Toward Privacy and Utility Preserving Image Representation

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

Arizona State University, Tempe AZ, USA
{amosalla,ysilva,mmanceni,huanliu}@asu.edu

Abstract. 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. In this paper, we study the novel problem of creating privacy-preserving image representations with respect to a given utility task by proposing a principled framework called the Adversarial Image Anonymizer (AIA). AIA first creates an image representation using a generative model, then enhances the learned image representations using adversarial learning to preserve privacy and utility for a given task. Experiments were conducted on a publicly available data set to demonstrate the effectiveness of AIA as a privacy-preserving mechanism for face images.

Keywords: Privacy · Utility · Adversarial Learning · Generative Model

1 Introduction

Security and surveillance systems, such as those found in private industries (e.g., biometric access control systems) and public domains (e.g., face recognition systems at airports and traffic thruways), acquire images of people's faces for verification and identification tasks. The ease of collecting data on private citizens raises concerns about violating privacy-preserving contracts or expectations [5], because organizations have been known to exercise their prerogative to sell information on individuals or have been subject to malicious attacks that compromised users' privacy [12]. For instance, in 2019, a malicious cyber-attack to a US Customs and Border Protection subcontractor exposed traveler's photos [1]. Due to such threats to individual liberties and privacy, one method that has been proposed to protect an individual's private information is to anonymize it before sharing. While some recent studies have shown that face images contain user-related private information such as gender or race [4, 10, 7], research on protecting face images from adversarial attacks has been limited [2].

Two general approaches have been proposed for preserving privacy on image data. The first method, known as visual privacy, perturbs images so that a

human cannot infer a user's private attributes. A second method creates a representation from the image data [2], which then replaces the original image in image-based applications. An advantage of visual privacy is that the perturbed images can easily be used in various image-based tasks. This approach, however, does not provide the same level of privacy-preserving effectiveness as the second method [2]. Furthermore, current methods do not guarantee that the purturbed image is still useful in a specific utility task.

In this paper, we address the challenge of creating a privacy-preserving image representation while simultaneously preserving the image's utility for a given image-base task. To address this challenge, we propose the AIA (Adversarial Image Anonymizer) framework composed of three main modules: (1) a component to encode images for representation learning; (2) a component for privacy-preserving representation learning; and (3) a component to preserve the utility of the learned representations. The main contributions of this paper are (1) the study of the novel problem of joint privacy and utility preserving image representation; (2) a principled framework (AIA) that integrates adversarial learning and a generative model to create a privacy-preserving image representations which can be used with a given utility task; and (3) extensive experiments on a publicly available data set to demonstrate the effectiveness of AIA for creating both privacy preserving image representations for a specific utility task.

2 Related Work

This section briefly describes relevant and related work on the following domains: (1) image generation and representation; and (2) adversarial learning on generative models.

Image Generation and Representation. Most of the previous work on image generation focuses on using generative models for image generation. For example, Cheung et al. proposed a method for generating images using auto-encoders by studying the hidden and latent factors in image representations [3]. Mathieu et al. proposed the use of adversarial training to create an image representation that can be used to generate sharp looking images [11]. More recently, Tran et al. proposed a face recognition model based on Generative Adversarial Networks (GAN) [17]. Despite not addressing privacy preservation, these contributions showed that latent factors in image representations could provide important information about the images such that one can generate the original images using the image representations. Some of these previous results were recently leveraged in the work of Chen et al. to build a model for privacy-preserving facial expression recognition through GAN-based image representation learning. This model is capable of generating privacy-preserving images [2]. A disadvantage of this model, however, is that the generated images are human recognizable.

Adversarial Learning on Generative Models. Generative Adversarial Networks (GAN) are neural networks that use adversarial training in a game environment to train two networks, a generator and a discriminator [9]. This method can generate high-quality images and image representations. Despite being ex-

tensively used, GANs could present some problems such as stability issues and mode collapse [16]. Makhzani et al. proposed the use of adversarial auto-encoders (GAN-based probabilistic auto-encoders) to create better image representations in order to generate high-quality images [8].

Our work is distinct from previous efforts in two ways. First, we consider generating a privacy-preserving image representation instead of generating an actual image. Second, we simultaneously optimize for both privacy and a given utility task during the creation of image representations. Our model can be particularly useful preserving privacy in systems that are designed for a specific task, e.g. 1-to-1 face matching.

3 Problem Statement

Let $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ denote a set of N grey-scaled images where each image x_i is composed of a matrix $x_i \in \mathbb{R}^{1 \times n \times m}$. Let p denote a private attribute that users do not want to disclose such as gender. We address the following problem:

Problem 1. Given a set of images $\mathcal{X}$ and private attribute p, learn an anonymizer f that can learn an image representation $\mathbf{g} \in \mathbb{R}^{1 \times k}$ such that: (1) [Privacy preservation] an adversary cannot infer the targeted user's private attribute p, and (2) [Utility preservation] the image representation $\mathbf{g}$ can be effectively used in a given task $\mathcal{T}$ such as 1-to-1 face matching. The problem can be expressed as:

$$\mathbf{g_i} = f(x_i, p, \mathcal{T}) \tag{1}$$

Due to the success of auto-encoders in learning image representations [8], we use auto-encoders with adversarial training to create both utility and privacy preserving image representations in this paper.

4 Proposed Method

The main components of the Adversarial Image Anonymizer (AIA) framework appear in Figure 1. The input to the system is a grey-scale image, while the output is a privacy-preserved vector $\mathbf{g} \in \mathbb{R}^{1 \times k}$. The model has 3 main components: (1) an auto-encoder composed of an encoder that learns the input image representation and a decoder that can reconstruct the input image given its representation, (2) an adversary which tries to infer the user's private attribute using the image representation, and (3) a task $\mathcal{T}$ that is used to preserve the utility of the image representation. In this framework, we first train each component individually and, then, use an adversarial loss function to enhance the overall model. We present the details of our proposed method next.

4.1 Learning Image Representations

The goal of this component is to generate an image representation that can be used in different tasks. The image representations in our model are created using

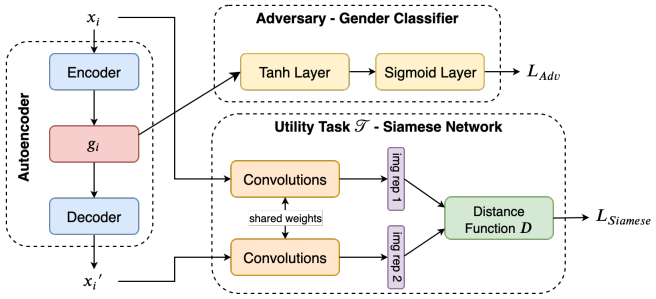


Fig. 1: Architecture of the Adversarial Image Anonymizer (AIA).

an auto-encoder. This is a generative model that is trained in an unsupervised way to learn latent representations of the input images. A key feature of auto-encoders is their dimensionality reduction ability which was an important reason to integrate it instead of using GANs. Another reason of using an auto-encoder instead of a GAN is to prevent some of the issues in GAN-based models such as stability problems and time-consuming training [14]. The auto-encoder consists of the following components:

Encoder. The encoder learns a latent representation of the input image and aims at reducing its dimensionality. x_i is a grey-scaled image with dimensions $n \times m$. To create a representation of the image we use Convolutional Neural Networks (CNN) to learn filters that can identify the important parts of an input image. In our model, we first use a convolution layer $Conv$ to create feature maps from the input image. Then, we apply an activation function F on the gathered features and use a pooling layer P. The pooling layer helps to select the parts of the features with strong correlation to the input image. We refer to the output of these three layers as a block L_j^{enc} where j represents the j^{th} encoder block:

$$L_j^{enc} = P(F(Conv(x))) \tag{2}$$

We use a stack of these blocks to create an image representation. Observe that in order to convert the output of the final pooling layer to a vector $\mathbf{g}$, we use a flattening layer that converts a matrix into a vector:

$$\mathbf{g} = Flatten(L_J^{enc}) \tag{3}$$

where J is the index of the final encoder's layer.

Decoder. The decoder tries to reconstruct the input image using the previously generated representation. The process of decoding the image representation $\mathbf{g}$ is similar to the one in the encoding phase, but in reverse. First, we use a convolution layer to create feature maps, then we apply an activation function to the features. As for the final layer, instead of using a pooling layer which generates a smaller matrix, we use an up-sampling layer U to increase the size of the reconstructed image:

$$L_l^{dec} = U(F(Conv(\mathbf{g}))) \tag{4}$$

where L_l^{dec} indicates the l^{th} decoder block. The final output of the decoder is the reconstructed image $\mathbf{x}'$ which is then used to train the auto-encoder.

After creating the auto-encoder, we train it using the input image x_i and the reconstructed image x_i'. This will result in learning a representation vector $\mathbf{g}$ which is expected to capture useful information of the input image.

4.2 Adversarial Training

Creating an image representation using an auto-encoder alone could result in privacy issues [2]. For example, a well-trained gender classifier could predict the gender of a person using the corresponding image representation. To prevent this issue, we integrate adversarial training to create privacy-preserving image representations. In this component, we use a powerful adversary, i.e. gender classifier, to further improve the learned representation of the auto-encoder. Because our goal is to create a privacy-preserving representation for a given task $\mathcal{T}$, we use the loss value of this task as a penalty to generate private learned representation.

4.3 Optimization Algorithm

The training process in the proposed model consists of two parts. In the first part, we train each component (auto-encoder, adversary, and the given task $\mathcal{T}$) separately:

- *Auto-encoder*: For the encoder component, we use stacked CNN blocks, each containing a 2D convolution layer with leaky ReLU as the activation function; and an average pooling which operates on the output of the activation function. The decoder has also two stacked CNN blocks. Each block has a convolution layer with a leaky ReLU activation function. The output of the activation function goes through an up-sampling layer to create an output image with similar size to the input image. We train the auto-encoder using the Binary Cross Entropy loss function between the input image x_i and the reconstructed image x_i'. This loss function calculates how well the auto-encoder has predicted the image:

$$L_{AE} = x' \cdot \log x + (1 - x') \cdot \log(1 - x) \tag{5}$$

- *Adversary*: In our model, we use a high-quality gender classifier as the adversary. The adversary acquires an image representation $\mathbf{g}$ as input and predicts whether the image corresponds to a female or male face. We use a three-layer neural network to output gender probablity $\mathbf{o}$:

$$\mathbf{o} = sigmoid(\mathbf{W_A^{(2)}}(\tanh(\mathbf{W_A^{(1)}}\mathbf{g} + \mathbf{b_A^{(1)}})) + \mathbf{b_A^{(2)}}) \tag{6}$$

  where $\mathbf{W_A^{(.)}}$, $\mathbf{b_A^{(.)}}$, are learnable weights. This classifier is also trained using the same Binary Cross Entropy (BCE) loss L_{Adv}.

– *Task* $\mathcal{T}$: In this paper we consider a 1-to-1 face matching task, which verifies if two input images are of the same individual. We use the well-known Siamese network for this task which acquires two images x_i and x'_i as input and returns their representations. The distance of the two representations is then calculated based on the similarity between x_i and x'_i. We train this model using the following loss function:

$$L_{Siamese} = (1-y)\frac{1}{2}D^2 + y\frac{1}{2}max(0, m-D)^2 \tag{7}$$

where D as the Euclidean distance and m a constant margin.

Adversarial Training: After training each component separately, we use the following loss function to enhance the autoencoder for generating privacy and utility preserving representations based on the feedback from the utility and the adversary components:

$$L_{Total} = L_{AE} + \alpha L_{Siamese} - \beta L_{Adv} \tag{8}$$

where α and β indicate the contribution of each loss value. In our model, we use a powerful attacker to ensure privacy even from other unseen attackers. In our experiments we show that our model can preserve privacy of user's private attributes from different attackers.

5 Experiments

We performed multiple experiments to evaluate the performance of the proposed model. We aim to answer the following questions: (**Q1**) how well does our method protect users' private attribute, i.e., gender?; (**Q2**) how well does our method preserve the utility of an image with respect to a given task $\mathcal{T}$?; and (**Q3**) what is the relation between privacy and utility? To answer **Q1**, we use an adversary to test if it can detect gender based on the perturbed representations. For **Q2**, we study the performance of the utility task before and after perturbing the learned image representations. Finally, to answer **Q3**, we study how the effectiveness of preserving users' privacy impacts the utility of the learned representations.

5.1 Data

In this study, we use two different publicly-available datasets, CelebA [6] and VGG face datasets [13]. CelebA consists of over 200K celebrity images with various metadata [6]. VGG consists of 2,622 identities where each identity has different images [13]. In both datasets, we use gender as the private attribute.

5.2 AIA Implementation Details

The adversary is a gender classifier which has 6 convolution layers with 32, 64, 64, 128, 128, and 256 channels, respectively. After the convolution layers, we use

a one-layer neural classifier with Softmax on the output layer. The auto-encoder is composed of an encoder and a decoder. The encoder has two convolution layers with 8 and 12 channels, while the decoder contains three convolution layers with $256, 128$, and 1 channels. The final convolution layer in the decoder converts the 128 features into 1 feature to generate the image corresponding to the associated input image. Finally, for the utility task $\mathcal{T}$, we use a Siamese network. This network has three convolution layers and three fully connected layers after flattening the output of the convolution layers. The three convolution layers have $4, 8$ and 8 channels, respectively. The fully connected network has an input layer, a hidden layer with 500 neurons, and an output layer with 128 neurons. The output of this network is an embedded image representation which is then used to calculate the distance between two input images.

5.3 Experimental Design

We use the following baseline methods for comparision:

- **Original:** this method does not use any anonymization and only outputs the learned image representation from the auto-encoder's output.
- **Random:** this method randomly changes 50% of each learned image representation to make it private. Because each value in the image representation is within $[0.0, 1.0]$, we randomly select 50% of the numbers and change each to a random number sampled from $[0.0, 1.0]$.
- **AIA\T:** this is a modified version of our proposed method that does not use the adversarial training for preserving the utility of the learned image representation.

5.4 Experimental Results

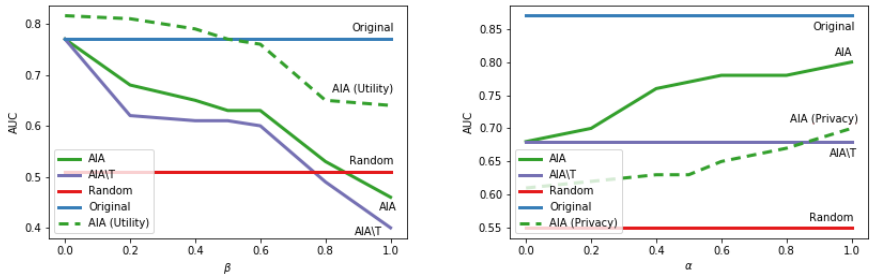
In this subsection, we evaluate AIA's performance using a gender classifier (adversary) and a similarity task (utility).

Privacy (Q1). Table 1 compares the accuracy of the different methods for gender-detection and the utility task. In the Gender Privacy column, lower accuracy indicates that the gender classifier for that method was less effective at predicting a user's gender based on the image representations. We observe that AIA is better at protecting privacy than the original approach. While AIA cannot preserve users' privacy as well as AIA\T and the random methods, it preserves utility significantly better than these methods. The random method, despite being more effective at privacy preservation, generated significantly lower utility, implying that it generated useless representations.

Utility (Q2). As illustrated in Table 1, our method performs better than the Random and AIA\T methods for preserving the utility of the image representations. While the Original method has the highest accuracy level, it provides the worst privacy-preserving guarantees. This highlights the need for changing image representations in order to preserve users' privacy. Changing image representations randomly will perverse users' privacy but it will also greatly decrease their utility. While AIA\T is relatively effective preserving privacy, it lacks the utility

Table 1: Accuracy of the private-attribute classifier and the utility task. In these results, $\alpha = \beta = 0.5$.

	CelebA Dataset		VGG Dataset	
Method	Privacy ($\downarrow$ *better*)	Utility ($\uparrow$ *better*)	Privacy ($\downarrow$ *better*)	Utility ($\uparrow$ *better*)
Original	%78.01	%88.87	%81.21	%91.31
Random	%52.12	%56.89	%51.34	%53.67
AIA\T	%62.53	%69.34	%65.67	%66.29
AIA	**%64.96**	**%78.64**	**%68.13**	**%77.96**



(a) Gender Inference ($\downarrow$ *better*) **(b)** Utility ($\uparrow$ *better*)

Fig. 2: AUC scores for private-attribute inference and utility tasks for different values of α and β.

benefits that AIA provides. This is because AIA\T does not have any component that forces the auto-encoder to preserve the utility of image representations.

Privacy Utility Trade-off (Q3). AIA has two main parameters, β controls the contribution of the gender classifier, while α controls the contribution of the utility task. In Figure 2 we show the performance of our model and the baselines for different values of α and β using the CelebA dataset. For each parameter, we hold one of them as 0.5 and vary the other one from 0 to 1. As shown in this figure, achieving higher levels of utility with AIA results in lower levels of privacy assurance and vice versa. Figure 2.a shows AIA's utility level using a dashed line for different values of β while $\alpha = 0.5$. We can observe that using a β value larger than α could result in substantial utility loss. Figure 2.b shows the inverse. The dashed line in this figure shows the gender inference AUC values (where higher values correspond to lower privacy) for different values of α while $\beta = 0.5$. These results indicate that reaching higher levels of utility can result in significant privacy loss. The problem of utility and privacy preservation is consequently a multi-objective, trade-off optimization problem where each objective antagonizes the other. From Figure 2, we conclude that using similar values for α and β provides a reasonable balance between both privacy and utility. In general, a judicious choice for the two tuning parameters depends on the application domain. If privacy is more important than utility, β should be higher than α, and vice versa.

Fig. 3: Visualization results. We used the decoder to reconstruct the image and used Grad-CAM to visualize the learned features from the gender classifier.

5.5 Visualization

To gain additional insights, we visualized and analyzed the learned image representations using a low value of $\alpha = 0.2$. We used Grad-CAM [15], a well-known CNN visualization technique. Given a label and an image, this method generates an activation map that can be used to identify the areas of the image that are most relevant to the label. Figure 3 shows both the reconstructed and Grad-CAM images for a sample photo and the result of the gender classifier. The reconstructed image in Figure 3 shows the outcome of the anonymization process. We used the decoder component from the auto-encoder to reconstruct the image using its representation. The Grad-CAM image shows the important parts of the image that resulted in the classifier predicting a female label. We can observe that the classifier focuses on the hair length and the eyes, as well as the location of the cheekbones and the shape of the chin. The adversarial training in AIA influences the auto-encoder to hide these pieces of information that result in more accurate gender labels.

6 Conclusions and Future Work

Protecting the privacy of citizens has been a widespread concern in an era where the intentional or unintentional propagation of private information has been an unfortunate byproduct of machine learning. We recognized that limited attention has been given to mechanisms that simultaneously preserve an individual's privacy, while preserving the intended utility of a machine or deep learning model. Thus, we proposed the AIA framework that uses an auto-encoder to learn image representations and then enhances these representations using adversarial training. Initial results showed an interesting trade-off between utility and privacy which results in outcomes that offer better privacy while having only a small impact on utility. The performance results showed that AIA performed well overall in comparison to the baseline methods. An immediate future work direction is the extension of the proposed framework to consider multiple private attributes. This work also has salient applications in the area of bias in deep learning models, a direction that is currently being pursued by the authors.

References

1. Malicious cyber-attack. https://www.theguardian.com/world/2019/jun/10/malicious-cyber-attack-exposes-travelers-photos-says-us-customs-agency
2. Chen, J., Konrad, J., Ishwar, P.: Vgan-based image representation learning for privacy-preserving facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018)
3. Cheung, B., Livezey, J.A., Bansal, A.K., Olshausen, B.A.: Discovering hidden factors of variation in deep networks. arXiv preprint arXiv:1412.6583 (2014)
4. Guo, S., Xiang, T., Li, X.: Towards efficient privacy-preserving face recognition in the cloud. Signal Processing **164**, 320–328 (2019)
5. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. pp. 193–204 (2011)
6. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (12 2015)
7. Ma, Z., Liu, Y., Liu, X., Ma, J., Ren, K.: Lightweight privacy-preserving ensemble classification for face recognition. IEEE Internet of Things Journal **6**(3) (2019)
8. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
9. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2017)
10. Mao, Y., Yi, S., Li, Q., Feng, J., Xu, F., Zhong, S.: A privacy-preserving deep learning approach for face recognition with edge computing. In: Proc. USENIX Workshop Hot Topics Edge Comput.(HotEdge). pp. 1–6 (2018)
11. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: Advances in neural information processing systems. pp. 5040–5048 (2016)
12. Mosallanezhad, A., Beigi, G., Liu, H.: Deep reinforcement learning-based text anonymization against private-attribute inference. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2360–2369 (2019)
13. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
14. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (2017)
16. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. In: Advances in Neural Information Processing Systems. pp. 3308–3318 (2017)
17. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1415–1424 (2017)