Synthesis(Generation) of adversarial (image) datasets to prove null hypothesis on DL systems (pick any SOA: OR, FR, Segmentation)

Computer Vision (CS6350)

TPA-12

1. Problem Statement

An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to misclassify it. In many cases, these modifications can be so subtle that a human observer does not even notice the modification at all, yet the classifier still makes a mistake. Refer [1] for more details about the concept of Adversarial Attack. In this problem, you are required to create a process/method/deep learning network to create adversarial examples for a state-of-the-art (SOTA) method for Image Classification and/or Object Recognition and/or Face Recognition and/or Image Segmentation.

2. Input

- SOTA deep learning methods for one or more tasks mention in problem statement (for instance: InceptionV3)
- Original Dataset that would be modified for generating adversarial samples

3. Expected Output

- A method/algorithm to generate samples that causes the pretrained SOTA to fail. Please note that the method to generate the samples should be generic such that all samples generated using it cause the pretrained network to fail.

4. Sample Input/Output

Figure 1: Target (Input Image)

Figure 2: Output of figure 1

Figure 3: Target (input)

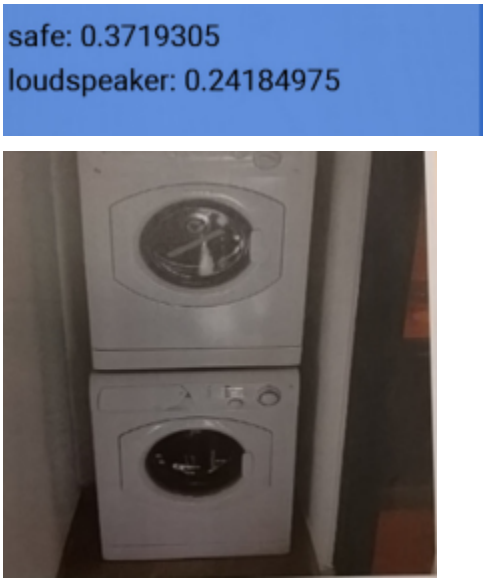

Figure 4: Output of figure 3

Generated Adversarial examples which are realistic but capable of fooling the Deep learning network

5. References

[1]. "Adversarial examples in the physical world" by Kurakin, Alexey, Ian Goodfellow, and Samy Bengio, In The International Conference on Learning Representations, Workshop, is, 2017.

[2]. "Physically realizable adversarial examples for lidar object detection" by Tu, James, et al. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[3]."Adversarial camouflage: Hiding physical-world attacks with natural styles", Duan, Ranjie, et al. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[4] "Benchmarking adversarial robustness on image classification". Dong, Yinpeng, et al. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[5] Wang, Zhibo, et al. "Feature importance-aware transferable adversarial attacks." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021

[6] Zhang, Jianping, et al. "Improving Adversarial Transferability via Neuron Attribution-Based Attacks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[7] Wang, Xiaosen, and Kun He. "Enhancing the transferability of adversarial attacks through variance tuning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021..

Aug, 2022