Image Enhancement and Adversarial Attack Pipeline for Scene Privacy Protection

Muhammad Bilal Sakha[1]
[1]Habib University, Pakistan
mbilal.sakha@gmail.com

ABSTRACT

In this paper, we propose approaches to prevent automatic inference of scene class by classifiers and also enhance (or maintain) the visual appeal of images. The task is part of the Pixel Privacy challenge of the MediaEval 2019 workshop. The fusion based approaches we propose apply adversarial perturbations on the images enhanced by image enhancement algorithms instead of the original images. They combine the benefits of image style transfer/contrast enhancement and the white-box adversarial attack methods and have not been previously used in the literature for fooling the classifier and enhancing the images at the same time. We also propose to use simple Euclidean transformations which include image translation and rotation and show their efficacy in fooling the classifier. We test the proposed approaches on a subset of the Places365-standard dataset and get promising results.

1 INTRODUCTION

Social media users unintentionally expose private information when sharing photos online [12], such as locations a user visited etc., which can be automatically inferred by state of the art methods [2]. The focus of Pixel Privacy task of MediaEval 2019 workshop [10] is to protect user uploaded multimedia data online. The task objective is to use image transformation algorithms for blocking the automatic inference of scene class by convolutional neural network (ConvNet) based ResNet50 classifier [6] trained on Places365-standard dataset [15]. The proposed methods should also either increase (or maintain) the visual appeal of an image. Additional details of the task can be found in [10].

We propose to combine image style-transfer and image enhancement with adversarial image perturbations to increase the visual appeal of the images, in addition to blocking the automatic inference of scene class information by the classifier. We also apply white-box (where the attacker has access to the model's parameters) adversarial perturbations alone to compare the performance to the fusion based approaches. Finally, we use simple euclidean operations like image translation and rotation to show how they are also able to fool the classifier. The proposed approaches are evaluated on the basis of reduction in the top-1 classifier accuracy and Neural Image Assessment (NIMA) [13] score is used to evaluate the image quality of the transformed images. The motivation behind proposing fusion based approaches is to incentivize the social media users to use such methods for not only protecting the privacy-sensitive information in the photos, but also to enhance their photos as an added bonus.

2 APPROACHES

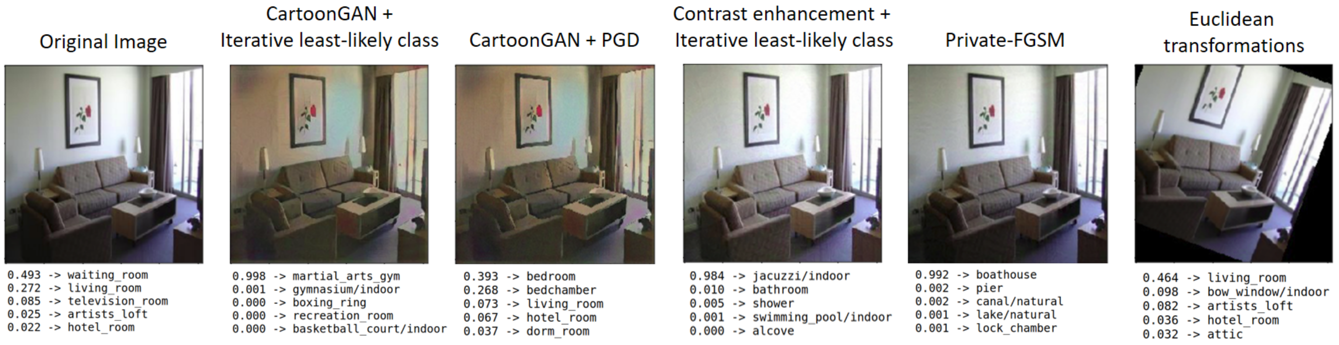
2.1 Fusion based approaches

CartoonGAN style transfer and Iterative least-likely class adversarial attack: In the first approach, we use an image style transfer method based on Generative Adversarial Networks (GANs) [4] called CartoonGAN [1], which enhances the image by applying cartoon style effects. On these enhanced set of images, we then apply a white-box targeted adversarial attack called the Iterative least-likely class method [7], which is a variant of the Fast Gradient Sign Method (FGSM) proposed by [5]. The Iterative least-likely class method tries to make an adversarial image by adding noise to the clean image, so that it will be classified as the class with the lowest confidence score for clean image. For choosing optimal ϵ (limit on the perturbation size), instead of doing binary search on each example because of the computational expense, we choose the value of ϵ to be $8/255$ on the basis of experimental results on a subset of validation set images. When enhancing the images using CartoonGAN, Hayao style is chosen because it results in the largest increase of mean aesthetic score among different CartoonGAN styles on the validation images.

CartoonGAN style transfer and PGD: In a slightly modified version, we now apply Projected Gradient Descent (PGD) [11] adversarial attack after enhancing the images with CartoonGAN style transfer. Here, we apply an untargeted adversarial attack, unlike in the previous method where the target class is the least-likely class of clean image. For the PGD adversarial attack, we chose the value of ϵ to be $2/255$ and the stepsize is chosen as $1/\epsilon$ on the basis of empirical results on a subset of validation images.

Image contrast enhancement & Iterative least-likely class: In this approach, we first enhance the contrast of the images using the method proposed by [14] and then perturb the enhanced images using the Iterative least-likely class adversarial method [7]. The reason for applying image processing to enhance the images initially is because the adversarial perturbation methods, reduce the visual appeal of the images, so enhancing the visual appeal of the images before applying adversarial perturbations will not only result in better performance on image quality metrics, but may also incentivize users to use this method over adversarial perturbations alone. In the image contrast enhancement approach by [14], the input image is fused with the synthetic image, which is obtained by finding the best exposure ratio to well-expose the under-exposed regions in the original image. Both the images are then fused according to the weight matrix, which is designed using illumination estimation techniques and the output is the contrast enhanced image. On these enhanced set of images, we then apply the Iterative least-likely class method, with the same parameters values as mentioned in the first approach.

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


Original Image
CartoonGAN + Iterative least-likely class
CartoonGAN + PGD
Contrast enhancement + Iterative least-likely class
Private-FGSM
Euclidean transformations

```
0.493 -> waiting_room
0.272 -> living_room
0.085 -> television_room
0.025 -> artists_loft
0.022 -> hotel_room
```

```
0.998 -> martial_arts_gym
0.001 -> gymnasium/indoor
0.000 -> boxing_ring
0.000 -> recreation_room
0.000 -> basketball_court/indoor
```

```
0.393 -> bedroom
0.268 -> bedchamber
0.073 -> living_room
0.067 -> hotel_room
0.037 -> dorm_room
```

```
0.984 -> jacuzzi/indoor
0.010 -> bathroom
0.005 -> shower
0.001 -> swimming_pool/indoor
0.000 -> alcove
```

```
0.992 -> boathouse
0.002 -> pier
0.002 -> canal/natural
0.001 -> lake/natural
0.001 -> lock_chamber
```

```
0.464 -> living_room
0.098 -> bow_window/indoor
0.082 -> artists_loft
0.036 -> hotel_room
0.032 -> attic
```

Figure 1: Original sample image from the Places356-standard dataset and the images transformed using different approaches with their corresponding top-5 classifier predictions.

2.2 White-box Private-FGSM adversarial attack

In order to compare the adversarial image perturbations with previous fusion based approaches, we use a more powerful variant of FGSM method, called Private-Fast Gradient Sign Method (P-FGSM) recently proposed by [8]. The values of ϵ and σ used for this method are set to 8/255 and 0.99 respectively.

2.3 Euclidean transformations

Inspired from the center crop and random crop operations in [9] to fool the classifier, we choose to explore other simple geometric operations on images, which are often overlooked in favor of adversarial attacks to fool the classifier. We consider two basic euclidean transformations i.e. image translation and rotation. To choose the optimal translation and rotation value to fool the classifier, we use the robust optimization method proposed by [3], instead of the computationally expensive grid-search. For majority of the images, we constrain translation to be within 20% of image size in each spatial direction and rotation up to 20°, and fill the resulting empty image spaces with zero pixel value.

3 RESULTS AND EVALUATION

In the Pixel Privacy task of the MediaEval 2019 workshop, the participants are allowed to submit five runs for the task, which are evaluated on the basis of top-1 classification accuracy (lower is better) and NIMA score [13] (higher is better), as shown in Table 1. Figure 1 shows the original image and the transformed images by different approaches and the corresponding top-5 class prediction.

Fusion based approaches: The performance of CartoonGAN + Iterative least-likely class adversarial method is good in terms of the top-1 accuracy, however it has the worst NIMA score of 4.37 among all runs. CartoonGAN + PGD adversarial method has the best NIMA score of 4.77 among all runs, but considerably higher classifier accuracy of 14%, which it is still less than 50%.

For Contrast Enhancement + Iterative least-likely class run, we get the lowest 0% top-1 accuracy and 4.47 NIMA score. The images enhanced using contrast enhancement method [14] look visually more appealing to the naked eye, however the NIMA score after applying only contrast enhancement is still slightly less than that of the clean images which is unexpected.

Private-FGSM adversarial attack: Private-FGSM attack reduces the top-1 classifier accuracy to 0%, at the cost of added noise in the submitted images, which is reflected in the reduced NIMA score of 4.49. Private-FGSM attack and previously used Iterative least-likely class methods are bounded by l_∞ norm, which results in small noise evenly distributed in the image, as can be seen by zooming the transformed images in Figure 1.

Table 1: Accuracy and NIMA score of different approaches

Run/Method	Accuracy	NIMA score
1. CartoonGAN + least-likely class	0.167%	4.37
2. CartoonGAN + PGD	14%	**4.77**
3. Contrast Enh. + least-likely class	**0%**	4.47
4. Private-FGSM	**0%**	4.49
5. Euclidean transformations [1]	6.667%	4.42
Original test images	100%	4.64

[1] Euclidean transformations evaluated on a smaller subset of test dataset consisting of 60 images.

Euclidean transformations: The final run of euclidean transformations which consists of translation and rotation operations achieves 6.667% top-1 classifier accuracy with a reasonable NIMA score of 4.42. For each image, finding the optimal translation and rotation value to fool the classifier is computationally expensive due to number of random transformations, therefore we test this approach on smaller subset of test dataset consisting of 60 images called test_manual, provided by the task organizers.

4 CONCLUSION AND OUTLOOK

In this paper, different approaches have been proposed for the Pixel Privacy task of MediaEval 2019 workshop. The fusion based approaches combining style transfer/image enhancement with adversarial attacks are chosen to increase the image appeal score beforehand, as reducing the classifier accuracy through adversarial perturbations decrease image appeal score, due to addition of noise.

In future, increasing image appeal by using the state of the art deep learning based image enhancement methods for image denoising, color/contrast/exposure adjustment etc. and then applying adversarial perturbation in our opinion will yield better results.

REFERENCES

[1] 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.

[2] 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.

[3] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. Exploring the Landscape of Spatial Robustness. In *International Conference on Machine Learning*. 1802–1811.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

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

[6] 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.

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

[8] C. Y. Li, A. S. Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro. 2019. Scene Privacy Protection. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Brighton, UK.

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

[10] 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*.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[12] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE International Conference on Computer Vision*. 3686–3695.

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

[14] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. 2017. A new image contrast enhancement algorithm using exposure fusion framework. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 36–46.

[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).