FACESEC:

A Fine-Grained Robustness Evaluation Framework for Face Recognition Systems

Liang Tong Zhengzhang Chen Jingchao Ni Wei Cheng

Dongjin Song Haifeng Chen Yevgeniy Vorobeychik

Washington University in St. Louis

NEC Laboratories America

Relentless passion for innovation

University of Connecticut

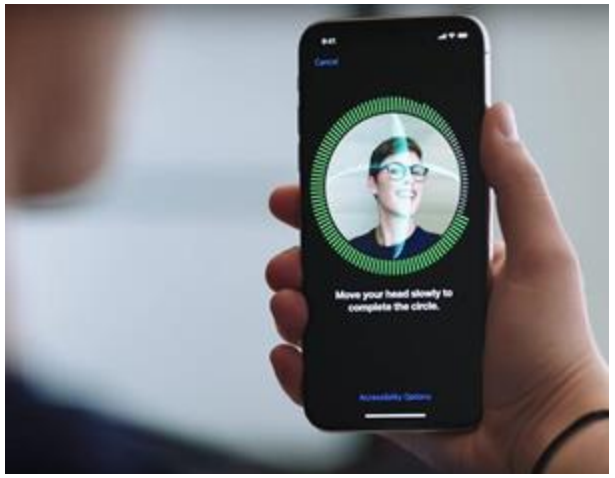
DL-Based Face Recognition in Daily Life

(Source: www.businessinsider.com)

Homeland security

(Source: www.nec.com)

Financial services

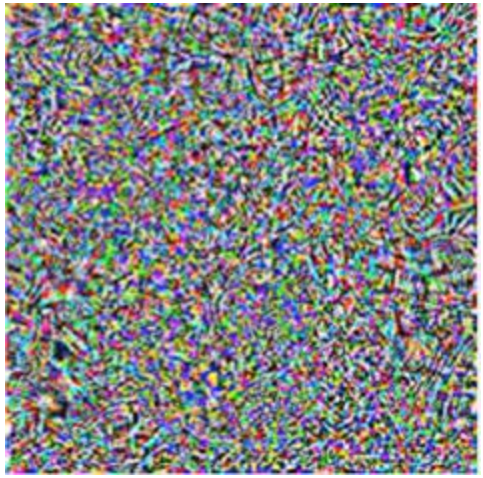


(Source: www.apple.com)

IoT devices

DL Models Are Not Robust

“Pig”



Small adversarial noise

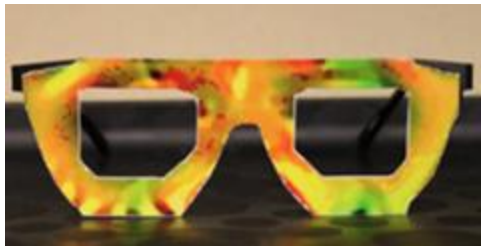
“Airliner”

(Goodfellow et al., ICLR'14)

DL models are inherently vulnerable to adversarial examples

Face Recognition Systems Are Not Secure

“Bob”



Adversarial eyeglass frame

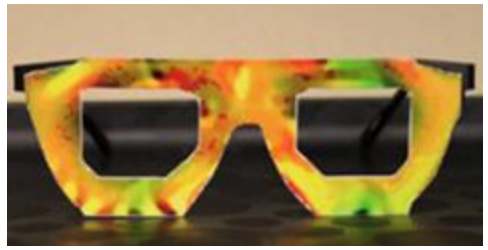
“Alice”

(Sharif et al., CCS'16)

Adversarial examples can be realizable in physical space

Face Recognition Systems Are Not Secure

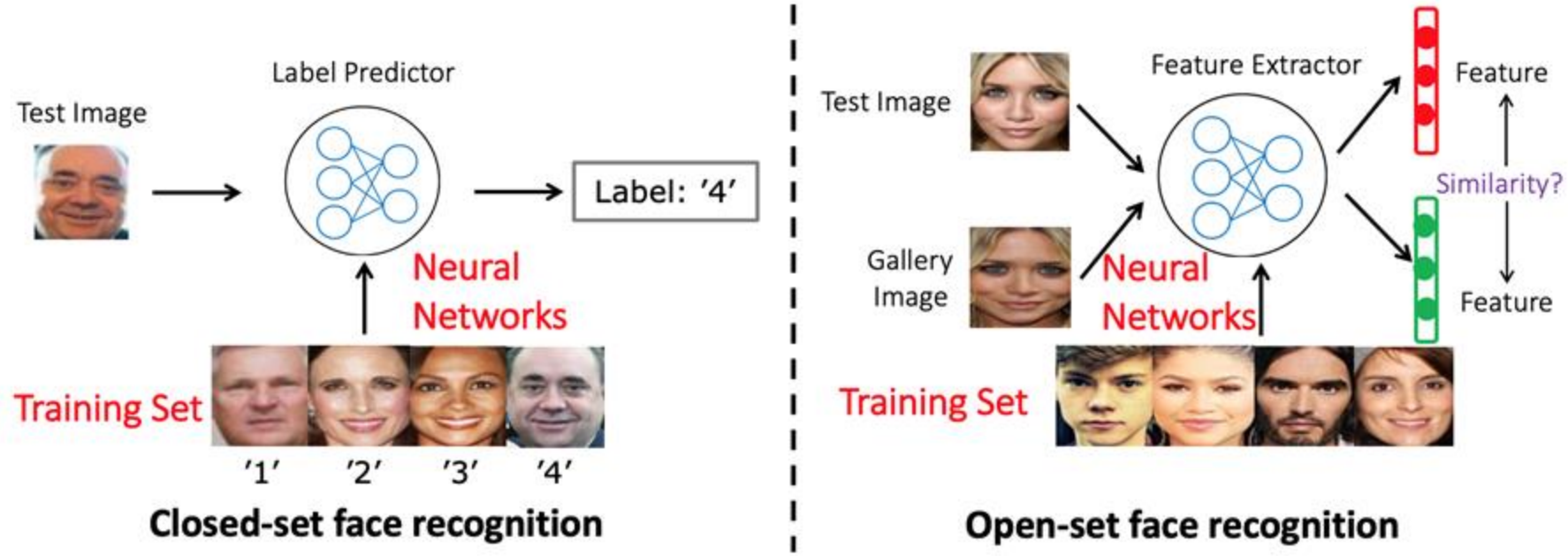
(Source: www.faception.com)



(Source: www. en.wikipedia.org)

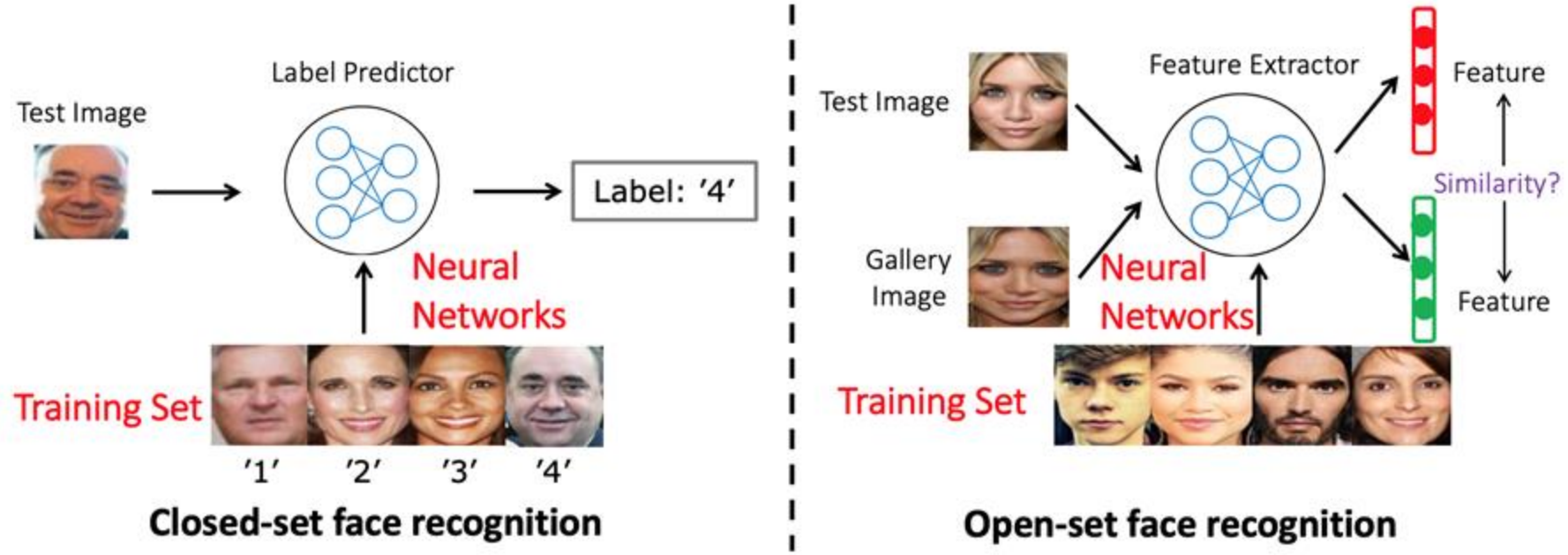
It is critical to evaluate robustness of face recognition systems in adversarial settings

Robustness Evaluation Is Challenging



- Lack of understanding which ***individual*** or ***combination*** of components is vulnerable to adversarial examples
  - *Different training sets and neural architectures result in different performance and robustness*
  - *Existing approaches: only use **white-box** or **black-box** attack for evaluation*

Robustness Evaluation Is Challenging



- Lack of understanding ***different levels of robustness*** corresponding to different types of attacks
  - *Attackers vary by perturbation types, goals, knowledge, and capabilities*
  - *Existing approaches: only use **a specific type** of attack (e.g., digital attack)*

Robustness Evaluation Is Challenging

Our main contribution:
A fine-grained robustness evaluation framework for face recognition systems

- Lack of understanding ***different levels of risks*** corresponding to different types of attacks
  - *Attackers vary by perturbation types, goals, knowledge, and capabilities*
  - *Existing work: only uses a certain type of attacks*

FaceSec - Framework

S: target system to be evaluated

G: attacker's goal
- *Dodging (non-targeted)*
- *Impersonation (targeted)*

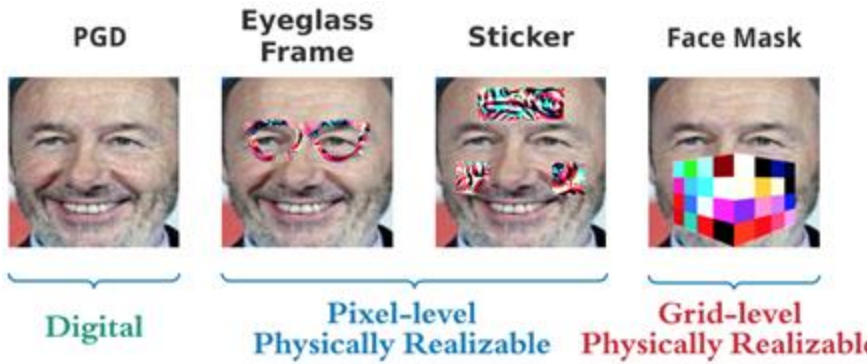
K: attacker's knowledge about S
- *Zero knowledge*
- *Training set*
- *Neural architecture*
- *Full knowledge*

Ensemble-based method w/momentum

Robustness = Evaluate(S, <P, K, G, C>)

C: attacker's capability
- *Individual attack for each image*
- ***Batch-based universal attack***

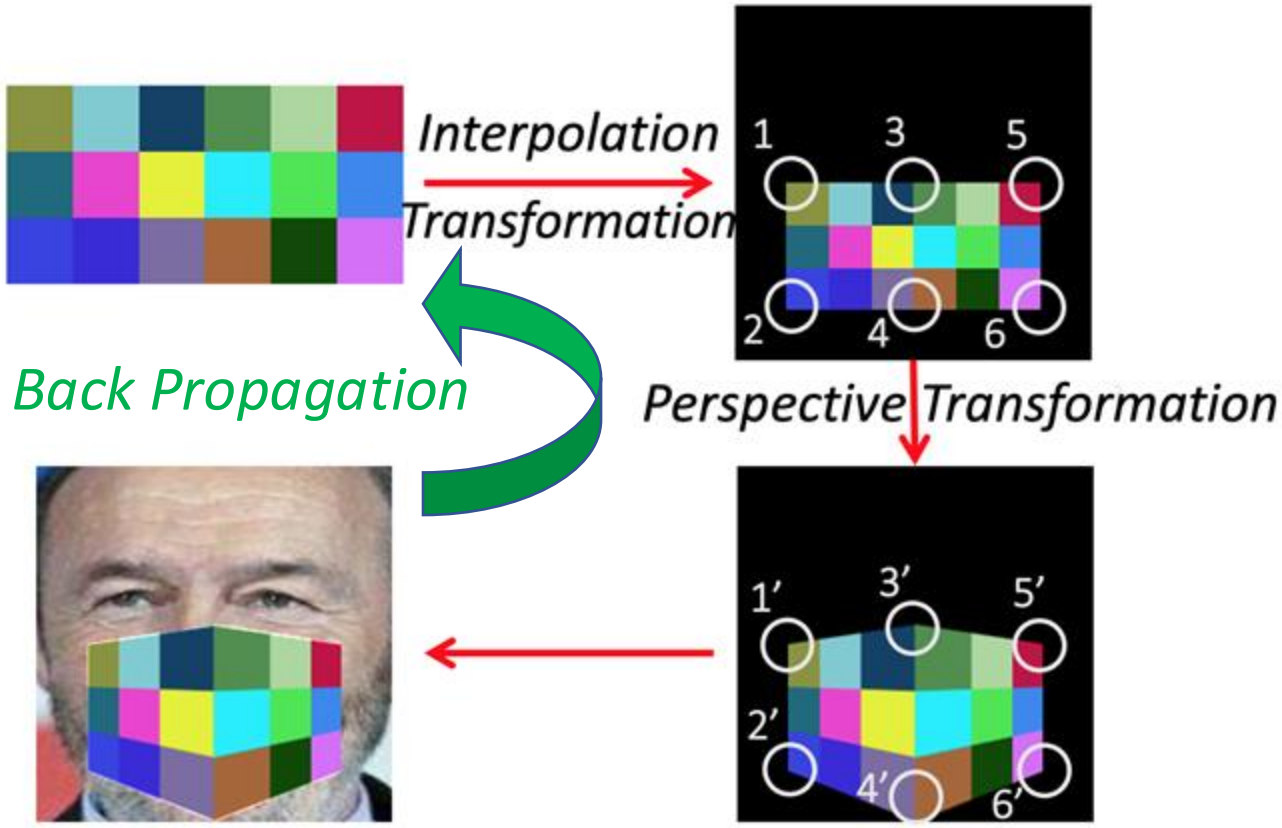
P: perturbation type



FaceSec – Face Mask Attack

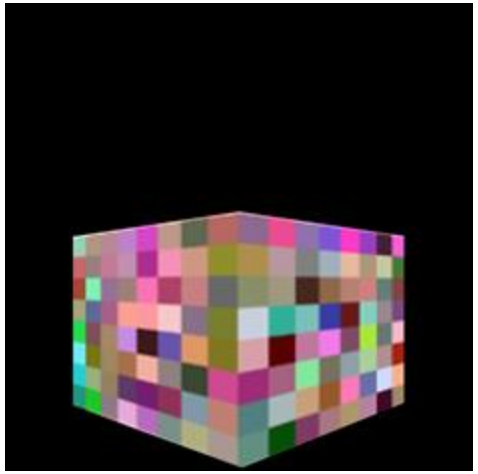
$$\max_{\delta} L(S(x * (1 - f) + T(\delta) * f), y)$$

- L: *loss function*
- S: *target system*
- x: *input image*
- y: *label*
- δ: *grid level color matrix*
- T: *a sequence of transformations*
- f: *areas where perturbation is allowed*



FaceSec – Universal Attack

Input batch



Face-agnostic perturbation

Output batch

$$\delta = arg \max_{\delta'} min\{L(S(x_i * (1 - f) + \delta' * f), y_i)\}_{i=1}^{N}$$

A general approach that works for both digital and physically realizable attacks

Experimental Results

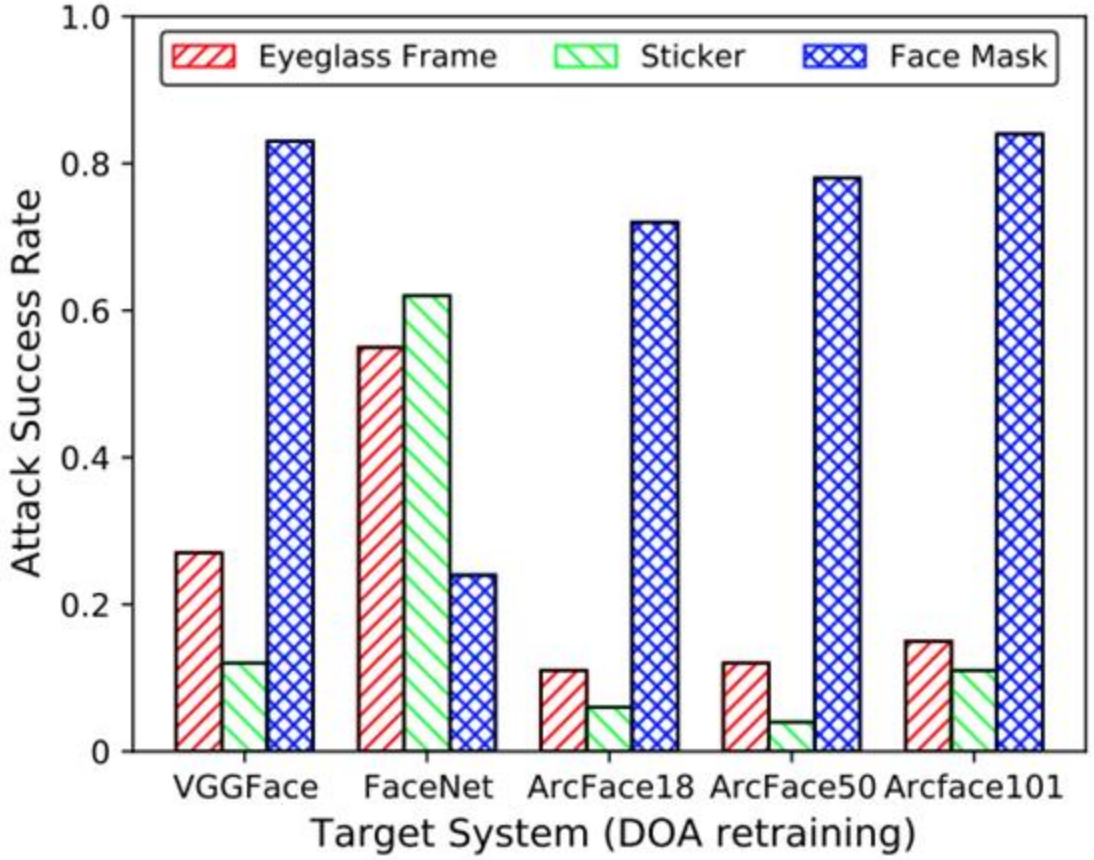
Attack Success Rate of Dodging Face Mask Attacks

Target System	Attacker's System Knowledge			
	Zero knowledge	Training set	Neural architecture	Full knowledge
VGGFace	0.26	0.32	0.63	1.00
FaceNet	0.30	0.42	0.83	1.00
ArcFace18	0.27	0.33	0.71	1.00
ArcFace50	0.29	0.36	0.67	0.99
ArcFace101	0.26	0.36	0.54	0.99

Accurate knowledge of neural architecture is significantly more important than training data in black-box attacks

Experimental Results

Dodging Attacks VS. SOTA Defense



1. ***Adversarial robustness highly depends on both the nature of perturbation and the neural architecture***
2. ***Adversarial robustness against one type of perturbation may not be generalized to other types***

Conclusion

- FACESEC is ***the first*** that supports to evaluate the risks of different components of face recognition systems from multiple dimensions and under various adversarial settings
- FACESEC can work for both ***closed-set*** and ***open-set*** systems
- Our systematic and comprehensive evaluations demonstrate that FACESEC can greatly help understand the robustness of face recognition systems

Thank you!

Liang Tong Zhengzhang Chen Jingchao Ni Wei Cheng
Dongjin Song Haifeng Chen Yevgeniy Vorobeychik