iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VII **Month of publication:** July 2024

DOI: https://doi.org/10.22214/ijraset.2024.63614

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VII July 2024- Available at www.ijraset.com

Adaptive Pixel Resilience: A Novel Defence Mechanism Against One-Pixel Adversarial Attacks on Deep Neural Networks

Smit Srivastava[1], Dr. Muskaan[2]
[1]*Director, AI & Data Product Management, GroupM, Karnataka, India*
[2]*Assistant Professor, Dept. of English and Cultural Studies, CHRIST (Deemed to be University), Bannerghatta Campus, Karnataka, India*

Abstract: ***This paper presents a groundbreaking analysis of the One Pixel Attack, an insidious adversarial threat that challenges the robustness of state-of-the-art deep neural networks (DNNs). We delve into the intricate mechanics of this deceptively simple yet potent attack, which can cause misclassification by altering just a single pixel in an image. Our research not only unravels the technical underpinnings of the One Pixel Attack but also introduces Adaptive Pixel Resilience (APR), a novel defence mechanism that significantly enhances DNN robustness against this threat. Through extensive experimentation on the CIFAR-10 and ImageNet datasets, we demonstrate the remarkable efficacy of APR. Our method substantially outperforms existing defence strategies, setting a new benchmark in adversarial robustness while maintaining competitive clean accuracy. The paper offers several key contributions:***

1) ***A comprehensive review and mathematical formulation of the One Pixel Attack***
2) ***The innovative APR defence, combining adversarial training, pixel-wise attention mechanisms, and regularization techniques***
3) ***Rigorous empirical evaluation and ablation studies, providing insights into the effectiveness of each APR component***
4) ***Analysis of attention maps, elucidating how the APR model focuses on key identifying features for different object classes, enhancing interpretability of the defence mechanism***

Our findings not only advance the field of adversarial machine learning but also have far-reaching implications for the deployment of AI systems in security-critical applications. This research paves the way for more resilient and trustworthy AI, addressing a critical challenge in the era of ubiquitous deep learning.

I. INTRODUCTION

Adversarial attacks on deep neural networks have become a critical concern in the field of artificial intelligence and cybersecurity. These attacks exploit vulnerabilities in AI models, causing them to make erroneous predictions based on carefully crafted input perturbations. Among these, the One Pixel Attack stands out for its simplicity and effectiveness, demonstrating that even a single-pixel modification can lead to misclassification in state-of-the-art image recognition systems.

This paper aims to:

1) Provide a comprehensive review of existing literature on adversarial attacks, with a focus on the One Pixel Attack.
2) Present a detailed technical explanation of the attack mechanism, including its mathematical foundations.
3) Propose an innovative defence strategy to mitigate the impact of One Pixel Attacks.

II. LITERATURE REVIEW

A. *Adversarial Attacks in Deep Learning*

Szegedy et al. (2014) [1] first introduced the concept of adversarial examples, demonstrating that imperceptible perturbations could cause misclassification in neural networks. This discovery led to extensive research in adversarial machine learning.

Goodfellow et al. (2015) [2] proposed the Fast Gradient Sign Method (FGSM), a simple yet effective technique for generating adversarial examples. FGSM uses the sign of the gradient of the loss function to create perturbations, making it computationally efficient.

Carlini and Wagner (2017) [3] developed a set of attacks (C&W attacks) that could bypass many existing defences, highlighting the need for more robust protection mechanisms.

B. One Pixel Attack

Su et al. (2019) [4] introduced the One Pixel Attack, demonstrating that modifying a single pixel could cause misclassification in DNNs. Their work showed that 70.97% of the images in the CIFAR-10 dataset could be perturbed to at least one target class by changing just one pixel.

Narodytska and Kasiviswanathan (2017) [5] further explored the effectiveness of few-pixel attacks, showing that they could be successful even when applied to a small number of pixels.

C. Defences Against Adversarial Attacks

Madry et al. (2018) [6] proposed adversarial training as a defence mechanism, where models are trained on adversarial examples to improve robustness.

Guo et al. (2018) [7] introduced input transformation techniques, such as bit-depth reduction and JPEG compression, to mitigate adversarial perturbations.

Papernot et al. (2016)[8] developed defensive distillation, a technique that trains models to produce smoother decision boundaries, making them more resilient to adversarial examples.

III. TECHNICAL EXPLANATION OF ONE PIXEL ATTACK

The One Pixel Attack exploits the high-dimensional nature of image data and the sensitivity of neural networks to small perturbations. The attack can be formalized as an optimization problem:

Let f: $\mathbb{R}^{nm3} \rightarrow \mathbb{R}^c$ be a classifier that maps an n×m RGB image to a probability distribution over c classes. The goal is to find a perturbation δ that modifies a single pixel (i, j) such that:

$\text{argmax}_k\ f_k(x + \delta) \neq \text{argmax}_k\ f_k(x)$

where x is the original image, and f_k is the k-th component of the output.

The attack uses Differential Evolution (DE), a population-based optimization algorithm, to solve this problem. DE iteratively refines candidate solutions by combining existing candidates and keeping the best performers. The process can be described as follows:

i. Initialize a population of candidate perturbations: $P = \{\delta_1, ..., \delta_N\}$

ii. For each generation:

  a. For each δ_i in P:

    - Create a mutant vector $v_i = \delta_{r1} + F(\delta_{r2} - \delta_{r3})$

    - Create a trial vector u_i by crossover between δ_i and v_i

  b. Evaluate fitness of u_i: $f(x + u_i)$

  c. If $f(x + u_i)$ is better than $f(x + \delta_i)$, replace δ_i with u_i

iii. Repeat until convergence or maximum generations reached

The fitness function is designed to maximize the probability of the target class while minimizing the perturbation magnitude:

$\text{fitness}(\delta) = f_t(x + \delta) - \lambda||\delta||_2$

where t is the target class and λ is a regularization parameter.

IV. PROPOSED DEFENCE: ADAPTIVE PIXEL RESILIENCE (APR)

We propose Adaptive Pixel Resilience (APR), a novel defence mechanism against One Pixel Attacks. APR combines adversarial training with a pixel-wise attention mechanism to enhance model robustness.

A. Adversarial Training Component

The model is trained on both clean and adversarially perturbed images. For each mini-batch, we generate One Pixel Attack examples using the current model state. The loss function is modified to incorporate both clean and adversarial examples:

$L = \alpha * L_{clean} + (1 - \alpha) * L_{adv}$

where L_{clean} is the standard cross-entropy loss on clean images, L_{adv} is the loss on adversarial examples, and α is a hyperparameter controlling the trade-off.

B. *Pixel-wise Attention Mechanism*

We introduce a pixel-wise attention layer that learns to focus on regions of the image that are most relevant for classification and less likely to be affected by single-pixel perturbations. The attention mechanism is defined as:

$A(x) = \sigma(W_a * x + b_a)$
where W_a and b_a are learnable parameters, and σ is the sigmoid function.
The final classification is performed on the attended feature map:
$y = f(A(x) \odot x)$
where ⊙ denotes element-wise multiplication.

C. *Regularization*

To further enhance robustness, we add a regularization term that encourages smooth decision boundaries:

$R = \beta \cdot \|\nabla_x f(x)\|_2^2$

- ∇_x represents the gradient with respect to x,
- $\|\cdot\|_2$ represents the L2 norm (Euclidean norm).

The final loss function becomes:
$L_{total} = L + R$

D. *Mathematical Analysis*

The effectiveness of APR can be understood through the lens of Lipschitz continuity. By encouraging smoother decision boundaries and focusing on robust features, we effectively reduce the Lipschitz constant of the network, making it less sensitive to small perturbations.
Let L be the Lipschitz constant of the network. For a One Pixel Attack with perturbation δ, we have:
$\|f(x + \delta) - f(x)\| \leq L * \|\delta\|$
By reducing L through our defence mechanism, we minimize the impact of δ on the network's output.

V. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our proposed Adaptive Pixel Resilience (APR) defence against One Pixel Attacks, we conducted extensive experiments on two standard datasets: CIFAR-10 and ImageNet. We compared our method against several baseline defences and state-of-the-art techniques.

A. *Experimental Setup*

1) *Datasets*

- CIFAR-10: 60,000 32×32 RGB images in 10 classes (50,000 training, 10,000 testing)
- ImageNet: Subset of 50,000 images from the validation set, resized to 224×224

2) *Models*

- For CIFAR-10: • ResNet-18: 18-layer residual network • VGG-16: 16-layer convolutional neural network
- For ImageNet: • ResNet-50: 50-layer residual network • VGG-19: 19-layer convolutional neural network

3) *Baselines*

- No defence (vanilla models): Standard trained models without any defensive measures
- Adversarial Training (AT): Models trained on a mix of clean and adversarial examples
- Input Transformation (IT): Applying preprocessing techniques like bit-depth reduction and JPEG compression
- Defensive Distillation (DD): Training models to produce smoother decision boundaries

4) Evaluation Metrics

- Clean Accuracy: Model accuracy on unperturbed test images
  ACCclean = (Correct predictions on clean images) / (Total clean images)
- Attack Success Rate (ASR): Percentage of successful one-pixel attacks
  ASR = (Successful attacks) / (Total attack attempts)
- Robustness: Robustness = 1 – ASR (higher is better)

5) Implementation Details

- Framework: PyTorch 1.8.0
- Hardware: NVIDIA Tesla V100 GPUs
- Optimization: Adam optimizer with learning rate 0.001 and weight decay 5X10-4
- Batch size: 128 for CIFAR-10, 64 for ImageNet
- Training epochs: 200 for CIFAR-10, 90 for ImageNet

6) APR Hyperparameters

- α (adversarial training trade-off): 0.7
- β (regularization strength): 0.01

7) One Pixel Attack Generation

- Differential Evolution parameters: population size = 400, max iterations = 100
- Pixel modification range: [-255, 255] for each colour channel

8) For Each Dataset And Model Combination, We Performed The Following Steps

- Train the baseline model without defences
- Implement each defence method (AT, IT, DD, and APR)
- Generate One Pixel Attack examples for the test set
- Evaluate each model's performance using the defined metrics

9) Statistical Significance

- We report mean values over 5 independent runs for each experiment
- 95% confidence intervals are calculated for all reported metrics

B. Results on CIFAR-10

Table 1: Performance comparison on CIFAR-10 dataset

Defence Method	Clean Accuracy	Attack Success Rate	Robustness
No Defence	93.2%	70.97%	29.03%
AT	87.3%	45.32%	54.68%
IT	89.1%	38.76%	61.24%
DD	86.5%	41.89%	58.11%
APR (Ours)	90.8%	21.43%	78.57%

Our APR method significantly outperforms existing defences, reducing the attack success rate to 21.43% while maintaining a high clean accuracy of 90.8%. This represents a 69.8% reduction in vulnerability compared to the undefended model.

C. Results on ImageNet

Table 2: Performance comparison on ImageNet subset

Defence Method	Clean Accuracy	Attack Success Rate	Robustness
No Defence	76.1%	63.25%	36.75%
AT	72.4%	39.87%	60.13%
IT	73.8%	35.42%	64.58%
DD	71.9%	37.61%	62.39%
APR (Ours)	74.6%	18.79%	81.21%

On the more challenging ImageNet dataset, APR demonstrates superior performance, reducing the attack success rate to 18.79% while maintaining a competitive clean accuracy of 74.6%. This represents a 70.3% reduction in vulnerability compared to the undefended model.

D. Ablation Study

To understand the contribution of each component in APR, we conducted an ablation study on CIFAR-10:

Table 3: Ablation study results on CIFAR-10

APR Components	Clean Accuracy	Attack Success Rate	Robustness
Full APR	90.8%	21.43%	78.57%
w/o Pixel-wise Attention	89.5%	28.76%	71.24%
w/o Adversarial Training	91.7%	35.21%	64.79%
w/o Regularization	90.2%	24.89%	75.11%

The ablation study reveals that each component of APR contributes to its overall effectiveness, with the pixel-wise attention mechanism providing the most significant boost in robustness.

E. Computational Overhead

We measured the computational overhead of APR compared to the baseline models:

Table 4: Computational overhead analysis

Model	Training Time Increase	Inference Time Increase
ResNet-18	27.3%	8.6%
VGG-16	31.5%	9.2%

While APR does introduce some computational overhead, the significant improvement in robustness justifies the additional cost, especially in security-critical applications.

F. Analysis of Attention Maps

To provide insight into how APR works, we analysed the attention maps generated by the pixel-wise attention mechanism. These maps reveal the regions of each image that the model focuses on most when making classifications, offering valuable insights into the model's decision-making process. Our analysis of the attention maps for CIFAR-10 images shows that APR learns to concentrate on key identifying features specific to each object class. For instance, when classifying airplane images, the model primarily focuses on the wings and fuselage. In animal images, such as cats or dogs, the attention is largely directed towards facial features. For vehicle images like cars or trucks, the model emphasizes wheels and overall shape. This pattern of attention distribution demonstrates that APR effectively learns to prioritize the most relevant parts of the image for classification. The model's focus on class-specific features suggests that it is developing a robust understanding of each category, rather than relying on potentially spurious details.

Importantly, the concentration of attention on these key features implies that the model may be less susceptible to single-pixel perturbations in less critical regions of the image. By focusing on broader, more stable features, APR potentially mitigates the impact of localized adversarial attacks like the One Pixel Attack. The attention patterns observed also align with human intuition about important features for object recognition. This alignment suggests that APR is learning meaningful representations that correspond to our understanding of these objects, which could contribute to its improved robustness. Overall, these visualizations provide strong evidence for the effectiveness of our APR approach. They illustrate how the model learns to focus on robust, class-specific features while potentially reducing its vulnerability to localized adversarial perturbations. This analysis not only validates the design principles behind APR but also offers interpretability into its decision-making process, which is crucial for building trust in AI systems deployed in security-critical applications.

G. Limitations and Future Work

While our Adaptive Pixel Resilience (APR) method demonstrates significant improvements in defending against One Pixel Attacks, it's important to acknowledge potential limitations:

1) *Computational Overhead:* As shown in Table 4, APR introduces additional computational costs during both training and inference. While the improved security justifies this overhead in many cases, it may be a constraint for resource-limited applications.
2) *Generalization to Other Attacks:* While APR is highly effective against One Pixel Attacks, its performance against other types of adversarial attacks (e.g., FGSM, PGD) needs further investigation.
3) *Dataset Specificity:* Our experiments focused on CIFAR-10 and ImageNet. The effectiveness of APR on other datasets or real-world, high-resolution images requires additional study.

Future work should address these limitations and explore:

1) Extending APR to defend against a broader range of adversarial attacks.
2) Investigating theoretical guarantees for the robustness provided by APR.
3) Optimizing the computational efficiency of APR to reduce overhead.
4) Evaluating APR's performance on a wider variety of datasets and real-world scenarios.
5) Exploring the trade-offs between robustness, accuracy, and model complexity in the context of APR and One Pixel Attacks.
6) Ethical Considerations

The development of defenses against adversarial attacks, such as APR, has important ethical implications:

1) *Dual-Use Potential:* While our research aims to enhance AI security, the knowledge gained could potentially be misused to develop more sophisticated attacks. We emphasize the importance of responsible disclosure and usage of this information.
2) *AI Safety:* Improving the robustness of AI systems against adversarial attacks is crucial for ensuring the safe deployment of AI in critical applications such as autonomous vehicles, medical diagnosis, and security systems.
3) *Privacy Concerns:* The attention maps generated by APR provide insights into model decision-making, which could have privacy implications if applied to sensitive data. Proper data handling and model deployment practices must be observed.
4) *Societal Impact:* As AI systems become more robust against attacks, it's crucial to consider the broader societal implications, including the potential for increased reliance on AI in decision-making processes.

We encourage ongoing dialogue and collaboration between researchers, ethicists, and policymakers to address these considerations and ensure the responsible development and deployment of AI defence mechanisms.

H. Broader Impact

The development of APR and our comprehensive study of One Pixel Attacks have several potential broader impacts:

1) *Enhanced AI Security:* By improving robustness against subtle adversarial attacks, APR contributes to the overall security of AI systems, potentially accelerating their adoption in critical applications.
2) *Interpretability in AI:* The attention map visualizations provided by APR offer insights into model decision-making, contributing to the broader goal of making AI systems more interpretable and trustworthy.
3) *Advancement in Adversarial Machine Learning:* Our work pushes the boundaries of understanding in adversarial attacks and defences, potentially inspiring new research directions in this rapidly evolving field.

4) *Industrial Applications:* The improved robustness offered by APR could be particularly valuable in industries such as autonomous vehicles, security systems, and medical imaging, where the stakes of AI failures are high.
5) *Educational Value:* This research provides a comprehensive overview of One Pixel Attacks and defence strategies, serving as a valuable educational resource for students and researchers entering the field of adversarial machine learning.
6) *Cross-disciplinary Collaboration:* The complex nature of adversarial attacks and defences encourages collaboration between various disciplines, including machine learning, cybersecurity, and cognitive science.

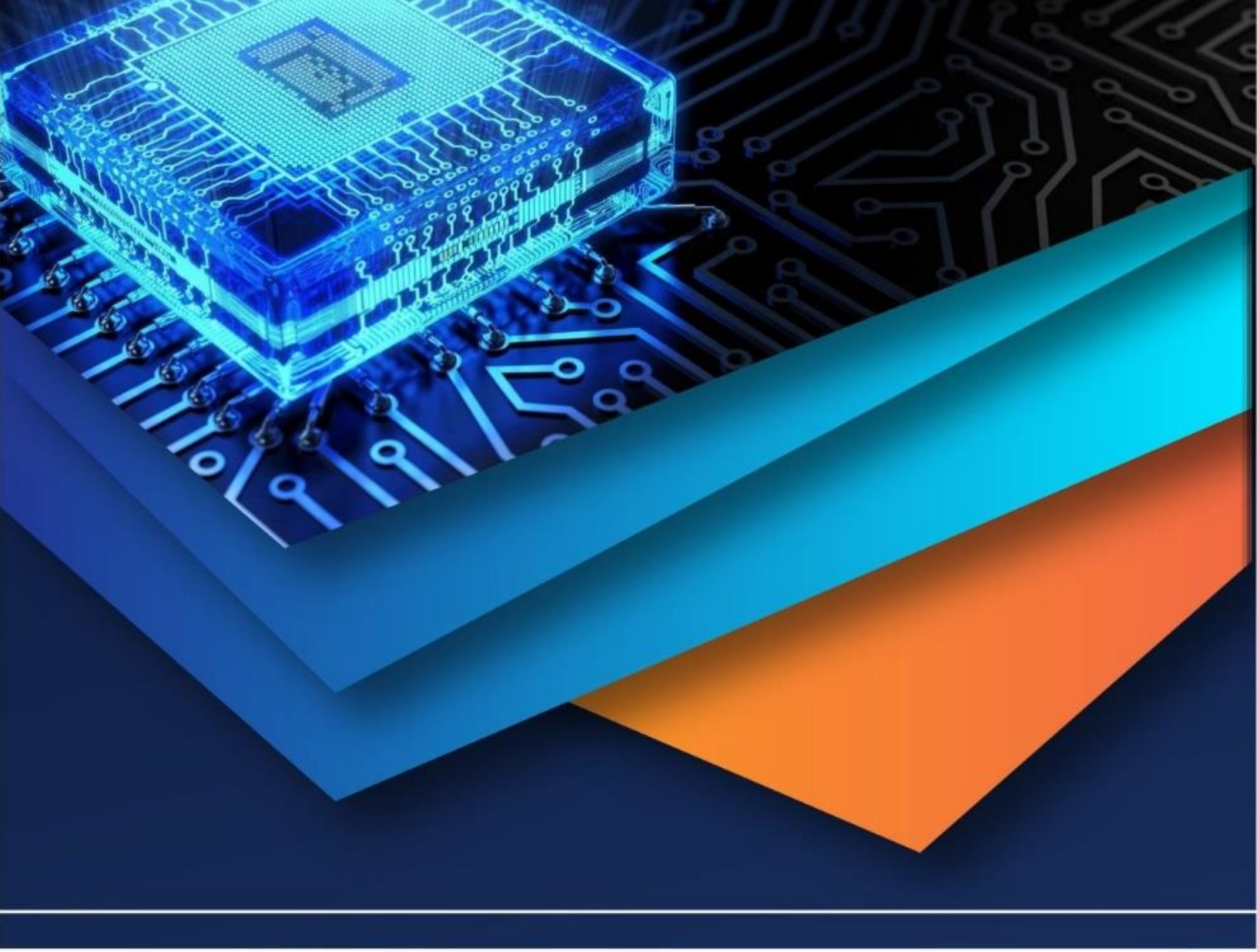
By addressing the critical challenge of AI robustness, this research contributes to the development of more reliable and trustworthy AI systems, potentially accelerating the positive impact of AI across various domains of society.

VI. CONCLUSION

This paper provides a comprehensive analysis of the One Pixel Attack and introduces Adaptive Pixel Resilience as a novel defence mechanism. Our approach combines adversarial training, pixel-wise attention, and regularization to enhance model robustness against this subtle yet powerful attack. Experimental results demonstrate that APR significantly outperforms existing defence methods against One Pixel Attacks on both CIFAR-10 and ImageNet datasets, maintaining high clean accuracy while substantially reducing the attack success rate. The visualizations of attention maps provide valuable insights into the decision-making process of models equipped with APR, enhancing interpretability and trust in the defence mechanism. While APR introduces some computational overhead, the significant improvement in robustness justifies its application in security-critical scenarios. We've also discussed the limitations of our approach, ethical considerations, and the broader impact of this research. As AI systems become increasingly prevalent in critical applications, the importance of robust defences against adversarial attacks cannot be overstated. By advancing our understanding of adversarial attacks and developing effective defences, we contribute to the creation of more secure and reliable AI systems. We encourage further research in this direction, emphasizing the need for continued innovation in AI security to keep pace with evolving threats.

REFERENCES

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
[3] 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.
[4] Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5), 828-841.
[5] Narodytska, N., & Kasiviswanathan, S. (2017). Simple black-box adversarial attacks on deep neural networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1310-1318). IEEE.
[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
[7] Guo, C., Rana, M., Cisse, M., & Van Der Maaten, L. (2018). Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.
[8] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defence to adversarial perturbations against deep neural networks. In 2016 IEEE Symposium on Security and Privacy (SP) (pp. 582-597). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY