Adversarial attacks

Паркин Александр

Adversarial examples

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Goodfellow et al., "*Explaining and Harnessing Adversarial Examples", 2015*

Типы атак

- White box
- Black box
- Промежуточный вариант

Мера изменения изображения

- ℓ_0
- ℓ_2
- ℓ_∞

White box methods

- L-BFGS

$$\min_{\boldsymbol{\rho}} \; c|\boldsymbol{\rho}| + \mathcal{L}(\mathbf{I}_c + \boldsymbol{\rho}, \ell) \;\; s.t. \; \mathbf{I}_c + \boldsymbol{\rho} \in [0,1]^m$$

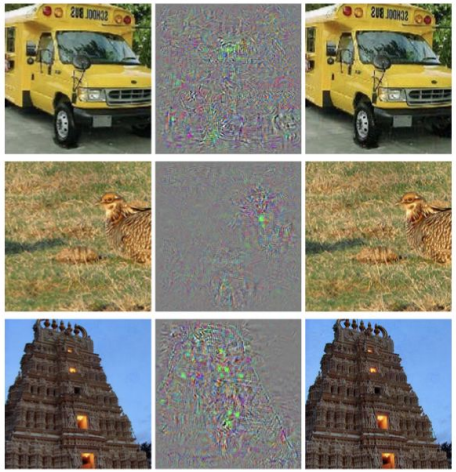
- FSGM (Fast gradient sign method)

$$\boldsymbol{\rho} = \epsilon \, \text{sign}\left(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)\right)$$

- I-FGSM (Iterative FGSM)

$$\mathbf{I}_{\boldsymbol{\rho}}^{i+1} = \text{Clip}_{\epsilon}\left\{\mathbf{I}_{\boldsymbol{\rho}}^{i} + \alpha \, \text{sign}(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_{\boldsymbol{\rho}}^{i}, \ell)\right\}$$

- JSMA(Jacobian-based Saliency Map Attack)



original image | noise | “ostrich”

Szegedy et al., “Intriguing properties of neural networks”, 2014

Goodfellow et al., “*Explaining and Harnessing Adversarial Examples”, 2015*

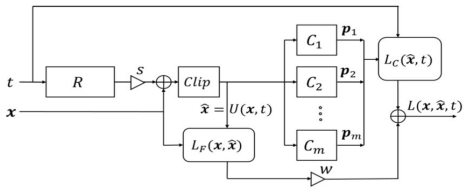
A. Kurakin et al., “*Adversarial examples in the physical world”, 2016*

Papernot et al., “*The Limitations of Deep Learning in Adversarial Settings”, 2016*

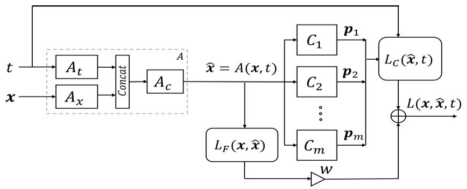
Black box methods

- Метод имитации отжига
- Эволюционные алгоритмы (One pixel attack)
- Дистилляция сети
- UPSET and ANGRI (Генеративные сети)

$$L(\mathbf{x}, \hat{\mathbf{x}}, t) \quad = \quad L_C(\hat{\mathbf{x}}, t) + L_F(\mathbf{x}, \hat{\mathbf{x}}) = -\sum_{i=1}^{m} log(C_i(\hat{\mathbf{x}})[t]) \; + \; w \parallel \hat{\mathbf{x}} - \mathbf{x} \parallel_k^k,$$



(a) Training scheme for UPSET(U).



(b) Training scheme for ANGRI(A).

Su et al., "*One pixel attack for fooling deep neural networks", 2017*

Sarkar et al., "*UPSET and ANGRI: Breaking High Performance Image Classifier*", 2017

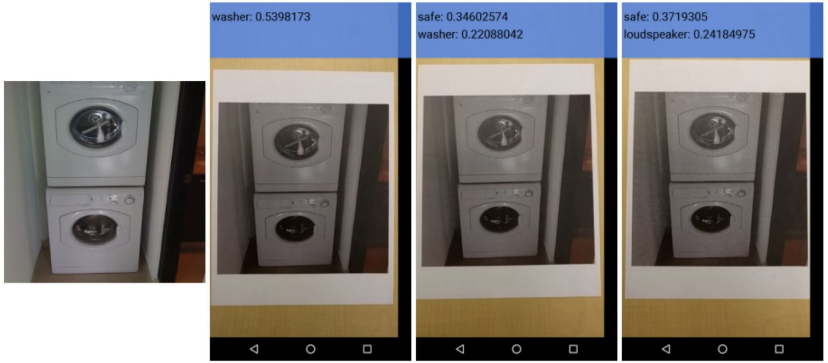
Общая таблица атак

Method	Black/White box	Targeted/Non-targeted	Specific/Universal	Perturbation norm	Learning	Strength
L-BFGS [22]	White box	Targeted	Image specific	ℓ_∞	One shot	$***$
FGSM [23]	White box	Targeted	Image specific	ℓ_∞	One shot	$***$
BIM & ILCM [35]	White box	Non targeted	Image specific	ℓ_∞	Iterative	$****$
JSMA [60]	White box	Targeted	Image specific	ℓ_0	Iterative	$***$
One-pixel [68]	Black box	Non Targeted	Image specific	ℓ_0	Iterative	$**$
C&W attacks [36]	White box	Targeted	Image specific	$\ell_0, \ell_2, \ell_\infty$	Iterative	$*****$
DeepFool [72]	White box	Non targeted	Image specific	ℓ_2, ℓ_∞	Iterative	$****$
Uni. perturbations [16]	White box	Non targeted	Universal	ℓ_2, ℓ_∞	Iterative	$*****$
UPSET [146]	Black box	Targeted	Universal	ℓ_∞	Iterative	$****$
ANGRI [146]	Black box	Targeted	Image specific	ℓ_∞	Iterative	$****$
Houdini [131]	Black box	Targeted	Image specific	ℓ_2, ℓ_∞	Iterative	$****$
ATNs [42]	White box	Targeted	Image specific	ℓ_∞	Iterative	$****$

Naveed Akhtar and Ajmal Mian. “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey“

Атаки в реальной жизни

- Распечатанные изображения
- Дорожные знаки
- Очки против модели распознавания лиц
- Adversarial 3D-объекты



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Kurakin et al. *“Adversarial examples in the physical world”, 2016*

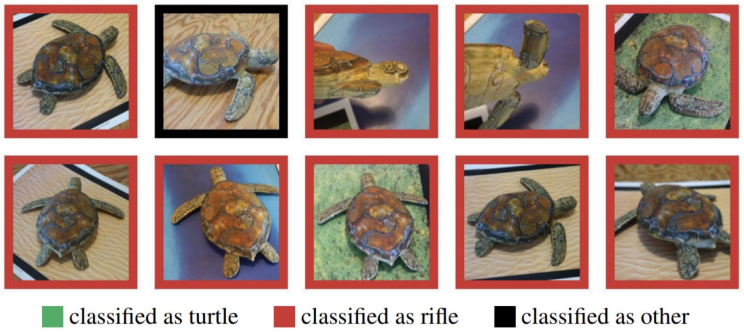
Etimov et al., *“Robust Physical-World Attacks on Deep Learning Models”, 2017*

Sharif et al., “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*”, 2016*

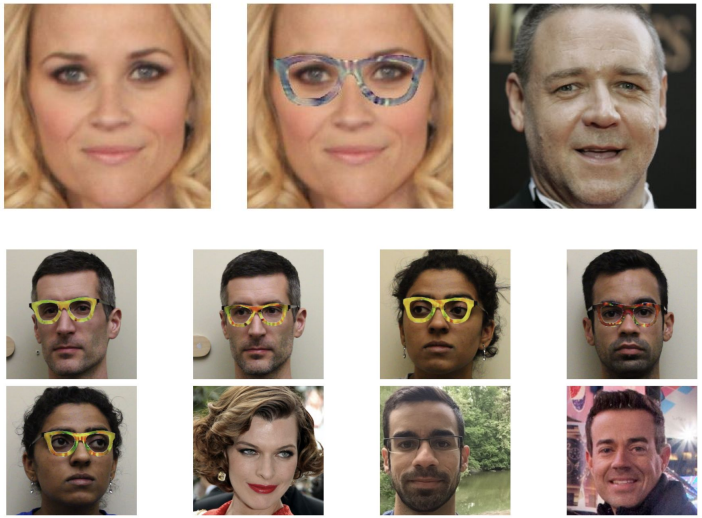
Athalye et al. *“Synthesizing Robust Adversarial Examples”, 2017*

Атаки в реальной жизни

- Распечатанные изображения
- Дорожные знаки
- Очки против модели распознавания лиц
- Adversarial 3D-объекты



https://www.youtube.com/watch?v=YXy6oX1iNoA



Kurakin et al. “*Adversarial examples in the physical world”, 2016*

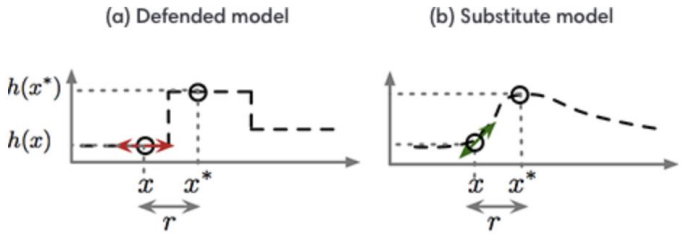
Etimov et al., *“Robust Physical-World Attacks on Deep Learning Models”, 2017*

Sharif et al., “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*”, 2016*

Athalye et al. *“Synthesizing Robust Adversarial Examples”, 2017*

Защита

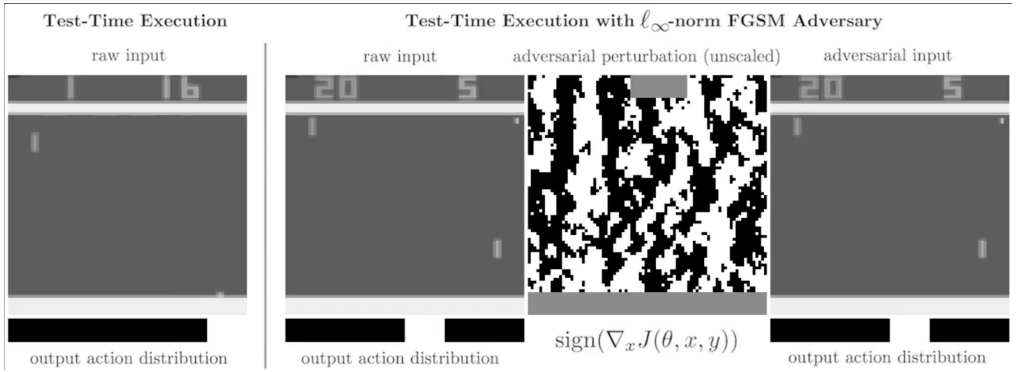
- Защита дистилляцией
- Добавление adversarial примеров в обучение
- Ансамбль классификаторов
- Обучение классификатора как часть GAN
- Детектирование атаки
- Сокрытие градиентов

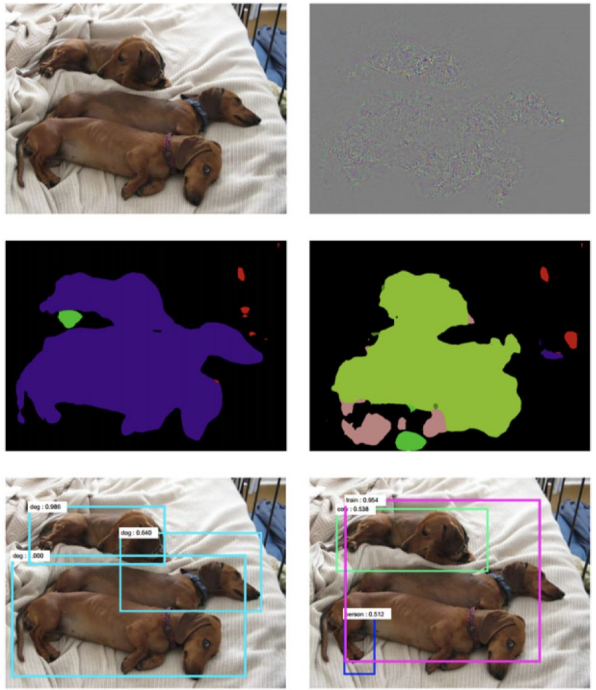


Papernot et al., *"Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples", 2016*

Атаки не только в классификации

- Детекция/сегментация изображения
- Распознавание голоса
- Классификация текста
- Deep Reinforcement Learning





Соревнования

NIPS 2017 (1 авг.- 1 окт. 2017)

Задачи:

		Топ1
● Non-targeted Adversarial Attack.	$score_{attack} = \sum_{defense \in D} \sum_{k=1}^{N} [defense(attack(Image_k)) \neq TrueLabel_k]$	0.782
● Targeted Adversarial Attack.	$score_{TargetedAttack} = \sum_{defense \in D} \sum_{k=1}^{N} [defense(TargetedAttack(Image_k)) = TargetLabel_k]$	0.402
● Defense Against Adversarial Attack.	$score_{defense} = \sum_{attack \in A} \sum_{k=1}^{N} [defense(attack(Image_k)) = TrueLabel_k]$	0.953

Submit: Docker-контейнер с исходным кодом и данными
Предоставляемые мощности:

- ОЗУ: 24 GB
- Место после разархивации: 16 GB
- Видеокарта: Tesla K40
- Ограничение по времени: 500 сек. на батч из 100 изображений

Соревнования

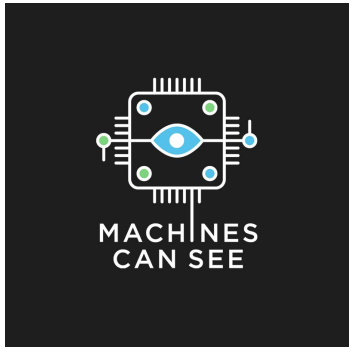
NIPS 2018 (25 авг - 15 нояб 2018)

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

Задачи:

- Non-targeted Adversarial Attack.
- Targeted Adversarial Attack.
- Defense Against Adversarial Attack.

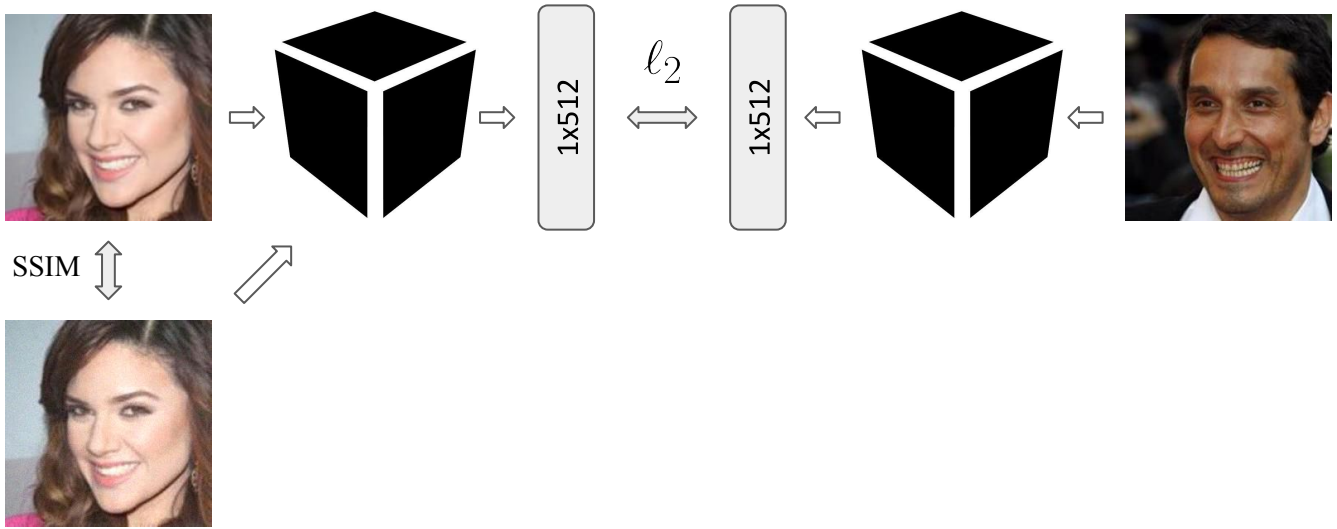
Метрика: ℓ_2 расстояние между оригинальным изображением и состязательным примером



MCS 2018. Adversarial Attacks on Black Box Face Recognition

Задача

1. Заставить модель распознавать изображение человека A как человека B
2. Атакованное изображение не вызывало подозрений и несильно отличалось от оригинального.



Схожесть изображений

SSIM (structure similarity) > 0.95

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Размер окна: 7

```
from skimage.measure import compare_ssim
```

Black box

```
import MCS2018
gpu_id = 0
net = MCS2018.Predictor(gpu_id)

def preprocess_img(img):
    MEAN = [0.485, 0.456, 0.406]
    STD = [0.229, 0.224, 0.225]
    preprocessing = transforms.Compose([
                    transforms.CenterCrop(224),
                    transforms.Scale(112),
                    transforms.ToTensor(),
                    transforms.Normalize(mean=MEAN, std=STD),
                    ])
    img_arr = preprocessing(img).unsqueeze(0).numpy()
    return img_arr

img_batch = preprocess_img(img)
descr = net.submit(img_batch)
```

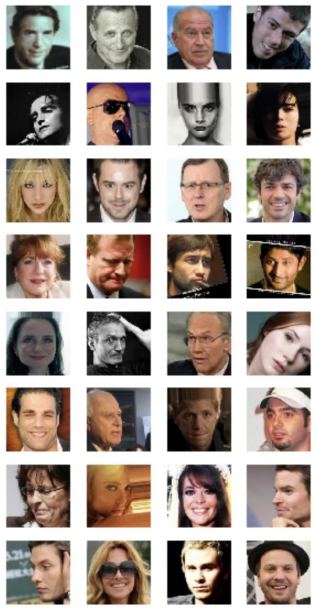
Время исполнения:
CPU: ~1.5 sec
GPU: ~7ms

OS	python 2.7	python 3.5	python 3.6
Ubuntu	CPU GPU(cuda8.0) GPU(cuda9.0) GPU(cuda9.1) GPU(cuda9.2)	CPU GPU(cuda8.0) GPU(cuda9.0) GPU(cuda9.1) GPU(cuda9.2)	CPU GPU(cuda8.0) GPU(cuda9.0) GPU(cuda9.1) GPU(cuda9.2)
CentOS	CPU GPU (cuda8.0)	CPU GPU(cuda8.0)	CPU GPU(cuda8.0)
Windows	CPU GPU (cuda 9.0)	CPU GPU (cuda 9.0)	CPU GPU (cuda 9.0)
MacOS	CPU	CPU	CPU

Данные

Source 5K

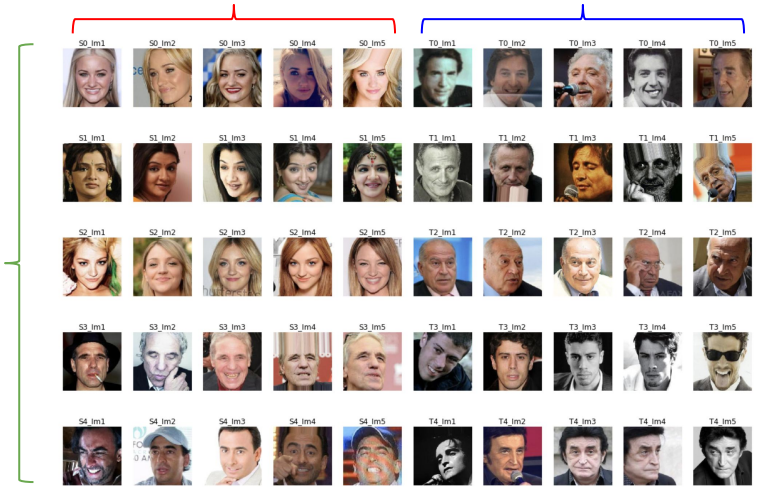
Target 5K



1M

Evaluation

Public leaderboard: 25%
Private leaderboard: 75%

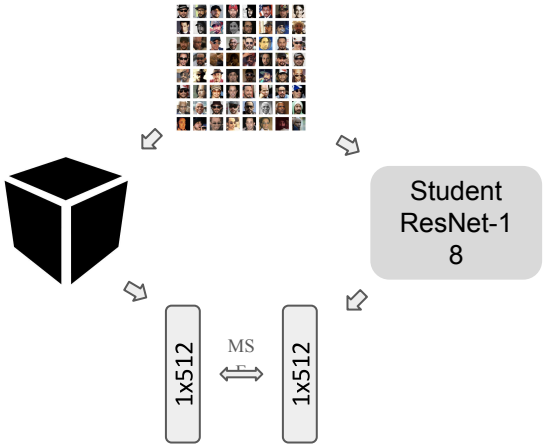


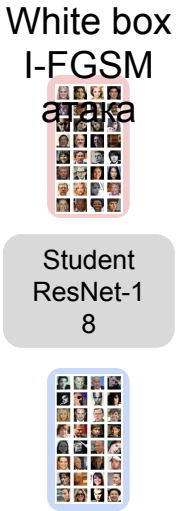
$$Score = \frac{1}{N}\frac{1}{25}\sum_{k=1..N}\sum_{i=1..5}\sum_{j=6..10}||D(G(I_s(k,i))) - D(I_t(k,j))||_2$$

Baseline code

Baseline 0

Baseline 1





https://github.com/AlexanderParkin/MCS2018.Baseline

Время проведения

14 мая 12:00 - начало соревнования

5 июня 23:59 - конец основного этапа

6 июня 23:59 - конец соревнования

Призы

1 место - 150 000 + 1080Ti

2 место - 75 000 + 1080Ti

3 место - 36 000 + 1080Ti

4 место - 24 000

5 место - 15 000

Results					
#	User	Entries	Date of Last Entry	Team Name	Score ▲
1	mortido	1	05/25/18		1.260 (1)
2	stalkermustang	3	05/24/18		1.362 (2)
3	alexey.grankov	3	05/25/18		1.403 (3)

CodaLab https://competitions.codalab.org/competitions/19090

https://github.com/AlexanderParkin/MCS2018.Baseline

bit.ly/mcs2018_telegram