Adversarial Photo Frame: Concealing Sensitive Scene Information of Social Images in a User-Acceptable Manner

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

ABSTRACT

Personal privacy protection has become more and more crucial in the era of big multimedia data and artificial intelligence. This paper presents our submission to pixel privacy task, where we propose to fool the deep visual classification model that is for recognition of sensitive scenes by adding adversarial frame to the image. Experimental results indicate that our method can achieve strong adversarial effects while maintaining the visual appeal and social function of the transformed images.

1 INTRODUCTION

Scene recognition is a hallmark topic in computer vision, and it provides global semantic information that facilitates different tasks. Leveraging large-scale scene datasets, deep learning-based scene recognition algorithms have made great progress in scene recognition [11]. But these algorithms also raised people's concerns on social multimedia privacy at the same time [6]. In the past, privacy protection algorithms mainly focused on leveraging adversarial machine learning, image style transfer [8] and image enhancement [1, 3]. Conventional adversary algorithms [2, 5] are effective for inducing misclassfication but not intentionally designed for increasing image appeal. In most cases, the resulting distortions even degrade the image quality. Methods of image enhancement and style transfer that are able to increase the visual appeal of images do not consider the adversarial function during the transformation process.

In the 2019 Pixel Privacy Task [9], we propose a baseline approach, called adversarial photo frame (APFrame). The approach strives for a balance between privacy protection and visual appeal of images. APFrame achieves privacy protection against network-based scene classifier based on adversarial machine learning techniques, while restricting the transformations to the edge of the image, i.e., the photo frame, in order to maintain the visual appeal. Experiments are conducted with different photo frame settings of APFrame. The algorithm details and experimental results are discussed in section 2 and section 3.

2 ADVERSARIAL PHOTO FRAME

The working diagram of APFrame is described in Figure 1. Given an image, a photo frame is generated randomly. We start from a white additive frame with all components equaling 1. This frame is fed into the classifier and the gradients of cross-entropy loss with respect to the original label is calculated by back propagation

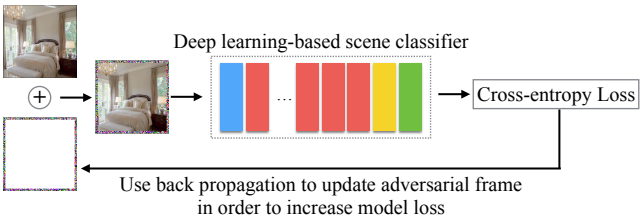

Figure 1: The working diagram for generating adversarial photo frame. The frame is randomly initialized and iteratively updated based on the classification loss. This process does not stop until the prediction of the classifier changes.

to update the frame. This process is repeated until the classifier outputs a different prediction.

Specifically, we consider two alternative constraints used for generating adversarial photo frames, namely, *full-flexibility* frame and *weight-constrained* frame. In the full flexibility APFrame, all values are admissible for pixels in the frame, and the resulting frame gives the impression of random couples (e.g., top row in Figure 2). In the weight-constrained frame, All the components in any of the three RGB color channels are required to change uniformly, i.e., only three parameters are learned for increasing the classification loss.

This setting enforces the adversarial frames to only have one color for more natural look. The width of the frame can be predefined to balance the protection effect and visual appeal. A narrow frame leads to less influence on the image content, but normally yields weaker protection effect.

The algorithm is summarized in Equation 1.

$$\begin{aligned} \underset{\boldsymbol{\delta}}{\text{minimize}} \quad & -\mathcal{L}(f(\boldsymbol{x}+\boldsymbol{\delta}), c_0) \\ \text{s.t.} \quad & c_0 \neq \underset{c}{\text{argmax}}(f(\boldsymbol{x}+\boldsymbol{\delta})), \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is the input image, which is correctly predicted into c_0 class, and $\boldsymbol{\delta}$ is the frame that only has values in the edge of the image with pre-defined width.

f represents the classifier, and $\mathcal{L}$ represents the cross-entropy loss function. The objective function is minimized until the predicted label of the classifier is different from the ground truth label. Obviously, narrow frame with the weight-constrained setting performs less effective due to the limited searching space of possible transformations.

Figure 2 shows some image examples achieved by APFrame.

3 EXPERIMENTS AND EVALUATION

We submit five runs to pixel privacy task for the official evaluation on our APFrame. Among them, three runs are using full-flexibility

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

Figure 2: Transformed images with different width of frame achieved by our APFrame. Images in top row (full-flexibility) and middle row (weight-constrained) look natural, while images in bottom row show the non-applicable cases of APFrame.

constraints with varied widths in the range [5,10,15], denoted as FfW5, FfW10 and FfW15. The other two runs are using weight-constrained settings with two different widths 15 and 20, denoted as WcW15 and WcW20.

We use Adam optimizer [4] with a learning rate of 1 to perform the gradient descent in Equation 1 on a Tesla P100 GPU.

Table 1: Evaluation results in terms of Top-1 accuracy and NIMA score on the scene images from the MEPP19test dataset for our five runs

	Top-1 acc. (%)	Aesthetics score.
Original	100.00	4.64
FfW5	0.00	4.72
FfW10	0.00	4.82
FfW15	**0.00**	**4.94**
WcW15	68.83	4.80
WcW20	64.67	4.82

Table 1 reports the evaluation results of the submitted five runs. The performance of privacy protection and visual appeal are quantified by top-1 accuracy and Aesthetics score obtained by NIMA [10], respectively. As shown, full-flexibility settings consistently outperform weight-constrained variants, by decreasing the top-1 accuracy of the classifier to 0.

In terms of visual appeal, we can observe that all the runs improve the original visual appeal. But different from our assumption, the best aesthetics result is also achieved by the full-flexibility variants. This may be due to the fact that NIMA was not really trained to be able to judge images with frames, so we cannot fully expect it can make decisions that correspond perfectly to human perception. Also, an aesthetics evaluation model trained with different kind of images may not be applicable to scene images.

On the other hand, the fact that NIMA scores are close to the original score is due to the high number of pixels are maintained from the original image.

4 DISCUSSION AND OUTLOOK

We proposed APFrame that can achieve adversarial effects against deep scene recognition networks for privacy protection, while maintaining image appeal. Compared with other adversarial machine learning-based techniques, APFrame maintains the visual appeal of images by avoiding modifications in the central part of images, which is arguably the part of the image most important to human viewers. In short, the social function and visual appeal of the images can be maintained to a large degree.

But it still has some disadvantages. For instance, it is not robust to preprocessing methods, e.g., center (random) cropping or resizing. The adversarial photo frames can be erased directly by center-cropping which is a common preprocessing step in deep learning-based models. To resolve this issue, developing the techniques that consider the semantics of image content is a direction to research in the future. APFrame can be extended to any shape that indicates different number of pixels in image, or implemented as QR code.

We also point out that due to the neural structure of NIMA, the generated adversarial frame patterns may have a transferable impact on the evaluation score [7]. This encourage us to have a non-neural evaluation scheme to better address human perception in the future.

ACKNOWLEDGMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

REFERENCES

[1] 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.*

[2] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.

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

[4] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR)*.

[6] Martha Larson, Zhuoran Liu, Simon Brugman, and Zhengyu Zhao. 2018. Pixel Privacy: Increasing Image Appeal while Blocking Automatic Inference of Sensitive Scene Information. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*

[7] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR).*

[8] Zhuoran Liu and Zhengyu Zhao. 2018. First Steps in Pixel Privacy: Exploring Deep Learning-based Image Enhancement against Large-scale Image Inference. In *Working Notes Proceedings of the MediaEval 2018 Workshop.*

[9] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 2019. Pixel Privacy 2019: Protecting Sensitive Scene Information in Images. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*

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

[11] 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* 40, 6 (2017), 1452–1464.