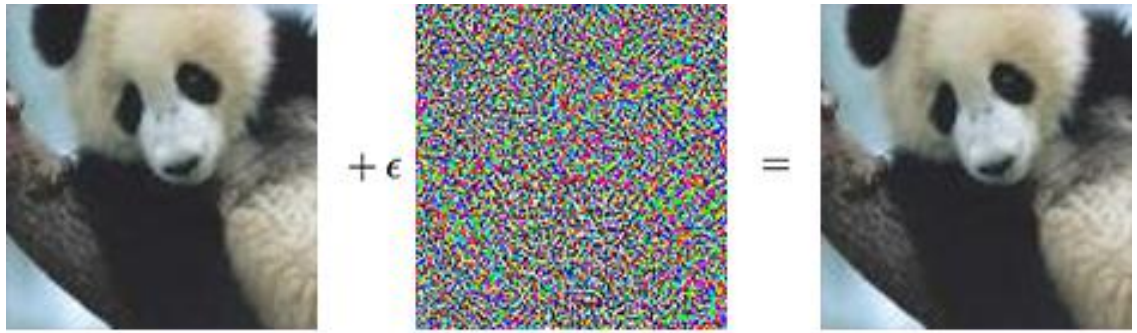
Adversarial examples

Каратаева Екатерина Владимировна
БПМИ182
НИУ ВШЭ

Определение

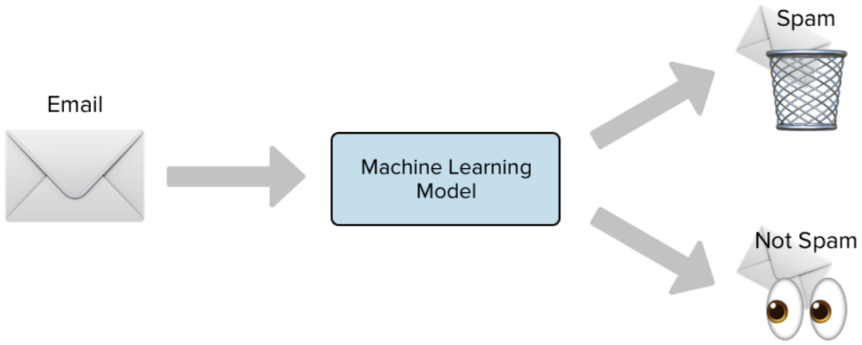
Adversarial examples - наборы данных, намеренно созданные, чтобы обмануть модель.

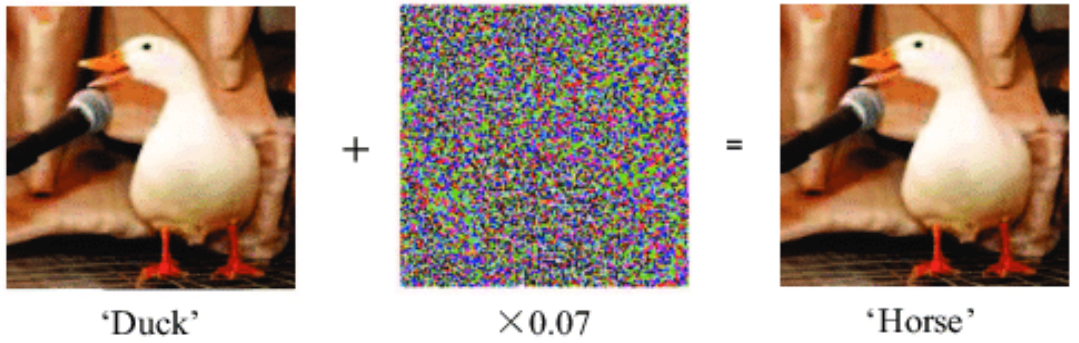
Изображения, звуки, тексты, выглядящие нормальными для людей, но воспринимаемые совершенно по-другому нейронными сетями.

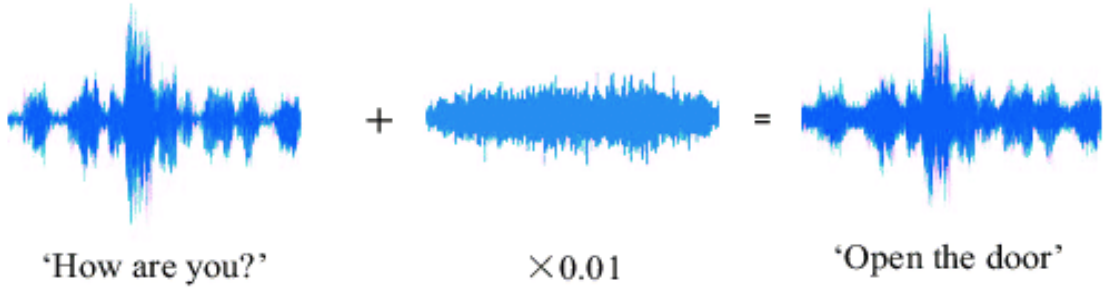


Примеры областей применения Adversarial examples:

- Ad-blockers
- Spam Classifiers
- Self Driving Cars
- Text Applications (Sentiment Analysis, Text classification)
- Computer vision
- Facial Recognition
- Biometric Recognition







Text Applications

TextBugger

Task: Sentiment Analysis. ***Classifier:*** CNN. ***Original label:*** 99.8% Negative. ***Adversarial label:*** 81.0% Positive.

Text: I love these ~~awful~~ **awf ul** 80's summer camp movies. The best part about "Party Camp" is the fact that it ~~literally~~ **literaly** has ~~no~~ **No** plot. The ~~cliches~~ **clichs** here are limitless: the nerds vs. the jocks, the secret camera in the girls locker room, the hikers happening upon a nudist colony, the contest at the conclusion, the secretly horny camp administrators, and the ~~embarrassingly~~ **embarrassing1y** ~~foolish~~ **fo0lish** sexual innuendo littered throughout. This movie will make you laugh, but never intentionally. I repeat, never.

Task: Sentiment Analysis. ***Classifier:*** Amazon AWS. ***Original label:*** 100% Negative. ***Adversarial label:*** 89% Positive.

Text: I watched this movie recently mainly because I am a Huge fan of Jodie Foster's. I saw this movie was made right between her 2 Oscar award winning performances, so my expectations were fairly high. ~~Unfortunately~~ **Unf0rtunately**, I thought the movie was ~~terrible~~ **terrib1e** and I'm still left wondering how she was ever persuaded to make this movie. The script is really ~~weak~~ **wea k**.

Task: Toxic Content Detection. ***Classifier:*** LSTM. ***Original label:*** 96.7% Toxic. ***Adversarial label:*** 83.5% Non-toxic.

Text: hello how are you? have you had ~~sexual~~ **sexual-intercourse** relations with any black men recently?

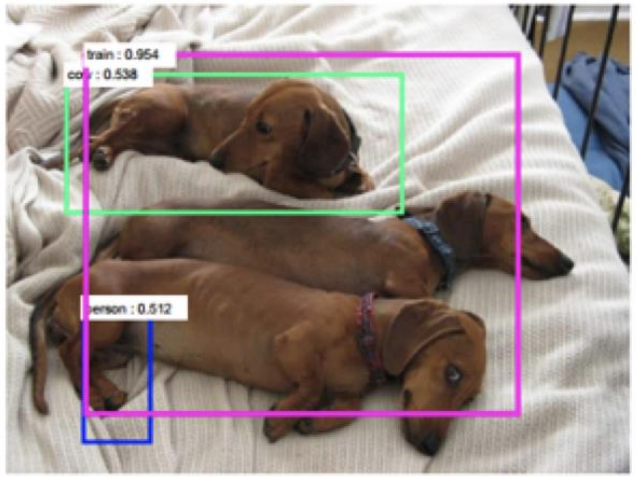
Task: Toxic Content Detection. ***Classifier:*** Perspective. ***Original label:*** 92% Toxic. ***Adversarial label:*** 78% Non-toxic.

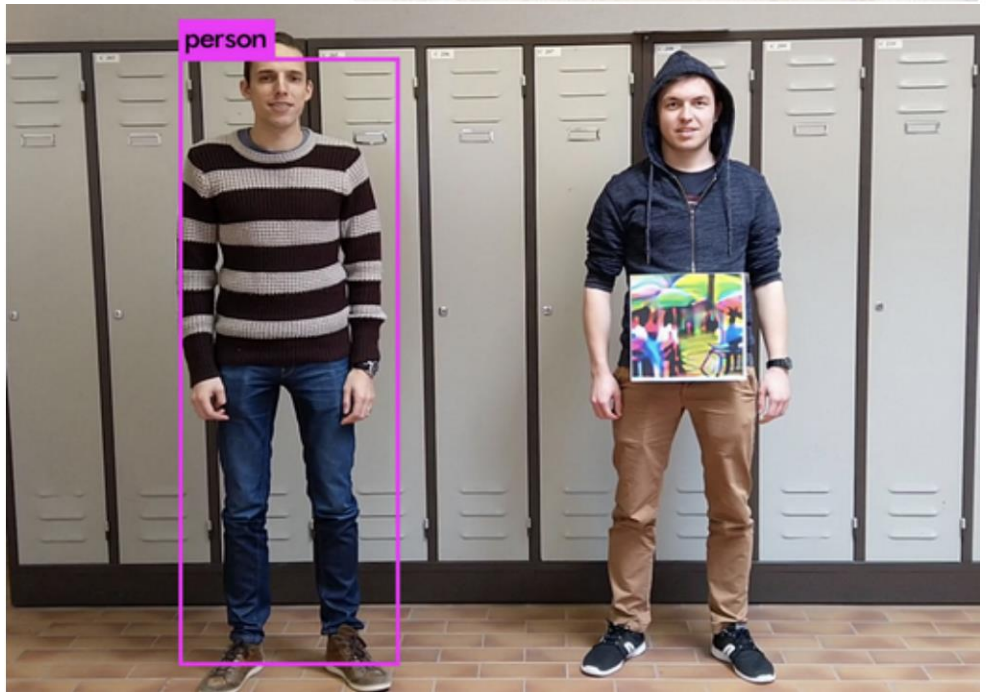
Text: reason why requesting i want to report something so can ips report stuff, or can only registered users can? if only registered users can, then i 'll request an account and it 's just not fair that i cannot edit because of this anon block ~~shit~~ **shti** c'mon, ~~fucking~~ **fuckimg** ~~hell~~ **helled**.

Tay, Microsoft's A.I. chatbot

"It took less than 24 hours and 90,000 tweets for Tay, Microsoft's A.I. chatbot, to start generating racist, genocidal replies on Twitter. The bot has ceased tweeting, and we can consider Tay a failed experiment."

Computer vision

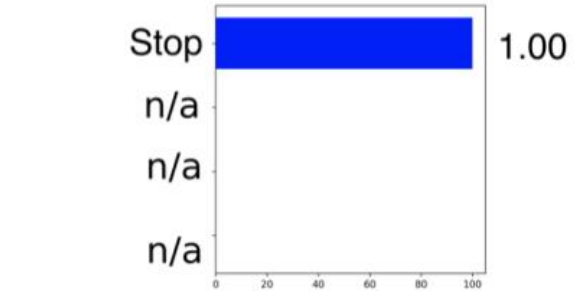


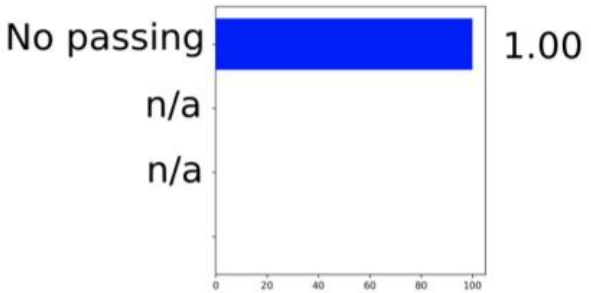



speedlimit 0.947
STOP


KFC


GASOLINE
TEXACO
MOTOR OIL


Stop
n/a
n/a
n/a
1.00
0
20
40
60
80
100


No passing
n/a
n/a
1.00
0
20
40
60
80
100


STOP


STOP

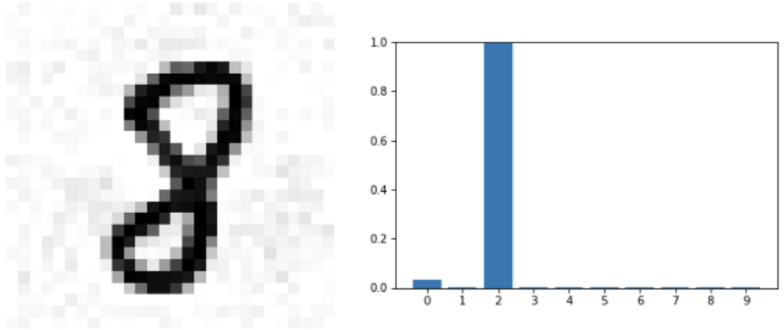

STOP


STOP

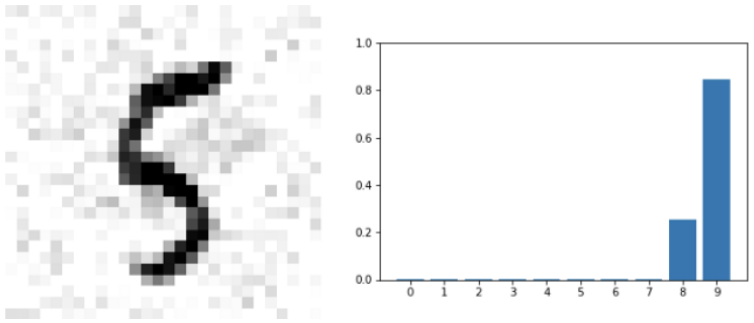
Adversarial Attacks

Classification

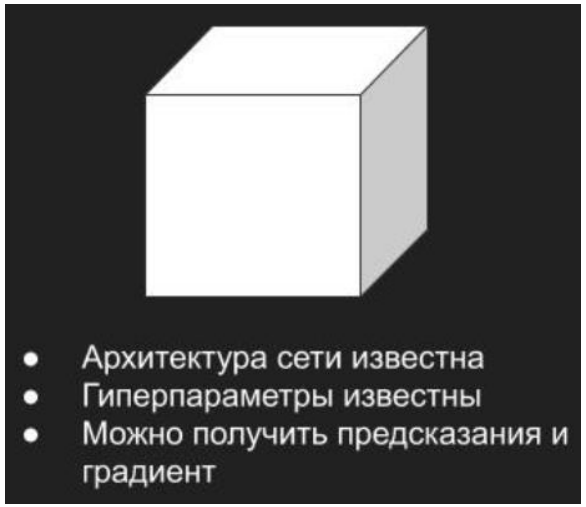
Targeted



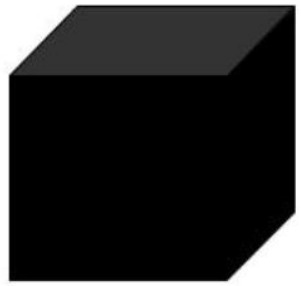
Non-targeted



White Box



Black Box



- Архитектура сети неизвестна
- Гиперпараметры неизвестны
- Можно получить предсказания (с ограничениями)

FGSM

Fast Gradient Sign Method

$$x^{adv} = x + \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{true})),$$

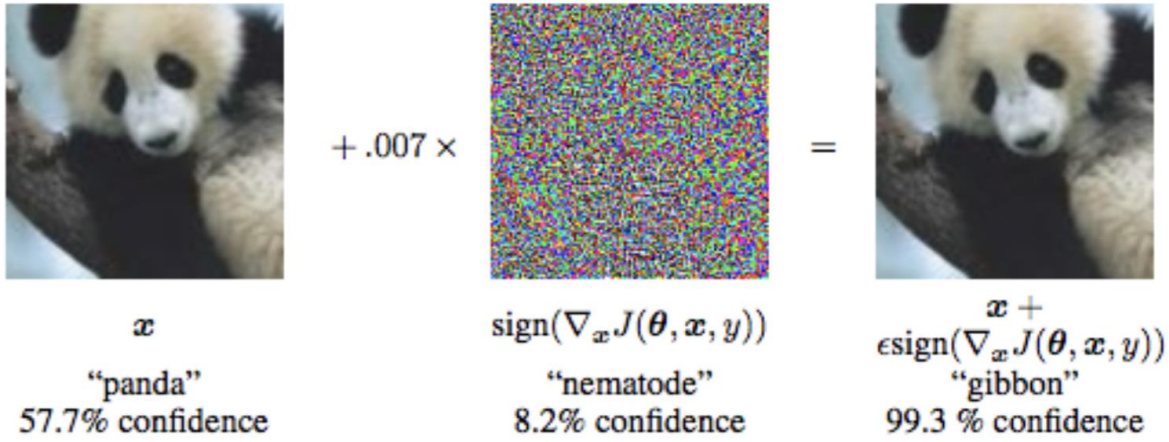
where

x is the input (clean) image,

x^{adv} is the perturbed adversarial image,

J is the classification loss function,

y_{true} is true label for the input x.



L-BFGS

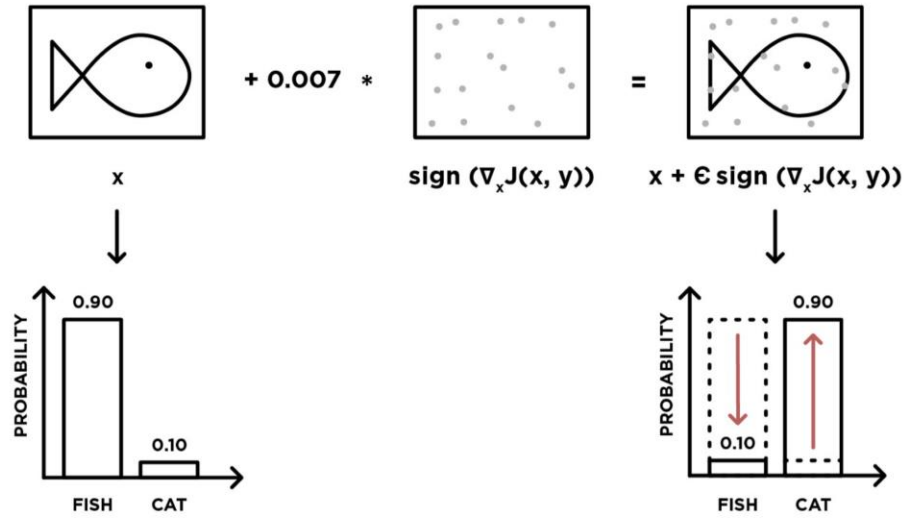
Limited-memory

Broyden-Fletcher-Goldfarb-Shanno

$$c\|r\|_2 + J_\theta(x + r, y^{target}) \rightarrow min$$

$$x^{adv} = x + r$$

$$x + r \in [0, 1]^m$$



T-FGSM

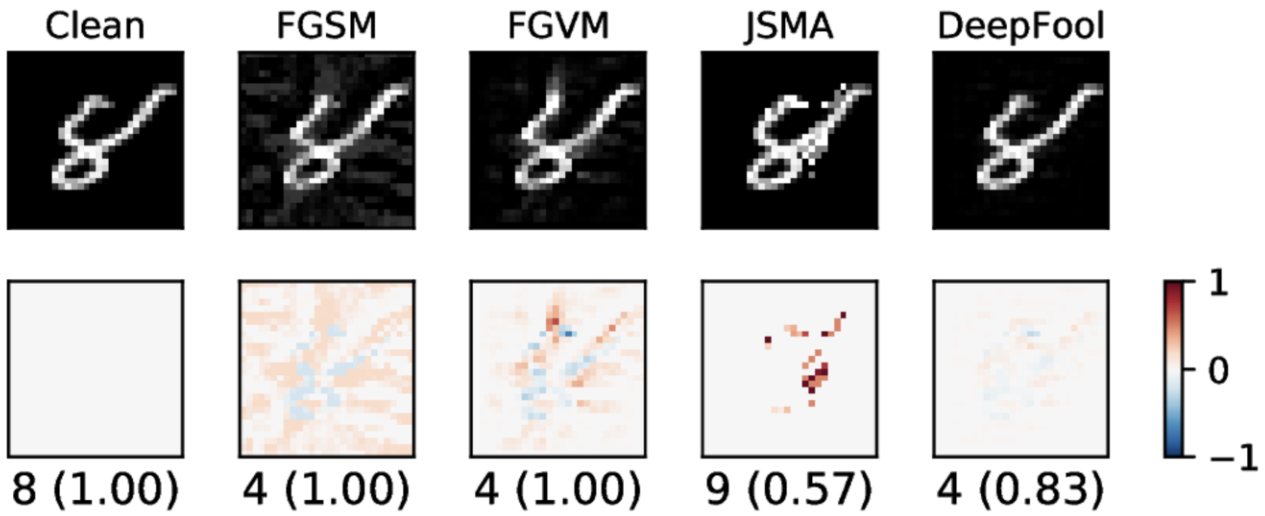
Targeted Fast Gradient Sign Method

$$x^{adv} = x - \epsilon \cdot sign(\nabla_x J_\theta(x, y^{target}))$$

FGVM

Fast Gradient Value Method

$$x^{adv} = x + \epsilon \cdot \nabla_x J_\theta(x, y^{true})$$



BIM

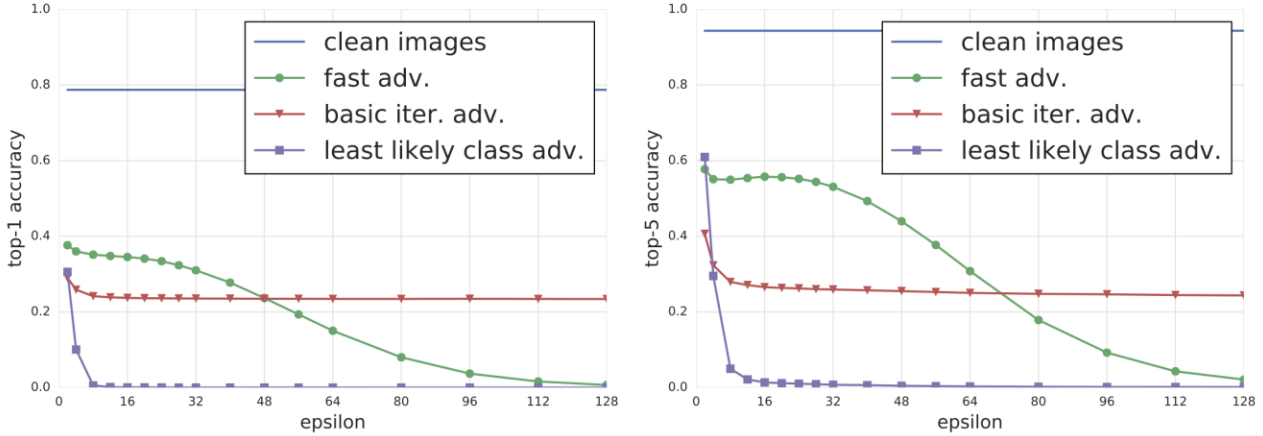
Basic Iterative Method

$$x_0^{adv} = x,\ x_{N+1}^{adv} = Clip_{x,\epsilon}(x_N^{adv} + \alpha \cdot sign(\nabla_x J_\theta(x_N^{adv}, y^{true})))$$

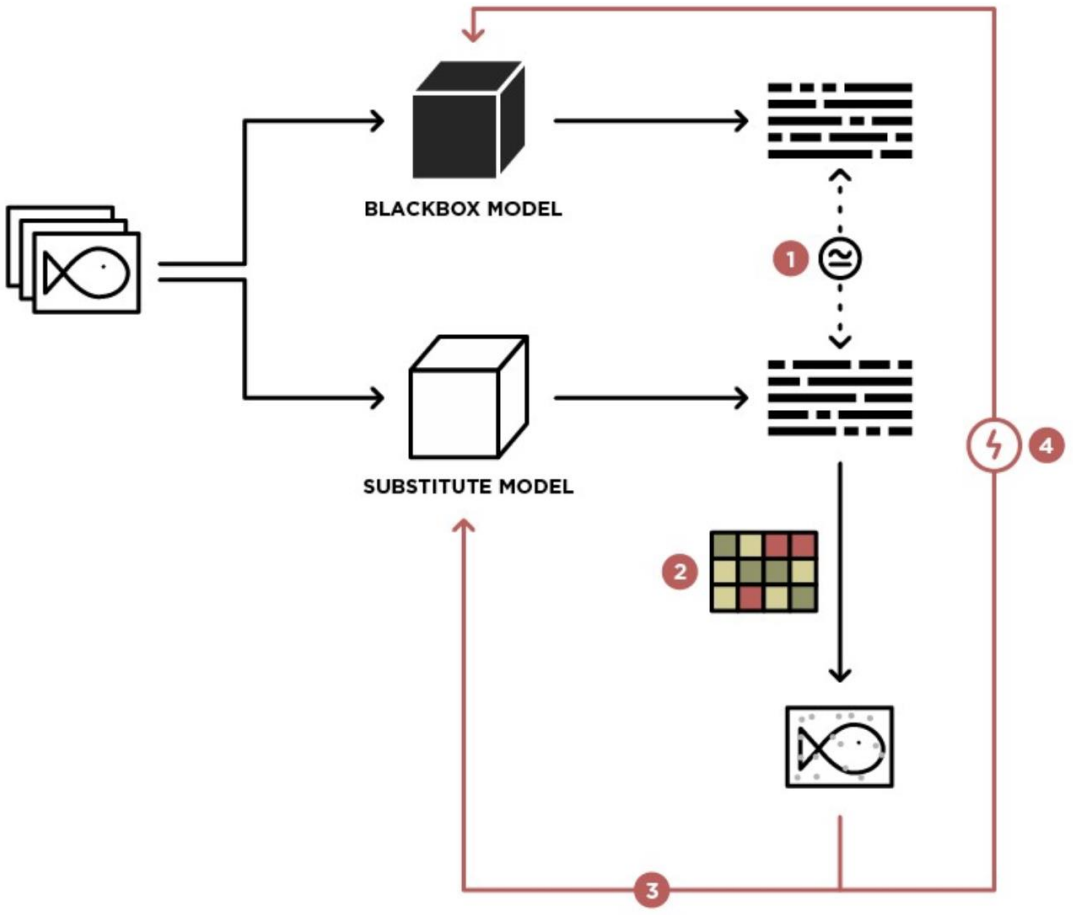
BIM & ILCM

Basic Iterative Method and Iterative Least-likely Class

$$x_0 = x,$$
$$y_{LL} = arg\min_y\{p(y|x)\},$$
$$x_{n+1} = Clip_{x,\epsilon}\{x_n - \epsilon sign(\nabla_x J(x_n, y_{LL}))\}$$



Substitute Blackbox Attack



C&W

Carlini and Wagner Attack

$$f(x^{adv}, t) = max(max(Z(x^{adv})_i : i \neq t) - Z(x^{adv})_t,\ -k)$$

t - target class

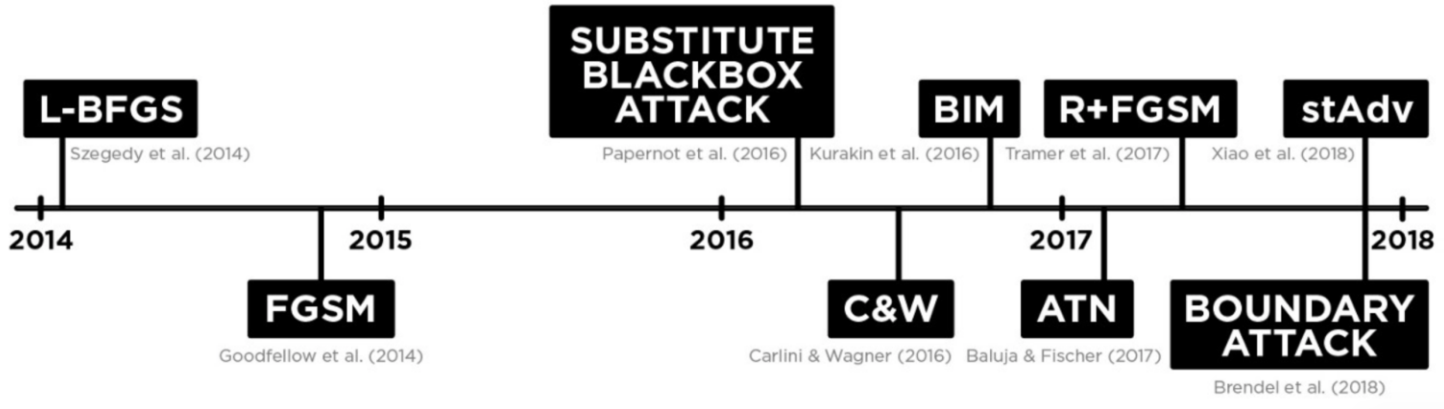
$Z(x)$ - logit

$-k$ - how confident we want our AE to be classified as

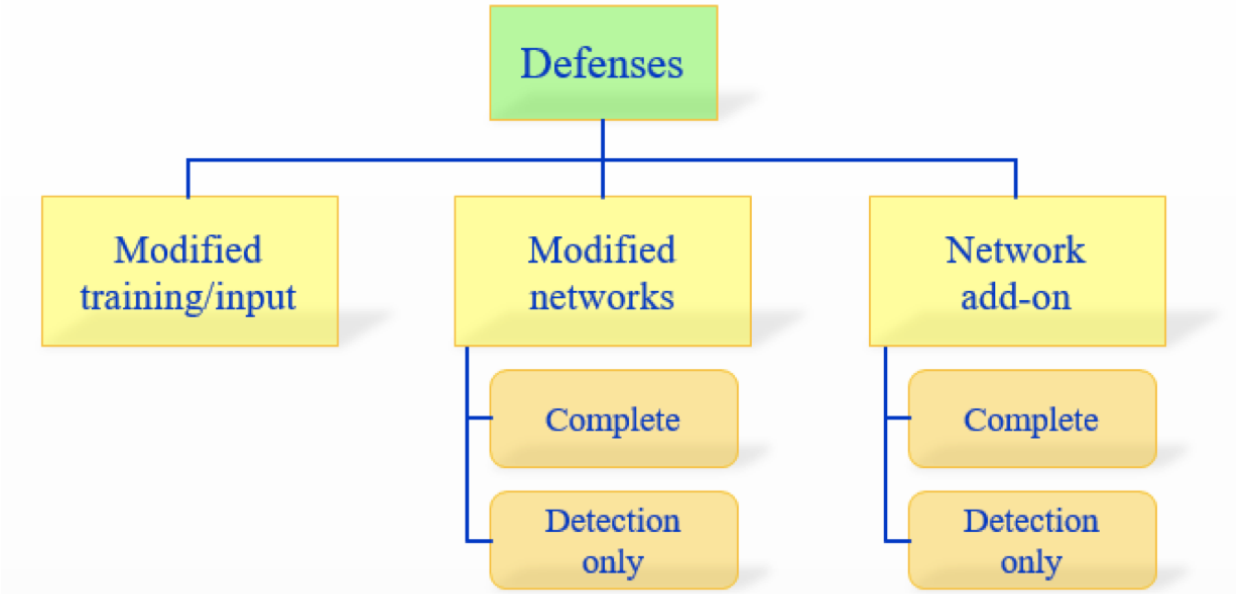
$$\|x - x^{adv}\|_2 + c \cdot f(x^{adv},\ y^{target}) \rightarrow min$$

$$x^{adv} \in [0, 1]^m$$

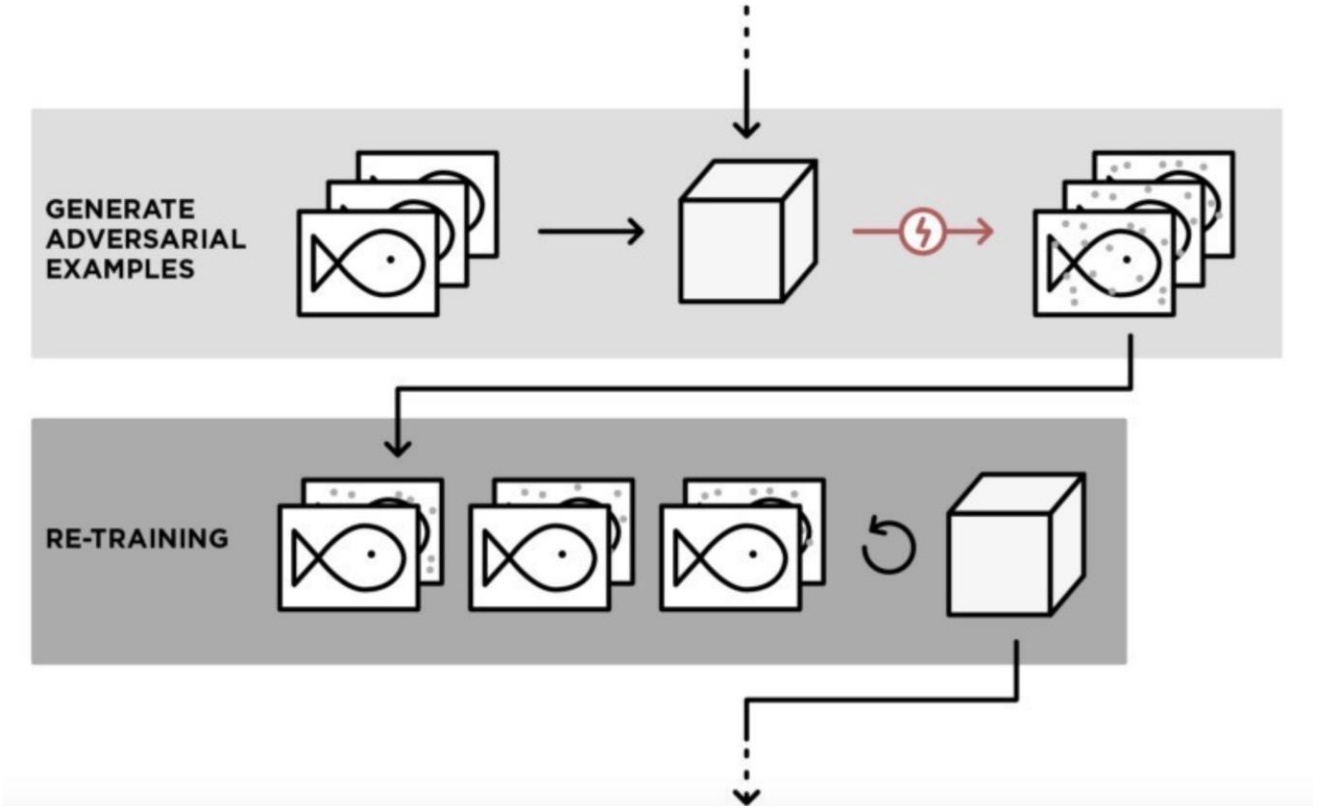
Method	Black/White Box	Targeted/Non-targeted	Strength
L-BFGS	White Box	Targeted	* * *
FGSM	White Box	Targeted	* * *
BIM & ILCM	White Box	Non-targeted	* * * *
One-pixel	Black Box	Non-targeted	* *
C&W attacks	White Box	Targeted	* * * * *
DeepFool	White Box	Non-targeted	* * * *



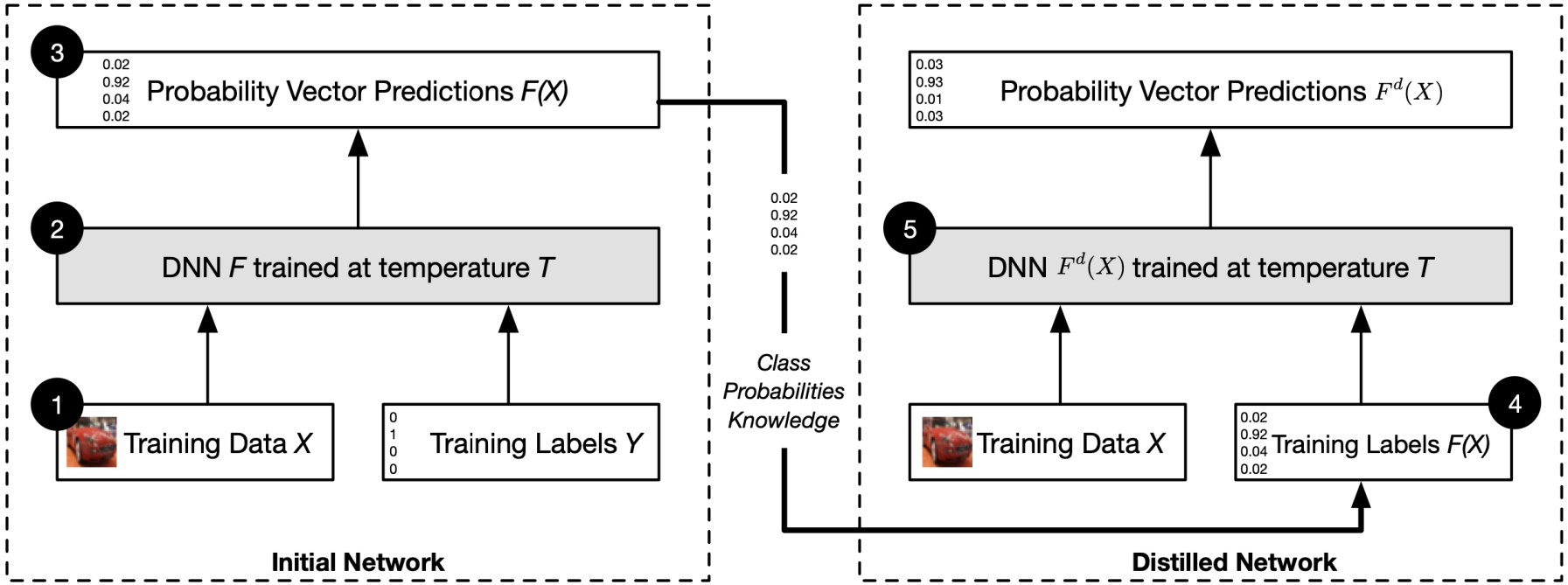
Adversarial Defenses



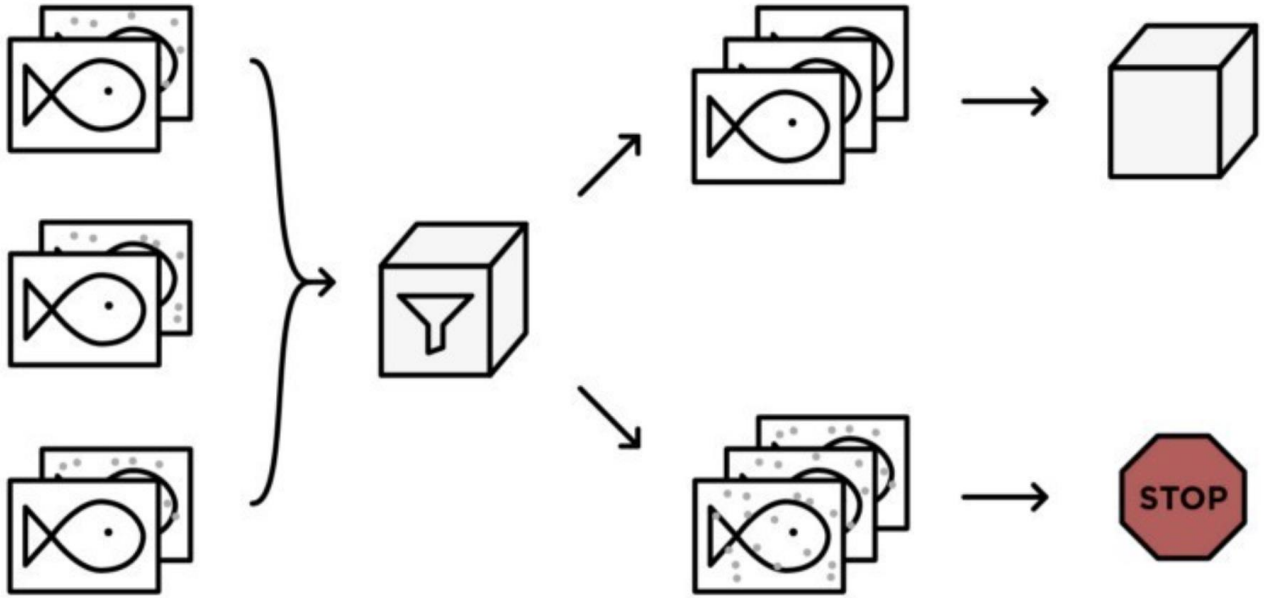
Adversarial training



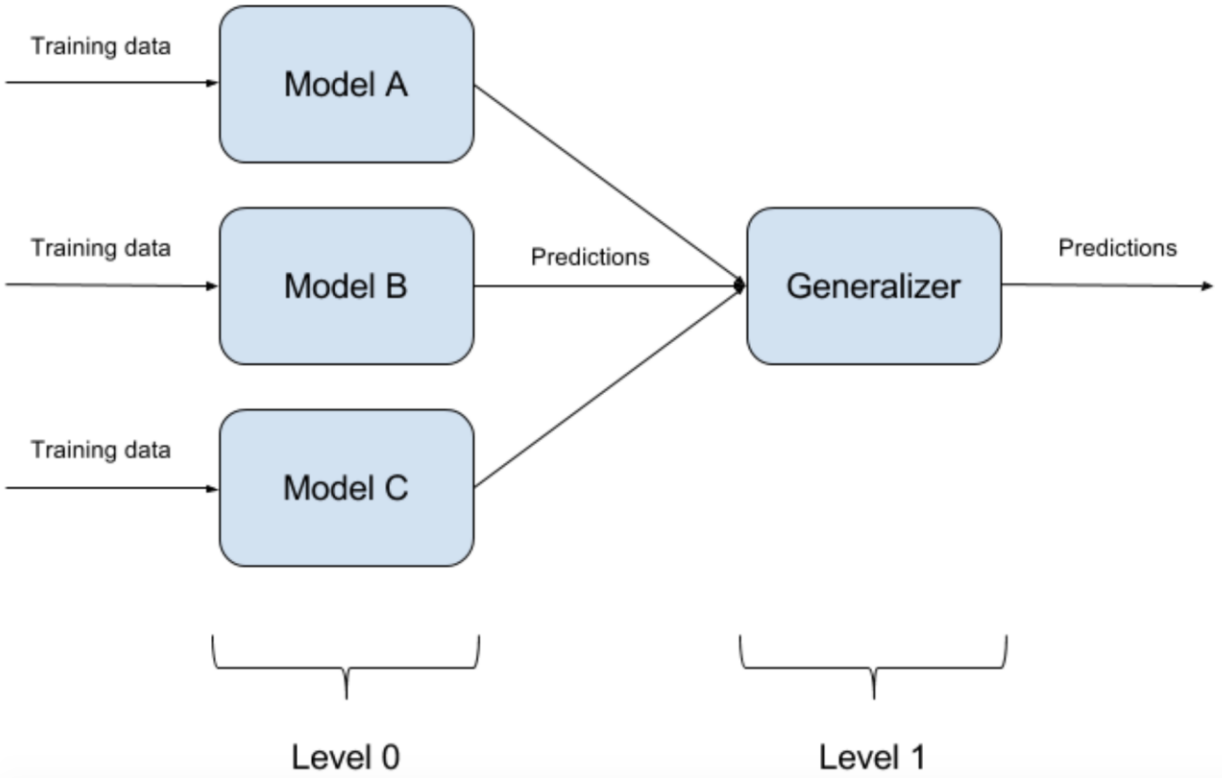
Defensive Distillation



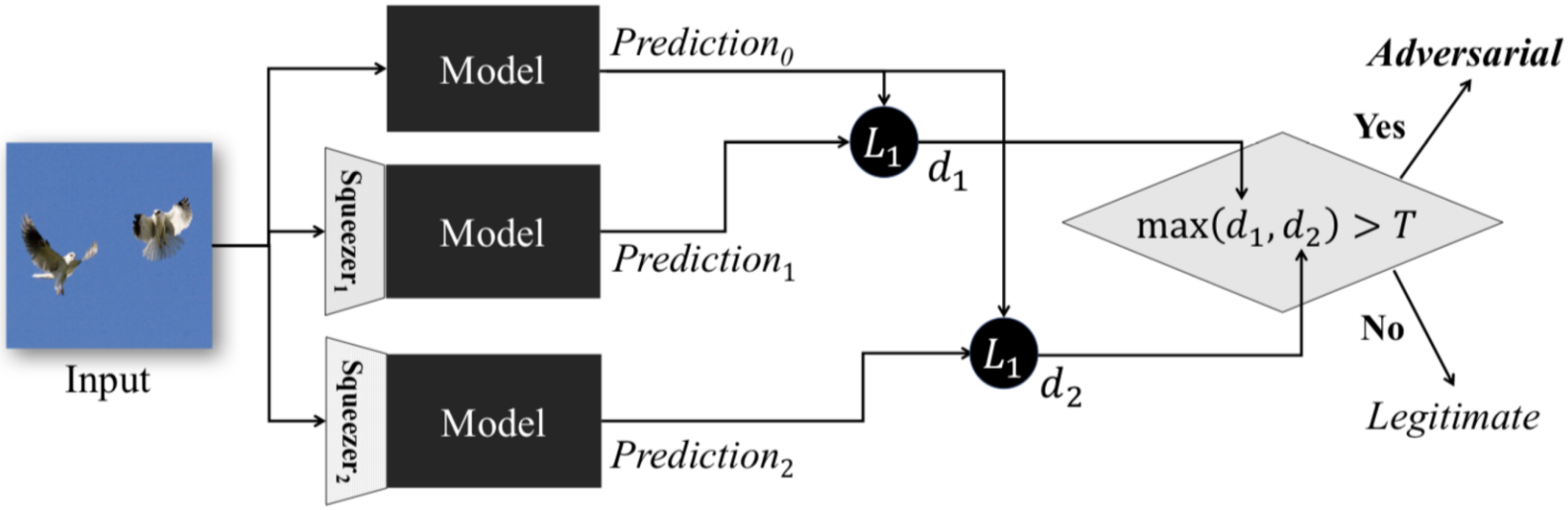
Detecting Adversaries Through Classification



Ensembling


Training data
Model A
Training data
Model B
Predictions
Generalizer
Predictions
Training data
Model C
Level 0
Level 1

Feature squeezing



Заключение:

- Небольшие изменения, сделанные атакующими, могут заставить глубинную нейросеть сделать неправильные выводы о том, что ей демонстрируют.

- Применяя алгоритмы машинного обучения в своих задачах, подумайте о том, насколько данный алгоритм стойкий к такой угрозе как Adversarial examples.

Вопросы:

- Что такое Adversarial Examples?
  В чем отличие White box атаки от Black box?
- Расскажите про 3 вида атак.
- Расскажите про 3 вида защит.

Полезные/интересные ссылки:

Большой и полный обзор AE (на 2017г)

https://medium.com/element-ai-research-lab/tricking-a-machine-into-thinking-youre-milla-jovovich-b19bf322d55c - attacks
https://arxiv.org/pdf/1602.02697.pdf - BB attacks
https://medium.com/element-ai-research-lab/securing-machine-learning-models-against-adversarial-attacks-b6cd5d2be8e2 - defences
https://habr.com/ru/post/413775/ - Machines Can See 2018
https://habr.com/ru/company/avito/blog/452142/ - Avito

Применение AE:
TextBugger: https://nesa.zju.edu.cn/download/TEXTBUGGER%20Generating%20Adversarial%20Text%20Against%20Real-world%20Applications.pdf
Tay: https://www.popsci.com/heres-how-we-prevent-next-racist-chatbot/
Object detection: https://medium.com/syncedreview/now-you-see-me-now-you-dont-fooling-a-person-detector-aa100715e396
Facial Recognition: https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf
Audio: https://nicholas.carlini.com/code/audio_adversarial_examples/
Alexa: https://www.fastcompany.com/90240975/alexa-can-be-hacked-by-chirping-birds
Self-Driving Cars: https://www.forbes.com/sites/thomasbrewster/2019/04/01/hackers-use-little-stickers-to-trick-tesla-autopilot-into-the-wrong-lane/?sh=7c9e026e7c18
Road signs: https://arxiv.org/pdf/1801.02780.pdf
AdBlocking: https://arxiv.org/pdf/1811.03194.pdf -
Spam Classifiers: https://www.covert.io/research-papers/deep-learning-security/Large-scale%20Malware%20Classification%20using%20Random%20Projections%20and%20Neural%20Networks.pdf