SCENE PRIVACY PROTECTION

CIS centre for intelligent sensing

Chau Yi Li*, Ali Shahin Shamsabadi*, Ricardo Sanchez-Matilla*, Riccardo Mazzon, Andrea Cavallaro

Queen Mary University of London

1. Introduction

Objective

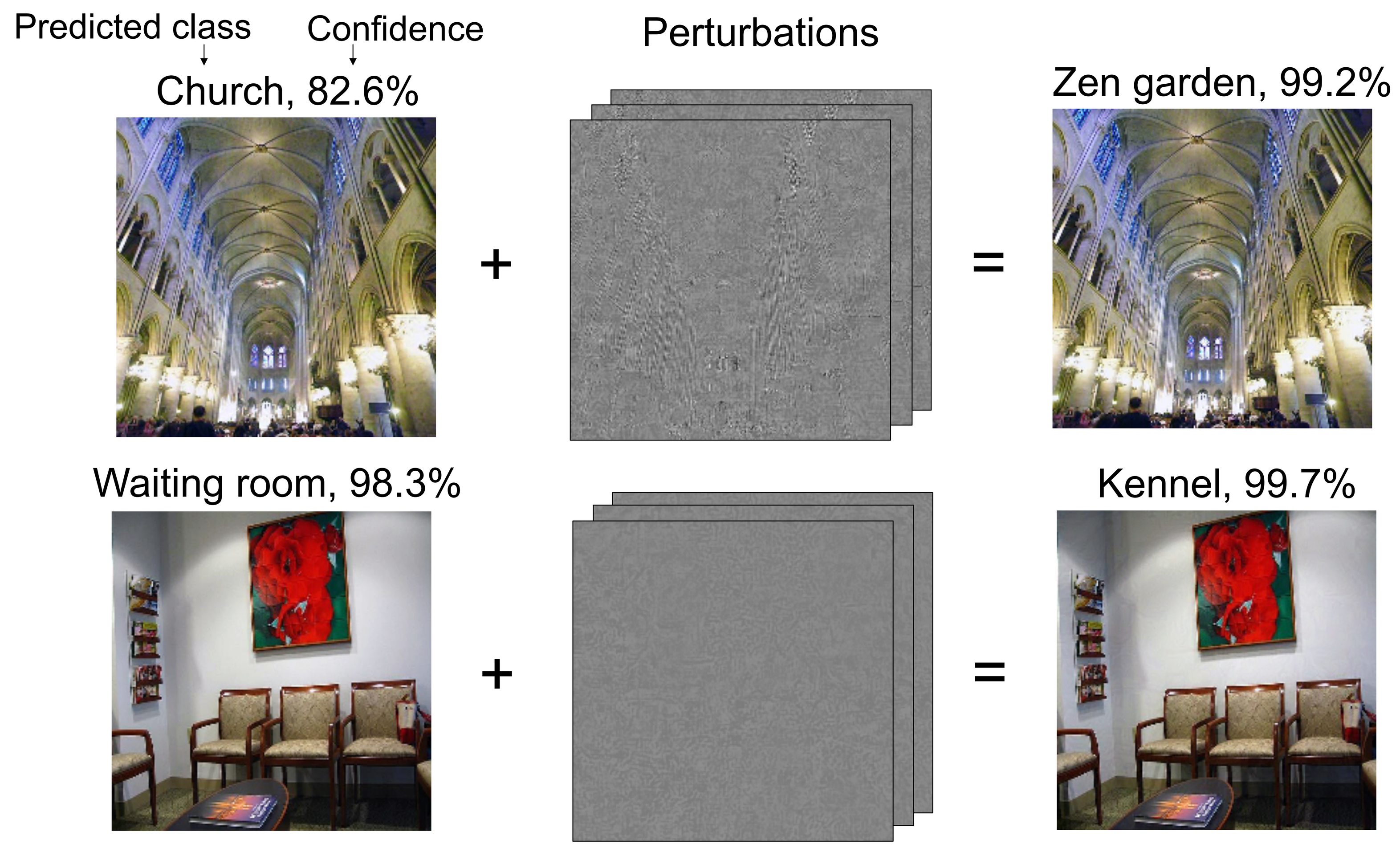
To design a transformation to protect private information in images against automatic inference prior to uploading to an online social media (privacy protection)

Motivation

Automatic inference of private information by online service providers for user profiling breaches privacy, e.g. scene

Properties

- unnoticeability
  distortion not perceived by humans
- irreversibility
  not possible to retrieve private information by automated method



2. Related work

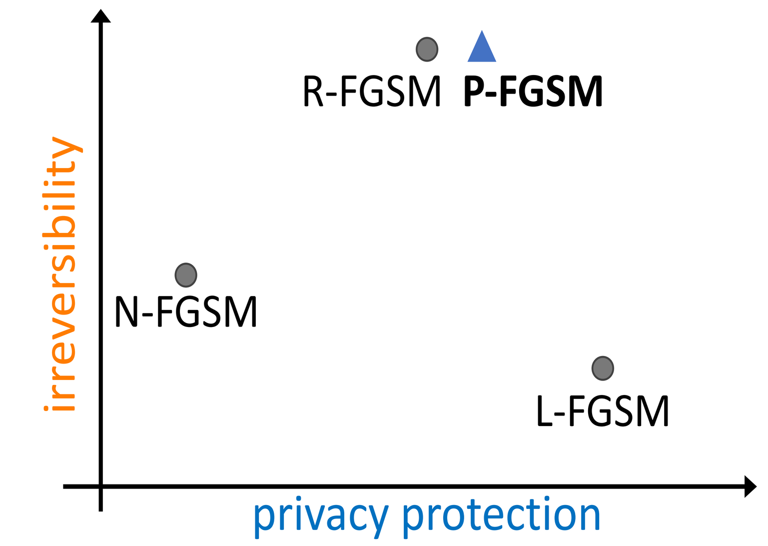
Traditional methods

distort the appearance of image regions containing private information
low unnoticeability with reduced image quality
e.g. redaction, cartooning, pixelation, single or multiple blurs, false colours, scrambling and warping

Adversarial methods

add small perturbations which mislead specific neural networks, used as classifiers
high unnoticeability
e.g. Fast Gradient Sign Method (FGSM) variants

- Non-targeted (N-FGSM) [1]
- Random (R-FGSM) [2]
- Least-likely (L-FGSM) [1]



3. Private Fast Gradient Sign Method (P-FGSM)

$$\underbrace{\dot{x}}_{\text{adversarial image}} = \underbrace{x}_{\text{original image}} + \underbrace{\delta_x^*}_{\text{adversarial perturbation}}$$

privacy protection $\quad M(\dot{x}) \neq M(x)$

unnoticeability $\quad \|\dot{x} - x\| \to 0$

irreversibility
the true class or $M(x)$ cannot be deduced from $M(\dot{x})$

Considers

- prediction probability $\boldsymbol{p} = (p_1, \dots, p_i, \dots, p_D)$ by M of each class $(y_1, \dots, y_i, \dots, y_D)$
- sort $\boldsymbol{p}$ in descending order as $\boldsymbol{p}' = (p'_1, \dots, p'_i, \dots, p'_D)$

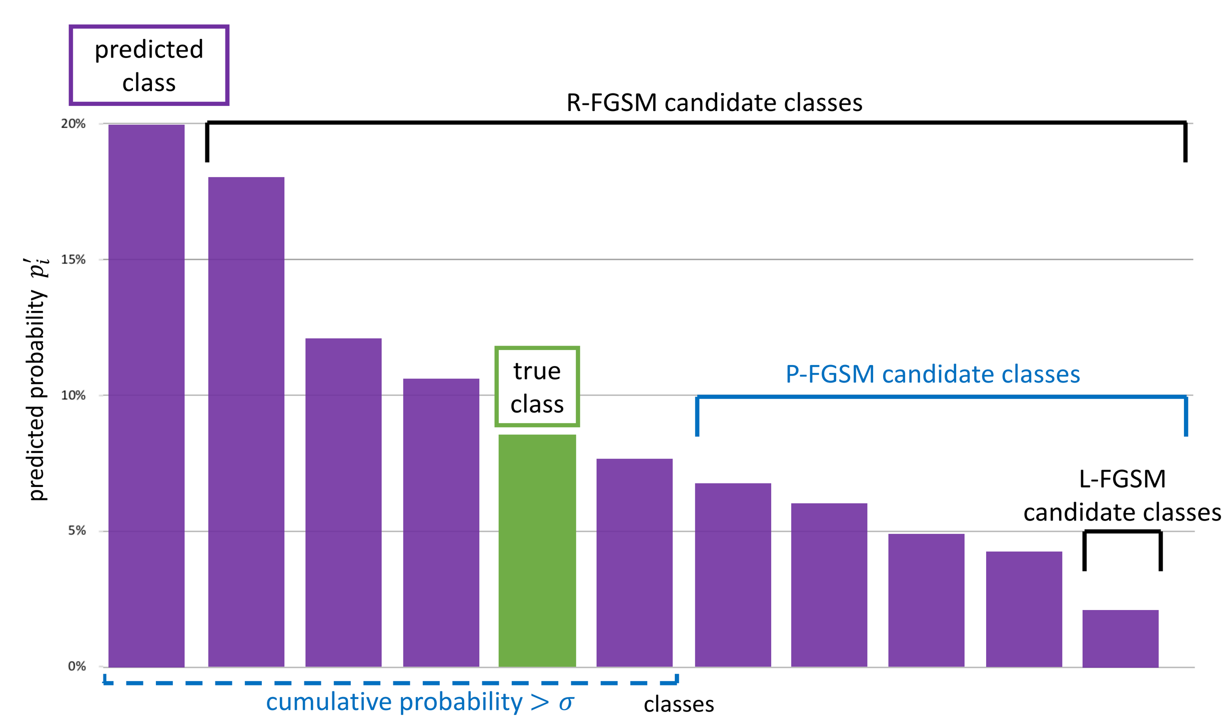


Illustration: when M is incorrect, predicted class ≠ true class

P-FGSM: iterative adversarial perturbation generation

$$\dot{x}_0 = x$$

$$\dot{x}_N = \dot{x}_{N-1} - \underbrace{\varepsilon}_{\text{magnitude of perturbation}} \text{sign}(\underbrace{\nabla_x}_{\text{gradient with respect to } x} \underbrace{J_M}_{\text{cost function of } M} (\underbrace{\theta}_{\text{parameters of } M}, \dot{x}_{N-1}, \underbrace{\tilde{y}}_{\text{target class}}))$$

Proposed target class $\tilde{y}$ selection

from classes with cumulative probability > threshold σ
avoid targeting true class even when M is incorrect

$$\tilde{y} = \underbrace{R}_{\text{random selection function}}\left(\underbrace{\left\{y_{j+1} : \sum_{i=1}^{j} p'_i > \sigma\right\}}_{\text{set of target candidate classes}}\right)$$

4. Experiments

Dataset: Mediaeval 2018 Pixel Privacy Challenge [3]

- a subset of Places365-Standard dataset [4]
- training/testing set: 3000/3000 images
- images from 60 private classes, defined in [3]

Classifier: ResNet50 365-class classifier
Preprocessing: resize to 224×224 pixels with bilinear interpolation
Parameters: σ = 0.99; ε = 0.007

Method	privacy protection: Accuracy (%) ↓		unnoticeability: PSNR		unnoticeability: BRISQUE [5]		irreversibility: Euclidean distance^ ↓
	Top-1	Top-5	avg. ↑	std. dev. ↓	avg. ↓	std. dev. ↓	
Original	56.40	86.47	-	-	26.72	8.66	-
N-FGSM	8.83	23.00	**40.62**	4.75	24.16	8.31	0.23
R-FGSM	0.17	7.00	40.24	2.87	23.99	8.29	**0.14**
L-FGSM	**0.00**	**0.17**	38.08	2.30	**23.67**	8.36	0.28
P-FGSM	**0.00**	5.60	39.99	2.72	23.85	8.28	**0.14**

^ between discrete uniform distribution and average discrete distribution of target class
↓: the smaller the better; ↑: the larger the better

5. Conclusions

P-FGSM: protects privacy against automatic inference

- by generating corresponding adversarial images
- misleads ResNet50 (always in its top-1 and 94.40% of the times in its top-5)
- higher degree of irreversibility compared to N-FGSM and L-FGSM
- comparable visual quality with other FGSMs

Code
bit.ly/pfgsm_github

Paper
bit.ly/pfgsm2019

References

[1] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in ICLR Workshops 2017
[2] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial learning at scale," in ICLR Workshops 2017
[3] M. Larson, Z. Liu, S.F.B. Brugman, and Z. Zhao, "Pixel Privacy: Increasing Image Appeal while Blocking Automatic Inference of Sensitive Scene Information" in MediaEval Workshop 2018
[4] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," in IEEE PAMI 2018
[5] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," in IEEE TIP 2012

* Equal contribution