UNIVERSITY OF CALIFORNIA MERCED

ADVERSARIAL LEARNING OF PRIVACY-PRESERVING AND TASK-ORIENTED REPRESENTATIONS

Taihong Xiao[1], Yi-Hsuan Tsai[2], Kihyuk Sohn[2],
Manmohan Chandraker[2,3], Ming-Hsuan Yang[1]

[1]University of California, Merced
[2]NEC Laboratories America
[3]University of California, San Diego

Content

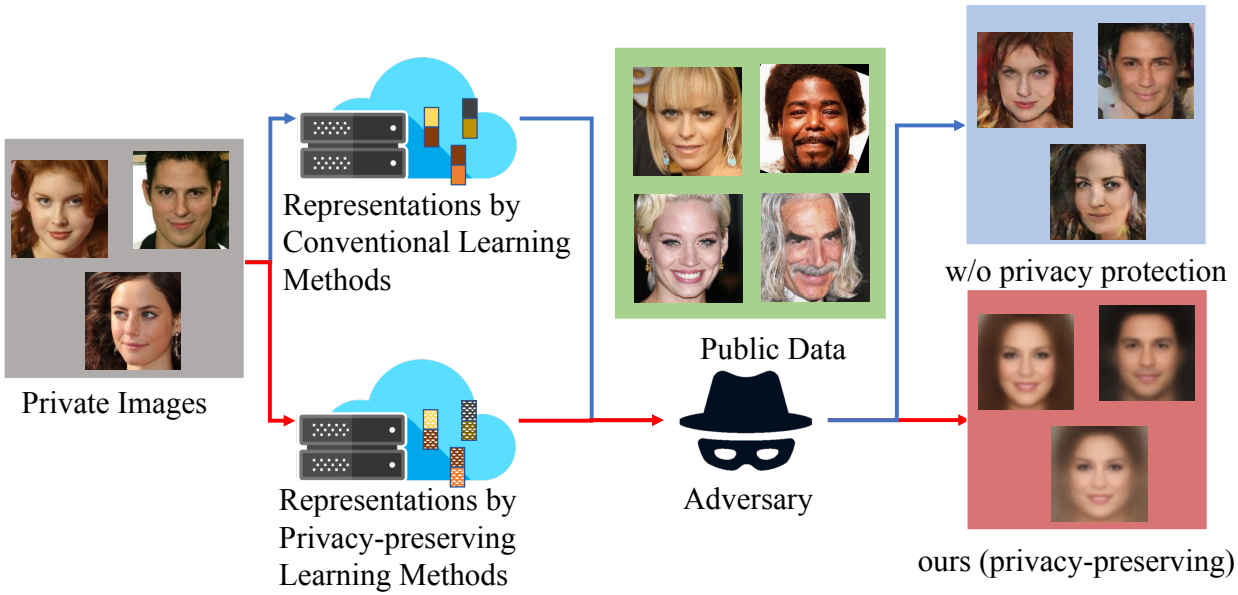
- Introduction
- Proposed Algorithm
- Experiments

Introduction

- Privacy risks in the machine learning cloud services
- Using deep features to protect the privacy
- Model inversion techniques
  - White-box: the utility model and its weights are fully transparent to the adversary
  - Black-box: the adversary can make unlimited inferences of their own data to recover input from acquired features of private user data.

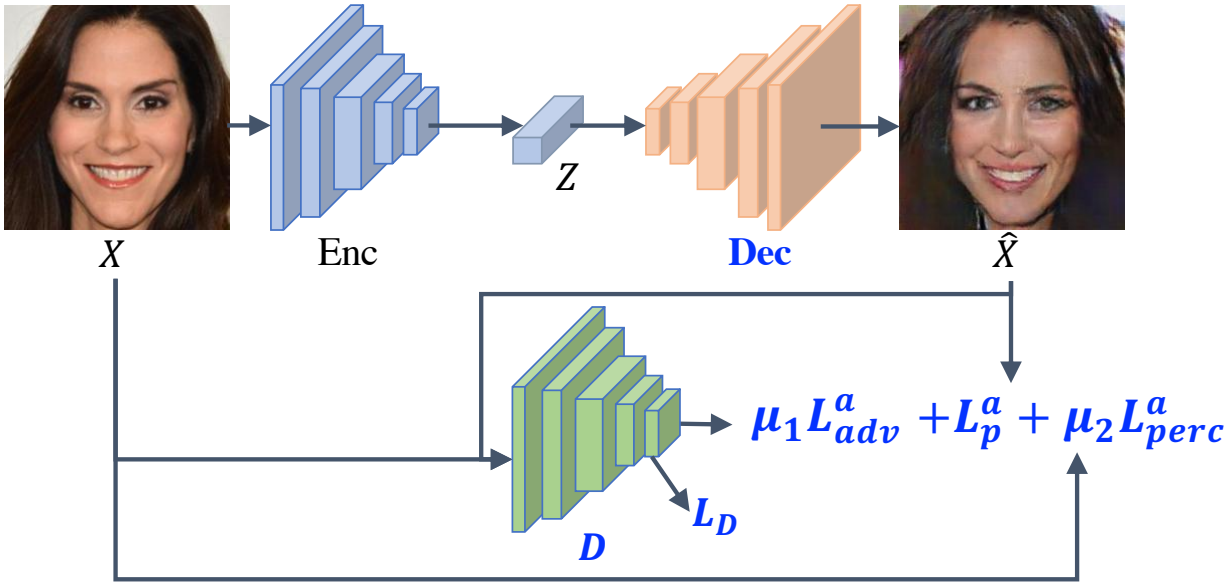
Introduction

- Focus on defense against a black-box model inversion attack in the context of face attribute analysis by adversarial learning.



Proposed Algorithm

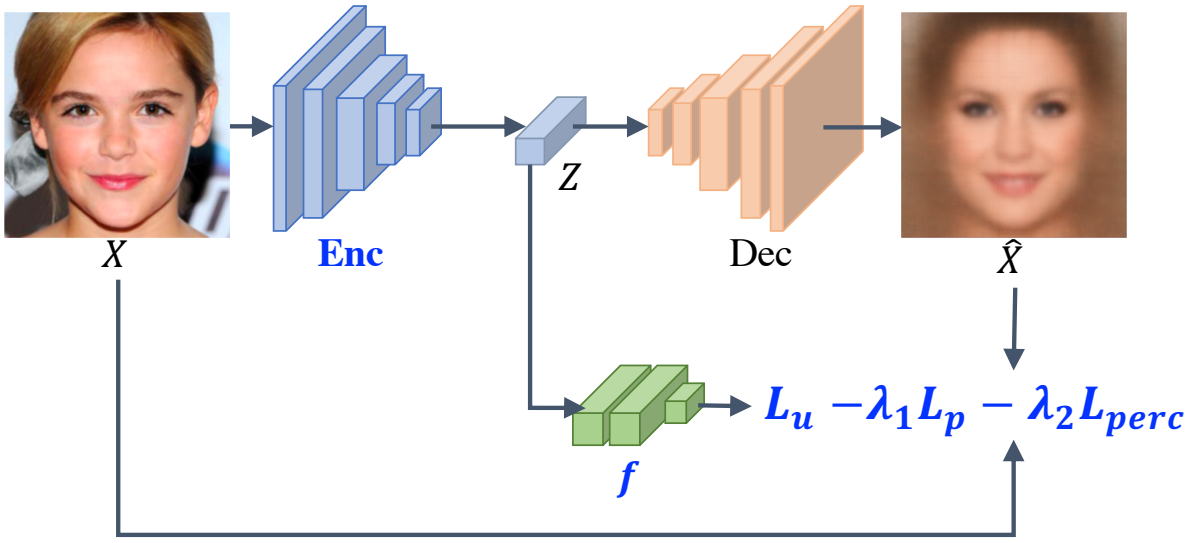
- Adversary: learn to invert



Update Dec and *D* using X ∈ $\mathcal{X}_2$ while fixing Enc and f .

Proposed Algorithm

- Protector: learn "not" to invert



Update Enc and f using X ∈ $\mathcal{X}_1$ while fixing Dec.

Experiments

- Utility Metric
  - Matthews correlation coefficient (MCC)

- Privacy Metric
  - Face Similarity
  - Feature Similarity
  - SSIM/PSNR

Experiments

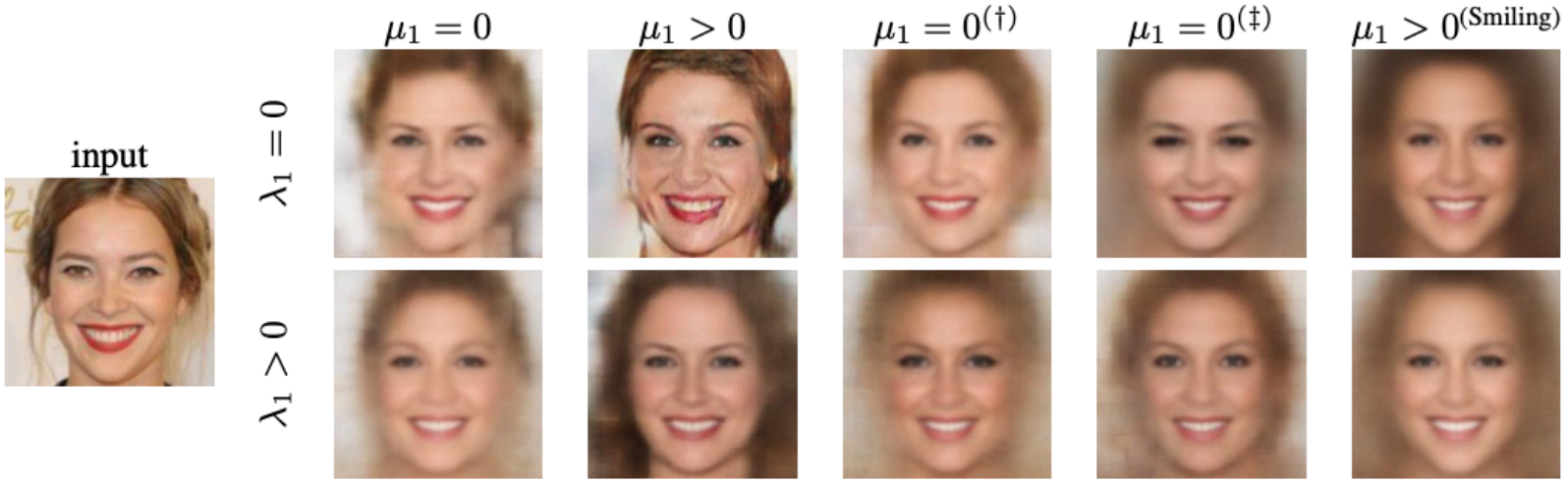
- The results on facial attribute prediction.

ID	Enc	Dec^a	Mean MCC ↑	Face Sim. ↓	Feature Sim. ↓	SSIM	PSNR
1	$\lambda_1 = 0$	$\mu_1 = 0$	0.641	0.551	0.835	0.231	13.738
2	$\lambda_1 > 0$	$\mu_1 = 0$	0.612	0.515	0.574	0.221	13.423
3	$\lambda_1 = 0$	$\mu_1 > 0$	0.641	0.585	0.835	0.240	14.065
4	$\lambda_1 > 0$	$\mu_1 > 0$	0.612	0.513	0.574	0.277	13.803
With more data for training Dec^a (ID #5 and #6) and both Enc and Dec^a (ID #7 and #8)							
5	$\lambda_1 = 0^\dagger$	$\mu_1 = 0$	0.641	0.594	0.864	0.250	14.132
6	$\lambda_1 > 0^\dagger$	$\mu_1 = 0$	0.612	0.541	0.633	0.222	13.703
7	$\lambda_1 = 0^\ddagger$	$\mu_1 = 0$	0.651	0.579	0.834	0.263	14.432
8	$\lambda_1 > 0^\ddagger$	$\mu_1 = 0$	0.630	0.550	0.591	0.231	13.334
Single (Smiling) attribute prediction. MCC for Smiling attribute is reported in the parenthesis.							
9	$\lambda_1 = 0$	$\mu_1 > 0$	0.001 (0.851)	0.460	0.494	0.204	13.214
10	$\lambda_1 > 0$	$\mu_1 > 0$	0.044 (0.862)	0.424	0.489	0.189	12.958

The rows with grey shadow are our results.

Experiments

- Visualization of reconstructions.



Conclusion

- An adversarial learning framework to protect privacy while maintaining utility performance.
- For more information, please check our paper Adversarial Learning of Privacy-Preserving and Task-Oriented Representations.