Review of Adversarial Attacks in Object Detection

Li Wang

University of British Columbia, 2329 West Mall Vancouver, BC Canada V6T 1Z4

liwang955@outlook.com

Abstract. Object detection, a fundamental element of computer vision and artificial intelligence, has experienced considerable advancements through the incorporation of deep learning-based techniques. Yet, despite the impressive strides in both accuracy and efficiency, object detection algorithms harbor inherent vulnerabilities to adversarial attacks. These well-crafted disruptions pose significant risks, especially considering the broad application of object detection across an array of safety-critical sectors such as autonomous transportation, medical imaging, and security systems. This comprehensive paper offers a thorough review of adversarial attacks against object detection systems, dissecting the methods employed, and scrutinizing the implications of their exploits. It dives deep into the mechanics and consequences of both white-box and black-box attacks on prevalent object detection networks, including but not limited to Faster R-CNN, YOLO, and SSD. Furthermore, this paper underscores an assortment of defense strategies developed to mitigate the effects of adversarial attacks. These include adversarial training, gradient masking, input transformations, and randomized defenses. While these strategies hold promise, it is acknowledged that they have their limitations and do not offer a universal solution against all adversarial attacks. As such, this paper underscores the urgent necessity for robust defense mechanisms and stimulates further discourse and investigation into developing truly resilient object detection systems, capable of withstanding the growing threat of adversarial attacks.

Keywords: Adversarial Attack, Object Detection, Deep Learning.

1. Introduction

In recent years, Artificial Intelligence (AI) has made significant progress, transforming a wide array of fields by automating tasks and enabling machines to process and learn from large volumes of data [1,2,3]. These systems are powered by intricate algorithms that simulate human-like decision-making capabilities, making them invaluable in domains such as healthcare, finance, and autonomous transportation.

Computer vision is a key component of AI, focusing on the automatic extraction, analysis, and understanding of useful information from digital images [1,2,3]. Object detection, a subfield of computer vision, is critical in equipping AI systems with the ability to identify, locate, and classify objects within images or videos. This functionality is essential for various applications, including but not limited to autonomous vehicles, surveillance systems, robotics, and medical imaging.

Over the years, numerous object detection methods have been proposed and developed over the years, with traditional approaches such as Haar cascades, HOG features, and sliding window techniques giving way to more accurate and efficient deep learning-based algorithms [1,2,3]. These contemporary methods primarily employ convolutional neural networks (CNNs) as the backbone for feature extraction and classification. Faster R-CNN, YOLO (You Only Look Once), and SSD (Single Shot MultiBox Detector) are among the most popular deep learning-based object detection techniques [4].

Despite these advancements in object detection, the vulnerability of these methods to adversarial attacks remains a significant concern [5]. Adversarial attacks involve carefully crafted perturbations to the input data, often imperceptible to humans, that can result in misclassification or even complete failure of an AI system [5,6]. The susceptibility of object detection algorithms to such attacks can have severe consequences, particularly in safety-critical applications like autonomous vehicles or medical imaging. This underscores the importance of studying and addressing adversarial attacks against object detection systems [7].

Various adversarial attack methods have been proposed to exploit the vulnerabilities of object detection algorithms [5,6,8]. These methods can be broadly categorized into white-box and black-box attacks. White-box attacks assume complete knowledge of the target model, including its architecture and parameters, while black-box attacks only have access to the model's input-output behavior [9]. Some of the prominent adversarial attack techniques include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) attack. It is imperative to analyze the effectiveness of these attacks and develop robust object detection systems capable of withstanding them [10].

This paper focuses on adversarial attacks against object detection. It analyzes both white-box and black-box attacks to explore their mechanisms and vulnerabilities in object detection networks. Furthermore, this paper evaluates their effectiveness against popular object detection networks and discusses potential strategies to mitigate their impact. This comprehensive analysis aims to provide valuable insights into the challenges posed by adversarial attacks and emphasizes the necessity of developing robust object detection systems capable of withstanding these threats.

The rest of the paper will be organized as follows: Section 2 will introduce the existing object detection methods and networks, while Section 3 will focus on classical adversarial attack methods. Section 4 will present a practical analysis of adversarial attacks.

2. Existing Object Detection Methods and Networks

This section aims to provide an overview of the existing object detection methods, with a focus on deep learning-based techniques that have significantly improved the accuracy and efficiency of object detection tasks [1,2,3]. These contemporary methods primarily rely on convolutional neural networks (CNNs) for feature extraction and classification, resulting in highly effective and robust object detection models.

2.1. Two-Stage Object Detection Methods

Two-stage object detection methods consist of a region proposal stage followed by a classification and bounding box regression stage. Although these methods offer high accuracy, they come at the cost of computational complexity.

2.1.1. R-CNN (Regions with CNN features). R-CNN, which was proposed by Girshick et al., is the pioneering method in the two-stage object detection paradigm [1]. It employs selective search to generate region proposals and then uses a CNN to extract features from these regions. These features are classified using support vector machines (SVMs), and bounding box regression refines the final object locations.

2.1.2. Fast R-CNN. Fast R-CNN, introduced by Girshick, addresses the computational inefficiency of R-CNN [11]. Instead of extracting features for each region proposal independently, Fast R-CNN processes the entire image with a CNN and generates a feature map. Region of Interest (RoI) pooling is then applied to the feature map to obtain fixed-size feature vectors corresponding to region proposals.

These vectors are subsequently passed through fully connected layers for classification and bounding box regression.

2.1.3. Faster R-CNN. Faster R-CNN, developed by Ren et al., further improves the efficiency of Fast R-CNN by replacing the selective search with a Region Proposal Network (RPN) for generating region proposals [2]. RPN shares the same feature map with the object detection network, enabling end-to-end training and significantly reducing the computation time.

2.2. One-stage Object Detection Methods

One-stage object detection methods, also known as single-shot detectors, combine the tasks of region proposal and classification into a single network. These methods typically sacrifice some accuracy for a considerable increase in speed [12].

2.2.1. YOLO (You Only Look Once). YOLO, proposed by Redmon et al., divides the input image into a grid and predicts bounding boxes and class probabilities for each grid cell [3]. YOLO treats object detection as a regression problem, enabling it to process images in a single pass through the CNN. The simplicity and speed of YOLO make it suitable for real-time applications.

2.2.2. SSD (Single Shot MultiBox Detector). SSD was introduced by Liu et al. and extends the YOLO framework by using multiple feature maps at different scales to detect objects of varying sizes [13]. For each feature map cell, SSD generates a fixed number of anchor boxes and performs classification and bounding box regression simultaneously. By doing so, it achieves a balance between accuracy and speed, making it a popular choice for various object detection tasks.

2.3. Other Object Detection Methods

RetinaNet, which was proposed by Lin et al., addresses the issue of class imbalance in one-stage detectors by introducing the Focal Loss function [14]. RetinaNet uses a Feature Pyramid Network (FPN) for multi-scale feature extraction and employs anchor boxes similar to SSD. The Focal Loss function improves RetinaNet’s performance by focusing on difficult-to-classify examples.

In summary, numerous object detection methods and networks have been proposed to improve the accuracy and efficiency of object detection tasks [1,2,3,13,14]. However, the vulnerability of these methods to adversarial attacks remains a critical concern [5]. As these object detection techniques become more prevalent in safety-critical applications, it becomes increasingly important to assess their robustness against adversarial attacks and develop countermeasures to ensure reliable performance.

3. Classical Adversarial Attack Methods

This section introduces classical adversarial attack methods and explores their mechanisms and the vulnerabilities they exploit in object detection networks [5,6,8]. These attacks can be broadly categorized into white-box and black-box attacks, with white-box attacks assuming complete knowledge of the target model, while black-box attacks only have access to the model's input-output behavior [15].

3.1. White-Box Attacks

3.1.1. Fast Gradient Sign Method (FGSM). FGSM, proposed by Goodfellow et al., is a straightforward one-step method for generating adversarial examples [5]. It computes the gradient of the loss function with respect to the input image and applies a small perturbation in the direction of the gradient's sign. The primary advantage of FGSM is its simplicity and computational efficiency, but the resulting adversarial examples may not be optimal.

3.1.2. Projected Gradient Descent (PGD). PGD, introduced by Madry et al., is an iterative variant of FGSM that refines adversarial examples by repeatedly applying FGSM-like updates followed by projection onto the allowed perturbation space [6]. PGD can generate more effective adversarial examples than FGSM at the cost of increased computational complexity.

3.1.3. Carlini & Wagner (C&W) Attack. The C&W attack, proposed by Carlini and Wagner, is an optimization-based method that minimizes the perturbation required to fool the target model while maintaining the perturbation's L2, L0, or Linf norm within specified bounds [8]. The C&W attack is known for generating high-quality adversarial examples but is computationally more expensive than FGSM or PGD.

3.2. Black-Box Attacks

3.2.1. Transferability-based Attacks. Transferability-based attacks leverage the phenomenon that adversarial examples generated for one model can often fool other models with similar architectures or trained on similar tasks. This property allows an attacker to generate adversarial examples using a surrogate model and then transfer these examples to the target model without direct access to the target model's architecture or parameters.

3.2.2. Zeroth Order Optimization (ZOO) Attack. The ZOO attack, introduced by Chen et al., is a black-box attack that estimates the gradients of the target model using numerical approximation techniques, such as finite differences [15]. This enables the attacker to generate adversarial examples without direct access to the model's gradients. Although the ZOO attack can be effective, it requires a large number of queries to the target model, making it less practical in some scenarios.

In conclusion, various classical adversarial attack methods have been developed to exploit the vulnerabilities of object detection algorithms. These attacks pose significant challenges to the robustness of object detection systems, particularly in safety-critical applications. The next section will present a practical analysis of adversarial attacks, evaluating their effectiveness against popular object detection networks and discussing potential strategies to mitigate their impact.

4. Practical Analysis of Adversarial Attacks

This section presents a practical analysis of adversarial attacks, evaluating their effectiveness against popular object detection networks and discussing potential strategies to mitigate their impact [4,12,15]. The analysis covers both white-box and black-box attack scenarios and considers various adversarial attack methods introduced in Section 3.

4.1. Effectiveness of Adversarial Attacks on Object Detection Networks

4.1.1. Two-stage Object Detection Networks. For two-stage object detection networks, such as Faster R-CNN, adversarial attacks can target both the region proposal and the classification stages. Studies have demonstrated that FGSM, PGD, and C&W attacks can successfully degrade the performance of these networks by causing misclassifications or suppressing true object detections [10]. Transferability-based black-box attacks have also shown to be effective in some cases, although their success rate is generally lower than white-box attacks [12].

4.1.2. One-stage Object Detection Networks. One-stage object detection networks, such as YOLO and SSD, are also susceptible to adversarial attacks. Similar to two-stage networks, these methods can be targeted by both white-box and black-box attacks. Research has shown that FGSM, PGD, and C&W attacks can cause significant performance degradation in these networks [15]. In general, one-stage networks tend to be more vulnerable to adversarial attacks than their two-stage counterparts due to their single-pass design and the use of anchor boxes.

4.2. Strategies for Mitigating Adversarial Attacks

4.2.1. Adversarial Training. Adversarial training is a widely used approach to improving a model's robustness against adversarial attacks [6]. It involves augmenting the training data with adversarial examples generated using various attack methods. This enables the model to learn robust features and enhances its ability to withstand adversarial attacks. However, adversarial training can be computationally expensive and may lead to a decrease in performance on clean (non-adversarial) data.

4.2.2. Gradient Masking and Input Transformations. Gradient masking techniques, such as defensive distillation, aim to reduce the model's susceptibility to adversarial attacks by making the gradients less informative for crafting adversarial examples [16]. Input transformations, such as image smoothing or JPEG compression, can be used to remove or reduce the impact of adversarial perturbations on the input data. However, these techniques may not always provide robust defense against adaptive attackers who are aware of the defense mechanisms.

4.2.3. Randomized Defenses. Randomized defenses introduce randomness into the model's architecture or input data, making it difficult for the attacker to generate effective adversarial examples [13]. Examples of randomized defenses include random resizing, random padding, or random dropout. While these techniques can provide some level of defense against adversarial attacks, they may not be sufficient to guarantee robustness against all attack scenarios.

In conclusion, adversarial attacks pose significant challenges to the robustness of object detection networks, and various strategies have been proposed to mitigate their impact. Although these defense mechanisms can improve the model's resilience against adversarial attacks, there is still no universal solution that guarantees complete robustness.

5. Conclusion

In this paper, a comprehensive review of adversarial attacks against object detection is presented to examine classical adversarial attack methods, their implications for popular object detection, and potential countermeasures. Our analysis reveals that both two-stage and one-stage object detection networks are susceptible to adversarial attacks, posing significant challenges to their robustness, particularly in safety-critical applications. This paper also reveals that the defense mechanisms often involve trade-offs between robustness, accuracy, and computational efficiency, and adaptive attackers may still exploit vulnerabilities in the presence of these defenses.

References

[1] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[2] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[4] Zou, X. (2019). A Review of Object Detection Techniques. In 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA) (pp. 251-254). Xiangtan, China. https://doi.org/10.1109/ICSGEA.2019.00065.

[5] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[7] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial Attacks and Defences: A Survey. ArXiv. /abs/1810.00069

[8] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE.

[9] Al-Shaer, R., Spring, J. M., & Christou, E. (2020). Learning the associations of MITRE ATT & CK Adversarial Techniques. In 2020 IEEE Conference on Communications and Network Security (CNS) (pp. 1-9). IEEE. https://doi.org/10.1109/cns48642.2020.9162207

[10] Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied Sciences, 9(5), 909. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app9050909

[11] Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[12] Lu, X., Li, Q., Li, B., Yan, J. (2020). MimicDet: Bridging the Gap Between One-Stage and Two-Stage Object Detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12359. Springer, Cham. https://doi.org/10.1007/978-3-030-58568-6_32

[13] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[14] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[15] Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 15-26).

[16] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In 2016 IEEE Symposium on Security and Privacy (SP) (pp. 582-597). San Jose, CA, USA. https://doi.org/10.1109/SP.2016.41.