Antoine Montier | Cindy Hartmann

Attaques Adversariales

sur les systèmes de détection de Deepfakes

Encadré par Christophe Charrier et Emmanuel Giguet

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

GREYC
Laboratoire de recherche en sciences du numérique

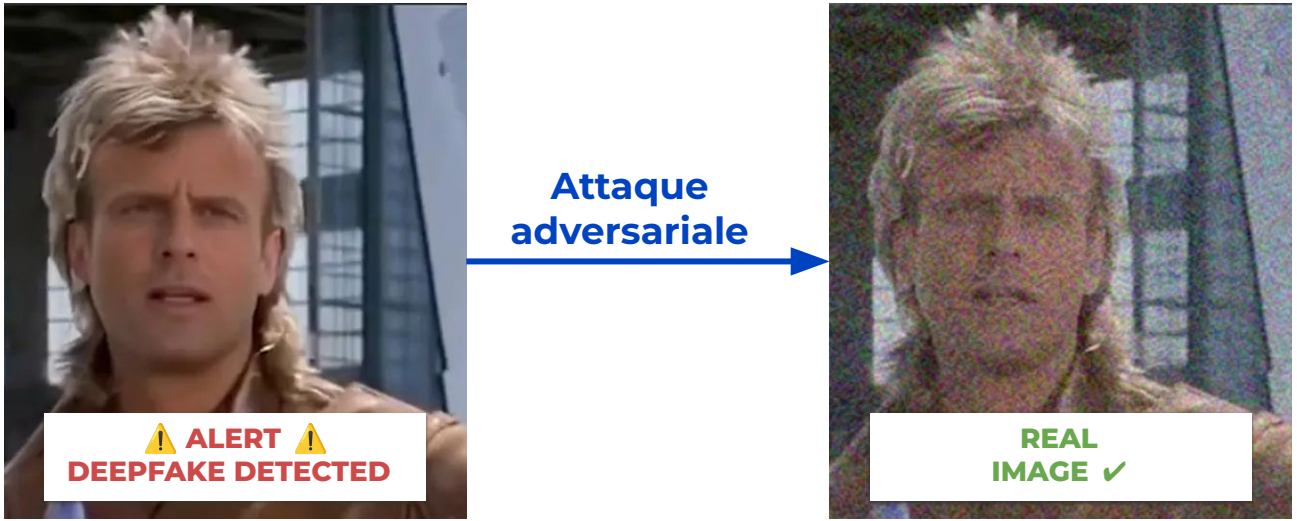
18/12/2025

Contexte

D’après une enquête de *Gartner* menée en 2024,

62% **des entreprises ont été victimes d’une attaque utilisant un deepfake .**

Systèmes de détection de deepfakes



Objectifs

Comprendre le fonctionnement des attaques adversariales

Appliquer ces attaques à un modèle de détection de deepfakes existant

Mesurer la dégradation des performances du modèle de détection choisi

Environnements de travail

Deepnote

Collaboration en temps réel
Peu de puissance
Mise en veille

Instance AWS

Calculs longs
Puissance moyenne
Stockage important

Google Colab

GPU disponible
Peu de stockage
Redémarrage fréquent du noyau

SOMMAIRE

1. Système de détection

- Jeu de données
- Modèle
- Traitement

Jeu de données :
FaceForensics++ (2019)

Extraction

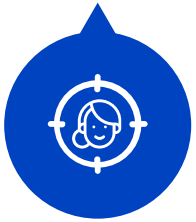
33 Go de vidéos
(réelles & fausses)
→ 2.1 To d'images

Migration

- Bucket AWS
- Organisation de l'architecture

Découpag

Mise en évidence
du visage



FaceForensics++

Exemple d'images réelles :




Exemple d'images truquées :


Choix du modèle de détection de deepfakes

ResNet-50

Classification au sens large
À affiner
GPU nécessaire

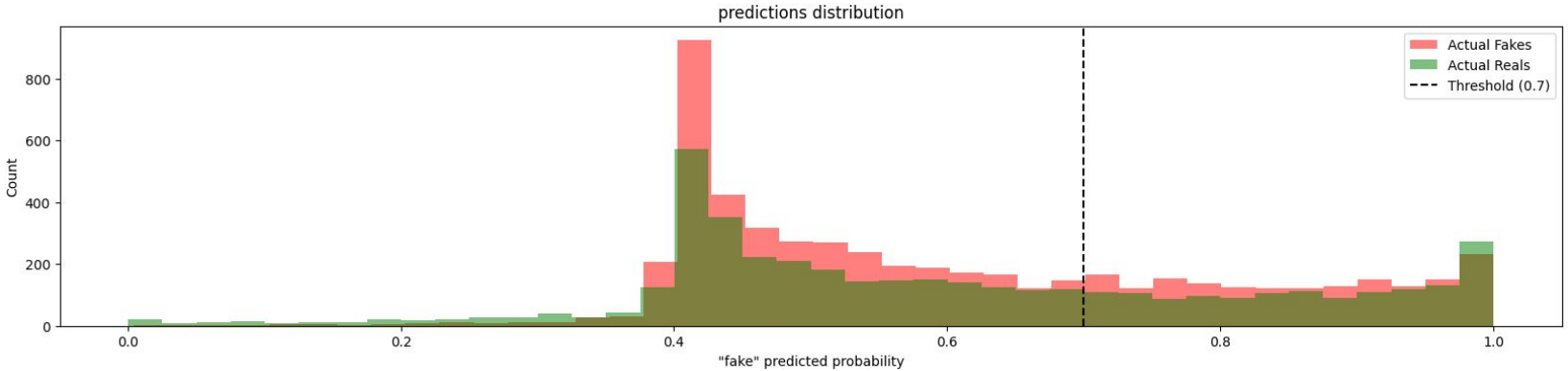
1er modèle DFDC

DeepFake Detection Challenge
Par *Selim Seferbekov*
obsolète (2019) & lourd

Xception

Affiné par *Fahim Faisal*
Kaggle Deepfake Challenge
léger

Evaluation

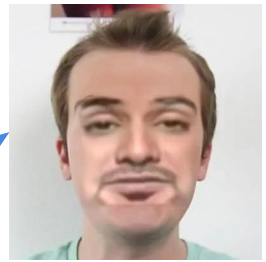


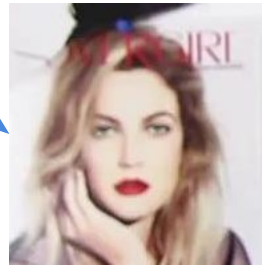
Métriques

Images (9876)	Précision	Recall
Réelle	0.44	0.67
Deepfake	0.55	0.32

Traitement

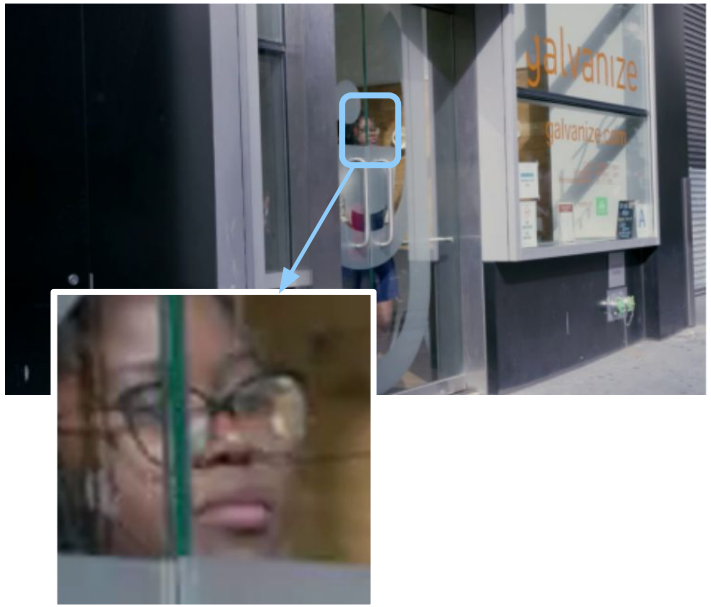
Outil d'extraction de visages de *freearhey*

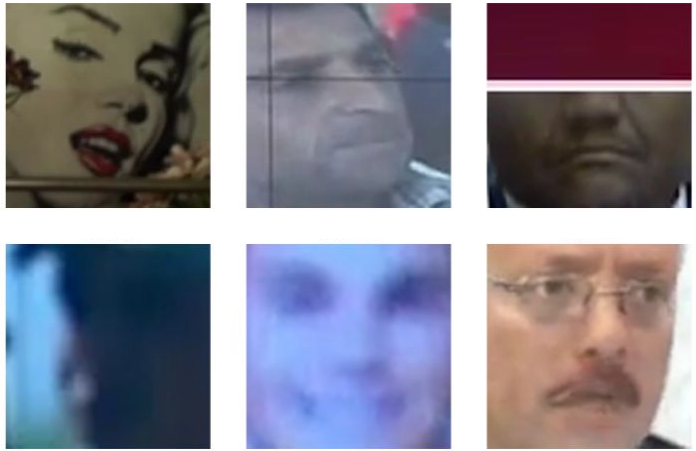




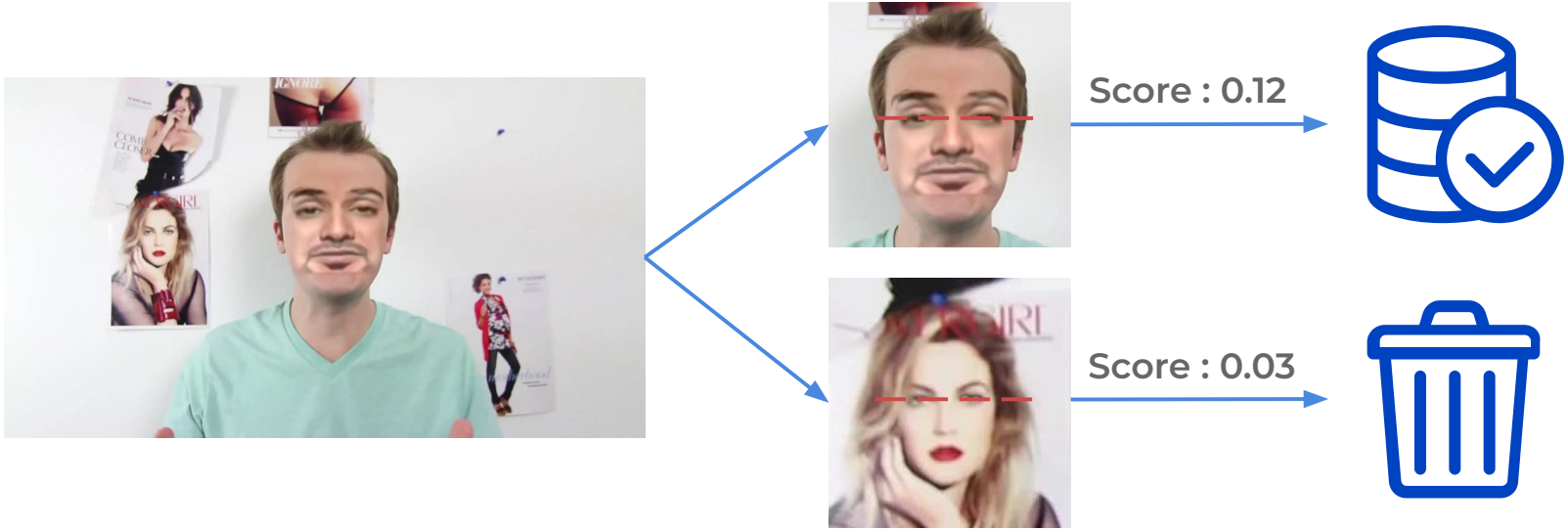
Traitement

Outil d'extraction de visages de *freearhey*





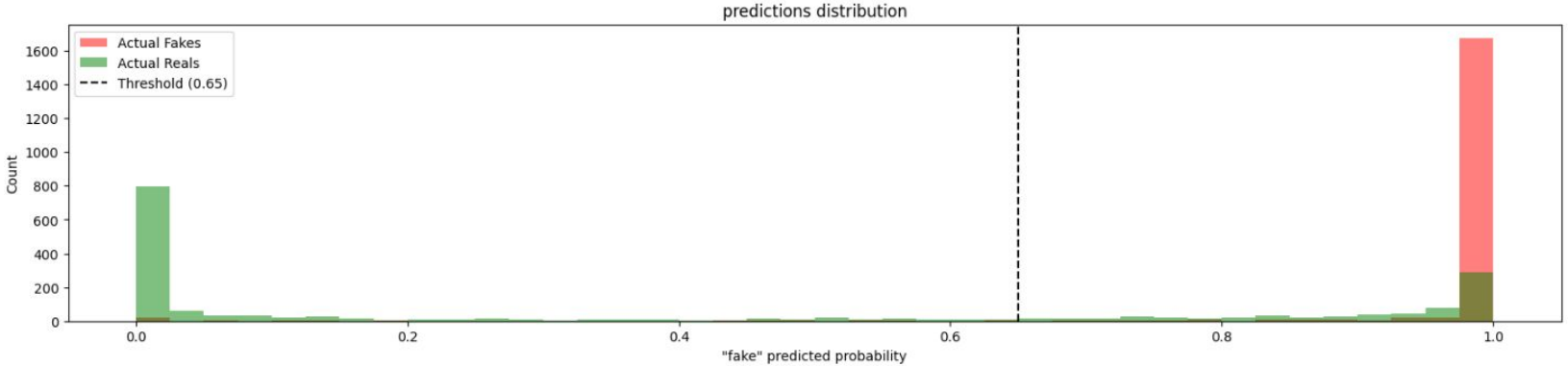
Traitement



1. Alignement selon les yeux
Eyeling (par *dullage*)

2. Score de qualité
Variance du Laplacien

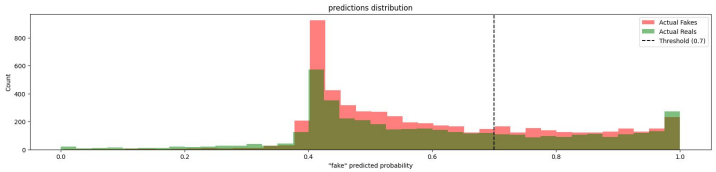
Evaluation



Métriques

Images (3644)	Précision	Recall
Réelle	0.96	0.64
Deepfake	0.73	0.97

Graphique de l'évaluation avant le changement de traitement :

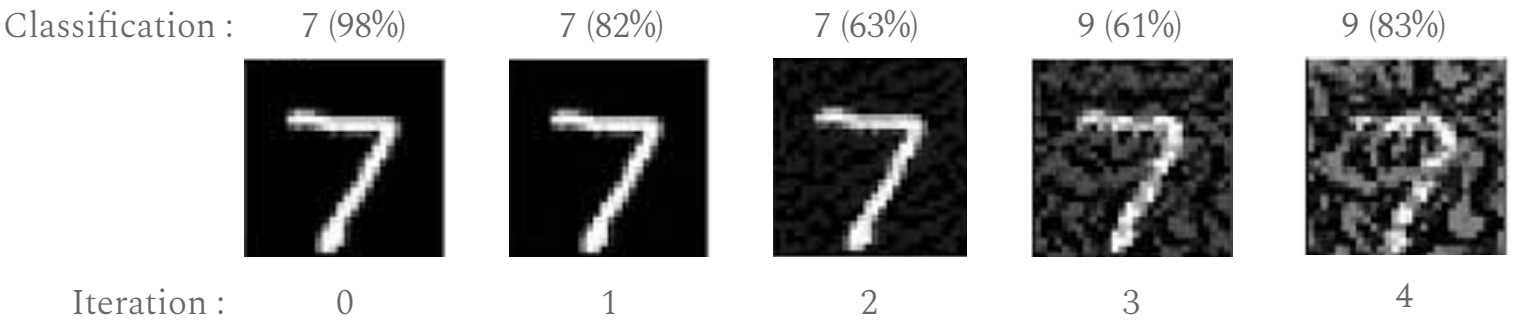


2. Attaques Adversariales

- PGD
- DeepFool

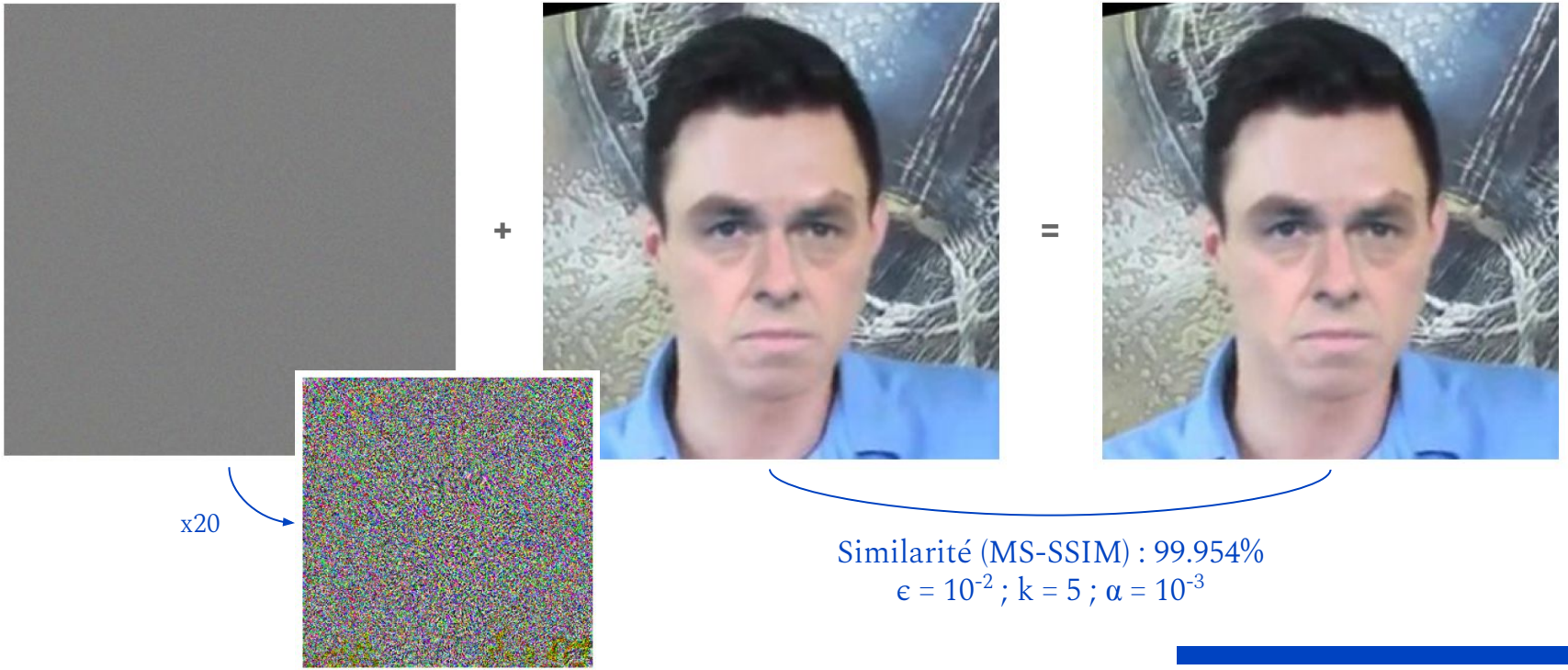
PGD - *Projected Gradient Descent*

- Fonction de coût : écart entre les prévisions du modèle et la classe réelle des données.

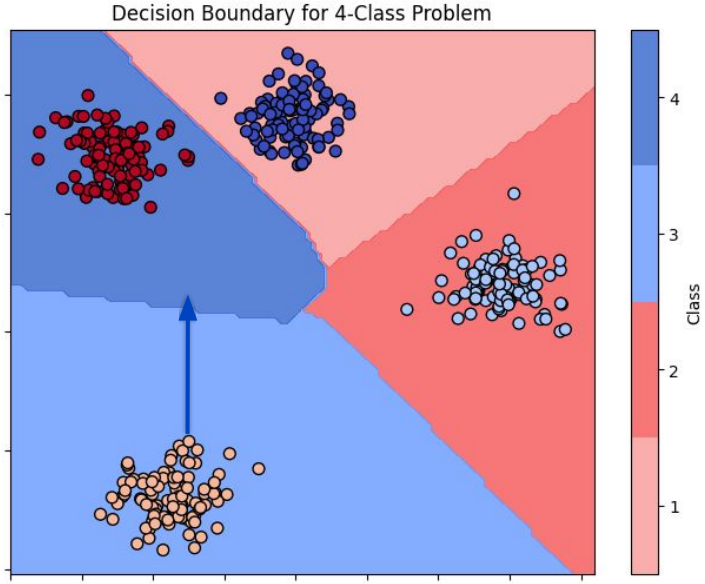


Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

PGD


Original
Réel: 0.01%
Faux: 99.99%
Perturbé
Réel: 100%
Faux: 0%
+
=
x20
Similarité (MS-SSIM) : 99.954%
$\epsilon = 10^{-2}$; k = 5 ; $\alpha = 10^{-3}$

DeepFool



medium.com/@Charles_Thiery/understanding-and-visualizing-decision-boundaries-in-a-deep-neural-network-50d7c82e6b5f

- Recherche du chemin vers la frontière de décision la plus proche
- Perturbation minimale pour changer la classe prédite

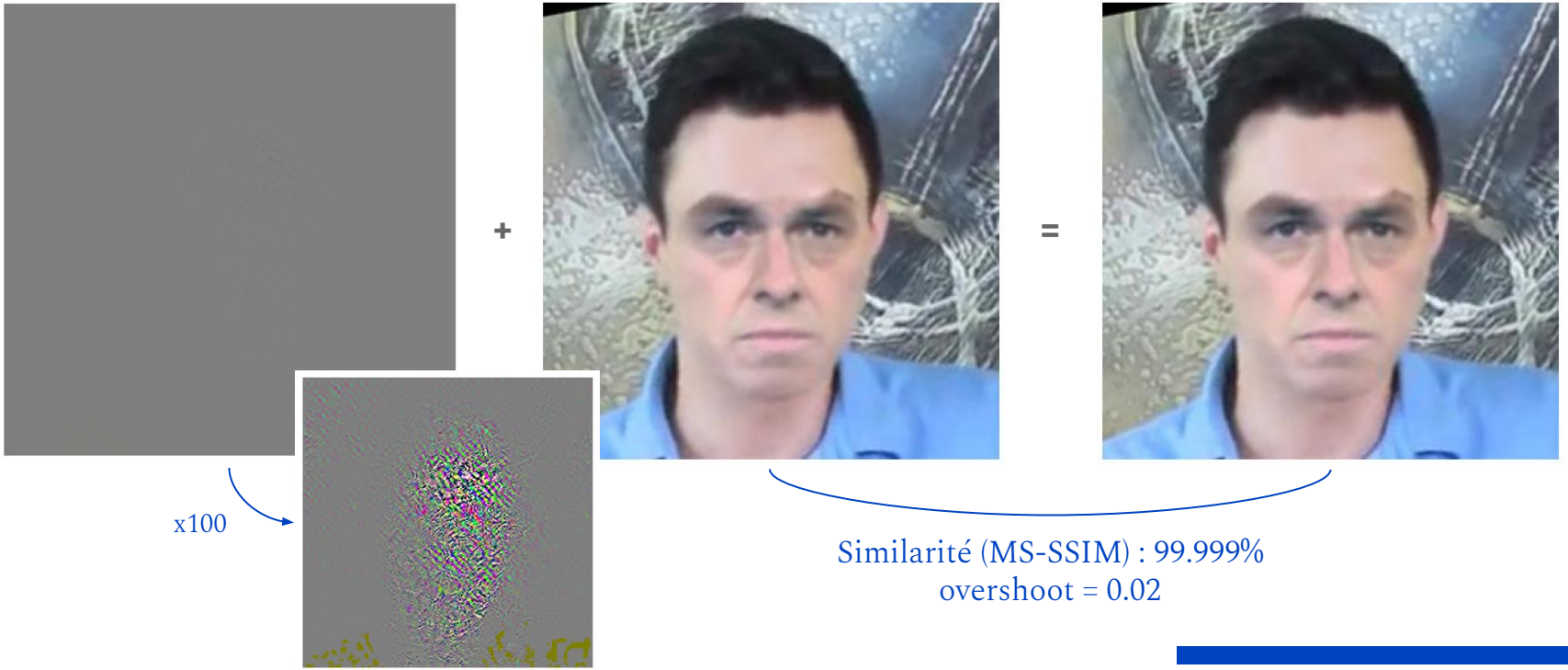
→ Moins de certitude sur la prédiction

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks.

DeepFool

Original
Réel: 0.01%
Faux: 99.99%

Perturbé
Réel: 50.13%
Faux: 49.87%

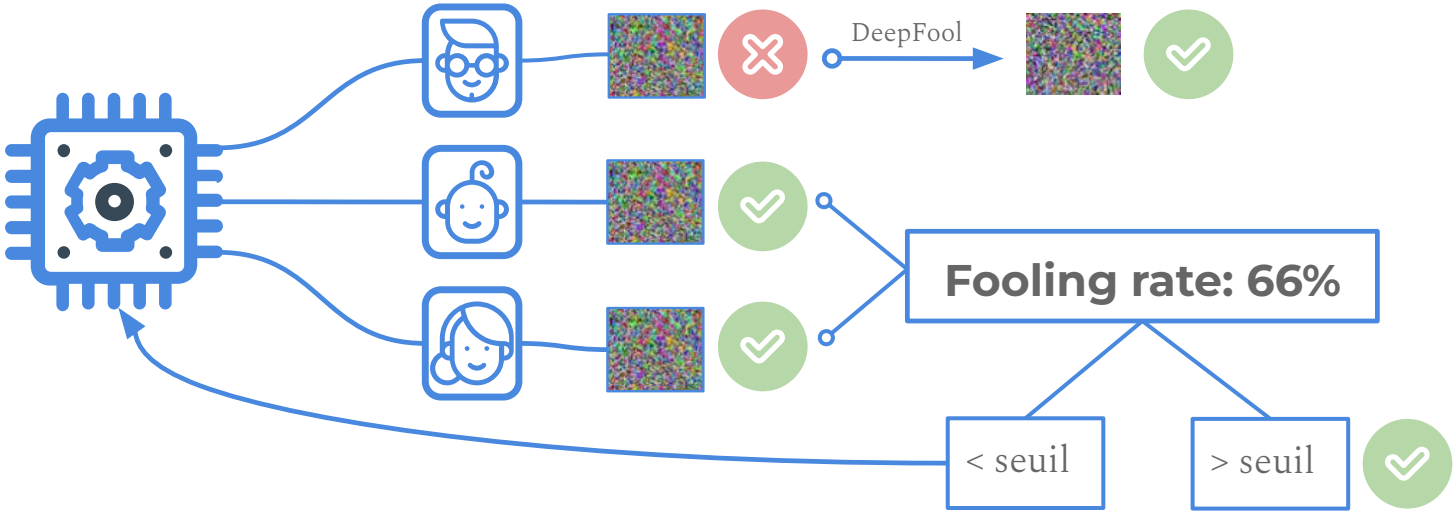


Similarité (MS-SSIM) : 99.999%
overshoot = 0.02

3.
Perturbations Universelles

- UAP-PGD
- UAP-DeepFool
- Transfert d’UAP

UAP - *Universal Adversarial Perturbations*

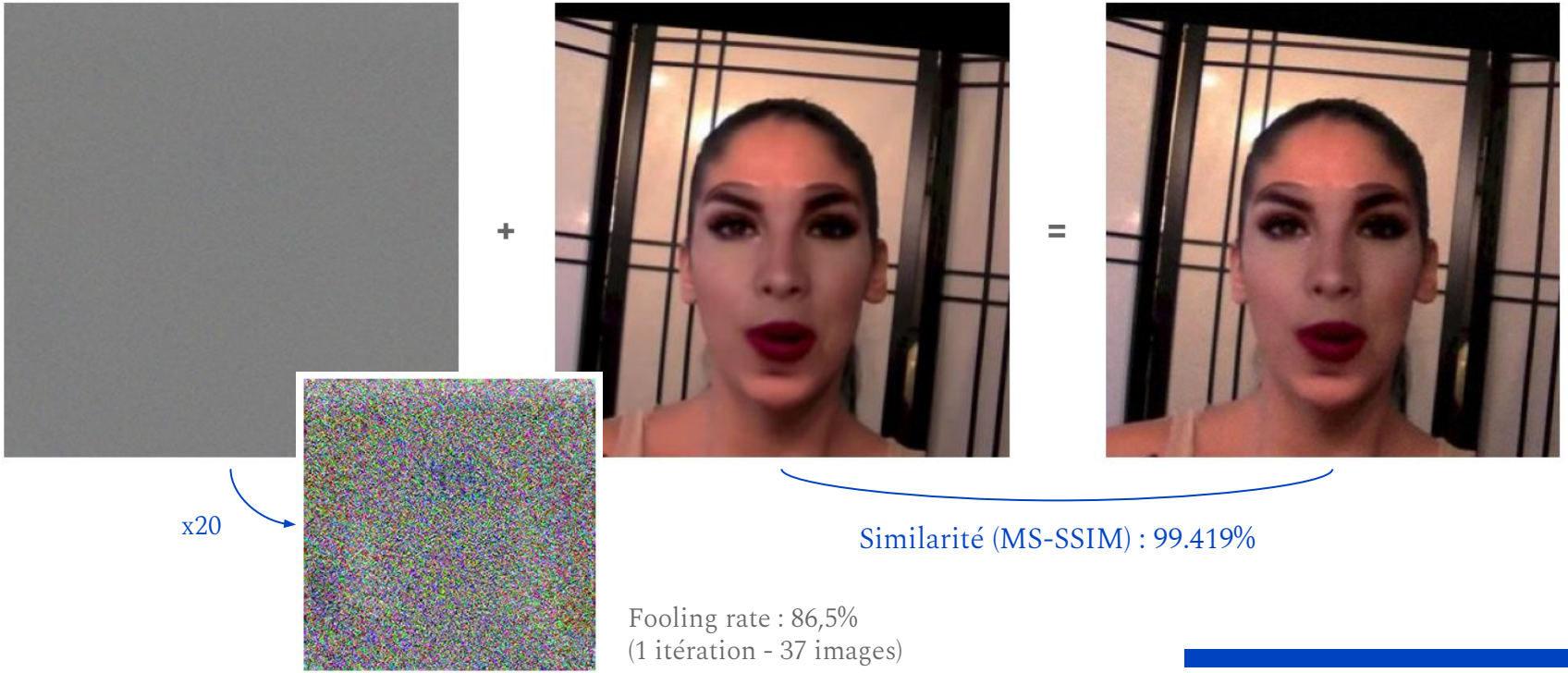


Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations.

PGD

Original
Réel: 0.02%
Faux: 99.98%

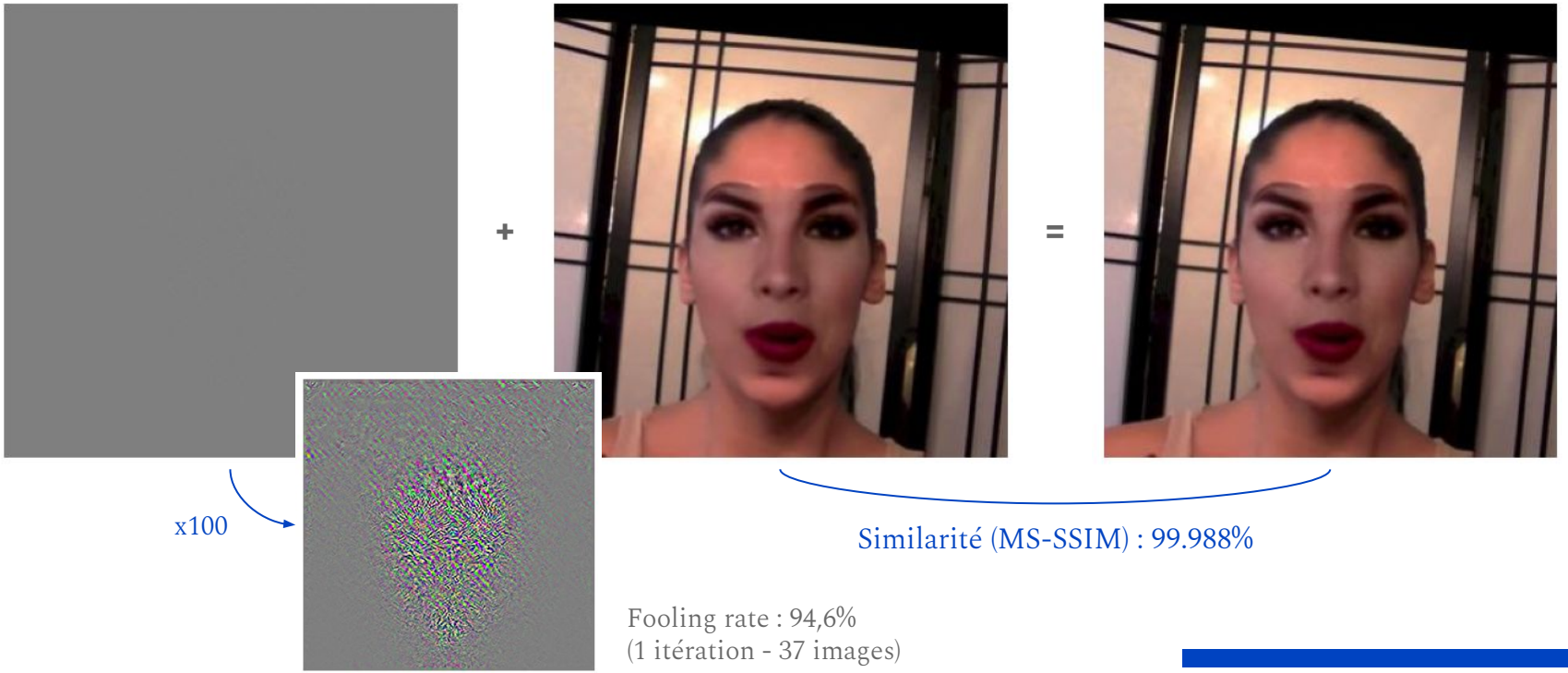
Perturbé
Réel: 100%
Faux: 0%



Similarité (MS-SSIM) : 99.419%

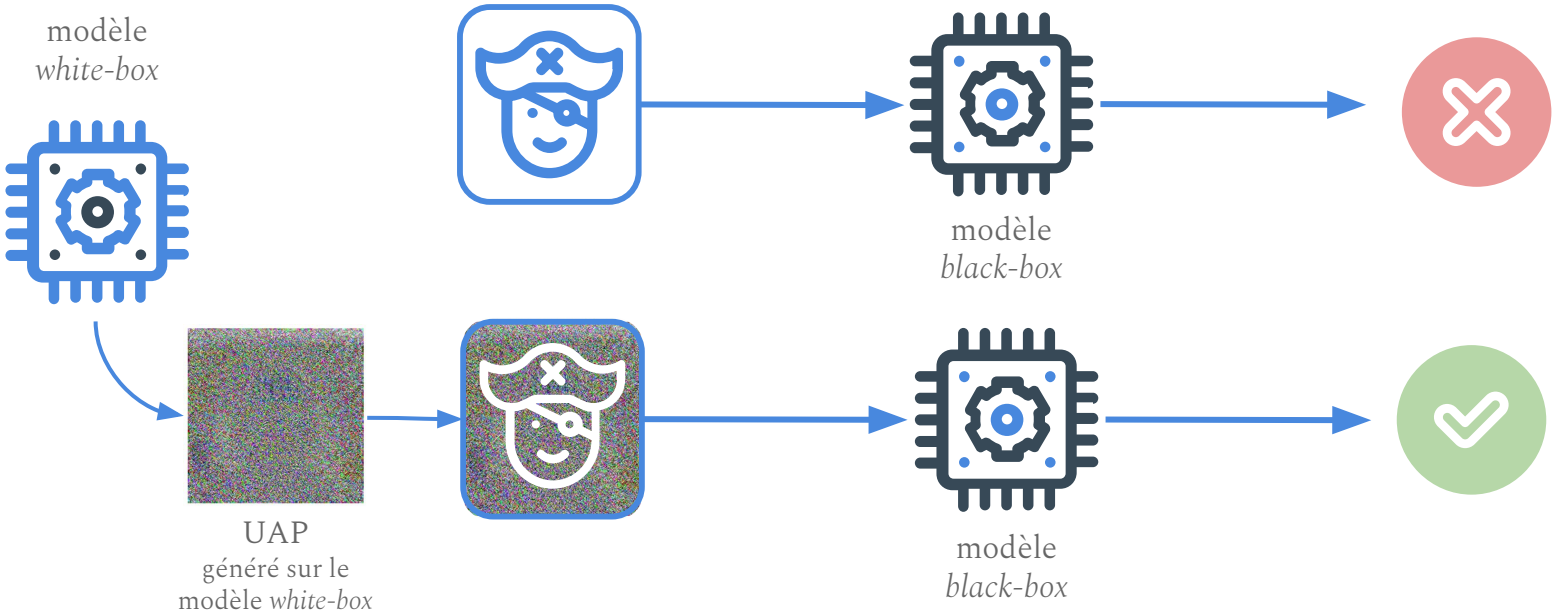
Fooling rate : 86,5%
(1 itération - 37 images)

DeepFool



Fooling rate : 94,6%
(1 itération - 37 images)

Transfert d’ UAP



Transfert d’ UAP

aifaceswap.io

Marina Kaye

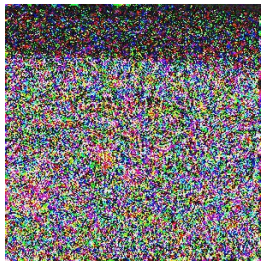
+

Sanae Takaichi

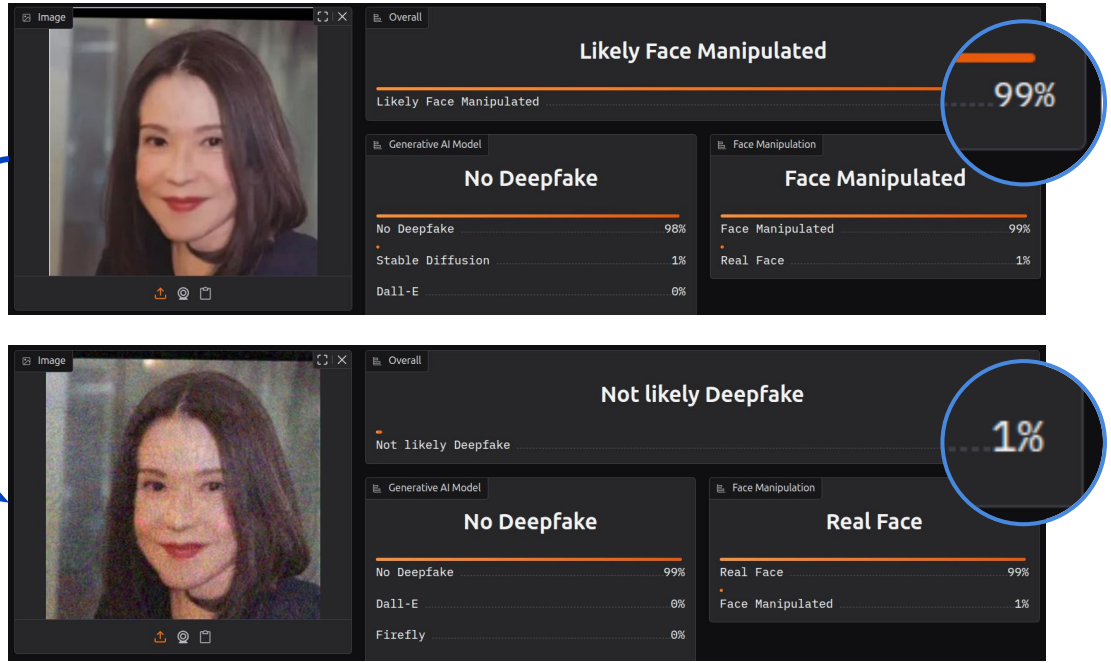
=

Transfert d’ UAP

faceonlive.com



UAP-PGD
(1632 images)



Conclusion

Perspectives d'amélioration

- jeu de données plus récent
- meilleur modèle de détection
- attaques sur les vidéos

→ Suite du projet : stratégies de défense

Antoine Montier | Cindy Hartmann

Merci !

Des questions ?

Bibliographie :

- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.
- Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks.
- Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations.

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

GREYC
Laboratoire de recherche en sciences du numérique