Adversarial Image Perturbation (AIP) for Privacy Protection
A Game Theory Perspective
Seong Joon Oh, Mario Fritz, Bernt Schiele. MPI Informatics, Germany.

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.

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

Equilibria
- Equilibrium: best strategy against worst opponent.
  \[ \theta^* \] := \arg \min_{\theta_u} \max_{\theta_r} \sum_{i,j} \theta^*_u \theta^*_r p_{ij} \]
- When \[ \theta^* \] is played, U's reward is lower bounded by \[ \psi^* \], independent of R's action. Independence!

Dynamics of the image perturbation game
User (U) wants to avoid recognition.
Recognizer (R) wants to re-enable recognition.
They do not know each other's strategy.

User-Recognizer Game over Privacy

User (U) : Applies a type of AIP i on her image to avoid recognition by model f.
Recognizer (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
\[ \Psi(\theta^*_u) \] = \arg \max_{\psi^*} \psi^* \]

Reward table

User-Recognizer Game and Guarantees
Equilibria:
\[ \theta^* \] is [B: 61%, /TNBC: 39%].
\[ \psi^* \] is [N: 52%, B: 48%].
Value of the game \[ \psi^* \] is 73%.

Interpretation:
If U mixes AIP types (/B, /TNBC) with probabilities (61%, 39%), then chance of recognition will be < 73%, no matter what R does.

References

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