ВСЕРОССИЙСКИЙ ФЕСТИВАЛЬ НАУКИ

NAUKA 0+

12 октября 2019 г.

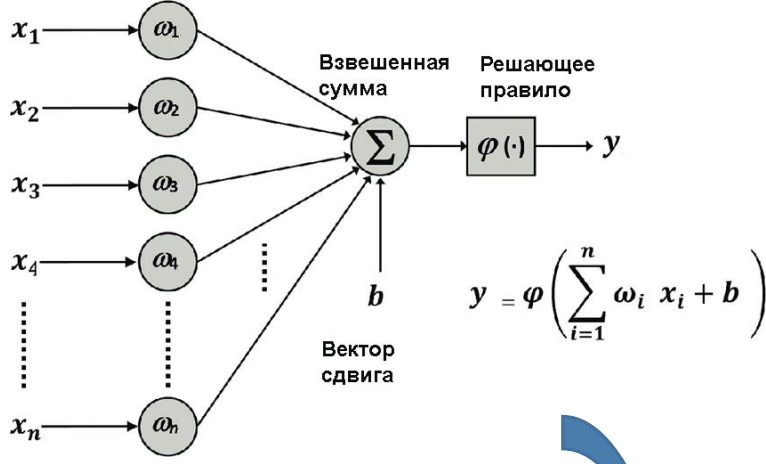
Как обманывают нейросети

Петюшко Александр

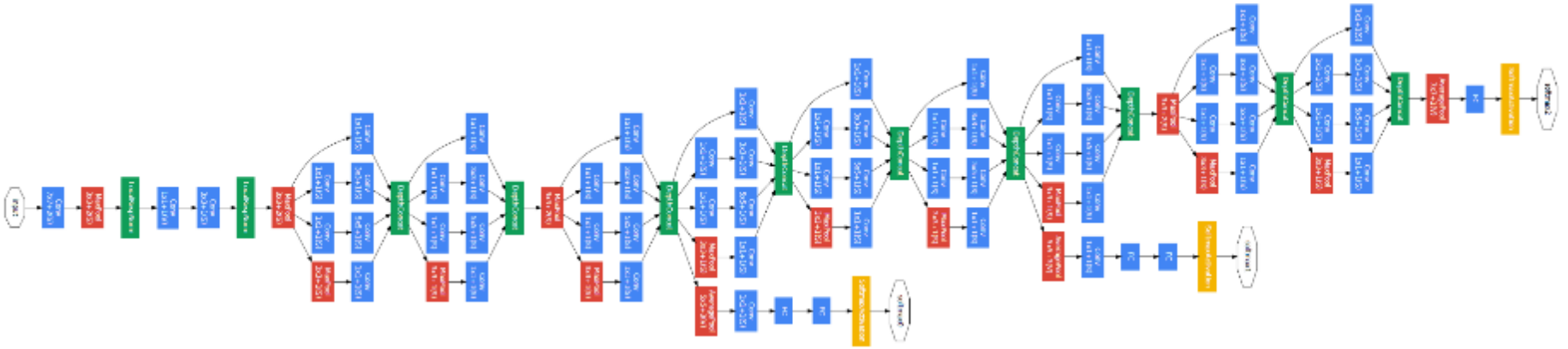
МГУ, мех-мат, к.ф.-м.н.

Huawei, ведущий инженер

Развитие нейросетей



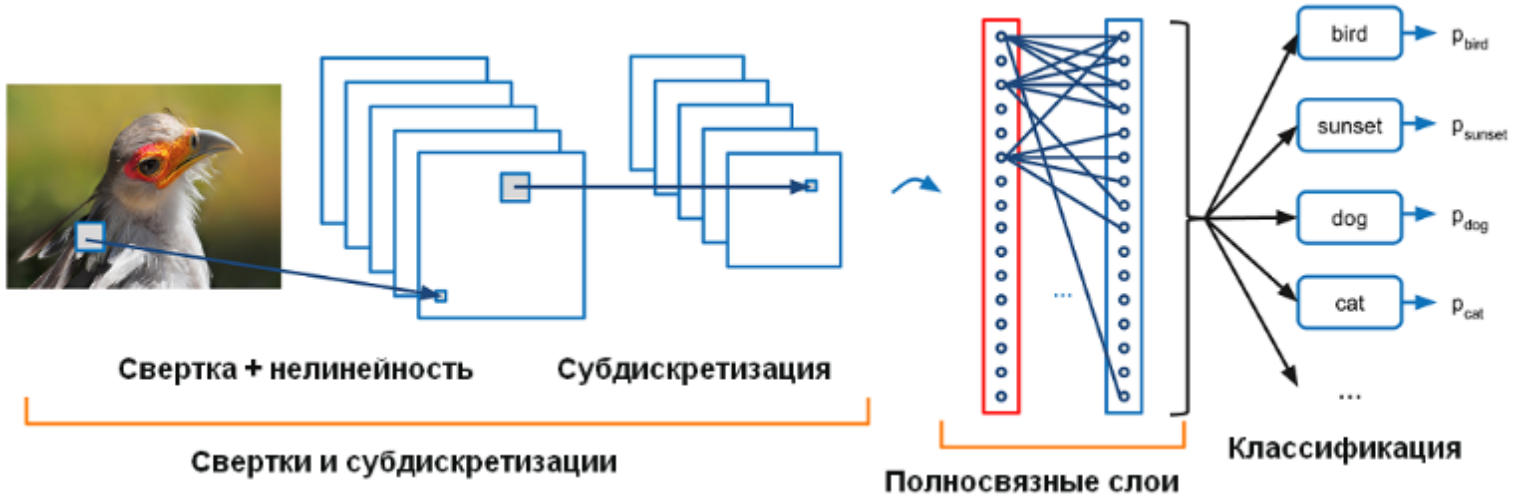
- Начиная с 1943 года, когда впервые была предложена математическая формализация МакКалоком и Питтсом понятия «искусственного нейрона», нейросети становились
  - Объемнее (содержали больше параметров)
  - Глубже (содержали больше блоков вычислений)
  - Лучше! (более правильно решали поставленные перед ними задачи)



Credits: https://arxiv.org/abs/1409.4842

Сверточные нейросети

- Для работы с фотографиями и видео лучше всего подходят сверточные нейронные сети (**СНС**)
- Например, позволяют выделять объекты и определять их класс
- Ну и отвечают на главный вопрос – кошка или собака?



Credits: https://adeshpande3.github.io/, https://stepupanalytics.com

СНС победили человека!..

- В конкурсе по распознаванию изображений **ImageNet** (классификация на 1000 классов) в настоящее время СНС дают ошибку top-5 порядка 2%, в то время как для подготовленного человека это всего лишь 5%
- В распознавании лиц на известной базе «**Labeled Faces in the Wild**» человеческий уровень в 97.5% был превзойден СНС еще в 2014 году на 1% (и составил 98.5%)
  - На данный момент сети соревнуются в точности между собой при уровне ошибки в 10^{-8} – 10^{-9} на гораздо более сложных базах

Credits: https://arxiv.org/abs/1404.3840, https://spectrum.ieee.org, http://www.image-net.org/, http://vis-www.cs.umass.edu/lfw/, http://karpathy.github.io

… Или все же нет?

- Оказывается, что можно внести практически **незаметные для глаза человека** возмущения во входные данные, которые, тем не менее, полностью поменяют выход нейронной сети
  - Например, результат классификации с «*панды*» поменяется на «*гиббона*»

Панда, 57.7%

$+ .007 \times$

$=$

Гиббон, 99.3%

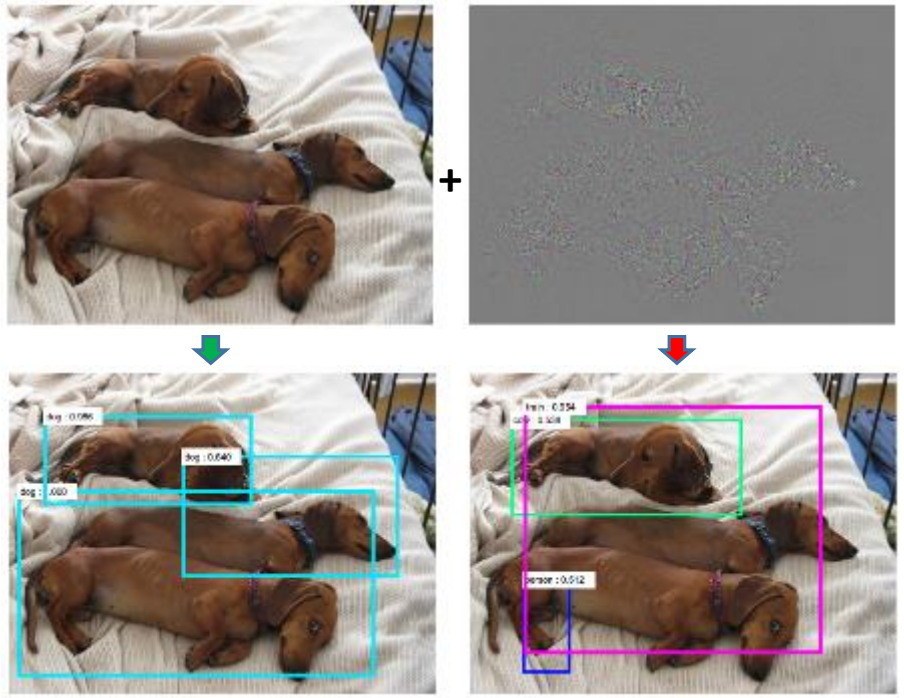
Называется такое возмущение состязательной атакой

Credits: https://arxiv.org/abs/1412.6572

Другие примеры «обмана»

- Точно так же можно «обманывать» не только СНС для классификации, но и
  - СНС для других задач (обнаружения и сегментирования объектов)
  - Нейросети для работы с текстом

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean



Собака, Собака, Собака

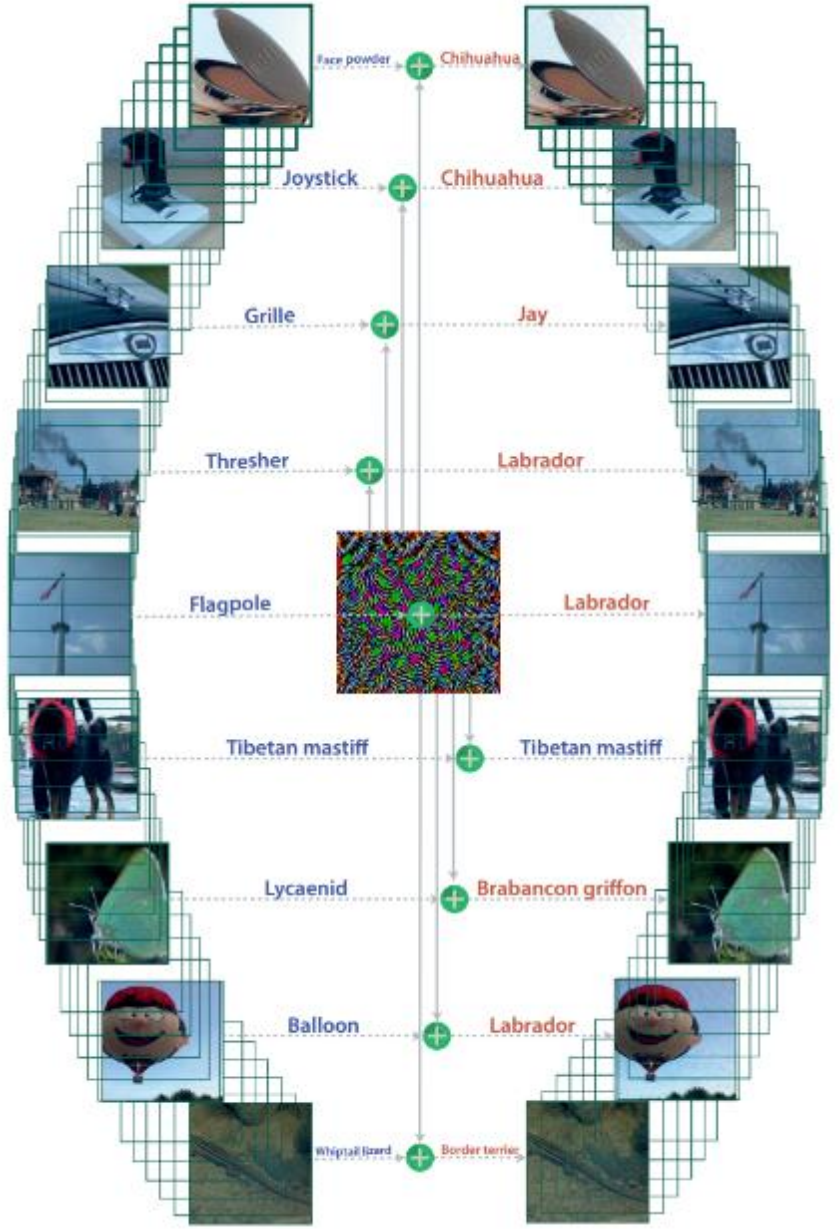
Поезд, Корова, Человек

Credits: https://arxiv.org/abs/1707.07328, https://arxiv.org/abs/1703.08603

Универсальная атака

- Более того, существуют **универсальные** атаки, которые работают для любых картинок в рамках заданной задачи

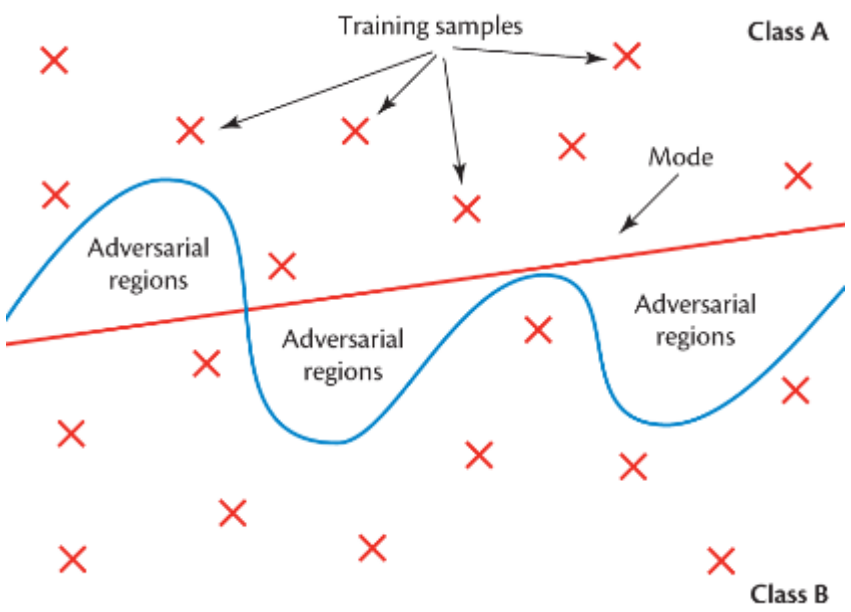
- *Пример справа:*
  - *универсальное возмущение для любого входа, которое ломает классификатор*

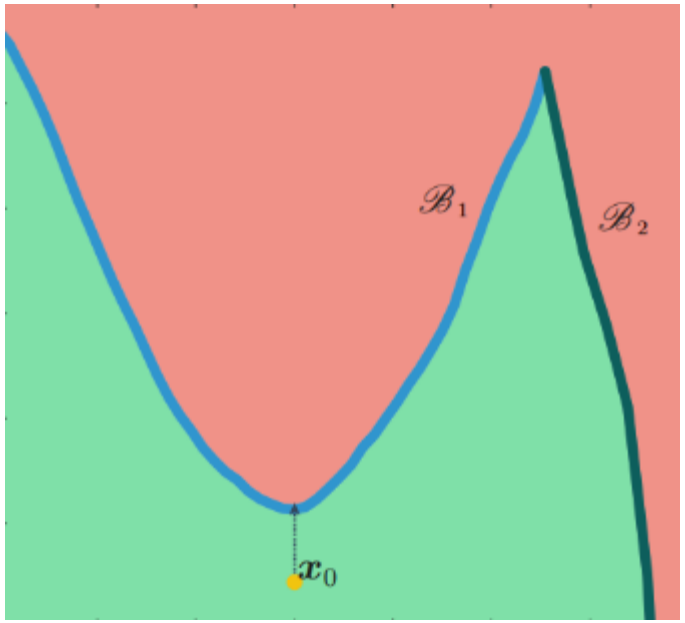


Credits: https://arxiv.org/abs/1610.08401

Причины «поломки»

- Одна из основных причин такого поведения нейросетей на близких изображениях – их метод обучения
  - А именно, разделяющие границы часто проходят очень близко к обучающим данным, и легко «заступить» за такую границу

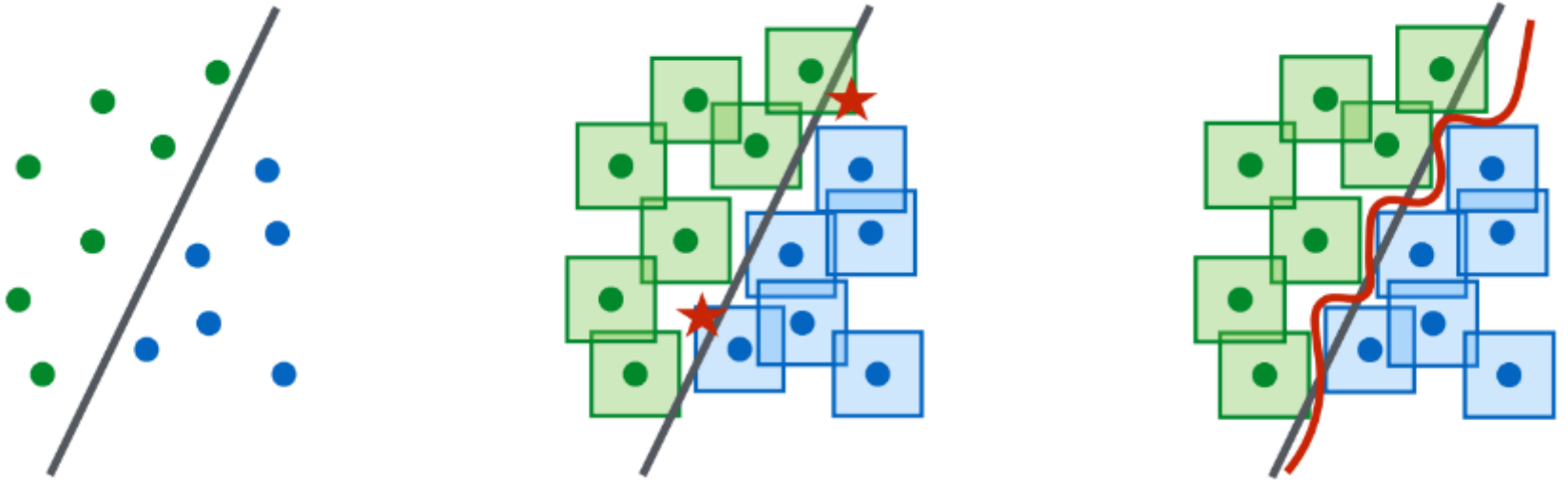




Credits: https://secml.github.io/, https://arxiv.org/abs/1608.08967

Давайте чинить!

- **Предложение**: А что если во время обучения к обучающему примеру добавить всю его «пиксельную» окрестность?



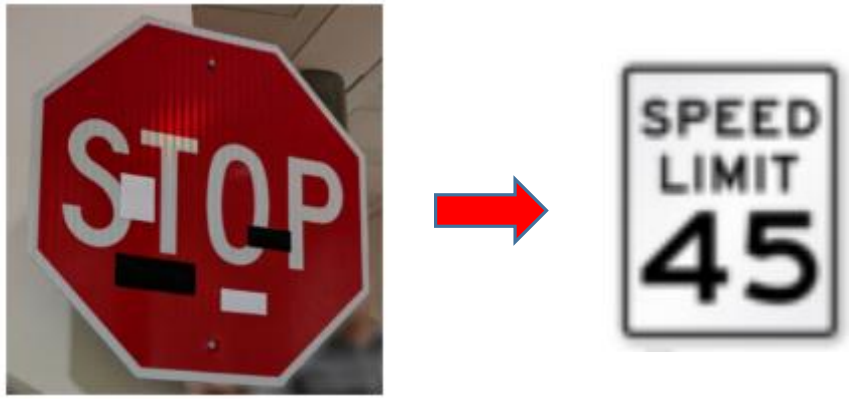
Credits: https://arxiv.org/abs/1706.06083

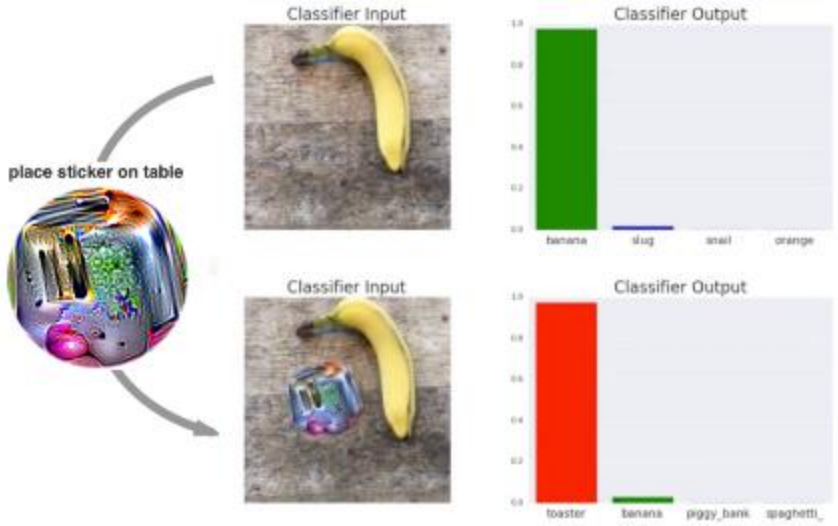
Давайте чинить...

- **Ответ**: в целом, так можно делать
- **Однако**:
  - Предположим, что исходная картинка размера 100 на 100, 3 цвета RGB
  - Предположим, что наш глаз не сильно различает колебания цвета пикселей в 2 градации (из 256)
  - Тогда для каждого обучающего примера нужно добавить:
    - $2^{3*100*100} = 2^{30000} = (2^{10})^{3000} \approx (10^{3})^{3000} = 10^{9000}$
  - Это гораздо больше числа атомов в видимой части Вселенной (10^{80})!
  - В общем, не очень реалистично

Переход в физическую реальность

- **Вопрос**: до этого все примеры были в цифровой области, исключительно менялись пиксели. А может, в физической реальности никаких состязательных атак не бывает?
- **Ответ**: бывает. И для простых объектов продемонстрированы
  - Наклейки на дорожные знаки
  - Разноцветный кусочек бумаги рядом с объектом и др.



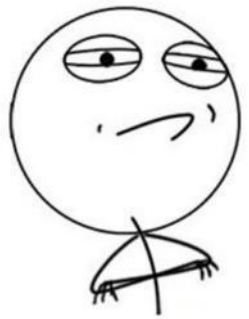


Credits: https://arxiv.org/abs/1712.09665, https://arxiv.org/abs/1707.08945, https://www.meme-arsenal.com

Вызов

- **Вызов**: а можно ли сделать атаку в физической реальности на сложный объект, с которым нейросети давно и уверенно справляются лучше человека?
- И еще чтобы атака была устойчивой к изменениям внешних условий (освещение, повороты)!
- Например, распознавание лиц.

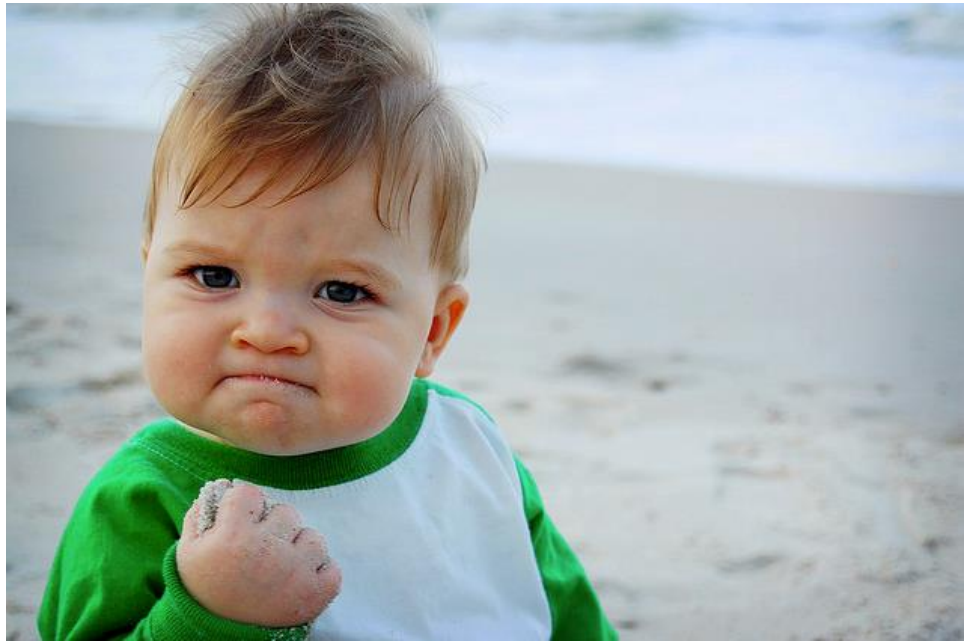
Challenge accepted!



Credits: https://knowyourmeme.com

Ии?

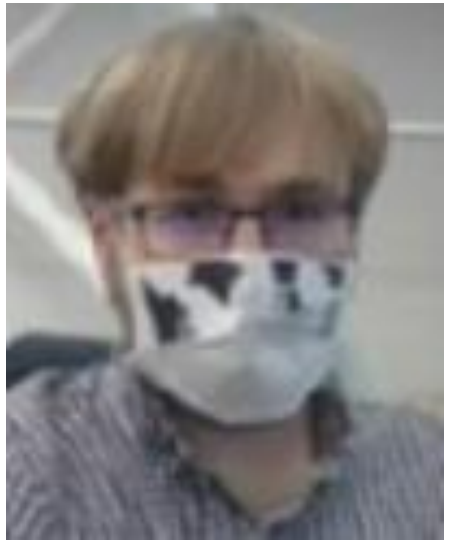
- Мы смогли найти успешную атаку для ведущей открытой системы распознавания лиц **ArcFace** в разных условиях
- Более того, заодно подобрали атаку и для одной из ведущих систем обнаружения лиц - **MTCNN**



Credits: https://arxiv.org/abs/1908.08705, https://arxiv.org/abs/1801.07698, https://arxiv.org/abs/1604.02878, http://reddit.com

Наши атаки

Реальная атака на Face ID



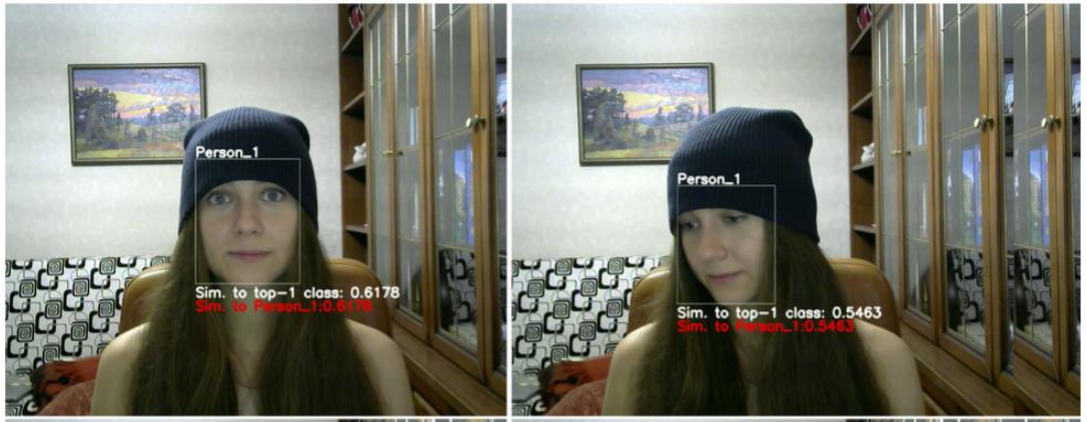
Реальная атака на Face Detection

Устойчивость

- И главное – атака устойчива к поворотам и разной освещенности

Фронтальное лицо (нет атаки)

Близость до своего эталона: 0.61



Поворот лица (нет атаки)

Близость до своего эталона: 0.54

Фронтальное лицо (атака)

Близость до своего эталона: 0.02
Близость до другого эталона: 0.23

Поворот лица (атака)

Близость до своего эталона: 0.11
Близость до другого эталона: 0.27

Credits: https://www.youtube.com/watch?v=a4iNg0wWBsQ

Спасибо за внимание!

(для человечества еще не все потеряно)