

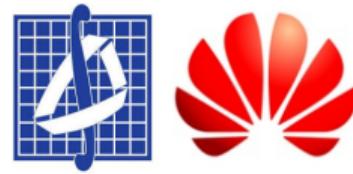
# Введение в искусственный интеллект. Современное компьютерное зрение

## Лекция 10. Состязательные атаки

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем

17 декабря 2019 г.



## ① Потрясающие успехи СНС в компьютерном зрении

- ① Потрясающие успехи СНС в компьютерном зрении
- ② (Не) устойчивость СНС в компьютерном зрении

- ① Потрясающие успехи СНС в компьютерном зрении
- ② (Не) устойчивость СНС в компьютерном зрении
- ③ Классификация состязательных атак

- ① Потрясающие успехи СНС в компьютерном зрении
- ② (Не) устойчивость СНС в компьютерном зрении
- ③ Классификация состязательных атак
- ④ Методы состязательных атак в цифровой области

- ① Потрясающие успехи СНС в компьютерном зрении
- ② (Не) устойчивость СНС в компьютерном зрении
- ③ Классификация состязательных атак
- ④ Методы состязательных атак в цифровой области
- ⑤ Методы состязательных атак в реальном мире

- ① Потрясающие успехи СНС в компьютерном зрении
- ② (Не) устойчивость СНС в компьютерном зрении
- ③ Классификация состязательных атак
- ④ Методы состязательных атак в цифровой области
- ⑤ Методы состязательных атак в реальном мире
- ⑥ Состязательные атаки на системы детекции и распознавания лиц в реальном мире

# Интересные вопросы

Давайте разберемся, так ли уж хороши сверточные нейросети,  
действительно ли оправдано все то внимание, которое им  
уделяют?

---

<sup>1</sup>Image credit: <https://spectrum.ieee.org>

Давайте разберемся, так ли уж хороши сверточные нейросети, действительно ли оправдано все то внимание, которое им уделяют?

## Вопрос1

Как сейчас соотносится качество распознавания человеком и СНС для известных баз данных?

<sup>1</sup>Image credit: <https://spectrum.ieee.org>

# Интересные вопросы

Давайте разберемся, так ли уж хороши сверточные нейросети, действительно ли оправдано все то внимание, которое им уделяют?

## Вопрос1

Как сейчас соотносится качество распознавания человеком и СНС для известных баз данных?

## Вопрос2

Насколько устойчивы СНС по отношению к входным данным?

Легко ли их сломать?

<sup>1</sup>Image credit: <https://spectrum.ieee.org>

Давайте разберемся, так ли уж хороши сверточные нейросети, действительно ли оправдано все то внимание, которое им уделяют?

## Вопрос1

Как сейчас соотносится качество распознавания человеком и СНС для известных баз данных?

## Вопрос2

Насколько устойчивы СНС по отношению к входным данным?  
Легко ли их сломать?

CNN vs Human<sup>1</sup>



<sup>1</sup>Image credit: <https://spectrum.ieee.org>

# Человек или СНС?

## ImageNet<sup>2</sup> (1000-классовая база данных изображений)

- Топ-5 ошибка для человека<sup>3</sup>: 5.1%
- Топ-5 ошибка для СНС<sup>4</sup>: 2.0%

---

<sup>2</sup><http://www.image-net.org/>

<sup>3</sup><http://karpathy.github.io/2014/09/02/>

[what-i-learned-from-competing-against-a-convnet-on-imagenet/](#)

<sup>4</sup>Touvron, Hugo, et al. "Fixing the train-test resolution discrepancy." 2019

<sup>5</sup><http://vis-www.cs.umass.edu/lfw/>

<sup>6</sup>Kumar, Neeraj, et al. "Attribute and simile classifiers for face verification." 2009

<sup>7</sup>Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



# Человек или СНС?

## ImageNet<sup>2</sup> (1000-классовая база данных изображений)

- Топ-5 ошибка для человека<sup>3</sup>: 5.1%
- Топ-5 ошибка для СНС<sup>4</sup>: 2.0%

## Labeled Faces in the Wild<sup>5</sup> (база данных лиц)

- Ошибка верификации для человека<sup>6</sup>: 2.47%
- Ошибка верификации для СНС<sup>7</sup>: 0.17%

<sup>2</sup><http://www.image-net.org/>

<sup>3</sup><http://karpathy.github.io/2014/09/02/>

[what-i-learned-from-competing-against-a-convnet-on-imagenet/](#)

<sup>4</sup>Touvron, Hugo, et al. "Fixing the train-test resolution discrepancy." 2019

<sup>5</sup><http://vis-www.cs.umass.edu/lfw/>

<sup>6</sup>Kumar, Neeraj, et al. "Attribute and simile classifiers for face verification." 2009

<sup>7</sup>Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



# Человек или СНС?

## ImageNet<sup>2</sup> (1000-классовая база данных изображений)

- Топ-5 ошибка для человека<sup>3</sup>: 5.1%
- Топ-5 ошибка для СНС<sup>4</sup>: 2.0%

lfw



## Labeled Faces in the Wild<sup>5</sup> (база данных лиц)

- Ошибка верификации для человека<sup>6</sup>: 2.47%
- Ошибка верификации для СНС<sup>7</sup>: 0.17%

<sup>2</sup><http://www.image-net.org/>

<sup>3</sup><http://karpathy.github.io/2014/09/02/>

[what-i-learned-from-competing-against-a-convnet-on-imagenet/](#)

<sup>4</sup>Touvron, Hugo, et al. "Fixing the train-test resolution discrepancy." 2019

<sup>5</sup><http://vis-www.cs.umass.edu/lfw/>

<sup>6</sup>Kumar, Neeraj, et al. "Attribute and simile classifiers for face verification." 2009

<sup>7</sup>Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



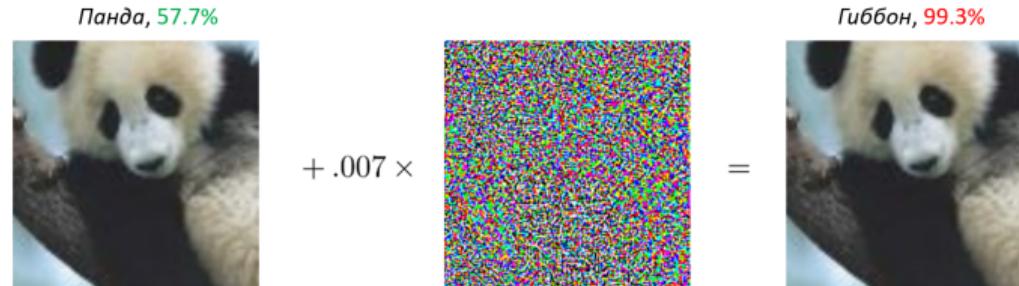
# Такие неустойчивые СНС

- Можно внести практически незаметные для глаза человека возмущения во входные данные, которые, тем не менее, полностью поменяют выход нейронной сети

<sup>8</sup>Image credit: <https://arxiv.org/pdf/1412.6572.pdf>

# Такие неустойчивые СНС

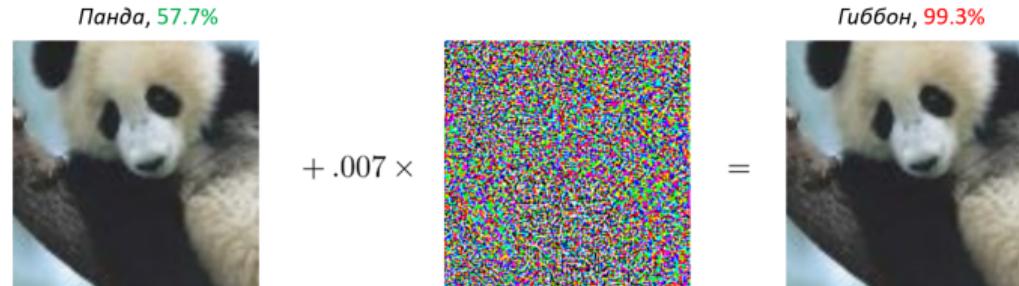
- Можно внести практически незаметные для глаза человека возмущения во входные данные, которые, тем не менее, полностью поменяют выход нейронной сети
- Например, результат классификации с “панды” поменяется на “гиббона”<sup>8</sup>



<sup>8</sup>Image credit: <https://arxiv.org/pdf/1412.6572.pdf>

# Такие неустойчивые СНС

- 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”<sup>8</sup>



Такое возмущение называется **состязательной атакой** (adversarial attack)

<sup>8</sup>Image credit: <https://arxiv.org/pdf/1412.6572.pdf>

# Атака СНС, предназначенных для сегментации или обнаружения

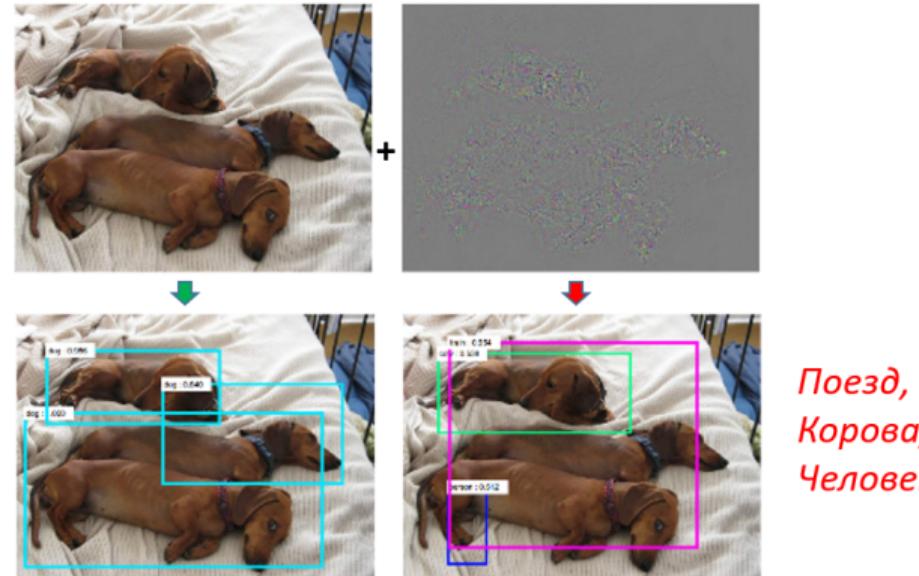
- Можно атаковать также СНС, которые не предназначены для классификации — например, для обнаружения и сегментации изображений<sup>9</sup>

<sup>9</sup>Xie, Cihang, et al. “Adversarial examples for semantic segmentation and object detection.” 2017



# Атака СНС, предназначенных для сегментации или обнаружения

- Можно атаковать также СНС, которые не предназначены для классификации — например, для обнаружения и сегментации изображений<sup>9</sup>



<sup>9</sup>Xie, Cihang, et al. “Adversarial examples for semantic segmentation and object detection.” 2017

# Атака нейросетей, не предназначенных для изображений

- Можно атаковать даже НС, которые вообще не работают с изображениями — например, НС для вопросно-ответных систем (QA, question answering systems)<sup>10</sup>

<sup>10</sup> Jia, Robin, and Percy Liang. "Adversarial examples for evaluating reading comprehension systems." 2017 

# Атака нейросетей, не предназначенных для изображений

- Можно атаковать даже НС, которые вообще не работают с изображениями — например, НС для вопросно-ответных систем (QA, question answering systems)<sup>10</sup>

**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*

<sup>10</sup> Jia, Robin, and Percy Liang. “Adversarial examples for evaluating reading comprehension systems.” 2017



# Одна из главных причин существования атак

- Одна из основных причин такого поведения СНС на похожих изображениях — неустойчивость СНС

---

<sup>11</sup>Image credit: <https://secml.github.io/>

<sup>12</sup>Fawzi, Alhussein, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. “Robustness of classifiers: from  adversarial to random noise.” 2016 

# Одна из главных причин существования атак

- Одна из основных причин такого поведения СНС на похожих изображениях — неустойчивость СНС
- А именно, разделяющие границы классификатора часто проходят очень близко к обучающим данным, и легко “заступить” за такую границу<sup>11,12</sup>

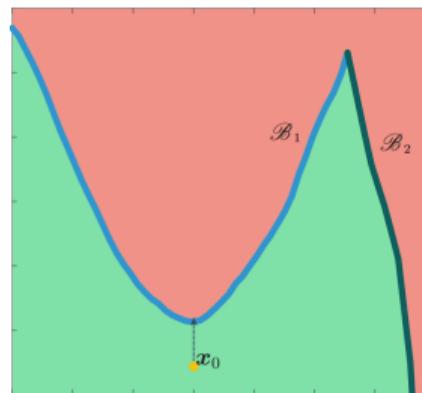
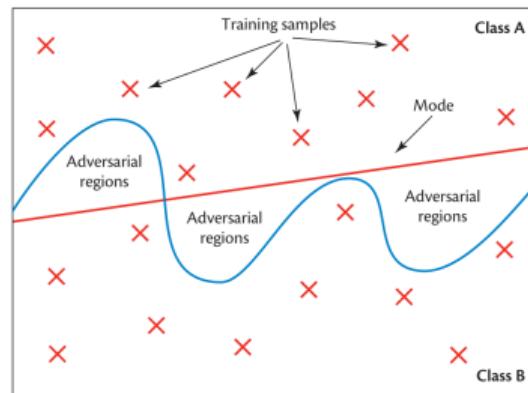
---

<sup>11</sup>Image credit: <https://secml.github.io/>

<sup>12</sup>Fawzi, Alhussein, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. “Robustness of classifiers: from  adversarial to random noise.” 2016 

# Одна из главных причин существования атак

- Одна из основных причин такого поведения СНС на похожих изображениях — неустойчивость СНС
- А именно, разделяющие границы классификатора часто проходят очень близко к обучающим данным, и легко “заступить” за такую границу<sup>11,12</sup>



<sup>11</sup>Image credit: <https://secml.github.io/>

<sup>12</sup>Fawzi, Alhussein, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Robustness of classifiers: from adversarial to random noise." 2016

# Простой метод защиты

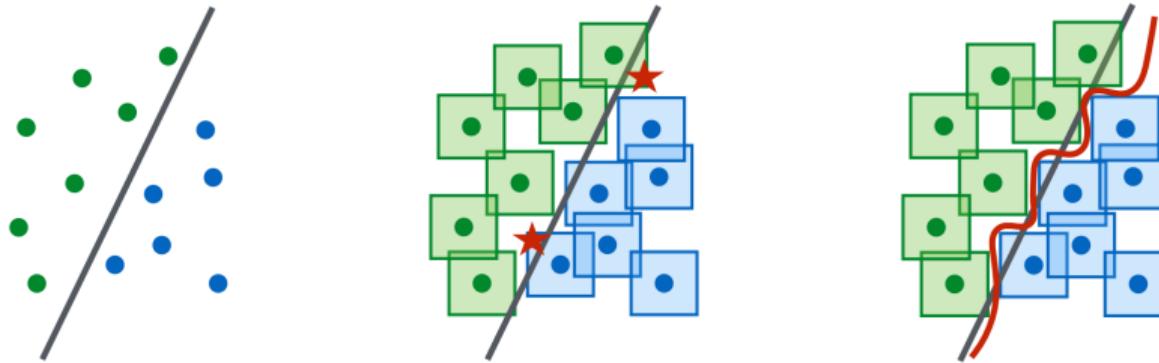
- Поскольку можно обмануть СНС путем небольшого пиксельного возмущения, то почему бы во время обучения для каждого обучающего примера не добавлять и всю его попиксельную окрестность (по некоторой норме, например,  $\ell_\infty$ )<sup>13</sup>

<sup>13</sup> Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." 2017



# Простой метод защиты

- Поскольку можно обмануть СНС путем небольшого пиксельного возмущения, то почему бы во время обучения для каждого обучающего примера не добавлять и всю его попиксельную окрестность (по некоторой норме, например,  $\ell_\infty$ )<sup>13</sup>



<sup>13</sup> Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." 2017

# Простой, но не работающий метод защиты

- Предположим, что исходная картинка размера  $100 \times 100$  пикселей, 3 цвета RGB

## Простой, но не работающий метод защиты

- Предположим, что исходная картинка размера  $100 \times 100$  пикселей, 3 цвета RGB
- Предположим, что наш глаз не сильно различает колебания цвета пикселей в 2 градации (из 256): в каждой точке для каждого цвета можем позволить  $\pm 1$  значение

## Простой, но не работающий метод защиты

- Предположим, что исходная картинка размера  $100 \times 100$  пикселей, 3 цвета RGB
- Предположим, что наш глаз не сильно различает колебания цвета пикселей в 2 градации (из 256): в каждой точке для каждого цвета можем позволить  $\pm 1$  значение
- Тогда для каждого обучающего примера нужно добавить следующее количество его пиксельных соседей:

$$2^{3 \times 100 \times 100} = 2^{30000} = (2^{10})^{3000} \approx (10^3)^{3000} = 10^{9000}$$

## Простой, но не работающий метод защиты

- Предположим, что исходная картинка размера  $100 \times 100$  пикселей, 3 цвета RGB
- Предположим, что наш глаз не сильно различает колебания цвета пикселей в 2 градации (из 256): в каждой точке для каждого цвета можем позволить  $\pm 1$  значение
- Тогда для каждого обучающего примера нужно добавить следующее количество его пиксельных соседей:

$$2^{3 \times 100 \times 100} = 2^{30000} = (2^{10})^{3000} \approx (10^3)^{3000} = 10^{9000}$$

- Это гораздо больше числа атомов в видимой части Вселенной ( $10^{80}$ )!

## Простой, но не работающий метод защиты

- Предположим, что исходная картинка размера  $100 \times 100$  пикселей, 3 цвета RGB
- Предположим, что наш глаз не сильно различает колебания цвета пикселей в 2 градации (из 256): в каждой точке для каждого цвета можем позволить  $\pm 1$  значение
- Тогда для каждого обучающего примера нужно добавить следующее количество его пиксельных соседей:

$$2^{3 \times 100 \times 100} = 2^{30000} = (2^{10})^{3000} \approx (10^3)^{3000} = 10^{9000}$$

- Это гораздо больше числа атомов в видимой части Вселенной ( $10^{80}$ )!
- В общем, не очень реалистично

# Работающий метод защиты

- Давайте не перебирать всю окрестность обучающего примера, а брать только те точки из окрестности, которые ближе всего к разделяющей поверхности

---

<sup>14</sup> Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



# Работающий метод защиты

- Давайте не перебирать всю окрестность обучающего примера, а брать только те точки из окрестности, которые ближе всего к разделяющей поверхности
- Такой метод обучения называется состязательным (adversarial training)<sup>14</sup>

---

<sup>14</sup> Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



# Работающий метод защиты

- Давайте не перебирать всю окрестность обучающего примера, а брать только те точки из окрестности, которые ближе всего к разделяющей поверхности
- Такой метод обучения называется состязательным (adversarial training)<sup>14</sup>

## Плюсы состязательного обучения

- Не нужно перебирать всю окрестность огромной мощности
- В целом, защищает от метода нахождения состязательных примеров

---

<sup>14</sup> Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



# Работающий метод защиты

- Давайте не перебирать всю окрестность обучающего примера, а брать только те точки из окрестности, которые ближе всего к разделяющей поверхности
- Такой метод обучения называется состязательным (adversarial training)<sup>14</sup>

## Плюсы состязательного обучения

- Не нужно перебирать всю окрестность огромной мощности
- В целом, защищает от метода нахождения состязательных примеров

## Минусы состязательного обучения

- Процедура нахождения хороших состязательных примеров работает медленно (гораздо медленнее одного градиентного шага)
- Защищает **только** от того метода нахождения состязательных примеров, который использовался в состязательном обучении

<sup>14</sup> Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



# Состязательные атаки: необходимые обозначения

- Пусть  $x \in B = [0, 1]^{C \times M \times N}$  — входная картинка  $C \times M \times N$ , где  $C$  — количество цветов (1 для ч/б, 3 для RGB)
- $y_{gt}$  — правильный класс для  $x$
- $\theta$  — параметры СНС-классификатора
- $L(\theta, x, y_{gt})$  — функция потерь
- $f(x)$  — выход классификатора (распознанный класс); при обучении мы добиваемся равенства  $f(x) = y_{gt}$

# Состязательные атаки: необходимые обозначения

- Пусть  $x \in B = [0, 1]^{C \times M \times N}$  — входная картинка  $C \times M \times N$ , где  $C$  — количество цветов (1 для ч/б, 3 для RGB)
- $y_{gt}$  — правильный класс для  $x$
- $\theta$  — параметры СНС-классификатора
- $L(\theta, x, y_{gt})$  — функция потерь
- $f(x)$  — выход классификатора (распознанный класс); при обучении мы добиваемся равенства  $f(x) = y_{gt}$
- $r \in B = [0, 1]^{C \times M \times N}$  — аддитивная добавка ко входу  $x$

## Цель состязательной атаки

Поменять выход классификатора  $f$  на неправильный путем добавления минимального по некоторой норме (на практике используются  $\ell_0$ ,  $\ell_1$ ,  $\ell_2$  и  $\ell_\infty$  — обозначим через  $\ell_p$ ) возмущения  $r$ , а именно:



## Цель состязательной атаки

Поменять выход классификатора  $f$  на неправильный путем добавления минимального по некоторой норме (на практике используются  $\ell_0$ ,  $\ell_1$ ,  $\ell_2$  и  $\ell_\infty$  — обозначим через  $\ell_p$ ) возмущения  $r$ , а именно: минимизировать  $\|r\|_p$  т.ч.

- ①  $f(x) = y_{gt}$
- ②  $f(x + r) \neq y_{gt}$
- ③  $x + r \in B$



## Нормы $\ell_p$ : напоминание

Напомним наиболее употребительные нормы  $\ell_p$  для  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ :

## Нормы $\ell_p$ : напоминание

Напомним наиболее употребительные нормы  $\ell_p$  для  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ :

- $\ell_2$ :  $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

## Нормы $\ell_p$ : напоминание

Напомним наиболее употребительные нормы  $\ell_p$  для  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ :

- $\ell_2$ :  $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- $\ell_1$ :  $\|x\|_1 = \sum_{i=1}^n |x_i|$

## Нормы $\ell_p$ : напоминание

Напомним наиболее употребительные нормы  $\ell_p$  для  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ :

- $\ell_2$ :  $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- $\ell_1$ :  $\|x\|_1 = \sum_{i=1}^n |x_i|$
- $\ell_\infty$ :  $\|x\|_\infty = \max_i |x_i|$

## Нормы $\ell_p$ : напоминание

Напомним наиболее употребительные нормы  $\ell_p$  для  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ :

- $\ell_2$ :  $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- $\ell_1$ :  $\|x\|_1 = \sum_{i=1}^n |x_i|$
- $\ell_\infty$ :  $\|x\|_\infty = \max_i |x_i|$
- $\ell_0$ :  $\|x\|_0 = \sum_{i=1}^n \mathbf{1}_{x_i \neq 0}$

## Нормы $\ell_p$ : напоминание

Напомним наиболее употребительные нормы  $\ell_p$  для  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ :

- $\ell_2$ :  $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- $\ell_1$ :  $\|x\|_1 = \sum_{i=1}^n |x_i|$
- $\ell_\infty$ :  $\|x\|_\infty = \max_i |x_i|$
- $\ell_0$ :  $\|x\|_0 = \sum_{i=1}^n \mathbf{1}_{x_i \neq 0}$

**Упражнение.** Доказать, что для  $0 < p < 1$  норма  $\ell_p$ , для которой  $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ , не является нормой

# Классификация состязательных атак

## По цели атаки

- Ненаправленная (untargeted): нужно просто сменить ответ классификатора
- Направленная (targeted): нужно сменить на заранее определенный класс  $y_t$

# Классификация состязательных атак

## По цели атаки

- Ненаправленная (untargeted): нужно просто сменить ответ классификатора
- Направленная (targeted): нужно сменить на заранее определенный класс  $y_t$

## По осведомленности атакующего

- Открытая (white-box): атакующий знает все о классификаторе (архитектуру и веса)
- Закрытая (black-box): атакующий имеет частичную информацию о классификаторе (обычно только информацию о выходе)



# Классификация состоятельных атак

## По цели атаки

- Ненаправленная (untargeted): нужно просто сменить ответ классификатора
- Направленная (targeted): нужно сменить на заранее определенный класс  $y_t$

## По осведомленности атакующего

- Открытая (white-box): атакующий знает все о классификаторе (архитектуру и веса)
- Закрытая (black-box): атакующий имеет частичную информацию о классификаторе (обычно только информацию о выходе)

## По условию применения

- Цифровая (digital): атака на фотографию
- Реальная (real-world): атака на реальный объект

## По универсальности

- Зависимая от входа (input-aware): возмущение  $r$  зависит от входа  $x$
- Универсальная (universal): возмущение  $r$  работает для любого входа  $x$

# Классификация состязательных атак

## По универсальности

- Зависимая от входа (input-aware): возмущение  $r$  зависит от входа  $x$
- Универсальная (universal): возмущение  $r$  работает для любого входа  $x$

## По переносимости

- Непереносимая (non-transferable): атака работает только для узкого класса классификаторов
- Переносимая (transferable): атака работает для широкого класса классификаторов (но при этом может быть не универсальной)
- Наиболее сложная атака — направленная закрытая реальная универсальная переносимая атака
- Для простоты будем рассматривать открытые атаки

## Эффективность состязательной атаки

Введем простой критерий успешности (success)  $S(A, Z)$  алгоритма  $A$  состязательной атаки  $r_A(x)$  на множестве  $Z \ni (x^i, y_{gt}^i)$ :

# Эффективность состязательной атаки

Введем простой критерий успешности (success)  $S(A, Z)$  алгоритма  $A$  состязательной атаки  $r_A(x)$  на множестве  $Z \ni (x^i, y_{gt}^i)$ :

- В случае ненаправленной атаки:

$$S(A, Z) = \frac{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\} \cdot \mathbf{1}\{f(x^i + r_A(x^i)) \neq y_{gt}^i\}}{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\}}$$

# Эффективность состязательной атаки

Введем простой критерий успешности (success)  $S(A, Z)$  алгоритма  $A$  состязательной атаки  $r_A(x)$  на множестве  $Z \ni (x^i, y_{gt}^i)$ :

- В случае ненаправленной атаки:

$$S(A, Z) = \frac{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\} \cdot \mathbf{1}\{f(x^i + r_A(x^i)) \neq y_{gt}^i\}}{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\}}$$

- В случае направленной атаки на класс  $y_t$ :

$$S(A, Z, y_t) = \frac{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\} \cdot \mathbf{1}\{f(x^i + r_A(x^i)) = y_t\}}{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\}}$$

# Эффективность состязательной атаки

Введем простой критерий успешности (success)  $S(A, Z)$  алгоритма  $A$  состязательной атаки  $r_A(x)$  на множестве  $Z \ni (x^i, y_{gt}^i)$ :

- В случае ненаправленной атаки:

$$S(A, Z) = \frac{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\} \cdot \mathbf{1}\{f(x^i + r_A(x^i)) \neq y_{gt}^i\}}{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\}}$$

- В случае направленной атаки на класс  $y_t$ :

$$S(A, Z, y_t) = \frac{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\} \cdot \mathbf{1}\{f(x^i + r_A(x^i)) = y_t\}}{\sum_i \mathbf{1}\{f(x^i) = y_{gt}^i\}}$$

**Замечание.** Очевидно, что  $S(A, Z, y_t) \leq S(A, Z)$

# Предтеча состязательных атак

- Изначально устойчивость СНС изучалась с точки зрения адекватной реакции на разные входы

---

<sup>15</sup>Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." 2014

# Предтеча состязательных атак

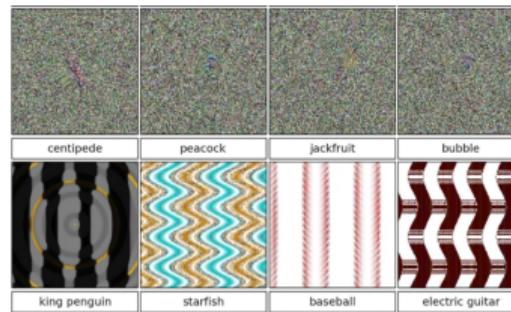
- Изначально устойчивость СНС изучалась с точки зрения адекватной реакции на разные входы
- Выяснилось, что существуют примеры (структурированные или нет), которые на выходе СНС могут давать с большой вероятностью любой класс

---

<sup>15</sup>Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." 2014

# Предтеча состязательных атак

- Изначально устойчивость СНС изучалась с точки зрения адекватной реакции на разные входы
- Выяснилось, что существуют примеры (структурированные или нет), которые на выходе СНС могут давать с большой вероятностью любой класс
- Такие примеры назывались “обманными изображениями”<sup>15</sup> (fooling images) и строились с помощью эволюционных алгоритмов



<sup>15</sup>Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." 2014

## Метод атаки: L-BFGS-B

- Первая предложенная атака<sup>16</sup> использовала  $\ell_2$ -норму для ограничения атаки
- Рассматривалась направленная атака на класс  $y_t \neq y_{gt}$
- Функционал для минимизации с ограничением  $x + r \in B$ ,  $c = const$ :

$$c\|r\|_2 + L(\theta, x, y_t) \rightarrow \min_r$$

- Для оптимизации использовался метод L-BFGS-B<sup>17</sup> (**L**imited memory **B**royden–**F**letcher–**G**oldfarb–**S**hanno algorithm with **B**ox constraints) — квази-Ньютоновский метод минимизации с ограничением на память и на переменные
- В какой-то мере атака была переносима на другие архитектуры

---

<sup>16</sup>Szegedy, Christian, et al. “Intriguing properties of neural networks.” 2013

<sup>17</sup>Byrd, Richard H., et al. “A limited memory algorithm for bound constrained optimization.” 1995

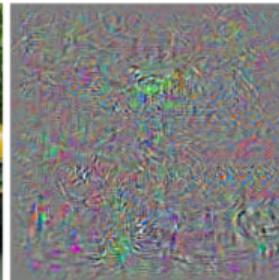


Пример работы:

Школьный  
автобус



10 \* г



Страус

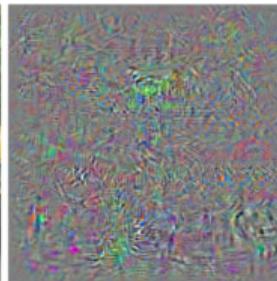


## Пример работы:

Школьный  
автобус



$10 * r$



Страус



## Переносимость:

	$FC10(10^{-4})$	$FC10(10^{-2})$	$FC10(1)$	$FC100-100-10$	$FC200-200-10$	$AE400-10$
$FC10(10^{-4})$	100%	11.7%	22.7%	2%	3.9%	2.7%
$FC10(10^{-2})$	87.1%	100%	35.2%	35.9%	27.3%	9.8%
$FC10(1)$	71.9%	76.2%	100%	48.1%	47%	34.4%
$FC100-100-10$	28.9%	13.7%	21.1%	100%	6.6%	2%
$FC200-200-10$	38.2%	14%	23.8%	20.3%	100%	2.7%
$AE400-10$	23.4%	16%	24.8%	9.4%	6.6%	100%

## Метод атаки: FGSM

- Несмотря на хорошую реализацию, метод L-BFGS-B не так быстр и требует внешнего (по отношению к исследуемой СНС) оптимизатора

---

<sup>18</sup>Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



## Метод атаки: FGSM

- Несмотря на хорошую реализацию, метод L-BFGS-В не так быстр и требует внешнего (по отношению к исследуемой СНС) оптимизатора
- **Предложение:** использовать линейную часть функции потерь в окрестности  $x$  и идти по градиенту — FGSM<sup>18</sup> (**F**ast **G**radient **S**ign **M**ethod):

$$r = \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y_t))$$

где  $0 < \epsilon < 1$  — некоторая константа

---

<sup>18</sup>Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



## Метод атаки: FGSM

- Несмотря на хорошую реализацию, метод L-BFGS-В не так быстр и требует внешнего (по отношению к исследуемой СНС) оптимизатора
- **Предложение:** использовать линейную часть функции потерь в окрестности  $x$  и идти по градиенту — FGSM<sup>18</sup> (**F**ast **G**radient **S**ign **M**ethod):

$$r = \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y_t))$$

где  $0 < \epsilon < 1$  — некоторая константа

- **Напоминание:** для оптимизации весов СНС мы применяли метод обратного распространения ошибок, где брали градиент по весам СНС, т.е.  $\nabla_{\theta} L(\theta, x, y_{gt})$

---

<sup>18</sup>Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



## Метод атаки: FGSM

- Несмотря на хорошую реализацию, метод L-BFGS-B не так быстр и требует внешнего (по отношению к исследуемой СНС) оптимизатора
- **Предложение:** использовать линейную часть функции потерь в окрестности  $x$  и идти по градиенту — FGSM<sup>18</sup> (**F**ast **G**radient **S**ign **M**ethod):

$$r = \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y_t))$$

где  $0 < \epsilon < 1$  — некоторая константа

- **Напоминание:** для оптимизации весов СНС мы применяли метод обратного распространения ошибок, где брали градиент по весам СНС, т.е.  $\nabla_{\theta} L(\theta, x, y_{gt})$
- Теперь исследуется норма возмущения  $\ell_{\infty}$  как наиболее близкая к тому, что использует человек

---

<sup>18</sup>Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." 2014



# Метод атаки: I-FGSM (PGD)

- Часто линейная оценка окрестности функции достаточно грубая, и один шаг FGSM порой не приводит к хорошей атаке

<sup>19</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. “Adversarial examples in the physical world.” 2016



## Метод атаки: I-FGSM (PGD)

- Часто линейная оценка окрестности функции достаточно грубая, и один шаг FGSM порой не приводит к хорошей атаке
- Для этого применяют итеративный метод I-FGSM<sup>19</sup> (Iterative FGSM), который позволяет двигаться в сторону границы классификатора более точно

<sup>19</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



## Метод атаки: I-FGSM (PGD)

- Часто линейная оценка окрестности функции достаточно грубая, и один шаг FGSM порой не приводит к хорошей атаке
- Для этого применяют итеративный метод I-FGSM<sup>19</sup> (Iterative FGSM), который позволяет двигаться в сторону границы классификатора более точно
- Если  $\Pi_B$  — проекция на  $B$ , то в случае ненаправленной атаки

$$x^{n+1} = \Pi_B(x^n + \text{sign } \nabla_x L(\theta, x, y_{gt})), \quad x^0 = x$$

<sup>19</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



# Метод атаки: I-FGSM (PGD)

- Часто линейная оценка окрестности функции достаточно грубая, и один шаг FGSM порой не приводит к хорошей атаке
- Для этого применяют итеративный метод I-FGSM<sup>19</sup> (Iterative FGSM), который позволяет двигаться в сторону границы классификатора более точно
- Если  $\Pi_B$  — проекция на  $B$ , то в случае ненаправленной атаки

$$x^{n+1} = \Pi_B(x^n + \text{sign } \nabla_x L(\theta, x, y_{gt})), \quad x^0 = x$$

- Если принять  $\|x - x_{adv}\|_\infty \leq \epsilon$ , то авторы предлагают делать  $n = \min(256\epsilon + 4, 320\epsilon)$  шагов

<sup>19</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



## Метод атаки: I-FGSM (PGD)

- Часто линейная оценка окрестности функции достаточно грубая, и один шаг FGSM порой не приводит к хорошей атаке
- Для этого применяют итеративный метод I-FGSM<sup>19</sup> (Iterative FGSM), который позволяет двигаться в сторону границы классификатора более точно
- Если  $\Pi_B$  — проекция на  $B$ , то в случае ненаправленной атаки

$$x^{n+1} = \Pi_B(x^n + \text{sign } \nabla_x L(\theta, x, y_{gt})), \quad x^0 = x$$

- Если принять  $\|x - x_{adv}\|_\infty \leq \epsilon$ , то авторы предлагают делать  $n = \min(256\epsilon + 4, 320\epsilon)$  шагов
- Этот метод также называется PGD (Projected Gradient Descent)

<sup>19</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



# Метод атаки: MI-FGSM

- **Замечание:** Методы атаки все больше похожи на шаги оптимизатора

<sup>20</sup>Dong, Yinpeng, et al. "Boosting adversarial attacks with momentum." 2017



# Метод атаки: MI-FGSM

- **Замечание:** Методы атаки все больше похожи на шаги оптимизатора
- **Идея:** давайте использовать сглаживание градиента — MI-FGSM<sup>20</sup> (Momentum I-FGSM)

<sup>20</sup>Dong, Yinpeng, et al. "Boosting adversarial attacks with momentum." 2017



# Метод атаки: MI-FGSM

- **Замечание:** Методы атаки все больше похожи на шаги оптимизатора
- **Идея:** давайте использовать сглаживание градиента — MI-FGSM<sup>20</sup> (Momentum I-FGSM)

---

## Algorithm 1 MI-FGSM

**Input:** A classifier  $f$  with loss function  $J$ ; a real example  $\mathbf{x}$  and ground-truth label  $y$ ;

**Input:** The size of perturbation  $\epsilon$ ; iterations  $T$  and decay factor  $\mu$ .

**Output:** An adversarial example  $\mathbf{x}^*$  with  $\|\mathbf{x}^* - \mathbf{x}\|_\infty \leq \epsilon$ .

- 1:  $\alpha = \epsilon/T$ ;
- 2:  $\mathbf{g}_0 = 0$ ;  $\mathbf{x}_0^* = \mathbf{x}$ ;
- 3: **for**  $t = 0$  to  $T - 1$  **do**
- 4:     Input  $\mathbf{x}_t^*$  to  $f$  and obtain the gradient  $\nabla_{\mathbf{x}} J(\mathbf{x}_t^*, y)$ ;
- 5:     Update  $\mathbf{g}_{t+1}$  by accumulating the velocity vector in the gradient direction as

$$\mathbf{g}_{t+1} = \mu \cdot \mathbf{g}_t + \frac{\nabla_{\mathbf{x}} J(\mathbf{x}_t^*, y)}{\|\nabla_{\mathbf{x}} J(\mathbf{x}_t^*, y)\|_1}; \quad (6)$$

- 6:     Update  $\mathbf{x}_{t+1}^*$  by applying the sign gradient as
- $$\mathbf{x}_{t+1}^* = \mathbf{x}_t^* + \alpha \cdot \text{sign}(\mathbf{g}_{t+1}); \quad (7)$$

7: **end for**

8: **return**  $\mathbf{x}^* = \mathbf{x}_T^*$ .

---

<sup>20</sup>Dong, Yinpeng, et al. "Boosting adversarial attacks with momentum." 2017



# Сравнение FGSM-like атак

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3 <sub>ens3</sub>	Inc-v3 <sub>ens4</sub>	IncRes-v2 <sub>ens</sub>
Inc-v3	FGSM	72.3*	28.2	26.2	25.3	11.3	10.9	4.8
	I-FGSM	<b>100.0*</b>	22.8	19.9	16.2	7.5	6.4	4.1
	MI-FGSM	<b>100.0*</b>	<b>48.8</b>	<b>48.0</b>	<b>35.6</b>	<b>15.1</b>	<b>15.2</b>	<b>7.8</b>
Inc-v4	FGSM	32.7	61.0*	26.6	27.2	13.7	11.9	6.2
	I-FGSM	35.8	<b>99.9*</b>	24.7	19.3	7.8	6.8	4.9
	MI-FGSM	<b>65.6</b>	<b>99.9*</b>	<b>54.9</b>	<b>46.3</b>	<b>19.8</b>	<b>17.4</b>	<b>9.6</b>
IncRes-v2	FGSM	32.6	28.1	55.3*	25.8	13.1	12.1	7.5
	I-FGSM	37.8	20.8	<b>99.6*</b>	22.8	8.9	7.8	5.8
	MI-FGSM	<b>69.8</b>	<b>62.1</b>	99.5*	<b>50.6</b>	<b>26.1</b>	<b>20.9</b>	<b>15.7</b>
Res-152	FGSM	35.0	28.2	27.5	72.9*	14.6	13.2	7.5
	I-FGSM	26.7	22.7	21.2	<b>98.6*</b>	9.3	8.9	6.2
	MI-FGSM	<b>53.6</b>	<b>48.9</b>	<b>44.7</b>	98.5*	<b>22.1</b>	<b>21.7</b>	<b>12.9</b>

# Метод атаки: DeepFool

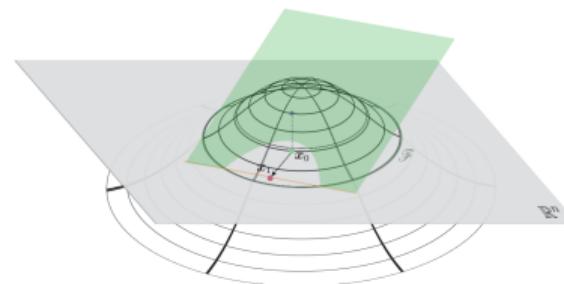
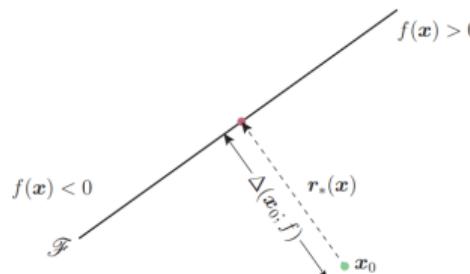
- Идея: проецировать точку  $x_0$  на разделяющую поверхность

# Метод атаки: DeepFool

- Идея: проецировать точку  $x_0$  на разделяющую поверхность
- В случае линейного бинарного классификатора  $\text{sign } f(x) = \text{sign}(w^T x + b)$ :
  - Направление:  $-\text{sign } f(x_0) \frac{w}{\|w\|_2}$
  - Длина:  $\frac{|f(x_0)|}{\|w\|_2}$
  - $\Rightarrow$  Атака:  $r = -\frac{f(x_0)}{\|w\|_2^2} w$

# Метод атаки: DeepFool

- Идея: проецировать точку  $x_0$  на разделяющую поверхность
- В случае линейного бинарного классификатора  $\text{sign } f(x) = \text{sign}(w^T x + b)$ :
  - Направление:  $-\text{sign } f(x_0) \frac{w}{\|w\|_2}$
  - Длина:  $\frac{|f(x_0)|}{\|w\|_2}$
  - $\Rightarrow$  Атака:  $r = -\frac{f(x_0)}{\|w\|_2^2} w$
- В случае нелинейной разделяющей поверхности  $f(x)$ :
  - применяем формулу Тейлора:  $f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0)$
  - и подставляем в формулу для  $r$  выражение  $w = \nabla f(x_0)$



# Метод атаки: DeepFool

- Итеративный алгоритм DeepFool<sup>21</sup> для произвольного классификатора

---

### Algorithm 1 DeepFool for binary classifiers

---

```
1: input: Image  $x$ , classifier  $f$ .  
2: output: Perturbation  $\hat{r}$ .  
3: Initialize  $x_0 \leftarrow x$ ,  $i \leftarrow 0$ .  
4: while  $\text{sign}(f(x_i)) = \text{sign}(f(x_0))$  do  
5:    $r_i \leftarrow -\frac{f(x_i)}{\|\nabla f(x_i)\|_2^2} \nabla f(x_i)$ ,  
6:    $x_{i+1} \leftarrow x_i + r_i$ ,  
7:    $i \leftarrow i + 1$ .  
8: end while  
9: return  $\hat{r} = \sum_i r_i$ .
```

---

- Существует естественное обобщение на случай многоклассового классификатора

---

<sup>21</sup>Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." 2015 

- JSMA<sup>22</sup> (**Jacobian-based Saliency Map Attack**) — одна из первых  $\ell_0$ -атак, когда важно количество задействованных в атаке пикселей, а не их значения

<sup>22</sup>Papernot, Nicolas, et al. “The limitations of deep learning in adversarial settings.” 2015



## Метод атаки: JSMA

- JSMA<sup>22</sup> (**Jacobian-based Saliency Map Attack**) — одна из первых  $\ell_0$ -атак, когда важно количество задействованных в атаке пикселей, а не их значения
- Идея: менять те пиксели, которые дают максимальный вклад в производную по входу для нужного класса для направленной атаки

<sup>22</sup>Papernot, Nicolas, et al. “The limitations of deep learning in adversarial settings.” 2015



## Метод атаки: JSMA

- JSMA<sup>22</sup> (**Jacobian-based Saliency Map Attack**) — одна из первых  $\ell_0$ -атак, когда важно количество задействованных в атаке пикселей, а не их значения
- Идея: менять те пиксели, которые дают максимальный вклад в производную по входу для нужного класса для направленной атаки
- Можно делать это итеративно, постепенно добавляя пиксели в область атаки  $r$

<sup>22</sup>Papernot, Nicolas, et al. “The limitations of deep learning in adversarial settings.” 2015



# Метод атаки: JSMA

- JSMA<sup>22</sup> (**Jacobian-based Saliency Map Attack**) — одна из первых  $\ell_0$ -атак, когда важно количество задействованных в атаке пикселей, а не их значения
- **Идея:** менять те пиксели, которые дают максимальный вклад в производную по входу для нужного класса для направленной атаки
- Можно делать это итеративно, постепенно добавляя пиксели в область атаки  $r$
- **Замечание:**  $F(x)$  — выход SoftMax слоя, пиксели добавляются парами (так проще)

---

<sup>22</sup>Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." 2015



# Метод атаки: JSMA

- JSMA<sup>22</sup> (**Jacobian-based Saliency Map Attack**) — одна из первых  $\ell_0$ -атак, когда важно количество задействованных в атаке пикселей, а не их значения
- **Идея:** менять те пиксели, которые дают максимальный вклад в производную по входу для нужного класса для направленной атаки
- Можно делать это итеративно, постепенно добавляя пиксели в область атаки  $r$
- **Замечание:**  $F(x)$  — выход SoftMax слоя, пиксели добавляются парами (так проще)

---

**Algorithm 3 Increasing pixel intensities saliency map**  
 $\nabla F(\mathbf{X})$  is the forward derivative,  $\Gamma$  the features still in the search space, and  $t$  the target class

**Input:**  $\nabla F(\mathbf{X})$ ,  $\Gamma$ ,  $t$

```
1: for each pair  $(p, q) \in \Gamma$  do
2:    $\alpha = \sum_{i=p,q} \frac{\partial F_t(\mathbf{X})}{\partial \mathbf{X}_i}$ 
3:    $\beta = \sum_{i=p,q} \sum_{j \neq t} \frac{\partial F_j(\mathbf{X})}{\partial \mathbf{X}_i}$ 
4:   if  $\alpha > 0$  and  $\beta < 0$  and  $-\alpha \times \beta > \max$  then
5:      $p_1, p_2 \leftarrow p, q$ 
6:      $\max \leftarrow -\alpha \times \beta$ 
7:   end if
8: end for
9: return  $p_1, p_2$ 
```

---

<sup>22</sup>Papernot, Nicolas, et al. “The limitations of deep learning in adversarial settings.” 2015



# Метод атаки: One pixel

- Однопиксельная атака<sup>23</sup> — предельный случай  $\ell_0$ -атаки

---

<sup>23</sup>Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." 2017

<sup>24</sup>Storn, Rainer, and Kenneth Price. "Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces." 1997



## Метод атаки: One pixel

- Однопиксельная атака<sup>23</sup> — предельный случай  $\ell_0$ -атаки
- Идея: применить эволюционный алгоритм (дифференциальной эволюции<sup>24</sup>)

---

<sup>23</sup>Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." 2017

<sup>24</sup>Storn, Rainer, and Kenneth Price. "Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces." 1997



# Метод атаки: One pixel

- Однопиксельная атака<sup>23</sup> — предельный случай  $\ell_0$ -атаки
- Идея: применить эволюционный алгоритм (дифференциальной эволюции<sup>24</sup>)
- Популяция состоит из 400 экземпляров, каждый из которых задается пятеркой: две координаты и три канала цвета

---

<sup>23</sup>Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." 2017

<sup>24</sup>Storn, Rainer, and Kenneth Price. "Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces." 1997



# Метод атаки: One pixel

- Однопиксельная атака<sup>23</sup> — предельный случай  $\ell_0$ -атаки
- Идея: применить эволюционный алгоритм (дифференциальной эволюции<sup>24</sup>)
- Популяция состоит из 400 экземпляров, каждый из которых задается пятеркой: две координаты и три канала цвета
- Генерация потомка — линейная комбинация трех случайных родителей

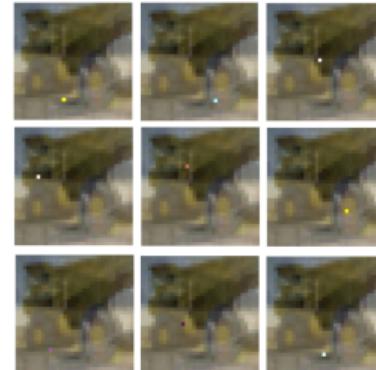
---

<sup>23</sup>Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." 2017

<sup>24</sup>Storn, Rainer, and Kenneth Price. "Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces." 1997

# Метод атаки: One pixel

- Однопиксельная атака<sup>23</sup> — предельный случай  $\ell_0$ -атаки
- Идея: применить эволюционный алгоритм (дифференциальной эволюции<sup>24</sup>)
- Популяция состоит из 400 экземпляров, каждый из которых задается пятеркой: две координаты и три канала цвета
- Генерация потомка — линейная комбинация трех случайных родителей



<sup>23</sup>Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." 2017

<sup>24</sup>Storn, Rainer, and Kenneth Price. "Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces." 1997

# Метод универсальной атаки

- До этого все атаки строились как функция от входа  $x$

<sup>25</sup> Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." 2016



# Метод универсальной атаки

- До этого все атаки строились как функция от входа  $x$
- Однако можно строить т.н. “универсальную” атаку<sup>25</sup>, которая будет уже функцией от всего обучающего множества  $X$

<sup>25</sup> Moosavi-Dezfooli, Seyed-Mohsen, et al. “Universal adversarial perturbations.” 2016



# Метод универсальной атаки

- До этого все атаки строились как функция от входа  $x$
- Однако можно строить т.н. “универсальную” атаку<sup>25</sup>, которая будет уже функцией от всего обучающего множества  $X$
- При построении атаки будем искать  $r$ , примерно одинаково ломающий все классы из  $X$

<sup>25</sup> Moosavi-Dezfooli, Seyed-Mohsen, et al. “Universal adversarial perturbations.” 2016

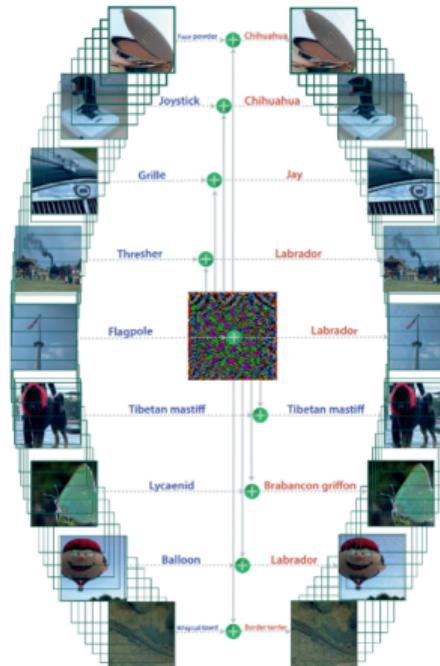
# Метод универсальной атаки

- До этого все атаки строились как функция от входа  $x$
- Однако можно строить т.н. “универсальную” атаку<sup>25</sup>, которая будет уже функцией от всего обучающего множества  $X$
- При построении атаки будем искать  $r$ , примерно одинаково ломающий все классы из  $X$
- Справа — универсальное возмущение для любого входа, которое ломает классификатор

<sup>25</sup> Moosavi-Dezfooli, Seyed-Mohsen, et al. “Universal adversarial perturbations.” 2016

# Метод универсальной атаки

- До этого все атаки строились как функция от входа  $x$
- Однако можно строить т.н. “универсальную” атаку<sup>25</sup>, которая будет уже функцией от всего обучающего множества  $X$
- При построении атаки будем искать  $r$ , примерно одинаково ломающий все классы из  $X$
- Справа — универсальное возмущение для любого входа, которое ломает классификатор



<sup>25</sup> Moosavi-Dezfooli, Seyed-Mohsen, et al. “Universal adversarial perturbations.” 2016

## Физические атаки

- Все атаки до этого работали в т.н. цифровой области (digital domain): изменяли картинку на уровне пикселей

<sup>26</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



# Физические атаки

- Все атаки до этого работали в т.н. цифровой области (digital domain): изменяли картинку на уровне пикселей
- Если нет возможности проатаковать изображение непосредственно перед подачей в СНС, то такая атака бесполезна

<sup>26</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



# Физические атаки

- Все атаки до этого работали в т.н. цифровой области (digital domain): изменяли картинку на уровне пикселей
- Если нет возможности проатаковать изображение непосредственно перед подачей в СНС, то такая атака бесполезна
- Поэтому атаки в реальном мире (real-world), или физические атаки, наиболее универсальны

<sup>26</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



- Все атаки до этого работали в т.н. цифровой области (digital domain): изменяли картинку на уровне пикселей
- Если нет возможности проатаковать изображение непосредственно перед подачей в СНС, то такая атака бесполезна
- Поэтому атаки в реальном мире (real-world), или физические атаки, наиболее универсальны
- Первый пример физической атаки<sup>26</sup> — атака на изображение в цифровой области, затем печать на физическом носителе (бумага), затем снимок цифровой камерой и последующая обработка СНС

<sup>26</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



# Физические атаки

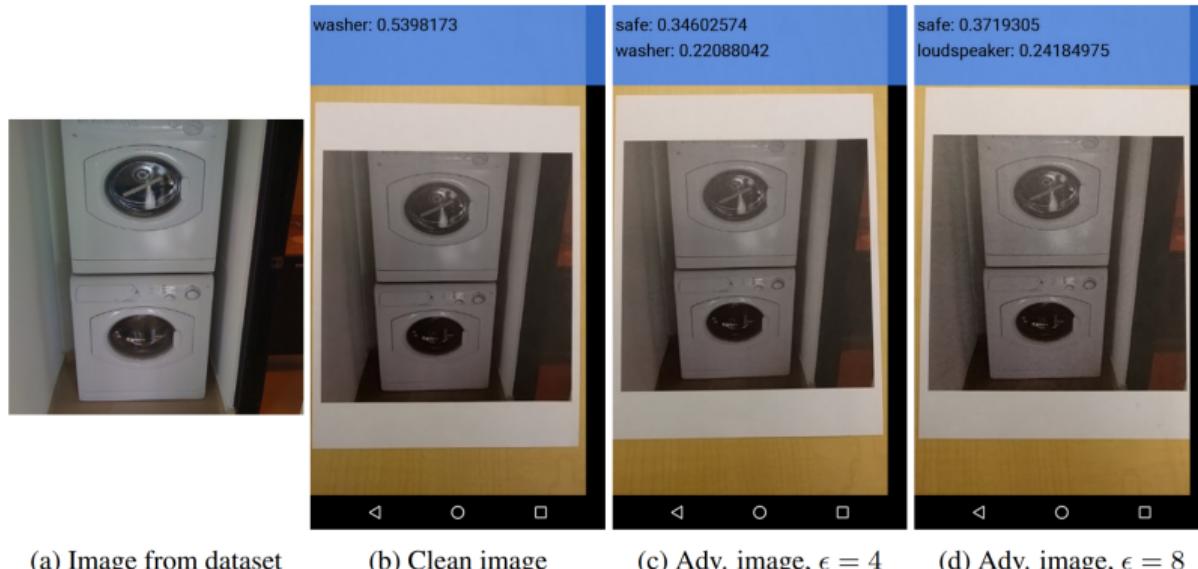
- Все атаки до этого работали в т.н. цифровой области (digital domain): изменяли картинку на уровне пикселей
- Если нет возможности проатаковать изображение непосредственно перед подачей в СНС, то такая атака бесполезна
- Поэтому атаки в реальном мире (real-world), или физические атаки, наиболее универсальны
- Первый пример физической атаки<sup>26</sup> — атака на изображение в цифровой области, затем печать на физическом носителе (бумага), затем снимок цифровой камерой и последующая обработка СНС
- Никакой специальной технологии для генерации таких атак еще не было, просто была показана их возможность

---

<sup>26</sup>Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." 2016



# Физические атаки



## Физические атаки: EOT

- Подход EOT<sup>27</sup> (**E**xpectation **O**ver **T**ransformation) учитывает, что объект в реальном мире обычно претерпевает ряд преобразований таких как:
  - Масштабирование
  - Трансляция (тряска)
  - Изменение яркости и/или контрастности

<sup>27</sup>Athalye, Anish, et al. "Synthesizing robust adversarial examples." 2017

# Физические атаки: EOT

- Подход EOT<sup>27</sup> (**E**xpectation **O**ver **T**ransformation) учитывает, что объект в реальном мире обычно претерпевает ряд преобразований таких как:
  - Масштабирование
  - Трансляция (тряска)
  - Изменение яркости и/или контрастности
- Поэтому задача — найти (направленную) состязательную атаку  $r$  с учетом множества преобразований  $T$ :

## EOT

Найти  $\arg \max_r \mathbb{E}_{g \sim T} P(y_t | g(x + r))$  при условии:

- ①  $f(x) = y_{gt} \neq y_t$
- ②  $\mathbb{E}_{g \sim T} \|g(x + r) - g(x)\|_p < \epsilon$
- ③  $x \in B$

<sup>27</sup>Athalye, Anish, et al. "Synthesizing robust adversarial examples." 2017

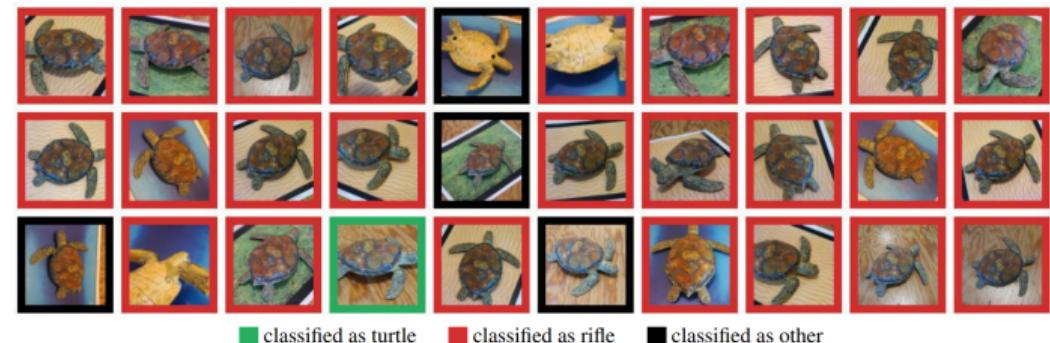
## Физические атаки: EOT

- В итоге, используя широкий ряд преобразований  $T$ , удалось сделать состязательный 3D-пример

# Физические атаки: EOT

- В итоге, используя широкий ряд преобразований  $T$ , удалось сделать состязательный 3D-пример

Transformation	Minimum	Maximum
Camera distance	2.5	3.0
X/Y translation	-0.05	0.05
Rotation	any	
Background	(0.1, 0.1, 0.1)	(1.0, 1.0, 1.0)
Lighten / Darken (additive)	-0.15	0.15
Lighten / Darken (multiplicative)	0.5	2.0
Per-channel (additive)	-0.15	0.15
Per-channel (multiplicative)	0.7	1.3
Gaussian Noise (stdev)	0.0	0.1



## Еще примеры физических атак

- Интересны примеры атак на объекты ImageNet<sup>28</sup>, дорожные знаки<sup>29</sup> и даже системы распознавания лиц<sup>30</sup>

---

<sup>28</sup>Brown, Tom B., et al. "Adversarial patch." 2017

<sup>29</sup>Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning models." 2017

<sup>30</sup>Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." 2016

## Еще примеры физических атак

- Интересны примеры атак на объекты ImageNet<sup>28</sup>, дорожные знаки<sup>29</sup> и даже системы распознавания лиц<sup>30</sup>
- Примечательно, что все эти атаки по существу  $\ell_0$ -атаки, а также используют NPS и TV-добавки в функцию потерь
  - NPS (Non Printability Score): штраф за использование цветов, которые не может воспроизвести данный принтер
  - TV (Total Variation): штраф за негладкость картинки

$$TV(x) = \sum_{i,j} \sqrt{(x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2}$$

---

<sup>28</sup>Brown, Tom B., et al. "Adversarial patch." 2017

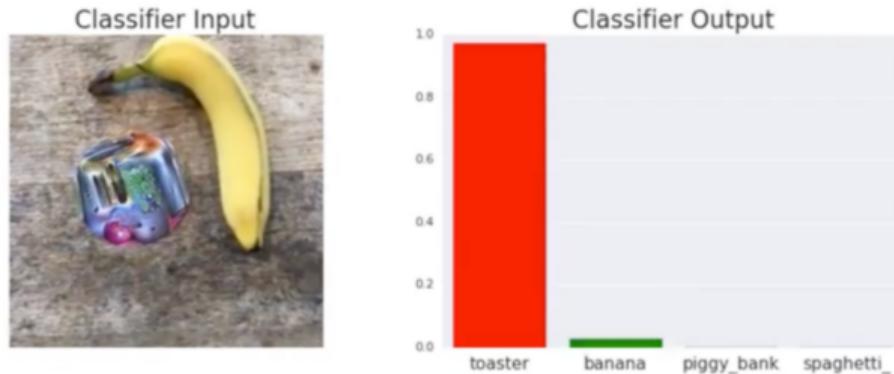
<sup>29</sup>Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning models." 2017

<sup>30</sup>Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." 2016

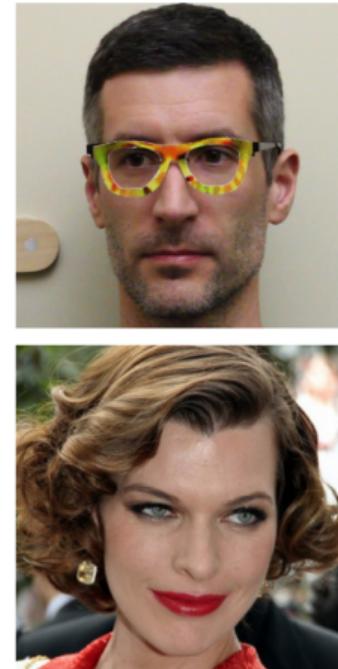


# Еще примеры физических атак

Атака на объекты ImageNet:



Атака на FaceID:



Атака на дорожные знаки:



# Атаки на ведущую систему распознавания лиц

- Обычно система распознавания содержит два важных элемента: детектор и извлекатель признаков (часто называемый FaceID)

---

<sup>31</sup>Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." 2016

<sup>32</sup>Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



# Атаки на ведущую систему распознавания лиц

- Обычно система распознавания содержит два важных элемента: детектор и извлекатель признаков (часто называемый FaceID)
- Использовались: крайне легкий нейросетевой детектор MTCNN<sup>31</sup> и ведущая открытая система извлечения признаков ArcFace<sup>32</sup>

---

<sup>31</sup> Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." 2016

<sup>32</sup> Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



# Атаки на ведущую систему распознавания лиц

- Обычно система распознавания содержит два важных элемента: детектор и извлекатель признаков (часто называемый FaceID)
- Использовались: крайне легкий нейросетевой детектор MTCNN<sup>31</sup> и ведущая открытая система извлечения признаков ArcFace<sup>32</sup>
- Атаки на FaceID: с цветным патчем и черно-белым

---

<sup>31</sup> Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." 2016

<sup>32</sup> Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



# Атаки на ведущую систему распознавания лиц

- Обычно система распознавания содержит два важных элемента: детектор и извлекатель признаков (часто называемый FaceID)
- Использовались: крайне легкий нейросетевой детектор MTCNN<sup>31</sup> и ведущая открытая система извлечения признаков ArcFace<sup>32</sup>
- Атаки на FaceID: с цветным патчем и черно-белым
- Атака на детектор: маска и черно-белый патч

---

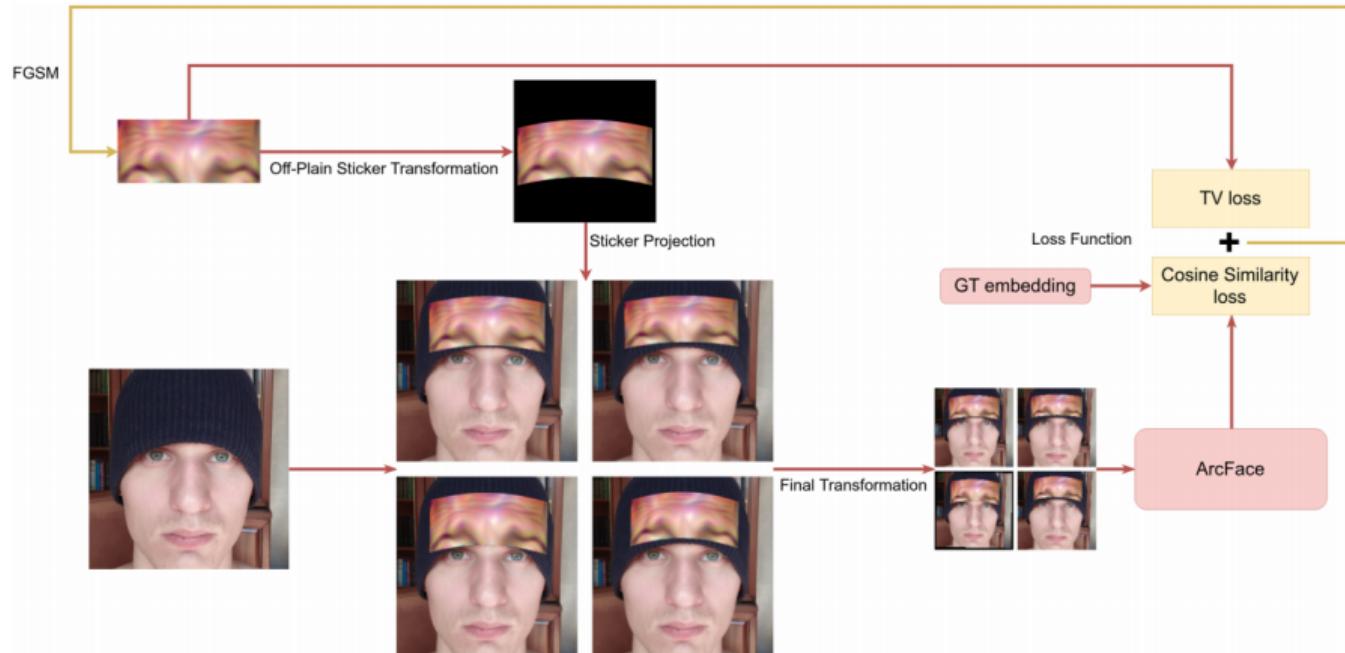
<sup>31</sup> Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." 2016

<sup>32</sup> Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." 2018



# AdvHat<sup>33</sup> — шапка-невидимка

Алгоритм обучения:



<sup>33</sup>Komkov, Stepan, and Aleksandr Petiushko. "AdvHat: Real-world adversarial attack on ArcFace Face ID" 2019  
system."

Устойчивость к поворотам и разной освещенности<sup>34</sup>:

**Фронтальное лицо  
(нет атаки)**

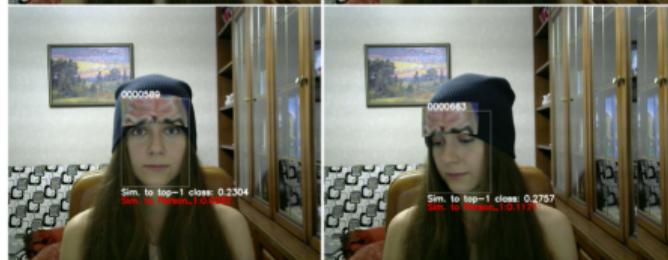
Близость до своего эталона: **0.61**



**Фронтальное лицо  
(атака)**

Близость до своего эталона: **0.02**

Близость до другого эталона: **0.23**



**Поворот лица  
(нет атаки)**

Близость до своего эталона: **0.54**

**Поворот лица  
(атака)**

Близость до своего эталона: **0.11**

Близость до другого эталона: **0.27**

<sup>34</sup><https://www.youtube.com/watch?v=a4iNg0wWBsQ>

# Adversarial patches<sup>35</sup> — черно-белые патчи

Дальнейшее развитие атак на FaceID:



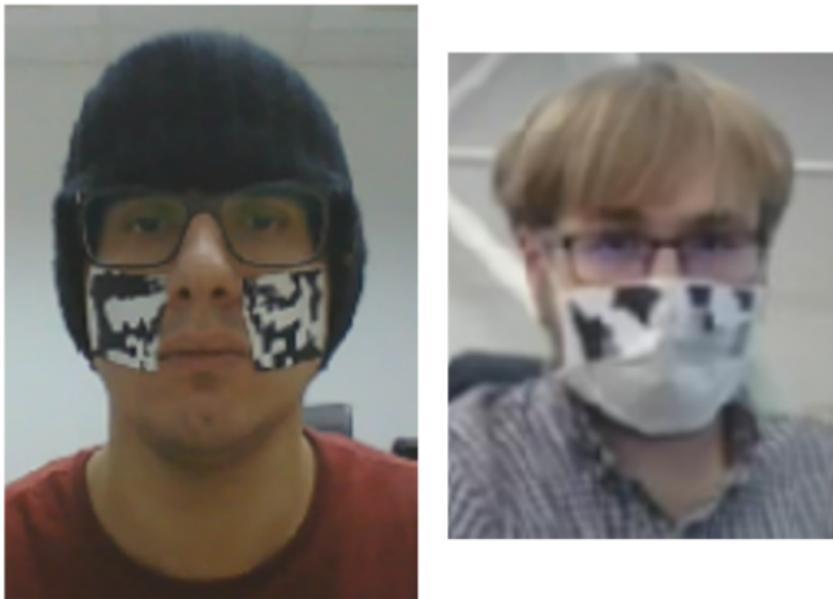
---

<sup>35</sup>Pautov, Mikhail, et al. "On adversarial patches: real-world attack on ArcFace-100 face recognition system" 2019



# Атака на детектор лиц<sup>37</sup> — черно-белые патчи

Физическая атака на неглубокий и поэтому крайне устойчивый к состязательным атакам детектор MTCNN<sup>36</sup>:



<sup>36</sup><https://www.youtube.com/watch?v=0Y700IS8bxs>

<sup>37</sup>Kaziakhmedov, Edgar, et al. "Real-world attack on MTCNN face detection system." 2019



## Заключительные выводы

- На данный момент СНС (в целом) работают лучше человека

<sup>38</sup>Image credit: <http://reddit.com>

## Заключительные выводы

- На данный момент СНС (в целом) работают лучше человека
- СНС легко “обмануть” используя их неустойчивость по входу

<sup>38</sup>Image credit: <http://reddit.com>

## Заключительные выводы

- На данный момент СНС (в целом) работают лучше человека
- СНС легко “обмануть” используя их неустойчивость по входу
- Наиболее распространенный прием атаки — производная по входу

<sup>38</sup>Image credit: <http://reddit.com>

## Заключительные выводы

- На данный момент СНС (в целом) работают лучше человека
- СНС легко “обмануть” используя их неустойчивость по входу
- Наиболее распространенный прием атаки — производная по входу
- Перенести атаку в реальный мир непросто

<sup>38</sup>Image credit: <http://reddit.com>

## Заключительные выводы

- На данный момент СНС (в целом) работают лучше человека
- СНС легко “обмануть” используя их неустойчивость по входу
- Наиболее распространенный прием атаки — производная по входу
- Перенести атаку в реальный мир непросто
- Однако можно сломать даже супер навороченные системы распознавания лиц, имея лишь обычный принтер

<sup>38</sup>Image credit: <http://reddit.com>

# Заключительные выводы

- На данный момент СНС (в целом) работают лучше человека
- СНС легко “обмануть” используя их неустойчивость по входу
- Наиболее распространенный прием атаки — производная по входу
- Перенести атаку в реальный мир непросто
- Однако можно сломать даже супер навороченные системы распознавания лиц, имея лишь обычный принтер
- Для человечества пока еще не все потеряно<sup>38</sup>!



<sup>38</sup>Image credit: <http://reddit.com>

Спасибо за внимание!