COMS 6730 Group-8 Project Abstract: Projected Gradient Descent Attack for Human Count Detection Error in Images on YOLOv8 Model

Group Members: Shuairan Chen, Zhou Fan, Karthik Hanumanthaiah

1. Problem Statement

In AI cybersecurity, deep learning models are relatively fragile and susceptible to adversarial attacks. These attacks exploit the model training process to introduce tiny, usually imperceptible perturbations to images, causing the model to misclassify images[4]. The human detection models, such as YOLO, perform well in most normal conditions. But It could be influenced by the attacks to incorrectly identify parts of an image where there are no people as people, or it could overestimate the number of people in the entire image. Existing research about Projected Gradient Descent (PGD) Attack[3] mainly focuses on defending against adversarial attacks or targeting misclassification, and it has limited effect on defensive models such as adversarial training[1] and defensive distillation[7]. Attacks such as Dense Adversary Generation (DAG)[9] and Universal Adversarial Perturbations (UAP)[6] typically focus on misclassifying existing objects or hiding them from detection and are less effective against models that employ defenses such as adversarial training or feature compression. In addition, the attack methods mentioned above are not specifically designed to induce false positives that detect humans when they are not present or overestimate the number of people, so we study and explore the vulnerabilities of this part of the task.

2. Goals

The project aims to enhance the PGD attack method to generate adversarial perturbations which will cause the YOLOv8 human detection model to misestimate the number of humans in an image. The attack will achieve two goals: (1) fooling the model to detect people in the images where no humans are present and (2) increasing the number of detected people in the images with humans. Current attacks typically focus on misclassifying objects as something other than people or hiding them from detection and are less effective against models that employ defenses such as adversarial training. The expected result is that the PGD attack can continuously deceive the human detection model in various situations, causing it to overestimate the number of people in the current image being processed, not just misclassifying other objects as people.

3. Methodology

We will use PGD to generate adversarial perturbations[5] to mislead YOLOv8 person detection. We will design a loss function for the human detection task to replace the regular loss function in the PGD attack CrossEntropyLoss[2, 3]. For the YOLOv8 model, the performance metrics have the box loss, class loss, the confidence and precision. So we will try to design the loss function that reduces the confidence score of the existing human area, and try to maximize the confidence of the background area which does not have humans, and amplify the bounding box prediction to increase the number of detected people. The difference with the general PGD attack, our method will try to combine the position and confidence of the detection box and the category probability in the loss function, so it is different from the loss function used in the standard PGD attack. Consequently, the misled YOLOv8 model will produce positive classification scores for the background areas without people not close to zero or generate more false localization boxes based on existing people. Our attack will limit the perturbations, such as the range from 0.05 to 0.3, to ensure that it is obvious to humans while affecting the model's judgment of the number of people in the image as much as possible. The novelty of our method is misleading the YOLOv8 human detection model by specially designed the loss function with the YOLOv8 performance metrics and using PGD to generate adversarial samples, causing it to overestimate the number of people in the image, which has not been explored in previous studies.

4. Evaluation

We will evaluate our attack efficiency on standard human detection datasets, such as COCO format, using YOLOv8 as the detection model. If the attack method we tried is effective, we will add defense-related tests to detect whether our attack method can effectively resist the two common defense methods of adversarial sample training and defensive distillation. The evaluating metrics include the False Positive Rate (FPR), the number of non-human images where YOLOv8 detects humans, the Over-detection Rate, and the number of extra people detected in the images that contain humans. The attack efficiency will be quantized by comparing the misguided and original models using such evaluating metrics. We expect a significant rise in false positives and over-detections under adversarial conditions.

5. Impact

This project will verify the vulnerability of YOLOv8 and human detection models to PGD attacks, and attempt to use this attack to mislead the model's prediction results[8]. This approach may affect certain functions of using the model to identify the location of people, such as overestimating the number and location of people, causing the AI program to stay in one state forever and unable to take the next action, or causing the model's focus target to constantly change due to too many people, in order to prevent the AI model from handling tasks related to human detection in the AI not allowed conditions. Our project can be used not only in attack but also in defense. We try to propose the need for improving the defense mechanism of this PGD attack in AI systems used for human detection in some key applications, such as autonomous driving, automatic detection of human position, etc.

References

1. Athalye, A., Carlini, N., & Wagner, D. (2018, July 31). *Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples*. arXiv.org. https://arxiv.org/abs/1802.00420
2. Chen, S.-T., Cornelius, C., Martin, J., & Chau, D. H. (2019, May 1). *Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector*. arXiv.org. https://arxiv.org/abs/1804.05810
3. Deng, Y., & Karam, L. J. (2020). Universal adversarial attack via enhanced projected gradient descent. *2020 IEEE International Conference on Image Processing (ICIP)*, 1241–1245. https://doi.org/10.1109/icip40778.2020.9191288
4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015, March 20). *Explaining and harnessing adversarial examples*. arXiv.org. https://arxiv.org/abs/1412.6572
5. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019, September 4). *Towards deep learning models resistant to adversarial attacks*. arXiv.org. https://arxiv.org/abs/1706.06083
6. Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., & Frossard, P. (2017, March 9). *Universal adversarial perturbations*. arXiv.org. https://arxiv.org/abs/1610.08401
7. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, March 14). *Distillation as a defense to adversarial perturbations against Deep Neural Networks*. arXiv.org. https://arxiv.org/abs/1511.04508
8. Wu, H., Yunas, S., Rowlands, S., Ruan, W., & Wahlström, J. (2023). Adversarial detection: Attacking object detection in Real time. *2023 IEEE Intelligent Vehicles Symposium (IV)*, *28*, 1–7. https://doi.org/10.1109/iv55152.2023.10186608
9. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., & Yuille, A. (2017). Adversarial examples for semantic segmentation and object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2017.153