Toward Privacy and Utility Preserving Image Representation

Ahmadreza Mosallanezhad, Yasin N. Silva, Michelle V. Mancenido, Huan Liu

CAOE Center for Accelerating Operational Efficiency
A DEPARTMENT OF HOMELAND SECURITY CENTER OF EXCELLENCE

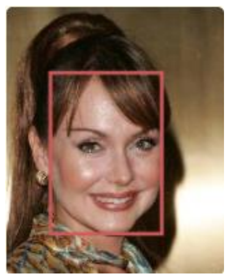
1. Introduction

Background: face images are rich data items that are useful and can easily be collected in many applications, such as in 1-to-1 face verification tasks in the domain of security and surveillance systems. Multiple methods have been proposed to protect an individual's privacy by perturbing the images to remove traces of identifiable information, such as gender or race. However, significantly less attention has been given to the problem of protecting images while maintaining optimal task utility

Aim: to learn a useful image representation which protects against private attribute inference attacks

Our contribution: proposing a principled framework, **A**dversarial **I**mage **A**nonymizer (AIA) capable of creating utility and privacy preserved image representations using adversarial learning

2. Example of Predicting Users' Private Attributes



https://imagerecognize.com

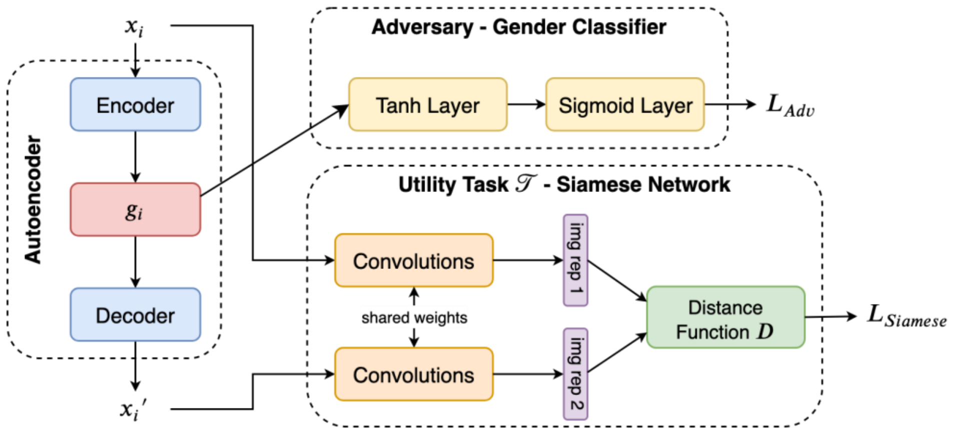
Attributes:
- Age: 22-34
- Gender: Female

Face Attributes:
- Smile
- Eyes open
- Mouth open

Missing Attributes:
- Sunglasses
- Beard
- Mustache
- Eyeglasses

Emotions:
- Happy

3. Model Architecture



AIA has three main components: **(1)** an auto-encoder for creating a general-purpose image representation, **(2)** an attribute inference attacker which predicts user's gender, and **(3)** a utility component which uses Siamese network to ensure the enhanced representation is still useful

Training: first, each model is trained separately using the following loss functions:

1. AutoEncoder:

$$L_{AE} = x'.\log x + (1 - x').\log(1 - x)$$

2. Gender Classifier:

$$L_{AE} = x'.\log x + (1 - x').\log(1 - x)$$

3. Siamese Network:

$$L_{Siamese} = (1 - y)\,\frac{1}{2}D^2 + y\,\frac{1}{2}\max(0, m - D)^2$$

Then, the whole model is fine-tuned using a combination of the three loss functions:

$$L_{Total} = L_{AE} + \alpha L_{Siamese} - \beta L_{Adv}$$

control parameters

4. Experiments and Results

We use two datasets **CelebA** and **VGG1** datasets to evaluate AIA. Both utility and privacy are considered in evaluation:

Privacy: the accuracy of gender classifier

Utility: the accuracy of Siamese network for 1-to-1 face matching task

4-1: CelebA Results

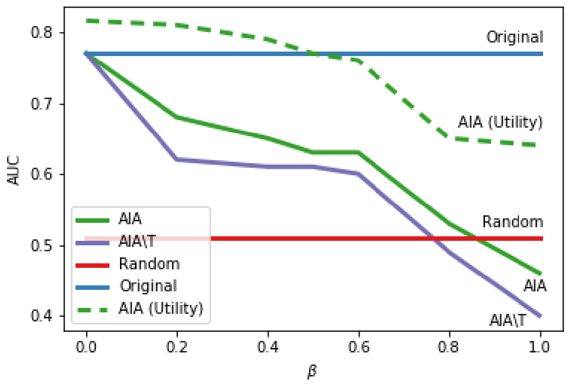
Method	Privacy (↓ better)	Utility (↑ better)
Original	%78.01	%88.87
Random	%52.12	%56.89
AIA\T	%62.53	%69.34
AIA	**%64.96**	**%78.64**

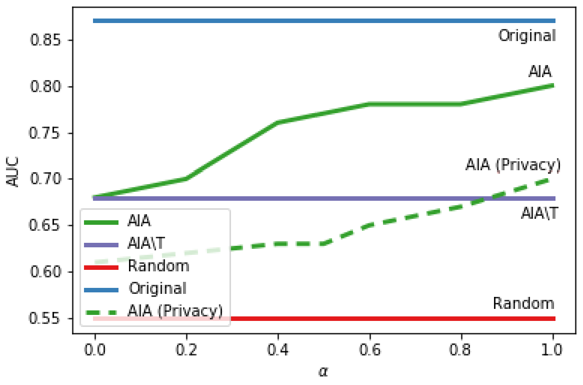
4-2: VGG1 Results

Method	Privacy (↓ better)	Utility (↑ better)
Original	%81.21	%91.31
Random	%51.34	%53.67
AIA\T	%65.67	%66.29
AIA	**%68.13**	**%77.96**

5. Parameter Analysis

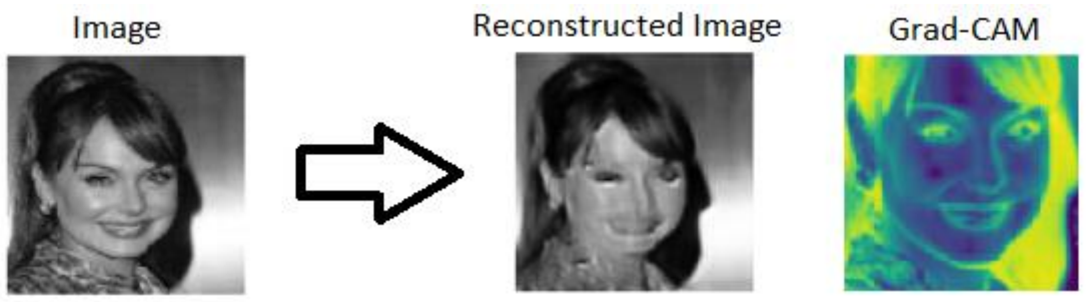
In this experiment, we test AIA using different values of α and β:





6. Visualization

In this part, we show a sample image which is created using our model's private representation:



Acknowledgement: "This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-04-00."

- Disclaimer: "The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security."