Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study

Zhenyu Wu [1] Zhangyang Wang [1] Zhaowen Wang [2] Hailin Jin [2]

[1]Texas A&M University

[2]Adobe Research

February 11, 2019

The Dilemma

- A smart home camera system is expected to
  - **Be able to** recognize important events and assist people's daily life by understanding videos
  - **Be unable to** obtain "too sensitive" visual information that can intrude people's privacy.

Would classical cryptographic solutions suffice? No.

- They secure the communication against unauthorized access from attackers
- ...but not applicable to preventing the abuse by authorized agents (e.g., the backend analytics)

What were there?

- Privacy Protection in Computer Vision Systems
  - Transmit feature descriptors to the cloud? Not safe
  - Homomorphic cryptographic solution? Expensive, working on only simple classifiers
  - Downsample the video aggressively, and strategically? Cheap, works empirically, but usually no competitive trade-off
  - A few game-theoretic or learning-based recent solutions ... IMPORTANT to distinguish between model-specific and model-agnostic privacy!

- Privacy Protection in Social Media and Photo Sharing
  - Add empirical obfuscations? Not safe
  - Deep learning-based adversarial perturbations? Model-specific privacy, also with a different goal with ours: they wish to cause minimum perceptual quality loss to those photos

A Formal Problem Definition

$$\min_{f_T, f_d} L_T(f_T(f_d(X)), Y_T) + \gamma L_B(f_d(X)), \qquad (1)$$

- X: raw visual data captured by camera:
  - target task $\mathcal{T}$, e.g., action recognition or visual tracking
  - privacy budget $\mathcal{B}$, e.g, leak of identity of other privacy attributes.
- f_T: a model to perform the target task $\mathcal{T}$ on its input data.
- Y_T: a label set provided on X for $\mathcal{T}$.
- L_T: cost function for the performance on $\mathcal{T}$, e.g., action recognition accuracy.
- L_B: budget cost function to evaluate the privacy leak risk of X: the larger L_B, the higher privacy leak risk.

A Formal Problem Definition (Cont.)

Our goal is to seek such an active degradation function f_d to transform the original X for both L_T and L_B, such that:

- The achievable target task performance L_T is minimally affected compared to when using the raw data, i.e., $\min_{f_T, f_d} L_T(f_T(f_d(X)), Y_T) \approx \min_{f'_T} L_T(f'_T(X), Y_T)$.
- The privacy budget L_B is greatly suppressed compared to raw data: $L_B(f_d(X)) \ll L_B(X)$.

How to Define Privacy Cost?

The definition of the privacy budget cost L_B is not straightforward.

- Privacy is subjective, and usually needs to be placed in concrete application contexts, often in a task-driven way.
- We denote the privacy-related annotations (such as identity label) as Y_B, and rewrite $L_B(f_d(X))$ as $L_B(f_b(f_d(X)), Y_B)$, where f_b denotes a budget model to predict the corresponding privacy information.
- *Different from* L_T, minimizing L_B will encourage $f_b(f_d(X))$ to diverge from Y_B as much as possible.

The ∃-∀ Challenge

Define a privacy prediction function family $\mathcal{P}$: $f_d(X) \rightarrow Y_B$, the ideal privacy protection of f_d should be **suppressing every possible model** f_b from $\mathcal{P}$ (worst-case guaranteed protection)

$$\min_{f_T, f_d} L_T(f_T(f_d(X), Y_T) + \gamma \underline{\max_{f_b \in \mathcal{P}} L_B(f_b(f_d(X)), Y_B)}. \quad (2)$$

For the solved f_d, the two goals should be simultaneously satisfied: (1) there **exists** ("∃") at least one f_T function that can predict Y_T from $f_d(X)$ well; (2) **for all** ("∀") f_b functions $\in \mathcal{P}$, **none of them** (even the best one) can reliably predict Y_B from $f_d(X)$.

(Naive) Adversarial Learning Implementation

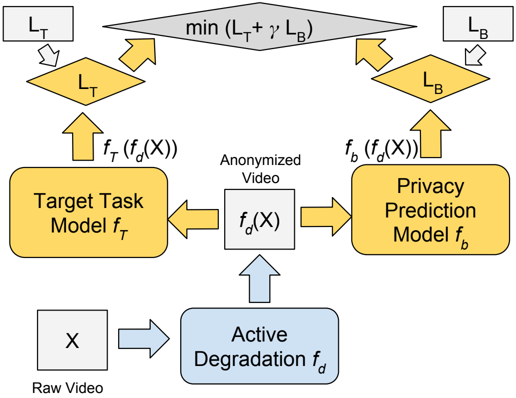


Figure 1: The basic adversarial training framework for privacy-preserving visual recognition.

Learning Model-Agnostic Privacy Protection

- Naive implementation by choosing a very strong f_b is insufficient (overfitting one only).
- **Improved Solution 1**: Budget Model Re-starting and Re-fitting
- **Improved Solution 2**: Budget Model Ensemble Training
  - We approximate the continuous $\mathcal{P}$ 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 of (2):

$$\min_{f_T, f_d} L_T(f_T(f_d(X), Y_T) + \gamma \max_{i \in \{1,2,\ldots,M\}} L_B(f_b^i(f_d(X))). \tag{3}$$

  - Sampling the meta model space and always suppressing the most confident model

Two-Fold Evaluation Protocol

- The double-sided problem calls for two-folds evaluation:
  - Is target task utility maintained after active degradation? (standard)
  - Is privacy protected against any possible (unseen) privacy prediction model? (non-standard)

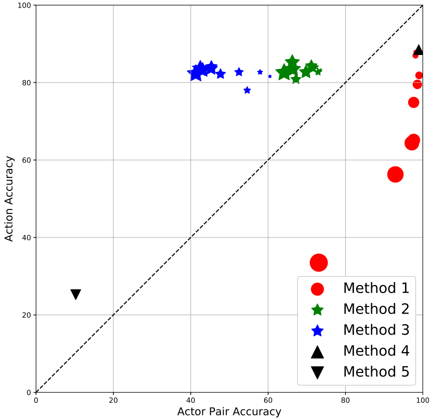
- For the second evaluation:
  - We first sample a different, unseen set of N privacy prediction models from $\mathcal{P}$
  - We then **train** each of them to predict privacy information, over the **degraded training set** X by applying the learned f_d
  - We finally apply them to the **degraded testing set** after applying the learned f_d, and the **highest accuracy** achieved among the N models is used to approximately represent the "worst-case privacy protection"

Experiments (i): SBU dataset

$\mathcal{T}$: action recognition
$\mathcal{B}$: actor pair identification

- **Method 1**: downsampling raw RGB frames under different ratios.
- **Method 2** (Proposed): applying the proposed adversarial training to RGB frames, using budget model ensemble *without* restarting.
- **Method 3** (Proposed): applying the proposed adversarial training to RGB frames, using budget model ensemble *with* restarting.
- **Method 4**: detect & crop out faces from RGB frames.
- **Method 5**: detect & crop out whole actor bodies.

Figure 2: The SBU trade-off plot.



Experiment (ii): UCF-101 + VISPR

- $\mathcal{T}$: action recognition
- $\mathcal{B}$: protection of multiple privacy attributes (VISPR-17/7)
- Cross-dataset training and evaluation
  - UCF101 dataset: 101 different action classes.
  - Visual Privacy (VISPR) dataset: 22, 167 images manually annotated with many privacy attributes, e.g. face, race, gender, skin color, age group...

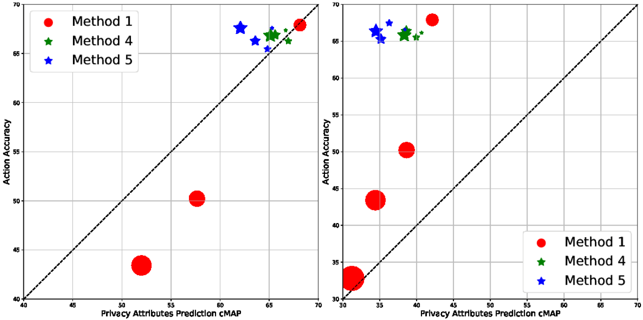


Figure 3: UCF-101/VISPR trade-off plot. Left: VISPR-17; Right: VISPR-7.

Open Questions: A lot

- Current budget model ensemble is a rough discretized approximation of $\mathcal{P}$. More elegant to tackle this $\forall$ optimization is desired.
- Stabilizing the adversarial training
- Need better theory – information theory or game theory?
- Collecting datasets with both $\mathcal{T}$ and $\mathcal{B}$ well defined