Decentralized Attribution of Generative Models

Published as a conference paper at ICLR 2021

Changhoon Kim*, Yi Ren*, and Yezhou Yang

Design Informatics Lab

Introduction


[1]

Malicious Personation:

Attackers ***create*** fake and disseminate as ***real***

Examples:

- Malicious Personation (LinkedIn profile)
- Deepfake pornography


[2]

Copyright Infringement:

Attackers claim **ownership** of generated content

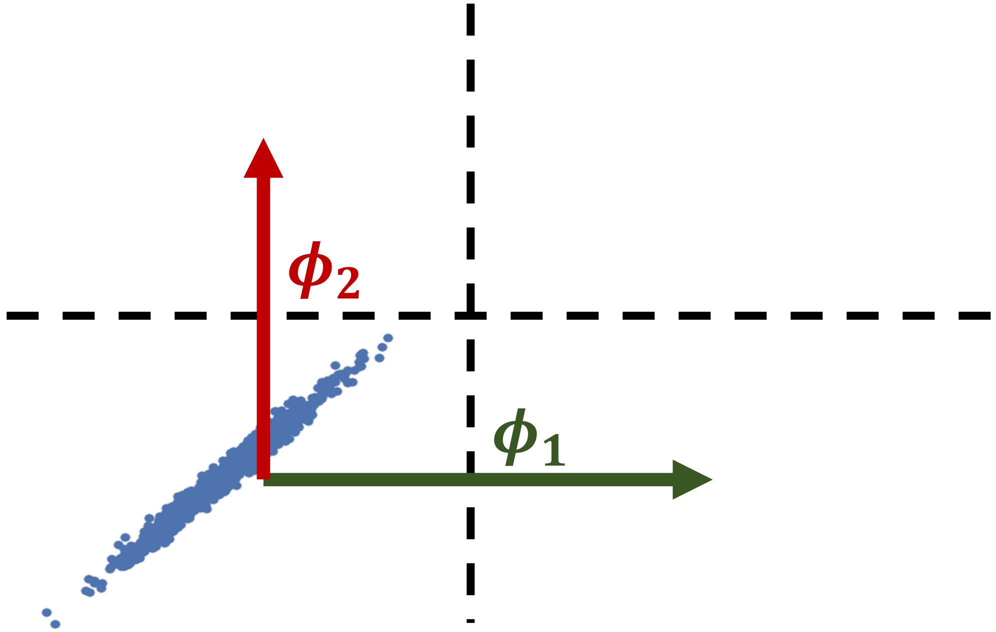
Examples:

- Replicates of arts

[1] Raphael S. (2019, Jul 14). Experts: Spy used AI-generated face to connect with targets. APNews.

[2] Portrait of Edmond Belamy, 2018, created by GAN (Generative Adversarial Network)

Introduction: Certified Model Attribution



Certified Model Attribution:

- Determine which model the content comes from

Problem:

- The set of models is growing.

Contribution:

- Proposed sufficient conditions of watermarking individual models

Methods: Criteria

- **Distinguishability:** Accuracy of key at classifying G_ϕ against $\mathbb{D}$.

$$D(G_\phi) \coloneqq \frac{1}{2}\mathbb{E}_{x \sim P_{G_\phi},\, x_0 \sim P_{\mathbb{D}}}\left[\mathbf{1}\big(f_\phi(x) = 1\big) + \mathbf{1}\big(f_\phi(x_0) = -1\big)\right]$$

, where ϕ is the key and $f_\phi(x) = sign(\phi^T x)$.

- **Attributability:** Averaged classification accuracy of each generators.

$$A(\mathfrak{G}) \coloneqq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x \sim P_{G_{\phi_i}}}\left[\phi_i^T x > 0,\ \phi_j^T x < 0, \forall j \neq i\right]$$

- **Lack of Generation Quality:**

- FID[1] score

- Norm of mean output perturbation

$$\Delta x(\phi) \coloneqq \mathbb{E}_{z \sim P_z}\left[G_\phi(z) - G_0(z)\right]$$

, where G_0 is the model owner's generator.

[1] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." *arXiv preprint arXiv:1706.08500* (2017).

Methods: Training

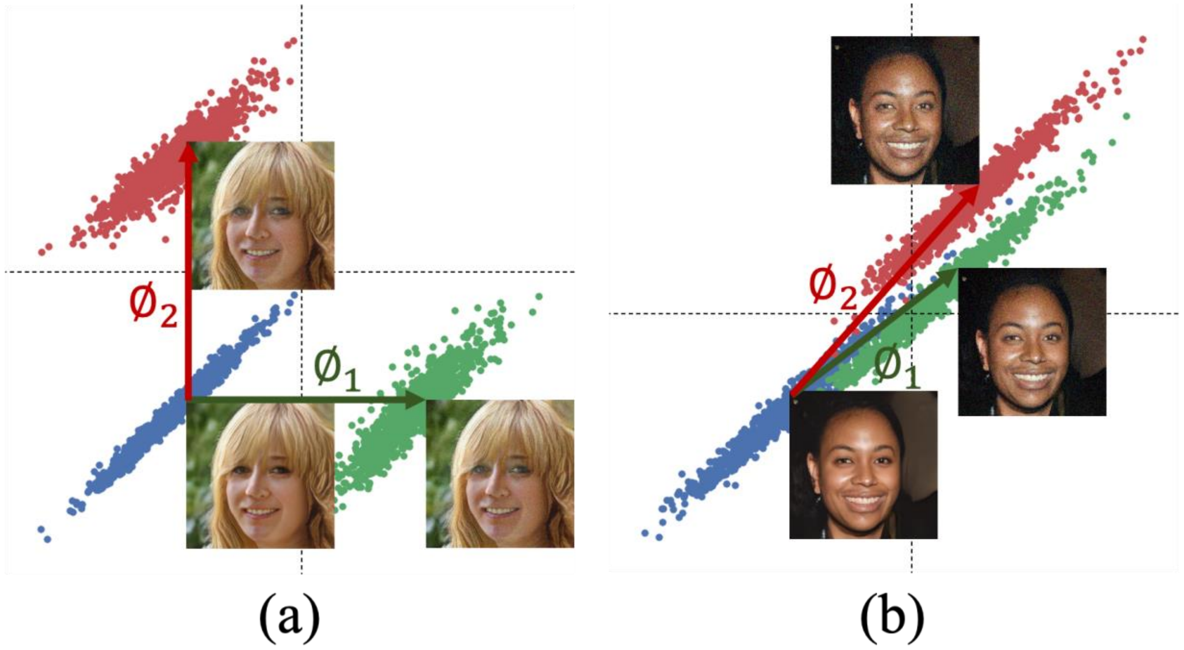


Fig. 2: (a) 90 deg. threshold. (b) Acute threshold

Step 1. Key generation to satisfy the sufficient conditions:

$$\phi_i = \ argmin_{\phi_i} \ \mathbb{E}_{x\sim\mathbb{D},G_0}\left[\max\{0, 1+\phi_i^T x\}\right] + \sum_{j=1}^{i-1} \max\{0, \phi_j^T \phi_i\}$$

Step 2. Train user-end generative models to be distinguishable:

$$Obj.\ min_\theta \mathbb{E}_{y\sim D_{\gamma,\phi}}\left[||G_\phi(z;\theta) - y||^2\right]$$

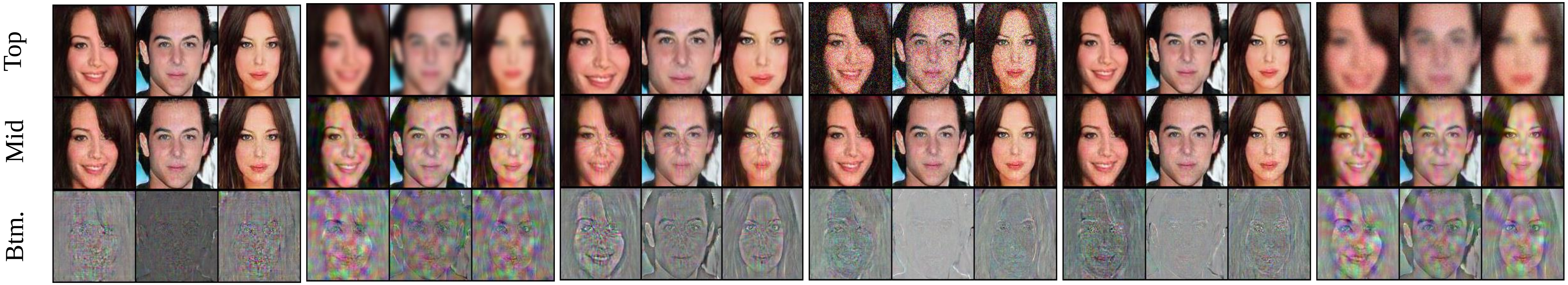
Results

Keys | **Corresponding Examples**

Fig. 4: Visualization of keys and contents

- Quality: $\| \Delta x \| = 36.04$, $FID_0 = 12.43 \Rightarrow FID = 35.23$

Robust Training



(a) Non-robust (b) Blurring (c) Cropping (d) Noise (e) JPEG (f) Combination

Fig. 5: Visualization of Robust Training

- Scenario: Adversary can modify the outputs of the generative models.
- Assumption: The post-processes are known.
- Trade-off: Robustness and Generation Quality
- Objective function of robust user-end model:

$$min_{\theta_i} \mathbb{E}_{z\sim P_z, T\in P_T}\left[max\left\{0, 1 - f_{\phi_i}\left(T\left(G_{\phi_i}(z;\theta_i)\right)\right)\right\} + C||G_0(z;\theta_0) - G_{\phi_i}(z;\theta_i)||^2\right]$$

Summary & Future Direction

- Sufficient conditions:
  - Keys satisfy distinguishability.
  - Keys are mutually orthogonal.
- This conditions relies on linear classification.
- Toward Nonlinear classification:
  - Sufficient conditions under nonlinear classifier
  - Increase the capacity of keys

Acknowledgement

- NSF Robust Intelligence Program (1750082)
- ONR (N00014-18-1-2761)
- Amazon AWS MLRA

Thank You

Email: kch@asu.edu

Poster session 3