MediaEval 2019: Concealed FGSM Perturbations for Privacy Preservation

Panagiotis Linardos, Suzanne Little, Kevin McGuinness
Dublin City University
linardos.akis@gmail.com

ABSTRACT

This work tackles the Pixel Privacy task put forth by MediaEval 2019. Our goal is to manipulate images in a way that conceals them from automatic scene classifiers while preserving the original image quality. We use the fast gradient sign method, which normally has a corrupting influence on image appeal, and devise two methods to minimize the damage. The first approach uses a map of pixel locations that are either salient or flat, and directs perturbations away from them. The second approach subtracts the gradient of an aesthetics evaluation model from the gradient of the attack model to guide the perturbations towards a direction that preserves appeal. We make our code available at: https://git.io/JesXr.

1 INTRODUCTION

The Pixel Privacy task, introduced by MediaEval [7], aims at developing methods for manipulating images in a way that fools automatic scene classifiers (referred to as attack models). As an added constraint, the images should not exhibit a decrease in aesthetic quality. The organizers made available the Places365-Standard data set [11] along with a pre-trained ResNet [3] attack model for the task.

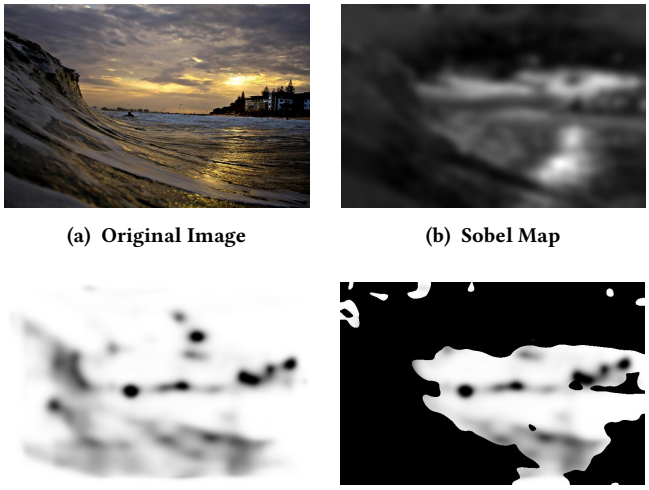
The contribution of image enhancement techniques in privacy protection has been previously explored [1], showing that even popular filters used in social media have a cloaking effect against geo-location algorithms. A more recent work by Liu et al. [6] proposed a perturbation-based approach (white-box) and a transfer style approach (black-box). Similar to the first module in that work, we propose two perturbation-based approaches and explore ways to localize the perturbations in a manner that does not reduce appeal.

2 APPROACH

We developed two approaches, both of which utilize FGSM [2]. FGSM uses the gradient of the attack model and changes the pixel values by nudging them towards the direction that maximizes the loss. Furthermore, the strength of these perturbations varies and is represented by the ϵ value.

2.1 Salient Defence

Our first approach combines two maps: one is a measure of saliency and the other a measure of flatness. Salient areas are the ones that are more likely to attract the eye of an observer, and are predicted by a DNN. In particular, we use SalBCE [5] trained on the SALICON dataset [4]. Furthermore, perturbations become more obvious when they are located in flat areas. For this reason, we also used a Sobel filter [9], which detects areas where edges are more prevalent. Gaussian blurring (σ=10) is applied to spread the detected edges, forming the final Sobel map. The saliency map is reversed so that the pixels corresponding to salient areas are zeroed out. Then, pixels where the mean value is below average on the Sobel map (hence more likely to be on a flat area) are also zeroed out. The resulting map M, the sign of the network's gradient g, and the value ϵ are multiplied and added to the original image I, completing the modification. Figure 1 illustrates an example of map generation.



(a) Original Image (b) Sobel Map (c) Reverse Saliency Map (d) Final Map

Figure 1: Maps used to constrain perturbations on less obvious areas. The reverse saliency map along with the Sobel map produce the final map.

$$I_{\text{modified}} = M \circ sgn(g(I)) \circ \epsilon + I \quad (1)$$

Additionally, we used a popular filter for image manipulation, namely tilt-shift to inspect how it affects the efficacy of our approach. Tilt-shift essentially blurs parts of the background while intensifying foreground. In our case we used the saliency maps as an estimate of the foreground to be intensified, blurring the rest.

2.2 Coupled Optimization

The second approach exploits the gradients of both the attack model and the aesthetics evaluation algorithm. The aesthetics evaluation in our case is the NIMA algorithm [10]. Since the networks differ significantly, the gradients are first scaled to be brought to the same range [0,1]. Afterwards, NIMA's gradient is subtracted from ResNet's and as a result we get the sign of the total gradient and

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

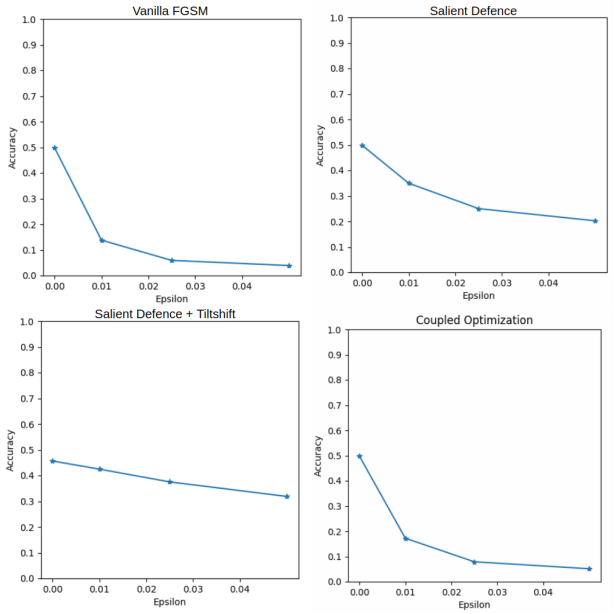


Figure 2: Attack model accuracy under differing values of ϵ.

multiply that by ϵ:

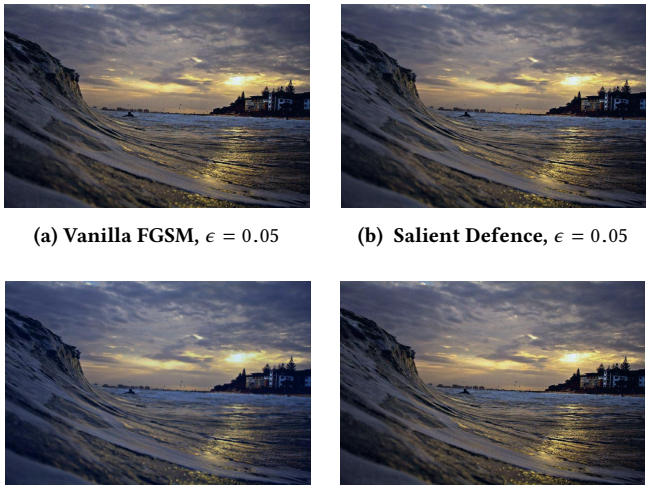
$$I_{\text{modified}} = sgn(g_{ResNet}(I) - g_{NIMA}(I)) \circ \epsilon + I \quad (2)$$

3 RESULTS AND ANALYSIS

In our initial experiments, we used the full-resolution images from Places365 and applied a variety of ϵ values to investigate how they affect the accuracy of the attack model (Figure 2). Salient Defence perturbs less pixels, which explains the lower impact on accuracy compared to the vanilla FGSM. We also note that the tilt-shift filter further reduces the efficacy of those perturbations. The coupled optimization approach, has a higher impact on the accuracy of the attack model, as it manipulates all the pixels of the image.

The test set, as evaluated by the MediaEval team (Table 1) was first downsampled to 256×256 and the algorithms were applied afterwards. Note that this set includes only images that ResNet predicts successfully, and so the initial accuracy ($\epsilon = 0$) is 100%. In that case it seems that the tilt-shift effect actually adds to the efficacy of the perturbations, bringing the accuracy of the attack model down while increasing the aesthetics score.

To test NIMA's sensitivity to perturbations, we used FGSM (vanilla) with a very high $\epsilon = 0.15$ on a small subset (100) of the validation images. This type of attack effectively ruins the visual appeal; however, the NIMA score drops by only a small amount (from 4.26 to 3.98). This indicates that NIMA has a low-sensitivity to adversarial perturbations. This could be explained by the fact that NIMA was trained on AVA [8], a dataset collected by photographers. The model is, therefore, sensitive to high-level concepts of aesthetic appeal, such as the rule of thirds, but has not been trained to be sensitive to the low-level corrupting influence of perturbations.



(a) Vanilla FGSM, $\epsilon = 0.05$ **(b) Salient Defence, $\epsilon = 0.05$**

(c) Salient Defence & tshift, $\epsilon = 0.01$ **(d) Coupled Optimization, $\epsilon = 0.05$**

Figure 3: The most promising configurations contrasted with the vanilla FGSM.

Methods	ϵ	Top-1 Acc.↓	NIMA Score↑
Salient Defence	0.01	0.937	4.63
	0.05	0.735	4.58
Salient Defence & tilt-shift	0.01	0.868	**4.75**
Coupled Optimization	0.01	0.917	4.63
	0.05	**0.458**	4.54
Original Test Set	-	1.0	4.64

Table 1: Results on MediaEval test set. Top-1 accuracy refers to the the prediction accuracy of the attack model (ResNet50 trained on Places365-standard data set). The *NIMA Score* column represents the average of the aesthetics scores.

4 DISCUSSION AND OUTLOOK

An obvious shortcoming of our salient defence algorithm is that saliency is subject to change after manipulations to the image. One way of improving this would be to predict the saliency of the perturbed image and reapply the modification on the original using this information. Also, the Sobel filter assigns gradients in the image such as that of the horizon as similar to edge-dense areas, resulting in a map where some flat areas are not obscured. Furthermore, we have shown that NIMA is not reliable when assessing the corrupting quality of low-level noise such as FGSM perturbations. We believe that aesthetic algorithms trained for low-level cues would improve the efficacy of our coupled optimization approach.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/15/SIRG/3283 and SFI/12/RC/2289

REFERENCES

[1] 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.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[4] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1072–1080.

[5] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Noel E O'Connor, Xavier Giro-i Nieto, and Kevin McGuinness. 2019. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869* (2019).

[6] Zhuoran Liu and Zhengyu Zhao. 2018. First Steps in Pixel Privacy: Exploring Deep Learning-based Image Enhancement against Large-Scale Image Inference.. In *MediaEval*.

[7] 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*.

[8] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[9] Irwin Sobel and Gary Feldman. 1968. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in* (1968), 271–272.

[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.