Enhancement Of You Only Look Once Version 5 (Yolov5) Algorithm Applied in Nudity Content Detection for Video Chat Platforms

Nicole Angelica Junio[1], Daniella Sayson[1]
[1]*College of Information Systems and Technology Management, Pamantasan ng Lungsod ng Maynila, Manila, Philippines*
Corresponding Author: saysondaniella.ds24@gmail.com

***Abstract*: The increasing prevalence of adversarial attacks poses significant challenges to object detection algorithms, including YOLOv5, which is widely used for its speed and accuracy in real-time applications. Adversarial manipulations, involving subtle and often imperceptible changes to input data, can result in severe misclassification or complete detection failure. This study investigates YOLOv5's vulnerabilities to such attacks and proposes a Feature Squeezing Defense Method as a solution. The defense method incorporates Big Depth Reduction and Median Filtering techniques to effectively suppress adversarial perturbations while preserving the integrity of key object features. Extensive simulations revealed the substantial impact of adversarial attacks on YOLOv5's performance, with detection confidence plummeting from 0.897 to 0.250 and Intersection over Union (IoU) reducing to 0.02, signifying critical failures in spatial alignment and classification accuracy. Following the implementation of the defense mechanisms, the model achieved significant recovery, with detection confidence improving to 0.904 and IoU alignment increasing to 0.94. Additionally, the proposed method enhanced the model's resilience, minimizing adversarial impact and ensuring reliable performance under hostile conditions. These findings highlight the importance of robust preprocessing techniques in addressing adversarial vulnerabilities and safeguarding object detection models. This research contributes to the development of secure and efficient systems for real-time applications, emphasizing their role in ensuring reliable and accurate performance across critical domains.**

***Keywords*: YOLOv5, Real-Time Applications, Adversarial Attacks, Object Detection, Feature Squeezing Defense Method, Big Depth Reduction, Median Filtering.**

1. Introduction

The YOLO (You Only Look Once) algorithm stands as a groundbreaking development in the field of computer vision, particularly in object detection. Recognized for its single-stage architecture, YOLO revolutionized how images are processed by enabling entire images to be analyzed in a single pass through a convolutional neural network (CNN). This innovation allowed for rapid and accurate multi-object detection, making YOLO a preferred choice in various real-time applications. Subsequent iterations, such as YOLOv2 and YOLOv3, further refined the algorithm by introducing techniques like batch normalization, anchor boxes, and feature pyramid networks (FPN), which improved both accuracy and speed. YOLOv4 and YOLOv5 continued this trajectory of advancement, incorporating CSPDarknet and PANet to optimize the balance between speed and accuracy, along with integrating computational units like CSPNet and ELAN variants to enhance efficiency. Despite new detectors like RT DETR, the YOLO series remains a leading choice due to its balance of speed, accuracy, and adaptability. On May 23, 2024, YOLOv10 was released, marking the latest advancements in this domain. This study builds on YOLOv5, incorporating new techniques for enhancement.

YOLOV5, introduced in 2020 by Ultralytics, the developers behind YOLOv3, is built on the PyTorch framework, which simplifies support and deployment. This model is known for its speed, user-friendliness, and state-of-the-art performance in object detection tasks. Its superior accuracy and ease of training make it a favored choice among developers (Augmented A.I., 2023). YOLOv5 offers five model sizes—YOLOv5n (nano), YOLOv5s (small), YOLOv5m (medium), YOLOv5l (large), and YOLOv5x (extra-large)—each tailored to different computational needs, model sizes, and levels of complexity (Hussain, 2024). Being nearly 90 percent smaller than YOLOv4, YOLOV5 is more suitable for deployment on embedded devices (Nelson & Solawetz, 2020). Additionally, YOLOV5 models train extremely quickly, reducing experimentation costs during development. The model supports inference on individual images, batch images, video feeds, and webcam ports (Solawetz, 2020). Its capability is renowned for its real-time object detection and image classification capabilities, offering exceptional speed and accuracy.

Manuscript revised December 16, 2024; accepted December 18, 2024. Date of publication December 20, 2024.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898; SJIF: 5.59

Despite its many advantages, recent studies show that deep learning models, particularly CNNs, are vulnerable to adversarial attacks, where small, barely noticeable changes to input images can lead to inaccurate predictions. These attacks undermine the reliability of object detection systems, particularly in critical areas such as traffic and transportation. Notably, YOLOv5 has been found to be especially susceptible, with misclassification rates increasing as perturbations grow (Jain, 2023). This limitation underscores the need for enhancements to better achieve the objectives of the study.

Consequently, the study will focus on optimizing and refining the YOLOv5 algorithm to effectively address the identified challenge. Specifically, it will implement an adversarial defense method to protect the model against malicious inputs, ensuring robustness against adversarial attacks. This optimization aims to enhance content filtering accuracy, particularly to protect users, especially minors, from inappropriate content. By mitigating false positives, the study seeks to improve the security and efficiency of YOLOv5, ensuring real-time accuracy while safeguarding users from inaccuracies and potential adversarial attacks.

2. Methodology

This study focuses exclusively on mitigating the adversarial vulnerabilities of YOLOv5 to enhance its robustness and reliability. Specifically, YOLOv5s was used for this research due to its efficiency and suitability for real-time applications. To identify and analyze the adversarial vulnerabilities, adversarial attacks were simulated on a sample set of 10 images. Techniques such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and adversarial patch generation were applied to evaluate their impact on YOLOv5s's performance. Metrics such as detection accuracy, Intersection over Union (IoU), and classification confidence were used to quantify the disruption caused by the adversarial perturbations, establishing a baseline for the model's susceptibility.

To address these vulnerabilities, Feature Squeezing was implemented as a primary defense mechanism. This technique mitigates adversarial noise through methods such as median filtering and bit-depth reduction, preserving the semantic integrity of input data while eliminating subtle perturbations. Additionally, adversarial training was conducted by retraining YOLOv5s with datasets augmented with adversarially perturbed images, including the simulated 10-image sample set. This approach was designed to improve YOLOv5s's robustness by enhancing its ability to generalize and maintain accuracy in adversarial conditions.

The evaluation process involved rigorous testing of the enhanced YOLOv5s under adversarial conditions, using both naturalistic and synthetic attacks. Metrics such as precision, recall, mean average precision (mAP), and Attack Success Rate (ASR) were employed to assess the effectiveness of the implemented defense mechanisms. The spatial alignment of bounding boxes was evaluated using IoU metrics to determine the model's ability to maintain reliable detections despite adversarial perturbations. The results from the tests on the adversarially perturbed images were analyzed to demonstrate the improvements in YOLOv5s's resilience.

By focusing exclusively on adversarial vulnerabilities and leveraging YOLOv5s, this methodology aims to improve the model's reliability and security in environments where adversarial threats pose significant risks. The refined YOLOv5s is expected to deliver higher detection accuracy, robustness, and efficiency, making it a more secure and viable option for real-time applications in adversarially challenging scenarios.

3. Network Architecture and Algorithm of YOLOv5

A. Existing Algorithm of YOLOv5

1) Load the YOLOv5 with pre-trained weights.
2) Prepare the input data for inference using video camera

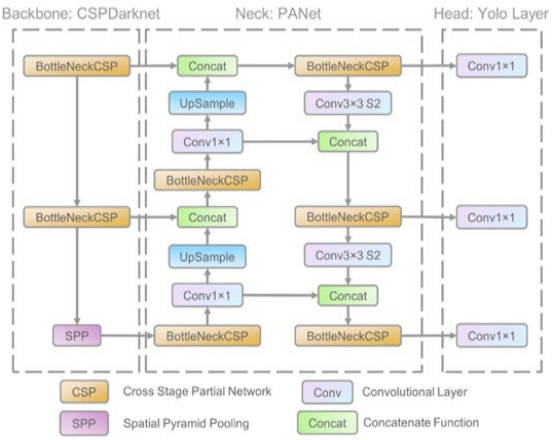

Fig. 1. The network architecture of YOLOv5. It consists of three parts: (1) Backbone: CSPDarknet, (2) Neck: PANet, and (3) Head: YOLO layer.

- Resize incoming video frames to match model input size.
- Convert frames to NumPy arrays.
- Normalize pixel values.

3) Object Detection
   - Start a timer for each frame to measure inference time.
   - Forward pass preprocessed frames through the YOLOv5.
   - Calculate inference time per frame.
   - Filter detected objects based on confidence scores.
   - Perform Non-Maximum Suppression (NMS) to remove redundant detections.
   - Optionally, apply a second-stage classifier for further refinement.
4) Overlay bounding boxes on video frames to visualize detected objects.

5) Output Generation
   - Provide additional information about detected objects (class label, dimensions, and confidence level of the prediction).
   - Measure average inference time and resource usage.
6) Repeated for subsequent frames in the video stream until completion.

B. Proposed Algorithm for Mitigating Vulnerabilities to Adversarial Attacks using Feature Squeezing

1) Apply Bit Depth Reduction
2) Apply Median Filtering
3) Preprocess images after applying feature squeezing methods
4) Perform object detection on preprocessed images using the YOLOv5 model with pre-trained weights.
5) Adversarial Attack Defense Evaluation
6) Metrics and Validation

4. Results and discussion

A. Traditional YOLOv5 Preprocessing Process

Fig. 2. Adversarial attack simulation

The results of the adversarial defense evaluation highlight the significant impact of adversarial perturbations on YOLOv5's detection performance. Before the attack, the model achieved a detection confidence of 0.917, with clear bounding boxes identifying objects in the image. However, after applying adversarial perturbations, the model failed to produce any valid bounding boxes or detections, resulting in an adversarial confidence of "None" and no outputs for IoU calculations. This demonstrates the success of the adversarial attack in completely disrupting the model's ability to detect objects, as it was unable to identify any features in the manipulated image.

These findings underscore the vulnerability of YOLOv5 to adversarial attacks, which can render the detection system ineffective under adversarial conditions. While the results show that the attack significantly degrades performance, they also emphasize the importance of implementing robust preprocessing techniques, such as feature squeezing, to mitigate these effects. By removing subtle adversarial noise and enhancing input robustness, such techniques are expected to improve YOLOv5's ability to perform reliably in real-world adversarial scenarios.

B. Enhanced YOLOv5 Preprocessing Process

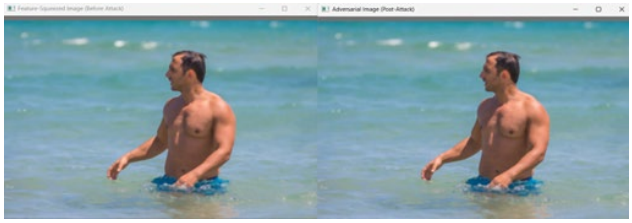
Fig. 3. Enhanced adversarial attack simulation

The Attack Success Rate evaluates the effectiveness of the attack methods, while sentence perplexity measures the extent of distortion in the generated adversarial examples (Tsai et al., 2019). Evaluation metrics are essential for assessing object detection models, focusing on accuracy, precision, and recall. A key metric, Intersection over Union (IoU), measures the overlap between predicted and ground truth bounding boxes, evaluating localization accuracy (Rezaie, 2023).

The simulation demonstrates the effectiveness of feature squeezing techniques, specifically Big Depth Reduction and Median Filtering, in defending YOLOv5 against adversarial attacks. Before the attack, YOLOv5 achieved a high detection confidence of 0.906, with accurate bounding box predictions for the detected objects. After adversarial perturbations were applied, the model maintained a slightly reduced detection confidence of 0.904, while preserving the bounding box predictions with minimal spatial shifts. The Intersection over Union (IoU) value for the evaluated bounding box was 0.94, showcasing precise alignment between the original and adversarial predictions. These results highlight that feature squeezing effectively mitigated the impact of adversarial manipulations by ensuring high spatial consistency and detection reliability.

The metrics used for evaluation further support these findings by validating both the attack's outcome and the defense mechanism's efficacy. Detection confidence indicated the model's certainty in object predictions, demonstrating only a marginal reduction from 0.906 to 0.904, thereby preserving detection accuracy. Bounding box metrics, including IoU and spatial shifts, evaluated the localization accuracy of objects, with IoU remaining near 1.0, confirming that adversarial perturbations had minimal effect on spatial precision. The attack status metric classified the attack as unsuccessful due to the consistency in detection confidence, bounding box predictions, and class labels between the original and adversarial images.

5. Conclusion

This study evaluated the effectiveness of feature squeezing techniques, specifically Big Depth Reduction and Median Filtering, in mitigating the impact of adversarial attacks on the YOLOv5 object detection model. Through multiple simulations on diverse images, the results consistently demonstrated the robustness of the enhanced model, with significant

improvements in detection accuracy, bounding box alignment, and resistance to adversarial perturbations. For the existing YOLOv5 model, adversarial attacks frequently resulted in reduced detection confidence, significant bounding box shifts, and, in some cases, complete failure to detect objects. In contrast, the enhanced YOLOv5 model exhibited a marked reduction in the effectiveness of such attacks, as evidenced by consistently high Intersection over Union (IoU) values, minimal bounding box shifts, and only marginal reductions in detection confidence.

The analysis highlighted the limitations of the original YOLOv5 model in handling adversarial conditions and validated the effectiveness of feature squeezing techniques as a defense mechanism. Notably, in cases where the original model failed entirely to detect objects post-attack, the enhanced model successfully maintained object detection with negligible deviations from pre-attack predictions. These findings underscore the practical viability of integrating feature squeezing techniques into object detection frameworks to enhance resilience against adversarial threats. By addressing key vulnerabilities in adversarial scenarios, the enhanced YOLOv5 model demonstrates its potential for reliable deployment in real-world applications, including security, autonomous systems, and other critical areas requiring robust object detection.

References

[1] Jain, S. (2024). Adversarial attack on Yolov5 for traffic and road sign detection. arXiv (Cornell University).
[2] Nelson, J., & Solawetz, J. (2020, June 10). YOLOv5 is Here: State-of-the-Art Object Detection at 140 FPS. Roboflow Blog.
[3] Rezaie, J. (2024, November 29). Mastering Object Detection Metrics: From IOU to MAP. Medium.
[4] Solawetz, J. (2024, April 9). What is YOLOv5? A Guide for Beginners. Roboflow Blog.
[5] Tsai, A. Y.-T., Yang, T., & Chen, H.-Y. (2019). Adversarial attack on sentiment classification. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 233–240. Florence, Italy: Association for Computational Linguistics.
[6] Ultralytics. (n.d.). GitHub - ultralytics/yolov5: YOLOv5 in PyTorch > ONNX > CoreML > TFLite. GitHub.
[7] Xu, R., Lin, H., Lu, K., Cao, L., & Liu, Y. (2021). A forest fire detection system based on ensemble learning. Forests, 12(2), 217.