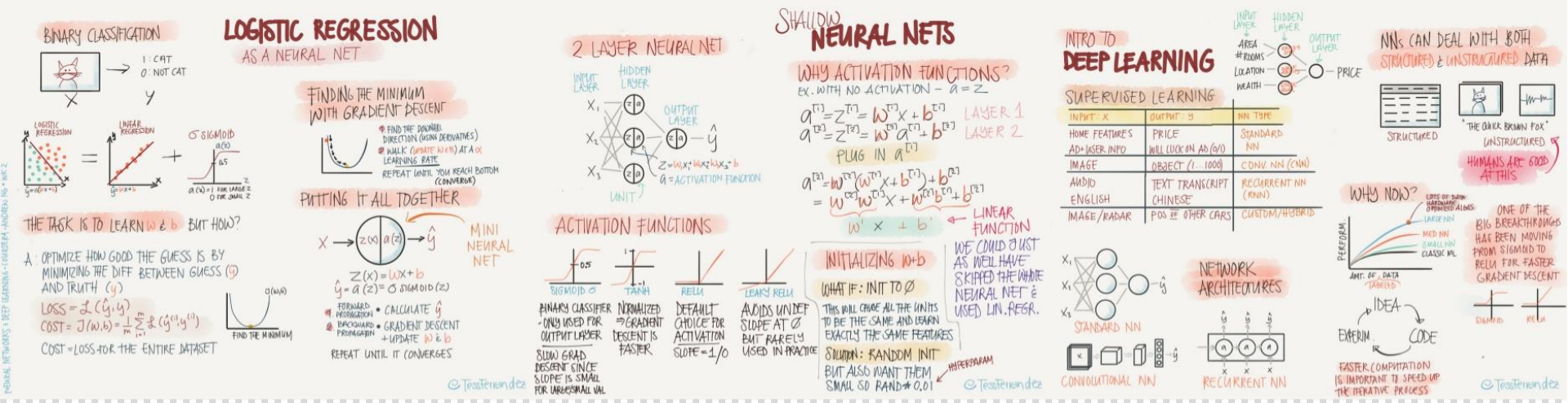


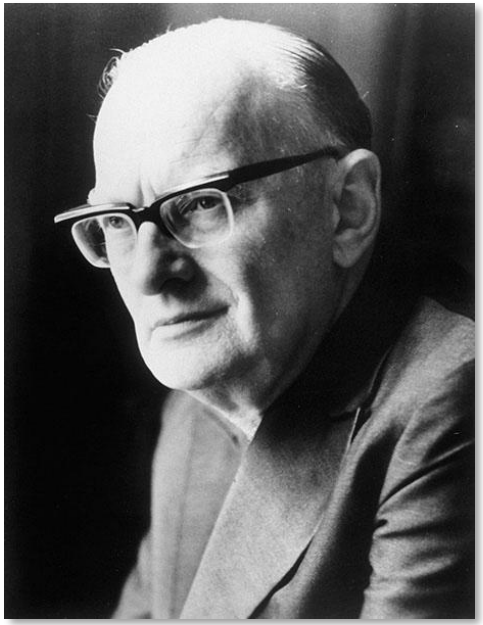
Уязвимости нейросетевых технологий

Афанасьев Сергей
КБ «Ренессанс Кредит»

8 сентября 2020 г
Москва

«Если знаменитый, но старый ученый утверждает, что нечто возможно, он почти определенно прав. Если он утверждает, что нечто невозможно, он, очень вероятно, ошибается»

Артур Кларк

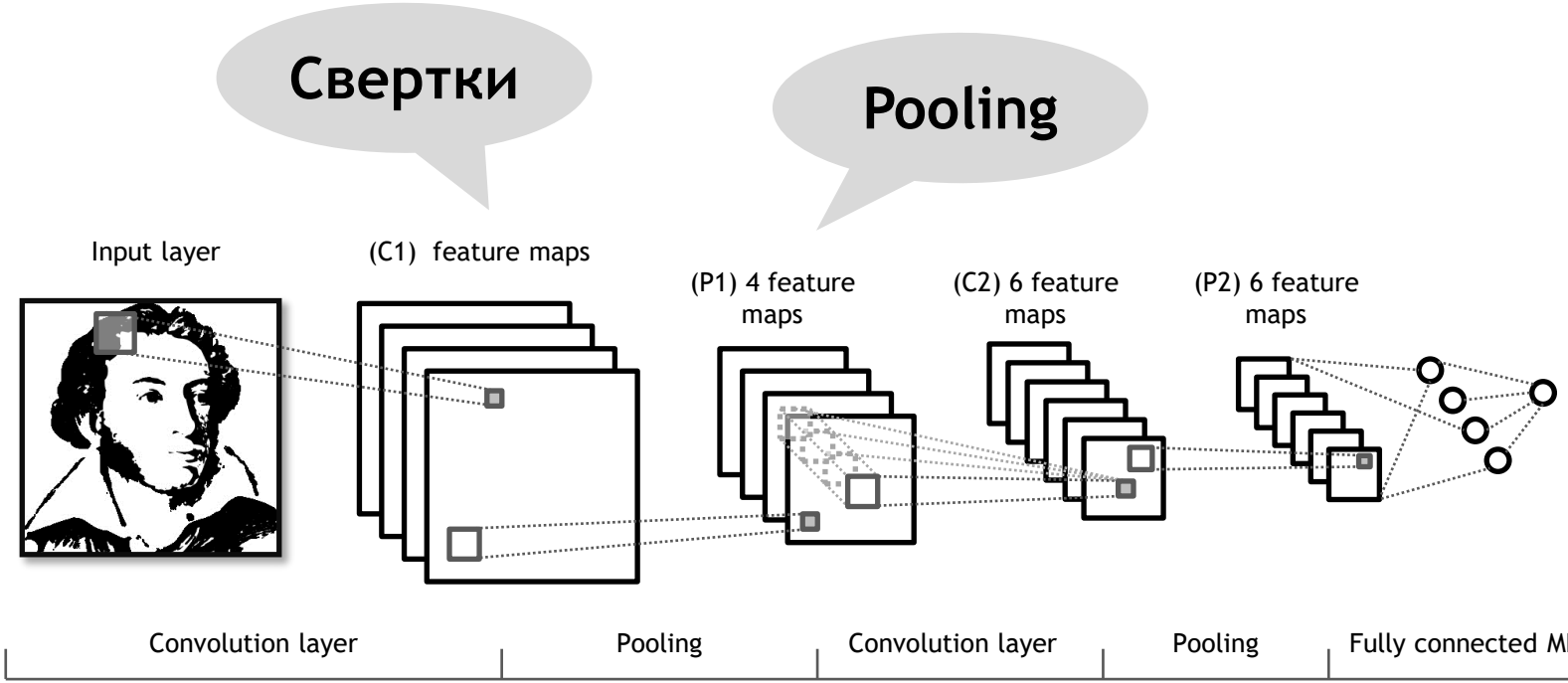


Как работает фотобиометрия?

Полиция американского города Ланкастер (штат Пенсильвания) смогла установить личность подозреваемого в краже благодаря портрету, который нарисовал один из свидетелей. При том, что портрет был, мягко говоря, схематичным. Уже на следующий день полиция установила личность предполагаемого преступника — им оказался бездомный по имени Хунг Фыок Нгуен.

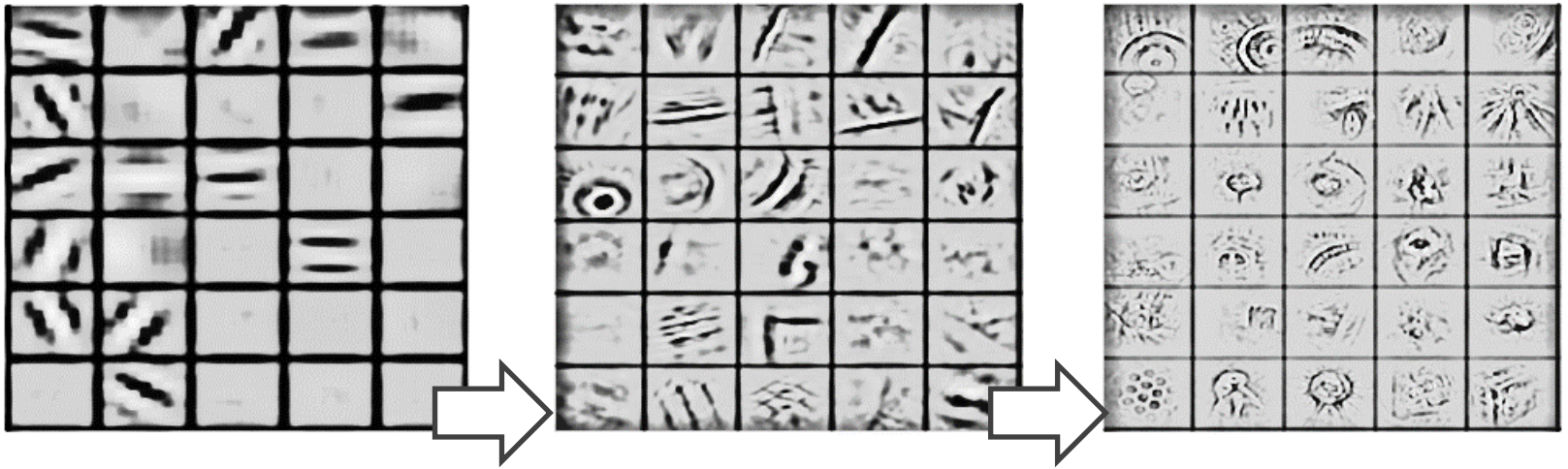
Есл вы мжте прчт эт
вы мжт плчит хошю рбту
с вскй зрплй!

Сверточные сети — это просто!



… выделяют палочки и крючки

Feature Map1 **Feature Map2** **Feature Map3**



Фотобиометрия уязвима

2017

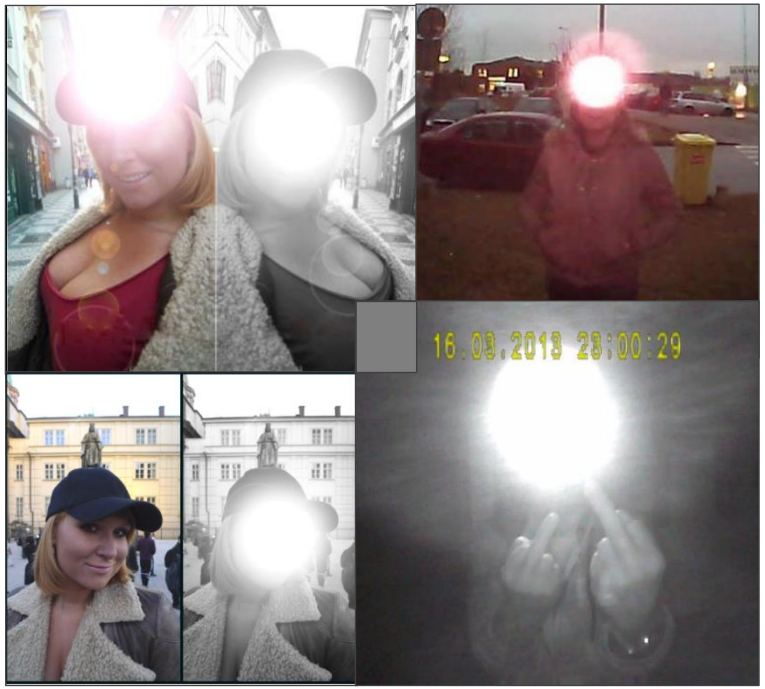
Макияж от Яндекса

Григорий Бакунов, занимающий в Яндексе должность директора по распространению технологий, рассказал о своем стороннем проекте — сервисе, проектирующем уникальный случайный макияж, который бы защищал лицо пользователя от идентификации системой распознавания лиц.

Фотобиометрия уязвима

2014

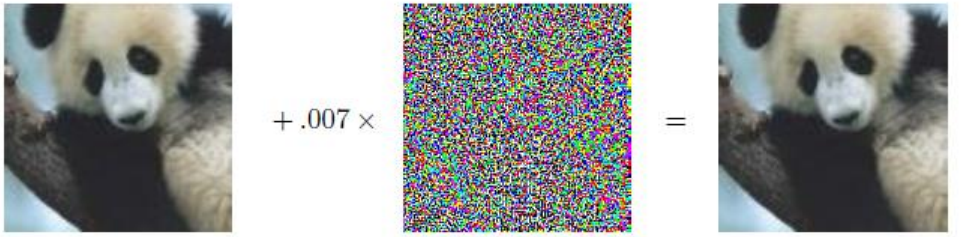
Бейсболка Justice Cap,
стоит всего $15.



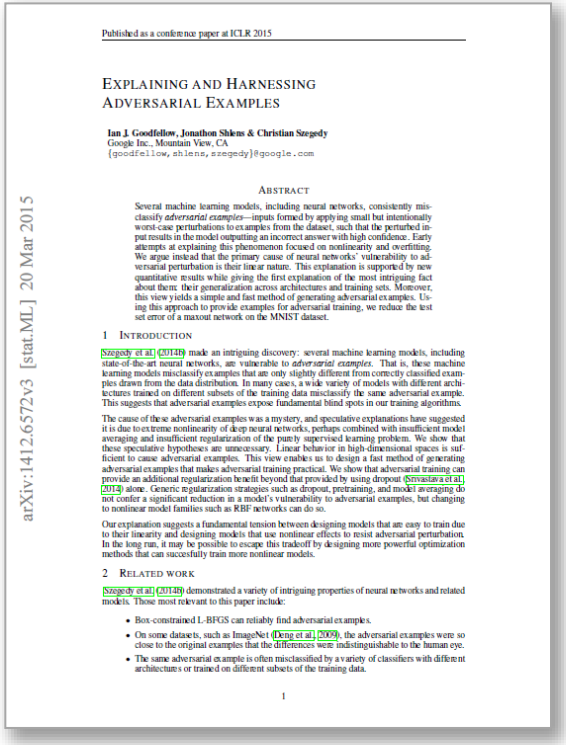
Атаки на нейронные сети

2015

Adversarial Attacks



Добавляем на картинку панды случайный шум и нейросеть определяет панду как гиббона



Published as a conference paper at ICLR 2015

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

ABSTRACT

Several machine learning models, including neural networks, consistently misclassify *adversarial examples*—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Early attempts at explaining this phenomenon focused on nonlinearity and overfitting. We argue instead that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This explanation is supported by new quantitative results while giving the first explanation of the most intriguing fact about them: their generalization across architectures and training sets. Moreover, this view yields a simple and fast method of generating adversarial examples. Using this approach to provide examples for adversarial training, we reduce the test set error of a maxout network on the MNIST dataset.

1 INTRODUCTION

Szegedy et al. (2014b) made an intriguing discovery: several machine learning models, including state-of-the-art neural networks, are vulnerable to *adversarial examples*. That is, these machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example. This suggests that adversarial examples expose fundamental blind spots in our training algorithms.

The cause of these adversarial examples was a mystery, and speculative explanations have suggested it is due to extreme nonlinearity of deep neural networks, perhaps combined with insufficient model averaging and insufficient regularization of the purely supervised learning problem. We show that these speculative hypotheses are unnecessary. Linear behavior in high-dimensional spaces is sufficient to cause adversarial examples. This view enables us to design a fast method of generating adversarial examples that makes adversarial training practical. We show that adversarial training can provide an additional regularization benefit beyond that provided by using dropout (Srivastava et al., 2014) alone. Generic regularization strategies such as dropout, pretraining, and model averaging do not confer a significant reduction in a model's vulnerability to adversarial examples, but changing to nonlinear model families such as RBF networks can do so.

Our explanation suggests a fundamental tension between designing models that are easy to train due to their linearity and designing models that use nonlinear effects to resist adversarial perturbation. In the long run, it may be possible to escape this tradeoff by designing more powerful optimization methods that can successfully train more nonlinear models.

2 RELATED WORK

Szegedy et al. (2014b) demonstrated a variety of intriguing properties of neural networks and related models. Those most relevant to this paper include:

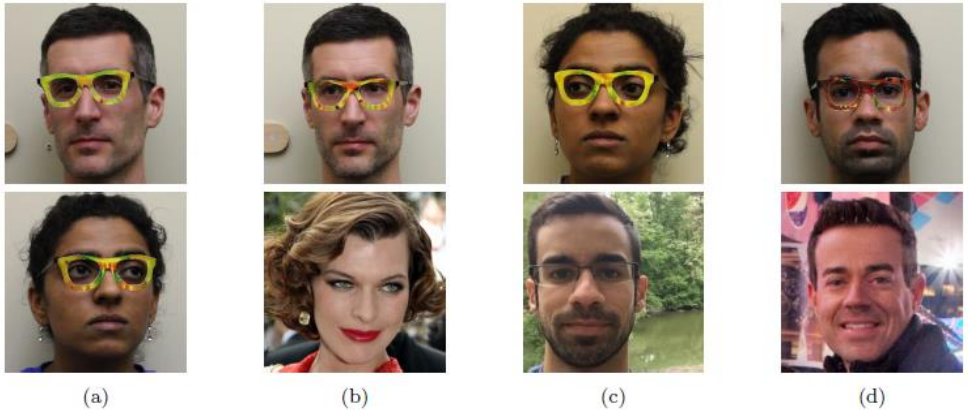
- Box-constrained L-BFGS can reliably find adversarial examples.
- On some datasets, such as ImageNet (Deng et al., 2009), the adversarial examples were so close to the original examples that the differences were indistinguishable to the human eye.
- The same adversarial example is often misclassified by a variety of classifiers with different architectures or trained on different subsets of the training data.

1

https://arxiv.org/pdf/1412.6572.pdf

Фотобиометрия уязвима

2016



(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Mahmood Sharif
Carnegie Mellon University
Pittsburgh, PA, USA
mahmoods@cmu.edu

Sruti Bhagavatula
Carnegie Mellon University
Pittsburgh, PA, USA
srutib@cmu.edu

Lujo Bauer
Carnegie Mellon University
Pittsburgh, PA, USA
lbauer@cmu.edu

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

ABSTRACT

Machine learning is enabling a myriad innovations, including new algorithms for cancer diagnosis and self-driving cars. The broad use of machine learning makes it important to understand the extent to which machine-learning algorithms are subject to attack, particularly when used in applications where physical security or safety is at risk.

In this paper, we focus on facial biometric systems, which are widely used in surveillance and access control. We define and investigate a novel class of attacks: attacks that are *physically realizable* and *inconspicuous*, and allow an attacker to evade recognition or impersonate another individual. We develop a systematic method to automatically generate such attacks, which are realized through printing a pair of eyeglass frames. When worn by the attacker whose image is supplied to a state-of-the-art face-recognition algorithm, the eyeglasses allow her to evade being recognized or to impersonate another individual. Our investigation focuses on white-box face-recognition systems, but we also demonstrate how similar techniques can be used in black-box scenarios, as well as to avoid face detection.

1. INTRODUCTION

Machine learning (ML) is a technology that is profoundly changing the world. Some of the transformative innovations that it enables, such as customized search results and automated translations, might seem minor. Other innovations are clearly revolutionary, enabling new applications or significantly increasing our quality of life—examples include algorithms for cancer diagnosis and self-driving vehicles.

The broad use of ML has also caused a rising interest to understand the extent to which ML algorithms and applications are vulnerable to attacks. These attacks are especially worrisome due to humans' increasing reliance on technology and ML (sometime trusting them even with their own lives [36]). The domains that received the majority of the attention have a common characteristic: the adversary is able to precisely control the digital representation of the input to the ML tools. For instance, in spam detection, the attacker can control the contents of the email and the address of the sender. Hence, unsurprisingly, it has been shown that attacks in many of these domains can be effective at evading ML classifiers designed to detect them [23, 32, 43].

In this paper, we explore the extent to which facial biometric systems are vulnerable to attacks. These systems are widely used for various sensitive purposes, including surveillance and access control [29, 27, 38]. Thus, attackers who mislead them can cause severe ramifications.

In contrast to domains previously explored, attackers who aim to mislead facial biometric systems often do not have precise control over the systems' input. Rather, attackers may be able to control only their (physical) appearance. The process of converting the physical scene into a digital input is not under the attackers' control, and is additionally affected by factors such as lighting conditions, pose, and distance. As a result, it is more difficult for attackers to manipulate or craft inputs that would cause misclassification than it would be, for example, in the domain of spam detection.

Another difficulty that attackers face in the case of facial biometric systems is that manipulating inputs to evade the ML classifiers might be easily observed from outside the systems. For example, attackers can wear an excessive amount of makeup in order to evade a surveillance system deployed at banks [14]. However, these attackers may draw an increased attention from bystanders, and can be deterred by traditional means, e.g., the police.

In the light of these two challenges, we define and study a new class of attacks: attacks that are physically realizable and at the same time are *inconspicuous*. In such attacks, the attacker manipulates the physical state that an ML algorithm is analyzing rather than the digitized representation of this state. At the same time, the manipulations of physical state needed by the attacks are sufficiently subtle that they are either imperceptible to humans or, if perceptible, seem natural and not representative of an attack.

Inconspicuousness We focus, unlike most related work (e.g., [7]), on attacks that are *inconspicuous*, i.e., a person who is physically present at a scene, or a person who looks at

CCS'16 October 24-28, 2016, Vienna, Austria

ACM ISBN 978-1-4503-4139-4/16/10
DOI: http://dx.doi.org/10.1145/2976749.2978392

Фотобиометрия уязвима

2017

Арт-проект дизайнера Дзинь-Цай Лю: проектор крепится на лбу и транслирует на лицо человека другое лицо

Проектор для лица (Источник: Jing-cai Liu)

Фотобиометрия уязвима

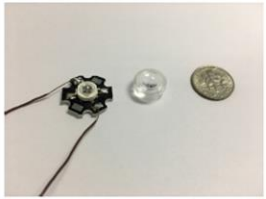
2018

Невидимая маска

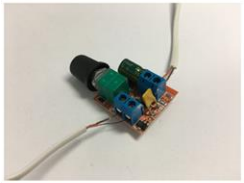
Китайские исследователи создали бейсбольную кепку, оснащенную миниатюрными инфракрасными светодиодами, которые размещены таким образом, что лучи, падающие на лицо владельца головного убора, помогают не только скрыть его личность, но и «выдать себя за другого человека для прохождения биометрической аутентификации».

1. Кепка со светодиодами

2. Батарейка



3. Светодиод и линза



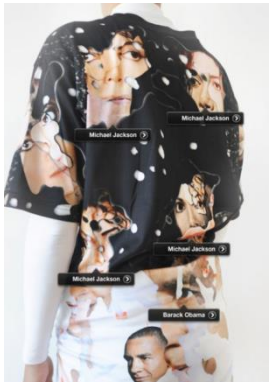
4. ШИМ-плата для управления яркостью

5. Три линзы для подбора диаметра луча

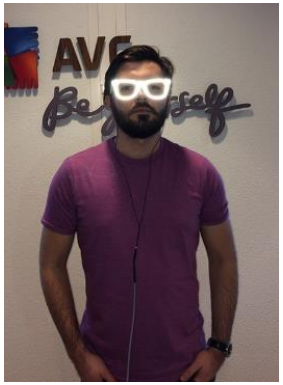
https://arxiv.org/pdf/1803.04683.pdf

Как еще обойти фотобиометрию?







Но это может стать незаконным

2020

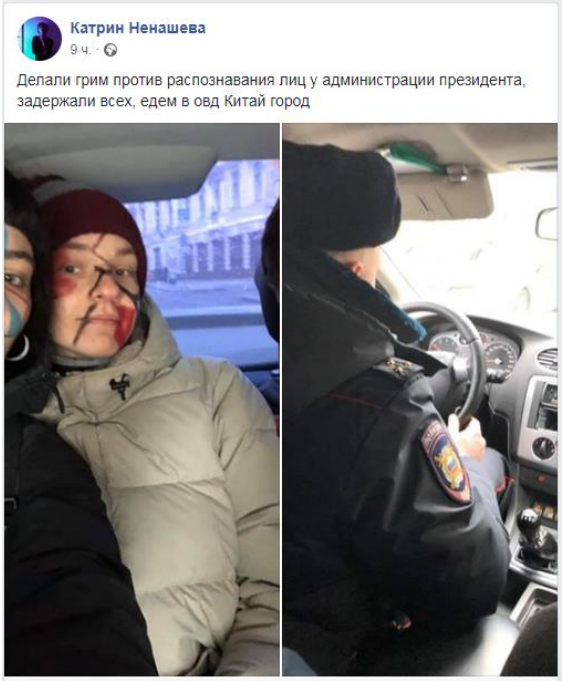
meduza

У администрации президента задержали четырех человек с гримом против системы распознавания лиц

13:07, 9 февраля 2020 · Источник: ОВД-Инфо · Надежный источник

У администрации президента задержали четырех активистов, которые делали грим против распознавания лиц, передает «ОВД-Инфо».

Им вменяют «несогласованную акцию».



https://meduza.io/news/2020/02/09/u-administratsii-prezidenta-zaderzhali-chetyreh-chelovek-s-grimom-protiv-raspoznayuschih-litsa-kamer?utm_source=facebook&utm_medium=main

GANы против фотобиометрии

2017

Deepfakes

Пользователь Reddit под ником deepfakes в декабре 2017 года опубликовал фейковые порнографические ролики с участием голливудских актрис, созданные с помощью нейронных сетей. Уже в январе 2018 вышло десктопное приложение FakeApp, которое позволяет создавать подобные ролики пользователям, не обладающих навыками программирования.

https://fakeapp.site/

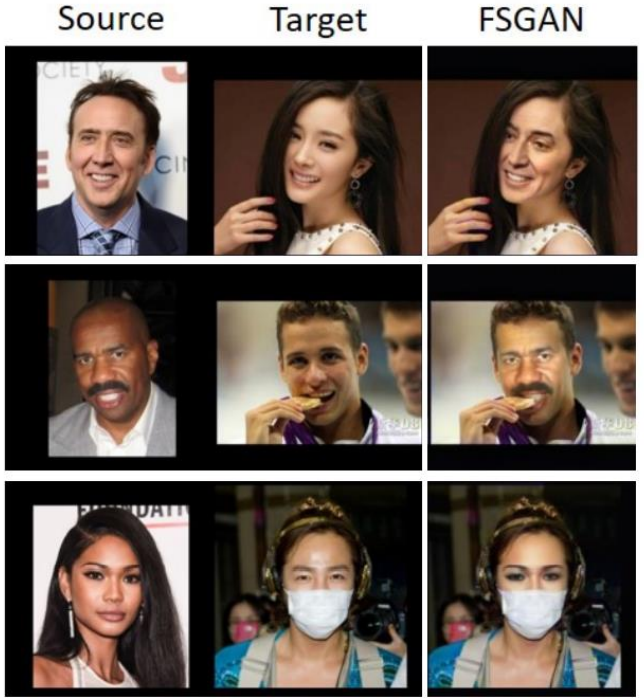
Deepfakes в реальном времени

2019

FSGAN

Новый метод для создания deepfakes создает реалистичные видео с заменой лиц в режиме реального времени, не требуя длительного обучения.

В отличие от предыдущих подходов deepfakes этот метод работает на любых двух людях без какой-либо специальной подготовки на их лицах.



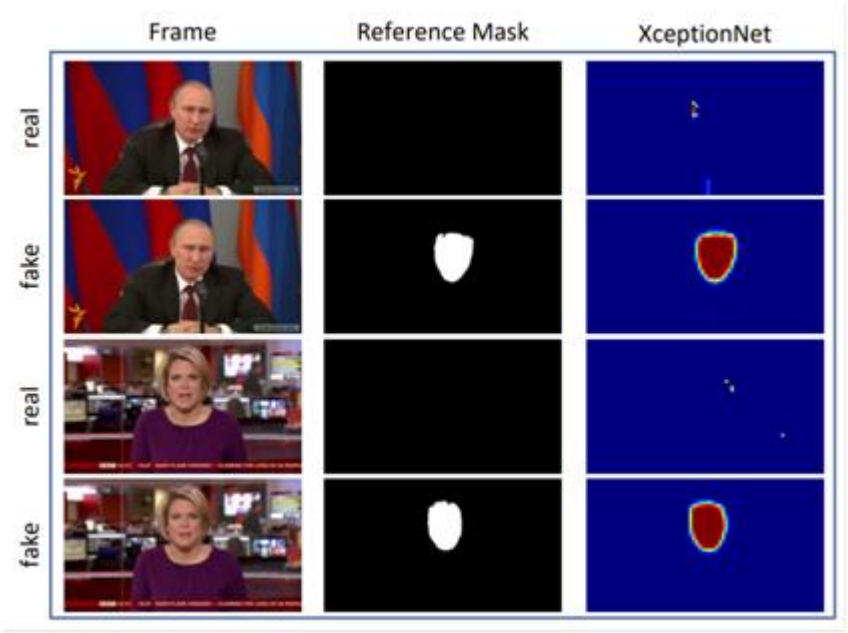
https://arxiv.org/pdf/1908.05932.pdf

А кто против GANов?

2018

XceptionNet

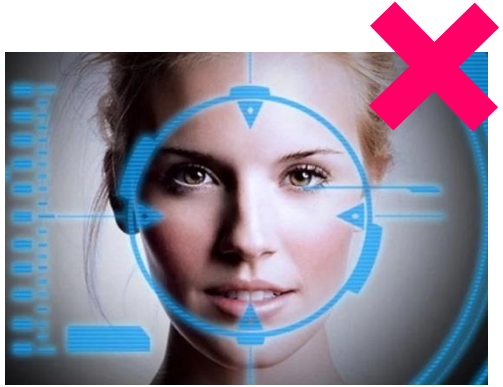
Европейские ученые научили алгоритм XceptionNet выявлять подмену лиц в видеороликах. Для обучения модели исследователи создали масштабный датасет FaceForensics из полумиллиона изображений, которые взяты из более чем тысячи видео, созданных с использованием алгоритма face2face.
По точности распознавания XceptionNet превосходит уже существующие алгоритмы в несколько раз даже при оценке сжатых видео.



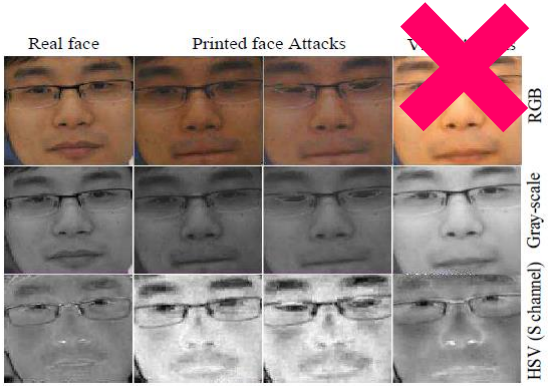
https://arxiv.org/pdf/1803.09179.pdf

А что насчет Liveness Detection?

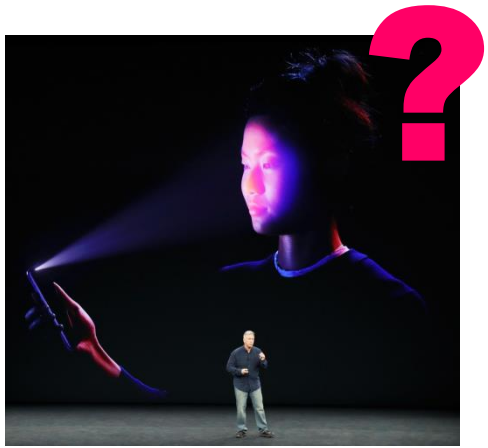
Интерактив



Текстуры



3D-камеры

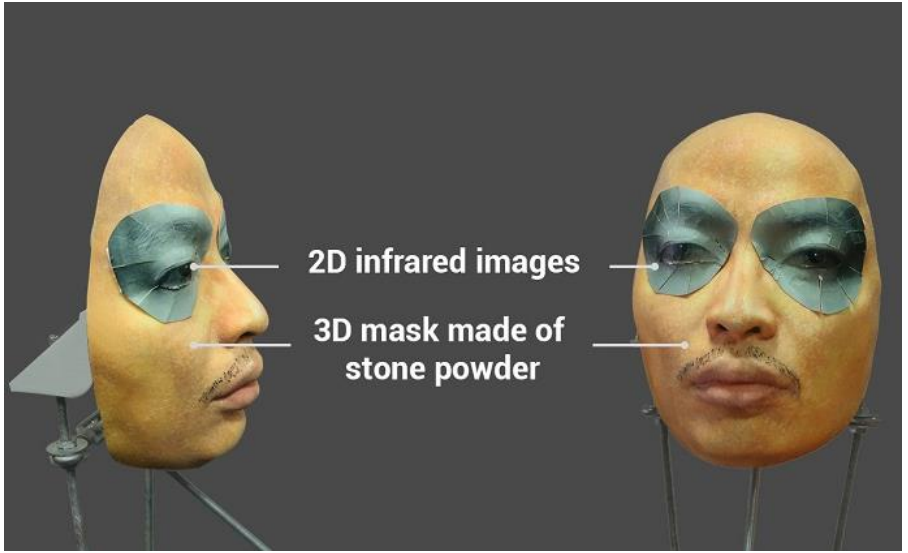


3D-биометрия не спасет

2017

FaceID взломан

Через несколько недель после выхода iPhone X вьетнамская компания Bkav создала маску для обхода защиты FaceID. Первая версия маски состояла из четырех компонентов: основы, напечатанной на 3D-принтере; силиконовой копии носа; двухмерных изображений глаз; а также «специально рассчитанных» областей лица. Во второй версии использовалась только 3D-основа из каменного порошка и двумерные изображения глаз.



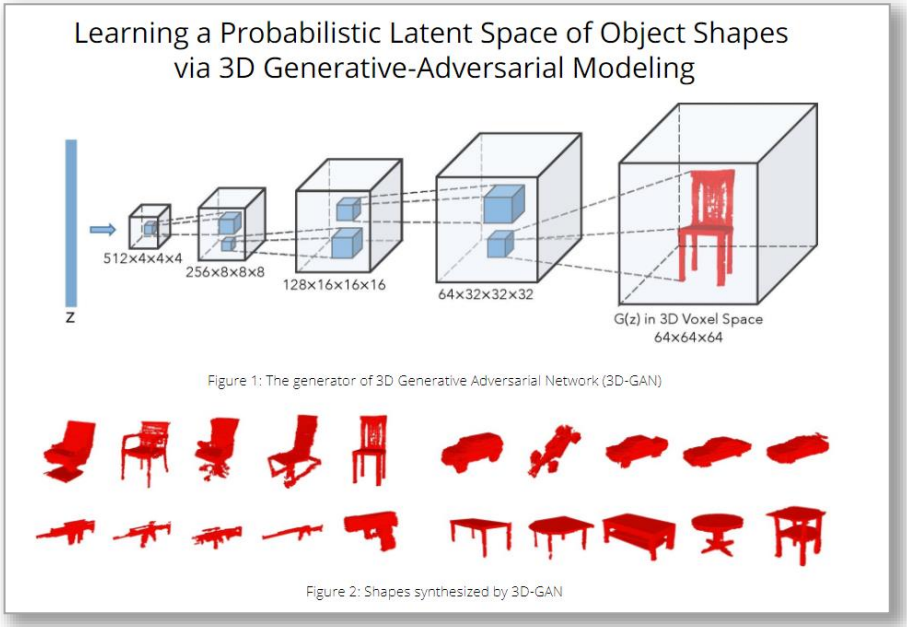
http://www.bkav.com

Генерация 3D-объектов

2016

3D-GAN vs. IM2CAD

Ученые из MIT CSAIL и Google Research разработали ИИ на основе генеративно-состязательных сетей (GAN), который умеет генерировать 3D-объекты высокого качества. 3D-GAN превосходит по качеству предшествующие CAD-модели, которые умеют восстанавливать 3D-объекты по фотографиям.



29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain. http://3dgan.csail.mit.edu/papers/3dgan_nips.pdf

Генерация 3D-масок

2017

VRN-маски

В 2017 году британские ученые разработали нейронную сеть VRN, которая создает 3D-маску человека всего по одной фотографии. Разработку выложили в открытый доступ, а демо-версию работы этой нейросети может протестировать любой желающий: http://www.cs.nott.ac.uk/~psxasj/3dme/

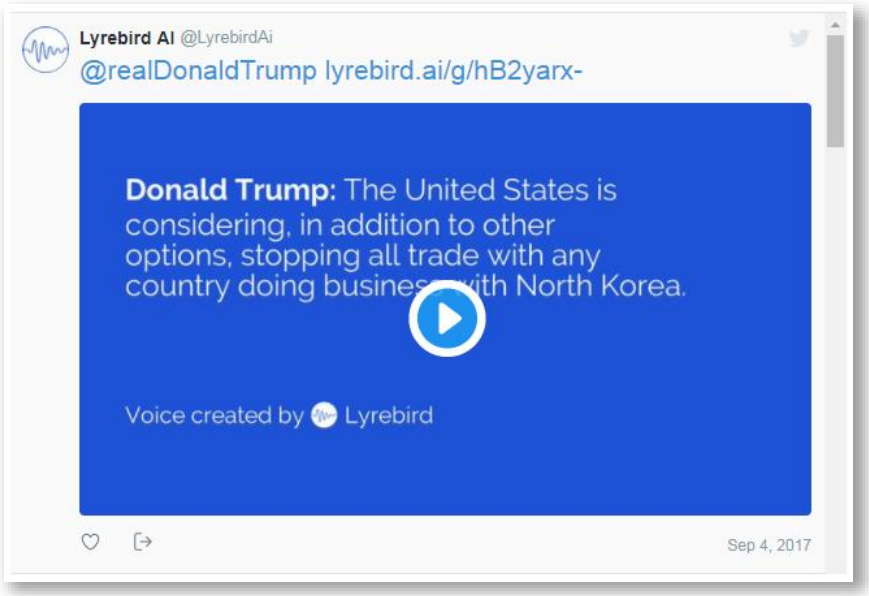
Голосовая биометрия не поможет

2017

Lyrebird AI

TNW с испугом пишет о потенциале проекта Lyrebird для преступников. Эта система позволяет загрузить минуту своего голоса, а потом получить текст-ту-спич с голосом, похожим на ваш. Автор прав, сделать из, скажем, 2-3 часов речи очень приличный TTS можно, причем для этого подойдут уже готовые опенсорсные технологии.

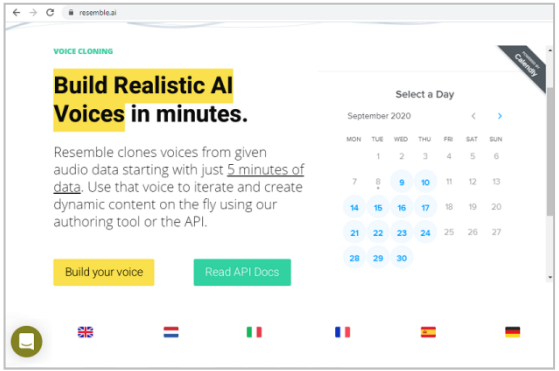
Григорий Бакунов, Яндекс



https://www.descript.com/lyrebird-ai?source=lyrebird

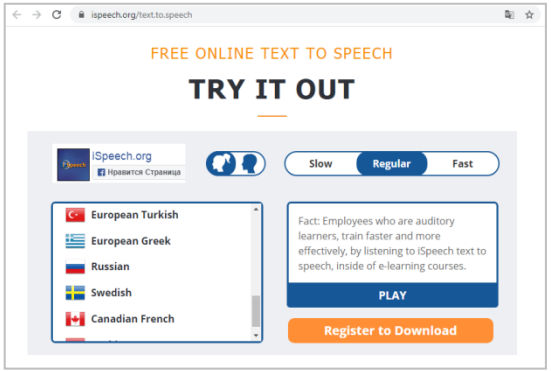
Еще генераторы речи…

Resemble.AI



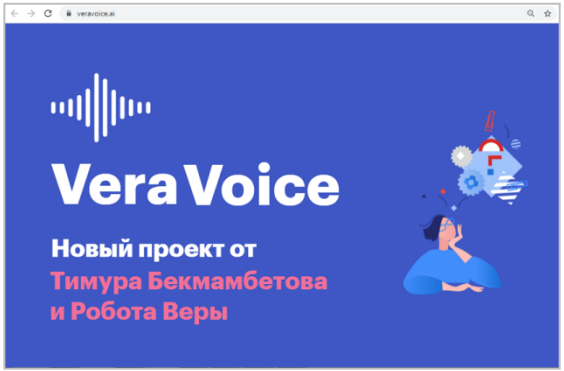
Предоставляется демоверсия программы

iSpeech



Есть демоверсия для 27 языков, включая русский;

Vera Voice



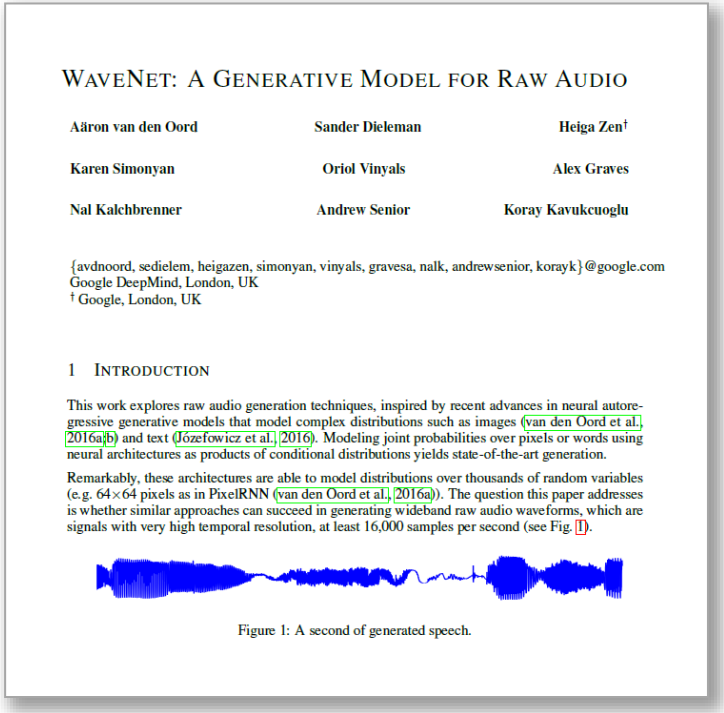
Проект от компании Screenlife Technologies и команды проекта «Робот Вера».

А как же Liveness Detection?

2016

WaveNet

Компания Google DeepMind разработала нейронную сеть WaveNet, позволяющую генерировать речь по технологии TTS (text-to-speech). В основе WaveNet лежат разработанные ранее в DeepMind сверточная и рекуррентная архитектуры PixelCNN и PixelRNN. Компания DeepMind отмечает, что качество сгенерированной WaveNet речи практически неотличимо от настоящей. В первых релизах 2016 года отмечалось, что генерация 1 секунды речи занимала 1-2 минуты. Но позже эту проблему решили и уже в сентябре 2017 года WaveNet была встроена в Google Assistant.



WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Aäron van den Oord Sander Dieleman Heiga Zen†
Karen Simonyan Oriol Vinyals Alex Graves
Nal Kalchbrenner Andrew Senior Koray Kavukcuoglu

{avdnoord, sedielem, heigazen, simonyan, vinyals, gravesa, nalk, andrewsenior, korayk}@google.com
Google DeepMind, London, UK
† Google, London, UK

1 INTRODUCTION

This work explores raw audio generation techniques, inspired by recent advances in neural autoregressive generative models that model complex distributions such as images (van den Oord et al., 2016a;b) and text (Józefowicz et al., 2016). Modeling joint probabilities over pixels or words using neural architectures as products of conditional distributions yields state-of-the-art generation.

Remarkably, these architectures are able to model distributions over thousands of random variables (e.g. 64×64 pixels as in PixelRNN (van den Oord et al., 2016a)). The question this paper addresses is whether similar approaches can succeed in generating wideband raw audio waveforms, which are signals with very high temporal resolution, at least 16,000 samples per second (see Fig. 1).

Figure 1: A second of generated speech.

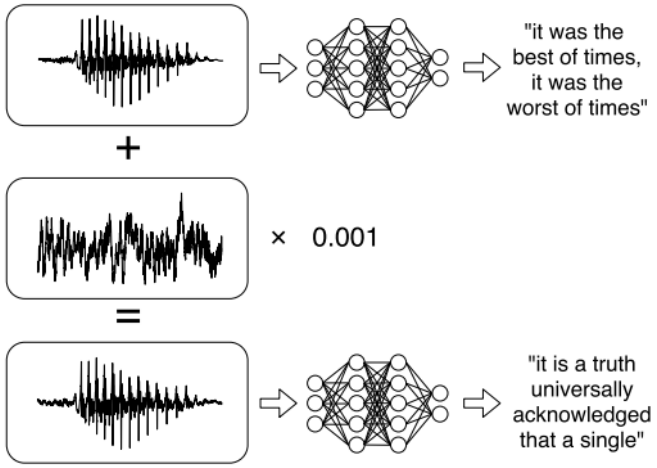
https://arxiv.org/pdf/1609.03499.pdf

Атака на голосовые помощники

2018

(Не)случайный шум

Исследователи из Китая и США показали, что они могут отправлять набор звуков, за пределами человеческой слышимости (или замаскированных среди музыки), которые Siri, Alexa или помощник Google воспринимают как команды. Исследователи смогли тайно активировать системы ИИ на смартфонах и смарт-динамиках, заставляя их набирать номера телефонов или открывать веб-сайты. В руках преступников технология может быть использована для разблокировки дверей, перевода денег или покупки вещей в интернете — просто с музыкой, играющей по радио.



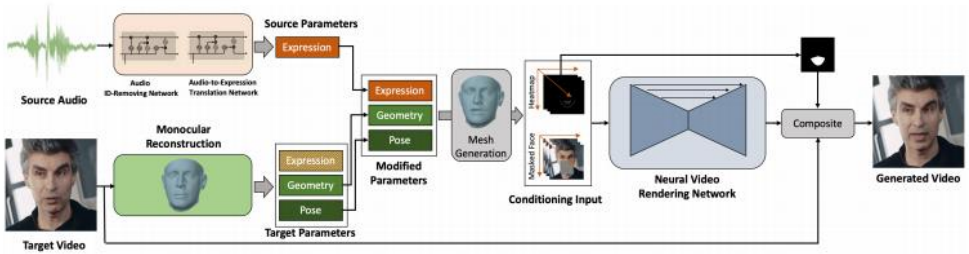
Пример атаки на голосовые помощники: учитывая форму сигнала, добавление небольшого «шума» заставляет результат транскрибироваться в любую желаемую фразу

https://arxiv.org/pdf/1801.01944.pdf

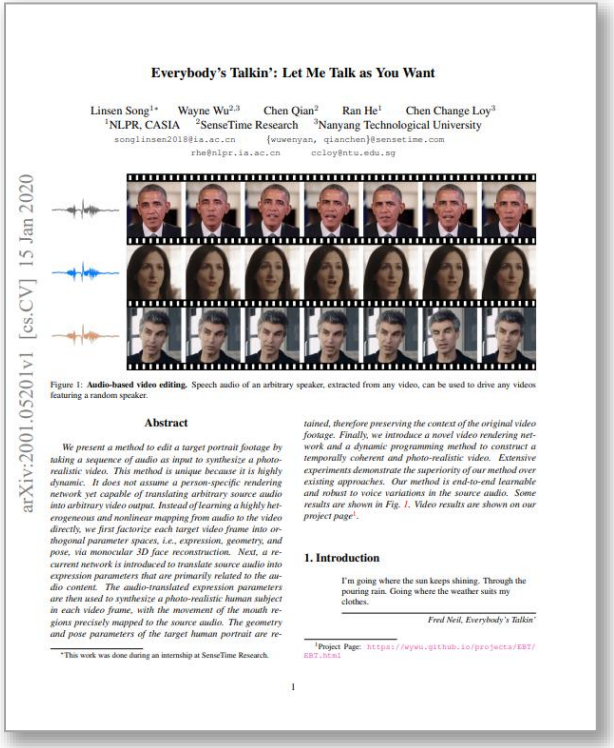
Генерация видео по голосу

2020

Речь + Фото = Видео



Исследователи из SenseTime разработали генеративную нейросеть, которая принимает на вход изображение целевой персоны и аудиозапись с речью, а на выходе отдает видеозапись с целевой персоной, на которой выражение лица персоны соответствует аудиодорожке.



Everybody's Talkin': Let Me Talk as You Want

Linsen Song[1*] Wayne Wu[2,3] Chen Qian[2] Ran He[1] Chen Change Loy[3]
[1]NLPR, CASIA [2]SenseTime Research [3]Nanyang Technological University
songlinsen2018@ia.ac.cn {wuwenyan, qianchen}@sensetime.com
rhe@nlpr.ia.ac.cn ccloy@ntu.edu.sg

arXiv:2001.05201v1 [cs.CV] 15 Jan 2020

Figure 1: **Audio-based video editing.** Speech audio of an arbitrary speaker, extracted from any video, can be used to drive any videos featuring a random speaker.

Abstract

We present a method to edit a target portrait footage by taking a sequence of audio as input to synthesize a photo-realistic video. This method is unique because it is highly dynamic. It does not assume a person-specific rendering network yet capable of translating arbitrary source audio into arbitrary video output. Instead of learning a highly heterogeneous and nonlinear mapping from audio to the video directly, we first factorize each target video frame into orthogonal parameter spaces, i.e., expression, geometry, and pose, via monocular 3D face reconstruction. Next, a recurrent network is introduced to translate source audio into expression parameters that are primarily related to the audio content. The audio-translated expression parameters are then used to synthesize a photo-realistic human subject in each video frame, with the movement of the mouth regions precisely mapped to the source audio. The geometry and pose parameters of the target human portrait are retained, therefore preserving the context of the original video footage. Finally, we introduce a novel video rendering network and a dynamic programming method to construct a temporally coherent and photo-realistic video. Extensive experiments demonstrate the superiority of our method over existing approaches. Our method is end-to-end learnable and robust to voice variations in the source audio. Some results are shown in Fig. 1. Video results are shown on our project page[1].

1. Introduction

I'm going where the sun keeps shining. Through the pouring rain. Going where the weather suits my clothes.

Fred Neil, Everybody's Talkin'

[*]This work was done during an internship at SenseTime Research.

[1]Project Page: https://wywu.github.io/projects/EBT/EBT.html

1

https://arxiv.org/pdf/2001.05201.pdf

Может тогда Behavioral Biometrics?

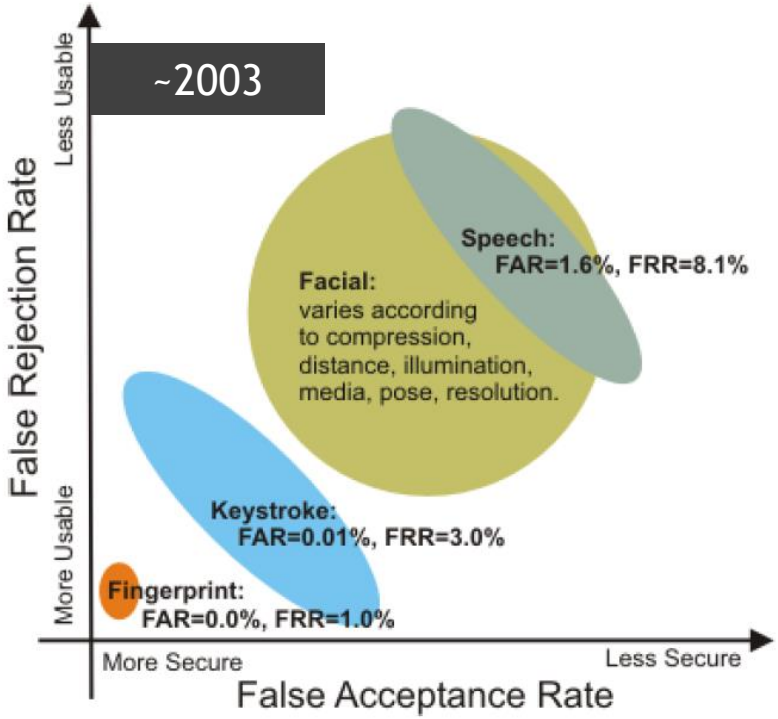
Classification of the various types of behavioural biometrics	Authorship	Direct human computer interaction		Indirect human computer interaction	Motor skill	Purely behavioral	Properties of behavioural biometrics			
		Input device interaction based	Software interaction based				Enrollment time	Verification time	Identification	Required hardware
Audit logs				■			D	D	N	Computer
Biometric sketch	■					■	M	S	N	Mouse
Blinking (моргание)					■		M	S	N	Camera
Call-stack				■			D	H	N	Computer
Calling behaviour						■	D	D	N	Phone
Car driving style (стиль вождения)						■	H	M	N	Car sensors
Command line lexicon			■			■	H	H	N	Computer
Credit card use (транзакционное поведение)						■	D	D	N	Credit card
Dynamic facial features					■		M	S	N	Camera
E-mail behaviour	■		■			■	D	M	N	Computer
Gait/Stride (походка)					■		M	S	N	Camera
Game strategy (игровая стратегия)			■				H	H	N	Computer
GUI interaction				■			D	H	N	Computer
Handgrip (рукоять пистолета)					■		M	S	N	Gun sensors
Haptic		■			■		M	M	N	Haptic
Keystroke dynamics		■			■		M	S	N	Keybord
Lip movements (движение губ)					■		M	S	N	Camera
Mouse dynamics		■			■		M	S	N	Mouse
Network traffic (сетевой трафик)				■			D	D	N	Computer
Painting style	■					■	D	D	N	Scanner
Programming style (стиль программирования)	■		■			■	H	H	N	Computer
Registry Access				■			D	H	N	Computer
Signature/Handwriting					■		M	S	Y	Stylus
Storage Activity			■				D	D	N	Computer
System Calls			■				D	H	N	Computer
Tapping				■			M	S	N	Sensor
Text Authorship (авторский стиль)	■					■	H	M	N	Computer
Voice/Speech/Singing				■			M	S	Y	Microphone

http://cecs.louisville.edu/ry/Behavioral.pdf

Клавиатурный почерк

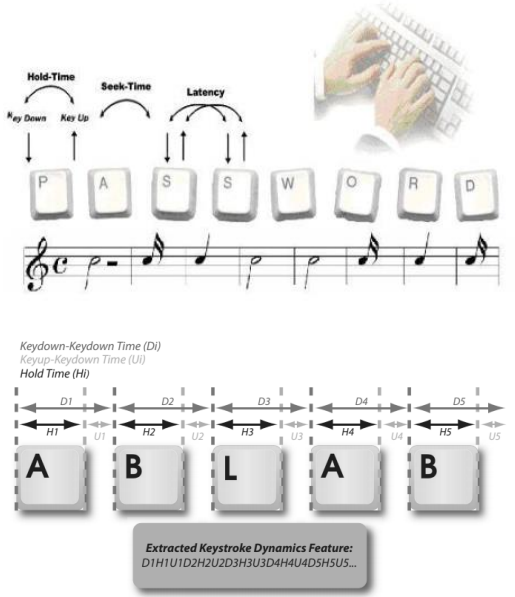
http://checco.com/about/john.checco/publications/2003_Keystroke_Biometrics_Intro.pdf

Keystroke был точнее лица и голоса



John C. Checco, Keystroke Dynamics And Corporate Security, 2003

Keystroke Dynamics



Keystroke с электромиографией

2015

Клавиатура + ЭМГ-сигналы

Ученые из университета Карнеги-Меллона разработали бимодальную биометрию на основе клавиатурного почерка и ЭМГ-сигналов (электромиография). Обычную технологию Keystroke, основанную на таймингах нажатия клавиш, легко обмануть, поэтому модальность в виде ЭМГ-сигналов позволит усилить защиту биометрии по клавиатурному почерку. ЭМГ-сигналы для данного вида биометрии можно собирать с пользовательских смарт-браслетов.

Figure 3. The electrodes from BioRadio attached to both arms of a participant measure the EMG signal as he/she types the given phrase during a trial

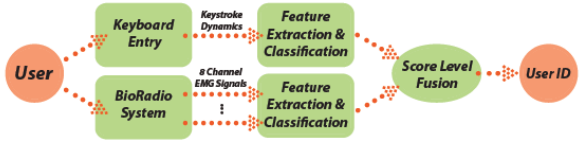


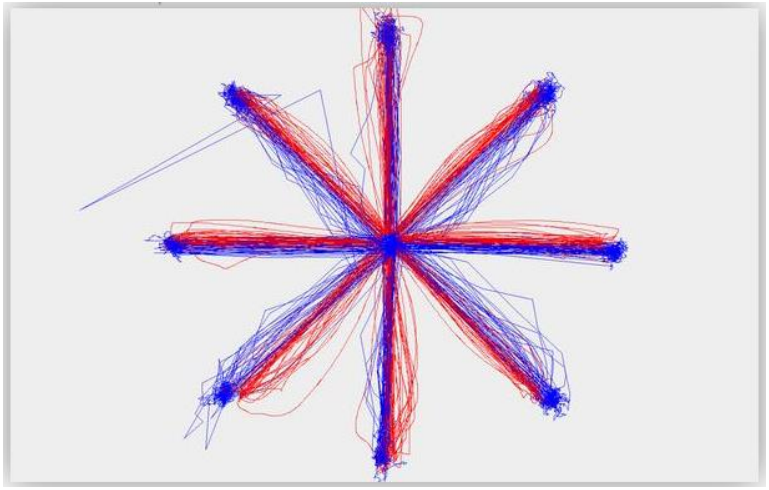
Figure 1. Our proposed experimental setup. We gather both EMG as well as keystroke dynamic data as the user types a fixed phrase on a keyboard. Various features are extracted from this data as detailed in later sections. We report user identification/verification scores when using either modality as a biometric.

http://openaccess.thecvf.com/content_cvpr_workshops_2015/W02/papers/Venugopalan_Electromyograph_and_Keystroke_2015_CVPR_paper.pdf

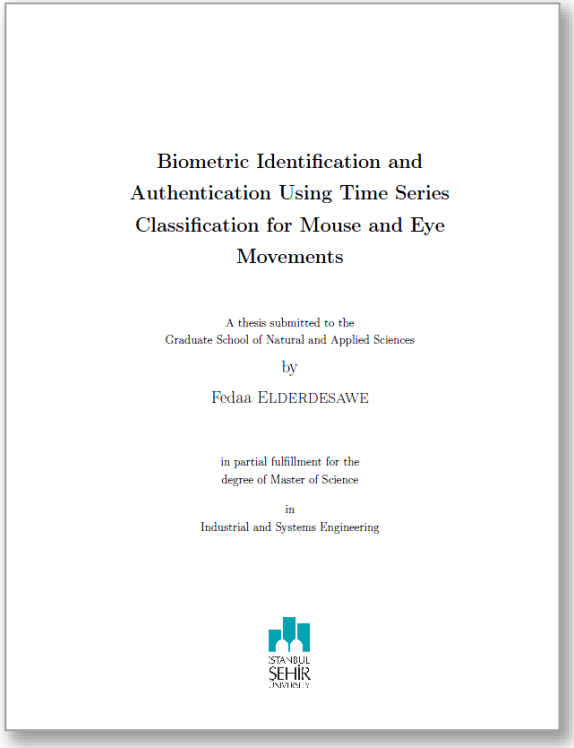
Биометрия по движению мышки

2017

Mouse + Eye movements



Синим – движение глаз. Красным – движение мышки



Biometric Identification and Authentication Using Time Series Classification for Mouse and Eye Movements

A thesis submitted to the
Graduate School of Natural and Applied Sciences

by

Fedaa ELDERDESAWE

in partial fulfillment for the
degree of Master of Science

in

Industrial and Systems Engineering

ISTANBUL SEHIR UNIVERSITY

Jamison Rose, Yudong Liu and Ahmed Awad
Biometric Authentication Using Mouse and Eye Movement Data

DOI: 10.1109/SPW.2017.18 Conference: 2017 IEEE Security and Privacy Workshops (SPW)

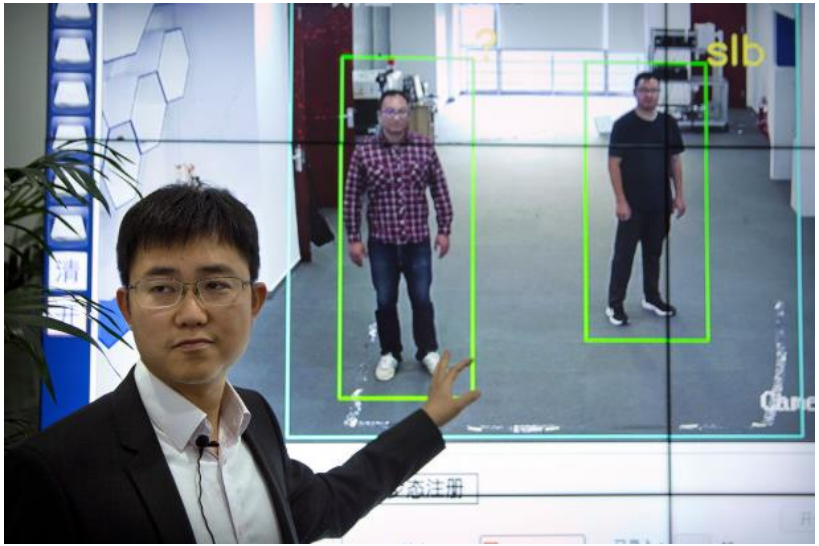
Биометрия по походке

2018

“Большой брат” от Watrix

Китайский стартап Watrix разработали технологи распознавания человека по походке. Программа может идентифицировать человека на расстоянии 50 метров, даже если его лицо закрыто или он повернут спиной к камере.

Полиция Пекина, Шанхая и Чунцина уже провела тестирования, а в 2019 году Watrix официально выпустили версию 2.0, которая поддерживает анализ изображений с камер в реальном времени в масштабах мегаполисов.



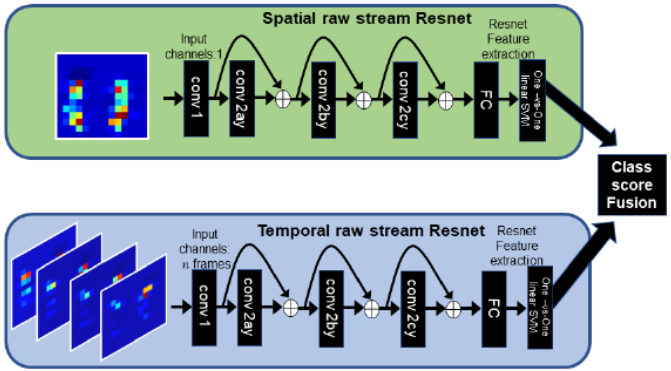
https://www.scmp.com/tech/start-ups/article/2187600/chinese-police-surveillance-gets-boost-ai-start-watrix-technology-can

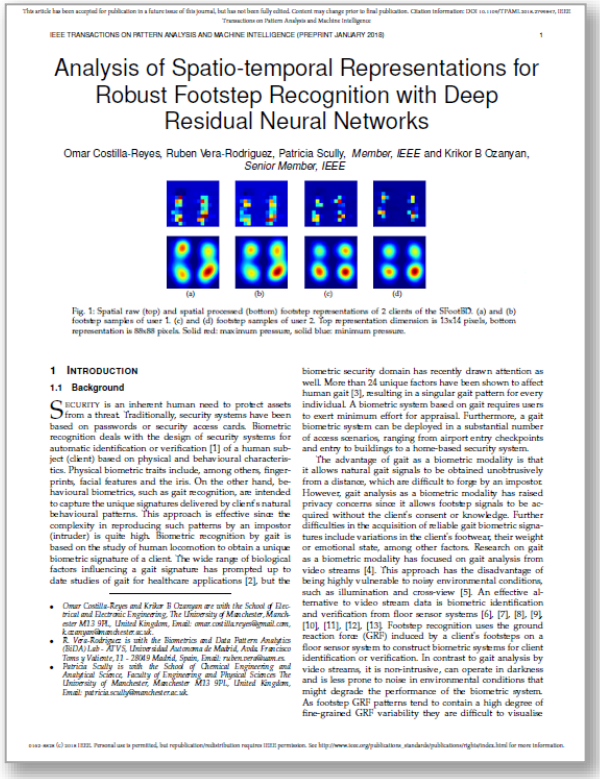
Биометрия по следам и походке

2018

Two-stream spatio-temporal ResNet

Команда британских и испанских ученых предложили нейросеть, которая позволяет распознавать человека по пространственным и временным характеристикам его следа, используя данные с пьезоэлектрических датчиков, рассчитывающих величину давления.





Analysis of Spatio-temporal Representations for Robust Footstep Recognition with Deep Residual Neural Networks

Omar Costilla-Reyes, Ruben Vera-Rodriguez, Patricia Scully, *Member, IEEE* and Krikor B Ozanyan, *Senior Member, IEEE*

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PREPRINT JANUARY 2018)

Fig. 1: Spatial raw (top) and spatial processed (bottom) footstep representations of 2 clients of the SFootBD. (a) and (b) footstep samples of user 1. (c) and (d) footstep samples of user 2. Top representation dimension is 13x14 pixels, bottom representation is 88x88 pixels. Solid red: maximum pressure, solid blue: minimum pressure.

1 Introduction

1.1 Background

Security is an inherent human need to protect assets from a threat. Traditionally, security systems have been based on passwords or security access cards. Biometric recognition deals with the design of security systems for automatic identification or verification [1] of a human subject (client) based on physical and behavioural characteristics. Physical biometric traits include, among others, fingerprints, facial features and the iris. On the other hand, behavioural biometrics, such as gait recognition, are intended to capture the unique signatures delivered by client's natural behavioural patterns. This approach is effective since the complexity in reproducing such patterns by an impostor (intruder) is quite high. Biometric recognition by gait is based on the study of human locomotion to obtain a unique biometric signature of a client. The wide range of biological factors influencing a gait signature has prompted up to date studies of gait for healthcare applications [2], but the biometric security domain has recently drawn attention as well. More than 24 unique factors have been shown to affect human gait [3], resulting in a singular gait pattern for every individual. A biometric system based on gait requires users to exert minimum effort for appraisal. Furthermore, a gait biometric system can be deployed in a substantial number of access scenarios, ranging from airport entry checkpoints and entry to buildings to a home-based security system.

The advantage of gait as a biometric modality is that it allows natural gait signals to be obtained unobtrusively from a distance, which are difficult to forge by an impostor. However, gait analysis as a biometric modality has raised privacy concerns since it allows footstep signals to be acquired without the client's consent or knowledge. Further difficulties in the acquisition of reliable gait biometric signatures include variations in the client's footwear, their weight or emotional state, among other factors. Research on gait as a biometric modality has focused on gait analysis from video streams [4]. This approach has the disadvantage of being highly vulnerable to noisy environmental conditions, such as illumination and cross-view [5]. An effective alternative to video stream data is biometric identification and verification from floor sensor systems [6], [7], [8], [9], [10], [11], [12], [13]. Footstep recognition uses the ground reaction force (GRF) induced by a client's footsteps on a floor sensor system to construct biometric systems for client identification or verification. In contrast to gait analysis by video streams, it is non-intrusive, can operate in darkness and is less prone to noise in environmental conditions that might degrade the performance of the biometric system. As footstep GRF patterns tend to contain a high degree of fine-grained GRF variability they are difficult to visualise

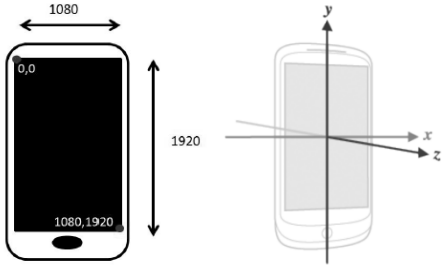
- *Omar Costilla-Reyes and Krikor B Ozanyan are with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, United Kingdom, Email: omar.costilla.reyes@gmail.com, k.ozanyan@manchester.ac.uk.*
- *R. Vera-Rodriguez is with the Biometrics and Data Pattern Analytics (BiDA) Lab - ATVS, Universidad Autonoma de Madrid, Avda. Francisco Tomas y Valiente, 11 - 28049 Madrid, Spain, Email: ruben.vera@uam.es.*
- *Patricia Scully is with the School of Chemical Engineering and Analytical Science, Faculty of Engineering and Physical Sciences The University of Manchester, Manchester M13 9PL, United Kingdom, Email: patricia.scully@manchester.ac.uk.*

https://ieeexplore.ieee.org/document/8275035

Биометрия по датчикам смартфона

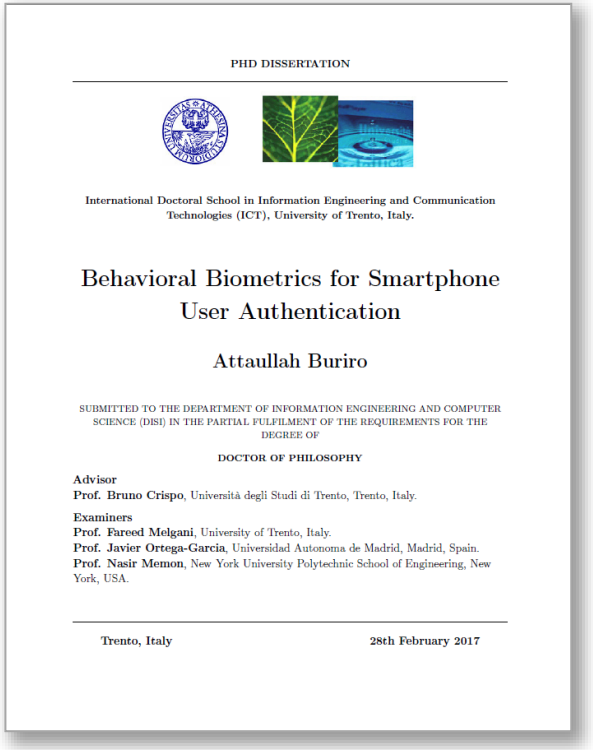
2017

Sensor Fusion



- Accelerometer
- Gravity
- Gyroscope
- Magnetometer
- Pressure
- Temperature
- Humidity
- Orientation
- Touchscreen

1. Поведенческие паттерны могут быть собраны прозрачно и даже без ведома пользователя
2. Сбор данных с датчиков смартфона не требует специального программного обеспечения



PHD DISSERTATION

International Doctoral School in Information Engineering and Communication Technologies (ICT), University of Trento, Italy.

Behavioral Biometrics for Smartphone User Authentication

Attaullah Buriro

SUBMITTED TO THE DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE (DISI) IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor
Prof. Bruno Crispo, Università degli Studi di Trento, Trento, Italy.

Examiners
Prof. Fareed Melgani, University of Trento, Italy.
Prof. Javier Ortega-Garcia, Universidad Autonoma de Madrid, Madrid, Spain.
Prof. Nasir Memon, New York University Polytechnic School of Engineering, New York, USA.

Trento, Italy 28th February 2017

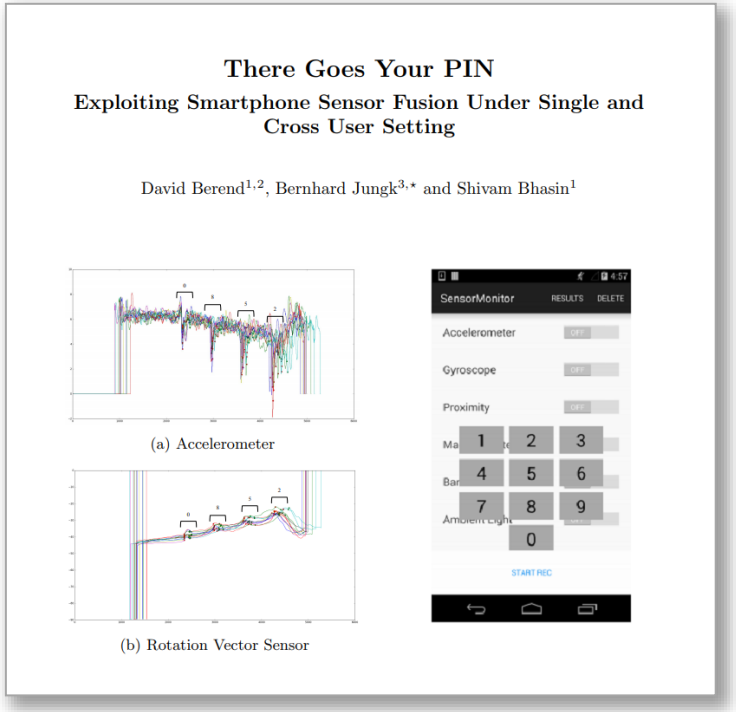
https://www.researchgate.net/publication/322295274_Behavioral_Biometrics_for_Smartphone_User_Authentication

Взлом PIN по датчикам смартфона

2017

PIN-коды не надежны

Нейросеть научили распознавать PIN-код пользователя с точностью 84%. Исследователи в области информационной безопасности использовали для определения PIN-кода смартфона данные с его датчиков. Они написали приложение для Android-смартфонов, которое собирает данные с шести датчиков (для их использования приложению не нужно получать разрешение пользователя): акселерометр, гироскоп, датчик вращения, магнитометр и датчик освещенности.



There Goes Your PIN

Exploiting Smartphone Sensor Fusion Under Single and Cross User Setting

David Berend[1,2], Bernhard Jungk[3,*] and Shivam Bhasin[1]

(a) Accelerometer

(b) Rotation Vector Sensor

https://eprint.iacr.org/2017/1169.pdf

Отпечатки взломали

2014

1

Необходимо сделать фото отпечатка в высоком разрешении – 2400 точек на дюйм. Фото отпечатка можно сделать, например, с оконного стекла

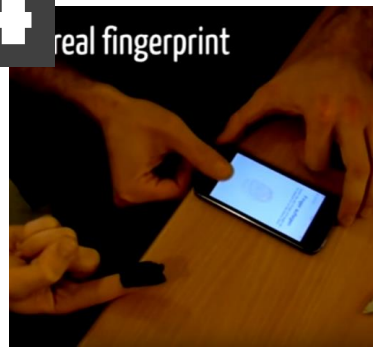
2

Изображение распечатывается в разрешением 1200 dpi на лазерном принтере на толстой бумаге

3

Отпечаток заливается жидким латексом, который после высыхания снимается и получается «факсимиле» отпечатка

4



Надетое на чужой палец «факсимиле» отпечатка воспринимается Touch ID как подушечка пальца настоящего владельца смартфона

https://www.youtube.com/watch?v=pIY6k4gvQsY

Радужку взломали

2017

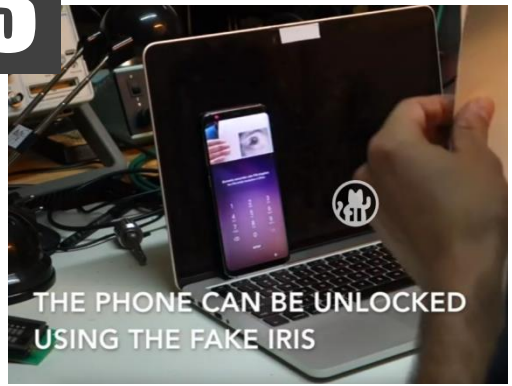
1

Необходимо сделать одну инфракрасную фотографию глаза со среднего расстояния

2

Поверх радужной оболочки на распечатке нужно поместить обычную контактную линзу

3



После этого смартфон опознает владельца и разблокирует экран

https://media.ccc.de/v/biometrie-s8-iris-en

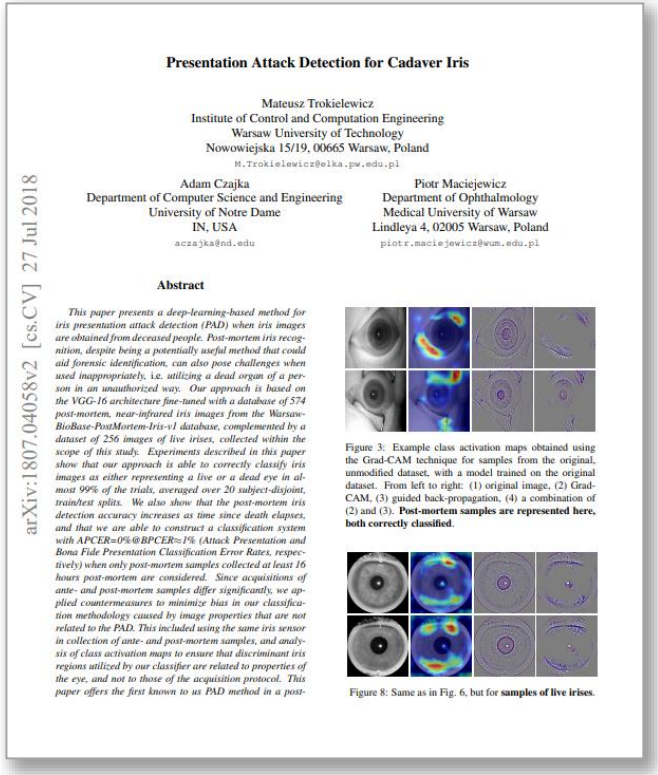
Liveness Detection для радужки

2018

DCNN

Польские программисты обучили нейросеть, которая отличает радужку мертвого человека от радужки живого с точностью до 99 процентов. Для этого они взяли предобученную популярную архитектуру VGG-16 и дообучили 3 последних слоя на 500 фотографиях.

По мнению разработчиков, в будущем это поможет избежать хакерских атак на системы идентификации пользователя по глазам.



Presentation Attack Detection for Cadaver Iris

Mateusz Trokielewicz
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska 15/19, 00665 Warsaw, Poland
M.Trokielewicz@elka.pw.edu.pl

Adam Czajka
Department of Computer Science and Engineering
University of Notre Dame
IN, USA
aczajka@nd.edu

Piotr Maciejewicz
Department of Ophthalmology
Medical University of Warsaw
Lindleya 4, 02005 Warsaw, Poland
piotr.maciejewicz@wum.edu.pl

Abstract

This paper presents a deep-learning-based method for iris presentation attack detection (PAD) when iris images are obtained from deceased people. Post-mortem iris recognition, despite being a potentially useful method that could aid forensic identification, can also pose challenges when used inappropriately, i.e. utilizing a dead organ of a person in an unauthorized way. Our approach is based on the VGG-16 architecture fine-tuned with a database of 574 post-mortem, near-infrared iris images from the Warsaw-BioBase-PostMortem-Iris-v1 database, complemented by a dataset of 256 images of live irises, collected within the scope of this study. Experiments described in this paper show that our approach is able to correctly classify iris images as either representing a live or a dead eye in almost 99% of the trials, averaged over 20 subject-disjoint, train/test splits. We also show that the post-mortem iris detection accuracy increases as time since death elapses, and that we are able to construct a classification system with APCER=0%@BPCER≈1% (Attack Presentation and Bona Fide Presentation Classification Error Rates, respectively) when only post-mortem samples collected at least 16 hours post-mortem are considered. Since acquisitions of ante- and post-mortem samples differ significantly, we applied countermeasures to minimize bias in our classification methodology caused by image properties that are not related to the PAD. This included using the same iris sensor in collection of ante- and post-mortem samples, and analysis of class activation maps to ensure that discriminant iris regions utilized by our classifier are related to properties of the eye, and not to those of the acquisition protocol. This paper offers the first known to us PAD method in a post-

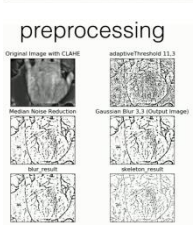
Figure 3: Example class activation maps obtained using the Grad-CAM technique for samples from the original, unmodified dataset, with a model trained on the original dataset. From left to right: (1) original image, (2) Grad-CAM, (3) guided back-propagation, (4) a combination of (2) and (3). **Post-mortem samples are represented here, both correctly classified**.

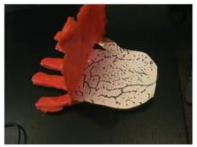
Figure 8: Same as in Fig. 6, but for **samples of live irises**.

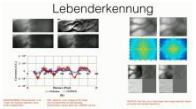
https://arxiv.org/pdf/1807.04058.pdf

Рисунок вен взломали

2019







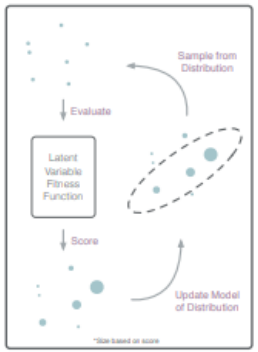
https://media.ccc.de/v/35c3-9545-venenerkennung_hacken#t=2377

Отпечатки не панацея...

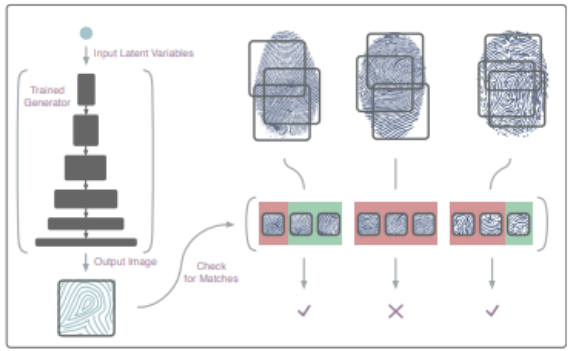
2018

Deep MasterPrints

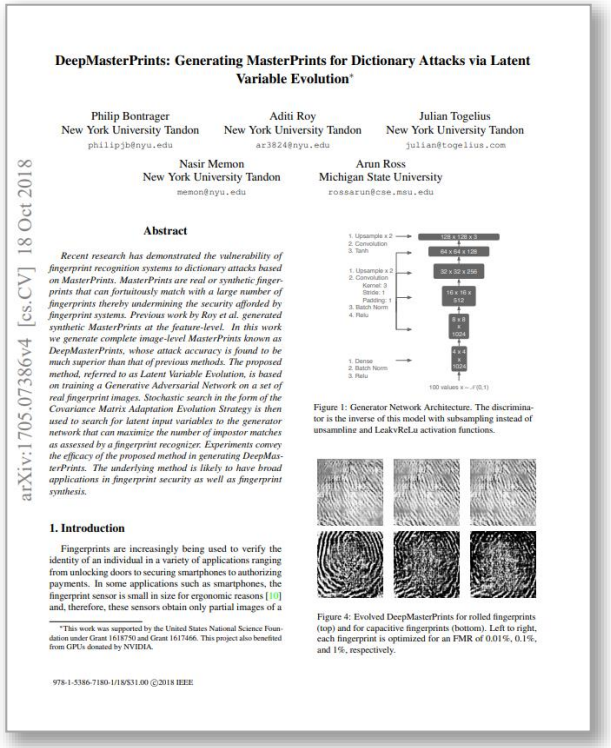
Ресерчеры из NYU разработали нейросеть, которая создает поддельные отпечатки пальцев. Метод работает по принципу «атаки словаря» против паролей, когда хакер подбирает пароль по заранее сгенерированному списку общих паролей к системе безопасности.



CMA-ES Optimization



Latent Variable Fitness Function



arXiv:1705.07386v4 [cs.CV] 18 Oct 2018

DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution*

Philip Bontrager, New York University Tandon, philip3b@nyu.edu
Aditi Roy, New York University Tandon, ar3824@nyu.edu
Julian Togelius, New York University Tandon, julian@togelius.com
Nasir Memon, New York University Tandon, memon@nyu.edu
Arun Ross, Michigan State University, rossarun@cse.msu.edu

Abstract

Recent research has demonstrated the vulnerability of fingerprint recognition systems to dictionary attacks based on MasterPrints. MasterPrints are real or synthetic fingerprints that can fortuitously match with a large number of fingerprints thereby undermining the security afforded by fingerprint systems. Previous work by Roy et al. generated synthetic MasterPrints at the feature-level. In this work we generate complete image-level MasterPrints known as DeepMasterPrints, whose attack accuracy is found to be much superior than that of previous methods. The proposed method, referred to as Latent Variable Evolution, is based on training a Generative Adversarial Network on a set of real fingerprint images. Stochastic search in the form of the Covariance Matrix Adaptation Evolution Strategy is then used to search for latent input variables to the generator network that can maximize the number of impostor matches as assessed by a fingerprint recognizer. Experiments convey the efficacy of the proposed method in generating DeepMasterPrints. The underlying method is likely to have broad applications in fingerprint security as well as fingerprint synthesis.

1. Introduction

Fingerprints are increasingly being used to verify the identity of an individual in a variety of applications ranging from unlocking doors to securing smartphones to authorizing payments. In some applications such as smartphones, the fingerprint sensor is small in size for ergonomic reasons [10] and, therefore, these sensors obtain only partial images of a

Figure 1: Generator Network Architecture. The discriminator is the inverse of this model with subsampling instead of unsampling and LeakyReLu activation functions.

Figure 4: Evolved DeepMasterPrints for rolled fingerprints (top) and for capacitive fingerprints (bottom). Left to right, each fingerprint is optimized for an FMR of 0.01%, 0.1%, and 1%, respectively.

*This work was supported by the United States National Science Foundation under Grant 1618750 and Grant 1617466. This project also benefited from GPUs donated by NVIDIA.

978-1-5386-7180-1/18/$31.00 ©2018 IEEE

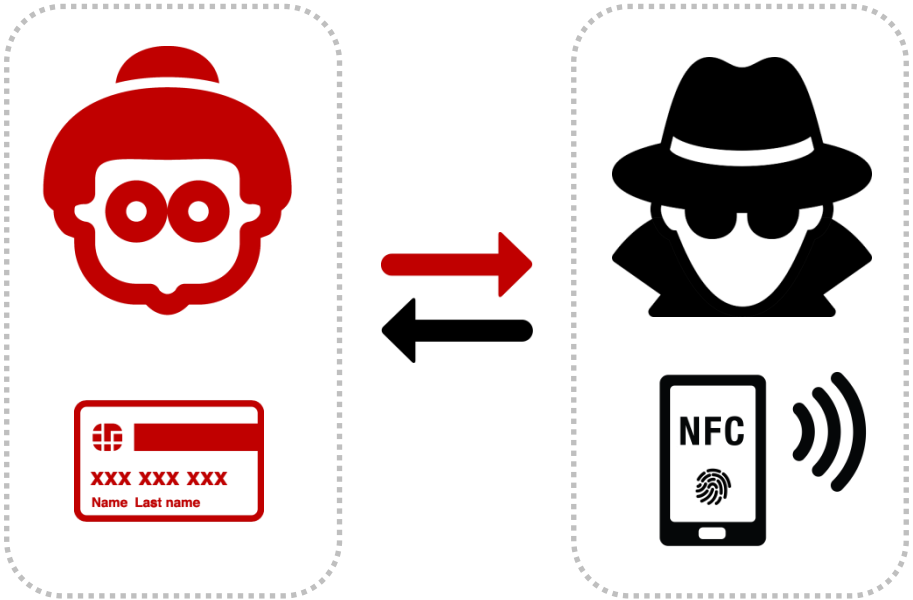
https://arxiv.org/pdf/1705.07386.pdf

… и помогают мошенникам

2015

Fraud Apple Pay

Мошенники привязывают к своим Apple-кошелькам скомпрометированные карты клиентов. Данные по картам, включая одноразовые SMS-пароли, злоумышленники получают методами социальной инженерии. После этого мошенники могут беспрепятственно использовать свой iPhone для оплаты товаров и услуг в магазинах.



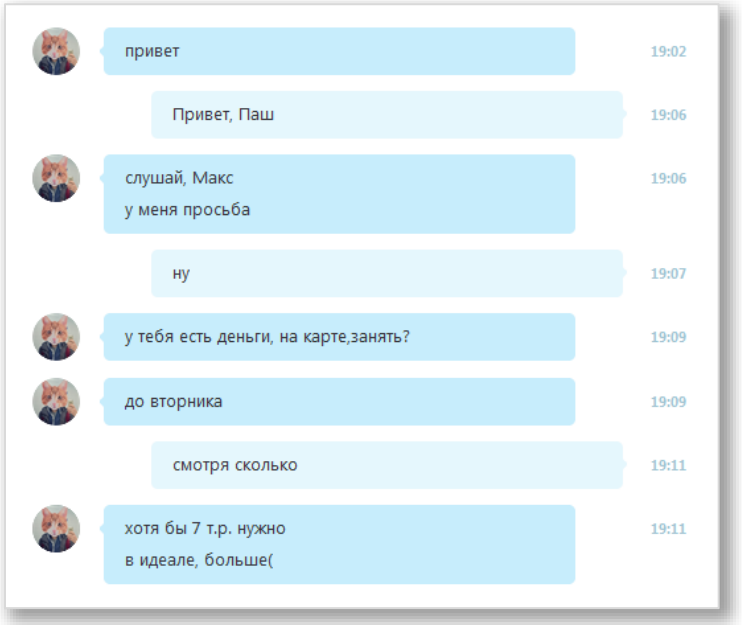
Социальная инженерия сейчас

2015

Интервью со Skype-мошенником

https://habrahabr.ru/post/255225/

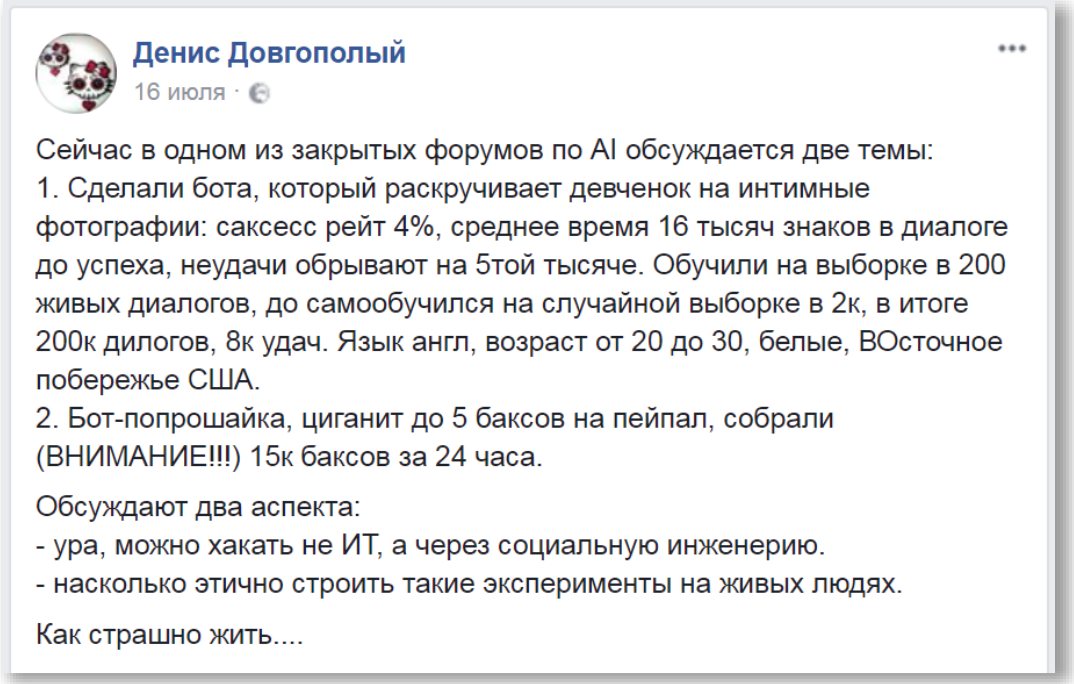
1. Нанимают студентов
2. Обучают их психологии
3. Выдают ворованные Skype-логины
4. Платят **50%** с выручки
5. **Profit**



Социальная инженерия — скоро

2017

… или сейчас?



Денис Довгополый
16 июля ·

Сейчас в одном из закрытых форумов по AI обсуждается две темы:
1. Сделали бота, который раскручивает девченок на интимные фотографии: саксесс рейт 4%, среднее время 16 тысяч знаков в диалоге до успеха, неудачи обрывают на 5той тысяче. Обучили на выборке в 200 живых диалогов, до самообучился на случайной выборке в 2к, в итоге 200к дилогов, 8к удач. Язык англ, возраст от 20 до 30, белые, ВОсточное побережье США.
2. Бот-попрошайка, циганит до 5 баксов на пейпал, собрали (ВНИМАНИЕ!!!) 15к баксов за 24 часа.

Обсуждают два аспекта:
- ура, можно хакать не ИТ, а через социальную инженерию.
- насколько этично строить такие эксперименты на живых людях.

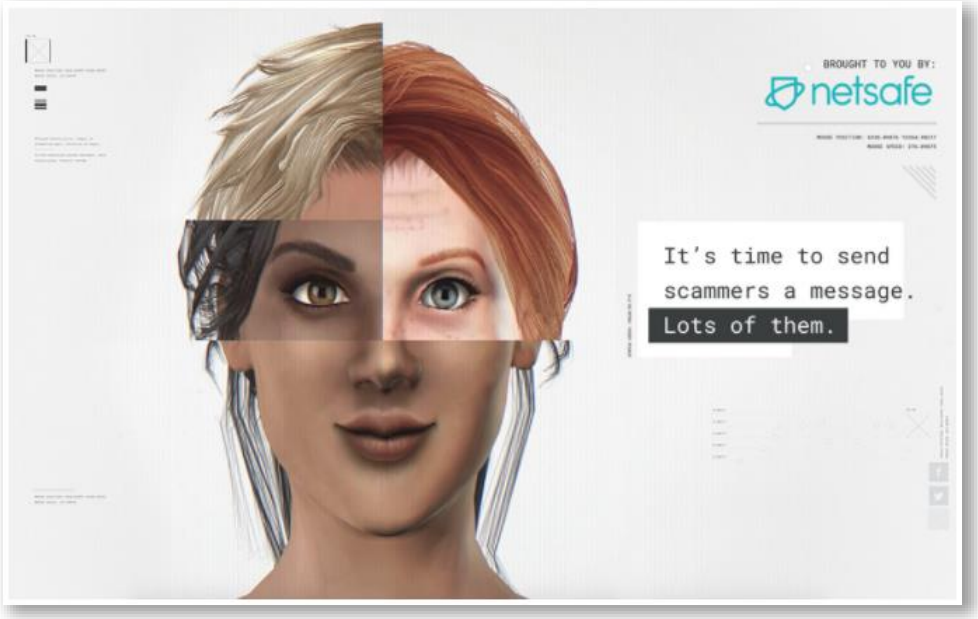
Как страшно жить....

Чат-бот «Отвлекайка»

2017

RE:SCAM

Чат-бот умеет распознавать тип фишинговых писем, подстраиваться под них и задавать мошенникам бесконечную серию вопросов. Чат-бот выражает заинтересованность в теме «нигерийских» писем, пытаясь узнать подробности, шутить и казаться «наивным». По мнению Netsafe, такое общение должно снизить эффективность мошенников — у них не будет времени на «живых» пользователей.

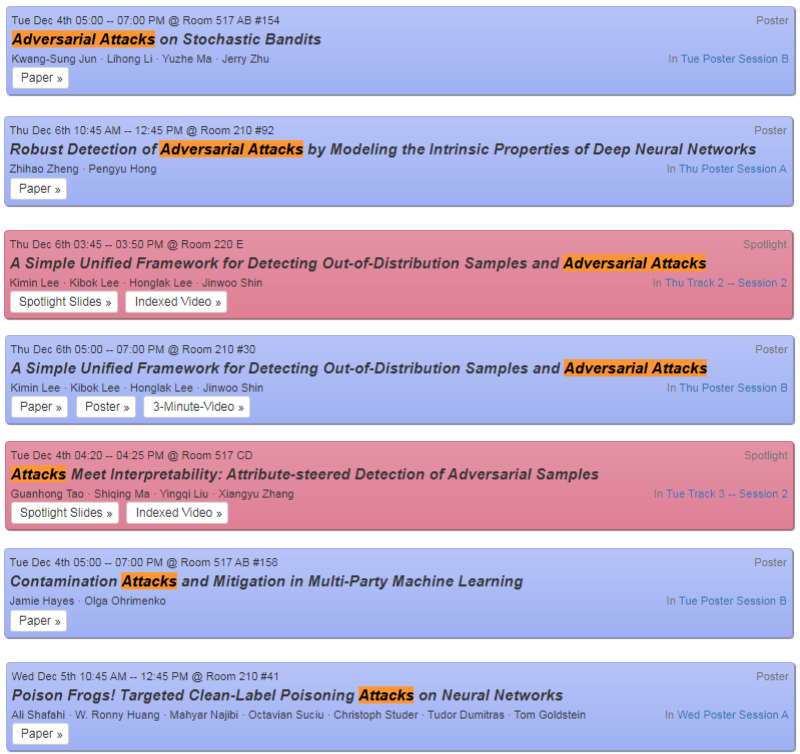


2017

«Следующие 10 лет мы будем не только учиться создавать искусственный интеллект, но и обманывать его, делать анти-обманные системы и так далее. Война предстоит не хуже, чем вирусно-антивирусная, готовьтесь!»

Григорий Бакунов
Яндекс

NeurIPS 2018-2019



Tue Dec 4th 05:00 -- 07:00 PM @ Room 517 AB #154 — Poster
Adversarial Attacks on Stochastic Bandits
Kwang-Sung Jun · Lihong Li · Yuzhe Ma · Jerry Zhu — In Tue Poster Session B
Paper »

Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 #92 — Poster
Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks
Zhihao Zheng · Pengyu Hong — In Thu Poster Session A
Paper »

Thu Dec 6th 03:45 -- 03:50 PM @ Room 220 E — Spotlight
A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks
Kimin Lee · Kibok Lee · Honglak Lee · Jinwoo Shin — In Thu Track 2 -- Session 2
Spotlight Slides » Indexed Video »

Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 #30 — Poster
A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks
Kimin Lee · Kibok Lee · Honglak Lee · Jinwoo Shin — In Thu Poster Session B
Paper » Poster » 3-Minute-Video »

Tue Dec 4th 04:20 -- 04:25 PM @ Room 517 CD — Spotlight
Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples
Guanhong Tao · Shiqing Ma · Yingqi Liu · Xiangyu Zhang — In Tue Track 3 -- Session 2
Spotlight Slides » Indexed Video »

Tue Dec 4th 05:00 -- 07:00 PM @ Room 517 AB #158 — Poster
Contamination Attacks and Mitigation in Multi-Party Machine Learning
Jamie Hayes · Olga Ohrimenko — In Tue Poster Session B
Paper »

Wed Dec 5th 10:45 AM -- 12:45 PM @ Room 210 #41 — Poster
Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks
Ali Shafahi · W. Ronny Huang · Mahyar Najibi · Octavian Suciu · Christoph Studer · Tudor Dumitras · Tom Goldstein — In Wed Poster Session A
Paper »

Thu Dec 12th 05:00 -- 07:00 PM @ East Exhibition Hall B + C #133 — Poster
Defending Against Neural Fake News
Rowan Zellers · Ari Holtzman · Hannah Rashkin · Yonatan Bisk · Ali Farhadi · Franziska Roesner · Yejin Choi — In Applications -- Natural Language Processing
Paper » Poster »

Tue Dec 10th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #8 — Poster
Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training
Haichao Zhang · Jianyu Wang — In Algorithms -- Adversarial Learning
Paper » Poster »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #15 — Poster
Functional Adversarial Attacks
Cassidy Laidlaw · Soheil Feizi — In Algorithms -- Adversarial Learning
Paper » Poster »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #16 — Poster
Improving Black-box Adversarial Attacks with a Transfer-based Prior
Shuyu Cheng · Yinpeng Dong · Tianyu Pang · Hang Su · Jun Zhu — In Algorithms -- Adversarial Learning
Paper » Slides »

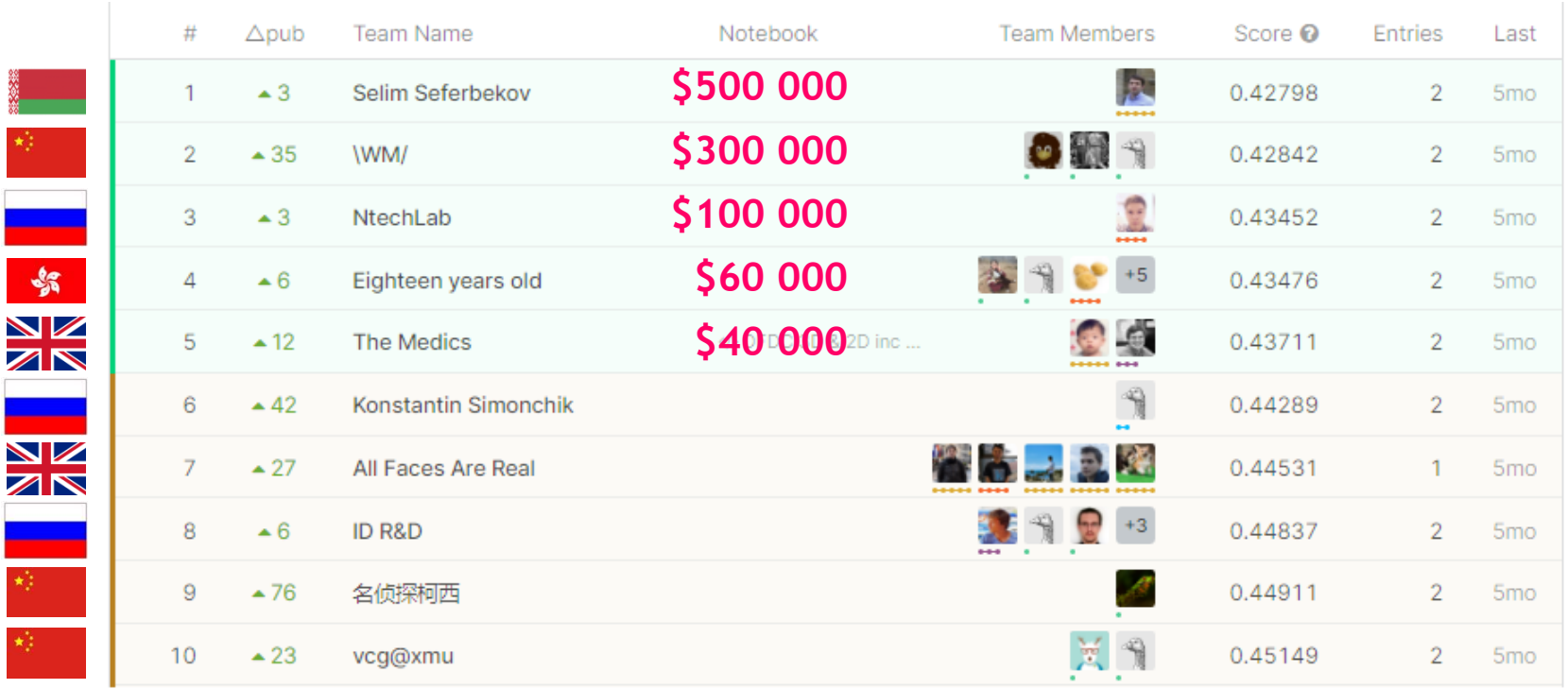
Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #23 — Poster
Provably robust boosted decision stumps and trees against adversarial attacks
Maksym Andriushchenko · Matthias Hein — In Algorithms -- Adversarial Learning
Paper » Poster »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #98 — Poster
Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks
Lixin Fan · Kam Woh Ng · Chee Seng Chan — In Applications -- Privacy, Anonymity, and Security
Paper » Poster » Slides »

Wed Dec 11th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #24 — Poster
A Unified Framework for Data Poisoning Attack to Graph-based Semi-supervised Learning
Xuanqing Liu · Si Si · Jerry Zhu · Yang Li · Cho-Jui Hsieh — In Algorithms -- Semi-Supervised Learning
Paper » Poster »

$1,000,000 Deepfake Detection Challenge

2020



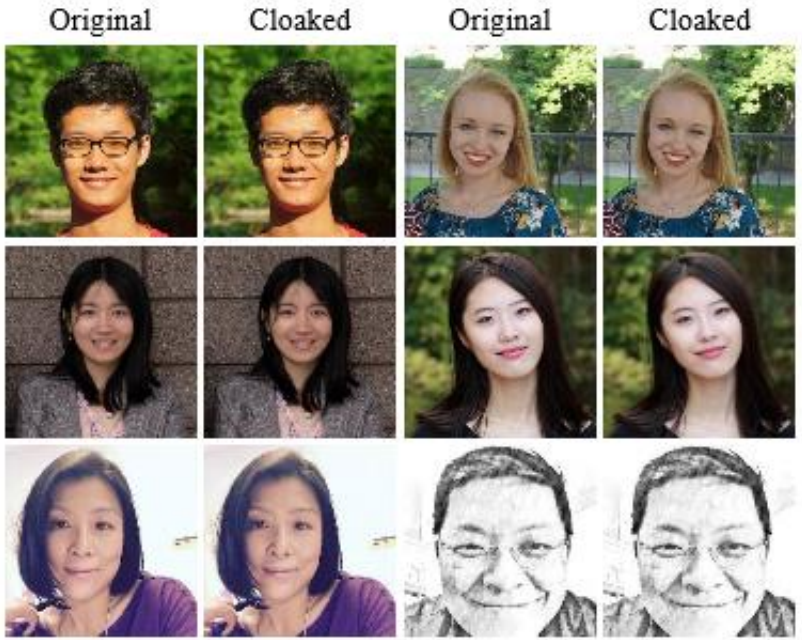
	#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
	1	▲ 3	Selim Seferbekov	$500 000		0.42798	2	5mo
	2	▲ 35	\WM/	$300 000		0.42842	2	5mo
	3	▲ 3	NtechLab	$100 000		0.43452	2	5mo
	4	▲ 6	Eighteen years old	$60 000	+5	0.43476	2	5mo
	5	▲ 12	The Medics	$40 000		0.43711	2	5mo
	6	▲ 42	Konstantin Simonchik			0.44289	2	5mo
	7	▲ 27	All Faces Are Real			0.44531	1	5mo
	8	▲ 6	ID R&D		+3	0.44837	2	5mo
	9	▲ 76	名侦探柯西			0.44911	2	5mo
	10	▲ 23	vcg@xmu			0.45149	2	5mo

Маскировщик фотографий

2020

Fawkes

Чикагские инженеры разработали инструмент под названием Fawkes (Фокс), который маскирует фотографии для защиты от фотобиометрических систем. Исследователи смогли обмануть системы распознавания лиц от Amazon, Microsoft и китайской технологической компании Megvii.



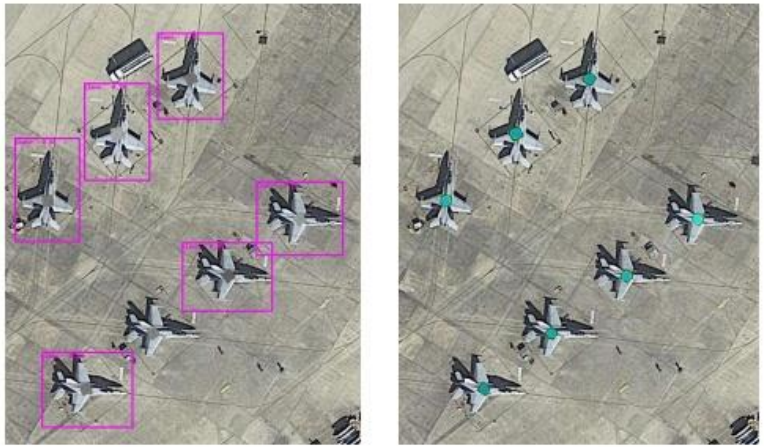
http://sandlab.cs.uchicago.edu/fawkes/

Маскировка военной техники

2020

YOLOv2 вместо камуфляжа

Нидерландские инженеры обучили нейросеть YOLOv2 создающую для самолетов специальную защиту (патчи) от систем распознавания военной техники по спутниковым снимкам.
Эксперименты на спутниковых снимках показали, что большие изображения снижают среднюю точность алгоритма обнаружения с 94% до 5,6%, а маленькие – до 37,8%. Изображение можно накладывать и на соседнюю часть взлетно-посадочной полосы, а не на сам самолет.



(a) Random patch (b) Adversarial patch

https://arxiv.org/pdf/2008.13671.pdf

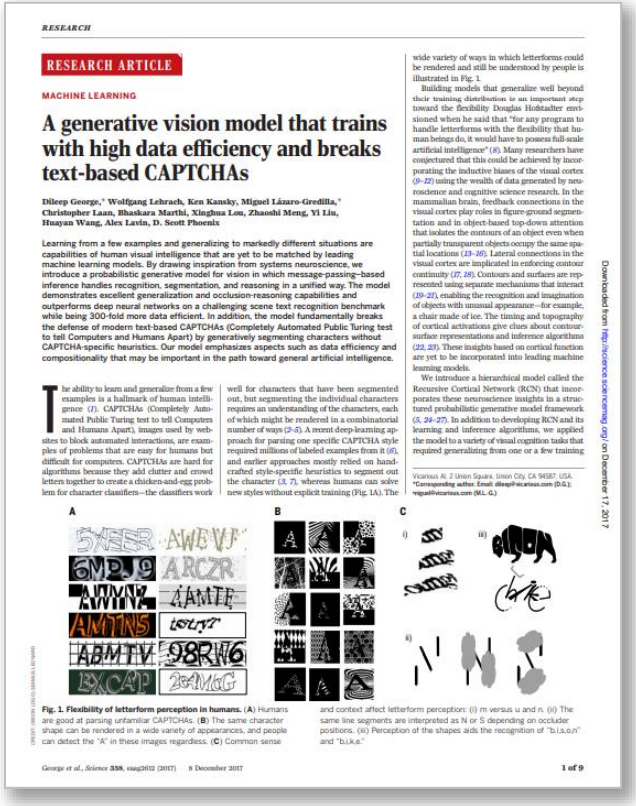
Мультимодальность – будущее биометрии?

Капчу взломали! Расходимся!

2017

RCN vs. CAPTCHA

Разработчики американской компании Vicarious создали алгоритм, эффективно расшифровывающий капчу — самый распространенный на сегодняшний день способ отличить человека от робота. Такой алгоритм работает на основе компьютерного зрения и рекурсивной кортикальной нейросети, и может расшифровать капчу на многих популярных интернет-платформах, в том числе на PayPal и Yahoo. Работа опубликована в журнале Science.



RESEARCH

RESEARCH ARTICLE

MACHINE LEARNING

A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs

Dileep George,* Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla,* Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, Alex Lavin, D. Scott Phoenix

Learning from a few examples and generalizing to markedly different situations are capabilities of human visual intelligence that are yet to be matched by leading machine learning models. By drawing inspiration from systems neuroscience, we introduce a probabilistic generative model for vision in which message-passing–based inference handles recognition, segmentation, and reasoning in a unified way. The model demonstrates excellent generalization and occlusion-reasoning capabilities and outperforms deep neural networks on a challenging scene text recognition benchmark while being 300-fold more data efficient. In addition, the model fundamentally breaks the defense of modern text-based CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) by generatively segmenting characters without CAPTCHA-specific heuristics. Our model emphasizes aspects such as data efficiency and compositionality that may be important in the path toward general artificial intelligence.

The ability to learn and generalize from a few examples is a hallmark of human intelligence (*1*). CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart), images used by websites to block automated interactions, are examples of problems that are easy for humans but difficult for computers. CAPTCHAs are hard for algorithms because they add clutter and crowd letters together to create a chicken-and-egg problem for character classifiers—the classifiers work well for characters that have been segmented out, but segmenting the individual characters requires an understanding of the characters, each of which might be rendered in a combinatorial number of ways (*2*–*5*). A recent deep-learning approach for parsing one specific CAPTCHA style required millions of labeled examples from it (*6*), and earlier approaches mostly relied on hand-crafted style-specific heuristics to segment out the character (*3*, *7*), whereas humans can solve new styles without explicit training (Fig. 1A). The wide variety of ways in which letterforms could be rendered and still be understood by people is illustrated in Fig. 1.

Building models that generalize well beyond their training distribution is an important step toward the flexibility Douglas Hofstadter envisioned when he said that "for any program to handle letterforms with the flexibility that human beings do, it would have to possess full-scale artificial intelligence" (*8*). Many researchers have conjectured that this could be achieved by incorporating the inductive biases of the visual cortex (*9*–*22*) using the wealth of data generated by neuroscience and cognitive science research. In the mammalian brain, feedback connections in the visual cortex play roles in figure-ground segmentation and in object-based top-down attention that isolates the contours of an object even when partially transparent objects occupy the same spatial locations (*13*–*16*). Lateral connections in the visual cortex are implicated in enforcing contour continuity (*17*, *18*). Contours and surfaces are represented using separate mechanisms that interact (*19*–*21*), enabling the recognition and imagination of objects with unusual appearance—for example, a chair made of ice. The timing and topography of cortical activations give clues about contour-surface representations and inference algorithms (*22*, *23*). These insights based on cortical function are yet to be incorporated into leading machine learning models.

We introduce a hierarchical model called the Recursive Cortical Network (RCN) that incorporates these neuroscience insights in a structured probabilistic generative model framework (*5*, *24*–*27*). In addition to developing RCN and its learning and inference algorithms, we applied the model to a variety of visual cognition tasks that required generalizing from one or a few training

Vicarious AI, 2 Union Square, Union City, CA 94587, USA.
*Corresponding author. Email: dilee@vicarious.com (D.G.); rajan@vicarious.com (M.L.-G.)

Fig. 1. Flexibility of letterform perception in humans. (A) Humans are good at parsing unfamiliar CAPTCHAs. (B) The same character shape can be rendered in a wide variety of appearances, and people can detect the "A" in these images regardless. (C) Common sense and context affect letterform perception: (i) m versus u and n. (ii) The same line segments are interpreted as N or S depending on occluder positions. (iii) Perception of the shapes aids the recognition of "b,i,s,o,n" and "b,i,k,e."

Спасибо за внимание!

Афанасьев Сергей

Исполнительный директор
Начальник управления
статистического анализа

КБ «Ренессанс Кредит»

svafanasev@gmail.com
safanasev@rencredit.ru