Adversarial examples

Гельван Кирилл
НИУ ВШЭ
15.11.19

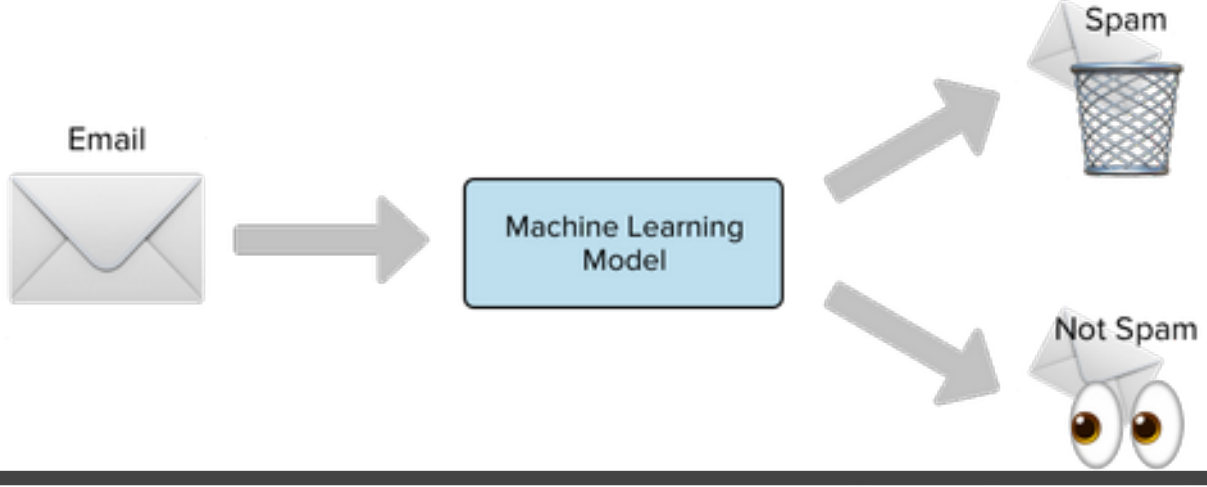
Что сейчас будет?

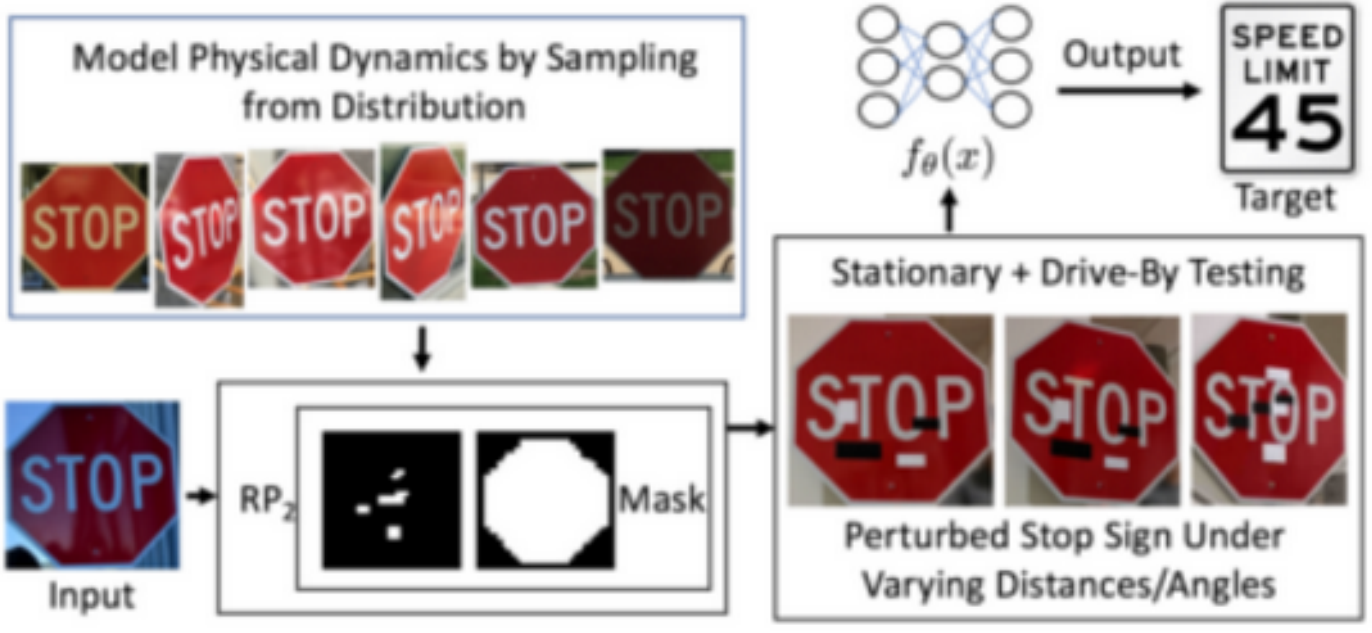
I. Что такое adversarial examples?

I. Как создавать adversarial examples?

I. Как защищаться от adversarial атак?

I. Что такое adversarial examples?

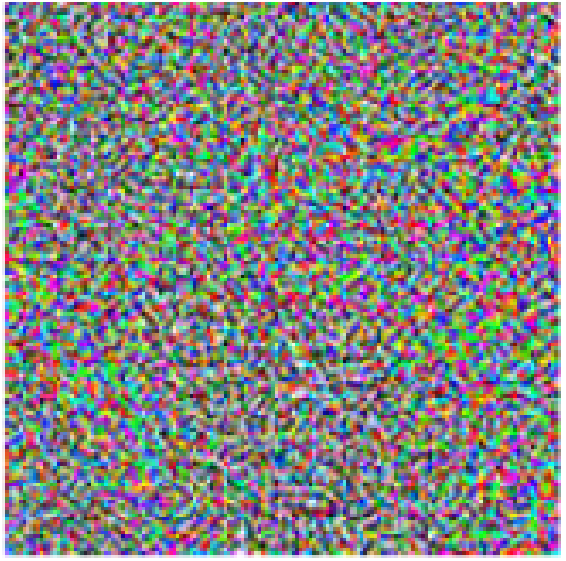

Email
Machine Learning Model
Spam
Not Spam


Model Physical Dynamics by Sampling from Distribution
STOP
Input
RP2
Mask
$f_\theta(x)$
Output
SPEED LIMIT 45
Target
Stationary + Drive-By Testing
Perturbed Stop Sign Under Varying Distances/Angles

Оригинал

Вредоносный экземпляр

$+\epsilon$



=

"panda"

57.7% confidence

?

"gibbon"

99.3% confidence

Наложение на картинку, может заставить классификатор определить панду в категорию «гиббон»

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\| \cdot \|\|_p$	l_p норма
∇_t	градиент по t

Постановка задачи

Дано:

Модель f, оригинал x

Задача:

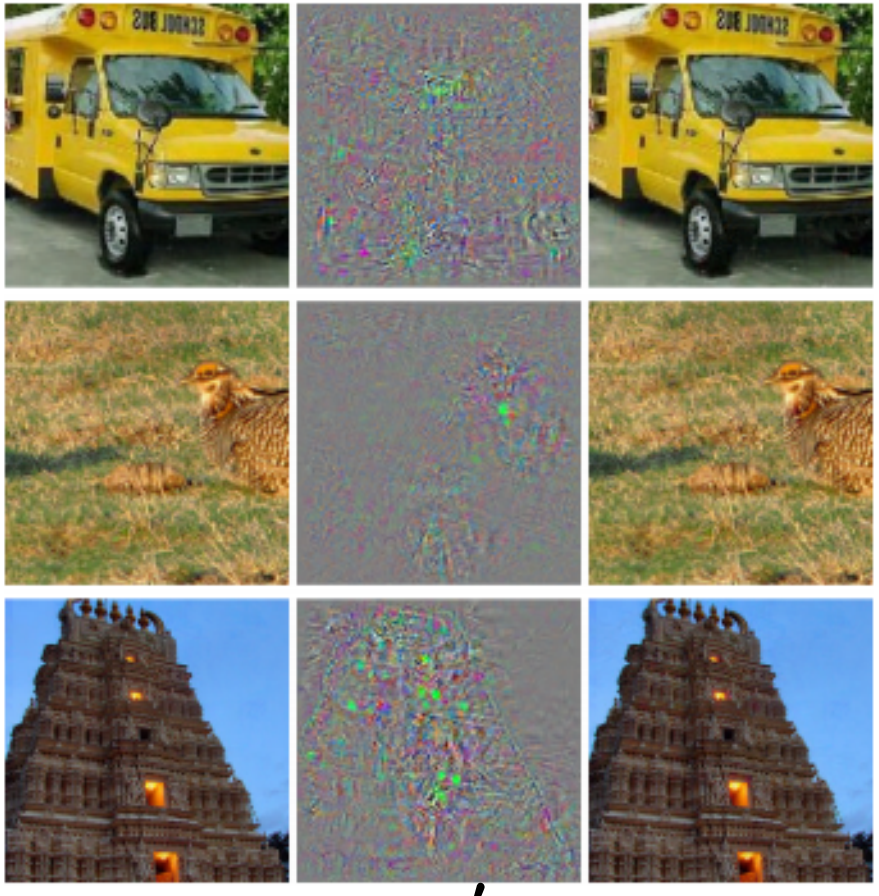
$$\min_{x'} \quad \|x' - x\|$$

$$s.t. \quad f(x') = l',$$
$$f(x) = l,$$
$$l \neq l',$$
$$x' \in [0, 1],$$

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\cdot\|_p$	l_p норма
∇_t	градиент по t

$$\eta = x' - x$$

Perturbation ~ разница, отклонение



$x + \epsilon x' =$ страус

$x + \epsilon x' =$ страус

Threat model

Adversarial Falsification

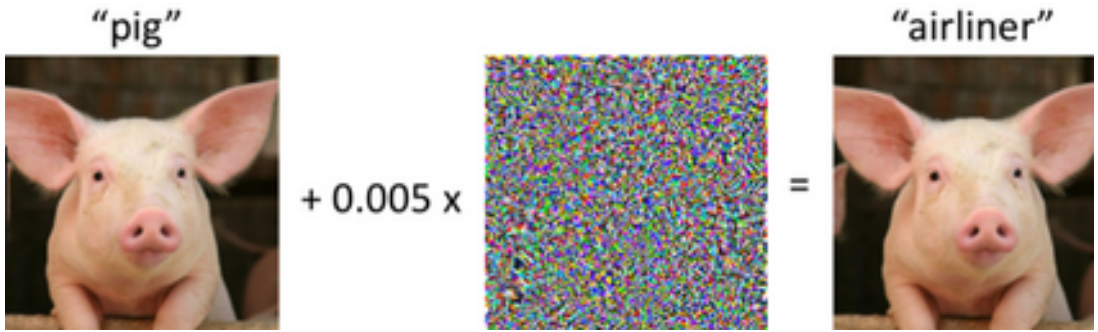
Adversarial Specificity

Adversary's Knowledge

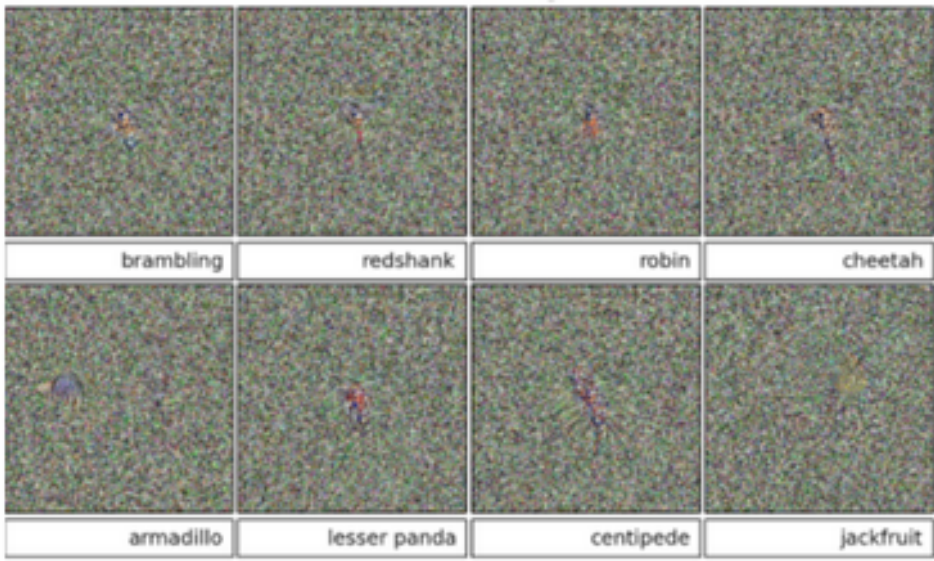
Attack Frequency

Adversarial Falsification

False Negative



False Positive



Уверенность >= 99%

Adversary’s Knowledge

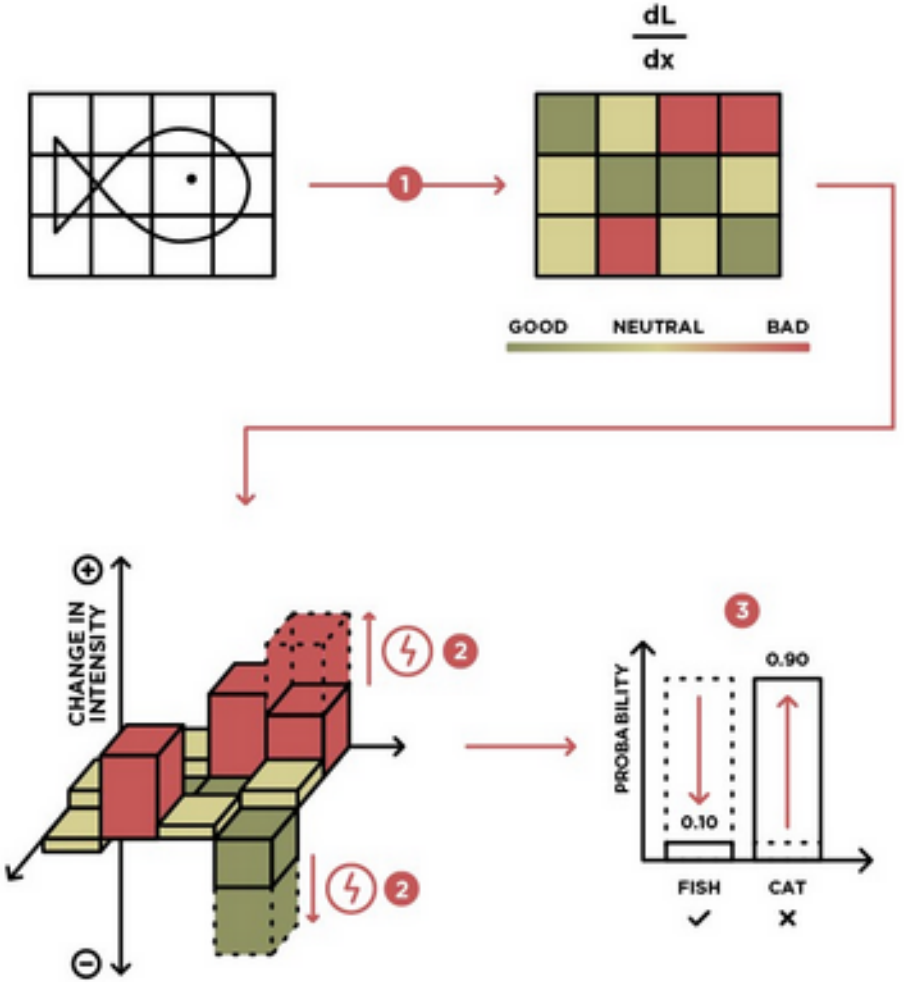
White-box

Всё

Black-box

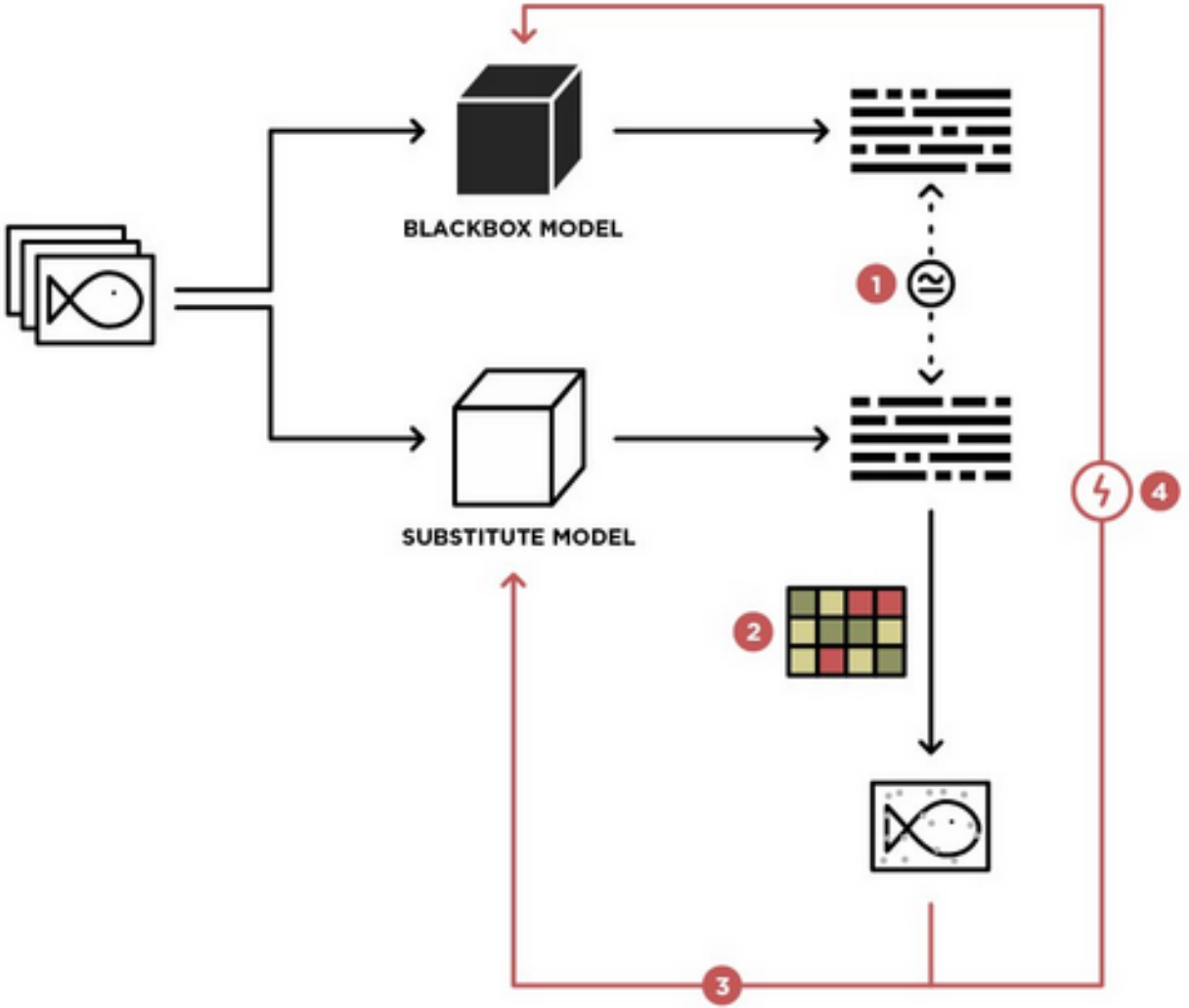
Уверенность

White-box



https://medium.com/element-ai-research-lab/tricking-a-machine-into-thinking-youre-milla-jovovich-b19bf322d55c

Black-box



https://medium.com/element-ai-research-lab/tricking-a-machine-into-thinking-youre-milla-jovovich-b19bf322d55c

Adversarial Specificity

Targeted attacks

$$P(f(x') = l') \rightarrow \max$$

Non-targeted attacks

a) $P(f(x) = l) \rightarrow min$

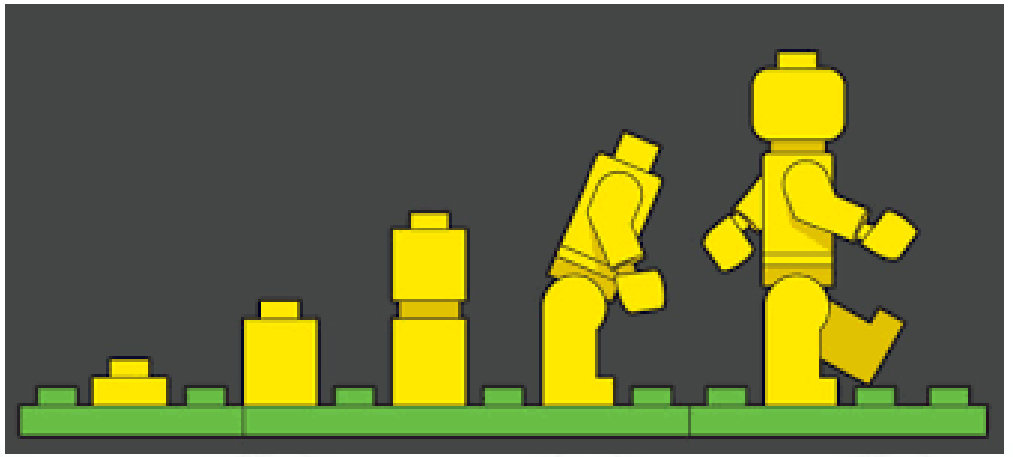
b) $l_k = \min(\eta), l_k \in l_1 \dots l_t$

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \rightarrow l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\| \cdot \|\|_p$	l_p норма
∇_t	градиент по t

Attack Frequency

One-time attacks

Iterative attacks



II. Как создавать adversarial examples?

L-BFGS

Limited-memory Broyden–Fletcher–Goldfarb–Shanno

$$\min_{x'} \quad c\|\eta\| + J_\theta(x', l')$$
$$s.t. \quad x' \in [0, 1].$$

$\Rightarrow$ Линейный поиск по c

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\cdot\|_p$	l_p норма
∇_t	градиент по t

Страус

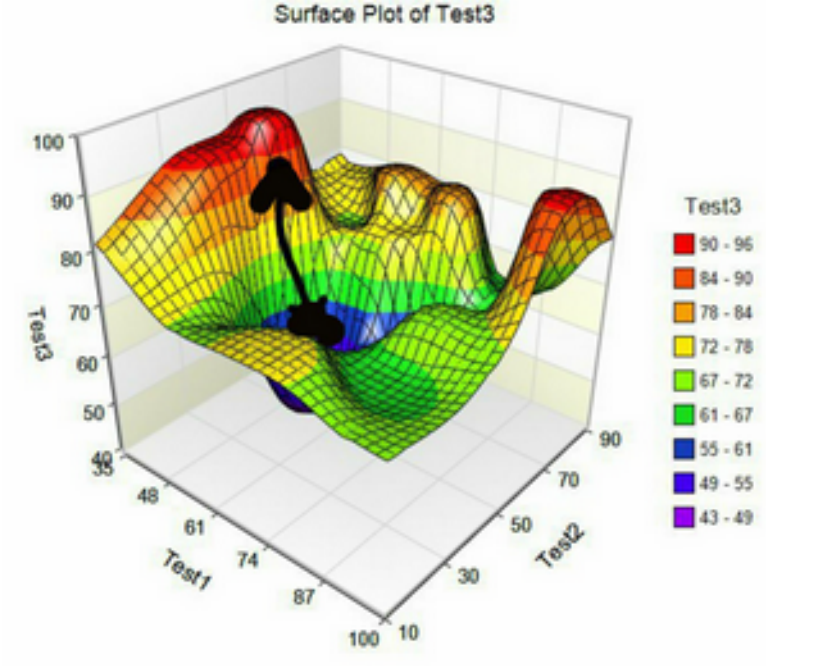
FGSM

Fast Gradient Sign Method

We are only interested in the sign of the slopes to know if we want to increase or decrease the pixel values

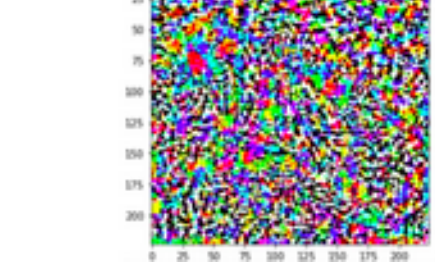
$$\eta = \epsilon sign(\nabla_x J_\theta(x, l)),$$
$$x' = x + \eta$$

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\| \cdot \|\|_p$	l_p норма
∇_t	градиент по t



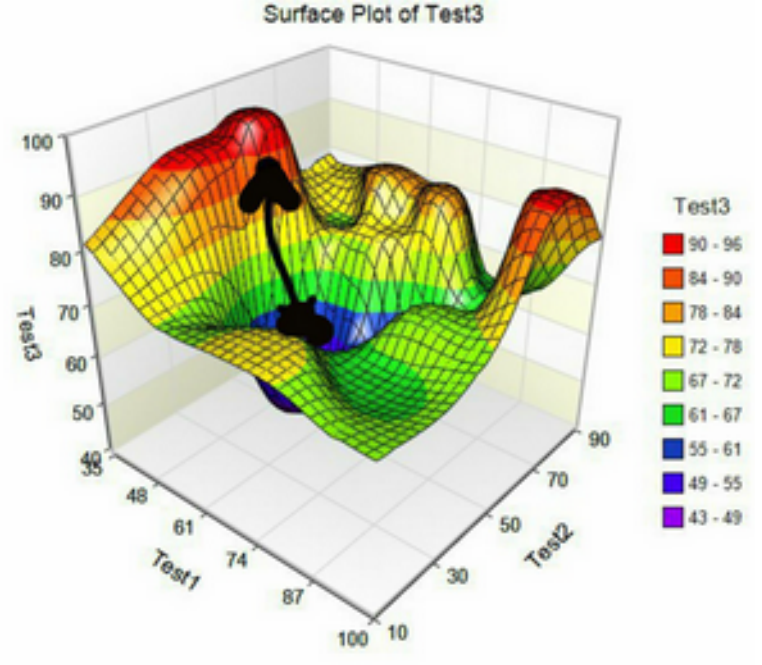
FGSM

Fast Gradient Sign Method



$$\eta = \epsilon sign(\nabla_x J_\theta(x, l)),$$
$$x' = x + \eta$$

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\| \cdot \|\|_p$	l_p норма
∇_t	градиент по t

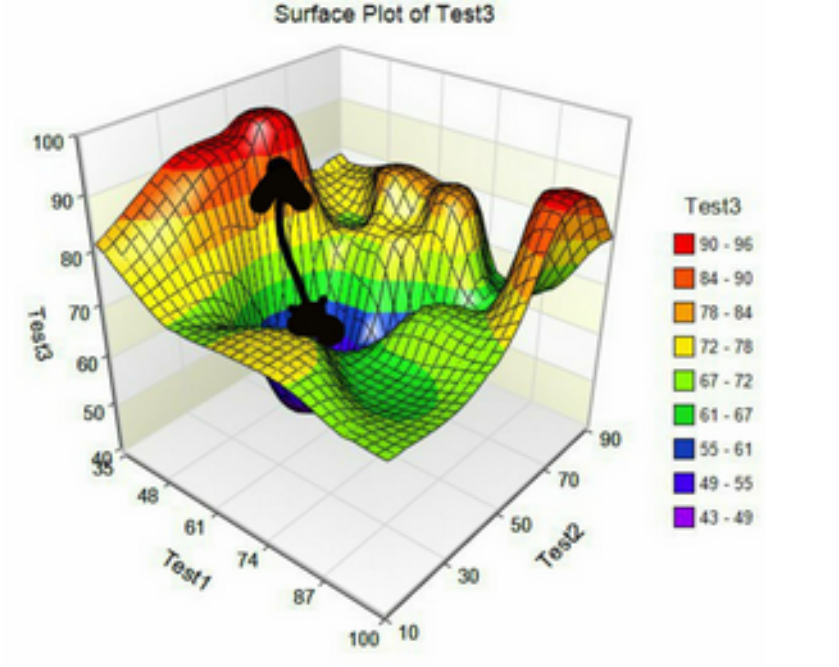


FGV

Fast Gradient Value Method

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\| \cdot \|\|_p$	l_p норма
∇_t	градиент по t

$$\eta = \nabla_x J(\theta, x, l)$$
$$x' = x + \eta$$



FGSM

$$\eta = \epsilon sign(\nabla_x J_\theta(x, l)),$$
$$x' = x + \eta$$

Обозначние	Значение
x	оригинальные данные
l	номер класса $(1 \dots m)$
x'	adversarial example
l'	номер класса (если adv. ex. направ.)
$f(\cdot)$	модель $(f \in F : \mathbb{R}^n \to l)$
θ	параметры модели f
$J_f(\cdot, \cdot)$	функция потерь для модели f
η	разница между оригиналом и adv. ex. $(\eta = x' - x)$
$\|\cdot\|_p$	l_p норма
∇_t	градиент по t

$$\mathbf{g}_{t+1} = \mu \mathbf{g}_t + \frac{\nabla_x J_\theta(x'_t, l)}{\|\nabla_x J_\theta(x'_t, l)\|},$$

FGSM + Momentum

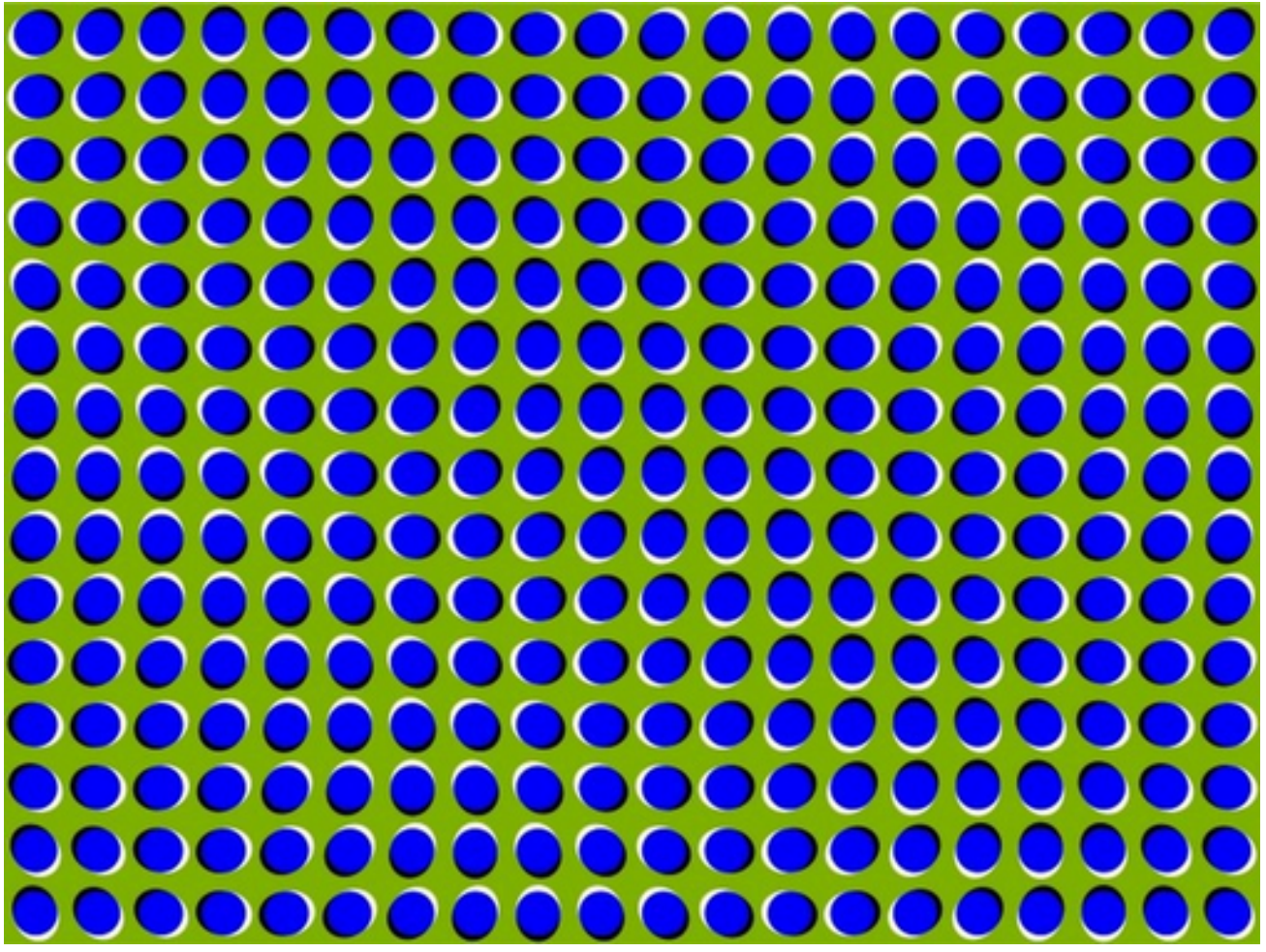
$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha\, \mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\right)\right\}$$

Basic Iterative Method (BIM)

$$x' = x - \epsilon sign(\nabla_x J(\theta, x, l')).$$

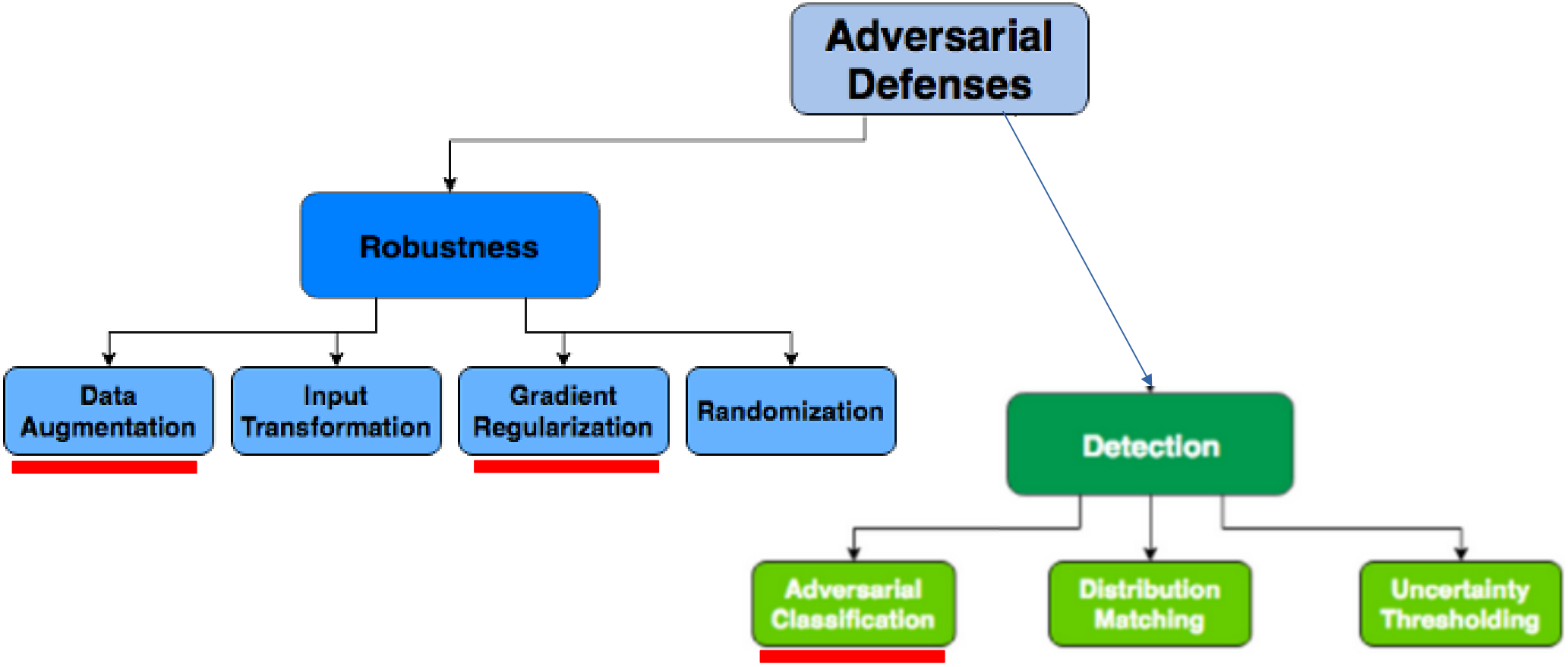
One-step Target Class Method (OTCM)

Почувствуй себя машиной

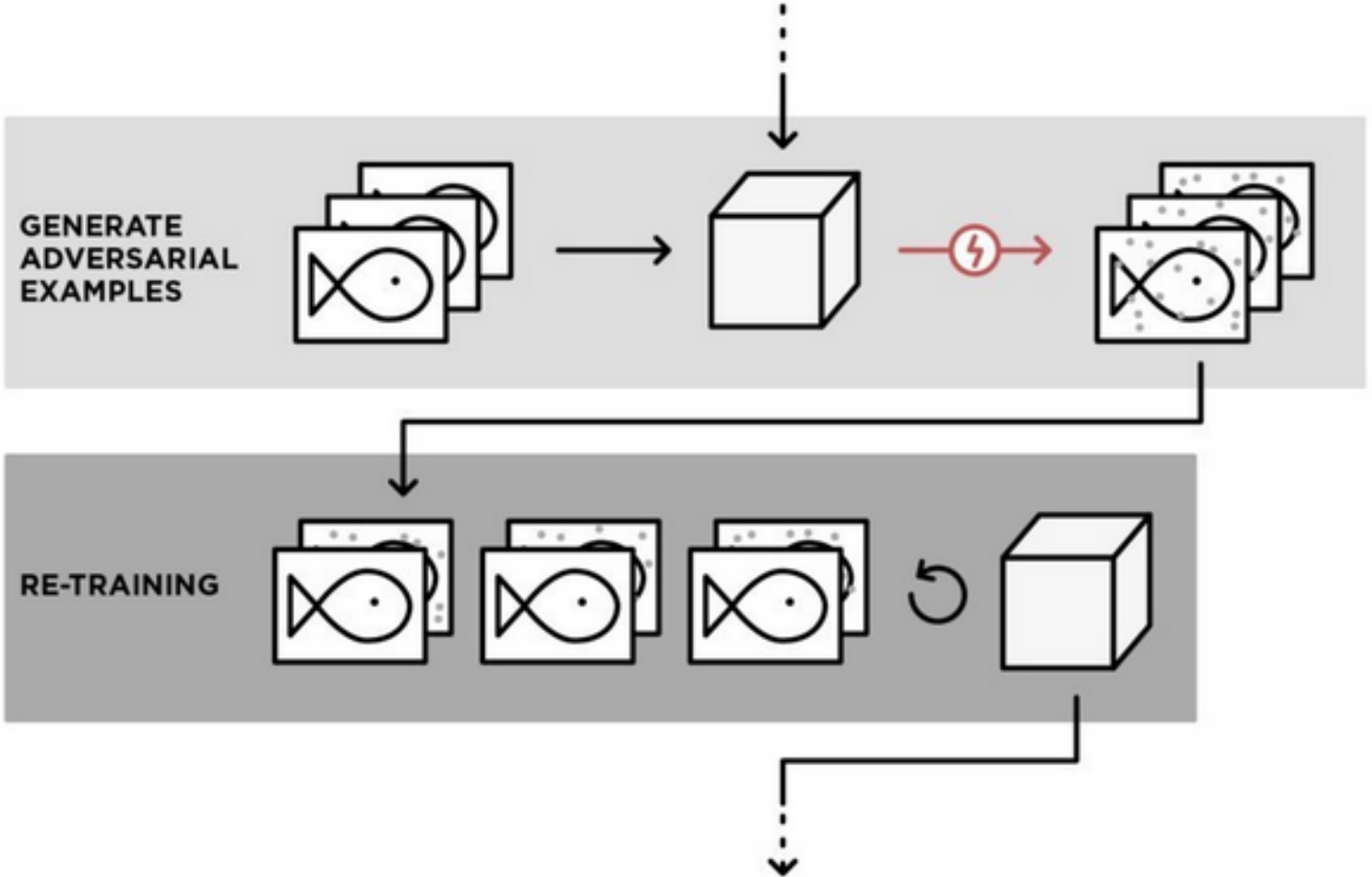


III. Как защищаться от adversarial атак?

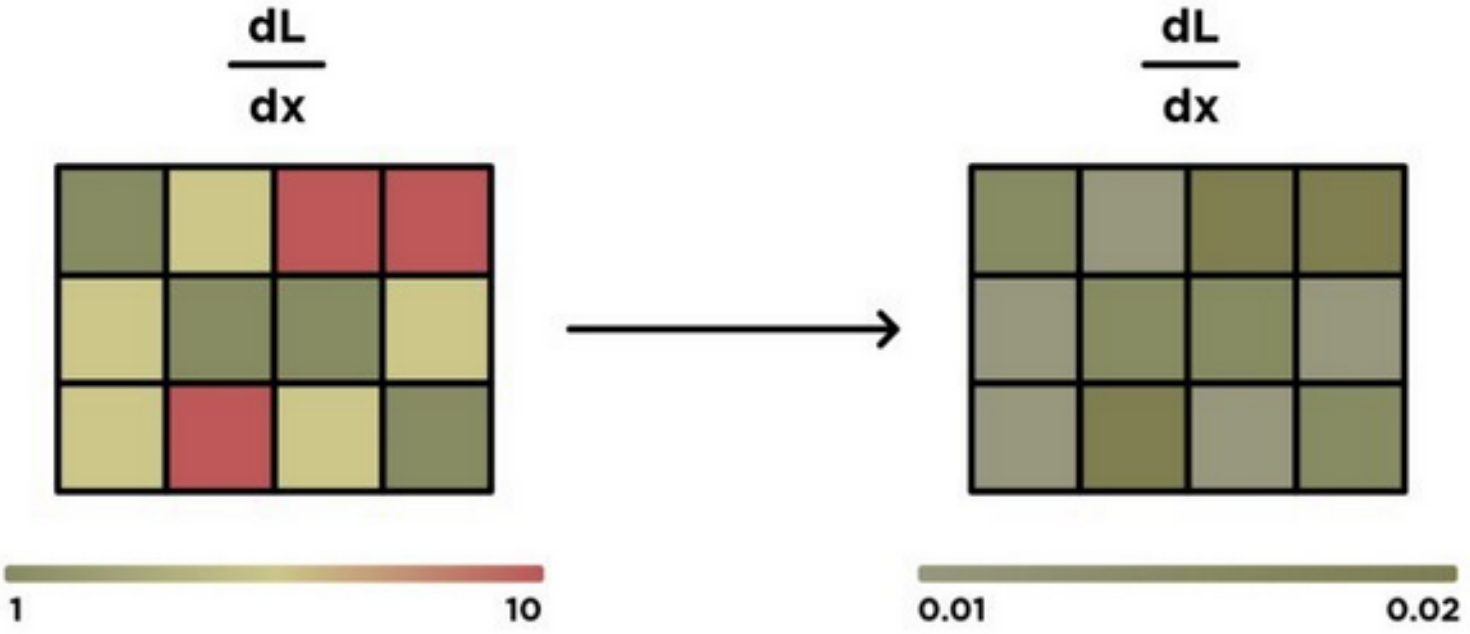
Adversarial Defences



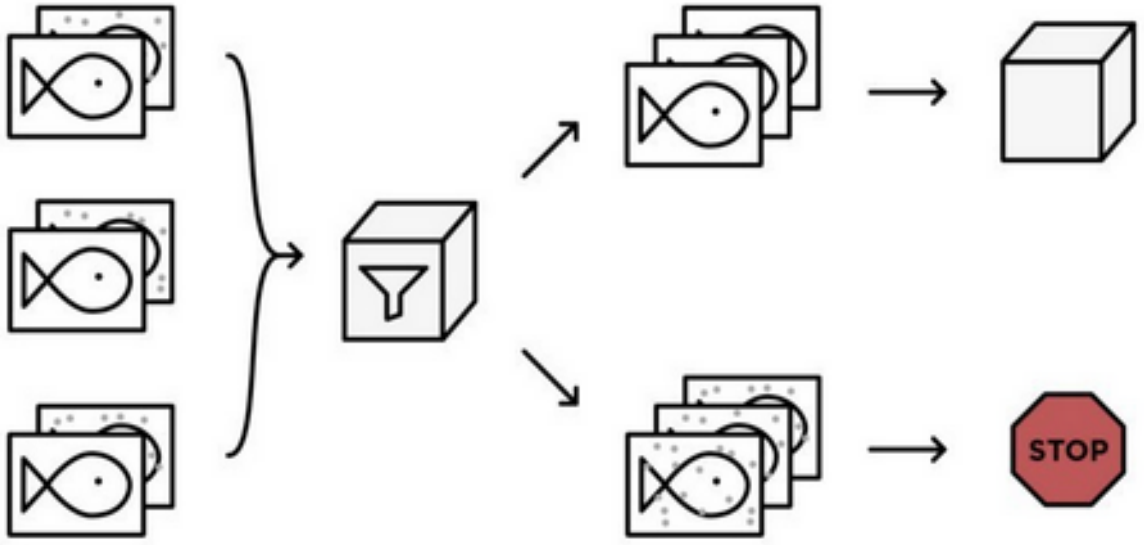
Adversarial Training



Regularizing the Gradient of the Model



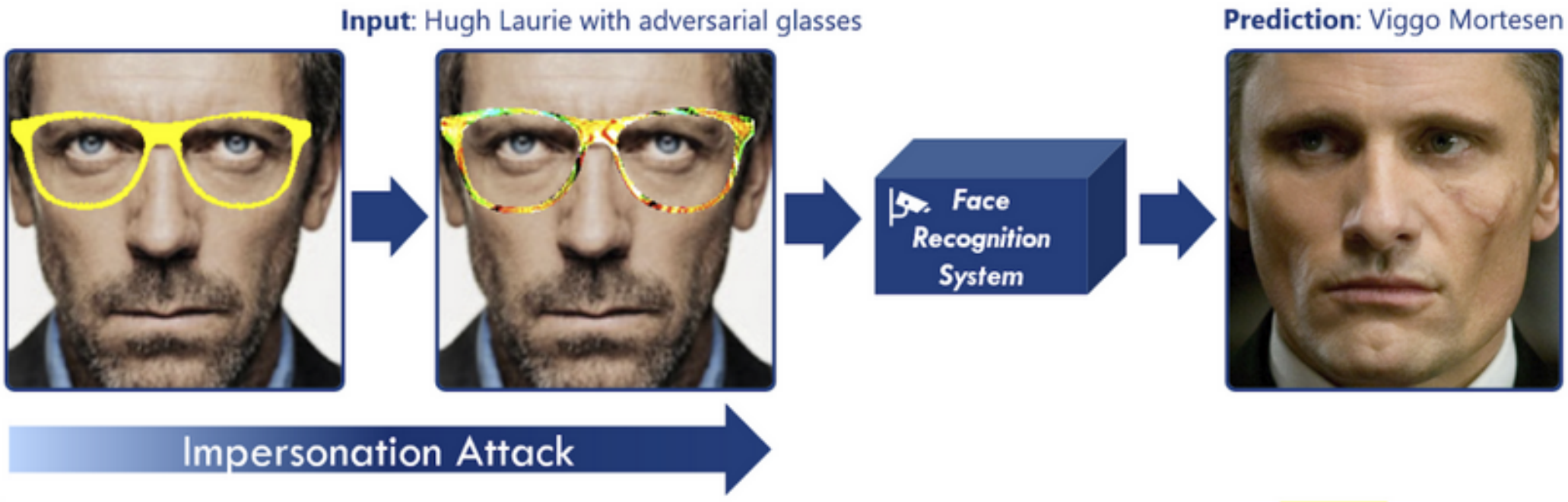
Detecting Adversaries Through Classification



Adversarival training with BIM

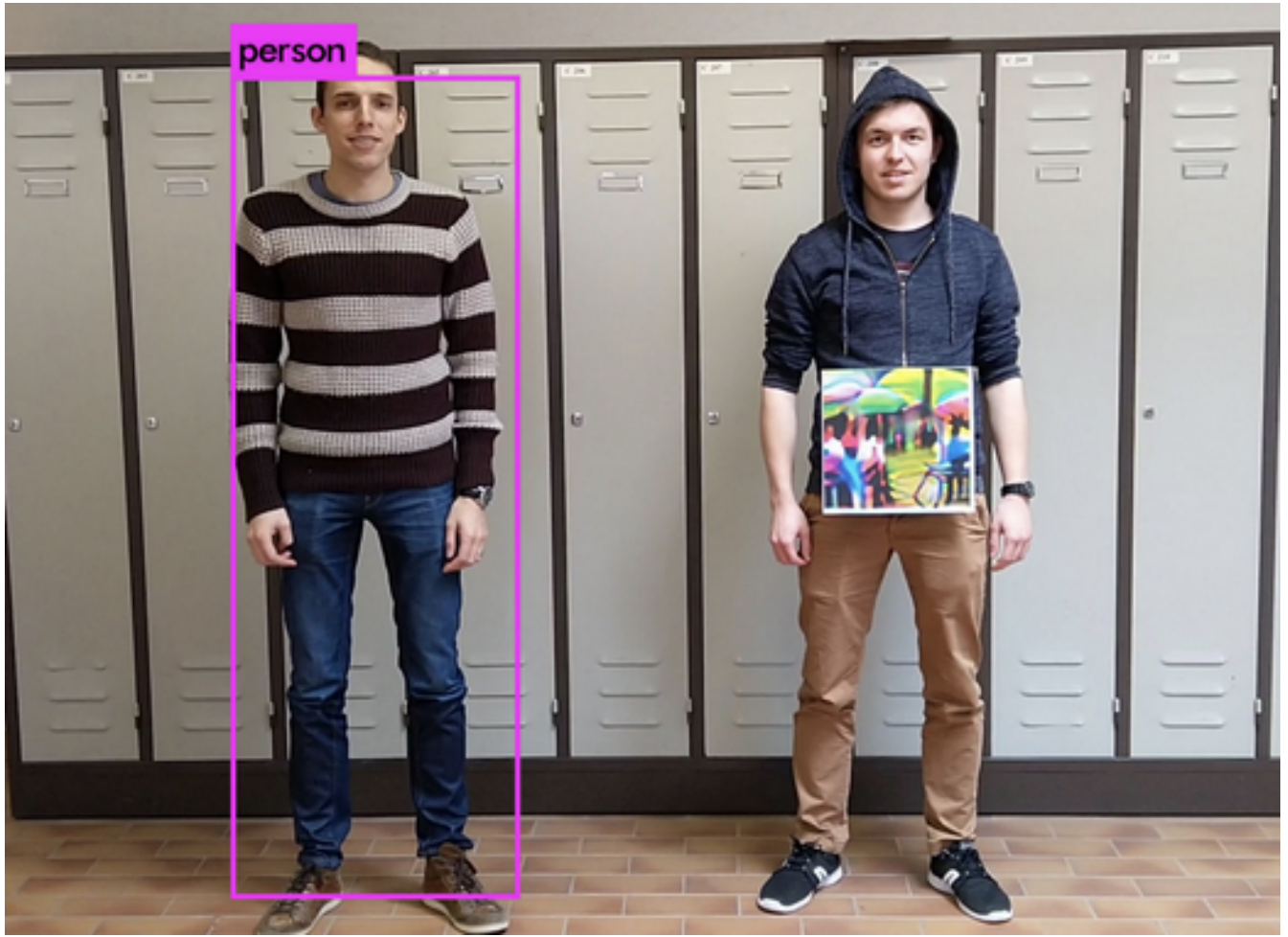
$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\right)\right\}$$

Face recognition



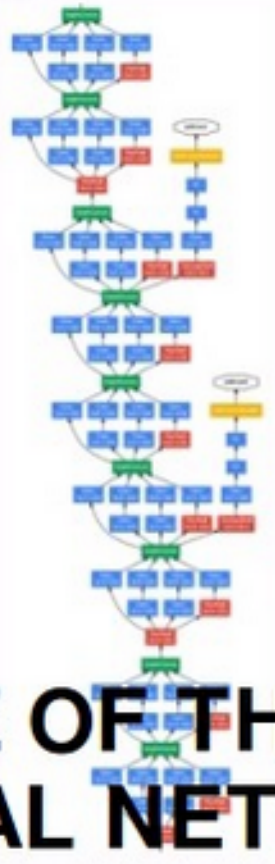
https://www.inovex.de/blog/machine-perception-face-recognition/

Object detection



https://medium.com/syncedreview/now-you-see-me-now-you-dont-fooling-a-person-detector-aa100715e396

WHO WOULD WIN?



STATE OF THE ART NEURAL NETWORK

ONE NOISY BOI

Вопросы:

- Чем отличаются white и black box атаки?
- Как работает метод FGSM?
- Опишите своими словами любой метод Adversarial Defence.

References:

https://arxiv.org/pdf/1312.6199.pdf
https://arxiv.org/pdf/1412.6572.pdf
https://arxiv.org/pdf/1712.07107.pdf
http://www.cleverhans.io/

Картинки:

https://medium.com/element-ai-research-lab/securing-machine-learning-models-against-adversarial-attacks-b6cd5d2be8e2
https://medium.com/@ml.at.berkeley/tricking-neural-networks-create-your-own-adversarial-examples-a61eb7620fd8
https://medium.com/element-ai-research-lab/tricking-a-machine-into-thinking-youre-milla-jovovich-b19bf322d55c
https://towardsdatascience.com/adversarial-examples-in-deep-learning-be0b08a94953
https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522?gi=324a60138bcc