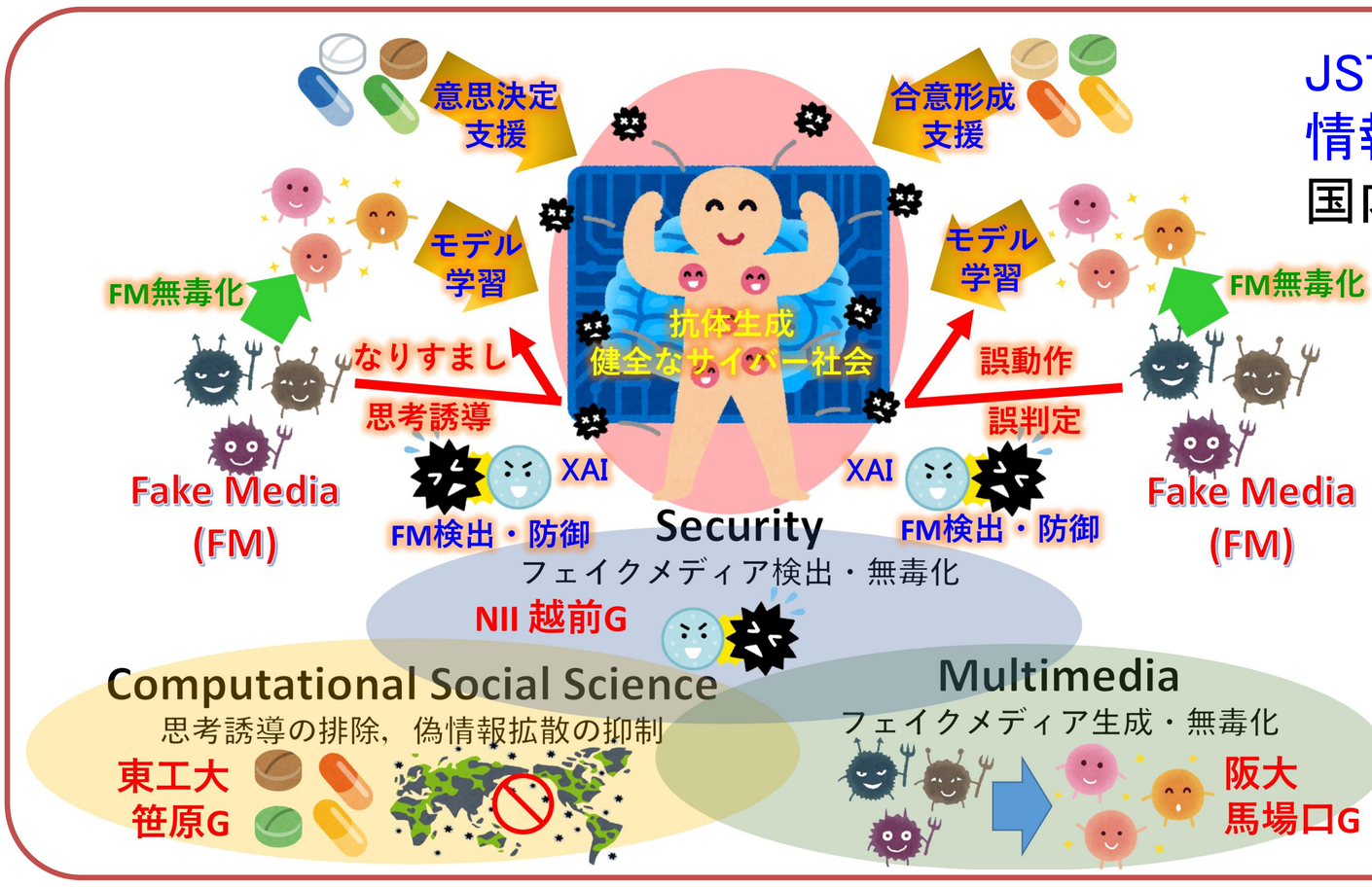
大学院 | 情報理工・電子情報学専攻 | 一ツ橋

越前研究室(Prof. Isao Echizen)
Echizen Laboratory

URL:http://research.nii.ac.jp/~iechizen/official/

ELAB Multimedia Security

越前研究室では，マルチメディアと情報セキュリティ，マルチメディアとプライバシー保護の融合領域に焦点をあてながら，基礎論から応用システムまでの幅広い研究分野に取り組んでいます．世界中から研究者を受け入れている国際色豊かな研究室です．



JST CREST インフォデミックを克服するソーシャル情報基盤技術（CREST FakeMedia）に参画しており，国内外の研究者と活発に交流を行っています．

研究実施項目

- 多様なモダリティによる高度なFM生成技術
- FM検出・防御技術
- FM無毒化技術
- インフォデミックを緩和し多様な意思決定を支援する情報技術

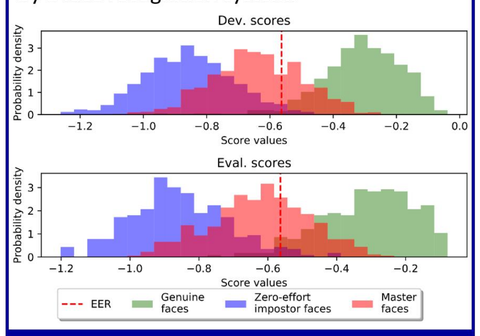
CREST FakeMedia

攻撃側 (Attacker side)

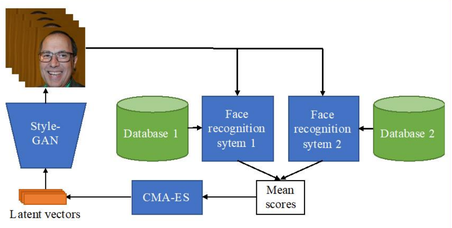
マスター顔（Master face）による生体認証システムへのウルフ攻撃

Background

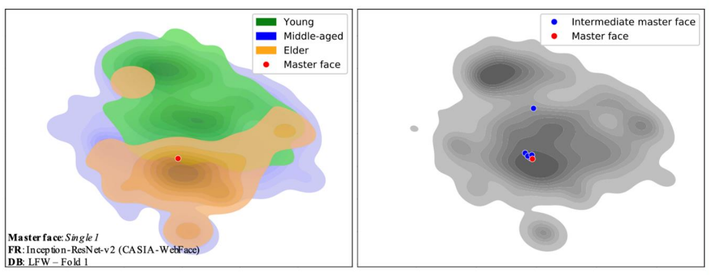
The **first work** to generate a **master face** (or a wolf face) which matches with multiple faces by a face recognition system.



Proposed Method



Results



Location of the master face in the latent space of a face recognition system

敵対的サンプルの転移性の分析

Background

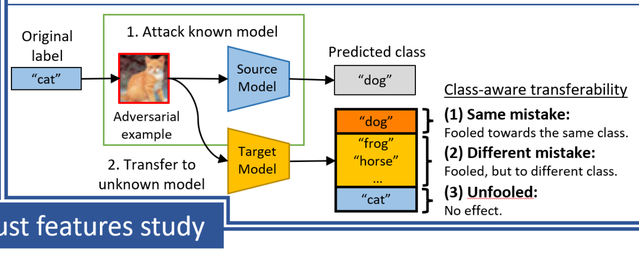
- Adversarial Examples (AEs) transfer among different models, i.e., **adversarial transferability**.
- This nature significantly increases the risk in our society.
- However, its mechanism is still not well understood.

Research questions

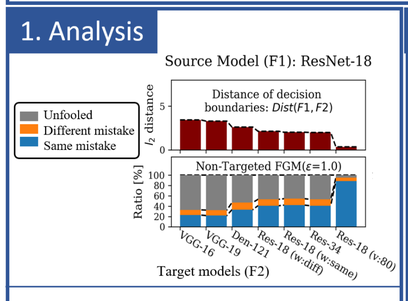
1. Towards which class the models' predictions are misled?
2. What are the mechanisms that AEs cause same mistakes or different mistakes?

Novel metric: Class-aware transferability

- We classify adversarial transferability into three cases, "same mistakes", "different mistakes" and "unfooled".



1. Analysis



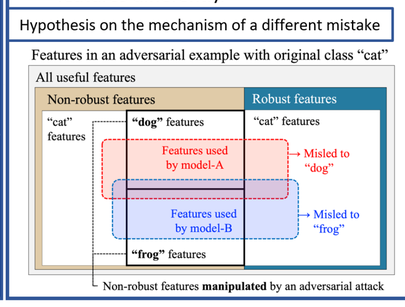
Finding 1: Majority of the fooled cases are "same mistakes". => AEs have effects to fool models towards specific classes.

Finding 2: "Different mistakes" occur even between very similar models.

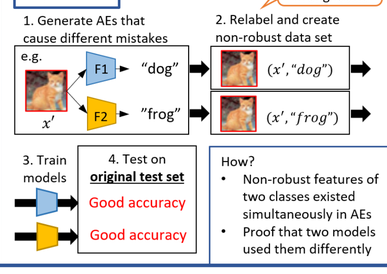
2. Non-robust features study

- AEs can cause "same mistakes" by containing "non-robust features", which are imperceptible but useful features (Ilyas et al. 2019).
- We show that "different mistakes" occur when different models use non-robust features differently.

Hypothesis on the mechanism of a different mistake



Experiment



防御側 (Defender side)

Cyber Vaccination: Deepfake映像の無毒化手法

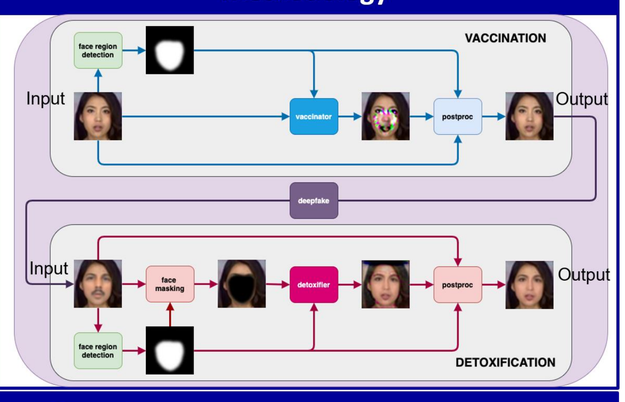
Background

- Face-swapping has imposed identity threats to the society and therefore it would be favourable if the modifications can be reversed.

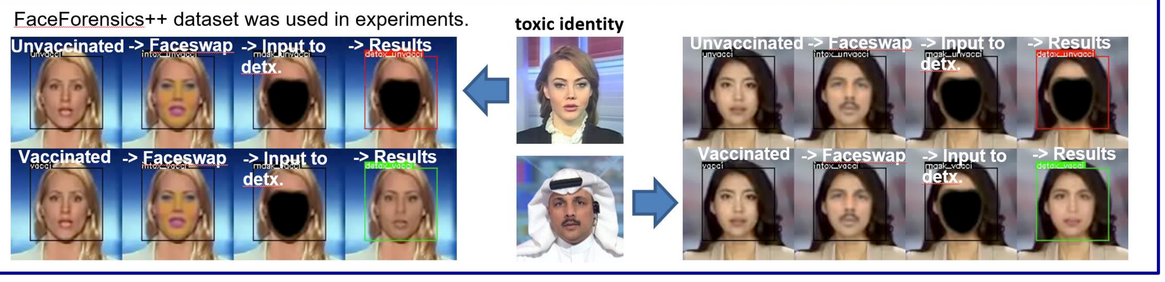
Contributions

- Attack-agnostic deep-learning framework for face recovery.
- Identification of unvaccinated and vaccinated images/videos.
- Detection of toxic images/videos.

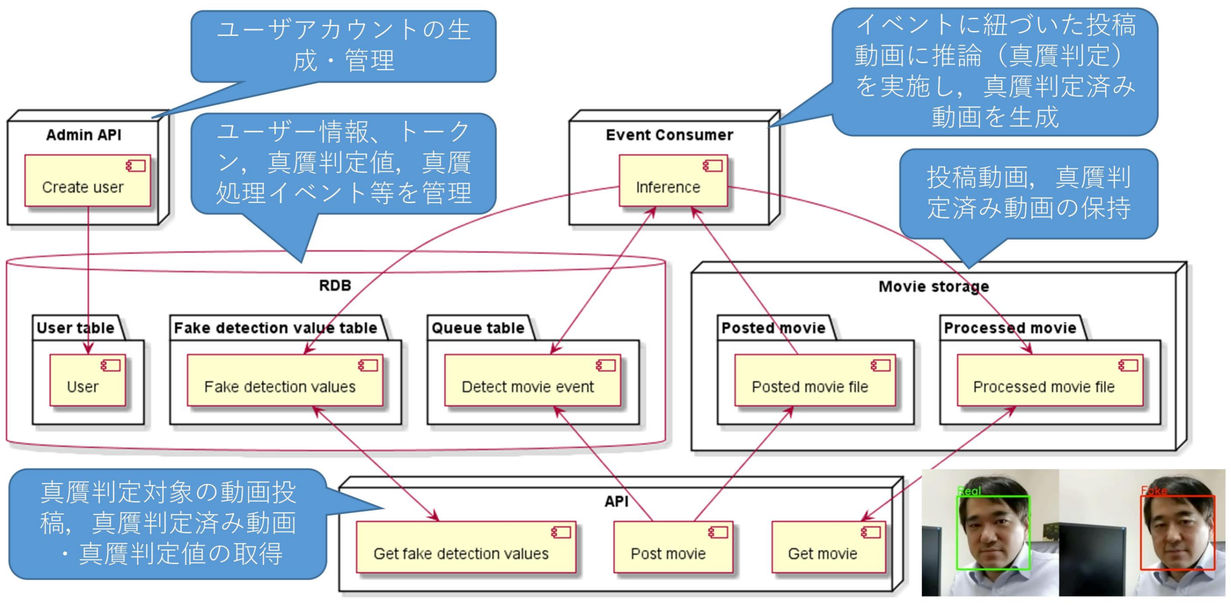
Methodology



Experimental Results



フェイク顔映像の真贋判定プログラム（SYNTHETIQ VISION）の実用化



Echizen Laboratory Multimedia Security Lab.

研究室連絡先:03-4212-2516 情報研16F 1603号室，教員連絡先:iechizen@nii.ac.jp