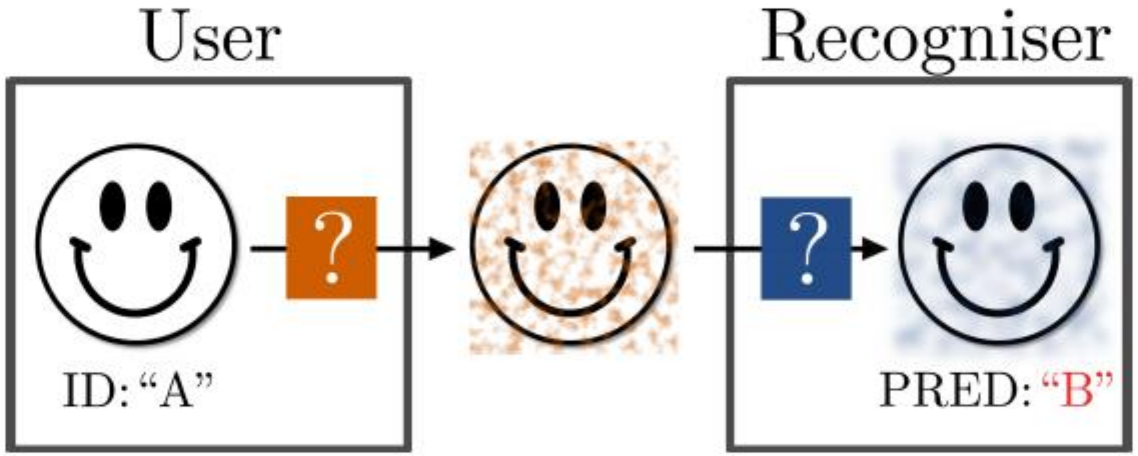
Adversarial Image Perturbation for Privacy Protection
A Game Theory Perspective

Seong Joon Oh, Mario Fritz, Bernt Schiele

Presenter: Yuxuan Mu

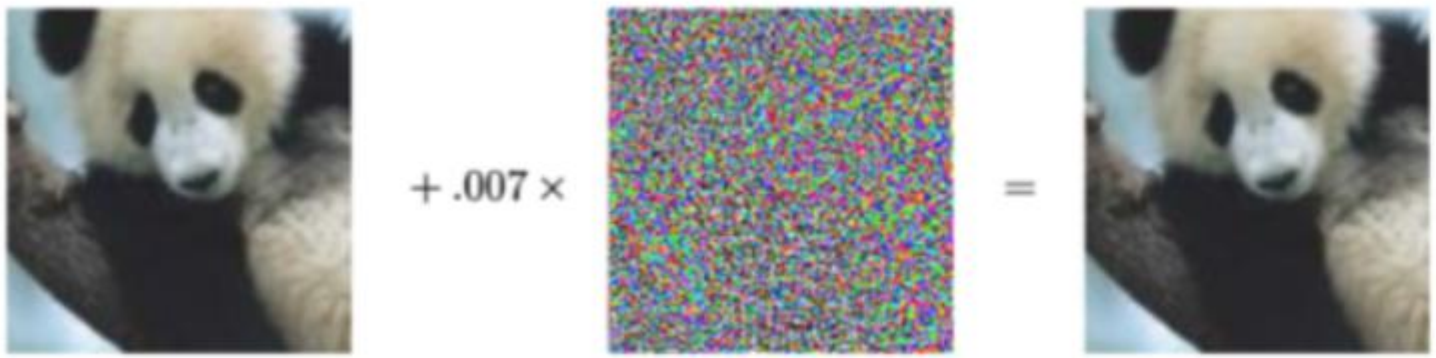
Background

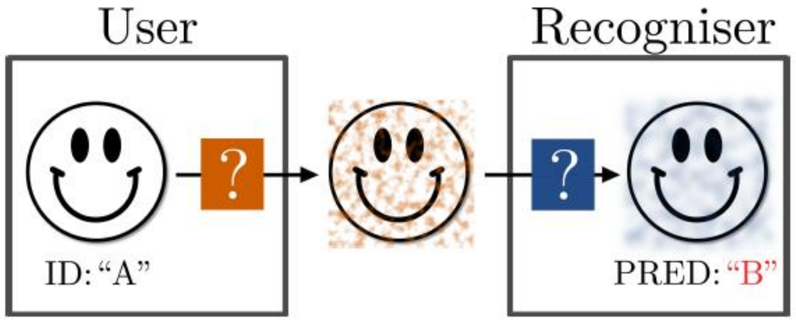
- Recent adversarial image perturbations (AIP) confuse recognition systems effectively without unpleasant artifacts
- However, how to **evaluate the AIP** in particular when the choice of **counter measure is unknown**.



Background

- AIP: Carefully crafted **additive perturbations** on the image that confuses a convnet while being nearly invisible to human eyes
- Counter measure: **Simple image processing tactics** to counter the AIP effects (e.g. blurring by small amount).





Motivations

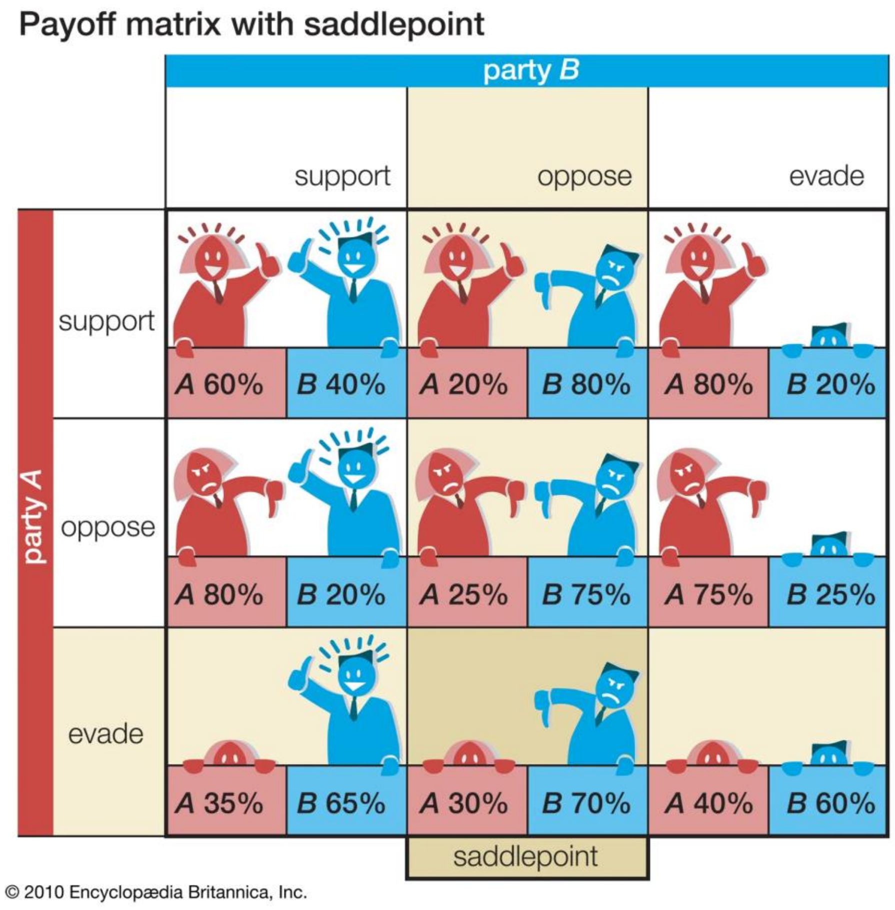
- Are AIPs still effective when counter measures are taken **?**
- Which is the best AIP strategy when the particular choice of counter measure is **unknown ?**

Game theory - Two Person Constant Sum Games

- The user-recognizer dynamics

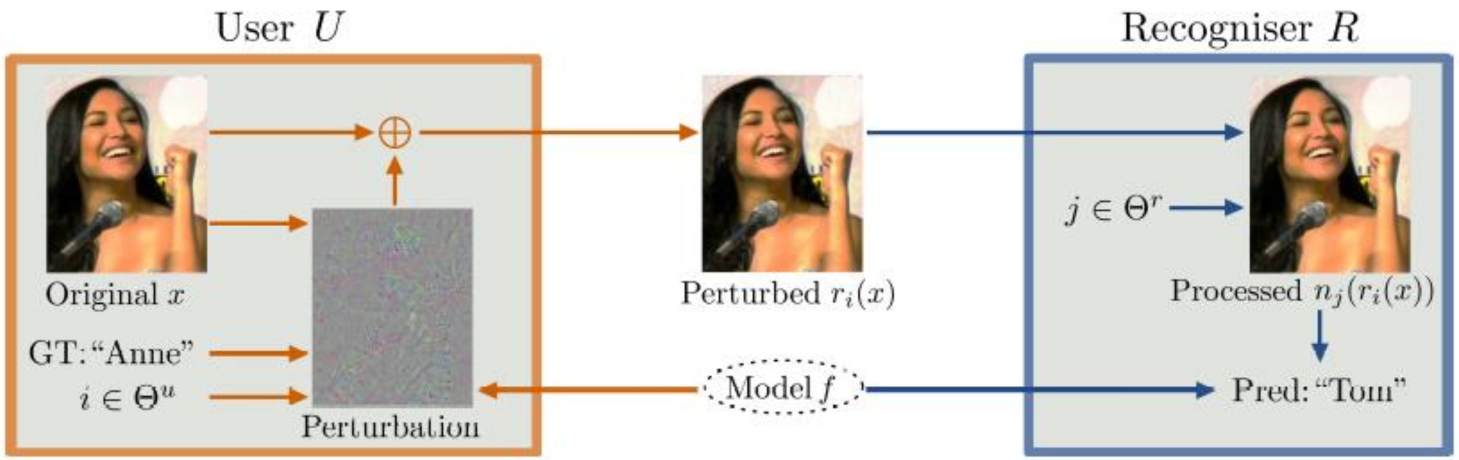
- The optimal strategy for the user that assures an **upper bound** on the recognition rate independent of the recognizer's counter measure

$$\arg\min_{\theta^u}\max_{\theta^r}\sum_{i,j}\theta_i^u\theta_j^r p_{ij}$$



User-Recogniser Game

- The user U and the recogniser R with designated strategy spaces, Θ^u and Θ^r.
- As a result of each player committing to strategies $i \in \Theta^u$ and $j \in \Theta^r$ respectively, R receives a payoff of p_{ij}, the recognition rate; U then receives a payoff of $1-p_{ij}$, the mis-recognition rate.



Known model. Each player is aware that the opponent uses f. This may be unrealistic, but provides a good starting point. Relaxation of this assumption is discussed in §3.3.

Payoff. When the players commit to strategies $i \in \Theta^u$ and $j \in \Theta^r$, R's payoff is the recognition rate on the test set:

$$p_{ij} = \mathop{\mathbb{P}}_{(\hat{x},\hat{y})\sim D}\left[\arg\max_y f^y\left(n_j\left(r_i\left(\hat{x}\right)\right)\right) = \hat{y}\right] \quad (3)$$

User-Recogniser Game

- Recogniser strategy

Translation	Gaussian additive noise	Blurring	Cropping & re-sizing	Combinations
T	N	B	C	TNBC

- Assume a finite strategy space, so we only consider a combination TNBC

User-Recogniser Game

- Adversarial Image Perturbation Strategies

Fast Gradient Vector	Fast Gradient Sign	Gradient Ascent	Basic Iterative	DeepFool
FGV	FGS	GA	BI	DF

$$\max_{t} \mathcal{L}\left(f\left(x+t\right), y\right) \qquad \text{s.t. } \|t\|_2 \leq \epsilon$$

User-Recogniser Game

- Adversarial Image Perturbation Strategies

Fast Gradient Vector	Fast Gradient Sign	Gradient Ascent	Basic Iterative	DeepFool	GA–Maximal Among Non-GT
FGV	**FGS**	**GA**	**BI**	**DF**	**GAMAN**

$$\max_{t} \mathcal{L}\left(f\left(x+t\right), y\right) \qquad \text{s.t.} \;\; ||t||_2 \leq \epsilon$$

- e.g. Fast Gradient Vector : one step gradient ascent

$$t^{\star} = -\gamma \nabla \mathcal{L}(x)$$

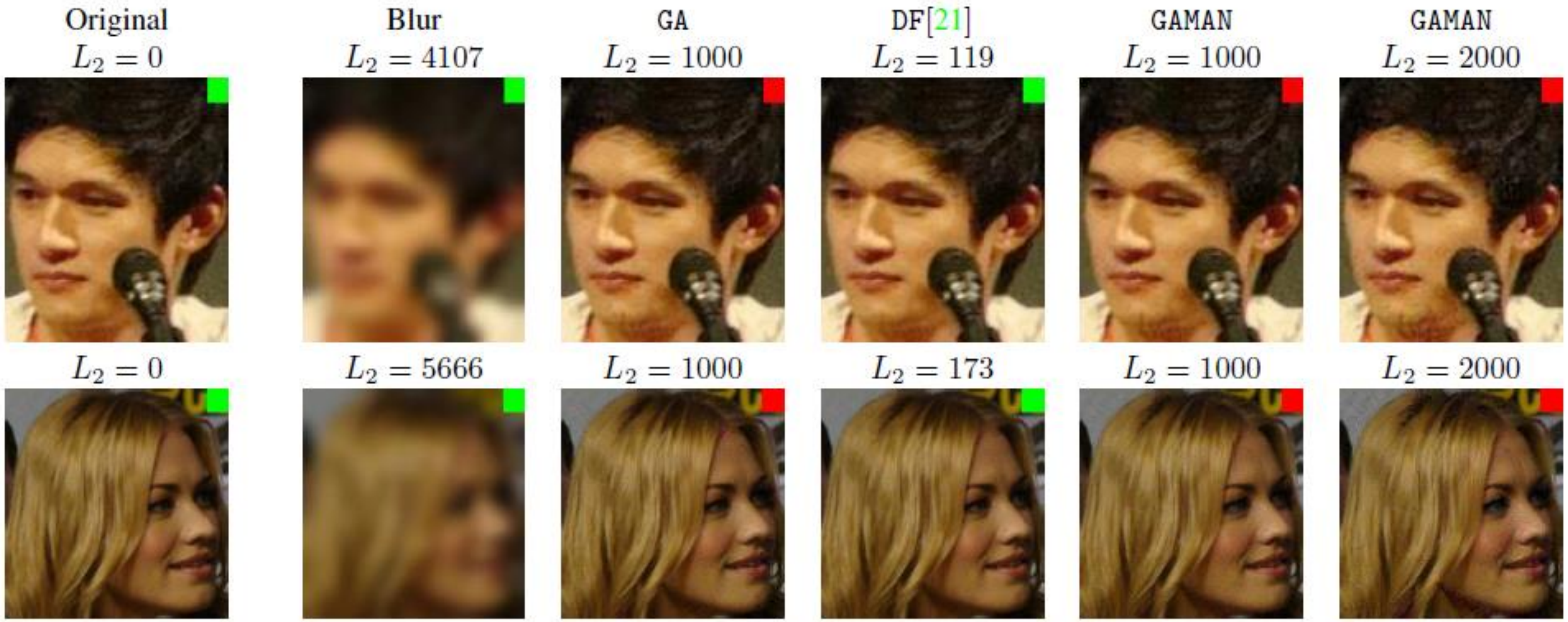
User-Recogniser Game

- Adversarial Image Perturbation Strategies

Variants	Loss $\mathcal{L}$	Stopping condition	Step size
FGS[6]	$-\log \hat{f}^y$	1 iteration	Fixed
FGV[31]	$-\log \hat{f}^y$	1 iteration	Fixed
BI[12]	$-\log \hat{f}^y$	K iterations	Fixed
GA	$-\log \hat{f}^y$	K iterations	Fixed
DF[21]	$f^{y^c} - f^y$	K it.$\vee$ fooled	Adaptive
GAMAN	$f^{y^\star} - f^y$	K iterations	Fixed

Experiments

Person identification



Experiments

- For each column (row), U's (R's) optimal strategy is marked orange (blue).

Perturb	∅	Proc	T	N	B	C	TNBC
None	87.8	87.8	87.6	64.0	81.2	85.4	87.3
BI[12]	0.0	8.3	15.8	16.8	28.6	27.4	17.6
GA	0.0	8.6	13.2	14.1	28.4	23.7	16.4
DF[21]	0.0	51.8	75.6	56.5	72.5	76.9	75.5
GAMAN	0.0	4.0	6.6	15.0	22.2	16.7	9.9

Table 3: Robustness analysis of AIPs on GoogleNet. AIPs are restricted to to $||\cdot||_2 \leq 1000$. Proc indicates the resizing and quantisation needed to convert AIP outputs to image files. (T, N, B, C) = (Translate, Noise, Blur, Crop).

Experiments - Vaccination

- For each column (row), U's (R's) optimal strategy is marked orange (blue).

$$\mathcal{L}(n_j(x+t))$$

$$\arg\min_{\theta^u} \max_{\theta^r} \sum_{i,j} \theta_i^u \theta_j^r p_{ij}$$

$$\theta^{u\star} = (\texttt{/B} : 61\%, \texttt{/TNBC} : 39\%)$$

User Θ^u	Recogniser Θ^r					
	Proc	T	N	B	C	TNBC
GAMAN	4.0	6.6	15.0	22.2	16.7	9.9
/T	2.5	2.3	11.6	18.5	7.2	4.9
/N	5.8	7.6	4.6	23.6	16.6	9.1
/B	0.4	0.8	8.6	5.8	3.1	1.4
/C	2.6	2.2	11.8	18.1	3.4	4.3
/TNBC	0.7	0.9	5.2	9.5	3.2	2.0

Table 4: Recogniser's payoff table p_{ij}, $i \in \Theta^u$ and $j \in \Theta^r$.

Experiments

- Selective AIP

Setup			$\mathcal{M}$ averaged		$\mathcal{B}$ averaged	
$\mathcal{M}$	$\mathcal{B}$	L_2	w/o AIP	w/ AIP	w/o AIP	w/ AIP
{G}	$\emptyset$	1000	87.8	4.0	-	-
{G}	{A}	1000	87.8	8.7	83.8	97.9
{A,R}	{V,G}	1000	87.4	17.7	87.0	97.7
{A,R}	{V,G}	2000	87.4	3.8	87.0	97.8

Table 5: Selective AIPs. AIPs are crafted to confuse $\mathcal{M}$ leaving $\mathcal{B}$ intact. [A,V,G,R] = [AlexNet, VGG, GoogleNet, ResNet152]. GAMAN has been used in all experiments. Reported performances are after Proc.

Thanks

Yuxuan Mu