VISUAL PRIVACY PROTECTION VIA MAPPING DISTORTION

Peidong Liu[1,⋆] *Yiming Li*[1,⋆] *Yong Jiang*[1,2] *Shu-Tao Xia*[1,2]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]PCL Research Center of Networks and Communications, Peng Cheng Laboratory
{lpd19, li-ym18}@mails.tsinghua.edu.cn; {jiangy, xiast}@sz.tsinghua.edu.cn

ABSTRACT

Data privacy protection is an important research area, which is especially critical in this big data era. To a large extent, the privacy of visual classification tasks is mainly in the one-to-one mapping between image and its corresponding label, since this relation provides a great amount of information and can be used in other scenarios. In this paper, we propose mapping distortion protection (MDP) and its augmentation-based extension (AugMDP) to protect the data privacy by modifying the original dataset. In MDP, the label of the modified image does not match the ground-truth mapping, yet DNNs can still learn the ground-truth relation even when the provided mapping is distorted. As such, this method protects privacy when the dataset is leaked. Extensive experiments are conducted on CIFAR-10 and restricted CASIA-WebFace dataset, which verifies the effectiveness and feasibility of the proposed method.

Index Terms— Privacy protection, Face recognition, Image classification, Deep learning

1. INTRODUCTION

Deep Learning, especially deep neural networks (DNNs), have been widely and successfully used in many fields, such as visual classification [1, 2, 3], object detection [4, 5, 6, 7], super-resolution [8, 9, 10], and visual tracking [11, 12, 13]. A large amount of training data is one of the key factors in the success of deep learning methods. Some of the well-developed methods, such as face recognition, have been used in real life.

While the massive amount of data dramatically improves the performance of the deep learning-based model, the collection of data from millions of users also raises serious privacy issues. For example, companies always store the training data on their servers, which may contain some sensitive items such as facial images used to log in to a specific system. However, this information has the risk of leakage, which harms the privacy of users.

To a large extent, the privacy of visual classification tasks is mainly in the ground-truth one-to-one mapping between the input image and its corresponding label, since this relation provides a significant amount of information and can be used in other scenarios. One of the most interesting related works, the k-anonymity [14], concentrates on hiding the ground-truth one-to-one mapping between the field-structure data and its corresponding label. Specifically, k-anonymity hides the mapping by guaranteeing that individuals who are the subjects of the data can not be re-identified while the private data remains practically available. However, this method mainly focuses on field-structured data, which can not be used in protecting images.

In this paper, we propose mapping distortion protection (MDP) and its augmentation-based extension (AugMDP). Our approaches aim at exploring a new possible way to protect the visual privacy in classification tasks by distorting the ground-truth one-to-one mapping between the image and its label. In other words, for a specific image in the modified dataset, its provided label does not match what the image looks like. For example, in the modified dataset, which contains only distorted mapping, a dog-like image may be labeled as the car. In this way, we can still protect the privacy when the dataset is leaked, under the condition that the hacker has no prior knowledge of the ground-truth mapping. Besides, models that use the modified dataset for training can still achieve good generalization on the standard test set, which guarantees the utility of the private dataset.

The mechanism behind MDP is that DNNs utilize lots of unstable yet useful features such as texture, as discussed in [15, 16]. It is precisely by hiding the information of the target image in the modified image that DNNs can learn the ground-truth relation even when the provided mapping is distorted. As such, companies can protect visual privacy by only storing the modified dataset instead of the original one with sensitive ground-truth mapping on their servers.

In addition to protecting data privacy, MDP has two extra latent advantages as follows: 1) The hackers may be insensible of the correctness of the dataset, and it is hard to denoise the modified dataset without prior knowledge even if they are aware of the distorted mapping. 2) The leak can be detected if someone applies the distorted mapping provided in the modified dataset.

The main contributions of our work can be summarized

⋆: Equal contribution.

as follows: 1) We propose a novel privacy protection method, dubbed mapping distortion protection (MDP), by modifying the ground-truth relation between image and the corresponding label. 2) An augmentation-based extension (AugMDP) is proposed, which further enhances the performance. 3) Extensive experiments are conducted, which verifies the feasibility and effectiveness of the proposed method.

2. PROPOSED METHOD

2.1. Preliminary

Suppose $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ is the original dataset needed to be protected, where N is the size of the dataset and the input-label pair $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$. We now define some key concepts for further discussion. The illustration of those concepts is shown in Fig. 1.

- **selected image**: For a given original dataset $\mathcal{D}$, the selected image $\boldsymbol{x}_{selected} \in \{\boldsymbol{x} | (\boldsymbol{x}, y) \in \mathcal{D}\}$.
- **target image**: The target image is also from the original dataset $\mathcal{D}$ and used for the modification of the selected image.
- **modified image**: The modified image is initialized with the selected image and updated by minimizing the distance between output of itself and output of the target image in the middle layer of a given pre-trained DNN.
- **original label**: The label of selected image.
- **target label**: The label of target image and modified image.

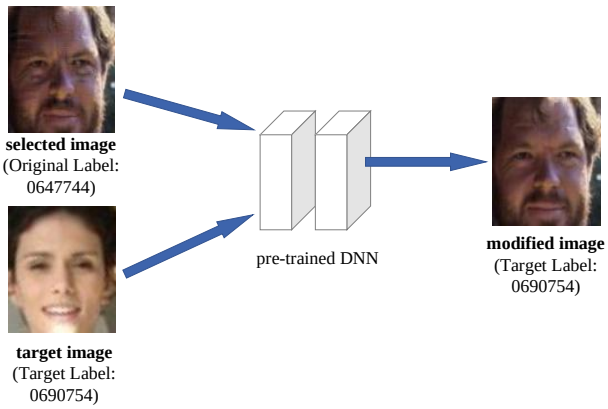


Fig. 1: The illustration of selected image, target image, modified image, and their corresponding label.

As suggested in [17], the features used by DNNs can be divided into stable and unstable two categories. Intuitively speaking, stable features are visible and interpretable to humans, such as the profile [18, 19], while the unstable features are usually invisible, such as the texture [15, 16]. Both stable and unstable features are strongly predictive for the classifier. Since unstable features can be replaced easily without being discovered by humans (small variance in image), these features can be utilized to construct a new dataset with distorted mapping, namely the **modified dataset**. The detailed construction procedure will be discussed in the following sections.

2.2. Mapping distortion protection

As discussed above, hiding the ground-truth mapping between image and its label is critical for privacy protection. Therefore, instead of storing the original dataset directly, we suggest keeping the modified version whose input-label mapping is distorted.

In this paper, we propose a novel method, dubbed mapping distortion protection (MDP), for constructing such a modified dataset. Specifically, in order to extract the useful yet unstable features from the target image, we utilize a standard pre-trained DNN. We first initialize the modified image with the selected image and then minimize the distance between the output of the modified image and the output of the target image in the middle layer of the DNN. We construct the modified training set $\mathcal{D}_{mod}$ via a one-to-one correspondence $\boldsymbol{x}_{selected} \rightarrow \boldsymbol{x}_{modified}$, where $\boldsymbol{x}_{selected}$ is the selected image and $\boldsymbol{x}_{modified}$ is the modified image. To be specific, for every target image $\boldsymbol{x}_{target}$ with label y_{target} in $\mathcal{D}$, MDP randomly chooses a selected image $\boldsymbol{x}_{selected}$ with original label $y_{original}$ from the $\mathcal{D}$ and initialize $\boldsymbol{x}_{modified}$ with $\boldsymbol{x}_{selected}$. Then, MDP updates $\boldsymbol{x}_{modified}$ gradually so that the output of $\boldsymbol{x}_{modified}$ and the output of $\boldsymbol{x}_{target}$ are similar in the middle layer of the DNN to guarantee that the features used by the DNN are relevant. MDP updates $\boldsymbol{x}_{modified}$ through the following optimization:

$$\boldsymbol{x}_{modified} = \underset{\boldsymbol{x}_{modified} \in [0,1]^d}{\operatorname{argmin}} ||f(\boldsymbol{x}_{modified}) - f(\boldsymbol{x}_{target})||_p, \quad (1)$$

where d is the dimension of the features, and f is the mapping from input to the output of a certain layer in DNN, such as the last layer or the penultimate layer. Specifically, we optimize $\boldsymbol{x}_{modified}$ in the input space using the projected gradient descent (PGD) [20].

As Fig. 2 shows, the modified image is obtained from the combination of the selected image and small perturbation related to unstable yet useful features. Since those invisible yet useful features in the modified dataset are still highly predictive, the distorted mapping can still lead to good generalization in the training of DNNs.

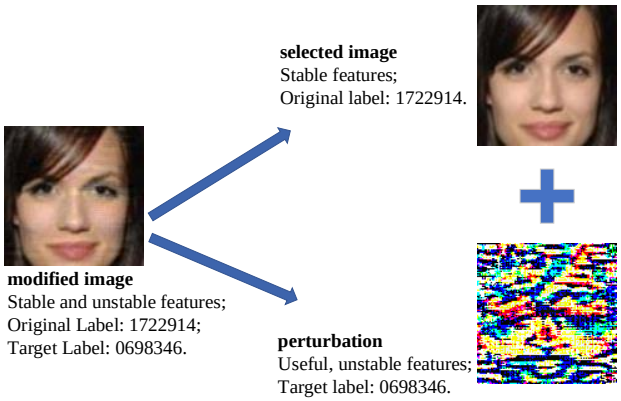


Fig. 2: The illustration of modified image.

Algorithm 1: Construction procedure of the augmented dataset.

Input: Original dataset $\mathcal{D}$, Augmentation-related hyper-parameter T.
Output: Augmented dataset $\mathcal{D}_{modAug}$
Initialize $\mathcal{D}_{modAug} = \{\}$
for *time in range* (T) **do**
  for $(\boldsymbol{x}_{target}, y_{target}) \in \mathcal{D}$ **do**
    Randomly pick a pair $(\boldsymbol{x}_{selected}, y_{original}) \in \mathcal{D}$.
    Calculate $\boldsymbol{x}_{modified}$ according to Equation (1).
    $\mathcal{D}_{modAug} = \mathcal{D}_{modAug} \cup \{(\boldsymbol{x}_{modified}, y_{target})\}$.
  end
end
return $\mathcal{D}_{modAug}$

2.3. Augmented mapping distortion protection

As mentioned in the previous section, we can use the modified dataset instead of the original dataset for storage to protect data privacy. However, due to the adverse effects of the incorrect stable features in the modified images, training with them will result in a certain decrease in the generalization accuracy. In this section, we introduce the augmented MDP (AugMDP), which enhances model performance effectively.

Specifically, in AugMDP, we first construct T different modified datasets $\{\mathcal{D}_{mod}^{(1)}, \cdots, \mathcal{D}_{mod}^{(T)}\}$ through MDP, where T is a non-negative hyperparameter to control the augmentation size. Then, the augmented dataset is obtained by combining all these datasets, *i.e.*, $\mathcal{D}_{modAug} = \mathcal{D}_{mod}^{(1)} \cup \mathcal{D}_{mod}^{(2)} \cup \cdots \cup \mathcal{D}_{mod}^{(T)}$. This augmented method is effective since the extra information carried by the unstable yet useful features in the augmented version is conducive to the learning of the model. The construction procedure of the augmented dataset is shown in Algorithm 1, and the effectiveness of such augmentation is verified in Section 3.3.

3. EXPERIMENTS

3.1. Settings

The experiments are conducted in CIFAR-10 [21] and (restrict) CASIA-WebFace dataset [22]. Note that instead of the whole CASIA-WebFace, we only focus on a subset of the dataset for the consideration of computational complexity. Restricted CASIA-WebFace has 46492 images with only 1000 classes, which are randomly chosen from the original one. For the model architecture, ResNet-50 [23] and IR-50 with ArcFace (an improved version of the vanilla ResNet-50) [24] are used in CIFAR-10 and restricted CASIA-WebFace datasets, respectively. To construct the modified dataset, we perform PGD [20] to optimize the objective function under ℓ^∞ norm, which aims to minimize the distance between the output of the modified image and output of the target image in the penultimate layer of the pre-trained model. Specifically, PGD-100 and PGD-40 with step size 0.1 are used in CIFAR-10 and restricted CASIA-WebFace dataset, respectively. The code and dataset will be publicly available.

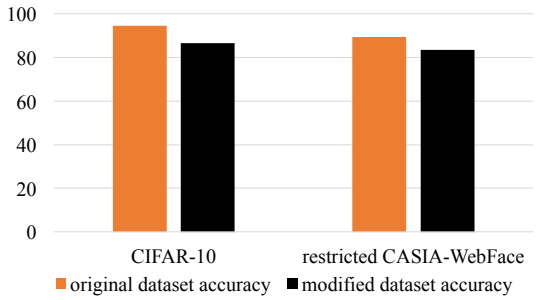


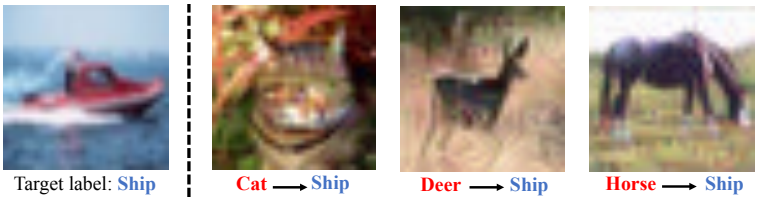
Fig. 3: Test accuracy in CIFAR-10, restricted CASIA-WebFace and their corresponding modified dataset.

	MDP	AugMDP (2)
CIFAR-10	85.52	**89.04**
restricted CASIA-WebFace	83.04	**84.67**

Table 1: Test accuracy of MDP and AugMDP in modified dataset constructed from CIFAR-10 and restricted CASIA-WebFace dataset.

3.2. Verification on CIFAR-10 and CASIA-WebFace

In this experiment, we construct two datasets $\mathcal{D}_{mod-CIFAR10}$ and $\mathcal{D}_{mod-CASIA}$ from CIFAR-10 and restricted CASIA-WebFace, respectively. The left-hand side of Fig.3 represents the test accuracy of the model trained on CIFAR-10, and of the model trained on $\mathcal{D}_{mod-CIFAR10}$. The right-hand side of Fig.3 indicates the generalization performance of model trained on restricted CASIA-WebFace and of model trained on $\mathcal{D}_{mod-CASIA}$. The result shows DNNs trained on the



(a) CIFAR-10

(b) restricted CASIA-WebFace

Fig. 4: The illustration of target images and modified images in CIFAR-10 and restricted CASIA-WebFace. **The first column**: target images from original datasets. **The next three columns**: modified images corresponding to the target images. The red and blue words indicate the original and target label, respectively.

Network Architecture	$\mathcal{D}_{mod-CIFAR10}$
ResNet-50	85.52
ResNet-154	86.08
DenseNet-154	83.15

Network Architecture	$\mathcal{D}_{mod-CASIA}$
IR-50	83.04
IR-152	84.37
IR-SE-50	81.37

Table 2: The transferability evaluation in $\mathcal{D}_{mod-CIFAR10}$ and $\mathcal{D}_{mod-CASIA}$.

modified dataset can generalize well in the standard testset, and therefore the utility of the dataset is verified.

Fig. 4 illustrates some target images and the modified images. To quantitively assess the similarity between the selected image and corresponding modified image, we calculate SSIM (Structural Similarity Index) [25] of the pairs. The mean SSIM for CIFAR-10 and restricted CASIA-WebFace are 96.9% and 96.1%, respectively. The results show that the modified image is highly similar to the selected image, therefore the invisibility of modification is guaranteed.

3.3. Augmentation effect

To verify the effectiveness of the augmentation in AugMDP, we compare AugMDP and MDP on both CIFAR-10 and restricted CASIA-WebFace datasets. Table 1 shows the test accuracy of these two methods on two datasets, where the number in the parenthesis following AugMDP is the value of augmentation-related hyper-parameter T. Particularly, AugMDP (1) is equivalent to MDP.

Table 1 shows AugMDP is better than MDP across different tasks even when T is relatively small. Besides, T should be adjusted according to specific requirements, since AugMDP brings additional computations and storage costs.

3.4. Transferability

In this experiment, we verify whether a modified dataset generated by a given network is also effective for training similar network architectures.

Table 2 shows the test accuracy of several architectures (ResNet-50, ResNet-154, and DenseNet-154 [26]) trained in $\mathcal{D}_{mod-CIFAR10}$ generated by standard ResNet-50, and the test accuracy of IR-50, IR-152, and IR-SE-50[1] trained on the $\mathcal{D}_{mod-CASIA}$ generated by IR-50 model. The result shows that the modified dataset is also effective for training similar network architectures, which indicates that unstable features in the visual images are probably general to similar classifiers.

4. CONCLUSIONS

In this paper, we propose mapping distortion protection (MDP) and its augmentation-based extension (AugMDP) to protect the data privacy by modifying dataset. This method is motivated by the fact that DNNs utilize some useful yet unstable features, which can be modified invisibly. Based on this method, we can protect privacy when the dataset is leaked and achieve good generalization when training on such a modified dataset. Extensive experiments are conducted to verify the feasibility and effectiveness of the method.

[1]IR-SE-50 combines IR-50 and SENet [27].

5. REFERENCES

[1] Xiao-Tong Yuan, Xiaobai Liu, and Shuicheng Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[2] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik, "Pairwise confusion for fine-grained visual classification," in *ECCV*, 2018, pp. 70–86.

[3] Naiyan Wang and Dit-Yan Yeung, "Learning a deep compact image representation for visual tracking," in *NeurIPS*, 2013, pp. 809–817.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

[6] Chenchen Zhu, Yihui He, and Marios Savvides, "Feature selective anchor-free module for single-shot object detection," in *CVPR*, 2019, pp. 840–849.

[7] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos, "Towards universal object detection by domain attention," in *CVPR*, 2019, pp. 7289–7298.

[8] Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer, "Deep learning massively accelerates super-resolution localization microscopy," *Nature Biotechnology*, vol. 36, no. 5, pp. 460, 2018.

[9] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *CVPR*, 2018, pp. 606–615.

[10] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *ICCV*, 2019.

[11] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders, "Siamese instance search for tracking," in *CVPR*, 2016, pp. 1420–1429.

[12] Hyeonseob Nam and Bohyung Han, "Learning multi-domain convolutional neural networks for visual tracking," in *CVPR*, 2016, pp. 4293–4302.

[13] Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, and Jianhua Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *CVPR*, 2019, pp. 4670–4679.

[14] Latanya Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[15] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *ICLR*, 2019.

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "Texture and art with deep neural networks," *Current Opinion in Neurobiology*, vol. 46, pp. 178–186, 2017.

[17] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," in *NeurIPS*, 2019.

[18] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

[19] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick, "Cognitive psychology for deep neural networks: A shape bias case study," in *ICML*, 2017, pp. 2940–2949.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[21] Alex Krizhevsky et al., "Learning multiple layers of features from tiny images," Tech. Rep., Citeseer, 2009.

[22] Yi Dong, Lei Zhen, Shengcai Liao, and Stan Z. Li, "Learning face representation from scratch," *Computer Science*, 2014.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[27] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.