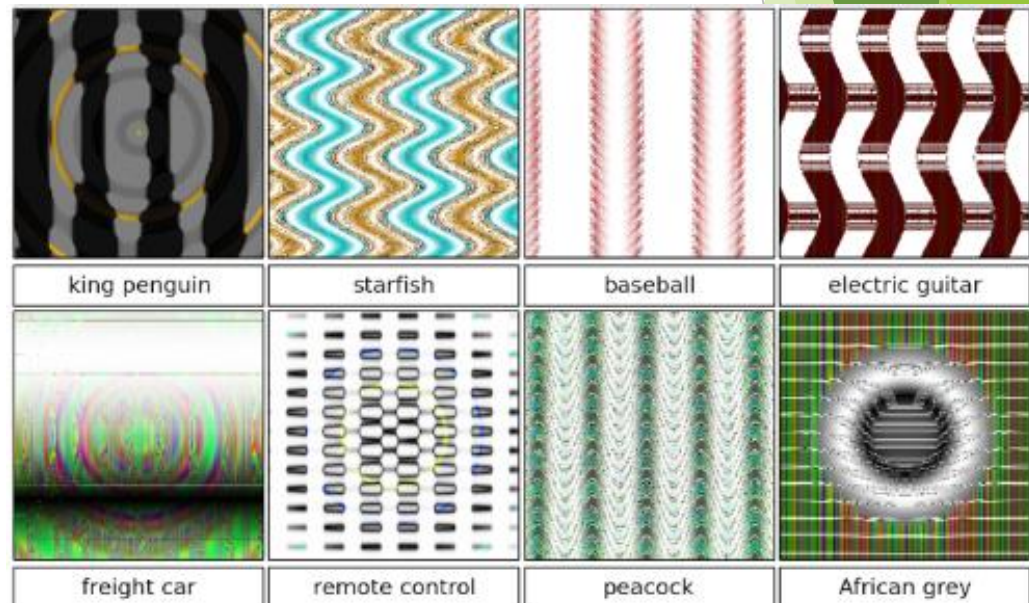
High confidence predictions for unrecognizable images

[Nguyen et al. 2015]

MNIST:



ImageNet:



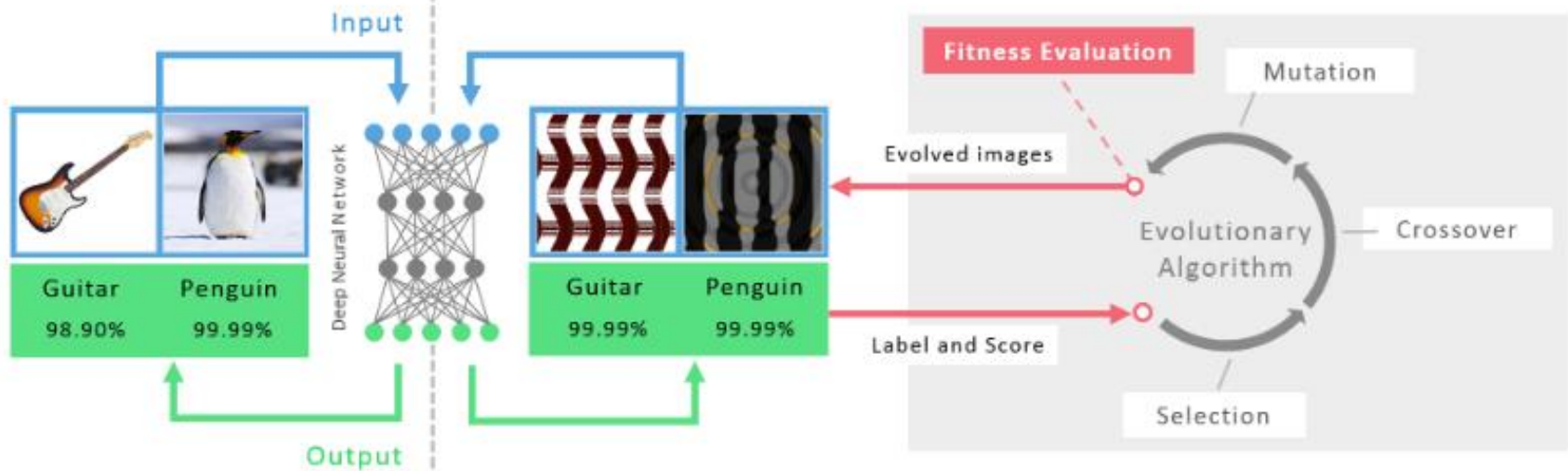
ponad 99% pewności

High confidence predictions for unrecognizable images

[Nguyen et al. 2015]

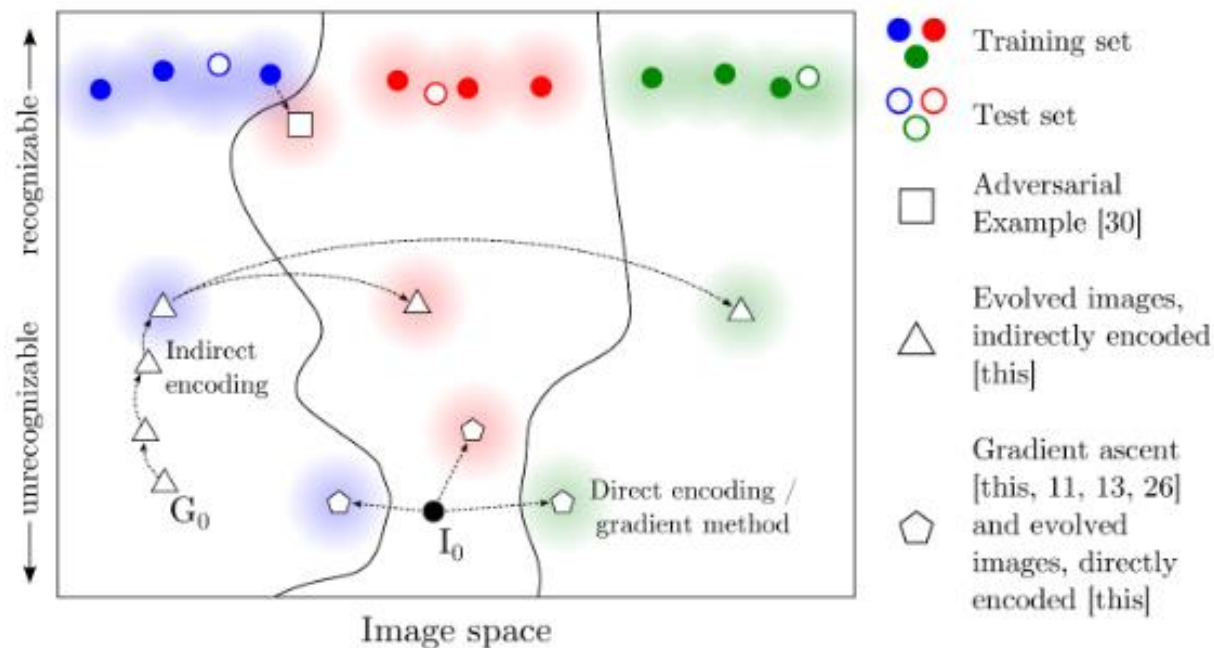
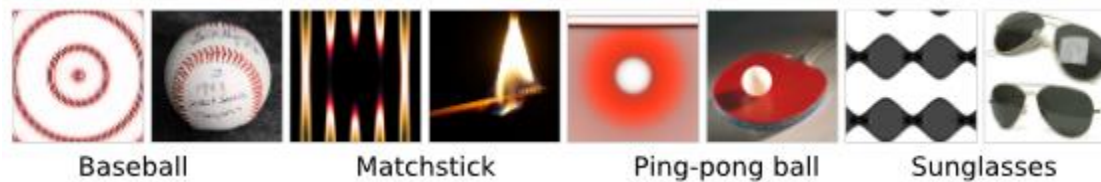
1 State-of-the-art DNNs can recognize real images with high confidence

2 But DNNs are also easily fooled: images can be produced that are unrecognizable to humans, but DNNs believe with 99.99% certainty are natural objects



High confidence predictions for unrecognizable images

[Nguyen et al. 2015]



Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

[Elsayed et al. 2018]

original



adv



perturbation size

8

16

24

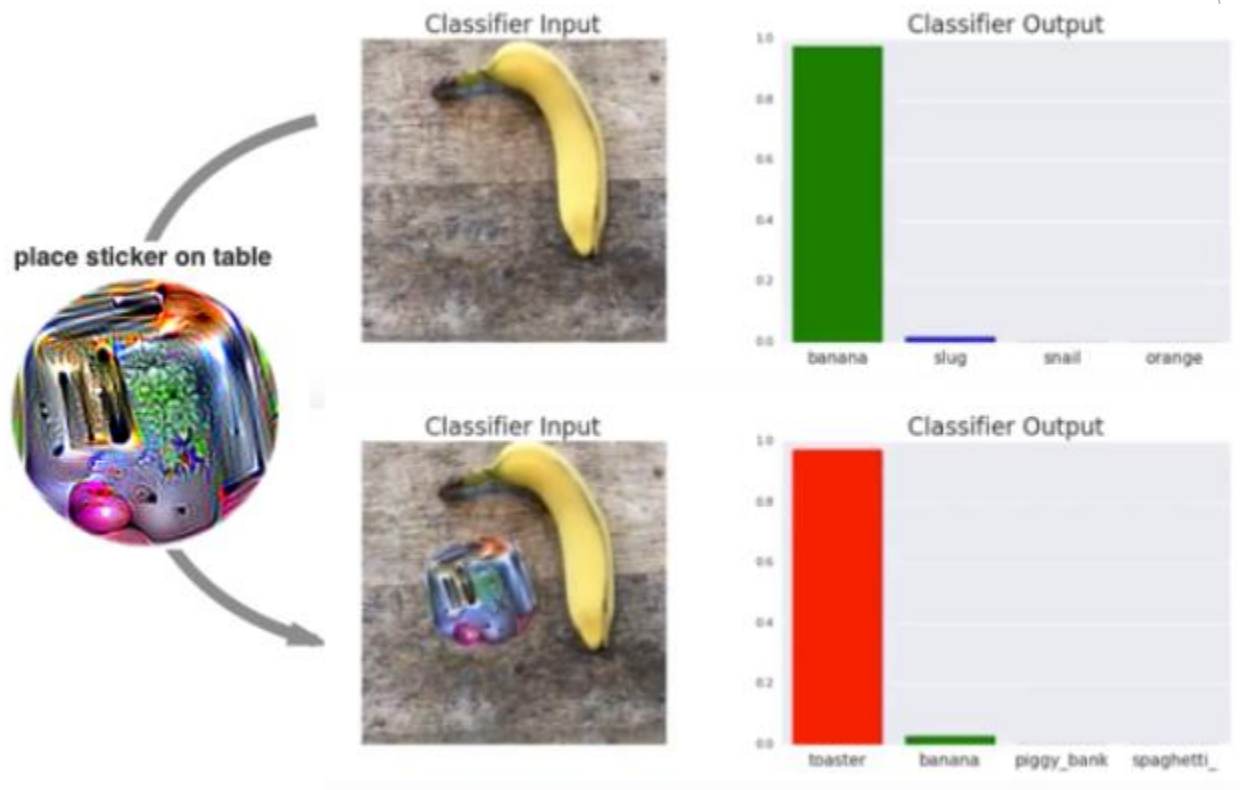
32

40



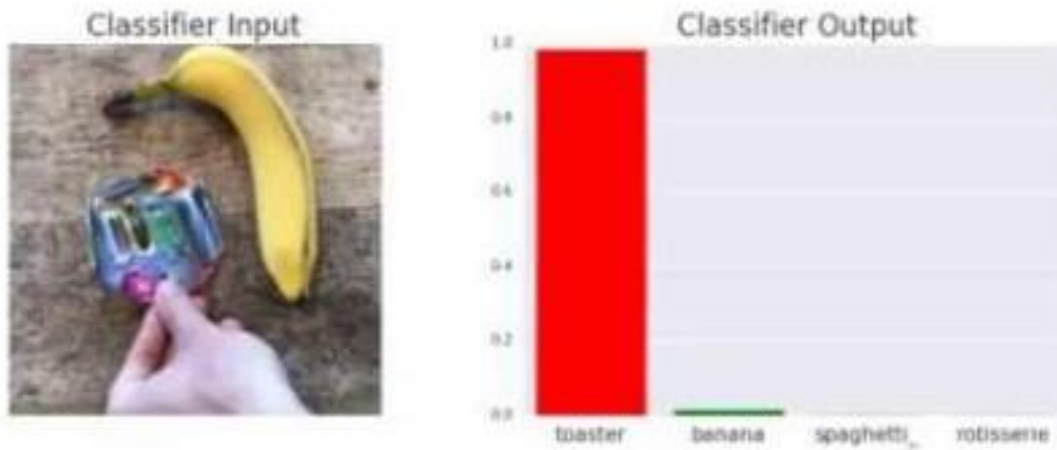
Adversarial Patch

[Brown et al. 2017]



Adversarial Patch

[Brown et al. 2017]

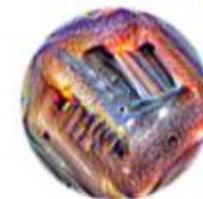
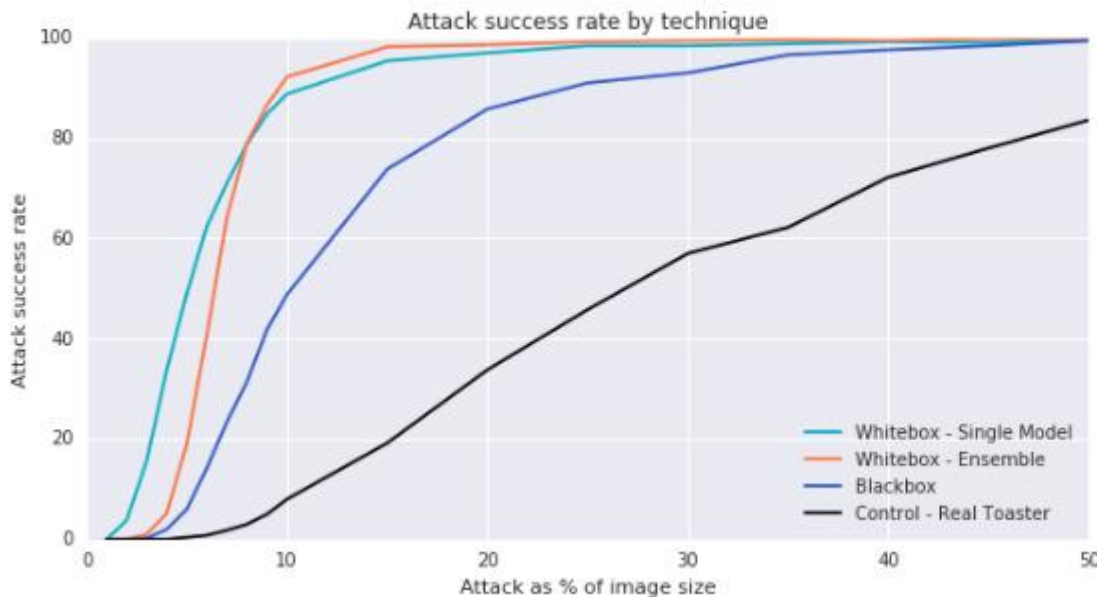


<https://youtu.be/i1sp4X57TL4>

Adversarial Patch

[Brown et al. 2017]

$$A(\text{Patch}, \text{Image}, \text{location, rotation, scale, ...}) =$$



Whitebox - Single Model



Whitebox - Ensemble



Control - Real Toaster



Blackbox

Adversarial Patch

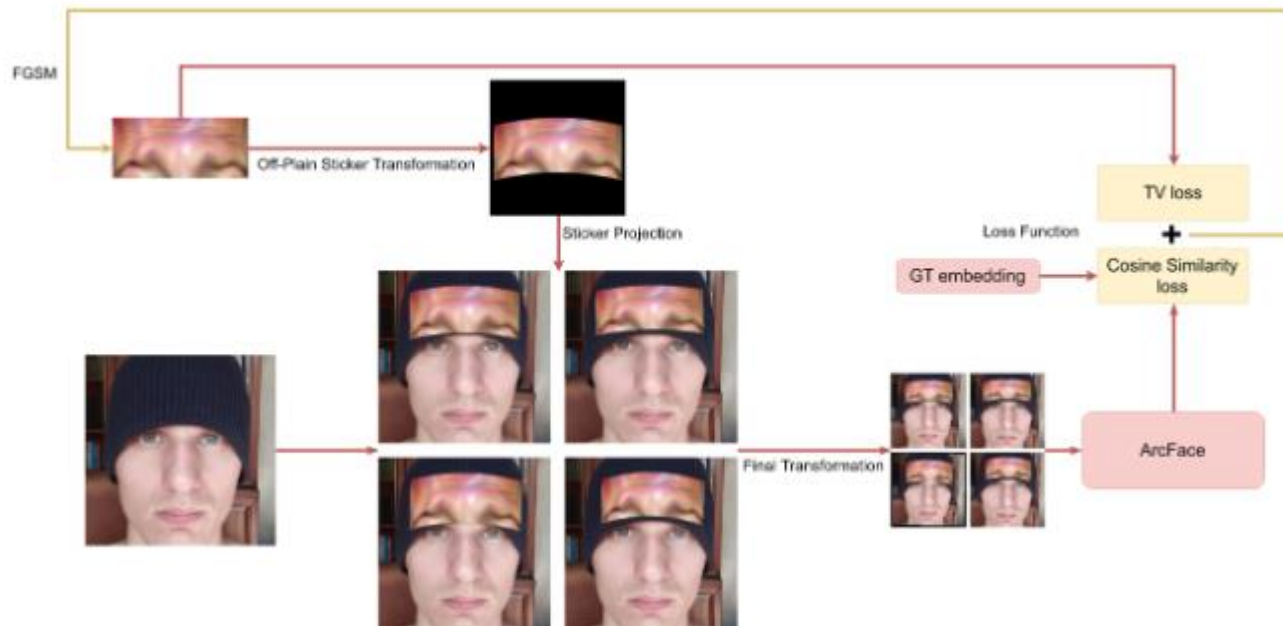
[Brown et al. 2017]

- ▶ nie wymaga precyzyjnej modyfikacji obrazu
- ▶ nie zależy od konkretnej sieci neuronowej
- ▶ łatwe do zastosowania i rozpowszechnienia

- ▶ działa tylko w przypadku klasyfikacji jednego obiektu (jednej klasy)

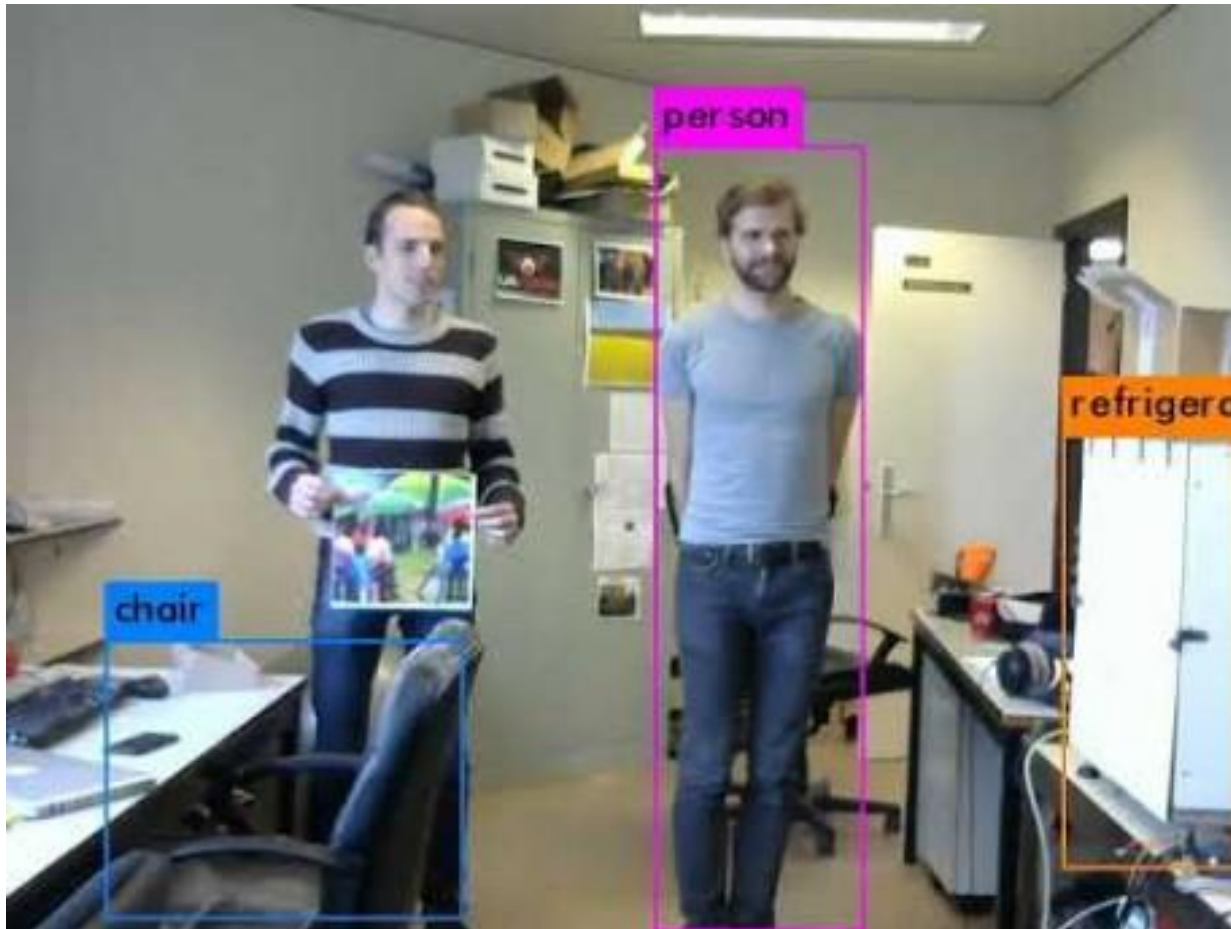
Adversarial Sticker (Face ID system)

[Komkov et al. 2019]



Adversarial Sticker (Face ID system)

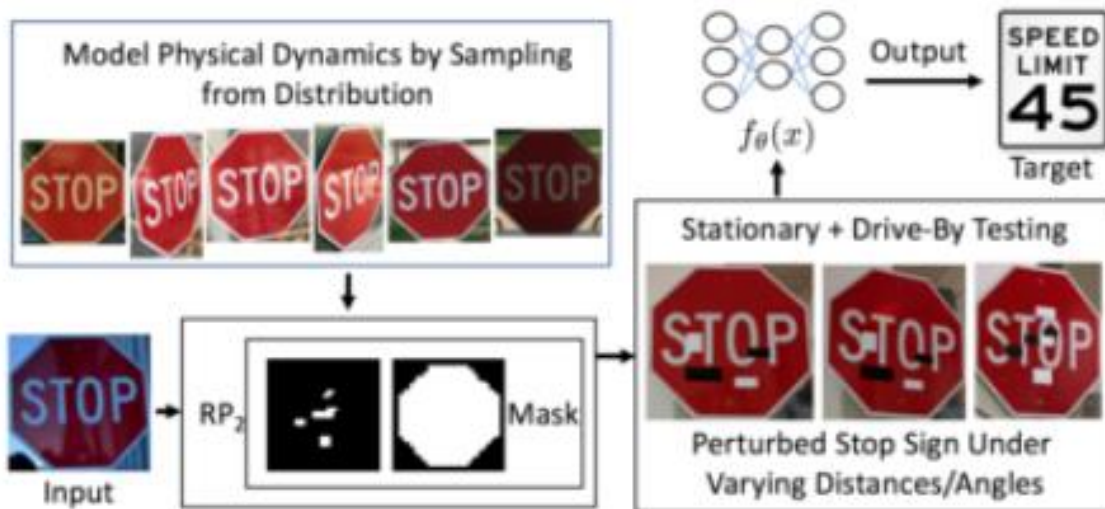
[Komkov et al. 2019]



<https://youtu.be/MlbFvK2S9g8>


























Physical-world attacks

[Eykholt et al. 2018]



Physical-world attacks

[Eykholt et al. 2018]

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5' 0°					
5' 15°					
10' 0°					
10' 30°					
40' 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Adversarial 3D objects

[Athalye et al. 2018]



Adversarial 3D objects

[Athalye et al. 2018]



■ classified as turtle ■ classified as rifle
■ classified as other

Adversarial 3D objects

[Athalye et al. 2018]



Original: speedboat



$P(\text{true}): 14\%$
 $P(\text{adv}): 0\%$



$P(\text{true}): 1\%$
 $P(\text{adv}): 0\%$



$P(\text{true}): 1\%$
 $P(\text{adv}): 0\%$



$P(\text{true}): 1\%$
 $P(\text{adv}): 0\%$



Adv: crossword
puzzle



$P(\text{true}): 3\%$
 $P(\text{adv}): 91\%$



$P(\text{true}): 0\%$
 $P(\text{adv}): 100\%$



$P(\text{true}): 0\%$
 $P(\text{adv}): 100\%$



$P(\text{true}): 0\%$
 $P(\text{adv}): 100\%$

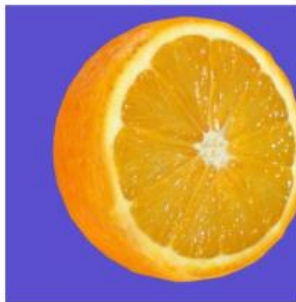
Adversarial 3D objects

[Athalye et al. 2018]

Original: orange



$P(\text{true}): 73\%$
 $P(\text{adv}): 0\%$



$P(\text{true}): 29\%$
 $P(\text{adv}): 0\%$



$P(\text{true}): 20\%$
 $P(\text{adv}): 0\%$

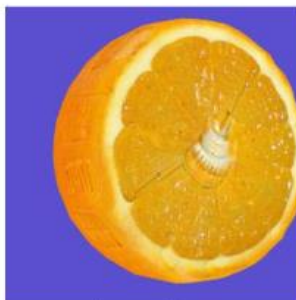


$P(\text{true}): 85\%$
 $P(\text{adv}): 0\%$

Adv: power drill



$P(\text{true}): 0\%$
 $P(\text{adv}): 89\%$



$P(\text{true}): 4\%$
 $P(\text{adv}): 75\%$



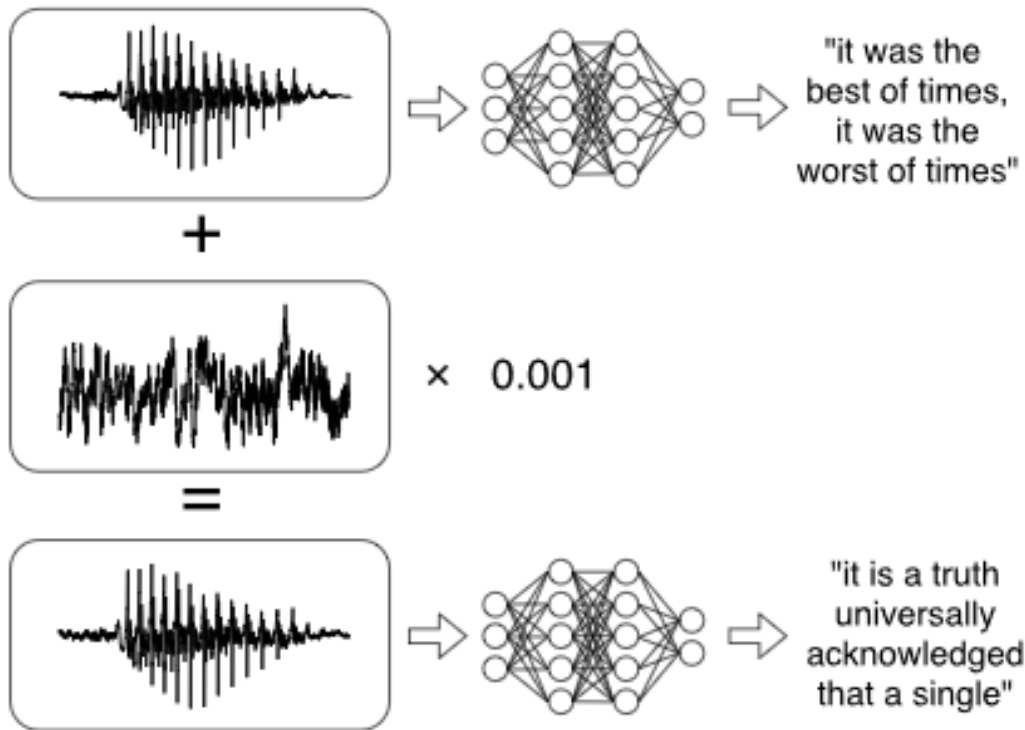
$P(\text{true}): 0\%$
 $P(\text{adv}): 98\%$



$P(\text{true}): 0\%$
 $P(\text{adv}): 84\%$

Audio Adversarial Examples

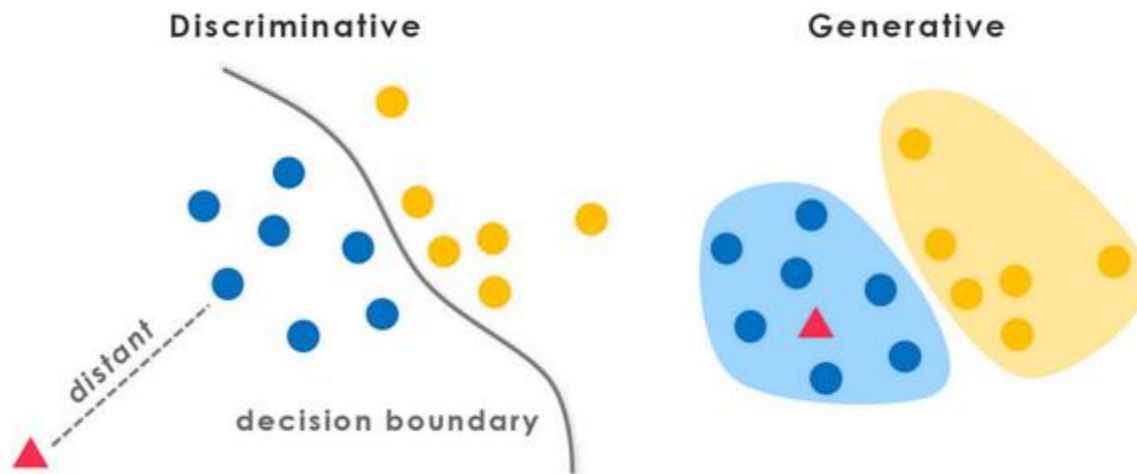
[Carlini et al. 2018]



https://nicholas.carlini.com/code/audio_adversarial_examples

Mam sieć neuronową. Co robić? Jak żyć?

- ▶ rozszerzenie zbioru uczącego
- ▶ „mieszanka” różnych architektur sieci
- ▶ ukrywanie gradientu (*gradient masking*)
- ▶ adversarial learning



Mam sieć neuronową. Co robić? Jak żyć?

- ▶ nadal klasyfikatory nie są doskonałe, więc ich zastosowanie jest ograniczone a projektowane systemy z ich użyciem są przygotowane na błędy
- ▶ większość ataków dedykowana dla konkretnej architektury sieci - potrzeba jej znajomości* lub setek tysięcy odpytań (klasyfikacji)

Bibliografia

- [1] Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).
- [2] Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016.
- [3] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation* (2019).
- [4] Elsayed, Gamaleldin, et al. "Adversarial examples that fool both computer vision and time-limited humans." *Advances in Neural Information Processing Systems*. 2018.
- [5] Brown, Tom B., et al. "Adversarial patch." *arXiv preprint arXiv:1712.09665* (2017).
- [6] Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Bibliografia

[7] Athalye, Anish, et al. "Synthesizing robust adversarial examples." *arXiv preprint arXiv:1707.07397* (2017).

[8] Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018.

[9] Thys, Simen, Wiebe Van Ranst, and Toon Goedemé. "Fooling automated surveillance cameras: adversarial patches to attack person detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[10] Papernot, Nicolas, et al. "Distillation as a defense to adversarial perturbations against deep neural networks." *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016.