Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study
Zhenyu Wu (TAMU), Zhangyang Wang (TAMU), Zhaowen Wang (Adobe Research), Hailin Jin (Adobe Research)

Motivation
A smart home camera system is expected to:
- Be able to recognize important events and assist daily life by understanding videos
- But unable to obtain too sensitive visual attributes that can intrude people’s privacy.

Problem Formulation: Adversarial Training

\[
\begin{align*}
\min_{f_d} & \ell_T(f_t(f_d(X)), Y_T) + \gamma \max_{f_b \in P} \ell_B(f_b(f_d(X)), Y_B) \\
\text{subject to} & \quad f_d: \text{a model to perform the target task} \\
& \quad f_b: \text{a model to perform the privacy attributes prediction task} \\
& \quad P: \text{privacy prediction function family} \\
& \quad \ell_T: \text{cost function for the performance on } T \\
& \quad \ell_B: \text{privacy budget cost function to evaluate the privacy leak risk of } X: \text{the higher privacy leak risk.}
\end{align*}
\]

For the solved \( f_d \) the two goals should be simultaneously satisfied: (1) there exists (3) at least one \( f_d \) function that can predict \( Y_T \) from \( f_d(X) \) well; (2) for all (4) \( f_b \) functions \( \in \) \( P \), none of them (even the best one) can reliably predict \( Y_b \) from \( f_d(X) \).

Naive Adversarial Learning Implementation

Experiments (i): Identity-Preserving Action Recognition on SBU

Method 1: downsampling raw RGB frames under different ratios
Method 2 (Proposed): adversarial training on RGB frames, budget model ensemble w/o restarting.
Method 3 (Proposed): adversarial training on RGB frames, budget model ensemble w/ restarting.
Method 4: detect & crop out faces from RGB frames.
Method 5: detect & crop out whole actor bodies.

Two-Fold Evaluation Protocol
1. Evaluate the target utility task: as standard
2. Evaluate the privacy protection effect: train a different set of unseen \( f_b \) models on \( f_d(X) \), see if they can generalize

Experiments (ii): Cross-Dataset Training on UCF=101 and VISPR

Codes + pre-trained models: [GitHub](https://github.com/wuzhenyusjtu/Privacy-AdversarialLearning/)

Towards Learning Model-Agnostic Privacy Protection

- Naive implementation by choosing a very strong \( f_b \) is insufficient (over-fitting one only)
- Improved Solution 1: Budget Model Re-starting and Re-fitting
- Improved Solution 2: Budget Model Ensemble Training

We approximate the continuous \( P \) with a discrete set of \( M \) sample functions. Assuming the budget model ensemble \( \{f_b^i\}_{i=1}^M \), we turn to minimizing the following discretized surrogate:

\[
\begin{align*}
\min_{f_d} & \ell_T(f_t(f_d(X)), Y_T) + \gamma \max_{i \in \{1, 2, \ldots, M\}} \ell_B(f_b^i(f_d(X)) \\
\end{align*}
\]