Исследование устойчивости сверточных нейросетей

на примере систем детекции и распознавания лиц

Петюшко Александр

МГУ им. М.В.Ломоносова, к.ф.-м.н.
Huawei, Video Intelligence Team Leader

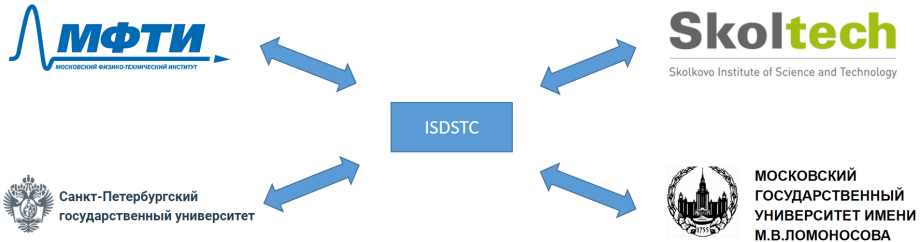
4 февраля, 2021

Plan

1. Лаборатория Интеллектуальных Систем и Науки о Данных
2. Потрясающие успехи CHC в компьютерном зрении
3. (Не) устойчивость CHC в компьютерном зрении
4. Методы для генерации состязательных примеров в цифровой области
5. ℓ_0-состязательные примеры
6. Методы для генерации состязательных примеров в реальном мире
7. Состязательные примеры для систем детекции лиц
8. Состязательные примеры для систем распознавания лиц
9. Защита от состязательных примеров в реальном мире
10. Black-box восстановление лица по вектору признаков

Научное сотрудничество: Лаборатория Интеллектуальных Систем и Науки о Данных

Российский исследовательский институт → Московский исследовательский институт → Лаборатория Интеллектуальных Систем и Науки о Данных



Человек или СНС?

ImageNet[1] (1000-классовая база данных изображений)

- Top-5 ошибка для человека[2]: 5.1%
- Top-5 ошибка для СНС[3]: 2.0%

Labeled Faces in the Wild[4] (база данных лиц)

- Ошибка верификации для человека[5]: 2.47%
- Ошибка верификации для СНС[6]: 0.17%

LFW

[1] http://www.image-net.org/

[2] http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

[3] Touvron, Hugo, et al. "Fixing the train-test resolution discrepancy." 2019

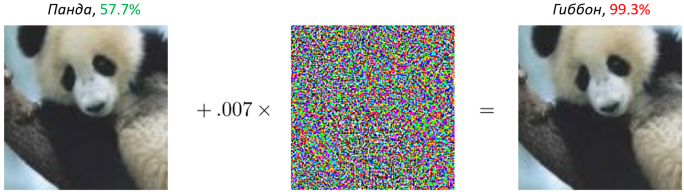
[4] http://vis-www.cs.umass.edu/lfw/

[5] Kumar, Neeraj, et al. "Attribute and simile classifiers for face verification." 2009

[6] 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
- Например, результат классификации с “панды” поменяется на “гиббона”[7]

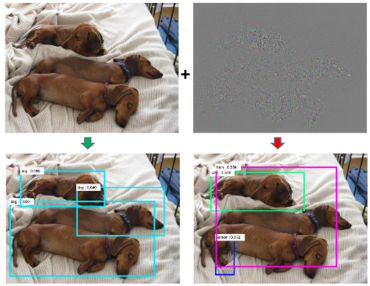


Такое возмущение называется **состязательным примером** (adversarial attack)

[7] Image credit: https://arxiv.org/pdf/1412.6572.pdf

Состязательные примеры в разных задачах

СНС для обнаружения и сегментации изображений[8]:



И даже НС для вопросно-ответных систем (QA, question answering systems)[9]:

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

[8] Xie C. et al. "Adversarial examples for semantic segmentation and object detection." 2017
[9] Jia R. et al. "Adversarial examples for evaluating reading comprehension systems." 2017

Состязательные примеры: 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 — правильный класс для x
- θ — параметры CHC-классификатора
- $L(\theta, x, y)$ — функция потерь
- $f(x)$ — выход классификатора (распознанный класс); при обучении мы добиваемся равенства $f(x) = y$
- $r \in B = [0,1]^{C \times M \times N}$ — аддитивная добавка ко входу x

Состязательный пример и устойчивость: формулировки

Цель состязательного примера

Поменять выход классификатора f на неправильный путем добавления минимального (по некоторой норме ℓ_p) возмущения r:

1. $||r||_p \to \min$
2. $f(x) = y$ (изначальный ответ СНС верный)
3. $f(x + r) \neq y$ (меняем выход с помощью возмущения r)
4. $x + r \in B$ (остаемся во множестве допустимых изображений)

Устойчивость классификатора

Найти такой класс возмущения $S(x, f) \subseteq B$, при котором классификатор не меняет свой выход:

$$f(x + r) = f(x) = y \quad \forall r \in S(x, f)$$

Состязательные примеры в цифровой области: FGSM, итеративные методы

В основном состязательные примеры для СНС формулируются в терминах ℓ_∞-нормы (что соответствует процессу восприятия человеческим глазом визуальной информации)

$$||x||_\infty = \max_i |x_i|, x = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

- **F**ast **G**radient **S**ign **M**ethod[10] (FGSM): $r = \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y))$
- Итеративный FGSM (I-FGSM)[11] / **P**rojected **G**radient **D**escent (PGD)[12] (где Π_B — операция проекции на B):
  $x^{t+1} = \Pi_B(x^t + \alpha \cdot \text{sign}\, \nabla_x L(\theta, x^t, y)), \quad x^0 = x, \alpha = \epsilon / T, t \leq T$
- Сглаживание градиента для I-FGSM (MI-FGSM):
  $g^{t+1} = \mu \cdot g^t + \frac{\nabla_x L(\theta, x^t, y)}{||\nabla_x L(\theta, x^t, y)||_1}, x^{t+1} = \Pi_B(x^t + \alpha \cdot \text{sign}(g^{t+1})), \quad x^0 = x, g^0 = 0$

[10] Goodfellow I. et al. "Explaining and harnessing adversarial examples." 2014

[11] Kurakin A. et al. "Adversarial examples in the physical world." 2016

[12] Madry A. et al. "Towards deep learning models resistant to adversarial attacks." 2017

Сравнение состязательных примеров на основе FGSM

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

Как видно, MI-FGSM — наиболее успешная методика генерации.

ℓ_0-состязательные примеры

- ℓ_∞-состязательные примеры “заточены” под визуальное вприятие, но нужно обычно менять все пиксели
- В реальном мире это нереалистично — мы можем менять только часть сцены
- ℓ_0-состязательные примеры более приспособлены для данной задачи: $||x||_0 = \sum_{i=1}^{n} \mathbf{1}_{x_i \neq 0}$
- **J**acobian-based **S**aliency **M**ap **A**ttack (JSMA)[13] и наиболее экстремальный случай — One Pixel[14] — примеры подобных ℓ_0-состязательных примеров, где минимизируется количество пикселей для изменения

[13] Papernot N. et al. “The limitations of deep learning in adversarial settings.” 2015
[14] Su J. et al. “One pixel attack for fooling deep neural networks.” 2017

Физические состязательные примеры: EOT

- Не имеем доступа к фото (и его пикселям) $\Rightarrow$ единственная возможность — это изменить внешний вид самого объекта
- Подход **E**xpectation **O**ver **T**ransformation (EOT)[15] учитывает, что объект в реальном мире обычно претерпевает ряд преобразований таких как:
  - Масштабирование
  - Трансляции и повороты
  - Изменение яркости и/или контрастности, шум и т.п.
- Т.о. для объекта x нужно найти состязательный пример r с учетом преобразований $g \in T$:

EOT

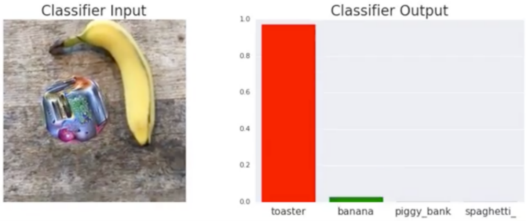
Найти $\arg\min_r \mathbb{E}_{g\sim T}[P(y|g(x+r))]$ при условии:

1. $\mathbb{E}_{g\sim T}[d(g(x+r), g(x))] < \epsilon$, где $d(a,b)$ – некоторая функция расстояние (например, $d(a,b) = ||a-b||_p$)
2. $x + r \in B$

[15] Athalye A. et al. "Synthesizing robust adversarial examples." 2017

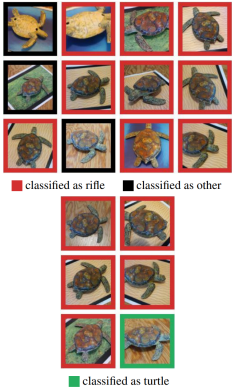
Физические состязательные примеры

Для ImageNet[16]:



Для дорожных знаков[17]:

3D-состязательные примеры:



[16] Brown T. et al. "Adversarial patch." 2017

[17] Eykholt K. et al. "Robust physical-world attacks on deep learning models." 2017

Физические состязательные примеры: основные ингридиенты

- ℓ_0 оптимизация (маска) + EOT: обязательно
- **T**otal **V**ariation (TV) функция потерь — штраф за негладкость примера (в реальном мире мало больших попиксельных градиентов):

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

- **N**on-**P**rintability **S**core (NPS) — штраф за использование цветов, которые не поддерживаются (принтером). Если $G \subset [0,1]^3$ — поддерживаемая цветовая палитра (gamut), то штраф за использование цвета пикселя $q_0 \in [0,1]^3$:

$$NPS(q_0) = \Pi_{q \in G} ||q - q_0||_2$$

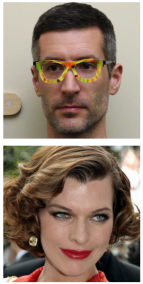
- Дополнительный маппинг цветов (например, принтер печатает не цвет c, а близкий $m(c)$, и это нужно учитывать)

Состязательные примеры для систем детекции и распознавания лиц

- Изначально т.н. **Camouflage Art**[18] использовался для обхода лидирующей системы детекции лиц Viola-Jones
- Это был вручную подобранный грим для обхода детектора Хаара
- Прорыв случился с работой Sharif et al.[19], где предложили использовать состязательные очки
- Использовано: ℓ_0-оптимизация + EOT + TV + NPS + цветовой маппинг
- Минусы: предобученные лица, системы распознавания лиц прошлого поколения

cvdazzle.com

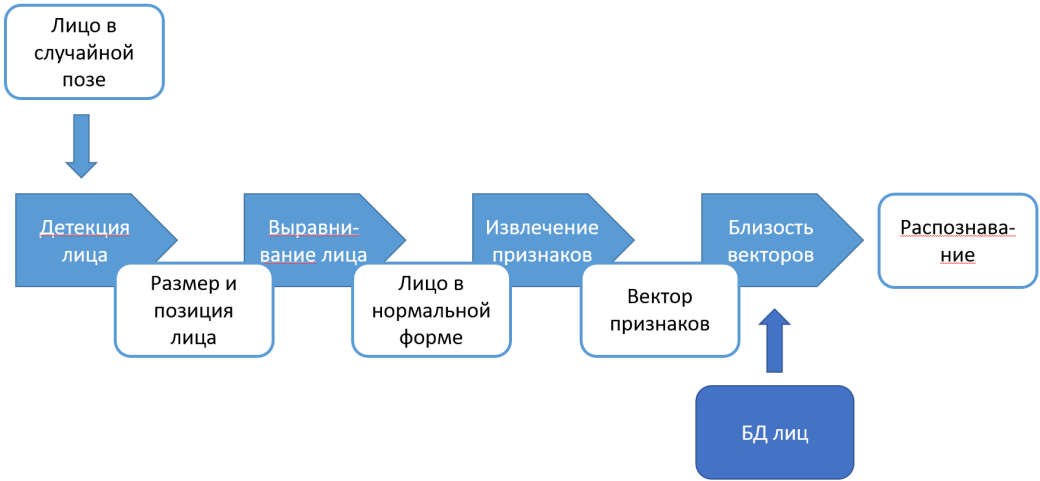
Состязательные очки



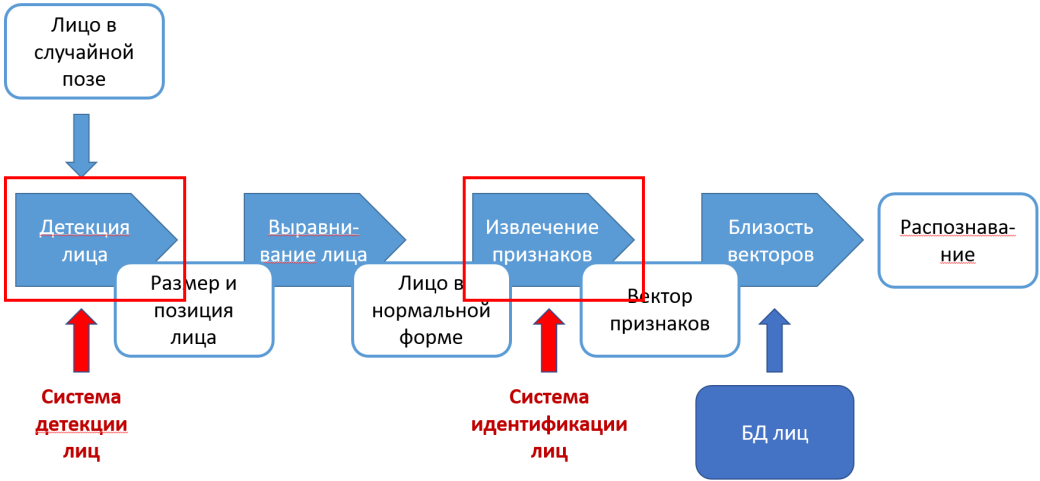
[18] Feng R. et al. "Facilitating fashion camouflage art." 2013

[19] Sharif M. et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." 2016

Общая схема детекции и распознавания лиц



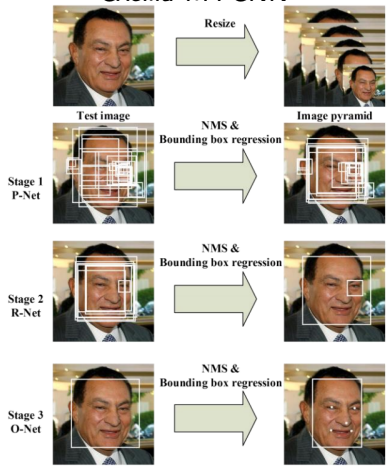
Общая схема детекции и распознавания лиц



Детектор лиц: MTCNN[20]

- В отличие от современных глубоких детекторов типа Faster RCNN или YOLO, MTCNN очень простой и неглубокий ⇒ меньше поле восприятия, сложнее поменять решение детектора
- В MTCNN каскадный подход: сначала грубое приближение (P-Net), а затем исправление (R-Net, O-Net)
- Решение: использовать для генерации состязательных примеров самый первый классификационный слой в P-Net (а не функции потерь для прямоугольников или ключевых точек)

Схема MTCNN

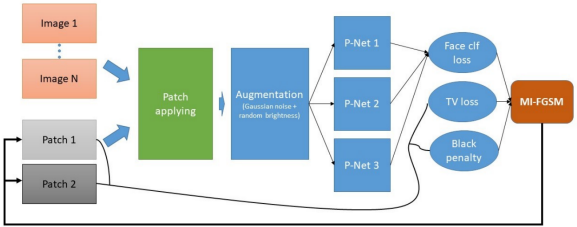


[20] Zhang K. et al. "Joint face detection and alignment using multitask cascaded convolutional networks." 2016

Состязательные примеры для детектора лиц MTCNN

- EOT: Гауссов шум, размер маски, яркость, набор разных фото лица
- TV: +, NPS: –
- Маппинг цветов: штрафуем за близость к черному цвету ($x_{i,j} = 1$) $\Rightarrow$ новая добавка в функцию потерь: $L_{BLK}(x) = \sum_{i,j} x_{i,j}$
- Оптимизатор: MI-FGSM

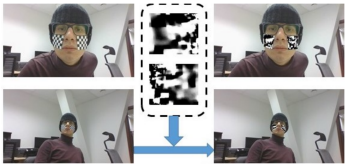
Схема генерациии состязательных примеров для MTCNN



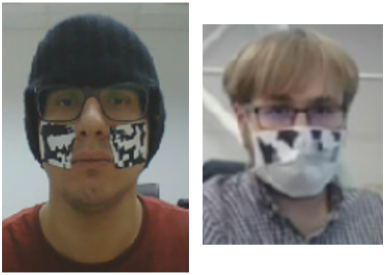
Состязательные примеры для детектора лиц MTCNN

- ℓ_0-оптимизация: 2 версии
  1. две раздельных маски на щеках
  2. цельная медицинская маска
- У MTCNN маленькое поле восприятия $\Rightarrow$ примеры не носят семантический характер
- Оценка параметров локальных аффинных проекций на основе заранее подготовленной специальной сетки для обучения

Проекция

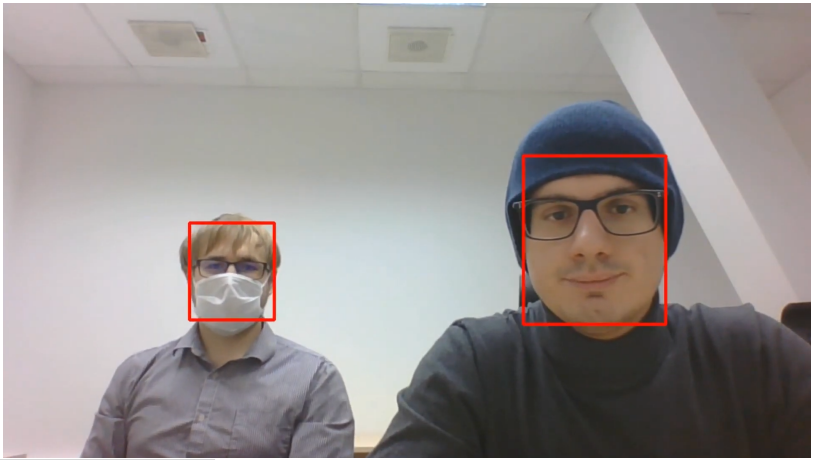


Примеры



Состязательные примеры для детектора лиц MTCNN: результат

Детали: статья[21] (IEEE-2019) и видео-демонстрация[22].

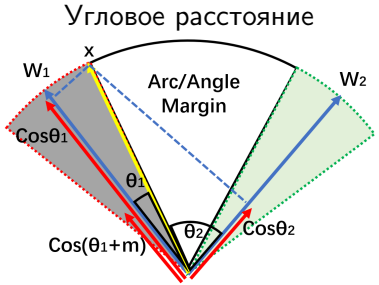


[21] Kaziakhmedov E. et al. "Real-world attack on MTCNN face detection system." 2019

[22] https://www.youtube.com/watch?v=0Y70OIS8bxs

Система распознавания лиц: ArcFace[23]

- Выбрана ведущая открытая система распознавания лиц признаков: ArcFace
- Главная идея ArcFace — использовать угловое расстояние между векторами признаков
- Огоромная обучающая база данных (MS1M) и глубокая СНС (ResNet-100)

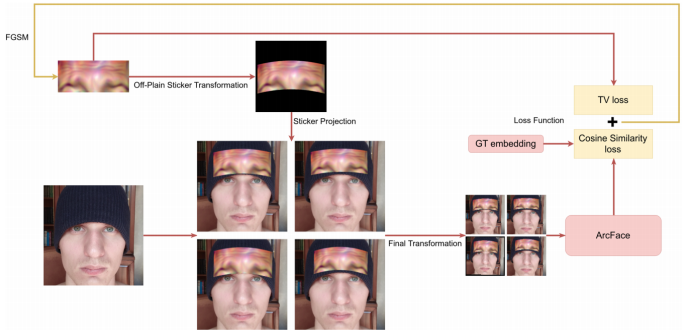


[23] Deng J. et al. "Arcface: Additive angular margin loss for deep face recognition." 2018

Состязательные примеры для распознавателя лиц ArcFace

- EOT: различные параметры проекций маски, одно изображение лица
- TV: +, NPS: –, цветовой маппинг: –
- Добавка к функции потерь $L_{sim}(x, x_{gt}) = cos(emb(x), emb(x_{gt}))$ для работы с любым лицом, где x_{gt} — фото лица, $emb(x)$ — вектор признаков x
- Оптимизатор: MI-FGSM

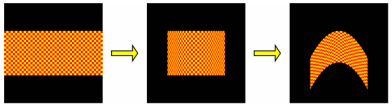
Схема генерациии состязательных примеров для ArcFace



Состязательные примеры для распознавателя лиц ArcFace

- ℓ_0-оптимизация: цветная связная область (патч) в области лба
- Глубокая СНС $\Rightarrow$ большое поле восприятия $\Rightarrow$ семантический характер примера
- Т.н. “off-plane” нелинейная проекция: $(x, y, 0) \to (x', y, z'), z' = a \cdot x'^2$

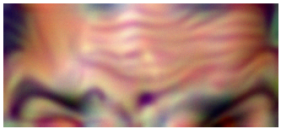
Off-plane проекция



- Дифференцируемый слой Spatial Transformer Layer[24]

$$x' = a \cdot \left(|x| \cdot \sqrt{x^2 + \frac{1}{4 \cdot a^2}} + \frac{1}{4 \cdot a^2} \cdot \ln \left(|x| + \sqrt{x^2 + \frac{1}{4 \cdot a^2}} \right) - \frac{1}{4 \cdot a^2} \cdot \ln \left(\frac{1}{2 \cdot a} \right) \right)$$

Семантический характер примера



[24] Jaderberg M. et al. “Spatial transformer networks.” 2015

AdvHat — шапка-невидимка

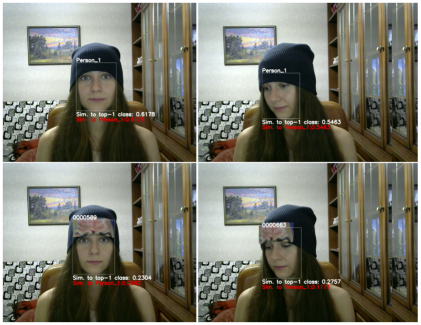
Благодаря улучшенной процедуре проекции и полного использования цветовой палитры, состязательный пример устойчив к разным поворотам и освещенности

Анфас
(AdvHat: –)
Близость (оригинал): **0.61**

|

|

Анфас
(AdvHat: +)
Близость (оригинал): **0.02**
Близость (другой): **0.23**



Профиль
(AdvHat: –)
Близость (оригинал): **0.54**

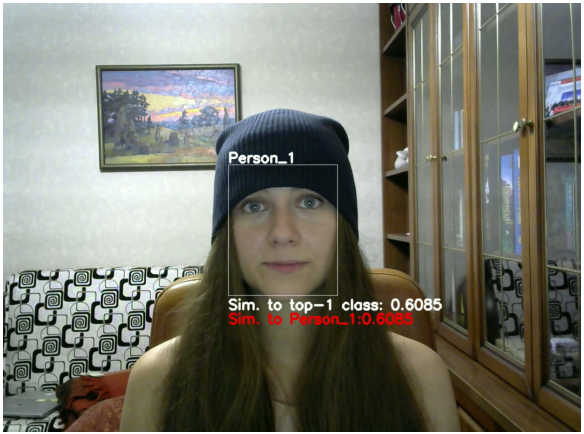
|

|

Профиль
(AdvHat: +)
Близость (оригинал): **0.11**
Близость (другой): **0.27**

Состязательные примеры для распознавателя лиц ArcFace: результат

Детали: статья[25] (ICPR-2020) и видео-демонстрация[26].



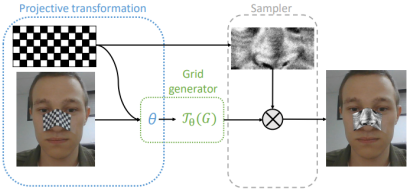
[25] Komkov S. et al. "Advhat: Real-world adversarial attack on arcface face id system." 2019

[26] https://www.youtube.com/watch?v=a4iNg0wWBsQ

Состязательные примеры для распознавателя лиц ArcFace: черно-белые патчи[27] (IEEE-2019)

- Комбинация двех предыдущих подходов:
  - Цветовой маппинг (штраф за черный цвет)
  - Локальные аффинные проекции на основе подготовленной сетки
- Пример носит семантический характер

Проекция



Состязательные примеры

[27] Pautov M. et al. "On adversarial patches: real-world attack on ArcFace-100 face recognition system." 2019

Защита от состязательных примеров для систем распознавания в реальном мире[29]

- Большинство состязательных примеров в реальном мире — патчи
  - Предложение: **A**dversarial **T**raining (AT)[28] в пиксельной области + прямоугольная аугментация
- Поэтапное AT:
  1. Позиция серого патча (max функции потерь) via:
     - Полный перебор
     - Max градиента по входу
  2. MI-FGSM внутри этого патча для поиска раскраски

- Обычная процедура обучения:

$$\min_{\theta} \mathbb{E}_{x,y}[L(\theta, x, y)]$$

- Состязательное обучение (AT):

$$\min_{\theta} \mathbb{E}_{x,y}[\max_{r \in \Delta} L(\theta, x + r, y)]$$

[28] Goodfellow I. et al. "Explaining and harnessing adversarial examples." 2014

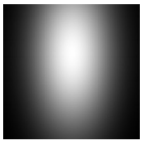
[29] Wu T. et al. "Defending Against Physically Realizable Attacks on Image Classification." 2019

Black-box восстановление лица по вектору признаков

- Black-box модель M: $M(x) = y$, где
  - x — входное фото лица
  - y — его векторное представление (эмбеддинг)
- **Задача**: восстановить x', сохраняя личность (ID)
- **Ранее**: использование функции потерь MSE и метрик на признаках / GAN (NBNet[30])
- **Здесь**: метод оптимизации нулевого порядка для нахождения x': $M(x') \approx M(x)$
  - В качестве M используется: ArcFace
  - Тест независимым критиком: FaceNET[31]
  - Функция потерь (близость): $1 - cos(M(x'), M(x))$
  - Добавка: $(||M(x')||_2 - ||M(x)||_2)^2$

- **Главная сложность**: огромное пространство поиска
- **Решение**: использование априорной информации о лице — 2D-гауссианы

$$G(x, y) = A \cdot e^{\frac{(x-x_0)^2}{2\sigma_1^2} + \frac{(y-y_0)^2}{2\sigma_2^2}}$$



$$(x_0, y_0, \sigma_1, \sigma_2, A) = (56, 72, 22, 42, 1)$$

[30] Mai G. et al. "On the reconstruction of face images from deep face templates." 2018

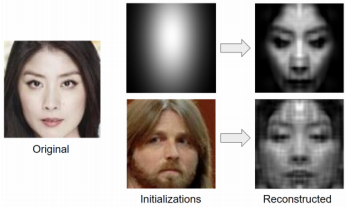
[31] Schroff F. et al. "Facenet: A unified embedding for face recognition and clustering." 2015.

Black-box восстановление лица по вектору признаков: трюки

- Даже использование априорной информации недостаточно
- **Трюк1**: Использовать вертикальную симметрию лица $\Rightarrow$ ищем только половину лица
- **Трюк2**: Для сохранения личности обычно цвет не нужен $\Rightarrow$ ищем только 1 цветовой канал вместо 3

Симметрия, без симметрии, в цвете

- **Инициализация**: Что взять за начальное приближение?
- **Обычно**: другое лицо (решение может быть смещено)
- **Можно**: оптимальный 2D-гауссиан (как раз нужна добавка на разность норм)



Black-box восстановление лица по вектору признаков: алгоритм и примеры

Алгоритм

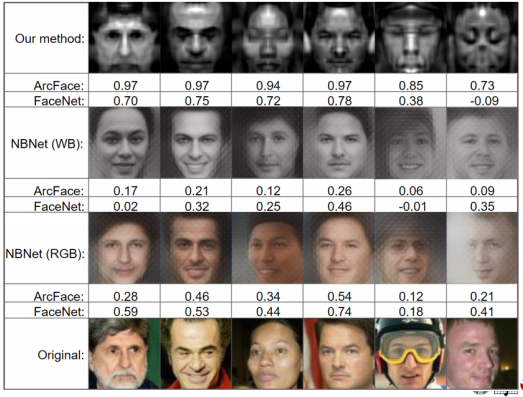
Algorithm 1 Face recovery algorithm

INPUT: target face embedding y, black-box model M, loss function L, $N_{queries}$

$X \leftarrow 0$
Initialize G_0
for $i \leftarrow 0$ to $N_{queries}$ **do**:
    Allocate image batch $\mathbf{X}$
    Sample batch $\mathbf{G}$ of random gaussians
    $\mathbf{X}_j = X + G_0 + \mathbf{G}_j$
    $\mathbf{y}' = M(\mathbf{X})$
    $\text{ind} = \text{argmin}\left(L(\mathbf{y}'_i, y)\right)$
    $X \leftarrow X + \mathbf{G}_{\text{ind}}$
    $G_0 \leftarrow 0.99 \cdot G_0$
    $i \leftarrow i + \text{batchsize}$
end for
$X \leftarrow X + G_0$

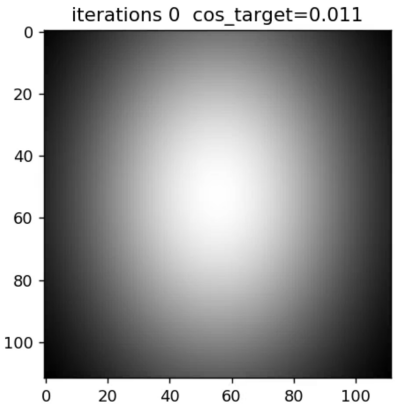
OUTPUT: reconstructed face X

Примеры



Black-box восстановление лица по вектору признаков: результат

Детали: статья[32] (ECCV-2020) и видео-демонстрация[33].



[32] Razzhigaev A. et al. "Black-Box Face Recovery from Identity Features." 2020

[33] https://www.youtube.com/watch?v=sOrTcqRTw2A

Заключительные выводы

- На данный момент CHC (в целом) работают гораздо лучше человека
- CHC неустойчивы (по входу)
- Перенести состязательный пример в реальный мир непросто
- Однако можно сломать даже супер навороченные системы распознавания лиц, имея лишь черно-белую бумажку с обычного принтера
- ℓ_0-оптимизация + EOT + TV обязательны
- Через проекции нужно уметь пропускать градиенты
- Состязательное обучение может помочь в защите от состязательных примеров
- Изображения лиц могут быть восстановлены по векторам признаков даже в black-box манере

Присоединяйтесь к нам!

Кого мы ждем:

- *Выпускники аспирантуры 2019-2021 годов*
- *Победители и призеры таких международных соревнований как ICPC, IMC, CTF, Kaggle, IMO, IOI, ICHO, IPHO etc.*
- *Техническое образование (информационные технологии, математика, физика, радиотехника, системы связи, информационная безопасность и др.)*
- *Английский на уровне "intermediate" и выше*

Наши направления:

- *Nonlinear algorithm development*
- *Wireless communication technologies*
- *Computer Vision with Deep Learning*
- *Math Library optimization*
- *Automatic program repair*
- *Compiler optimizations*
- *Automatic speech recognition*
- *AI databases and AI enabled systems*
- *Distributed and Parallel software*
- *Image/Video signal processing*
- *Software engineering and innovation*
- *Automated machine learning & Model optimization*
- *Computer architecture*

Резюме можно выслать на почту:

rrihr@huawei.com

Москва Санкт-Петербург Нижний Новгород Новосибирск

https://career.huawei.ru/rri/

Inspired by science to connect the world!

Looking forward to seeing your application

Спасибо за внимание!