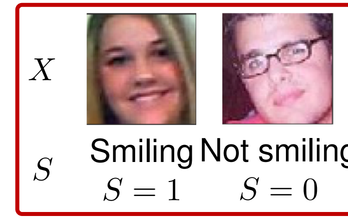
Discovering Information-Leaking Samples and Features

Hsiang Hsu, Shahab Asoodeh, and Flavio P. Calmon, School of Engineering and Applied Science, Harvard University



- Context-aware privacy, e.g., $\begin{cases}\text{Information-theoretic privacy}\\ \text{Generative adversarial privacy (GAP)}\end{cases}$

  $\Rightarrow$ Having data X and private attribute S
- A natural step in designing a privacy mechanism $\Rightarrow$ discovering information-leaking samples and features
- Private attributes $\mathbf{s} \in \mathcal{S}$, samples $\mathbf{x} \in \mathcal{X}$, features $\mathbf{x}^j \in \mathcal{X}$

  the information density $\begin{cases} i(\mathbf{s};\mathbf{x}) \triangleq \log \frac{P_{S,X}(\mathbf{s},\mathbf{x})}{P_S(\mathbf{s})P_X(\mathbf{x})} \Rightarrow \text{information-leaking score of samples}\\ i(\mathbf{s};\mathbf{x}^j) \triangleq \log \frac{P_{S,X}(\mathbf{s},\mathbf{x}^j)}{P_S(\mathbf{s})P_X(\mathbf{x}^j)} \Rightarrow \text{information-leaking score of features}\end{cases}$
- Thresholded Information Density Estimator (TIDE)
  - Donsker-Varadhan (DV) representation of KL Divergence

    $D(P_{S,X}\|P_S P_X) = \sup_{g:\mathcal{S}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}] \Rightarrow g^*(s,x) = i(s;x)$
  - Restricted g to $\mathcal{G}(\Theta)$: continuous functions g_θ $\begin{cases}\text{Bounded by } M\\ \text{Parameterized by } \theta \text{ in a compact domain } \Theta \subset \mathbb{R}^d\end{cases}$
  - TIDE: $\hat{g}_n(s,x) = \operatorname{argmax}_{g_\theta \in \mathcal{G}(\Theta)} \mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[e^{g_\theta(S,X)}]$
- Experiments

Original Images →

Information-leaking patches by TIDE →

Gaussian noise on those patches
$\Rightarrow$ preserves gender
$\Rightarrow$ hides emotion →