MediaEval 2020: Maintaining Human-Imperceptibility of Image Adversarial Attack by using Human-Aware Sensitivity Map

Zhiqi Shen[1], Muhammad Furqan Habibi[1], Shaojing Fan[1], Mohan Kankanhalli[1]
[1]National University of Singapore
dcsshenz@nus.edu.sg,furqan.habibi@u.nus.edu
dcsfs@nus.edu.sg,mohan@comp.nus.edu.sg

ABSTRACT

With the rapid rise of big data with developments in artificial intelligence, privacy has come under the spotlight. Adversarial attacks using image perturbation have recently been introduced to fool machines on pattern recognition tasks. They also have been successfully employed to protect privacy of images. However, only a few works consider the imperceptibility of perturbations for humans. This report presents our submission to the pixel privacy task, where we improve the imperceptibility of image perturbations by using a human-aware sensitivity map, while protecting image privacy via adversarial attack techniques.

1 INTRODUCTION

The Pixel Privacy task [7] of MediaEval aims to protect personal privacy by embedding human-imperceptible noise on images that fools the BIQA classifiers. The attack models use InceptionResNetV2 structure and are pre-trained on KonIQ-10k dataset. The organizers evaluated the performance in terms of success attack rate (accuracy) and imperceptibility of perturbation.

Prior work usually applies L_2 norm [1, 5, 6] to the loss function to improve the imperceptibility of perturbed images. However, L_2 norm only guarantees the overall noise to be small without considering the perceptual characteristics of regions. For example, observers will perceive differently when we add the same noise to a flat background versus a content-rich background. With this insight, we can apply a sensitivity map to the loss function that indicates which regions' changes are least sensitive to observers, so that the algorithms know where to add the noise. Recent works [2, 4, 11] published after our earlier work [9] do take human imperceptibility of perturbations into account. Unlike our deep learning-based method, most of them compute human imperceptibility based on texture information.

Our method is an optimization-based approach based on the CW attack [1]. We manipulate each input image's model logits to its target class. We then optimize the attack to minimize the loss function by modifying the input image. To improve human imperceptibility, we improve the loss function by integrating human sensitivity maps learned from [9]. Experimental evaluation indicates our approach achieves good results in terms of human imperceptibility.

MediaEval'20, December 14-15 2020, Online

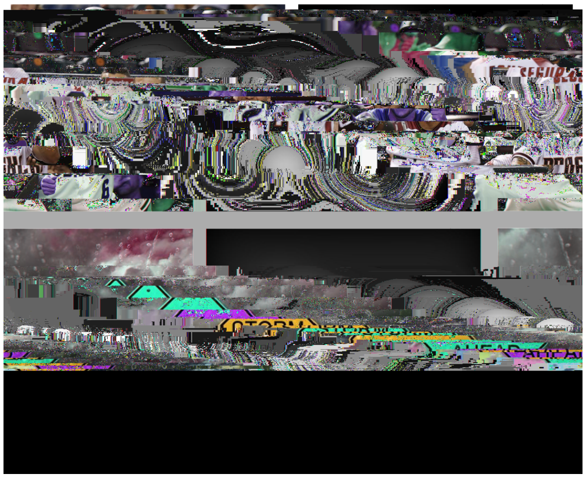

Figure 1: The figure shows sensitivity map examples. The left column has the original images and the right column are the corresponding sensitivity maps. For example, in the first image, its sensitivity map highlights the humans, indicating that noise added to the human region will be perceived more easily than when the noise is added to the background.

2 APPROACH

2.1 Preliminaries

We denote an image by $I \in \mathbb{R}^{H \times W \times C}$, where H, W, C is the frame height, frame width, and the number of channels, respectively. The BIQA classifier is denoted by $f(X, \theta) = logits$ that takes an image as input and produces the corresponding logits $\mathbb{R}^{K}$, which K is the class number. A softmax layer is followed to the network to transfer the logits to each class's probability y. The whole BIQA classifier is represented by $softmax(f(X, \theta)) = y$.

The image adversarial attack approach aims to find an image perturbation I_{adv} that maximizes the classification error. We denote $I_s = I_{adv} - I$ the adversarial image perturbation.

We propose an optimization-based approach. The general idea of generating perturbation for an image is by using the following optimization equation.

$$\arg\min_{I_s} \quad \alpha D(I_s) - \ell(f(I + I_s, \theta), \hat{l}) \tag{1}$$

where $D(.)$ is the perception regularization to keep the perturbation to be small and imperceptible to humans. $\hat{l}$ is the target logits. $\ell(., .)$ is the loss function to measure the difference between the actual prediction and the target prediction. To obtain a high attack

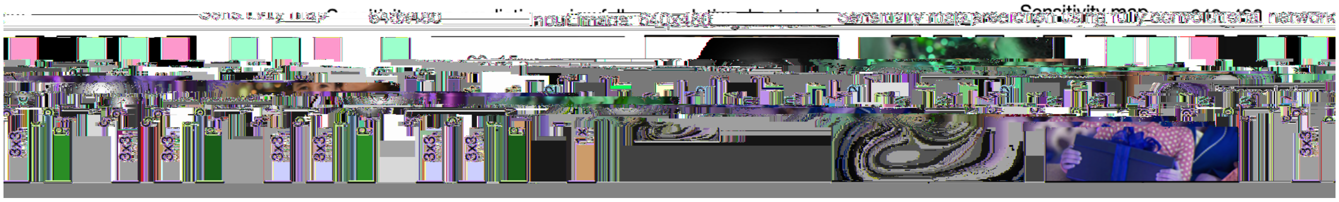

Figure 2: The figure shows the sensitivity map prediction network. The network bases on FCN network and use VGG-16 as its backbone network.

rate success, we minimize the distance between actual logits and the target logits. α is a hyper-parameter to balance these two terms.

2.2 Loss to fool machines

We follow the loss in [1] to fool machines. For the sake of clarity, we use $L_C = \ell¹f¹I ¸ I_s, \thetaº, \hat{l}º$, the detailed formulation is as follows:

$$L_C = \begin{cases} j max¹f¹I ¸ I_sºº - max¹\hat{l}ºj, & \text{if} \quad argmax\ f¹I ¸ I_sº \neq argmax\ \hat{l} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Where $f¹I ¸ I_sº$ and $\hat{l}$ are the one-hot vectors representing the current logits and desired logits. The losses consist of two parts. The first part represents the situations when the perturbed image has not been into our desired class. The loss value is the absolute distance between the most trusted class in current logits and the desired class. The second part depicts the situation when the perturbed image has been classified into our desired class, so we set the loss value to zero.

2.3 Loss to fool humans

We observed that the traditional norms (e.g., L_0, L_2, L_{inf}) consider all pixels in the images to be equal, while humans have different priorities when viewing different image regions. More specifically, even adding the same perturbation noise to different regions will lead to different humans' perceptibility. For quantifying humans' perceptibility of each pixel, we integrate a sensitivity map with our loss function. The value of each pixel in the sensitivity map ranges from 0 to 1. The larger value indicates more chance to be perceived when adding noise on such pixels.

Human-aware sensitivity map Human perception is a complex phenomenon which is not easily captured in a neat mathematical formulation. Therefore, we train a neural network to generate the spatially dense prediction of each pixel with human sensitivity scores. The network is designed based on a fully convolutional network (FCN) [8]. The backbone network is a VGG-16 [10] model pre-trained ImageNet dataset. A 1*1 convolutional layer is used to combine all feature maps extracted from VGG-16 to obtain the final sensitivity map. The architecture of our DNN is illustrated in Figure 2.

Embed sensitivity maps to attack approach For this workshop, we train the sensitivity map generation model on the EMOd dataset [3] and then test it on the given Place365 testing set. In order to integrate the human perceptual sensitivity, we extend the L2 norm by multiplying it with the sensitivity map, as shown below.

$$D¹I_sº = \|\beta_s I_s\|_2^2 \tag{3}$$

3 RESULTS AND ANALYSIS

We submitted five runs towards the Pixel Privacy task. The organizers selected 20 images with the largest BIQA variance for human evaluation. They then put the same image of all qualified runs in one folder and let 7 experts select the most appealing (i.e., "Best") three runs out of 17 runs. A run can be selected as "Best" for at most 140 times.

From Table 1, we can observe that the accuracy of our first run (with parameter $\alpha = 10$) has dropped to lower than random guess (50%), meaning that our perturbed images have fooled machines' prediction. More importantly, more than half of the images are selected as the best three images out of 17 runs. From the trend of parameters, we can see the potential of our algorithm. If we can try more parameters (e.g,. smaller than 10), the performance might be even better than the current one. For the other runs, we have not achieved a good attack rate. This is because the parameter α is too large that forces the perturbed images to focus more on image quality during back-propagation.

Table 1: The table shows the evaluation of our five runs. The first run with parameter $\alpha = 10$ has a high attack rate success with more than half of the perturbed images selected as best.

Parameter (α)	Accuracy	Number of times selected as "Best"
10	42.73	74
20	52.91	Not qualified
30	62.36	Not qualified
40	75.10	Not qualified
50	93.82	Not qualified

4 CONCLUSION AND FUTURE WORKS

This report introduces our approach for privacy protection, which integrates the human-aware sensitivity map to the loss function to improve the quality of perturbed images'. The results demonstrate the effectiveness of the sensitivity map in maintaining noise imperceptibility. However, some aspects can be further improved. The current sensitivity map prediction network is trained on the EMOd dataset, which has only 698 images. Another problem is that the network structure (FCN) is rudimentary. We can foresee that with a more sophisticated structure, trained on a larger data-set, can improve the performance.

REFERENCES

[1] 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.

[2] Francesco Croce and Matthias Hein. 2019. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4724–4732.

[3] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. 2018. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*. 7521–7531.

[4] Diego Gragnaniello, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. 2019. Perceptual Quality-preserving Black-Box Attack against Deep Learning Image Classifiers. *arXiv preprint arXiv:1902.07776* (2019).

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).

[6] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016).

[7] Zhuoran Liu, Zhengyu Zhao, Martha Larson, and Laurent Amsaleg. 2020. Exploring Quality Camouflage for Social Images. In *Working Notes Proceedings of the MediaEval Workshop*.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[9] Zhiqi Shen, Shaojing Fan, Yongkang Wong, Tian-Tsong Ng, and Mohan Kankanhalli. 2019. Human-imperceptible privacy protection against machines. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1119–1128.

[10] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[11] Eric Wong, Frank R Schmidt, and J Zico Kolter. 2019. Wasserstein adversarial examples via projected sinkhorn iterations. *arXiv preprint arXiv:1902.07906* (2019).