15ème
édition de la

Journée Française des Tests Logiciels

13 juin 2023

Beffroi de Montrouge

Conférence

Retour d'expérience: test logiciel de la robustesse d'un système industriel de détection d'objets basé IA

KEREVAL

LACROIX

ENSTA BRETAGNE

UNIVERSIDAD DE SEVILLA

CFTL
Comité Français des Tests Logiciels

Anne-Laure Wozniak

Contexte : Pourquoi tester l'IA ?

- Développement de modèles d'IA qui dépassent les performances humaines,
- Intégration de modèles aux systèmes critiques (e.g., véhicule autonome).

En 2018, un véhicule autonome Uber ne détecte pas correctement un piéton et son vélo, conduisant à un accident fatal.

> "The system design did not include a consideration for jaywalking pedestrians"
>
> *National Transportation Safety Board (NTSB)*

Processus de développement d'une IA

LES 7 GRANDES ÉTAPES DU **MACHINE LEARNING**

Test ou évaluation ?

L'évaluation du modèle couvre l'ensemble des métriques et graphiques qui résument la performance sur un jeu de données de validation ou de test.

Le test du modèle implique des vérifications explicites du comportement attendu de la part du modèle.

Les limites de l'évaluation des performances

L'évaluation des performances n'est pas suffisante

Elle…

- vérifie que le modèle **généralise bien** (pas de surajustement ou de sous-ajustement),
- assure que la **performance globale est satisfaisante**

mais…

- ne localise pas et ne caractérise pas les **erreurs**,
- ne traque pas les **régressions comportementales**,
- ne détecte pas les **biais**,
- etc.

C'est un détecteur de neige…

Quoi tester ?

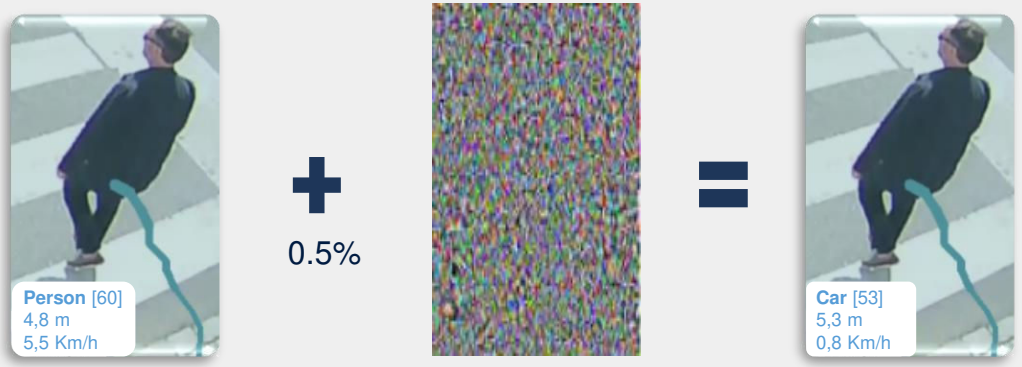
Quelques caractéristiques d'une IA de confiance

- Qualité des données
- Ethique et équité (*fairness*)
- Explicabilité
- Performance (*accuracy*)
- Cybersécurité
- Robustesse

Comment tester ?

Quelques problématiques actuelles…

- Le comportement n'est pas programmé explicitement
- Problème de l'oracle
- Absence de métriques de couverture
- Besoin de connaissances métier
- Problématiques spécifiques au domaine d'application : *adversarial examples*, éthique…



Expérimentation

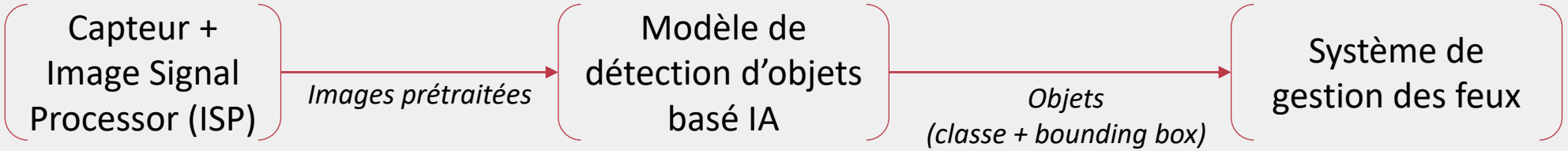
Le test de la robustesse d'un système industriel de détection d'objets basé IA

Références :

[1] A.-L. Wozniak, S. Segura, R. Mazo and S. Leroy, "Robustness Testing of a Machine Learning-based Road Object Detection System: An Industrial Case," *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, Pittsburgh, PA, USA, 2022, pp. 9-12, doi: 10.1145/3526073.3527592.

[2] Article en préparation : A.-L. Wozniak, R. N. Q. K. Duong, I. Benderitter, S. Leroy, S. Segura, R. Mazo.

Système industriel



- Système de **gestion de la circulation**, développé par Lacroix Impulse.
- Six classes détectées : **voitures**, **piétons**, **camions**, **bus**, **vélos** et **motos**.
- Le modèle de détection d'objets est une **boite noire** : pas d'information sur l'architecture et les paramètres.

Image issue du dataset propriétaire Django

Robustesse

"Ability of an AI system to **maintain its level of performance under any circumstances**."

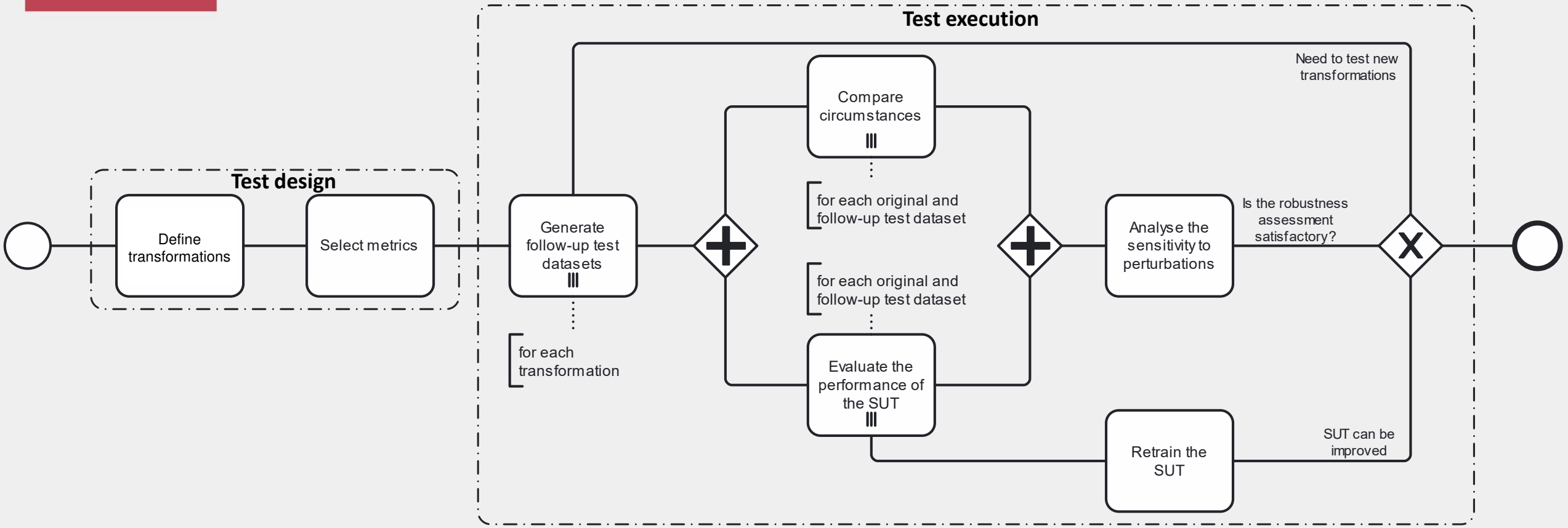
ISO/IEC TR 24029-1:2021

Notre objectif :

tester le système face à des **perturbations communes**, rencontrées dans son **fonctionnement normal**

- Changements dans l'environnement (e.g., météo, jour/nuit, pays)
- Attaques (intentionnelles ou non) par patches (e.g., le motif du t-shirt d'un piéton)
- **Changements dans l'équipement pour l'acquisition et le traitement des images** (e.g., paramètres de l'ISP, qualité de la camera…)

Processus de test



Transformations

Acquisition (caméra)

Paramètre	Catégorie
Matériau (lentille)	Flou
Distorsion de la lentille	Distorsion
Focus	Flou
Résolution	Flou
Ouverture	Couleur, Flou
Vitesse d'obturation	Flou
Sensibilité ISO / Gain	Bruit
Pixels morts	Pixel

Prétraitement (ISP)

Paramètre	Catégorie
Dématriçage	Couleur
Mappage tonal	Couleur
Netteté	Flou
Contraste	Couleur
Luminosité	Couleur
Réduction du bruit	Bruit
Balance des blancs	Couleur
Vignettage	Couleur
Configuration des couleurs	Couleur
Encodage/Compression	Flou
Niveau de noir	Bruit

Transformations

(a) Original

(b) Pixelate

(c) Zoom blur

(d) Gamma correction darker

(e) Speckle noise

(f) Chromatic Aberration

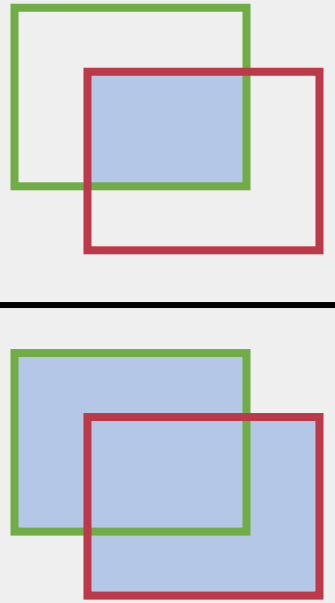
Métriques

Mesurer la performance : mean Average Precision

Deux composantes : classification et localisation.

- Vrai Positif : correctement classé et "assez bien" localisé.
- Faux Positif : incorrectement classé ou mal localisé.
- Faux Négatif : non détecté.

$$\text{Intersection over Union (IoU)} = \frac{\text{Aire de l'intersection}}{\text{Aire de l'union}} =$$



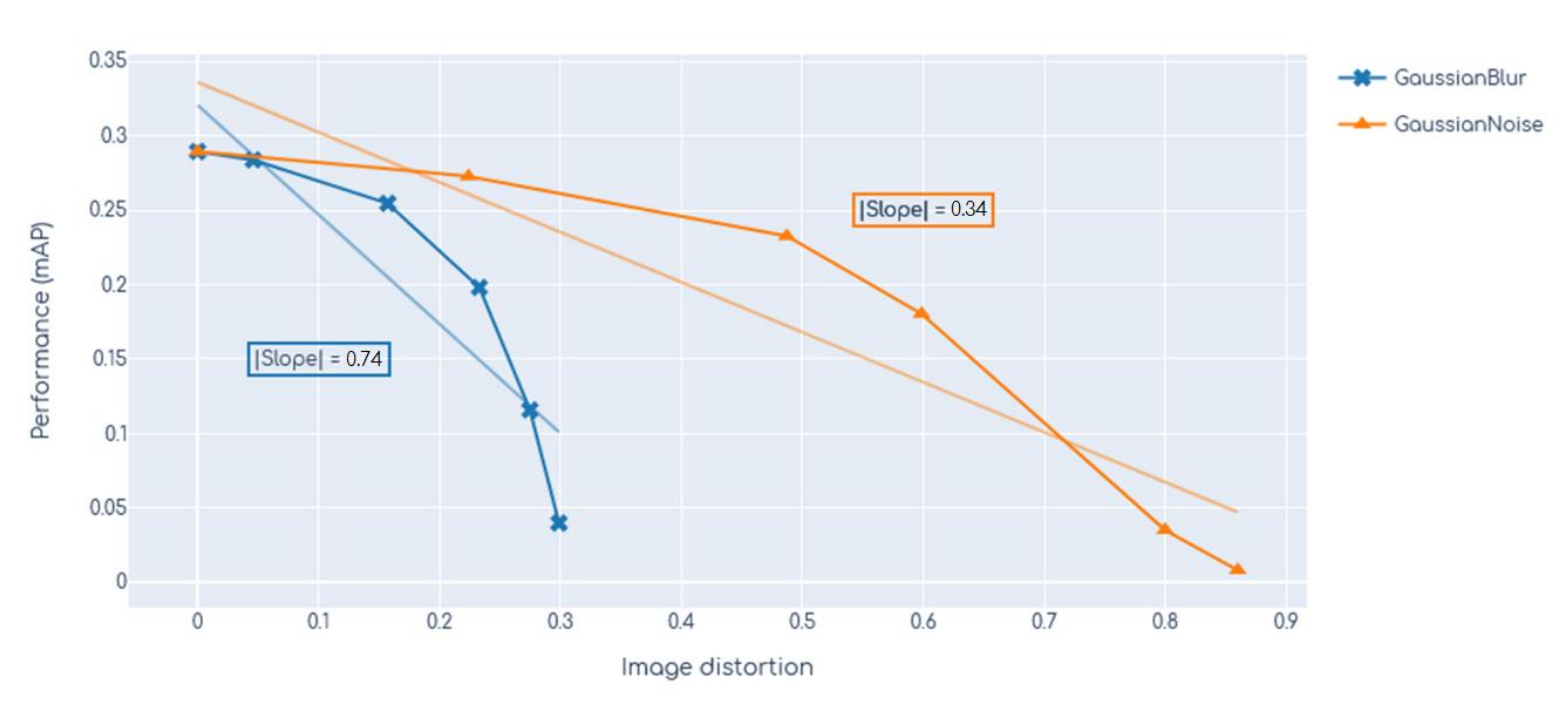
Métriques

Comparer les circonstances : Structural SIMilarity index (SSIM)

- Mesure la qualité d'image vis-à-vis d'une référence.
- Critère basé sur les changements de contraste, de luminance et de structure.
- Plus en adéquation avec la perception humaine que d'autres distances (e.g., MSE).

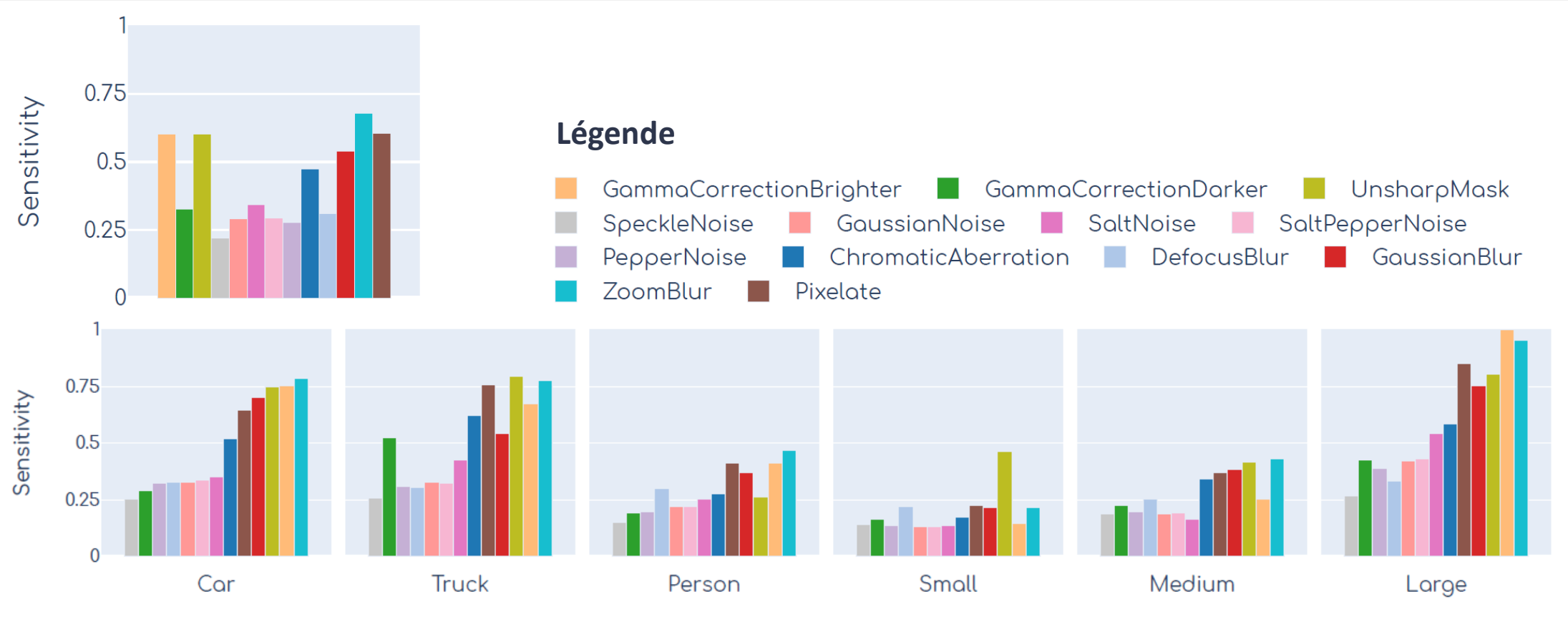
Métriques

Analyser la sensibilité / robustesse



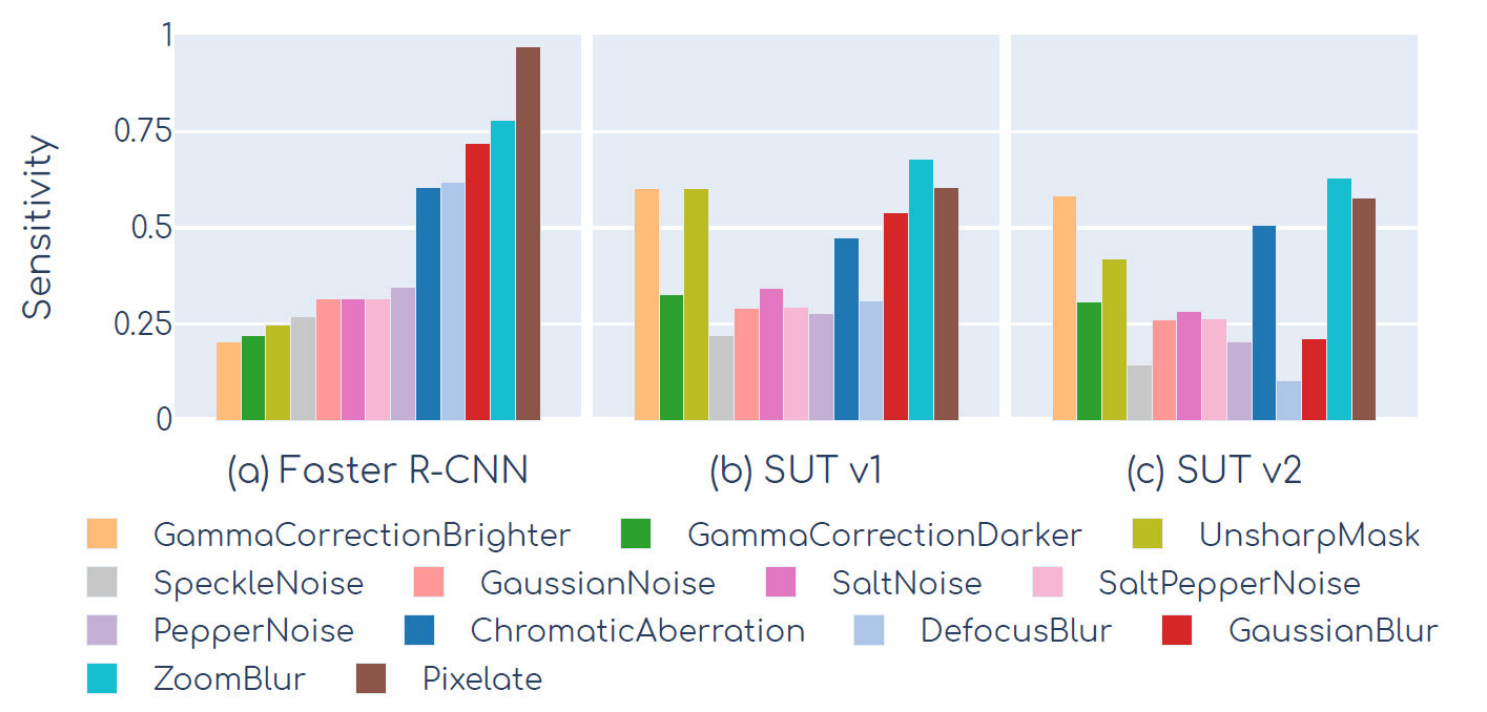
Résultats

Sensibilité du modèle aux transformations



Résultats

Comparaison à d'autres modèles



Faster R-CNN[1] : modèle de l'état de l'art, entrainé sur BDD100k

SUT v1 : modèle sous test initial, entrainé sur Django

SUT v2 : modèle sous test réentraîné sur Django mais incluant des images perturbées

[1] S. Ren, K. He, R. Girshick, and J. SUn, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in neural information processing systems*, 28, pp. 91-99, 2015.

15ème édition de la Journée Française des Tests Logiciels

Conférence

Retour d’expérience

Leçons tirées de l’expérimentation

CFTL
Comité Français des Tests Logiciels

1. Efficacité de la méthode pour le test

La méthode est efficace en pratique… mais des indications supplémentaires sont nécessaires aux étapes de la conception des tests et de l'analyse des résultats.

- Besoin d'approches systématiques pour le **choix des métriques** (notamment celles de comparaison des circonstances).
- Besoin d'indications supplémentaires sur la **manière d'analyser la sensibilité** du système aux perturbations, avec pour objectif de fournir un indicateur de robustesse global.

2. Connaissances métier

Des connaissances métiers sont absolument nécessaires.

- A l'étape de la **définition des transformations** : pour qu'elles soient représentatives de circonstances réelles.
  - ➔ **Les benchmarks de transformations issus de la littérature tendent à être trop génériques (transformations peu réalistes ou incomplètes en fonction des cas d'usage).**

- A l'étape de l'**analyse de la sensibilité** : pour interpréter correctement les résultats, comprendre l'impact sur le système, déterminer les causes des défaillances et définir des seuils pertinents afin de décider de la robustesse du système.
  - ➔ **Il y a actuellement un flou sur les seuils acceptables en termes de robustesse et sur la façon de les définir.**

3. Pertinence des métriques

Il y a besoin de définir de nouvelles métriques, au-delà des métriques conventionnelles pour l'évaluation de la performance.

- Les **métriques d'évaluation de la performance** les plus utilisées (e.g., mAP) ne suffisent pas à analyser la robustesse d'un système.
  Exemple : le cas du nombre de faux négatifs.
- Comment définir un **score de robustesse** ? Comment statuer sur les tests (définition de seuils acceptables) ?

Conclusion

Une méthodologie est nécessaire à mi-chemin entre l'agnostique et le spécifique.

- Contraintes et particularités propres à chaque système.
- Besoin d'adapter la méthode de test (perturbations ou métriques utilisées), sans tomber dans la sur-spécification.

➔ Il est donc nécessaire de guider les praticiens sur la méthode et les paramètres à utiliser, sans fermer la porte aux contraintes des cas d'utilisation spécifiques.

Conférence

15ème
édition de la

Journée Française des Tests Logiciels

13 juin 2023

Beffroi de
Montrouge

Merci de votre écoute !

CFTL

Comité Français
des Tests Logiciels