First Steps in Pixel Privacy: Exploring Deep Learning-based Image Enhancement against Large-scale Image Inference

Zhuoran Liu, Zhengyu Zhao
Radboud University, Netherlands
{z.liu,z.zhao}@cs.ru.nl

ABSTRACT

In this paper, we present several enhancement approaches for the Pixel Privacy Task of MediaEval 2018. The goal of this task is to use image enhancement techniques to fool the state-of-the-art convolutional neural network (ConvNet) classifiers in scene classification problem, and maintain the visual appeal of images. Our proposed approaches are based on image crop, adversarial perturbations and style transfer, respectively. Firstly, we showed the potential influence of easy-to-use image processing operations, i.e., cropping (center cropping and random cropping). In perturbation-based approach, we apply a white-box technique, which makes use of the information of ConvNet classifiers. Based on the experiments, we observed some limitations of this approach, caused by, for example, image preprocessing. In addition, we demonstrated the style transfer-based approach, which was not developed for privacy protection, could be also used to reduce the effectiveness of the classifiers for Large-scale Image Inference. Specifically, we implement black-box techniques based on the Generative Adversarial Network. Experimental results showed that style transfer-based approach could address privacy protection and appeal improvement simultaneously.

1 INTRODUCTION

Multimedia data is generated every day and accumulated as large-scale datasets. Based on large-scale image data and the development of artificial intelligence, privacy-sensitive information, e.g., daily patterns and locations, can be efficiently inferred by some state-of-the-art techniques [5]. The objective of the Pixel Privacy Task of MediaEval 2018 is to protect privacy-sensitive scene images against large-scale image inference algorithms, and at the same time, maintain or even increase the visual appeal of the image.

Commonly used approaches to protection are based on hiding visible privacy-sensitive information of images. In the visual privacy task of MediaEval [2], some approaches are proposed to protect information in video sequences. For example, [6] proposes an approach based on false colour within the JPEG architecture to prevent revealing of sensitive information of video surveillance to viewers. Although these approaches can protect sensitive information in images, they are not applicable in the context of this task, because users would not like to share these social images, which have been blurred or changed obviously.

In the scenario of social images, we found two main categories of techniques can be used to protect privacy-sensitive scene images. One of them is based on generating adversarial examples. There are already some techniques which could generate adversarial examples, e.g., L-BFGS method [13], fast gradient sign method [8] and so on. These generated adversarial examples could fool the ConvNet-based classifiers to protect privacy information. The adversarial examples software library *cleverhans* [12] collects some construction techniques to generate adversarial examples. Given the condition that perturbation-based approach may need information from classifier and the resulting images may look not good. We propose to use style transfer approach to protect image privacy. This category of approaches protects sensitive information in images by transferring social images to another style, and at the same time , improves the image appeal. There are plenty of methods to do style transfer. By making use of image representations from ConvNets, [7] renders the semantic content of an image in different styles. Some generative models are also applied for style transfer, for example, conditional adversarial networks for image-to-image translation [9] and cycle-consistent adversarial networks (CycleGAN) for unpaired image-to-image translation [16].

In this paper, we explore these two categories of approaches to image privacy and image appeal, and show their effectiveness base on the experiments.

2 APPROACHES

In this section, we describe our perturbation-based approach and show the potential limitations of it. Then, style-transfer-based approach was proposed to achieve more effective protection and generate images with better quality with respect to human perception.

2.1 Perturbation-based approach

Our perturbation-based approach generates a fixed 2-d perturbation vector for each image. After adding this quasi-imperceptible vector to original images, the performance of the ConvNets-based classifier will decreased by a large margin. Our implementation of image perturbation refers to *Universal Adversarial Perturbation (UAP)* [10], which makes use of *DeepFool* [11] and generalizes well across different ConvNets. This approach follows a white-box setting. In other words, the calculation of perturbation need explicit information from both training dataset and the ConvNet model. Specifically, for each image in the dataset, it computes a minimal perturbation which sends the perturbed image to the decision boundary. Then, this perturbation will be updated iteratively with a constraint, e.g., $L_\infty \leq \xi$, to make the final perturbation as small as possible.

In our implementation, we calculate the perturbation vector on the basis of 3000 images from the validation data set provided by 2018 Pixel Privacy Task and a ConvNets model (ResNet50), which was pretrained on the Places-Standard dataset [15]. In the preprocessing step, we resize input images to 256 with respect to its short edge and keep the original ratio. Then we crop a 224 square in the

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

center of images. After training, we add the resulting perturbation vector in the resized test images. We summarized the potrntial limitations of UAP as follows. Firstly, in most practical cases of social images, the information of inference models and the traing set is not available. It is hard to generate optimal perturbation vector without this explicit information. Secondly, quasi-imperceptible artifacts added in the perturbed images are still not satisfying in the context of social images. In addition, the generated perturbation is vulnerable to image preprocessing [1]. Exploratory experiments showed potential influences of image preprocessing with cropping operations have potential influence on this approach.

With additional scaling and cropping of the perturbed images, the top-1 accuracy of the classification drops from 46.4% to 41.9%.

2.2 Style transfer-based approach

We propose a style transfer-based approach to protect image privacy and increase appeal. In particular, we apply GANs-based methods to change images to some certain styles, such as *Ukiyo-e* style with CycleGAN [16] and *Hayao* style with CartoonGAN [4]. Both of two above GAN-based methods are used for unpaired image-to-image translation. Given a source domain X (input images) and a target domain Y (styled images), a mapping $G : X \rightarrow Y$ is learned such that $G(X)$ is indistinguishable from Y. The objective of the learning is a summation of adversarial losses and cycle consistency loss. After training on source set and target set, we learn a mapping function G, which can transfer the style of any input images to a target style.

3 EVALUATION RESULTS

We submitted five runs for the Pixel Privacy Task of MediaEval 2018. Fig. 1 shows image examples enhanced by these five runs. In social multimedia, it is common that users crop images and videos to improve the appeal before sharing them. Due to different settings of ConvNet-based classifiers, image cropping may also have some influences on the classification. For example, in the preprocessing stage of the classification, the input images are scaled and cropped by default. So we submitted two runs based on central cropping and random cropping to explore potential influence of image cropping. In addition, we submit one run using our perturbation-based approach, and another two runs using our style transfer-based approach.

Table 1 presents evaluation results of our five runs in terms of Top-1 and Top-5 classification accuracy. We see that style transfer-based and image cropping approaches yield obvious decrease of accuracy compared with the original performance of the classifier. Perturbation-based approach shows less decrease, due to image preprocessing as we discussed in 2.1. In addition, the number of training images and selection of hyper-parameters in UAP may also influence the evaluation performance.

We also get aesthetics results of our submissions [3]. We also evaluate the aesthetic quality of the images enhanced by our runs, using *NIMA (Neural Image Assessment)* [14]. Hayao style with CartoonGAN shows the largest increase of mean aesthetics score (5.09), compared with the score (4.472) of original images.

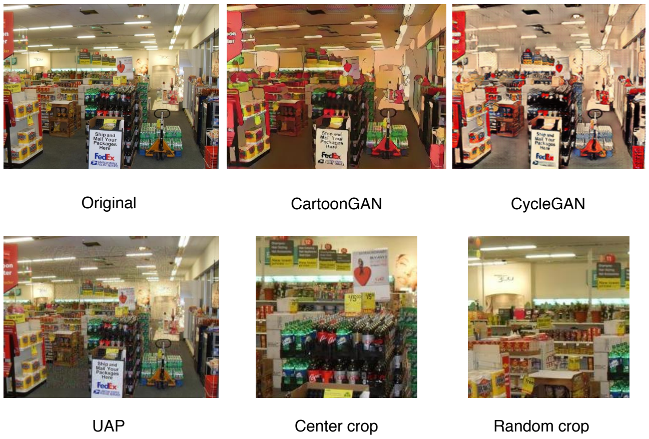
Original
CartoonGAN
CycleGAN
UAP
Center crop
Random crop

Figure 1: An example of resulting images by different protection approaches.

Table 1: Evaluation results in terms of Top-1 and top-5 prediction accuracy on the scene images from the MEPP18test dataset.

	Top-1 acc.	Top-5 acc.
Original	60.23%	88.63%
Hayao	41.57%	69.97%
Ukiyo-e	34.23%	62.00%
C-crop	39.33%	70.57%
R-crop	34.27%	63.17%
UAP	45.33%	76.97%

4 CONCLUSIONS AND FUTURE WORK

In this paper, we propose perturbation-based approach (white-box) and style transfer-based approach (black-box) to protect privacy and improve appeal in scene sensitive images. The proposed style transfer-based approach has a good evaluation performance in both image privacy and image appeal.

From the exploratory experiments and evaluation results, we find that the perturbation-based approach generally works well, but is vulnerable to image preprocessing. In addition, added perturbation vector decreases the image appeal. For style transfer-based approach, the prediction accuracy decreases significantly. In addition, Hayao style with shows an increase in appeal score by NIMA evaluation.

In the future, we will combine the proposed two approaches to simultaneously achieve effectiveness of privacy protection based on optimal computation of the perturbation and improve the aesthetic quality of images.

ACKNOWLEDGMENTS

This work is part of the Open Mind research program, financed by the Netherlands Organization for Scientific Research (NWO). The experiments were carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

REFERENCES

[1] Anish Athalye and Ilya Sutskever. 2017. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397* (2017).

[2] Atta Badii, Mathieu Einig, Tomas Piatrik, and others. 2014. Overview of the MediaEval 2013 Visual Privacy Task.. In *MediaEval*.

[3] Simon Brugman, Maciej Wysokinski, and Martha Larson. 2018. MediaEval 2018 Pixel Privacy Task: Views on image enhancement. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[4] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. 2018. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9465–9474.

[5] Jaeyoung Choi, Martha Larson, Xinchao Li, Kevin Li, Gerald Friedland, and Alan Hanjalic. 2017. The Geo-Privacy Bonus of Popular Photo Enhancements. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 84–92.

[6] Serdar Çiftçi, Ahmet Oğuz Akyüz, and Touradj Ebrahimi. 2018. A reliable and reversible image privacy protection based on false colors. *IEEE Transactions on Multimedia* 20, 1 (2018), 68–81.

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2014). arXiv:1412.6572 http://arxiv.org/abs/1412.6572

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint* (2017).

[10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1765–1773.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 2574–2582.

[12] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *arXiv preprint arXiv:1610.00768* (2018).

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[14] Hossein Talebi and Peyman Milanfar. 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[15] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.