ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/348920165

Visual Privacy Protection via Mapping Distortion

Conference Paper · January 2021

DOI: 10.1109/ICASSP39728.2021.9414149

CITATIONS
8

READS
66

4 authors, including:

Yiming Li
Zhejiang University
76 PUBLICATIONS **1,173** CITATIONS

SEE PROFILE

Peidong Liu
Tsinghua University
9 PUBLICATIONS **68** CITATIONS

SEE PROFILE

All content following this page was uploaded by Yiming Li on 03 February 2021.

The user has requested enhancement of the downloaded file.

VISUAL PRIVACY PROTECTION VIA MAPPING DISTORTION

Yiming Li[1,⋆] *Peidong Liu*[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
{li-ym18, lpd19}@mails.tsinghua.edu.cn; {jiangy, xiast}@sz.tsinghua.edu.cn

ABSTRACT

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 data is mainly in the mapping between the 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 the mapping distortion based protection (MDP) and its augmentation-based extension (AugMDP) to protect the data privacy by modifying the original dataset. In the modified dataset generated by MDP, the image and its label are not consistent ($e.g.$, a cat-like image is labeled as the dog), whereas the DNNs trained on it can still achieve good performance on benign testing set. As such, this method can protect privacy when the dataset is leaked. Extensive experiments are conducted, which verify the effectiveness and feasibility of our method. The code for reproducing main results is available at `https://github.com/PerdonLiu/Visual-Privacy-Protection-via-Mapping-Distortion`.

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

1. INTRODUCTION

Deep learning, especially deep neural networks (DNNs), have been successfully adopted in many fields, such as object detection [1, 2, 3], super-resolution [4, 5, 6], and visual tracking [7, 8, 9]. A large amount of training data is one of the key factors in the success of DNNs. While the massive amount of data dramatically improves the performance of the DNNs, the collection of data from millions of users also raises serious privacy issues. For example, the collected data has the risk of being leaked, which harms the privacy of users. Accordingly, how to protect the privacy of data is of great significance.

In this paper, we focus on protecting data privacy in image classification tasks. To a large extent, the privacy of the tasks is mainly in the ground-truth 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. To the best of our acknowledge, there is only one research, $i.e.$, k-anonymity [10], in this research area. 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 focused on the field-structured data, which can not be adopted in protecting visual data.

To address this problem, we propose a mapping distortion based protection (MDP) and its augmentation-based extension (AugMDP). Our approaches aim at exploring a new possible way to protect the visual privacy by distorting the ground-truth mapping between the image and its label. In other words, for a specific image in the modified dataset generated by the proposed method, its provided label does not match what the image truly is. 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 trained with the modified dataset can still achieve good performance on the benign testing set, which guarantees the utility of the generated dataset.

Specifically, the modified image is generated by minimizing the distance with a random image with the provided label in the feature space. The mechanism behind MDP is that the DNNs utilize lots of unstable yet useful features such as texture, as discussed in [11, 12, 13]. It is precisely by hiding the information of the given label in the modified image that DNNs can learn the ground-truth relation even when the provided mapping seems to be distorted. Besides, MDP has two extra latent advantages, including **(1)** The hackers can hardly realize the incorrectness of the dataset since the perturbation in the modified image is invisible. **(2)** The leakage can be detected if the specific distorted mapping appears.

The main contributions of our work can be summarized as follows: **(1)** We explore a novel angle of privacy protection by modifying the ground-truth relation between image and its corresponding label. **(2)** We provide a simple yet effective method, the MDP, and its augmentation-based extension (AugMDP) for privacy protection. **(3)** Extensive experiments

⋆ indicates equal contribution.

This work is supported by the National Science Foundation of China under Grant 61771273, the Natural Science Foundation of Zhejiang Province (LSY19A010002), and the R&D Program of Shenzhen (JCYJ20180508152204044). Corresponding author: Yong Jiang.

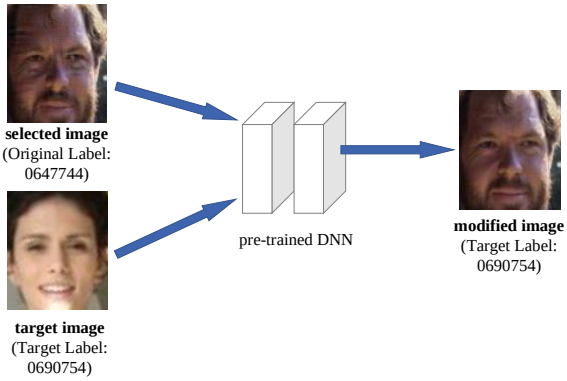


Fig. 1: The illustration of selected image, target image, modified image, and their corresponding label. The modified image looks similar to its corresponding selected image, whereas its label is the label of the corresponding target image (which is different from that of the selected image). Accordingly, the ground-truth mapping between image and label of the original dataset is hidden and therefore is protected.

are conducted, which verifies the feasibility and effectiveness of our privacy protection method.

2. THE 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 $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$. Before the detailed discussion, we first define some key concepts, as visualized in Fig. 1.

- **Selected Image** $\boldsymbol{x}_{selected}$: For a given original dataset $\mathcal{D}$, the selected image $\boldsymbol{x}_{selected} \in \{\boldsymbol{x}|(\boldsymbol{x}, y) \in \mathcal{D}\}$.
- **Original Label** $y_{original}$: The label of the selected image $\boldsymbol{x}_{selected}$.
- **Target Image** $\boldsymbol{x}_{target}$: The target image is also from the original dataset $\mathcal{D}$, $i.e.$, $\boldsymbol{x}_{target} \in \{\boldsymbol{x}|(\boldsymbol{x}, y) \in \mathcal{D}\}$, while its label $y_{target} \neq y_{original}$.
- **Target Label** y_{target}: The label of $\boldsymbol{x}_{target}$.
- **Modified Image**: The modified image is visually similar to the selected image. It is obtained through minimizing the distance between the output of the image initialized with the selected image and that of the target image in the middle layer of a given pre-trained DNN. Note that its label is the same as that of the target image, $i.e.$, $y_{modified} = y_{target}$, therefore the relation within the image-label pair $(\boldsymbol{x}_{modified}, y_{modified})$ is distorted to hide the ground-truth relation within $(\boldsymbol{x}_{selected}, y_{selected})$.

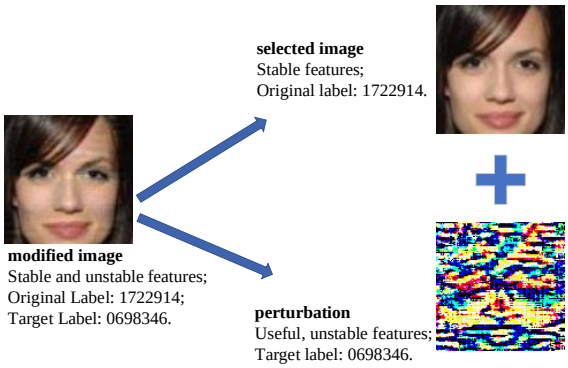


Fig. 2: The illustration of a modified image. The modified image contains two types of features, including stable features from the selected image and unstable features from the perturbation. The information about ground-truth mapping will be hidden in the unstable features associate with the target label.

As suggested in [14], the features used by DNNs can be divided into two categories, including stable and unstable features. Intuitively, stable features ($e.g.$, the profile [15, 16]) are visible and interpretable to humans, while the unstable features, such as the texture [11, 12], are usually invisible. Both stable and unstable features are all useful to the classifier. Since unstable features can be modified easily without being discovered by humans, these features can be utilized to construct a new dataset with distorted mapping, namely the **modified dataset** to hide to ground-truth mapping for privacy protection. The detailed construction procedure will be discussed in Section 2.2.

2.2. Mapping Distortion based 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 its modified version where the input-label mapping is distorted.

In this paper, we propose a simple yet effective method, dubbed mapping distortion based protection (MDP), for constructing the modified dataset. To obtain useful yet unstable features from the target image, we use a standard pre-trained DNN. Specifically, we first initialize the modified image with the selected image and then minimize the distance between its output and that of the target image in the feature space. 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.

Specifically, 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 $y_{original}$ ($y_{original} \neq y_{target}$) in $\mathcal{D}$. After that, the 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. The update is through the following optimization process:

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 optimization (1).
    $\mathcal{D}_{modAug} = \mathcal{D}_{modAug} \cup \{(\boldsymbol{x}_{modified}, y_{target})\}$.
  end
end
return $\mathcal{D}_{modAug}$

$$\boldsymbol{x}_{modified} = \underset{\boldsymbol{x} \in [0,1]^D}{\operatorname{argmin}}\, d\left(f(\boldsymbol{x}) - f(\boldsymbol{x}_{target})\right), \tag{1}$$

where D is the dimension of the features, f is the mapping from input to the output of a certain layer in DNN, and $d(\cdot)$ is a distance metric. We adopt the most widely used ℓ^∞ distance, *i.e.*, $d\left(f(\boldsymbol{x}_{modified}) - f(\boldsymbol{x}_{target})\right) = ||f(\boldsymbol{x}_{modified}) - f(\boldsymbol{x}_{target})||_\infty$, for simplicity. More different distance metrics will be discussed in our future work.

Specifically, to solve the optimization problem (1), $\boldsymbol{x}$ is initialized with $\boldsymbol{x}_{selected}$ and we optimize the problem with the classical projected gradient descent (PGD) method [17]. As shown in Fig. 2, the modified image is obtained from the combination of the selected image and a small perturbation related to unstable yet useful features. Since those invisible yet useful features contained in the images are still highly predictive, DNNs trained with the modified dataset can still have a good performance on the benign testing test.

2.3. Augmented Mapping Distortion based Protection

As mentioned in the previous section, we can store the modified dataset instead of the original one to protect data privacy. However, due to the adverse effects of the incorrect stable features in modified images, training with them will result in a certain decrease in the performance. In this section, we introduce an MDP extension, dubbed augmented MDP (AugMDP), to further enhance the effectiveness of our method.

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 positive integer hyper-parameter 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 data is conducive to the learning of the model. The construction procedure of $\mathcal{D}_{modAug}$ is shown in Algorithm 1, and its effectiveness is verified in Section 3.3.

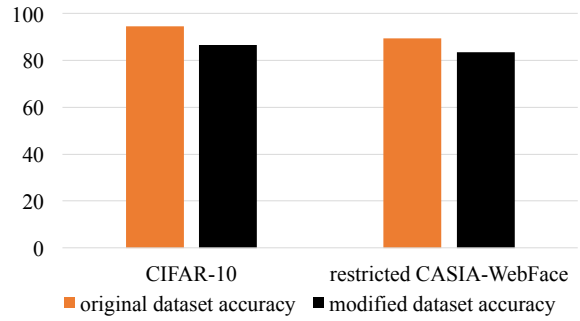


Fig. 3: Test accuracy on CIFAR-10, restricted CASIA-WebFace, and their corresponding modified dataset.

3. EXPERIMENTS

3.1. Settings

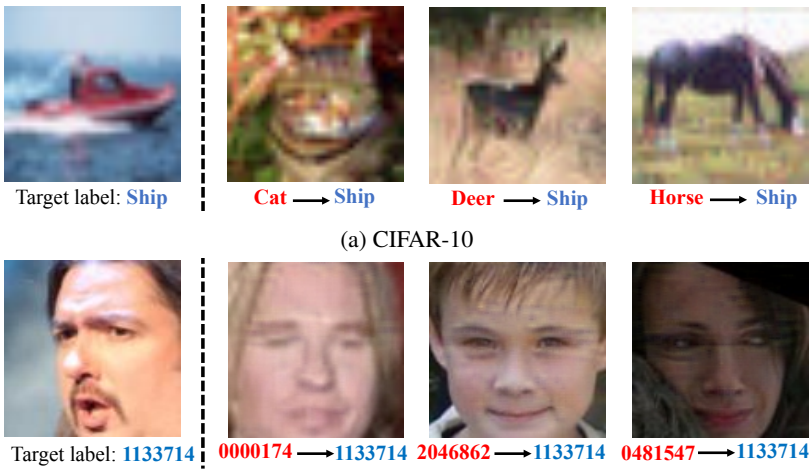
The experiments are conducted on the CIFAR-10 [18] and (restrict) CASIA-WebFace [19] dataset. Instead of the whole CASIA-WebFace, we only use a subset of the dataset for the consideration of computational complexity. Restricted CASIA-WebFace has 46,492 images with only 1,000 classes, which are randomly chosen from the original dataset. For DNNs, we use ResNet-50 [20] and IR-50 with ArcFace (an improved version of the vanilla ResNet-50) [21] on CIFAR-10 and restricted CASIA-WebFace datasets, respectively.

To construct the modified dataset, we perform PGD [17] to optimize the objective function under ℓ^∞ norm, which aims to minimize the distance between the output of the modified image and that 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 on CIFAR-10 and restricted CASIA-WebFace dataset, respectively. The performance of the trained model is examined on the benign testing set to verify the effectiveness of the modified dataset.

3.2. Verification on CIFAR-10 and CASIA-WebFace

In this experiment, we construct two datasets $\mathcal{D}_{mod-CIFAR10}$ and $\mathcal{D}_{mod-CASIA}$ based on 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 $\mathcal{D}_{mod-CIFAR10}$, while the right-hand side indicates the performance of the model trained on restricted CASIA-WebFace and $\mathcal{D}_{mod-CASIA}$. The result shows that DNNs trained on the modified dataset can generalize well on the benign testing set, and therefore the effectiveness of MDP is verified.

Fig. 4 illustrates some target images and their correspondingly modified images. To quantitively assess the similarity between the selected image and corresponding modified image, we also calculate their structural similarity index (SSIM) [22]. 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



(a) CIFAR-10

(b) restricted CASIA-WebFace

Fig. 4: Some examples of the target image and modified image on the CIFAR-10 and restricted CASIA-WebFace dataset. **The first column**: target images from the original datasets. **The next three columns**: correspondingly modified images generated by the MDP. The red and blue words indicate the original and the target label, respectively.

Table 1: The transferability evaluation on the $\mathcal{D}_{mod-CIFAR10}$ and $\mathcal{D}_{mod-CASIA}$ dataset.

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: Test accuracy of MDP and AugMDP on modified dataset constructed based on CIFAR-10 and restricted CASIA-WebFace dataset, respectively.

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

image, therefore the invisibility of modification is guaranteed.

3.3. Discussion

Augmentation Effects. To verify the effectiveness of the augmentation effects in AugMDP, we compare AugMDP and MDP on both CIFAR-10 and restricted CASIA-WebFace datasets. Table 2 shows the test accuracy of these two methods on two datasets, where the number in the parenthesis following AugMDP in the table is the value of augmentation-related hyper-parameter T. Particularly, AugMDP (1) is equivalent to MDP. As demonstrated in Table 2, AugMDP is better than MDP across different tasks even when T is relatively small ($i.e., T = 2$). Note that T should be adjusted according to specific requirements since AugMDP brings additional computation and storage costs.

Transferability. In this experiment, we verify whether the modified dataset generated by a given network is also effective for training other network architectures. Table 1 shows the test accuracy of several architectures (ResNet-50, ResNet-154, and DenseNet-154 [23]) trained on $\mathcal{D}_{mod-CIFAR10}$ generated by ResNet-50, and the test accuracy of IR-50, IR-152, and IR-SE-50[1] trained on $\mathcal{D}_{mod-CASIA}$ generated by IR-50. The result shows that the modified dataset is also effective for training different (especially similar) network architectures. It is probably because the unstable features learned by similar classifiers share certain similarities. Detailed reasons will be further explored in our future work.

4. CONCLUSIONS

In this paper, we propose the mapping distortion based protection (MDP) and its augmentation-based extension (AugMDP) to protect the visual data privacy in classification tasks by modifying the training set. 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, while still achieve good performance on the benign testing set when the model is trained on the modified dataset. Extensive experiments are conducted, which verify the feasibility and effectiveness of the proposed methods.

[1] IR-SE-50 combines IR-50 and SENet [24].

5. REFERENCES

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

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

[3] Syeda Mariam Ahmed and Chee Meng Chew, "Density-based clustering for 3d object detection in point clouds," in *CVPR*, 2020.

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

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.

[6] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho, "Meta-transfer learning for zero-shot super-resolution," in *CVPR*, 2020.

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

[8] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe, "Siam r-cnn: Visual tracking by re-detection," in *CVPR*, 2020.

[9] Bingyan Liao, Chenye Wang, Yayun Wang, Yaonong Wang, and Jun Yin, "Pg-net: Pixel to global matching network for visual tracking," in *ECCV*, 2020.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[16] 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.

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

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

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

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

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

[22] 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.

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

[24] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

View publication stats