Adversarial Image Perturbation (AIP) for Privacy Protection
A Game Theory Perspective

Seong Joon Oh, Mario Fritz, Bernt Schiele. MPI Informatics, Germany.

max planck institut informatik

github.com/coallaoh/AIP

Motivation

Privacy is becoming a greater concern.

- Social media photos contain private information.
- Improvement of ML and CV makes it easier for malicious users to extract such information.

Image blurring doesn't work.

- ML systems can adapt & use context [2].

AIP is superb – with caveats.

- Works well for fixed, fully known target model.
- But what if target is uncertain?
- Active research on AIP defense mechanisms.

Recognised by machine

Avoided machine recognition

Game Theory to Model Uncertainty

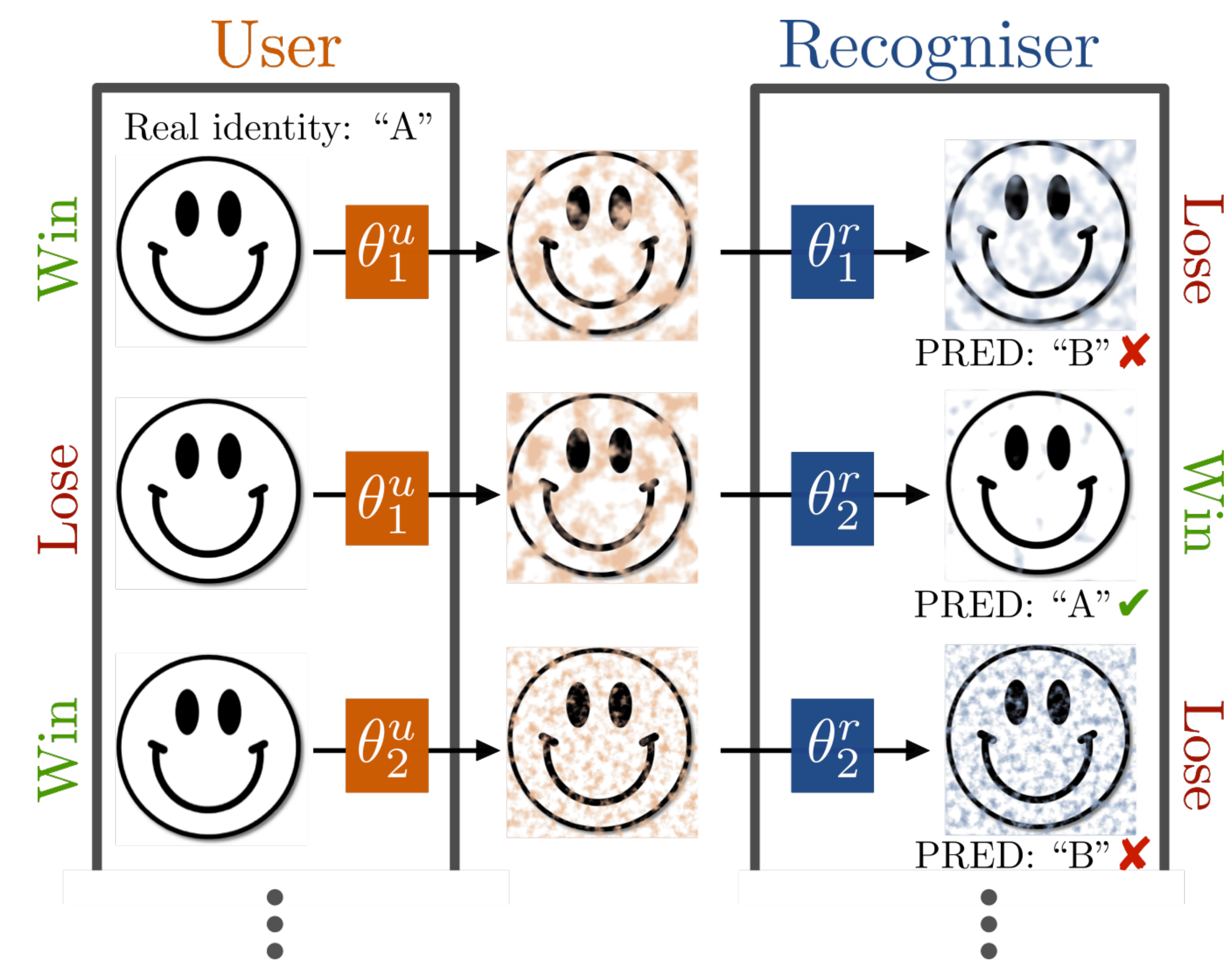
GT is a tool for systematically linking
Input: Players with explicit goals (rewards) and possible choices of actions (strategies).
to
Output: Guarantee on each player's reward, *independent* of the others' actions.

Equilibria

- Equilibrium: best strategy against worst opponent.

$$\theta^{u\star} := \arg\min_{\theta^u} \max_{\theta^r} \sum_{i,j} \theta^u_i \theta^r_j p_{ij}$$

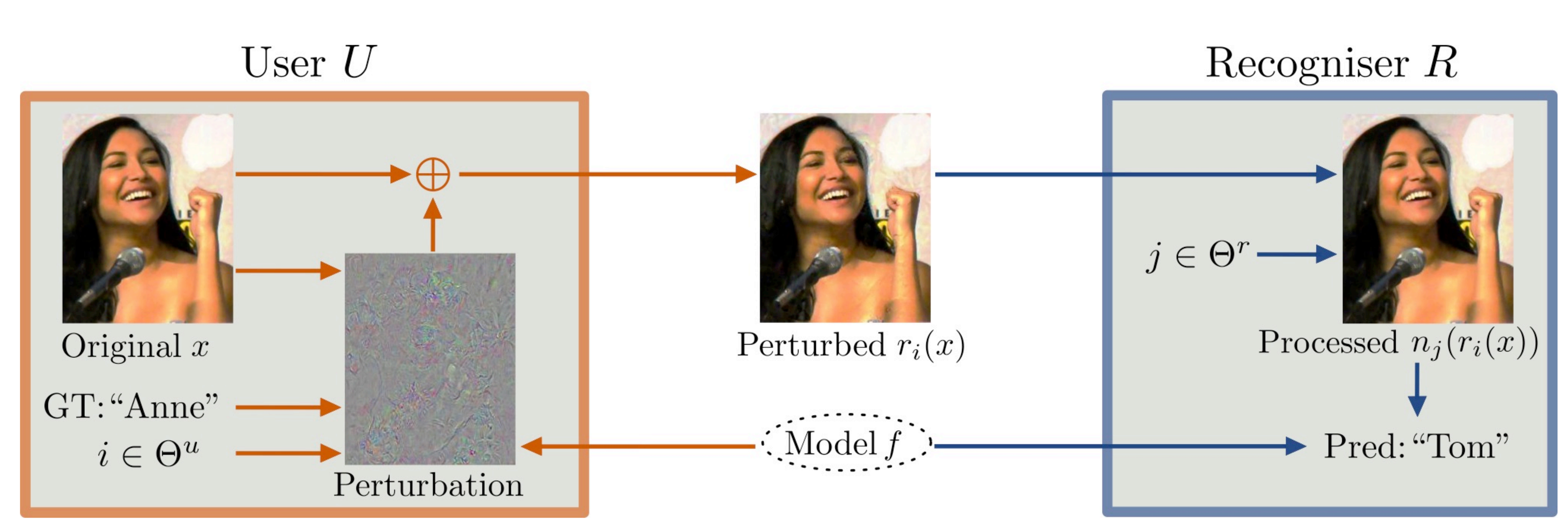
- When $\theta^{u\star}$ is played, *U*'s reward is lower bounded by v, independent of *R*'s action. Independence!



Dynamics of the image perturbation game

User (*U*) wants to avoid recognition.
Recogniser (*R*) wants to re-enable recognition.
They do not know each other's strategy.

User-Recogniser Game over Privacy



- **User (*U*)** : Applies a type of AIP *i* on her image to avoid recognition by model *f*.
- **Recogniser (*R*)** : Applies a type of image transformation *j* on the image to nullify the effect of AIP; then pass it to model *f*.
- **Rewards** : Recognition *success* (*failure*) rate for *R* (*U*).

Extensions for future work

- *R* can change the model *f* – AIP against black-box models needed.
- Non-constant sum game: Nash equilibria.

Takeaways

1. AIPs can protect privacy while preserving image aesthetics.
2. Derive explicit privacy guarantees via GT.
3. Schemes for robust AIPs.

Case Study: Person Recognition [1]

R's strategy space

AIPs are brittle; small translation (T), Gaussian noise (N), blurring (B), or cropping & resizing (C) is already nullifying. [3]
R chooses his image transformation from {None, T, N, B, C, TNBC}.

U's strategy space

GAMAN: Our reformulation of DeepFool [4] as gradient ascent optimisation. Superior robustness.

Perturb	∅	T	N	B	C	TNBC
None	87.8	87.6	64.0	81.2	85.4	87.3
BI	0.0	15.8	16.8	28.6	27.4	17.6
GA	0.0	13.2	14.1	28.4	23.7	16.4
DF[4]	0.0	75.6	56.5	72.5	76.9	75.5
GAMAN	0.0	6.6	15.0	22.2	16.7	9.9

Vaccination: Adapt GAMAN against each of R's image transofrmation strategy by backpropagating through each transformation.
U chooses her AIP from {GAMAN, /T, /N, /B, /C, /TNBC}.

Reward table

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

- *R*'s transformation strategies do re-enable recognition.
- *U*'s vaccination strategies do work against the speicifc *R* strategy.

User-Recogniser Game and Guarantees

Equilibria:
$\theta^{u\star}$is [/B: 61%, /TNBC: 39%].
$\theta^{r\star}$ is [N: 52%, B: 48%].
Value of the game v is 7.3%.

Interpretation:
If *U* mixes AIP types (/B, /TNBC) with probabilities (61%, 39%), then chance of recognition will be < 7.3%, no matter what *R* does.

References

[1] Person Recognition in Personal Photo Collections. Oh et al. ICCV'15.
[2] Faceless Person Recognition; Privacy Implications in Social Media. Oh et al. ECCV'16.
[3] Assessing Threat of Adversarial Examples on Deep Neural Networks. Graese et al. ICMLA'16.
[4] DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. Moosavi-Dezfooli et al. CVPR'16.

Acknowledgement: This research was supported by the German Research Foundation (DFG CRC 1223).