ST Engineering Electronics

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

CYBER SECURITY LABORATORY

# Adaptive Noise Injection for Training Stochastic Student Networks from Deterministic Teachers

*Yi Xiang Marcus Tan, Yuval Elovici, Alexander Binder*

SUPPORTED BY

NATIONAL RESEARCH FOUNDATION
PRIME MINISTER'S OFFICE
SINGAPORE

# Background - Preamble

- Machine learning models are widely used to automate decision making processes
  - E.g. image classification

Image Classifier

"Otter"

# Background - Preamble

- However, such methods are known to be susceptible to adversarial attacks.

+ Specially crafted perturbation →

"Otter" "Monkey"

*Simple illustration of the effects of an adversarial attack*

# Background – Attacks Routines Used

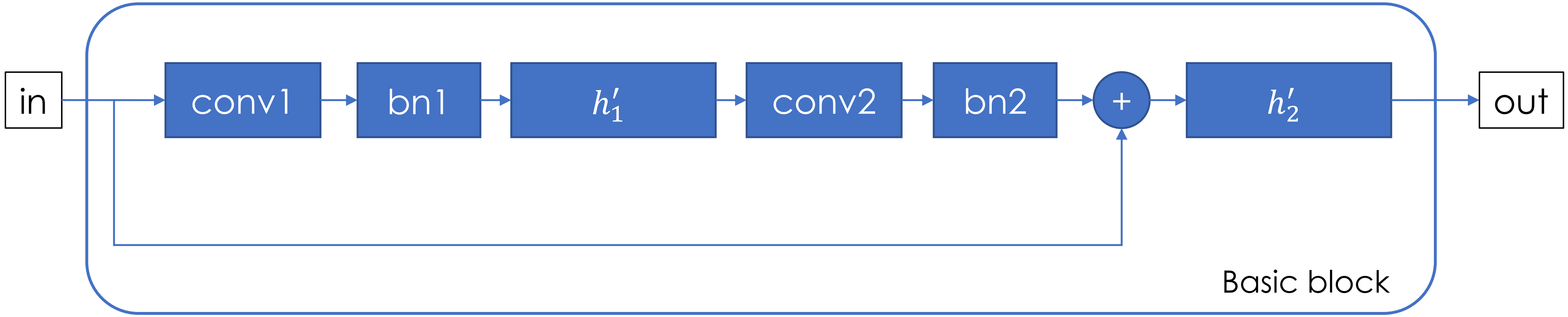
- We used several popular white-box attack routines
  1. Basic Iterative Method (BIM)
  2. Projected Gradient Descent (PGD)
  3. Momentum Iterative Method (MIM)
  4. Carlini & Wagner Attack (CW)
- A black-box attack routine was also used
  1. Boundary Attack (BA)

# Proposed Method

- Propose Adaptive Noise Injection Stochastic Students ($ANIS^2$) mechanism
  - Fine-tunes a deterministic network (teacher) to a stochastic variant (student)
  - Injects noise within *activation functions* with adaptive *stochasticity* during training
  - Using input data statistics based on Exponential Moving Average (EMA)
- *Different degrees of noise* are used at *different parts* of the network
  - Different hidden activation values across the network
- Trained in conjunction with Adversarial Training

# Proposed Method

- Denote our proposed activation block as StocReLUEMA, $h'(\cdot)$
- For an exemplary ResNet18 basic block:



# Proposed Method

- Let our StocReLUEMA be $h'(\cdot)$ and vanilla ReLU be $h(\cdot)$. At Layer $i$:

$$h'\left(x^{(i)}\right) = h(x^{(i)} + \delta^{(i)})$$

$$such\ that\ \delta^{(i)} \sim N(0, \gamma \cdot {\sigma^{(i)}}^2)$$

  - $\gamma$ increases as training epochs increases

- Adaptive noise injection tuned during training, updated after each batch $t$ via:

$$\sigma_{t+1}^{(i)} = (1 - \alpha) \cdot \sigma_t^{(i)} + \alpha \cdot STD_{chnwise}(x^{(i)})$$

# Proposed Method

- $\alpha$ set as 0.5
- Recall that StocReLUEMA:

$$h'(x^{(i)}) = h(x^{(i)} + \delta^{(i)})$$

$$such\ that\ \delta^{(i)} \sim N(0, \gamma \cdot {\sigma^{(i)}}^2)$$

**Algorithm 1:** Training with adaptive noise injector

**Input:** Teacher network's weights, $\theta_{teach}$; Max epochs, $T$; Initial $\gamma_{init}$; Max $\gamma_{max}$; Gamma update interval, $r$

**Output:** Student network's weights, $\theta_{student}$

Initialise stochastic student network with $\theta_{teach}$ and $\gamma_{init}$;

$k = r * (\gamma_{max} - \gamma_{init})/T$;

**for** $t = 1, ..., T$ **do**

  Get mini-batch from training data $B = \{(x_1, y_1), ..., (x_m, y_m)\}$;

  **for** $j = 1, ..., m$ **do**

    Perform standard training routine with adversarial training on mini-batch;

    Update $\sigma$ in each stochastic layer, $\sigma_{new} = (1-\alpha)*\sigma_{old} + \alpha * STD_{chnwise}(input)$

  **end**

  **if** $t$ mod $r = 0$ **then**

    $\gamma = \gamma + k$

  **end**

**end**

# Baselines Used

1. Adversarial Training (AT)
   - Trains model on adversarial samples generated with correct labels
2. TRADES
   - Introduce a regularisation term that encourages adversarial robustness
3. Learn2Perturb (L2P)
   - Introducing noise parameters as learnable parameters for the network
   - Trained with AT

# Baseline Classification Results

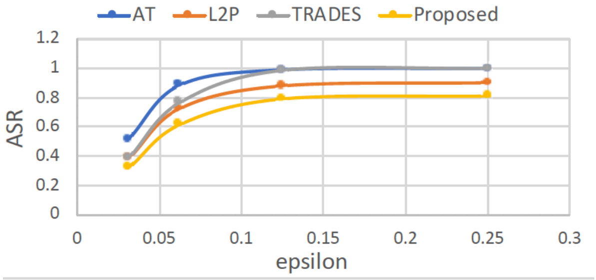
| Defence Methods | CIFAR-10 | CIFAR-100 |
|---|---|---|
| None | 0.940 | 0.760 |
| AT | 0.846 | 0.574 |
| L2P | 0.859 | 0.566 |
| TRADES | 0.809 | 0.594 |
| **$ANIS^2$ *(Proposed)*** | **0.829** | **0.575** |

Classification accuracy of the respective approaches on clean CIFAR-10 and CIFAR-100 test data. "None" indicates standard training without any defence introduced. Higher is better.

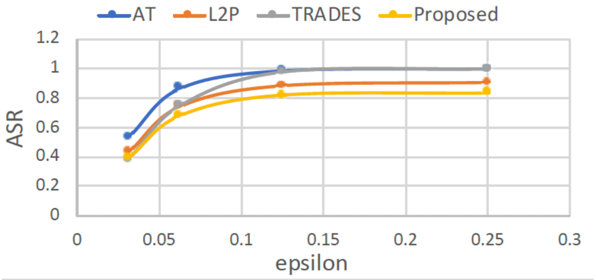
# White-box Attack Results

- We report the Adversarial Success Rate (ASR). More specifically:

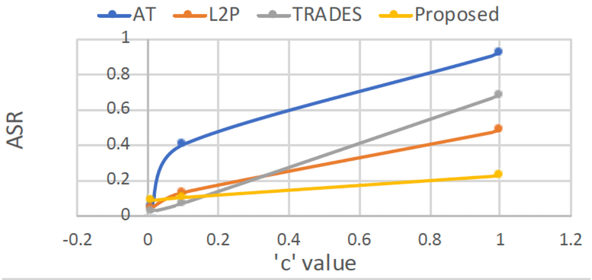
$$ASR = \mathbb{E}_{X,Y\sim D}\{P(f(x+\delta) \neq Y \,||f(x) = Y\}$$



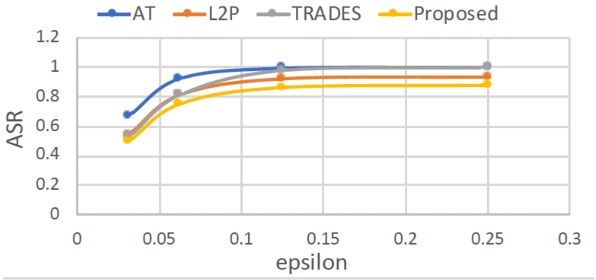

(a) BIM attack; CIFAR-10



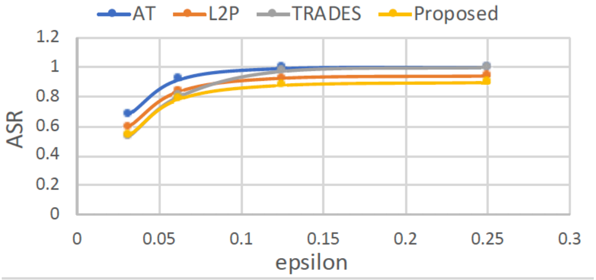

(b) MIM attack; CIFAR-10



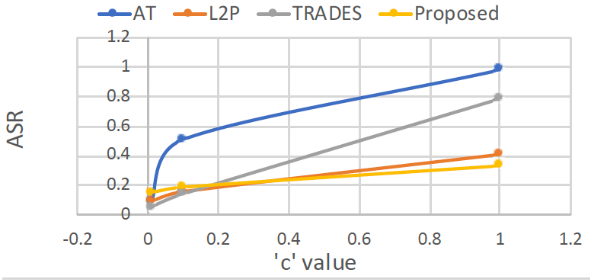

(c) CW attack; CIFAR-10




(d) BIM attack; CIFAR-100

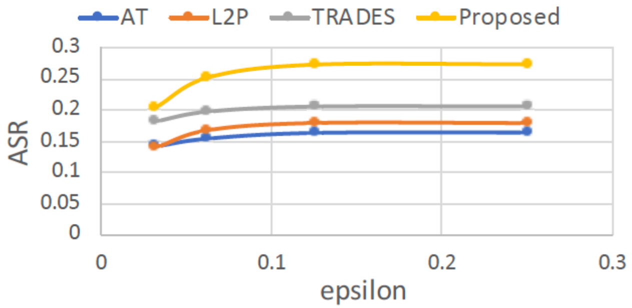



(e) MIM attack; CIFAR-100

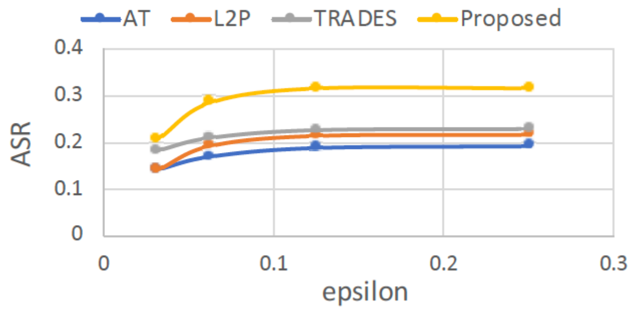



(f) CW attack; CIFAR-100

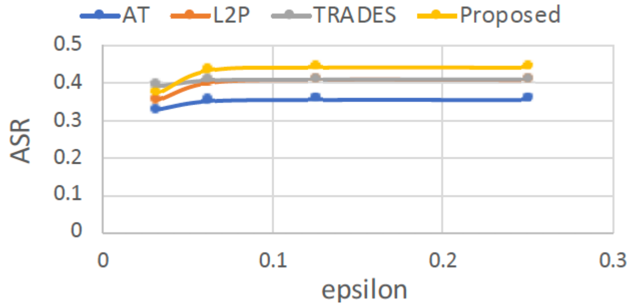
# Black-box Attack Results



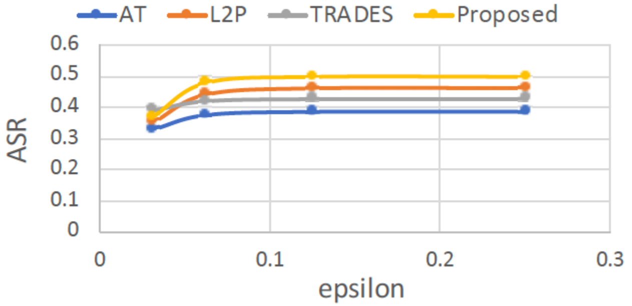

BIM attack on teacher model; CIFAR-10




MIM attack on teacher model; CIFAR-10




BIM attack on teacher model; CIFAR-100




MIM attack on teacher model; CIFAR-100

- Black-box *transferability* attack
  - Generate on teacher, launched against student
- Due to weights initialization policy
  - Proposed VS the rest

# Black-box Attack Results

- L2P and $ANIS^2$ show high robustness to decision-based black-box attacks

| Defence Method | CIFAR-10 | CIFAR-100 |
|---|---|---|
| AT | 0.758 | 0.818 |
| L2P | **0.022** | **0.036** |
| TRADES | 0.942 | 0.768 |
| $ANIS^2$ *(Proposed)* | **0.048** | **0.064** |

ASR against the various defence methods when launching BA across CIFAR-10 and CIFAR-100. 500 samples were used. Lower is better.

# Decision Boundary Evolution During Training

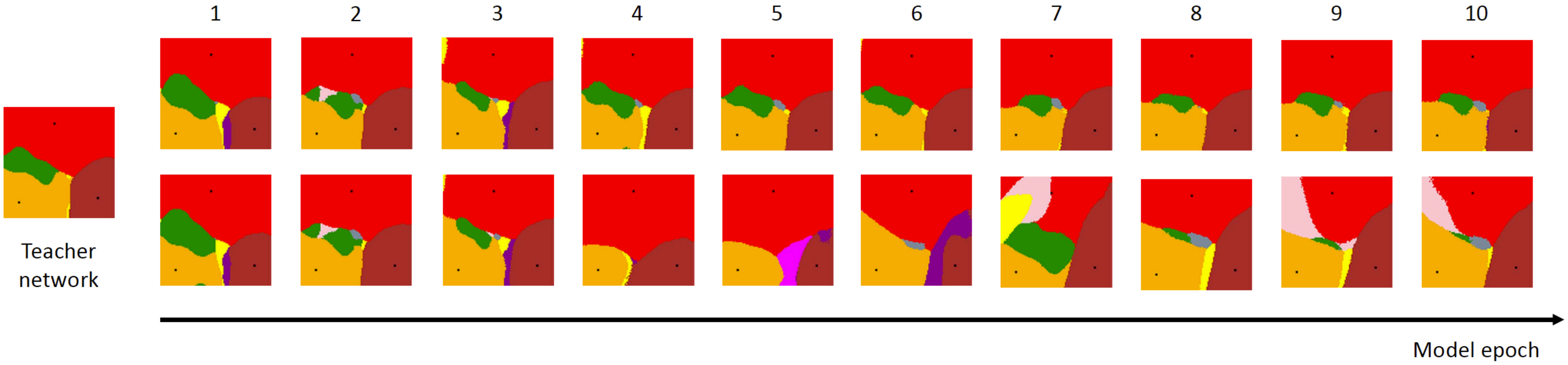
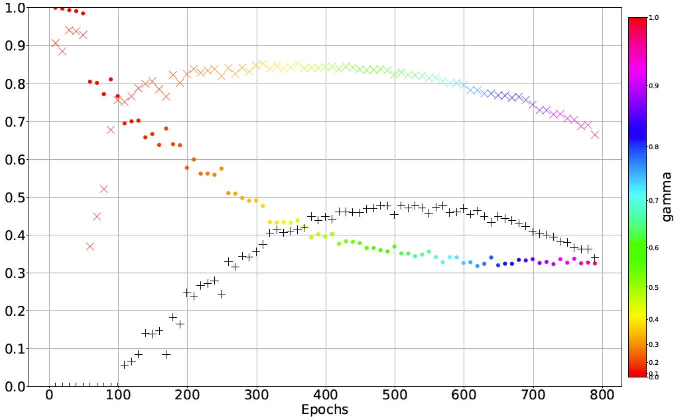


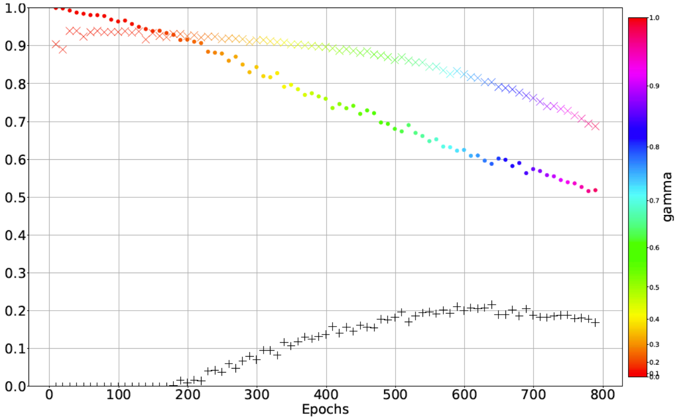

Illustration of prediction labels of an exemplary region based on three data points (black dots) of CIFAR-10. **Top row:** our stochastic student trained with $ANIS^2$ *WITHOUT* adversarial training. **Bottom row:** our stochastic student trained with $ANIS^2$ *WITH* adversarial training, introduced from the fourth image onward.

# Ablation Study

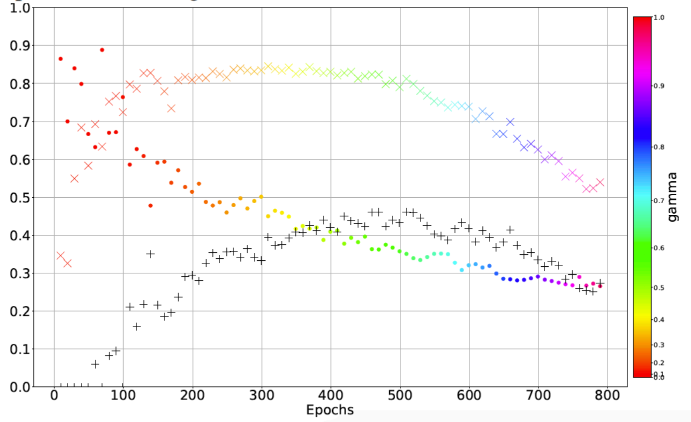
- Varied the following factors:
    - Presence of AT
    - Presence of EMA
    - Presence of teacher-initialisation
- Coloured ‘x’ – clean accuracy
- Coloured ‘.’ – ASR
- Black ‘+’ - max(ACC – ASR, 0)



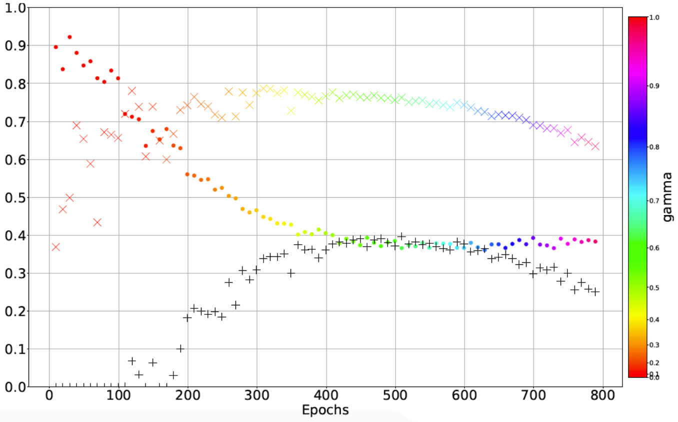

(a) Original proposed method with adversarial training, EMA-based adaptive noise tuning mechanism and teacher-initialised student.




(b) Without adversarial training.




(c) Without EMA-based adaptive noise tuning mechanism.




(d) Without teacher-initialisation of student.

# Conclusion

- Propose $ANIS^2$, conceptually simple EMA-based adaptive noise injection mechanism
  - Can be applied to any layer
- Able to outperform baselines in robustness under white-box attack settings
- AT as finetuning allows adaptation to new features
  - Exemplified by evolution of decision boundary
- EMA to adapt noise prevents sharp degradation in clean accuracy while providing smooth trends
- Stochasticity should be used as a complement instead of a substitute

# Selected References


1. A. Jeddi, M. J. Shafiee, M. Karg, C. Scharfenberger, and A. Wong, "Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1241–1250.
2. H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically Principled Trade-off between Robustness and Accuracy," Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp 7472-7482
3. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
4. A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.
5. Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.
6. N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 IEEE symposium on security and privacy (sp). IEEE, 2017, pp. 39–57.
7. A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in International Conference on Machine Learning, 2018, pp. 284–293.
8. W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in International Conference on Learning Representations, 2018.