An Autoencoder-Based Image Anonymization Scheme for Privacy Enhanced Deep Learning

David Rodriguez(✉) and Ram Krishnan(✉)

Electrical and Computer Engineering Department,
University of Texas at San Antonio, San Antonio, TX 78249, USA
david.rodriguez3@my.utsa.edu, ram.krishnan@utsa.edu

Abstract. The development of deep learning (DL) technology is dependent on the availability of large-scale image datasets to train deep neural networks (DNNs) for image classification. However, many raw image datasets contain sensitive identity feature information that prohibit entities from disclosing data due to privacy regulations. For example, an image dataset may include age or gender information that could be used to identify an individual. Furthermore, medical images may include additional disease information that could lead to patient re-identification. To address this problem, we propose an image transformation scheme using a convolutional autoencoder and multi-output classification model for privacy enhanced deep learning. The proposed scheme obfuscates image visual information while retaining useful attribute features that are required for model utility. Additionally, the proposed method enhances privacy by generating encoded images that exclude sensitive identity feature information. First, we train a multi-output convolutional neural network (CNN) to classify identity features and image attributes. Second, we use the pre-trained multi-output classifier for regularization in training a standard convolutional autoencoder to generate obfuscated versions of the original images that exclude identity feature information and preserve attribute features that are useful for classification. Our results on CelebA and Cifar-100 datasets illustrate that the proposed method successfully degrades classification accuracy of sensitive image data while maintaining model utility for non-sensitive data features.

Keywords: deep neural networks · convolutional neural networks · convolutional autoencoder · privacy · utility

1 Introduction

Cloud-based services have become an extremely popular option for data owners to outsource large computationally expensive deep learning tasks due to flexibility and cost saving [1]. Cloud providers offer a full range of services including

Research supported in part by NSF CREST Grant HRD-1736209 (RK) and NSF CAREER Grant CNS-1553696 (RK).

Published by Springer Nature Switzerland AG 2023
V. Atluri and A. L. Ferrara (Eds.): DBSec 2023, LNCS 13942, pp. 302–316, 2023.
https://doi.org/10.1007/978-3-031-37586-6_18

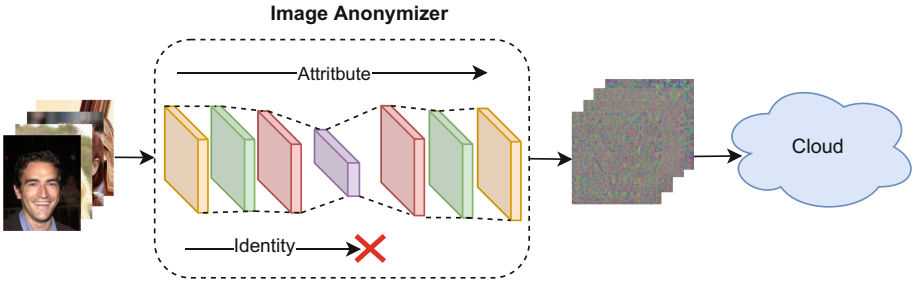


Fig. 1. Image anonymization overview.

storage, servers, virtual desktops, full applications and development platforms. Many organizations have access to large amounts of data but very limited computational resources and storage which prevent them from performing feature extraction tasks locally. Therefore, a large amount of data owners have opted for cloud services to allocate resources as needed for the given task at hand [2]. Typically, an entity will send its raw data such as images to a machine learning as a service (MLaaS) provider for the purpose of developing a DL algorithm directly using the raw images. However, image data may contain sensitive information that the data owner wishes keep private while preserving model utility.

There are several privacy risks that accompany the disclosure of raw image data containing sensitive information. Raw images consist of features that are useful for a specific classification task such as classifying facial attributes which may include if an individual is smiling or wearing glasses, etc. On the other hand, raw images may also include additional feature information that is not useful for the specific classification task such as gender or age which could be used to reveal the identity of an individual. For example, previous work [3] has shown that person identification can be accomplished with as little as a human ear, so given a dataset of raw human faces an attacker could gain access to a victims personal identity by simply possessing an image of the human ear. Furthermore, [4] demonstrated that DNNs could be trained to recover patient identity from chest X-ray data by identifying if two frontal chest X-ray images belong to the same individual even if they were taken years apart. Attackers could potentially leak patient information or analyze the identified images to gain access to additional sensitive information. Consequently, we aim to increase the privacy and security of sensitive data by transforming the original images such that identity information or sensitive attributes are excluded from encoded versions while maintaining classification accuracy.

Several visual information protection methods have been proposed to preserve privacy of image data such as pixelation, blurring and P3 [5]. Visual information protection methods encrypt data such that visible feature information of an image is concealed while making sure that the transformed version remains useful for classification [6–10]. However, these methods do not exclude identity feature information from the encoded version of the original image. Our

proposed method not only transforms the image such that it is longer recognizable to humans but it also excludes specific sensitive feature information from the encoded data.

One of the major challenges in developing algorithms to anonymize sensitive image data is known as the trade-off between privacy and utility [11–13]. The goal is to anonymize image data such that an attacker could not learn any sensitive identity feature information while authorized users could perform useful statistics. Eliminating the entire dataset provides perfect privacy but this is not useful. On the other hand, publishing raw unaltered data is statistically useful but may be detrimental to the privacy of sensitive data. We propose to publish transformed versions of the original data that maintain model utility by retaining useful attribute features that are beneficial for classification while increasing privacy by removing sensitive identity features from the data. An overview of the image anonymization process is depicted in Fig. 1.

In this paper, we propose an image data anonymization scheme using a DL approach to increase data privacy while maintaining model utility. Specifically, we train a multi-output DL model to increase classification accuracy of identity feature information and image attributes. Then we train an anonymization network consisting of a convolutional autoencoder attached to the input of a pre-trained multi-output classifier to generate obfuscated versions of the original images. The encoded images exclude identity feature information and preserve attribute features that are useful for classification. In our results, we demonstrate that our image anonymization method increases data privacy while maintaining model utility using CelebA [14] and Cifar-100 [15] datasets.

In summary our contributions are as follows:

- We develop an autoencoder-based image anonymization method for privacy enhanced deep learning.
- We increase privacy of image identity feature information while maintaining model utility.

The remainder of this paper is organized as follows. In Sect. 2, we review related works of privacy protection methods in machine learning. In Sect. 3, the proposed image data anonymization method formulation and loss function are discussed. In Sect. 4, the dataset, network architecture and training procedure are described. In Sect. 5, we evaluate our image anonymization method by assessing the trade-off between privacy-utility and robustness to attacks. Finally, we discuss and conclude our paper in Sects. 6 and 7, respectively.

2 Related Works

Privacy protection in machine learning typically address the privacy of a model's input, the privacy of the model, or the privacy of the model's output. Several privacy preserving techniques have been proposed in the literature, some of which utilize secure multi-party computation, homomorphic encryption, federated learning, visual image protection and learnable image encryption. Secure

multi-party computation is a set of cryptographic protocols that allow multiple parties to evaluate a function to perform computation over each parties private data such that only the result of the computation is released among participants while all other information is kept private [16]. Secure multi-party computation methods have been applied in machine learning among multiple parties by computing model parameters using gradient descent optimization without revealing any information beyond the computed outcome [17–20]. Our proposed method does not require multiple parties to perform gradient descent individually but instead allows all users to anonymize private data individually and share the transformed images.

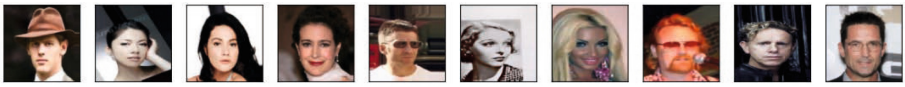
Homomorphic encryption is a type of encryption that allows multiple parties to perform computations on its encrypted data without having access to the original data. It provides strong privacy but is computationally expensive requiring significant overhead to train machine learning models [21–26]. Our proposed encoding scheme does not require expensive encryption operations or specialized primitives for the training process.

Federated learning allows multiple parties to train a machine learning model without sharing data [27–29]. For example, in centralized federated learning a central server sends a model to multiple parties to train locally using their own data, then each participant sends it's own model update back to the central server to update the global model which is again sent to each party to obtain the optimal model without access to the local data by iterating through this process [30]. Essentially, federated learning builds protection into the model. Nevertheless, federated learning suffers from the privacy-utility trade-off [31]. Our proposed encoding scheme enables entities to share encoded data which do not reveal sensitive feature information and maintain model accuracy.

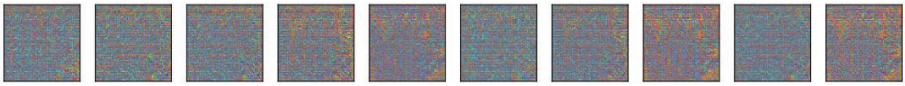
Visual image protection methods transform original images to unrecognizable versions of the image while maintaining the ability to perform useful statistics. A few examples of visual image protection methods are pixelation, blurring, P3 [5], InstaHide [32] and NueraCrypt [33] which aim at preserving privacy and utility—a model trained on an encoded dataset should be approximately as accurate as a model trained on the original dataset [34,35]. InstaHide mixes multiple images together with a linear pixel blend and randomly flips the pixel signs. NeuraCrypt encodes data instances through a neural network with random weights and adds position embeddings to keep track of image structure then shuffles the modified output in blocks of pixels. Our proposed encoding scheme removes the unnecessary complexity of NeuraCrypt's positional embeddings and permutations while maintaining privacy and utility.

Learnable image encryption methods encrypt images such that the encoded versions are useful for classification [6–10]. However, in some cases network adjustments are required to process learnable image encryptions such as block-wise adaptation [6]. Our method does not require any special modifications to the network and excludes identity information from the obfuscated samples while maintaining usability for classification. Our work is most closely related to [36] which removes user identity information from mobile sensor data while

training a network to classify user activities. In our work, we develop a deep learning classification model using image attribute features while removing image identity features.



(a) Original images



(b) Anonymized Images

Fig. 2. Examples of anonymized images from Celeba dataset using the proposed scheme.The bottom row are the corresponding anonymized images of the top row.

3 Image Data Anonymization

Our goal is to transform image data such that all visual feature information is unrecognizable to humans as depicted in Fig. 2 but remains useful for classification. Additionally, we aim to remove identity feature information from the transformed images while preserving attribute features. Our method enables entities to share encoded versions of the original data that exclude sensitive feature information while maintaining model utility. We consider identity features that can be collected from an image as sensitive data. On the other hand, we consider attribute features in a given image as non-sensitive data. We wish to preserve attribute features in the transformed images while removing sensitive identity features. Additionally, we aim to maintain similar attribute classification performance on the transformed images as the original images.

3.1 Formulation

Our goal is to anonymize image data by removing identity feature information while preserving attribute features for model utility. Let $\mathcal{X}$ be the set of all possible 8-bit images in the data domain, $X_a \subseteq \mathcal{X}$ is the data owner's private subset and Y_a is the corresponding label set. Given the private image dataset $\{x_{a_i}\}_{i=1}^{N}$ where $x_{a_i} \in X_a$, the data owner encodes all images using a private image anonymization function $z_a = E(x_a)$ and shares the encoded set $\{z_{a_i}\}_{i=1}^{N}$ and corresponding attribute labels $\{y_{a_i}\}_{i=1}^{N}$ where $y_{a_i} \in Y_a$ with a third party cloud service provider without revealing sensitive identity feature information. The proposed image anonymization function is similar to [36] but instead of anonymizing mobile sensor data we develop our encoding function to anonymize image data. The proposed method consists of a multi-output classifier to distinguish between attribute and identity features. In addition, the network consists of

an autoencoder to anonymize images. The objective of training the network is to obtain the image anonymizer E^* which transforms raw images into anonymized images.

In the multi-output classification model training phase, a resnet50 model is trained to classify identity features and attribute features using the same input images $\{x_{a_i}\}_{i=1}^{N}$ with their respective class labels. Our objective function for the multi-output classification model has two loss terms: identity loss for classifying identity features; and attribute loss for classifying attribute features. We aim to classify identity features and attribute features of a given image with high classification accuracy for the multi-output network. After training, the multi-output classification model is used to develop the anonymization network for the purpose of transforming original images into anonymized images. Our anonymization objective function also contains two loss terms: identity suppression loss for removing identity features; and attribute preservation loss for preserving attribute features. We aim to degrade the identity feature classification accuracy while preserving the attribute feature classification accuracy.

3.2 Multi-Output Classification Loss Function

The multi-output network is trained using a multi-objective loss function for image classification which consists of an identity and attribute loss function. The identity loss is used to minimize the error between the true identity and identity classifier's predicted identity. The attribute loss is used to minimize the error between the true attribute and the attribute classifier's predicted attribute. The aim is to classify identity features and attribute features with high classification accuracy.

Identity Loss. The identity loss function L_i uses cross-entropy to measure the performance of identity classifier $I(\cdot)$ which is trained to classify image identity features.

$$L_i(I, X, Y) = -\frac{1}{N}\sum_{i=1}^{N} Y_i log(I(x_i)) \tag{1}$$

where x_i is the i^{th} image and Y_i is the corresponding ground truth identity label. $I(x_i)$ is the identity classifier's predicted output for the i^{th} image.

Attribute Loss. The attribute loss function L_a uses categorical cross-entropy to measure the performance of the attribute classifier $A(\cdot)$ which is trained to classify image attribute features.

$$L_a(A, T, X) = -\frac{1}{N}\sum_{i=1}^{N} T_i log(A(x_i)) \tag{2}$$

where T_i is the ground truth N-dimensional one hot encoded vector attribute label for the i^{th} image and $A(x_i)$ is the attribute classification function predicted

softmax output which is an N-dimensional vector consisting of the attribute label probabilities for the i^{th} image.

3.3 Multi-output Classification Objective

Our multi-output classification objective is:

$$L(I, A) = L_a(A, T, X) + L_i(I, X, Y) \tag{3}$$

we aim to solve:

$$I^*, A^* = \underset{I,A}{\operatorname{argmin}} L(I, A) \tag{4}$$

3.4 Image Anonymization Loss Function

The image anonymization network is trained using a multi-objective loss function for image classification which consists of an identity suppression and attribute preservation loss function. The aim is to remove identity features while preserving attribute features that are useful for classification.

Identity Suppression Loss. The identity suppression loss function L_s uses mean squared error to remove identity feature information from sensitive data.

$$L_s(\xi, I^*, E) = -\frac{1}{N}\sum_{i=1}^{N}(\xi - I^*(E(x_i)))^2 \tag{5}$$

where E is the anonymization function and I^* is a pre-trained identity classification function. ξ is a positive value between 0–1. We maximize the difference between the predicted identity label and the true identity label by minimizing the mean squared error between ξ and the predicted identity label given the i^{th} encoded image. The anonymization network is penalized if the transformed image contains identity feature information.

Attribute Preservation Loss. The attribute preservation loss function L_p uses categorical cross-entropy to preserve attribute feature information.

$$L_p(A^*, E) = -\frac{1}{N}\sum_{i=1}^{N} T_i log(A^*(E(x_i))) \tag{6}$$

where A^* is the pre-trained attribute classification function. The aim is to minimize the preservation loss given the i^{th} encoded image. We minimize the difference between the predicted attribute label and the true attribute label by minimizing the crossentropy between T_i and the predicted attribute label.

3.5 Image Anonymization Objective

Our image anonymization objective is:

$$L(\xi, I^*, A^*, E) = \lambda_1 L_p(A^*, E) + \lambda_2 L_s(\xi, I^*, E) \tag{7}$$

where the regularization parameters λ_1 and λ_2 are positive values that regulate the trade-off between privacy and utility.

we aim to solve:

$$E^* = \underset{E}{\operatorname{argmin}}\, L(\xi, I^*, A^*, E) \tag{8}$$

Our anonymization function E^* generates encoded images that retain useful attribute features by penalizing the autoencoder network using crossentropy if the output does not contain attribute features. In addition, the autoencoder network is penalized using mean squared error if the output contains identity features. Thus our objective is used to preserve attribute features by applying L_p while removing identity features by applying L_s.

4 Methods

4.1 Dataset

In this work, we use the publicly available CelebA [14] and Cifar-100 [15] image datasets to develop anonymization networks. The CelebA dataset is a large-scale face attribute dataset that consists of approximately 200K celebrity face images. It includes gender and 40 attributes per image with a variety of poses and backgrounds. However, in our experiments we select images of 4 mutually exclusive attribute labels consisting of pale skin, smiling, eye glasses and wearing hat. Increasing the number of attributes significantly reduces the amount images per class. Consequently, we include 10K images per attribute label. Our goal is to train the anonymization network to generate encoded images that include attribute label features while removing gender label features. The Cifar-100 dataset consists of 60,000 32×32 color images. It consists of 100 classes containing 600 images each which are referred to as the fine label set. It is also available with 20 superclasses containing 3,000 images each which are referred to as the coarse label set. We consider the fine label set to be private. Thus we aim to remove image features associated with the fine label set. Our goal is to train the anonymization network to generate encoded images that include coarse label feature while removing fine label features.

4.2 Anonymization Network Architecture

The anonymization network architecture depicted in Fig. 3 consists of two parts: a multi-output Resnet50 for image classification and a standard convolutional autoencoder (CAE) for image transformation. Resnets are large state-of-the-art

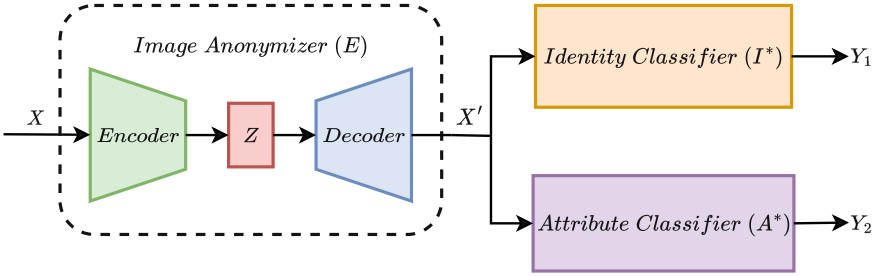


Fig. 3. Proposed anonymization model architecture.

DL architectures that consist of several blocks of residual modules and skip connections [37]. The multi-output architecture consists of one resnet50 feature extraction network with two separate classifiers at the output. The CAE encoder network consists of three convolution layers with 32, 64 and 128 filters, respectively. The kernel size is 3×3 with a stride of 2 and a latent space of 128. Each convolution layer consists of a leaky relu activation function with alpha 0.2 followed by a batch normalization layer. The decoder network consists of three transposed convolution layers with 128, 64 and 32 filters, respectively. The kernel size is 3×3 with a stride of 2 and output size of $224 \times 224 \times 3$. Each transposed convolution layer consists of a leaky relu activation function with alpha 0.2 followed by a batch normalization layer.

4.3 Training Procedure

Our training procedure consists of a feature extraction phase for classification and an identity removal phase for anonymization. First, in the feature extraction phase we train a multi-output resnet50 model from randomly initialized parameters for two different classification tasks given the same images. We train one classifer to predict the gender identity for a given image using binary crossentropy loss function for the CelebA dataset. In our Cifar-100 experiments we train the identity classifier using the fine label set which includes 100 classes. Simultaneously, we train another classifier to predict the attribute of the same image using categorical crossentropy loss function. In our Cifar-100 experiments we train the attribute classifier using the coarse label set which includes 20 classes. The coarse label set is the superclass of the fine label set, e.g., the fish label is the superclass of aquarium fish, flatfish, ray, shark, trout. We wish to classify fine label features and coarse label features for a given image set.

Second, in the identity removal phase we randomly initialize the CAE parameters and attach its output to the previously trained multi-output resnet50 classification model input. We freeze the resnet50 classifier model parameters to ensure that the weights do not change during CAE training for the identity removal phase. During training we aim to learn a CAE that retains useful attribute feature information to reconstruct an unrecognizable version of the

Table 1. Image Classification Accuracy of identity and attribute classifier for CelebA and Cifar-100 datasets

Encryption	Identity Acc (%)		Attribute Acc (%)	
	CelebA	Cifar-100	CelebA	Cifar-100
Plain Images	95.85	82.96	87.02	88.13
Proposed Scheme	50.33	20.71	85.96	83.45

original image for classification while removing the identity feature information. To remove identity feature information we optimize the identity classifier with a modified version of the mean absolute error loss function. To assure that the anonymized images retain attribute feature information we continue training the attribute classifier with the categorical crossentropy loss function.

Both networks were trained using the adam optimizer with a batch size of 128. Check points were used to save the model with the highest validation accuracy during the training procedure. All images were resized to 224×224 and normalized between 0 and 1. The dataset was randomly shuffled and split to generate the train, test and validation set. Minor data augmentation was applied during training using keras image data generator which include zoom range 0.2 and horizontal flip. All training was completed using a tesla v100 graphical processing unit.

5 Evaluation

5.1 Evaluating the Privacy/Utility Trade-Off

We train the anonymization network using the proposed method and examine the trade-off between privacy and utility, i.e. we measure the change in identity and attribute classification accuracy. First, we transform the original images using our anonymization method. Second, we compare the identity and attribute classification accuracy of original images and the transformed images.

The identity classification accuracy of anonymized images significantly decreased compared to original images. Additionally, the attribute classification accuracy of anonymized images is similar to original images. To quantify the trade-off between privacy and utility we measure the reduction in identity and attribute classification accuracy for the anonymized dataset compared to original dataset. In our experiments, we demonstrate that the proposed image anonymization method enables us to maintain high image attribute classification accuracy of 85.96% & 83.45% for CelebA and Cifar-100 datasets, respectively, which is similar to original images. It also enables us to reduce image identity classification accuracy from 95.85% & 82.96% to 50.33% & 20.71% for CelebA and Cifar-100 datasets, respectively, as shown in Table 1.

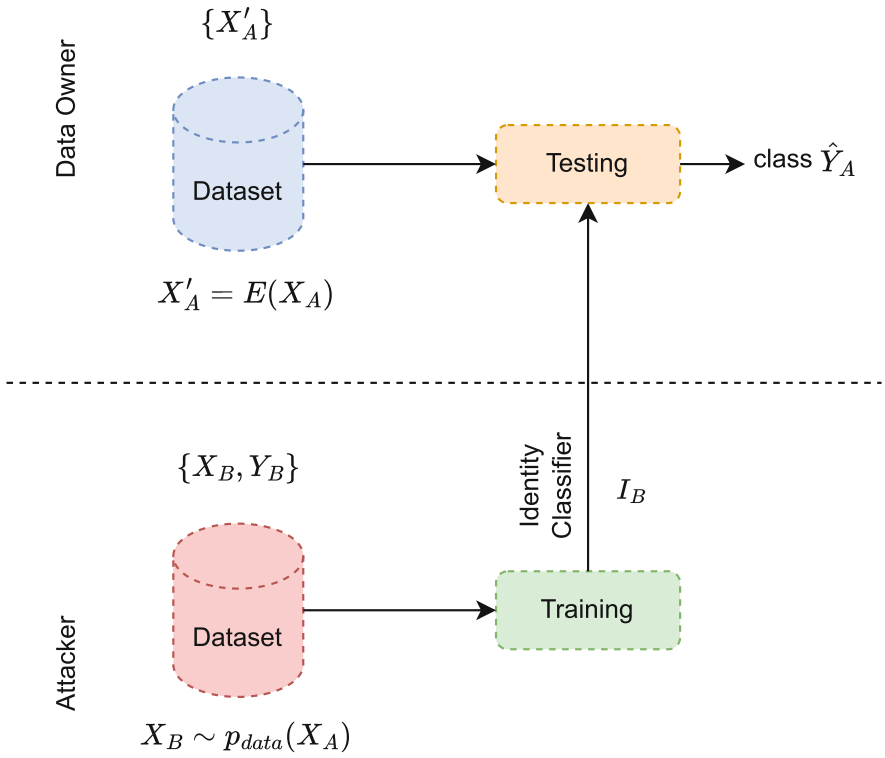


Fig. 4. Classifier transfer attack diagram. Where X'_A is the data owner's encoded dataset and X_B, Y_B are the attackers raw image dataset and identity labels which follows the probability distribution of the data owner's original dataset. I_B is the attacker's identity classifier. The attacker trains I_B with X_B, Y_B and uses the classifier to predict the identity label of the data owner's encoded dataset.

5.2 Evaluating Robustness to Attacks

Classifier Transfer Attack. We evaluate the robustness of our image anonymization approach against attacks that aim to learn an identity feature classifier and transfer it onto the data owners encoded set for classification. We conduct experiments on CelebA and Cifar-100 datasets using gender and coarse labels, respectively. We assume that the attacker is able to construct a dataset that follows a similar probability distribution as the data owner's original dataset and corresponding labels. First, the attacker trains his own identity classifier using the constructed dataset to achieve high classification accuracy. Then he attempts to classify the data owners encoded set using his pre-trained identity classifier. An overview of the classifier transfer attack is depicted in Fig. 4.

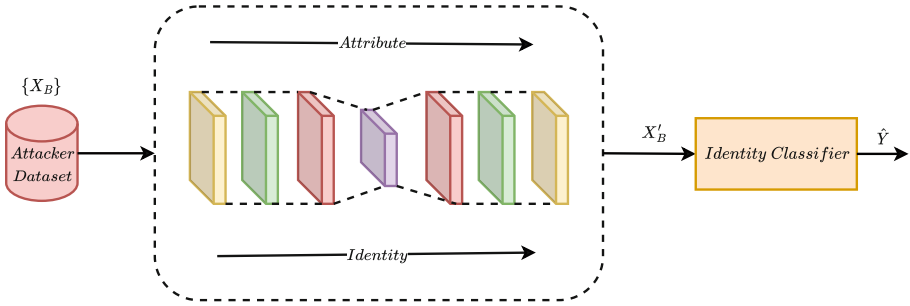


Fig. 5. Encoding transfer attack diagram. Where X_B is the attacker's dataset which follows the probability distribution of the data owner's original dataset. X'_B is the attacker's encoded dataset which consists of attribute features and identity features. The data owners identity classifier is used to predict the identity label of the attacker's encoded dataset to verify if X'_B captures the data owner's identity features.

We evaluate the performance of the attacker's pre-trained identity classifier using the data owner's encoded set. The goal of the attack is to classify identity features given the data owner's encoded dataset. Our experimental results demonstrate that the proposed method is resistant against classifier transfer attacks as shown in row 1 of Table 2 the classification accuracy is 23.49% and 17.01% for CelabA and Cifar-100, respectively.

Encoding Transfer Attack. We also consider the scenario where the attacker aims to learn a representation of the data owner's encoded set to classify identity features. Again, we assume the attacker constructs a dataset that follows a similar distribution as the data owner's original dataset and corresponding labels. First, the attacker trains a multi-output classification model for identity and attribute features similar to the proposed method. Then a randomly initialized autoencoder is trained to generate encoded samples such that identity and attribute information are both preserved. This is accomplished by freezing the weights of the pre-trained identity and attribute classifier and updating the autoencoder parameters based on the gradients of the classification loss. The attacker's modified anonymization network is trained to maintain high classification accuracy for both the identity and attribute classifiers. The encoding transfer attack is depicted in Fig. 5.

Finally, we evaluate the effective of the proposed method against encoding transfer attacks by assessing the performance of the data owner's identity classifier given the attacker's generated encoded set. The goal of the attack is to generate encoded images that include exploitable identity features. The data owner's identity classifier is used to verify if identity features are present in the attackers encoded set. Our experimental results show that the proposed method is resistant to encoding transfer attacks as shown in row 2 of Table 2, the classification accuracy is poor 25.59% and 20.58% for CelebA and Cifar-100, respectively.

Table 2. Classifier and encoding transfer attack performance on CelebA and Cifar-100 datasets

Attack Scheme	Identity Acc. (%)	
	CelebA	Cifar-100
Classifier Transfer	23.49	17.01
Encoding Transfer	25.59	20.58

6 Discussion

In this work, we train an attribute and identity classification model on raw image data and use the network to update a convolutional autoencoder to generate anonymized image data. In our experiments, we evaluate the trade-off between privacy and utility of our image anonymization method by measuring the identity and attribute classification accuracy before and after encoding the dataset. Our results show that the attribute classification accuracy remains high for the transformed images while the identity classification accuracy is significantly reduced for the transformed images. Also, we evaluate the robustness to attacks against the proposed method. In our experiments, we demonstrate that the classifier transfer attack and encoding transfer attack are unsuccessful at inferring the identity of the original images. The identity suppression loss function could be modified as an extension to our proposed method by minimizing the error between the incorrect class and the predicted label using cross-entropy, which we leave for future work. Also, we suspect that the autoencoder's encoder latent space may be sufficient to develop a DL classification model for attribute features while excluding identity features as compared to our proposed method in which we develop our anonymization network with obfuscated reconstructed images (decoder output), we leave this for future work.

7 Conclusion

We proposed an image anonymization method using a standard convolutional autoencoder and multi-ouput resnet50 model to enhance the privacy of raw image data. The images were transformed into unrecognizable versions of the original input data. Highly relevant feature information that is useful for classification was captured in the encoded images. Additionally, we increase privacy through the reduction of identity classification accuracy using the transformed images. In this paper, we demonstrated that the proposed method was able to protect raw data features in the original images and enhance privacy of identity feature information while maintaining model utility with high attribute classification accuracy. In our experiments, we evaluated the effectiveness of our image anonymization method by measuring the reduction of attribute and identity classification accuracy. The experimental results confirm that our proposed method not only enables us to maintain high image attribute classification accuracy but also to reduce image identity classification accuracy.

References

1. Atallah, M.J., Pantazopoulos, K.N., Rice, J.R., Spafford, E.E.: Secure outsourcing of scientific computations. Adv. Comput. **54**, 215–272 (2002)
2. Yuan, X., Wang, X., Wang, C., Squicciarini, A., Ren, K.: Enabling privacy-preserving image-centric social discovery. In: Proceedings of the 2014 IEEE 34th International Conference on Distributed Computing Systems, ser. ICDCS '14. USA: IEEE Computer Society, pp. 198–207 (2014). https://doi.org/10.1109/ICDCS.2014.28
3. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Trans. Pattern Anal. Mach. Intell. **39**(2), 209–226 (2016)
4. Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., Maier, A.: Is medical chest x-ray data anonymous? arXiv preprint arXiv:2103.08562 (2021)
5. McPherson, R., Shokri, R., Shmatikov, V.: Defeating image obfuscation with deep learning. arXiv preprint arXiv:1609.00408 (2016)
6. Tanaka, M.: Learnable image encryption. In: 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), pp. 1–2 (2018)
7. Sirichotedumrong, W., Maekawa, T., Kinoshita, Y., Kiya, H.: Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 674–678 (2019)
8. Sirichotedumrong, W., Kinoshita, Y., Kiya, H.: Pixel-based image encryption without key management for privacy-preserving deep neural networks. IEEE Access **7**, 177:844–177:855 (2019)
9. Sirichotedumrong, W., Kiya, H.: A gan-based image transformation scheme for privacy-preserving deep neural networks (2020). https://arxiv.org/abs/2006.01342
10. Chen, Z., Zhu, T., Xiong, P., Wang, C., Ren, W.: Privacy preservation for image data: a Gan-based method. Int. J. Intell. Syst. **36**(4), 1668–1685 (2021)
11. Rastogi, V., Suciu, D., Hong, S.: The boundary between privacy and utility in data publishing. In: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 531–542 (2007)
12. Li, T., Li, N.: On the tradeoff between privacy and utility in data publishing. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 517–526 (2009)
13. Yonghao, G., Weiming, W.: A quantifying method for trade-off between privacy and utility. In: IET International Conference on Information and Communications Technologies (IETICT 2013). IET, pp. 270–273 (2013)
14. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
15. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
16. Yao, A.C.: Protocols for secure computations. In: 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982), pp. 160–164. IEEE (1982)
17. Chase, M., Gilad-Bachrach, R., Laine, K., Lauter, K., Rindal, P.: Private collaborative neural network learning. Cryptology ePrint Archive (2017)
18. Mohassel, P., Zhang, Y.: Secureml: a system for scalable privacy-preserving machine learning. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 19–38 (2017)
19. Wagh, S., Gupta, D., Chandran, N.: Securenn: 3-party secure computation for neural network training. Proc. Priv. Enhancing Technol. **2019**(3), 26–49 (2019)

20. Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D., Taft, N.: Privacy-preserving ridge regression on hundreds of millions of records. In: 2013 IEEE Symposium on Security and Privacy, pp. 334–348 (2013)
21. Aono, Y., Hayashi, T., Trieu Phong, L., Wang, L.: Scalable and secure logistic regression via homomorphic encryption. In: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, pp. 142–144 (2016)
22. Bonte, C., Vercauteren, F.: Privacy-preserving logistic regression training. BMC Med. Genomics **11**(4), 13–21 (2018)
23. Crawford, J.L.H., Gentry, C., Halevi, S., Platt, D., Shoup, V.: Doing real work with FHE: the case of logistic regression. Cryptology ePrint Archive, Paper 2018/202 (2018). https://eprint.iacr.org/2018/202
24. Graepel, T., Lauter, K., Naehrig, M.: ML confidential: machine learning on encrypted data. In: Kwon, T., Lee, M.-K., Kwon, D. (eds.) ICISC 2012. LNCS, vol. 7839, pp. 1–21. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37682-5_1
25. Kim, M., Song, Y., Wang, S., Xia, Y., Jiang, X., et al.: Secure logistic regression based on homomorphic encryption: design and evaluation. JMIR Med. Informat. **6**(2), e8805 (2018)
26. Nandakumar, K., Ratha, N., Pankanti, S., Halevi, S.: Towards deep neural network training on encrypted data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
27. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Sig. Process. Mag. **37**(3), 50–60 (2020)
28. Bonawitz, K., et al.: Towards federated learning at scale: system design. Proc. Mach. Learn. Syst. **1**, 374–388 (2019)
29. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)
30. Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: strategies for improving communication efficiency (2016). https://arxiv.org/abs/1610.05492
31. Kairouz, P., et al.: Advances and open problems in federated learning. Found. Trends® Mach. Learn. **14**(1–2), pp. 1–210 (2021)
32. Huang, Y., Song, Z., Li, K., Arora, S.: InstaHide: instance-hiding schemes for private distributed learning. In: Proceedings of the 37th International Conference on Machine Learning, Ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul, pp. 4507–4518 (2020). https://proceedings.mlr.press/v119/huang20i.html
33. Yala, A., et al.: Neuracrypt: hiding private health data via random neural networks for public training (2021). https://arxiv.org/abs/2106.02484
34. Carlini, N., et al.: Is private learning possible with instance encoding? (2020). https://arxiv.org/abs/2011.05315
35. Raynal, M., Achanta, R., Humbert, M.: Image obfuscation for privacy-preserving machine learning (2020). https://arxiv.org/abs/2010.10139
36. Malekzadeh, M., Clegg, R.G., Cavallaro, A., Haddadi, H.: Mobile sensor data anonymization. In: Proceedings of the International Conference on Internet of Things Design and Implementation, pp. 49–58 (2019)
37. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). https://arxiv.org/abs/1512.03385