Maintaining perceptual faithfulness of adversarial image examples by leveraging color variance

Sam Sweere, Zhuoran Liu[1], Martha Larson[1]
[1]Radboud University, Netherlands
samsweere@gmail.com,z.liu@cs.ru.nl,m.larson@cs.ru.nl

ABSTRACT

With the popularity of social networks, large scale user-generated data is accumulated online. Possible misappropriation of these data may arise severe personal privacy problems. In this paper, we propose an image transformation method that protects images against scene classifier by exploiting the knowledge of adversarial examples. At the same time, our method maintains the perceptual faithfulness of protected images by leveraging color variance.

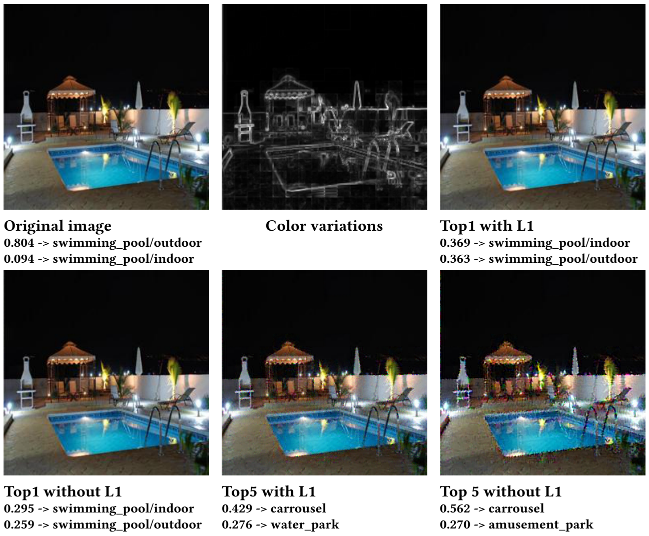
1 INTRODUCTION

Modern machine learning algorithms are able to extract privacy sensitive information from user-generated multimedia data that is available online, such as geo-location [1, 4]. The objective of the Pixel Privacy Task of MediaEval 2019 [5] is to find solutions that could protect privacy-sensitive information in images against scene classifier, and at the same time keep or increase the visual appeal of these adversarial images. The provided test-dataset is a subset of the Places365-Standard dataset [11]. This paper proposes a method to create adversarial examples by perturbing the pixels in the image based on color variation, which is better aligned with human perception.

2 RELATED WORK

Neural network-based algorithms have weaknesses that make them susceptible to adversarial examples [2]. By perturbing specific pixels in the image, the resulting outcome of the network can be changed without a substantial change to the image. The knowledge of adversarial examples can also be applied to protect privacy sensitive information. PIRE [6] is an iterative method to generate adversarial images, but it does not consider the human noticeability of these perturbed pixels. One solution to this problem is to consider the human perception of these perturbations. [7] suggests that perturbing pixels in low variance regions (i.e. white walls or blue skies) is more noticeable than perturbing pixels in high variance regions (i.e. a brick wall). In [7] this is implemented by calculating the standard deviation around a pixel based on its intensity (greyscale value). However, in this case, the color-specific information is discarded. The CV-PIRE method [9] suggests an approach that takes the human perception of colors into account, using high and low color variance regions.

Pixel color variance can be calculated by using CIEDE2000 [8] difference between a specific pixel and the surrounding pixels. The CIEDE2000 algorithm gives a better numerical distance between two colors based on human perception than for example the Euclidean distance. A Gaussian distribution can be used to weight the contribution of the surrounding pixels in order to smooth the color variance between two neighboring pixels. Pixels further away contribute less to the total color difference.



Original image
0.804 -> swimming_pool/outdoor
0.094 -> swimming_pool/indoor

Color variations

Top1 with L1
0.369 -> swimming_pool/indoor
0.363 -> swimming_pool/outdoor

Top1 without L1
0.295 -> swimming_pool/indoor
0.259 -> swimming_pool/outdoor

Top5 with L1
0.429 -> carrousel
0.276 -> water_park

Top 5 without L1
0.562 -> carrousel
0.270 -> amusement_park

Figure 1: Example of the protected images that meet different protective conditions. Top 2 predicted labels and probabilities are listed below each example.

3 APPROACH

In this section, we describe our perturbation-based protection algorithm in detail. In particular, we discuss the construction of model loss, hyper-parameter selection and training details.

3.1 Protection by perturbation

Pixel values of images can be perturbed in a certain way to influence the predicted label by the threat model when generating adversarial examples. Given the ground truth label, we could optimize the perturbations iteratively until the predicted class meets our adversarial condition (i.e. fall out of top 2 or top 5). To minimize the noticeability of the perturbations, we include a threshold that determines the maximum perturbation of one pixel. If we set the same threshold for every pixel in the image we may get very noticeable perturbations in low color variance regions, while the perturbations in high color variance regions are less noticeable. In our method, we follow the threshold method [9] in which CIEDE2000-based color variation is used to generate threshold map (e.g., second figure of top row in Figure 1).

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

Algorithm 1: Color Variance based perturbations

```
Input: Image $\boldsymbol{img}$ with ground truth label $\boldsymbol{c_0}$; Color variance
       matrix $\boldsymbol{cV}$; Classification model $\boldsymbol{f}$; Threshold
       multiplier $\boldsymbol{m}$; Weight of L1-loss $\boldsymbol{\alpha}$; Maximum iterations
       $\boldsymbol{T}$; Cross-entropy loss function $\boldsymbol{l_{ce}}$ and protection
       $\boldsymbol{condition}$;
Output: Generated adversarial example image;
Set the final per pixel threshold matrix;
$\boldsymbol{pT} = \boldsymbol{m} \cdot \boldsymbol{cV}$;
Set the initial random perturbations and iteration counter;
$\boldsymbol{v} = random(-0.01, 0.01) * \boldsymbol{pT}$;
$t = 0$;
while $t < \boldsymbol{T}$ do
    $logits = \boldsymbol{f}((\boldsymbol{img} + \boldsymbol{v}).clamp(0, 1))$;
    if $\boldsymbol{condition}$ is met then
        return $\boldsymbol{img} + \boldsymbol{v}$ ;
    $loss = -1 \cdot \underbrace{\boldsymbol{l_{ce}}(logits, \boldsymbol{c_0})}_{\text{Cross-entropy loss}} + \boldsymbol{\alpha} \cdot \underbrace{\frac{||\boldsymbol{v}||_1}{||\boldsymbol{pT}||_1}}_{\text{L1 norm loss}}$ ;
    Update the perturbations with an optimizer;
    $\boldsymbol{v} = \underset{\boldsymbol{v}}{argmin}\ loss$;
    Clip and round to stay within threshold and image
      relevance boundaries;
    $\boldsymbol{v} = \frac{round(clip(\boldsymbol{v}, -\boldsymbol{pT}, \boldsymbol{pT}) \cdot 255)}{255}$;
    $t = t + 1$;
Failed to generate an adversarial example in $\boldsymbol{T}$ iterations;
return False;
```

3.2 Color variance-based approach

Algorithm 1 demonstrates how the adversarial examples are generated. Here $\boldsymbol{cV}$ is the pixel color variance matrix for the image, $\boldsymbol{f}$ the classification model, in our case a ResNet-50 [3] classifier, $\boldsymbol{m}$ threshold multiplier that determines how big the actual threshold is relative to the color variance, $\boldsymbol{T}$ the maximum iterations, $\boldsymbol{\alpha}$ the weight of the L1-norm loss in the total loss function and $\boldsymbol{condition}$ the adversarial condition which will later be discussed.

Using the color variance matrix $\boldsymbol{cV}$ as in [9], we define the per pixel threshold that represents the maximum perturbation per pixel. The initial perturbations are randomly set based on this pixel threshold. We update the perturbations until the condition is met or the maximum amount of iterations is reached. Iteratively calculate the logits of the image with perturbations, the clamping is done such that image stays within the valid image range. If the condition is met based on the logits then we have a successfully adversarial example and we stop the loop. If not, we calculate the loss, this consists of the cross-entropy loss l_{ce}, where the goal is to minimize loss function given the ground truth label, and the L1 norm loss, which minimizes the total amount of perturbations. The perturbations are updated by back-propagation to minimize this loss. Finally, the updated perturbations are clipped to make sure they stay within the pixel value range of image and rounded such that the perturbations would remain when saved in a uint8 image format. If it is not possible to meet the condition within the maximum interactions we return *False*.

Table 1: Evaluation results for five submitted runs.

	Top-1 acc. (%)	Top-5 acc. (%)	Aesthetics score.	SSIM
Original	1.00	1.00	4.64	1.00
Top1-L1	0.00	0.99	4.63	0.9994
Top1-L1-50	0.00	0.93	4.62	0.9992
Top1-noL1	0.00	0.52	4.61	0.9975
Top5-L1	0.00	0.00	4.65	0.9614
Top5-noL1	0.00	0.00	4.81	0.8955

4 SUBMISSION RESULTS AND ANALYSIS

We submitted five runs for the Pixel Privacy task: Top1 with L1-norm (Top1-L1), Top1 with L1-norm and a 50 percent relative difference in prediction confidence compared to the top1 label (Top1-L1-50), Top1 without the L1-norm (Top1-noL1), Top5 with L1-norm (Top5-L1) and Top5 without the L1-norm (Top5-noL1). We noticed two potential flaws in the top1 with L1-norm, first average difference in the confidence between the ground-truth label and the new top1 label is often small and the new top1 label is often similar to the ground-truth label, this can also been seen in Figure 1. To counter these potential weaknesses we included the Top1-L1-50 run to increase the distance in confidence and the Top5-L1 and Top5-noL1 runs, that make sure the ground-truth label is not in the top5. We included the Top2-noL1 and Top5-noL1 runs since these need less computational resources.

Table 1 presents the results of the different conditions. As can be observed, all the adversarial images achieved their goal of top-1 or top-5. Following the official evaluation rule, Top5-noL1 achieves the best result. We also include the Structural similarity (SSIM) [10] score, this measures the perceptual difference between the original image and its adversarial counterpart, this could be interperted as how faithfull the adversarial image is to the original.

5 DISCUSSION AND OUTLOOK

The perturbations of especially the Top1 are small, as can also been seen in the SSIM score. This could cause the robustness of the adversarial images under image transformations such as compression or filters to be weak.

As can been seen in Table 1 there is a clear difference between the aesthetics score and SSIM. SSIM score could represent how noticeable the perturbations are, where in the Top1-L1 the perturbations are least to barely noticeable and the Top5-noL1 have the most noticeable perturbations. However, the aesthetics score is the highest on Top5-noL1, which could mean that the locations where the perturbations take place in our method are creating somewhat of an adversarial example to the aesthetics score method.

Further research could look at possible conditions where all the top predictions would be of a different scenes that are not closely related to the ground-truth scene. To increase the usability of this approach in practice, the robustness of protected images should be improved. Data augmentation of original image and ensemble training against different threat models can be considered in the future.

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 *ACM International Conference on Multimedia Retrieval (ICMR).* ACM, 84–92.

[2] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR).*

[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 (CVPR).* 770–778.

[4] Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman, and Gareth JF Jones. 2011. Automatic tagging and geotagging in video collections and communities. In *ACM International Conference on Multimedia Retrieval (ICMR).* ACM.

[5] 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.*

[6] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 2019. Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR).* ACM.

[7] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. 2018. Towards imperceptible and robust adversarial example attacks against neural networks. In *AAAI Conference on Artificial Intelligence (AAAI).*

[8] M Ronnier Luo, Guihua Cui, and Bryan Rigg. 2001. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application* 26, 5 (2001), 340–350.

[9] Sam Sweere. 2019. *Increasing the Perceptual Image Quality of Adversarial Queries for Content-based Image Retrieval.* Bachelor Thesis. Radboud University Nijmegen, the Netherlands. Available at: https://github.com/SamSweere/CV-PIRE.

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[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.