PRIC: A Privacy-Respecting Image Crowdsourcing Framework for Deep Learning with Anonymized Intermediate Representations

Ang Li[1], **Yixiao Duan**[2], **Huanrui Yang**[1], **Yiran Chen**[1], **Jianlei Yang**[2],
[1]Department of Electrical and Computer Engineering, Duke University
[2]School of Computer Science and Engineering, Beihang University
{ang.li630,huanrui.yang,yiran.chen}@duke.edu, {jamesdyx,jianlei}@buaa.edu.cn

1 Introduction

In the past decade, deep learning has achieved unprecedented success in computer vision. Such success benefits from various large-scale datasets, such as ImageNet (1), MS COCO (2), etc. These datasets that are crowdsourced from individual users for deep learning applications often contain private information such as gender, age, etc. The data breach of Facebook (3), for example, raises users' severe concerns about sharing their personal data. These emerging privacy concerns hinder generation or use of large-scale crowdsourcing datasets and lead to hunger of training data of many new deep learning applications.

Many countries are also establishing laws to protect data security and privacy. As a famous example, the new European Union's General Data Protection Regulation (GDPR) (4) is an example, require companies to not store personal data for a long time, and allow users to delete or withdraw their personal data within thirty days. However, such regulation cannot be applied if the data is anonymized. The need of collecting large-scale crowdsourcing dataset under strict requirement of data privacy motivates us to design a privacy-respecting image crowdsourcing framework: the raw image from the users is locally transformed into an intermediate representation that can remove the private information while retaining the discriminative features for primary learning tasks.

A few studies have been done to protect private information of intermediate representation extracted from images. Osia *et al.* (5) design an approach to hide privacy information from extracted features by maximizing mutual information between the feature and primary variable while while minimizing mutual information between the feature and sensitive variable. Feutry *et al.* (6) also propose an image anonymization approach to hide sensitive features related to private attributes. We also designed an adversarial training framework (7) to prevent attackers from reconstructing raw images and inferring the private attributes from the extracted features, while retaining useful information for an intended learning task. However, all the above solutions are designed for known primary learning tasks, which limits their applicability when the primary learning task is unknown. Therefore, it is necessary to design a more general method to protect private information without target specific learning tasks.

In this paper, we propose *PRIC* - a privacy-respecting image crowdsourcing framework with anonymized intermediate representation. The goal of this framework is to learn a feature extractor that can hide the privacy information from the intermediate representations while maximally retaining the original information embedded in the raw data for primary learning tasks. As Figure 1 shows,Participants will be able to locally run the feature extractor and submit only those intermediate representations to the data collector instead of submitting the raw data. The data collector can then train deep learning models using these collected intermediate representations. Existing adversarial training methods (6; 7) for anonymizing features usually require to determine the primary learning task before training. On the contrary, PRIC does not require the knowledge of the primary learning task. It is challenging to remove all the private information that needs to be protected while keeping

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

everything else for primary learning tasks. To address this issue, we design a hybrid training method to learn the anonymized intermediate representation. The training purpose is two folded: **hiding private information from features (goal 1)** and **maximally retaining original information (goal 2)**. Specifically, the goal 1 is achieved by performing our proposed adversarial training algorithm, which simulates the game between an attacker who makes efforts to infer private attributes from the extracted features and a defender who aims to protect user privacy. The goal 2 is realized by maximizing the mutual information between the feature of the raw image and the combination of the hidden privacy feature and the retained feature.

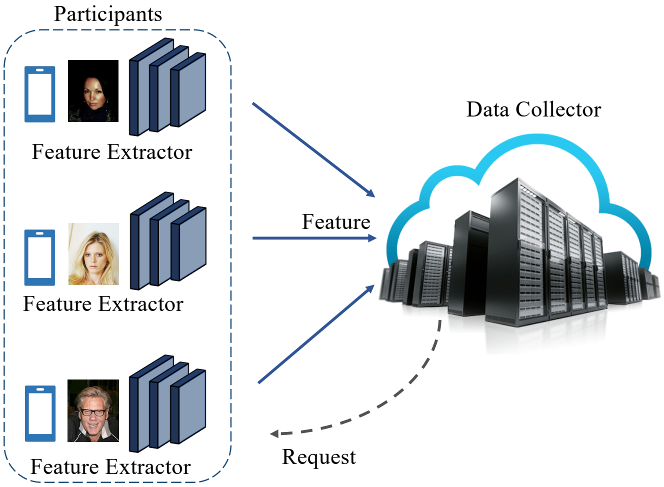

Figure 1: The design of PRIC.

2 Problem Formulation

Let x be the raw image, z the original information need to be retained and u the privacy information requires to hide. We aim to extract a feature z so that u can be hidden from x, i.e, there is less information overlapped between z and u. We also expect z can maximally retain the original information from x, so that the information carried by the combination of z and u is maximally overlapped with the information embedded in x.

Formally, we aim to find a feature extractor with parameters φ which defines the distribution of feature z given raw data x and user defined privacy information u.

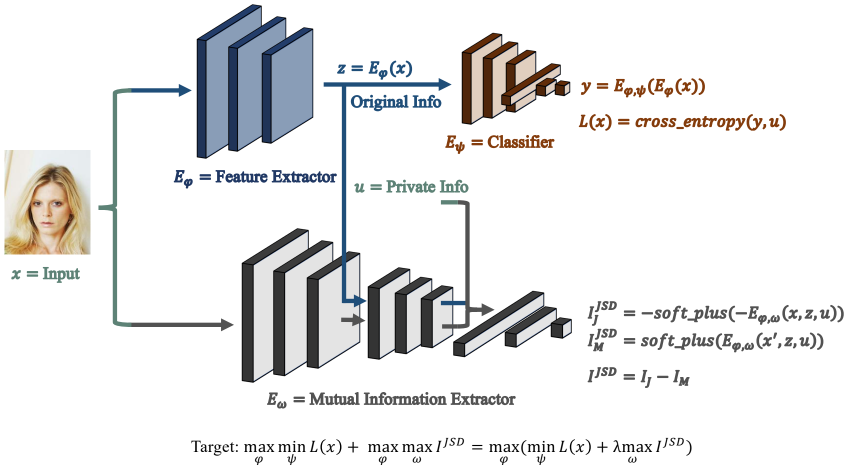
$z = E_\varphi(x)$
Original Info
$y = E_{\varphi,\psi}(E_\varphi(x))$
$L(x) = cross\_entropy(y,u)$
E_ψ = Classifier
E_φ = Feature Extractor
u = Private Info
x = Input
$I_J^{JSD} = -soft\_plus(-E_{\varphi,\omega}(x,z,u))$
$I_M^{JSD} = soft\_plus(E_{\varphi,\omega}(x',z,u))$
$I^{JSD} = I_J - I_M$
E_ω = Mutual Information Extractor
Target: $\max_\varphi \min_\psi L(x) + \max_\varphi \max_\omega I^{JSD} = \max_\varphi(\min_\psi L(x) + \lambda \max_\omega I^{JSD})$

Figure 2: The hybrid training procedure for learning the feature extractor.

Goal 1: hide privacy information. The goal 1 can be formulated as $\min_{\varphi} I(z;u)$. Since we can upper bound $\min_{\varphi} I(z;u)$ as $\min_{\varphi} I(z;u) < \mathbb{E}_{q_\varphi}\left[\log p_\Psi(u|z) - \log p(u)\right]$ with any distribution p_Ψ, we can parameterize p_Ψ with a neural network. Then the goal 1 is therefore converted to an adversarial training process, which simulates the game between an attacker p_Ψ who tries to predict u from z and a defender q_φ who aims to protect the user privacy. As Figure 2 shows, the classifier E_ψ is jointly trained with the feature extractor E_φ as a complete CNN model. Given the input image x, the prediction can be expressed as $y = E_\psi(E_\varphi(x))$. The performance of the classifier is measured using the cross-entropy loss function, which is expressed as:

$$\mathcal{L}(C) = cross\_entropy(y, u), \tag{1}$$

where u is label of the private attribute we aim to protect. Therefore, when play as an attacker, the feature extractor can be trained using Eq. 2:

$$\arg\min_{\psi} \mathcal{L}(C). \tag{2}$$

On the contrary, the obfuscator can be trained using Eq. 3 when simulating a defender:

$$\arg\min_{\varphi,\psi} -\mathcal{L}(C). \tag{3}$$

Goal 2: retain original information. The objective of goal 2 can be formulated as $\max_{\varphi} I(z;x)$. However, considering the fact that u and x may have some correlation, directly maximizing $I(z;x)$ may conflict with the objective in goal 1. Therefore, for goal 2 we propose to maximize $I(x;z,u)$: the mutual information between the raw image and the union of the feature and the private information. Since we tend to remove the information of u from z in goal 1, keeping u within the objective of goal 2 will lead the extracted features to focus on only covering the information of x uncorrelated to u, therefore mitigating the potential conflict between our two goals. Similar to the method proposed in (8), a mutual information estimator that can be parametrized as a deep neural network will be used to provide a lower bound of the mutual information, and the feature extractor will be trained to maximize such a lower bound. Specifically, we adopt Jensen-Shannon mutual information estimator (9; 10) to express the lower bound of the mutual information x and the combination of z and u, which is defined as follows:

$$\mathcal{I}(x;z,u) \geq \mathcal{I}^{(JSD)}_{\varphi,\omega}(x;z,u) := -sp(-E_{\varphi,\omega}(x, E_\varphi(x), u)) - sp(E_{\varphi,\omega}(x', E_\varphi(x), u)), \tag{4}$$

where x' is an input sampled from the same dataset of x, $sp(z) = log(1 + e^z)$ is the softplus functions and E_ω is a discriminator function modeled by a neural network with parameters ω as shown in Figure 2. Therefore, to maximally retain the original information, the feature extractor and the mutual information estimator can be optimized using Eq. 5:

$$\arg\max_{\varphi}\max_{\omega} \mathcal{I}^{(JSD)}_{\varphi,\omega}(x;z,u). \tag{5}$$

Finally, based on Eq. 2, Eq. 3 and Eq. 5, the objective function of the proposed hybrid training procedure can be summarized as:

$$\arg\max_{\varphi}(\min_{\psi} \mathcal{L}(C) + \max_{\omega} \lambda \mathcal{I}^{(JSD)}_{\varphi,\omega}(x;z,u)), \tag{6}$$

where λ is a parameter to tune the utility-privacy trade-off.

3 Evaluation

3.1 Experiment Setup

We implement PRIC with PyTorch, and train it on a server with 6×NVIDIA TITAN RTX GPUs. We apply mini-batch technique in training with a batch size of 128, and adopt the AdamOptimizer (11) with an adaptive learning rate in the adversarial training procedure.

We adopt CelebA (12) for the training and testing of PRIC. CelebA consists more than 200K face images. Each face image is labeled with 40 binary facial attributes. The dataset is split into 160K images for training and 40K images for testing.

Given a specific private attribute to protect, we first train the feature extractor using PRIC. Then, we train a classifier using features that are extracted by the pre-trained feature extractor. To compare with our design in experiments, we also implement a baseline model. The baseline model has the same architecture as the PRIC but is trained without using our proposed adversarial training procedure.

3.2 Utility-Privacy Trade-Off

We quantitatively evaluate the effectiveness of privacy protection and the accuracy of primary learning tasks using PRIC. Specifically, we first choose detecting 'gray hair' and 'smiling' as the primary classification tasks, and 'young' as the private attribute we aim to protect. Figure 4 shows the average accuracy of primary learning tasks and private attribute using the classifier which is trained in the way adopted by PRIC and the baseline model, respectively. With the proposed adversarial training, PRIC can effectively prevent private attributes from being inferred by an attacker while only incurring a small accuracy drop on the primary classification tasks. For example, when we set $\lambda = 0.1$, the accuracy of 'young' dramatically decreases from 83.03% to 65.63%, while the accuracy of 'gray hair' and 'smiling' only drop by 5.08% and 2.16%, respectively. With a larger λ we can achieve higher accuracy on primary classification tasks, but the protection for private attribute will be weakened. For instance, when we set $\lambda = 10$, the accuracy of 'gray hair' and 'smiling' increase to 92.15% and 91.7% from 87.97% and 90% when we set $\lambda = 0.1$, but the accuracy of 'young' increases to 71.1%. The reason is that the larger λ enforces the feature extractor to retrain more original information from the image, which is correlated to the private attribute.

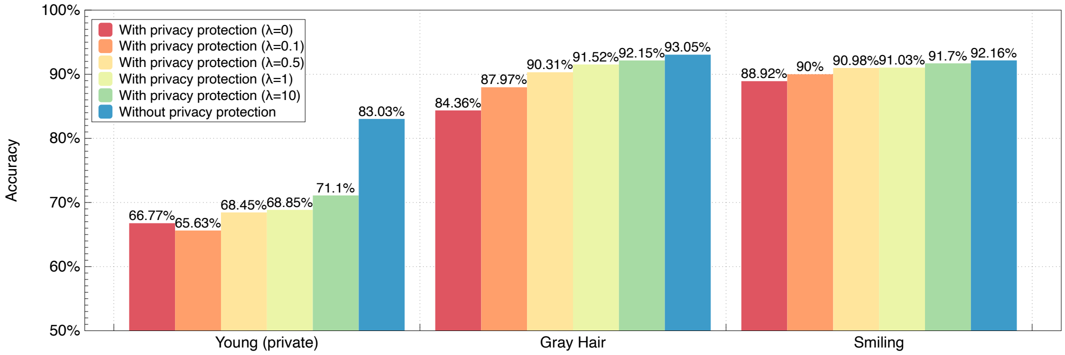

Figure 3: Accuracy of primary learning tasks and private attributes on CelebA using PRIC ('young' is set as the private attribute).

We also conduct another experiment by choosing detecting 'heavy makeup' and 'smiling' as the primary classification tasks, and 'gender' as the private attribute we aim to protect. Generally, we observed the result as same as that of the above experiment - the private attribute can be effectively protected will an appropriate λ while only incurring a small performance drop on the primary learning tasks. For example, when we set $\lambda = 0.1$, the accuracy of 'gender' dramatically decreases from 96.64% to 57.41%, while the accuracy of 'heavy makeup' and 'smiling' only drop by 3.55% and 2.83%, respectively.

4 Conclusion

We proposed an privacy-respecting image crowdsourcing framework PRIC with learning the anonymized intermediate representations. This is done by training a feature extractor to hide privacy information from features while maximally retaining the original information embedded in the raw image. The feature extractor is trained using our proposed hybrid training procedure, including an adversarial training process by simulating between an attacker who makes efforts to infer private attributes from the extracted features and a defender who aims to protect user privacy, and maximizing the mutual information between the raw image and the union of the feature and the private information. Our experiments on CelebA dataset shows the classification accuracy of the protected private attribute drops by around 18% and 40% on 'young' and 'gender', respectively. But the accuracy of the primary

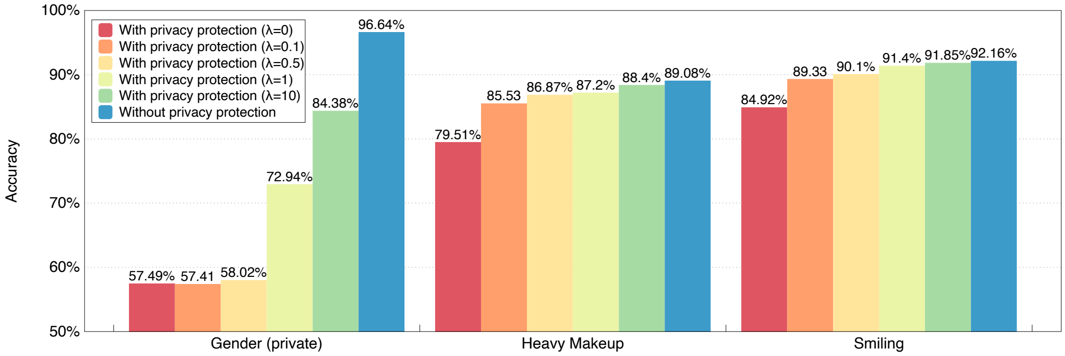

Figure 4: Accuracy of primary learning tasks and private attributes on CelebA using PRIC ('gender' is set as the private attribute).

learning task drops by only 2%. Although PRIC is applied to image crowdsourcing in this paper, it can be easily extended to many other applications, such as crowdsourcing voice data, federated learning, etc. Next, we proposed to conduct extensive experiments to evaluate the performance of PRIC, including protecting multiple private attributes, inferring different primary classification tasks from the anonymized features, etc.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[3] S. Meredith, "Facebook-cambridge analytica: A timeline of the data hijacking scandal," 2018.

[4] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[5] S. A. Osia, A. Taheri, A. S. Shamsabadi, M. Katevas, H. Haddadi, and H. R. Rabiee, "Deep private-feature extraction," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[6] C. Feutry, P. Piantanida, Y. Bengio, and P. Duhamel, "Learning anonymized representations with adversarial neural networks," *arXiv preprint arXiv:1802.09386*, 2018.

[7] A. Li, J. Guo, H. Yang, and Y. Chen, "Deepobfuscator: Adversarial training framework for privacy-preserving image classification," *arXiv preprint arXiv:1909.04126*, 2019.

[8] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[9] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Advances in neural information processing systems*, pp. 271–279, 2016.

[10] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[11] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv.org*, p. arXiv:1412.6980, Dec. 2014.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.