Projet Image M2 : Evaluation de la sécurité visuelle d'images obscures par CNN
HAI918I : Image, sécurité et deep learning

Université de Montpellier - FDS
2ième année Master IMAGINE
Oren AMSALHEM - Thomas CARO

24 Novembre 2024

1 Introduction

Cette semaine, nous avons intégrer une nouvelle méthode d'obscuration appelée FGSM(Fast Gradient Sign Method), et nous avons mis en oeuvre une application simple permettant de charger une image, de l'obscurcir et de faire une prédiction en choisissant le modèle de prédiction parmi ceux que l'on a mis en place.

2 FGSM

FGSM est une méthode adverseriale qui a pour but de créer des images modifiés spécifiquement pour qu'un CNN ne puisse pas reconnaitre l'image. Elle est donc spécifique au model car elle utilise le gradient de l'erreur issue de la prédiction d'image. Dans un contexte réel ou l'on souhaite juste obscurcir une image ou d'une application simple, cela n'est pas véritablement faisable. Voici le principe :

Soit $\mathbf{x}$ l'entrée (par exemple une image), $\mathbf{y}$ la sortie correcte (l'étiquette), et $J(\theta, \mathbf{x}, \mathbf{y})$ la fonction de perte du modèle. L'exemple adversarial $\mathbf{x}'$ est généré comme suit :

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, \mathbf{y})\right)$$

où :

- $\mathbf{x}'$ est l'image modifiée (l'exemple adversarial),
- ϵ est un petit facteur d'échelle qui contrôle la taille de la perturbation,
- $\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, \mathbf{y})$ est le gradient de la fonction de perte par rapport à l'entrée $\mathbf{x}$,
- $\text{sign}(\cdot)$ désigne l'opération qui prend le signe du gradient, appliquant ainsi la perturbation dans la direction de la plus grande pente.

Nous avons codé une méthode FGSM qui prend un des modèles que l'on a crée et permet à partir d'un dossier d'image d'effectuer la perturbation sur l'image. Nous l'avons tester sur des images issues d'une distorsion de niveau 1 et sur du flou de mouvement de niveau 1. Les résultats sont du type suivant :

FIGURE 1 – FGSM avec epsilon faible à partir d'une distorsion de niveau 1

FIGURE 2 – FGSM avec epsilon élevé à partir d'une distorsion de niveau 1

FIGURE 3 – FGSM avec epsilon faible à partir d'un flou de mouvement de niveau 1

FIGURE 4 – FGSM avec epsilon élevé à partir d'un flou de mouvement de niveau 1

On observe déja que les perturbations ne sont pas faites aux mêmes endroits. Pour le flou la perturbation est faite sur l'intégralité de l'image alors que pour la distorsion la perturbation est faite dans les zones caractéristiques de l'image tels que le contour du visage, les cheveux, les yeux et le nez.

Les détails sont gardés pour la distorsion, alors que pour une image ou les caractéristiques sont déja moins distinctes à cause d'un flou, les perturbations sont faites sur l'ensemble de l'image.

Pour tester l'efficacité de cette perturbation sur le CNN nous avons fait des tests avec l'ensemble obscurci de base et celui perturbé en plus. Voici les résultats avec le flou de mouvement :

```
Matrice de confusion:
[[   0    0]
 [  17 4273]]

Rapport de classification:
              precision    recall  f1-score   support

    Classe 0       0.00      0.00      0.00         0
    Classe 1       1.00      1.00      1.00      4290

    accuracy                           1.00      4290
   macro avg       0.50      0.50      0.50      4290
weighted avg       1.00      1.00      1.00      4290
```

FIGURE 5 – Ensemble obscurci de base

```
Matrice de confusion:
[[   0    0]
 [4290    0]]

Rapport de classification:
              precision    recall  f1-score   support

    Classe 0       0.00      0.00      0.00       0.0
    Classe 1       0.00      0.00      0.00    4290.0

    accuracy                           0.00    4290.0
   macro avg       0.00      0.00      0.00    4290.0
weighted avg       0.00      0.00      0.00    4290.0
```

FIGURE 6 – Ensemble obscurci et perturbé

La FGSM du flou de mouvement fonctionne très bien comme on peut le voir. Cependant la FGSM pour la distorsion est mauvaise :

```
Matrice de confusion:
[[   0    0]
 [ 438 3827]]

Rapport de classification:
              precision    recall  f1-score   support

    Classe 0       0.00      0.00      0.00         0
    Classe 1       1.00      0.90      0.95      4265

    accuracy                           0.90      4265
   macro avg       0.50      0.45      0.47      4265
weighted avg       1.00      0.90      0.95      4265
```

FIGURE 7 – Ensemble obscurci de base

```
Rapport de classification:
              precision    recall  f1-score   support

    Classe 0       0.00      0.00      0.00         0
    Classe 1       1.00      1.00      1.00      4265

   micro avg       1.00      1.00      1.00      4265
   macro avg       0.50      0.50      0.50      4265
weighted avg       1.00      1.00      1.00      4265
```

FIGURE 8 – Ensemble obscurci et perturbé

La FGSM aide même le modèle a trouvé que les images étaient obscurcies dans ce cas. Cela nous semble bizarre que le programme qui génère les images perturbées ne fonctionne pas avec la distorsion mais bien avec le flou de mouvement. On dirait même que les perturbations sont effectuées dans le mauvais sens par rapport au gradient de la perte sachant que le CNN reconnait encore plus facilement les images issues de distorsions perturbées.

3 Application

Nous avons réalisé une application dans le cadre de nos travaux pratiques pour mettre en œuvre les concepts étudiés au cours des dernières semaines.

Cette application permet d'ouvrir une image et de charger un modèle préentraîné afin de le tester avec l'une des quatre méthodes d'obscuration de départ que nous avons étudiées :

- **Pixelisation**
- **Flou Gaussien**
- **Flou de Mouvement**
- **Distorsion**

Une fois une méthode appliquée, le modèle prédit avec précision (en pourcentage) quelle méthode a été utilisée. Cela nous permet de mesurer la robustesse et la performance du modèle face aux modifications apportées aux images.

4 Présentation de l'Application Python

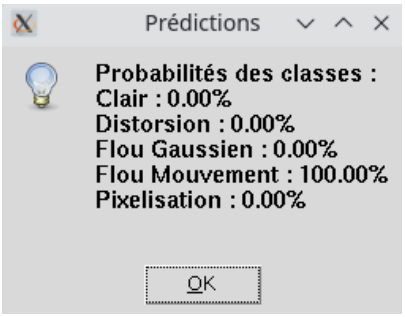
L'application que nous avons développée est interactive et offre une interface utilisateur simple et intuitive. Voici les étapes principales de son fonctionnement :

1. **Chargement de l'image et du modèle :** L'utilisateur peut sélectionner une image et un modèle depuis son système pour effectuer les tests.

2. **Choix de la méthode d'obscuration :** Une des quatre méthodes d'obscuration peut être choisie via l'interface.

3. **Application du filtre :** L'image est modifiée en fonction du filtre choisi et selon la zone de l'image choisi.

4. **Prédiction :** Le modèle préentraîné analyse l'image modifiée et affiche sa prédiction sur la méthode appliquée.



5. **Évaluation :** La qualité de la modification est mesurée par des indicateurs tels que le **PSNR** (Peak Signal-to-Noise Ratio).