When AI fails to see: The challenge of adversarial patches

M. ZIMOŃ, R. KASPRZYK

michal.zimon@wat.edu.pl, rafal.kasprzyk@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Kaliskiego Str. 2, 00-908 Warsaw, Poland

Object detection, a key application of machine learning in image processing, has achieved significant success thanks to advances in deep learning [6]. In this paper, we focus on analysing the vulnerability of one of the leading object detection models, YOLOv5x [14], to adversarial attacks using specially designed interference known as "adversarial patches" [4]. These disturbances, while often visible, have the ability to confuse the model, which can have serious consequences in real world applications. We present a methodology for generating these interferences using various techniques and algorithms, and we analyse their effectiveness in various conditions. In addition, we discuss potential defences against these types of attacks and emphasise the importance of security research in the context of the growing popularity of ML technology [13]. Our results indicate the need for further research in this area, bearing in mind the evolution of adversarial attacks and their impact on the future of ML technology.

Keywords: object detection, adversarial patches, YOLO model, machine learning.

DOI: 10.5604/01.3001.0054.0092

1. Introduction

Object detection is one of the key applications of machine learning in the field of image processing. It enables automatic recognition and localization of objects in an image or video [18].

In recent years, thanks to the progress in the field of deep learning, object detection techniques have achieved significant success, becoming more precise and efficient [9]. One of the leading models in this field is YOLOv5 (You Only Look Once version 5). It features high-performance real-time detection, making it an attractive choice for a wide range of applications, from security monitoring to medical image analysis. However, with the development of machine learning technology, new challenges also arise. One of them is the so-called adversarial attack [17]. These attacks involve introducing small, intentional disturbances to the input data that can confuse the model, causing false predictions. In the past, many studies focused on creating interference invisible to the human eye, which made them particularly dangerous in the context of digital images. However, these days, in an era of growing concern for real-world privacy, visible disruptions such as patches are increasingly accepted. What matters is not whether the disturbance is visible but whether it is effective in deceiving the model. In the context of object detection, an adversarial attack could, for example, make a person "invisible" to the surveillance system, even if they wear a visible patch [16]. The importance of adversarial attacks in the context of the security of ML models is enormous [3]. These models are increasingly used in critical applications such as autonomous cars and medical systems [21]. Misprediction in such situations can have serious consequences. Therefore, understanding and countering adversarial attacks has become one of the priorities in the field of machine learning security research [19], [20].

2. Adversarial patches

Adversarial patches are specific types of disturbance that are intentionally designed to confuse machine learning models [4]. Unlike traditional adversarial attacks, which involve minor modifications to the entire image, adversarial patches are usually visible to the human eye and can be placed anywhere in the image. Their main purpose is to mislead the model by "obscuring" or "interfering" with correct detection. The operation of adversarial patches is based on exploiting the weaknesses of the model's internal representations. Through appropriate manipulations, the patch can effectively "confuse" the model, causing

correctly recognised objects to become invisible or misclassified [7].

Application examples:

- Protection of privacy: to avoid detection by surveillance systems, people may wear clothing with adversarial patches.
- Research: researchers use adversarial patches to test the robustness and understanding of how deep learning models work [15].

Potential threats:

- Security: in the context of surveillance systems or autonomous cars, patches can be used to evade detection, creating a security risk [5].
- Unauthorised access: patches can be used to mislead facial recognition systems, allowing unauthorised access to protected areas or systems.
- Misinformation: in the context of image analysis, patches can be used to mislead algorithms, leading to misinterpretations and misinformation.

3. YOLOv5x model

The model was trained on the popular COCO collection [10]. However, to focus on human detection, photos were filtered out of the original collection so that only those containing people remained. The COCO set is widely recognised in the research community as one of the best sets for training object detection models, making it an ideal choice for this experiment.

The YOLOv5x architecture is used, which is one of the newest and most advanced variants of the YOLO model. It is characterised by high precision and detection speed, which makes it an ideal choice for real-time applications.

The YOLOv5x model was evaluated against various metrics to thoroughly understand its human detection performance. The most important metrics used to evaluate the model are:

- Precision: precision determines the model's ability to correctly identify only relevant instances. High precision indicates that false positives are rare.
- Recall: sensitivity determines the model's ability to correctly identify all relevant instances. High recall indicates that false negatives are rare.
- mAP (average precision across all feature classes): this is the average precision value for various IoU thresholds. mAP is one of the most important metrics used to evaluate object detection models.

Results for YOLOv5x model:

Tab. 1. Results for the YOLOv5x model

Metric	Value
Highest precision	0.95709
Highest recall	0.93942
mAP_{50}	0.97787
mAP_{50-95}	0.83472

These results indicate that the YOLOv5x performs excellently in human detection, making it an ideal choice for applications that require accurate, real-time human detection.

4. Adversarial patch generation methodology

In the context of studying the influence of adversarial interference on the detection efficiency of the YOLOv5x model, advanced image manipulation techniques and algorithms were used. The aim of these manipulations was to create disturbances capable of disorienting the model while preserving certain aesthetic and structural properties of the image. Below are the key methods used in the interference generation process:

- Adding Gaussian Noise: this technique introduces random noise into the image, which can interfere with the activation functions of the inner layers of the model.
- Brightness and Contrast Modification: subtle changes in brightness and contrast can affect the model's perception of objects, especially in the context of edge and feature detection.
- Random rotation and scaling: by changing the orientation and size of the disturbances, you can test the model's resistance to various geometric transformations.
- Affine Transforms: advanced transformation techniques such as rotation and scaling that can be applied simultaneously to produce more complex distortion effects.

Each of the above techniques was used to investigate how different types of image manipulation can affect the model's ability to detect objects correctly.

5. Metrics for patches

To ensure that patches can fool the model but still stay as printable aspossible, the following losses were introduced:

- Maximum Probability Extraction (MPE): focuses on minimising the probability of detection for a specific class. In practice, this means that patches are designed to maximally interfere with the model's ability to correctly classify objects:

$$L_{MPE} = 1 - \max\big(p(y|x)\big) \quad (1)$$

where $p(y \mid x)$ is the probability of the object belonging to a specific class.

- Saliency Loss (SL) (Hasler and Susstrunk 2003) used to minimise the colour of the image. In practice, this means that the patches are less expressive and more inconspicuous to the human eye:

$$L_{SL} = \sqrt{\left(\sigma_{rg^2} + \sigma_{yb^2}\right)} + \\ + 0.3\sqrt{\left(\mu_{rg^2} + \mu_{yb^2}\right)} \quad (2)$$

where σ and μ are the standard deviation and mean value of the colours, respectively.

- Total Variation (TL): it focuses on minimising patch variation, which helps create more uniform and less noticeable patches:

$$L_{TL} = \sum\left|x_{(i+1,j)} - x_{(i,j)}\right| + \\ + \sum\left|x_{(i,j+1)} - x_{(i,j)}\right| \quad (3)$$

where x is the pixel value in the image.

- Non-printability Score (NPS) (Sharif et al. 2016): it aims to minimize the "printability" of patches, which means patches are designed to be harder to detect once printed:

$$L_{WN} = \min\left(\left|\left|x - c\right|\right|_2\right) \quad (4)$$

where x is the colour of the pixel per patch and c is a colour from a predefined set of printable colours.

6. Results

The experiment focused on evaluating the effectiveness of adversarial patches in fooling the YOLOv5x model. The results are presented for three different scenarios: images without patches (as a reference), images with correct patches, and images with random noise.

a. Images without patches (for reference):
   - The average precision (AP) for different IoU values was about 0.901, which indicates the high efficiency of the model in detecting people without the presence of patches.
   - The average recall was 0.929, suggesting that the model correctly identified most people in the images.

Fig. 1. Images without patches

b. Images with the correct patches:
   - The average precision (AP) dropped drastically to 0.139, indicating that the adversarial patches were effective in misleading the model.
   - The average recall was only 0.143, suggesting that the model had difficulty identifying people correctly in patched images.
   - The success rate of the attack was 0.571 for all areas, which indicates that in more than half of the cases, the model was fooled by patches.

Fig. 2. Images with proper patches

c. Images with random patches:
   - The average precision (AP) was 0.521, indicating that random noise was not as effective as specially designed patches in fooling the model.
   - The average recall was 0.679, suggesting that the model was able to correctly identify people in most images with random noise.
   - The attack's success rate was 0.000 in every area, showing that random noise was not the cause of the model's deception.

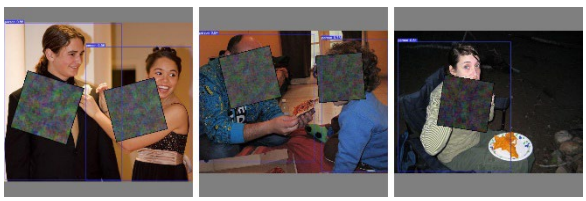

Fig. 3. Images with random patches

Adversarial patches have been effective at fooling the YOLOv5x model, especially when specifically designed for this purpose as shown in Figure 2. Compared to random noise, generated patches had a much higher attack success rate. However, it is worth noting that the model still had some difficulty detecting people even in the presence of random noise, indicating the potential sensitivity of the model to various types of interference.

7. Generate patches for use in the real world

Attacks based on adversarial patches in the digital environment differ significantly from those in reality. In the real world, many factors must be taken into account that can affect the effectiveness of an attack.
Challenges of moving an attack from the digital world to the real world:

- Lighting: lighting is a variable that is difficult to control. Changing lighting conditions can affect the appearance of the patch, which may affect its effectiveness.
- Perspective and distance: the camera can view the patch from various angles and distances, which may affect how the model perceives the patch.
- Interactions with the environment: a patch can be covered, distorted, or destroyed by interactions with the environment, such as wind, rain, or other objects.

The patch generation methodology to use in reality:

- Simulation of various lighting conditions: patches are tested in various lighting conditions to ensure their effectiveness in various scenarios.
- Different perspectives and distances: patches are tested from different angles and distances to ensure they are effective regardless of camera position.
- Optimised for interaction with the environment: patches are designed to resist covering, distortion and damage.

8. Defence against adversarial patches

Adversarial attacks are a serious challenge for machine learning-based systems, especially in the context of real-world object detection. However, the development of defences against these attacks is an active area of research [1], [2], [11], [12]. Below are some potential defences against attacks based on adversarial patches and ways to improve the model's resistance to such attacks.

a. Potential defences:
   - Anomaly Detection: one defence is to detect unusual patterns in the image that may indicate the presence of an adversarial patch. If the system detects such an anomaly, it may ignore the area or take additional verification actions.
   - Data Augmentation: various types of noise can be introduced into the training data during model training, which can help make the model more resilient to adversarial attacks.
   - Ensemble models: using several models in an ensemble can help increase resiliency, as an attack that works against one model may not work against another.
   - Countering adversarial training: this involves deliberately adding adversarial patches to the training data and teaching the model to correctly classify images despite their presence.

b. Ways to improve model resilience:
   - Regular updates: regular updates to the model, including adapting it to new attack techniques, can help maintain its resistance to adversarial attacks.
   - Deeper Layer Analysis: activation analysis of the inner layers of the model can help detect unusual activation patterns that are characteristic of adversarial attacks.
   - Exposure restriction: restricting access to the model and its parameters to third parties can help prevent attacks that require accurate knowledge of the model.

9. Summary

During the research and experiments related to the generation and application of "adversarial patches" on the YOLOv5x model, several important conclusions were obtained:

- Effectiveness of attacks: adversarial patches can significantly disrupt the operation of object detection models, even as advanced as YOLOv5x. This indicates potential security risks for applications using ML technologies in real-world environments.
- Complexity of the attack: while generating effective disruptions requires careful

selection of techniques and algorithms, these attacks have been proven to be feasible even for people with limited resources.

- Visibility of Disruptions: unlike many other adversarial attacks that aim to be invisible, this disruption was intentionally visible. This underscores the fact that in some scenarios, the attacker may not care about the stealth of the attack but rather its effectiveness.

Reflections on the future of adversarial attacks:

- Evolution of ML technology: as our research shows, the development of ML technology goes hand in hand with the emergence of new attack vectors. As models become more advanced, adversarial attacks will also evolve, becoming more sophisticated.
- Importance of Defence: considering the potential threats of adversarial attacks, intensive research and development of defensive techniques is essential. Protection against adversarial attacks will be a key element in future ML-based systems.
- Ethical and social implications: adversarial attacks can have serious consequences in real-life applications such as autonomous cars and surveillance systems. As ML technology becomes more and more integrated into our daily lives, it is important that the research community, industry and policymakers are aware of the potential risks and work to minimise them.

10. References

[1] Alshahrani E., Alghazzawi D., Alotaibi R.M. Rabie O., "Adversarial attacks against supervised machine learning based network intrusion detection systems", *PLoS ONE*, 17(10): e0275971 (2022).

[2] Apruzzese G., Conti M., Yuan Y., "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning", *ACM Digital Library*, 2022.

[3] Biggio B., Roli F., "Wild patterns: Ten years after the rise of adversarial machine learning", *Pattern Recognition*, vol. 84, 317–331 (2018).

[4] Brown T.B., Man'e, D., Roy A., Abadi M., Gilmer J., "Adversarial patch", arXiv preprint arXiv:1712.09665, 2017.

[5] Evtimov I., Eykholt K., Fernandes E., Kohno T., Li B., Prakash A., Rahmati A., Song D., "Robust physical-world attacks on deep learning visual classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, IEEE, 2018.

[6] Girshick R., Donahue J., Darrell T., Malik J., "Rich feature hierarchies for accurate object detection and semantic segmentation", *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, IEEE, 2014.

[7] Goodfellow I.J., Shlens J., Szegedy C., "Explaining and harnessing adversarial examples", arXiv preprint arXiv:1412.6572, 2014.

[8] Hasler D., Susstrunk S., "Measuring colorfulness in natural images", *Proceedings of SPIE*, vol. 5007 (2003).

[9] He K., Gkioxari G., Dollár P., Girshick R., "Mask R-CNN", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, IEEE, 2017.

[10] Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C.L., "Microsoft COCO: Common Objects in Context", [in:] *Computer Vision – ECCV 2014*, pp. 740–755, Springer, 2014.

[11] Lou W., "Fortifying your defenses: Techniques to thwart adversarial attacks and boost performance of machine learning-based intrusion detection systems", *ACM Digital Library*, 2023.

[12] Nowroozi E., Mohammadi M., Savas, E., Mekdad Y., Conti M., "Employing deep ensemble learning for improving the security of computer networks against adversarial attacks", *IEEE Transactions on Network and Service Management*, PP(99):1-1, 2022.

[13] Papernot N., McDaniel P., Goodfellow I., Jha S., Celik Z.B., Swami A., "Practical black-box attacks against machine learning", pp. 506–519, *ACM*, 2016.

[14] Redmon J., Divvala S., Girshick R., Farhadi A., "You only look once: Unified, real-time object detection", *2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016.

[15] Sharif M., Bhagavatula S., Bauer L., Reiter M.K., "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition", *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1528–1540, ACM, 2016.

[16] Song C., Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Xiao C., Prakash A., Kohno T., "Physical adversarial

examples for object detectors", *12th Workshop on Offensive Technologies* (*WOOT'18*), Baltimore, USA, 2018.
[17] Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R., "Intriguing properties of neural networks", arXiv preprint arXiv:1312.6199, 2013.
[18] Tarsała P., Kasprzyk R., "Deep Learning Algorithms in Computer Vision", *Proceedings of the 37th International Business Information Management Association* (*IBIMA*), pp. 11147–11155, ISBN: 978-0-9998551-6-4, 30–31 May 2021, Cordoba, Spain.
[19] Tymoszuk Ł, Kasprzyk R., "Adversarial Machine Learning as A Forerunner of Future Wars on Algorithms", *Proceedings of the 37th International Business Information Management Association* (*IBIMA*), pp. 11165–11176, ISBN: 978-0-9998551-6-4, 30–31 May 2021, Cordoba, Spain.
[20] Zimoń M., Kasprzyk R., "Yet another research on GANs in cybersecurity", *Cybersecurity and Law*, vol. 9(1), 61–72 (2023).
[21] Zimoń M., Kasprzyk R., "Digital revolution and cyber threats as its consequence", *Proceedings of the 38th International Business Information Management Association* (*IBIMA*), pp. 7750–7755, ISBN: 978-0-9998551-7-1, 23–24 November 2021, Seville, Spain.

Kiedy sztuczna inteligencja nie widzi: wyzwanie antagonistycznych wstawek

M. ZIMOŃ, R. KASPRZYK

Wykrywanie obiektów to kluczowe zastosowanie algorytmów uczenia maszynowego w przetwarzaniu obrazu, które odniosło znaczący sukces dzięki postępom w głębokim uczeniu. W artykule przedstawiono analizę podatności jednego z wiodących modeli wykrywania obiektów, YOLOv5x, na ataki z wykorzystaniem specjalnie zaprojektowanych zakłóceń, znanych jako antagonistyczne wstawki. Omówiono metodę generowania antagonistycznych wstawek z wykorzystaniem różnych algorytmów i ich skuteczność w różnych warunkach. Ponadto przedstawiono potencjalne mechanizmy obronne przed tego typu atakami. Uzyskane wyniki wskazują na potrzebę dalszych badań w tym obszarze, w szczególności biorąc pod uwagę rozwój obszaru antagonistycznego uczenia maszynowego.

Słowa kluczowe: detekcja obiektów, antagonistyczne wstawki, model YOLO, uczenie maszynowe.