Математика нейронных сетей

Иван Оселедец,
Сколковский институт
науки и технологий

Skoltech
Skolkovo Institute of Science and Technology

Задача машинного обучения

U — Множество объектов

X(U) – Множество признаков

Y – множество классов

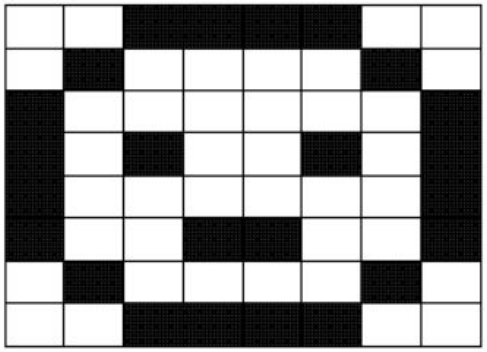
Нужно по X определить класс

Skoltech
Skolkovo Institute of Science and Technology

Пример

U

$X(U)$



Y

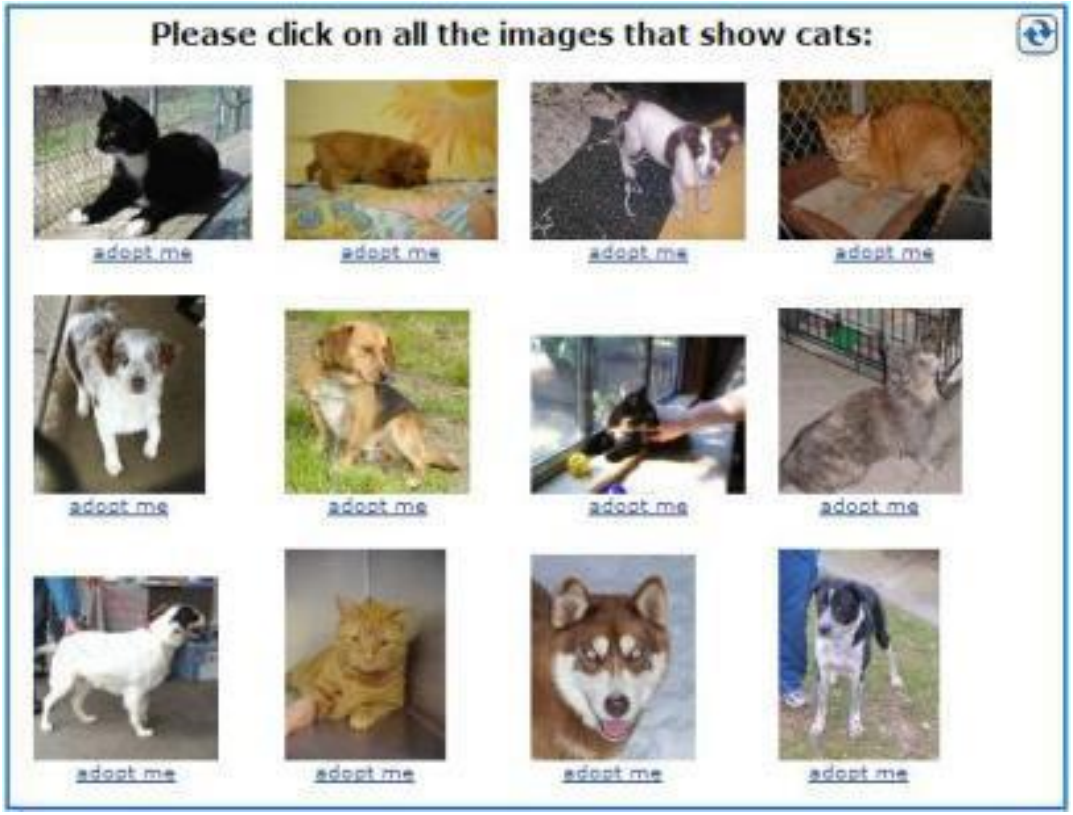
Человек

Птица

Дом

Skoltech
Skolkovo Institute of Science and Technology

2006 год: соревнование Asirra



2006 год: 60% точности, 0.6^12 = 0.00217

2014 год: соревнование Asirra

Completed • Swag

Dogs vs. Cats

Wed 25 Sep 2013 – Sat 1 Feb 2014 (2 years ago)

Dashboard

Private Leaderboard - Dogs vs. Cats

This competition has completed. This leaderboard reflects the final standings.

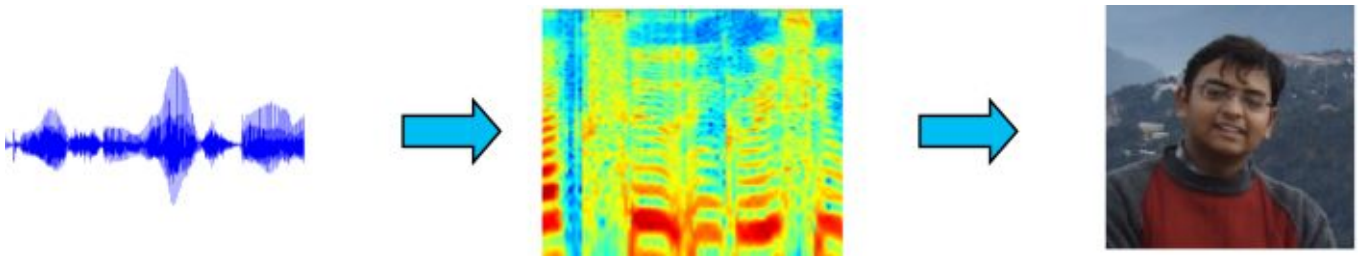
See someone using multiple accounts?
Let us know.

#	Δrank	Team Name • in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	Pierre Sermanet *	0.98914	5	Sat, 01 Feb 2014 21:43:19 (-1.2h)
2	↑4	orchid *	0.98309	17	Sat, 01 Feb 2014 23:52:30
3	–	Owen	0.98171	15	Sat, 01 Feb 2014 17:04:40 (-1.3h)
4	–	Paul Covington	0.98171	3	Sat, 01 Feb 2014 23:05:20
5	↓3	Maxim Milakov	0.98137	24	Sat, 01 Feb 2014 18:20:58
6	↓1	we've been in KAIST	0.98103	8	Sat, 01 Feb 2014 21:15:30

$$0.989^{12} = 0.875$$

Классический подход

Изображение Признаки Распознавание

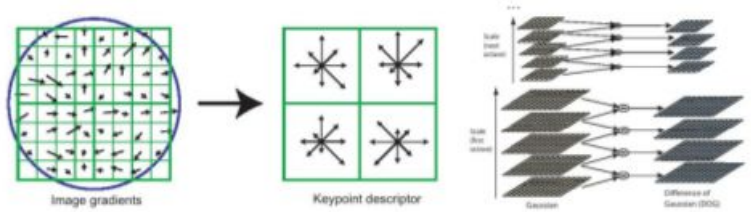


Звук Признаки Распознавание

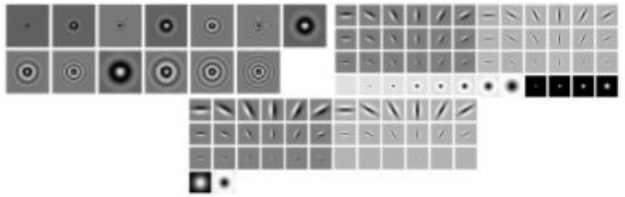
Skoltech
Skolkovo Institute of Science and Technology

Искусство построения признаков

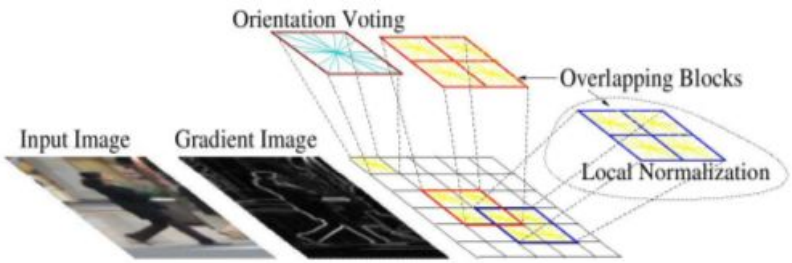
.



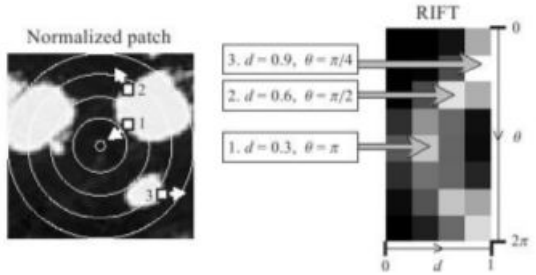
SIFT



Textons

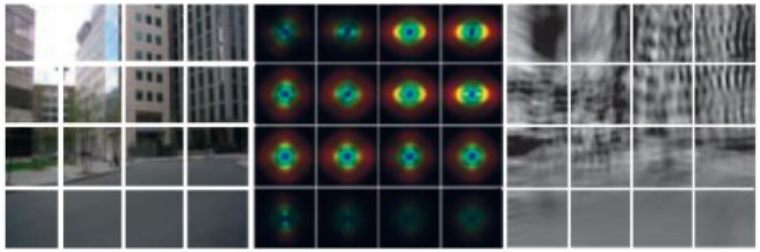


HoG



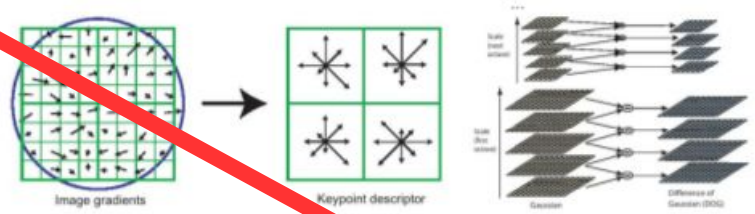
RIFT

GIST

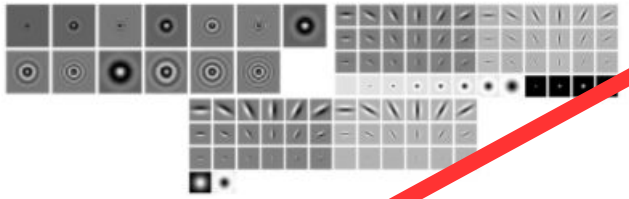


Skoltech
Skolkovo Institute of Science and Technology

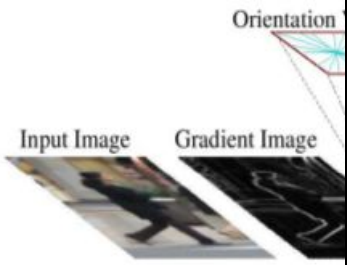
Искусство построения признаков

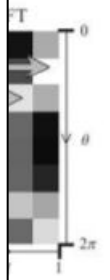


SIFT



Textons



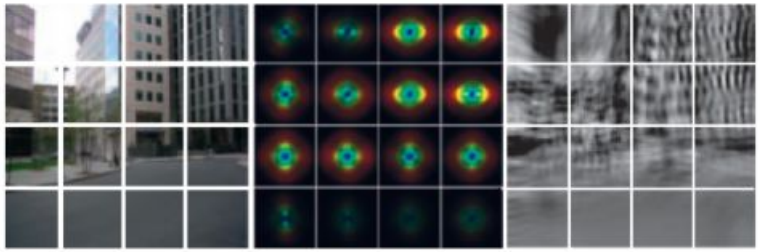


Глубинные нейронные сети:
Одна концепция

HoG

RIFT

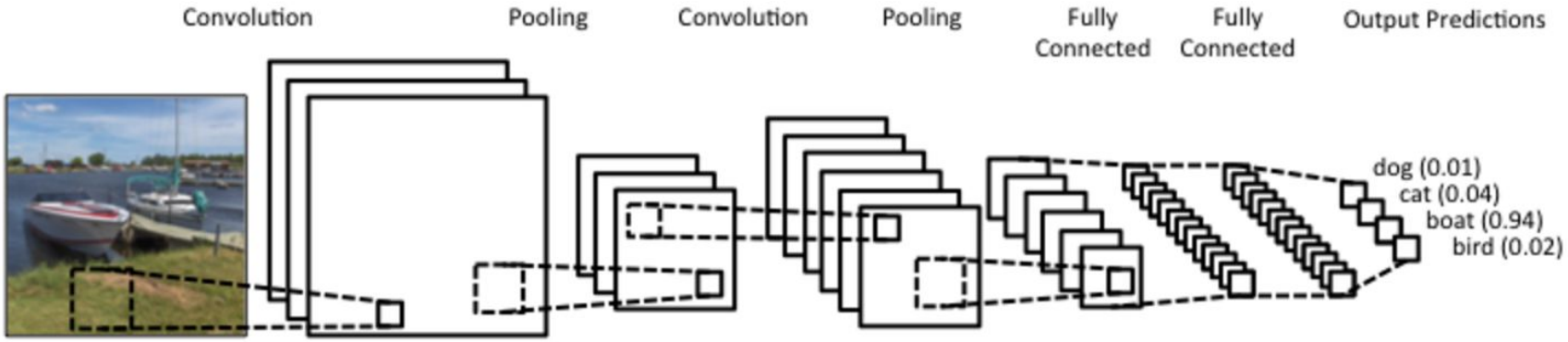
GIST



Skoltech
Skolkovo Institute of Science and Technology

Сверточная нейронная сеть

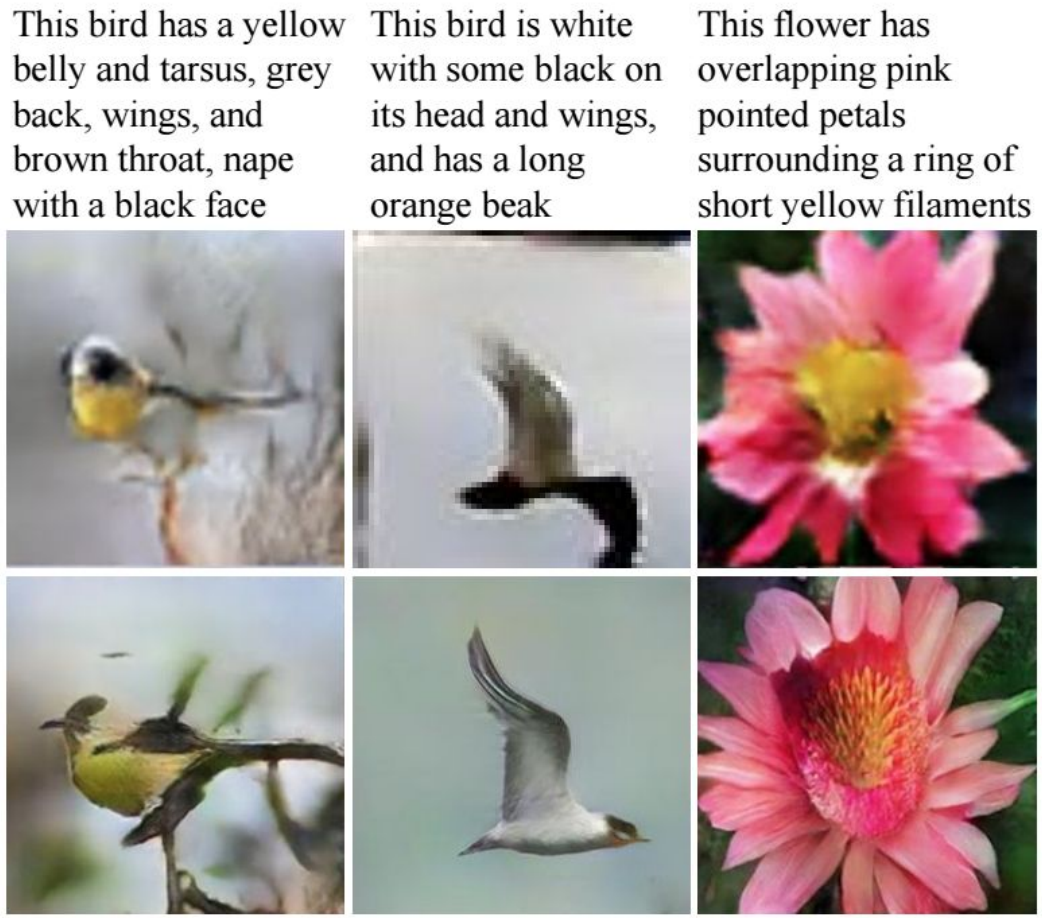
- Математически, нейронная сеть – комбинация простых преобразований (линейные/нелинейные)
- Глубина архитектуры
- Сверточные слои



Skoltech
Skolkovo Institute of Science and Technology

Пример работы нейронной сети

У этой птицы Желтый живот, серая спинка, крылья, коричневое горло.



Skoltech
Skolkovo Institute of Science and Technology

ImageNet (Princeton, Stanford, Google)

Более 10 миллионов аннотированных “вручную” изображений с помощью сервиса Amazon Mechanical Turk

Простота доступных технологий

Tensorflow - открытый фреймворк глубоких нейронных сетей, разрабатываемый Google (>15 инженеров, работающих на полную ставку)

> 700 тысяч строк кода

> 600 участников разработки

PyTorch - другой открытый фрейморк, разрабатывается при поддержке Facebook

PYTORCH

> 90 тысяч строк кода

http://pytorch.org/2018/01/19/a-year-in.html

> 400 участников разработки

Skoltech
Skolkovo Institute of Science and Technology

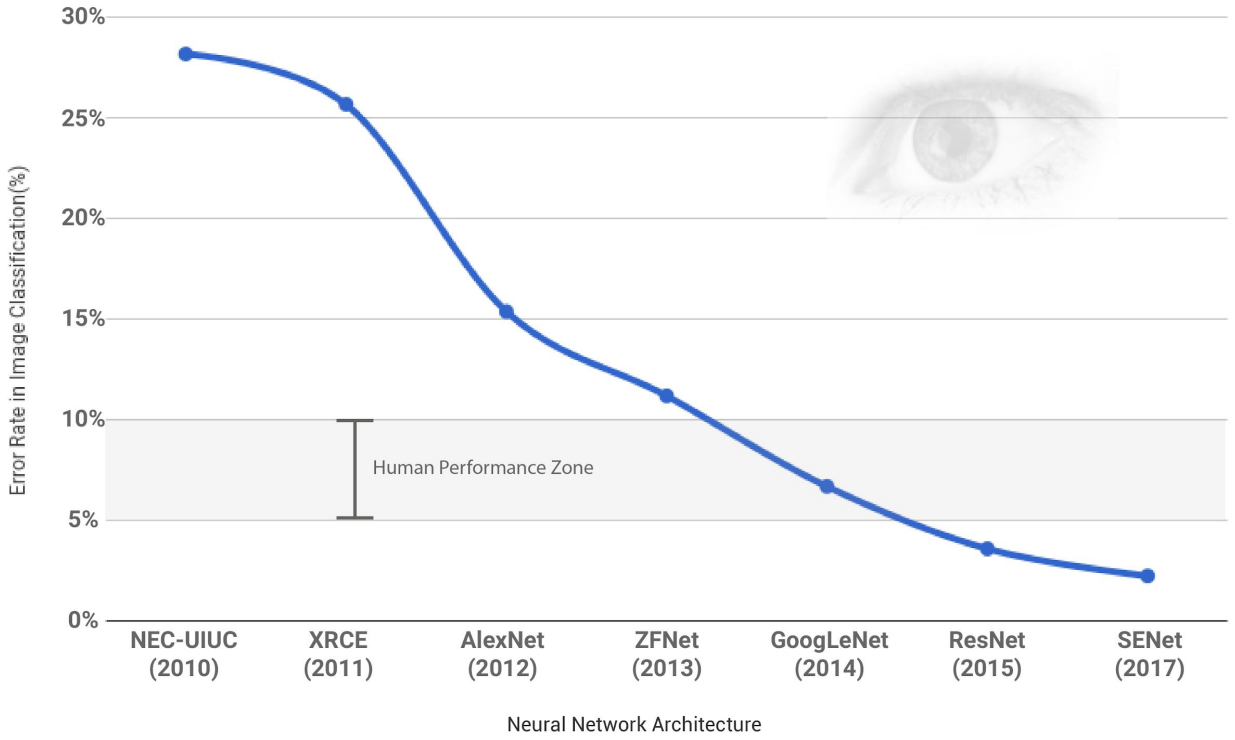
Причины прорыва

Причины прорыва в сочетании факторов!

- Рост мощности и доступности железа (GPU)
- Новые объемы открытых, размеченных данных
- Открытость исследований и исходных кодов фреймворков

Как результат, точечная «фокусировка» исследований в узкой области

Глубина нейронных сетей



Глубокие сети позволяют приближать функции экспоненциально лучше, чем “мелкие”

Открытые вопросы

- Почему нет переобучения?
- Почему более глубокие сети лучше?

Skoltech
Skolkovo Institute of Science and Technology

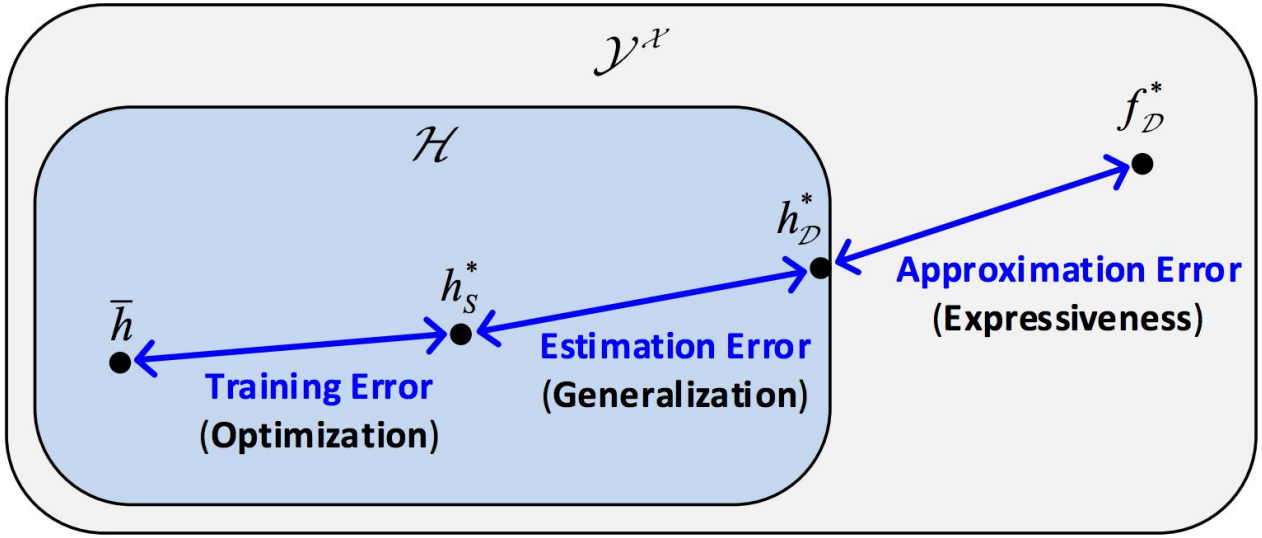
Статистическая теория обучения

Выразительность-Обобщение-Оптимизация

- Теория для классических моделей
- Нет для Deep Learning

Skoltech
Skolkovo Institute of Science and Technology

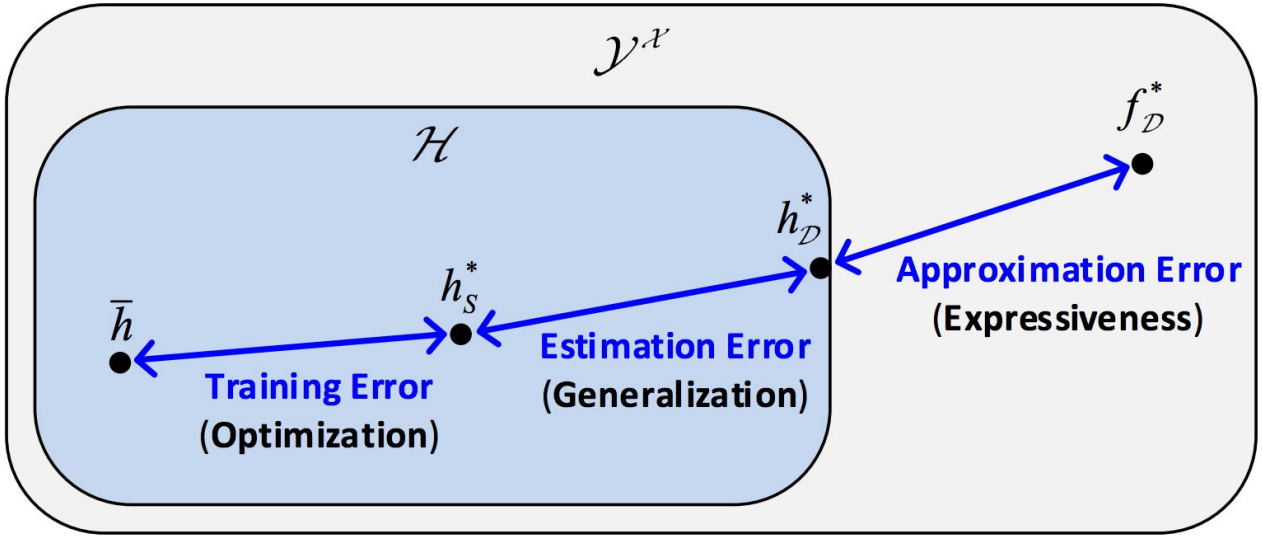
Точность/обобщение: классика



Число параметров модели	Аппроксимация	Оценка параметров
увеличивается	↘	↗
падает	↗	↘

Skoltech
Skolkovo Institute of Science and Technology

Точность/обобщение (сети)



Число параметров модели	Аппроксимация	Оценка параметров
увеличивается	↗	↗
падает		

Skoltech
Skolkovo Institute of Science and Technology

Rethinking generalization

Zhang et. al, ICLR 2017

Нейронная сеть может “запомнить” случайные метки

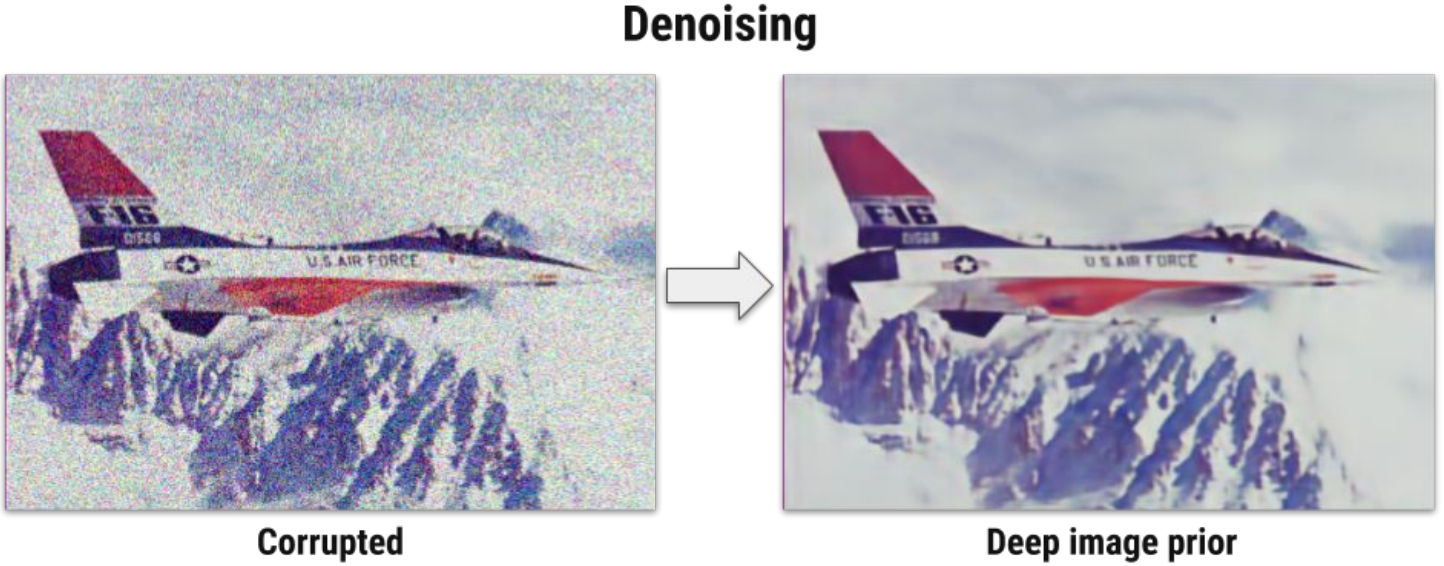
VC-размерность огромная!

Skoltech
Skolkovo Institute of Science and Technology

Deep Prior: обучение по 1 примеру

D. Ulyanov, V. Lempitsky, A. Vedaldi

https://dmitryulyanov.github.io/deep_image_prior



Deep Prior

Идея:

Считаем, что картинка сгенерирована нейронной сетью

$$\min_{\theta} E(f_\theta(z); x_0) .$$

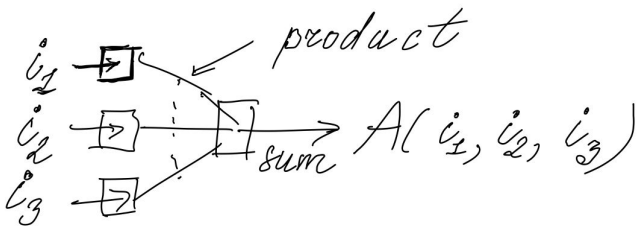
Skoltech
Skolkovo Institute of Science and Technology

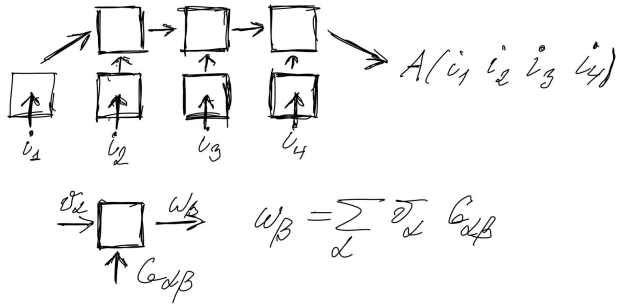
Тензорные разложения и нейронные сети

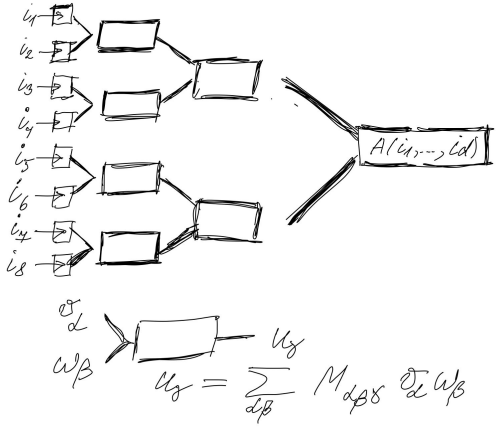
- Связь между тензорами и специальными нейронными сетями
- Мы можем исследовать связь с помощью тензорного анализа

Expressive power of recurrent neural networks
V **Khrulkov**, A Novikov, I **Oseledets** - ICLR 2018
On the expressive power of deep learning: A tensor analysis
N **Cohen**, O Sharir, A **Shashua** - Conference on Learning Theory, 2016 - jmlr.org

Тензорные разложения и нейронные сети







Обучение без учителя

- Unsupervised learning становится ключевой технологией

Учим вероятностные распределения

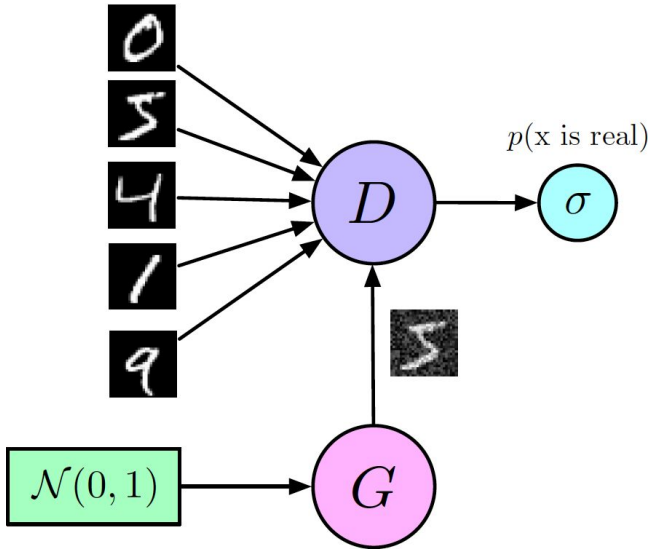
- Есть данные $x_1, \ldots, x_N$
- Generative adversarial network (GAN) стали основным подходом

Generative adversarial networks

- Предложены by Goodfellow et. al in 2014
- Идея состоит в том чтобы параметризовать распределение как образ заданного
- Дополнительная функция (дискриминатор) отличает правду от фейка

Иллюстрация

- Теория игр
- **Генератор** учится генерить реалистичные **изображения**
- **Дискриминатор** учится отличать реальные от синтетических



Как оценить качество

Оценка качества работы GAN непростая: нет правдоподобия

Что лучше?

Существующие подходы

- Люди
- Inception score
- Frechet Inception Distance (FID)

$$I(\{x_n\}_{n=1}^{N}) \triangleq \exp \mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{KL}}(p(y|\mathbf{x})||p(y)))$$

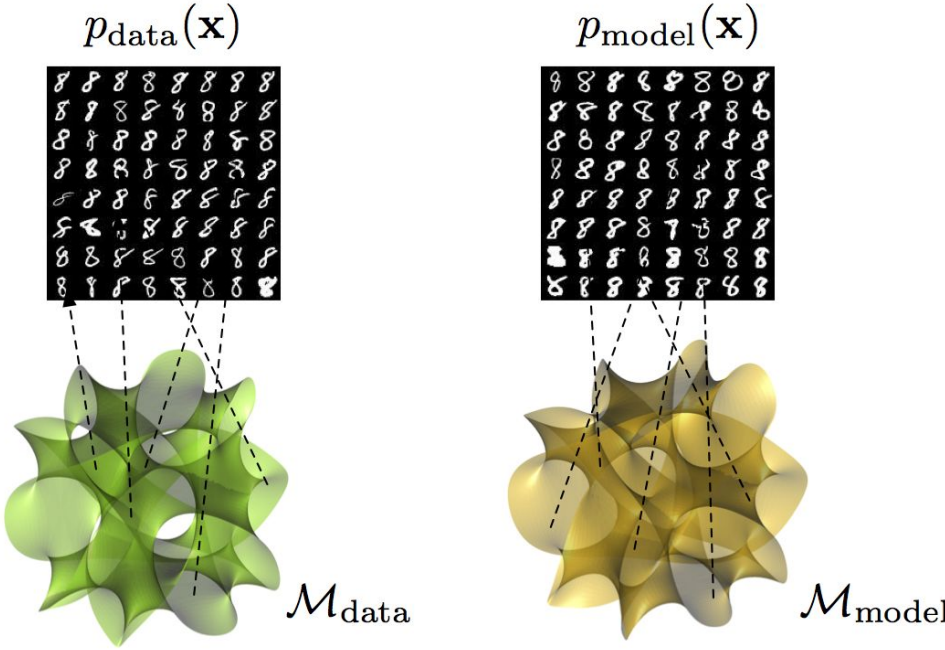
- Сложно использовать вне **ImageNet**
- Необязательно коррелирует с человеческой оценкой

“A Note on the Inception Score” Shane Barratt and Rishi Sharma

“Improved techniques for training GANs” Salimans et al.

Гипотеза о многообразии

Многомерные данные (такие как картинки) живут на **многообразиях небольшой размерности**

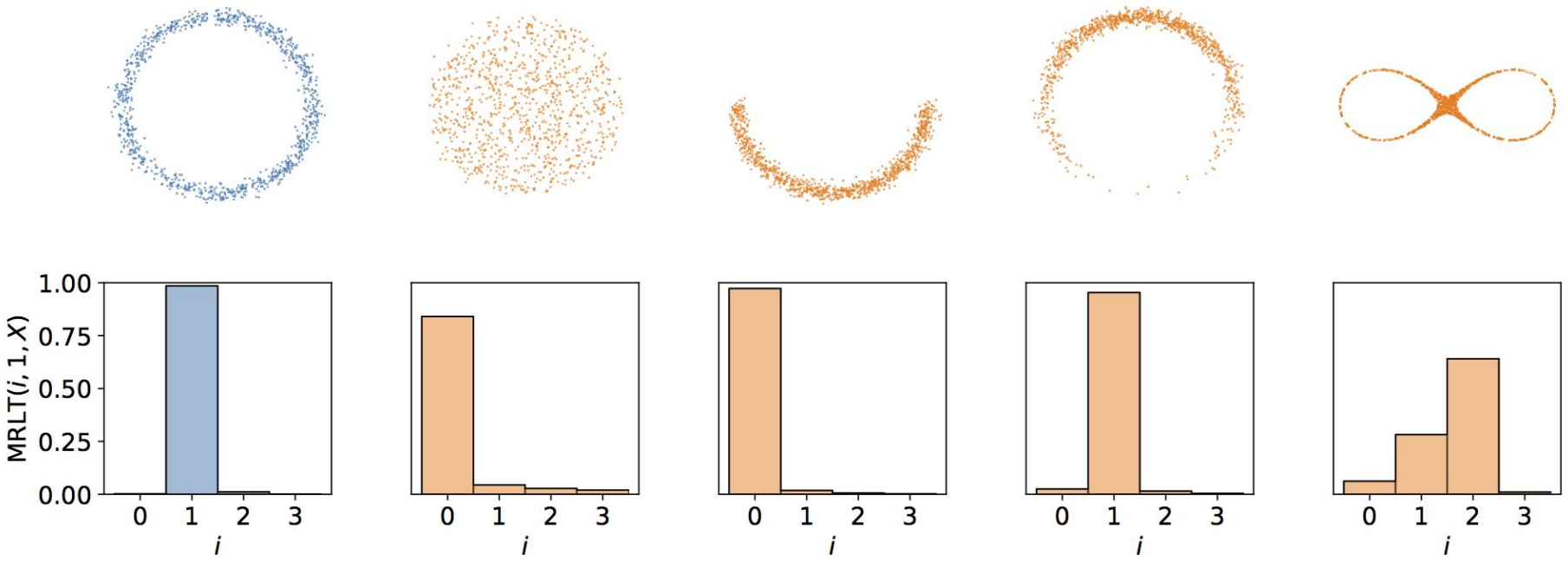


Похожи?

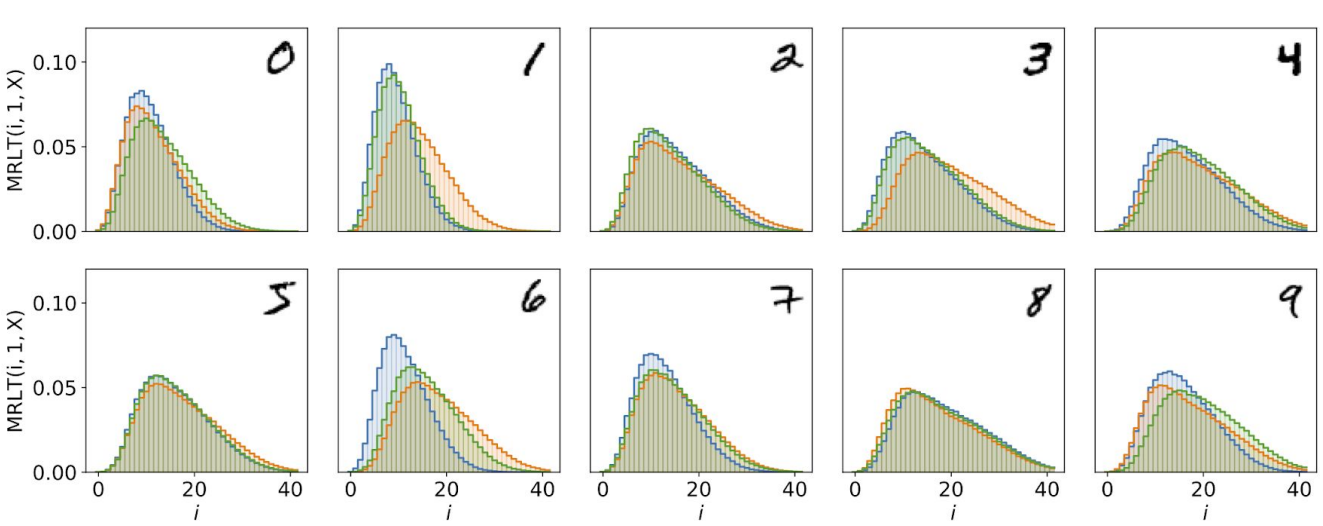
Наша идея

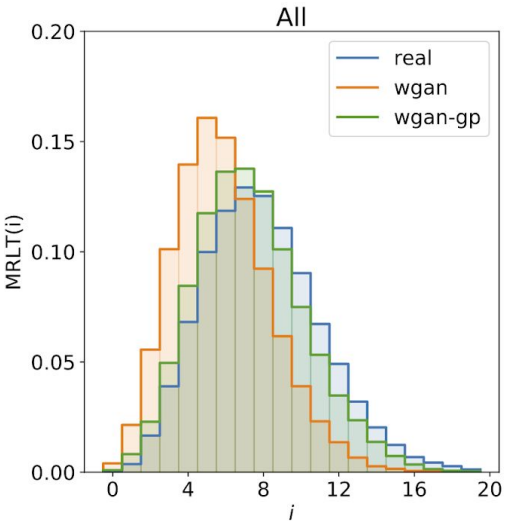
Мы используем **топологические свойства** многообразий

Игрушечные данные



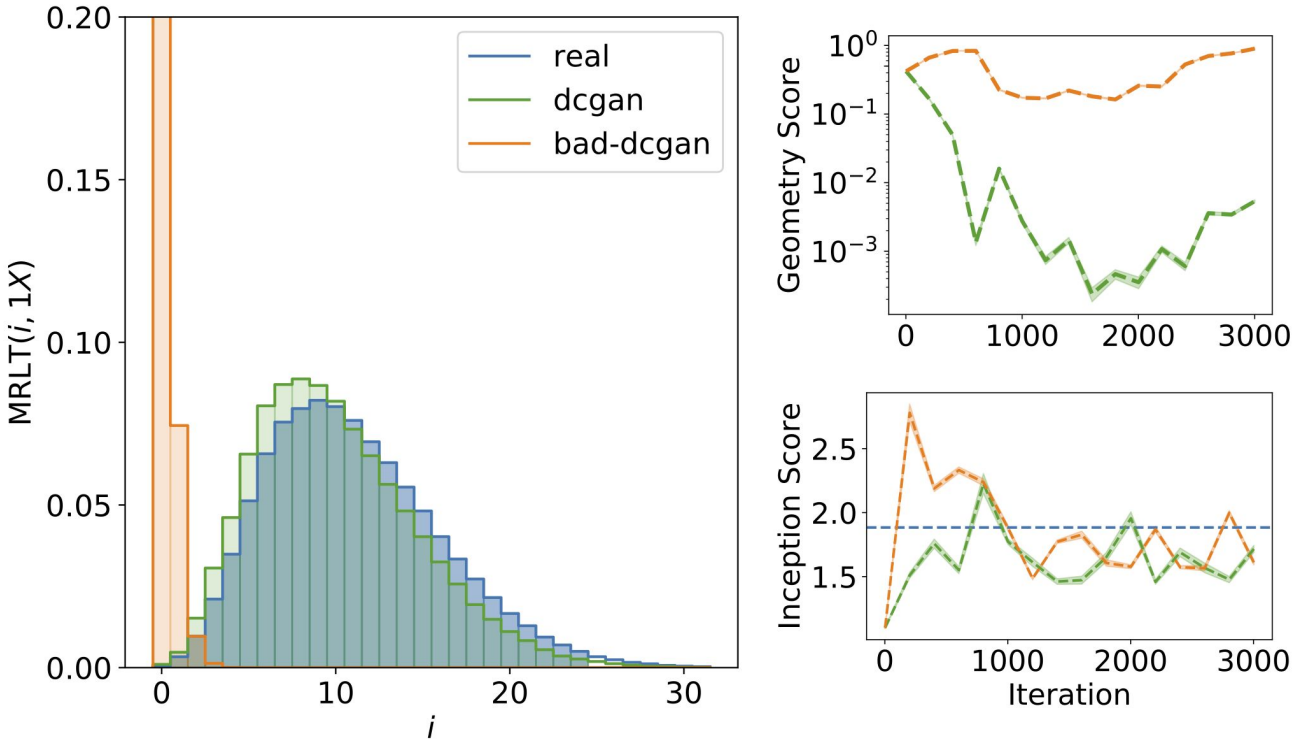
MNIST





Label	0	1	2	3	4	5	6	7	8	9	all
WGAN	**0.85**	21.4	0.60	7.04	**1.52**	0.47	22.8	2.20	0.76	**1.27**	26.1
WGAN-GP	5.19	**1.44**	**0.54**	**0.27**	2.16	**0.03**	**13.5**	**1.38**	**0.14**	5.00	**2.04**

Лица (CelebA)


real
dcgan
bad-dcgan
MRLT(i, 1X)
i
0.20
0.15
0.10
0.05
0.00
0
10
20
30
Geometry Score
10^0
10^{-1}
10^{-2}
10^{-3}
0
1000
2000
3000
Inception Score
2.5
2.0
1.5
0
1000
2000
3000
Iteration

Злонамеренные возмущения (adversarial)

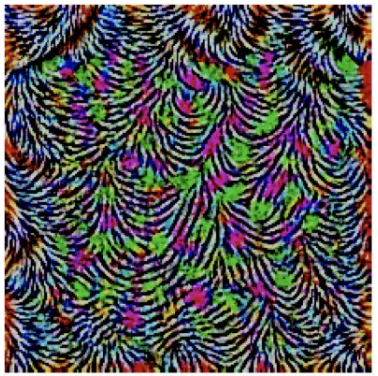
- **Злонамеренные возмущения** легко обманывают нейронные сети
- Добавка небольшой нормы приводит к неправильной классификации

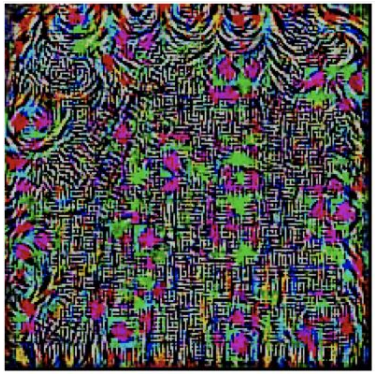
$$\arg\max p(y|x;\theta) \neq \arg\max p(y|x+\varepsilon;\theta)$$

This work is supported by Ministry of Education and Science of the Russian Federation (grant 14.756.31.0001)

Универсальные возмущения

- Mosaavi et al (2017) предложил **универсальные возмущения**: добавляем ровно одну картинку ко всем, обманываем сеть в (~70%) случаем
- Хорошо обобщаются **между сетями**



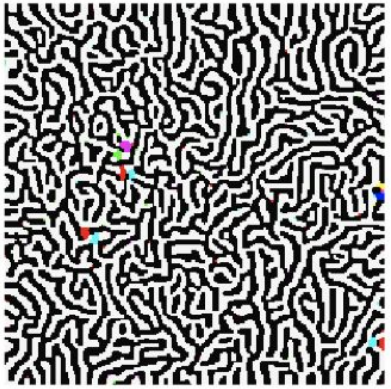


Universal adversarial perturbations, Moosavi et al, CVPR 2017

Наши результаты (Khrulkov, Oseledets, CVPR 2018)

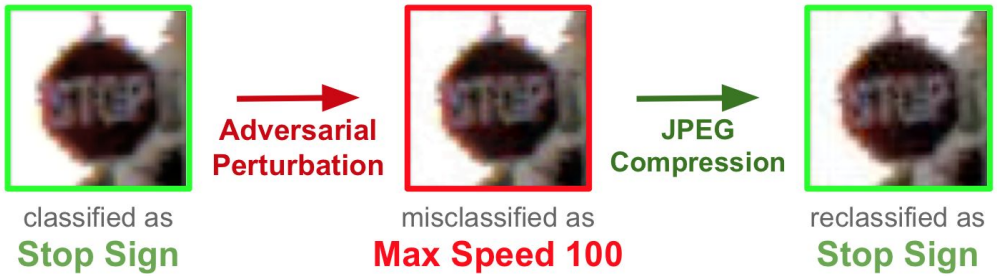
- Простой алгоритм
- Быстрый
- Мы атакуем **низкоуровневые признаки**

	VGG-16	VGG-19	ResNet50
VGG-16	0.52	**0.60**	0.39
VGG-19	0.48	**0.60**	0.38
ResNet50	0.41	**0.47**	0.44



Было предложено несколько методов защиты, но универсального нет

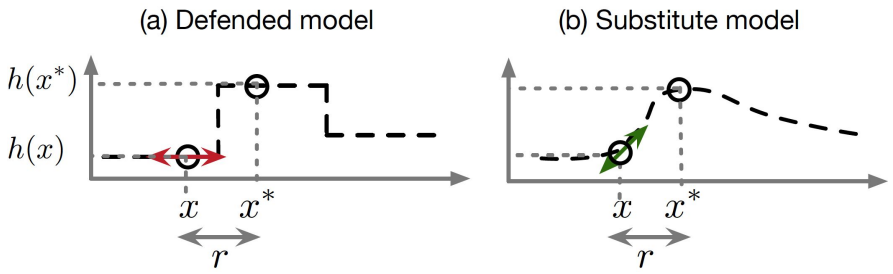
1. Определять и удалять шум с изображений
2. Во время обучения добавлять случайный шум
3. Контролировать свойства сети в процессе обучения



Соревнования по поиску лучших методов атаки и защиты:

NIPS 2017: Targeted Adversarial Attack

NIPS 2017: Defense Against Adversarial Attack



Skoltech
Skolkovo Institute of Science and Technology

Итоги

Мы не знаем:

почему глубина улучшает.

как они работают (интерпретируемость)

их легко обмануть (устойчивость)

Архитектуры сложные, тяжело учить

Skoltech
Skolkovo Institute of Science and Technology