SCIENCE CHINA
Information Sciences

• **Supplementary File** •

MAD: Meta Adversarial Defense Benchmark

Xiaoxu PENG[1], Dong ZHOU[1*], Guanghui SUN[1], Jiaqi SHI[1] & Ligang WU[1]

[1]*Department of Control Science and Engineering, Harbin Institute of Technology, Harbin* 150001*, China*

* Corresponding author (email: dongzhou@hit.edu.cn)

Appendix A Attack IDs

Table A1 Adversarial attack algorithms

ID	Name	Measurement	Knowledge	Introduction
0	JSMA [53]	L_2	White-box	Constructing an adversarial saliency map involves generating input features that have the most significant influence on the output.
1	Deep-Fool [54]	L_2	White-box	Calculating the minimum distance between normal examples and the decision boundary of the model in order to generate perturbations.
2	Universal-Perturbation [55]	L_∞	White-box	Searching for general perturbations at training points, aggregate perturbation vectors that send successful data points to decision boundaries.
3	Newton-Fool [56]	L_2/L_0	White-box	Generating AEs based on model gradient.
4	Boundary-Attack [57]	L_2	Black-box	Reducing the size of disturbances gradually starting fromlarge disturbances while maintaining aggressiveness based on decision boundary.
5	Elastic Net [58]	L_1	White-box	Expressing the adversarial attack problem as the elastic network regularization optimization problem.
6	Zoo-Attack [59]	L_0	Black-box	Generating AEs based on the gradient of the zero-order optimization estimation target model.
7	Spatial-Trans-formation [60]	L_N [1)]	Black-box	Generating perturbation based on natural perturbation categories such as translation.
8	Hop-Skip-Jump [61]	L_∞/L_2	Black-box	Generating perturbation gradient direction estimation at decision boundaries based on binary information.
9	Sim-BA [62]	L_2	Black-box	Adding sample vector from a predefined orthogonal offset on the original image.
10	Shadow-Attack [63]	L_N [1)]	White-box	Keeping the adversarial example far away from the decision-making boundary under the premise that it is not perceivable.
11	GeoDA [64]	L_∞	Black-box	Black-box iterative attack algorithm based on query.
12	Wasserstein [65]	L_N [1)]	White-box	Searching for adversarial perturbations of Wasserstein distance based on Sinkhorn iteratively.
13	FGSM [41]	L_∞	White-box	Adding perturbations in reverse based on the gradient direction of normal examples.
14	BIM [66]	L_∞	White-box	Adversarial sample attacks against the physical world.
15	CW [67]	L_∞	White-box	Transforming the problem of finding the smallest perturbation that causes a neural network to misclassify into a convex optimization problem.
16	MIFGSM [68]	L_∞	White-box	Generating perturbations by an iterative approach based on momentum to find counter perturbations.
17	TIFGSM [69]	L_∞	White-box	Generating a more transferable perturbations by optimizing the perturbation of the image transformation set.
18	PGD [13]	L_∞	White-box	Generating adversarial perturbation based on gradient projection direction iterative algorithm.
19	PGD-L2 [13]	L_2	White-box	Generating L2 perturbation based on gradient projection direction iterative algorithm.
20	TPGD [44]	L_∞	White-box	Generating adversarial perturbations by an iterative algorithm of gradient projection direction based on KL-Divergence loss.
21	RFGSM [70]	L_∞	White-box	A single-step gradient attack method based on small random step size.
22	APGD [38]	L_∞/L_2	White-box	Generating adversarial perturbation an iterative algorithm of gradient projection direction based on variable step size with loss “ce”.
23	APGD2 [38]	L_∞/L_2	White-box	Generating adversarial perturbation an iterative algorithm of gradient projection direction based on variable step size with loss “dlr”.
24	FFGSM [22]	L_∞	White-box	Gradient single-step attack method based on random initialization.
25	Square [71]	L_∞/L_2	Black-box	An iterative algorithm to find feasible set boundaries against perturbations by updating local squares at random locations.
26	TIFGSM2 [69]	L_∞	White-box	Generating a more transferable perturbations by optimizing the perturbation of the image transformation set with different resize rate.
27	EOTPGD [72]	L_∞	White-box	An iterative method for generating AEs by estimating the direction of gradient projection by multiple random vectors.
28	One-Pixel [73]	L_0	Black-box	Generating single-pixel perturbations based on differential evolution.
29	FAB [74]	$L_\infty/L_2/L_1$	White-box	Generating minimal perturbation based on fast adaptive boundaries.

1) L_N indicates the unconventional perturbation measurement.

Appendix B Attacked Datasets

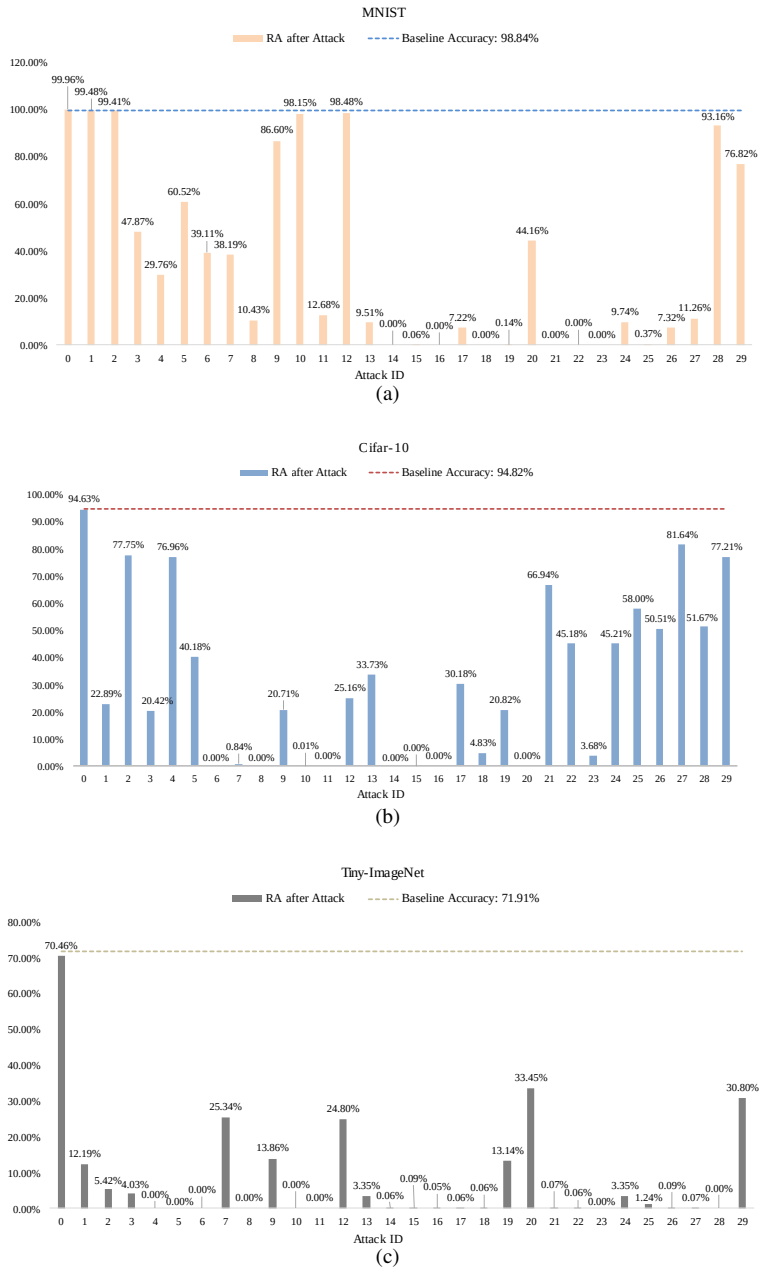

Figure B1 Robust accuracy (RA) of the initial validation datasets under various attacks. Dashed lines indicate the baseline standard accuracy (SA) for each dataset.

References

1 Chen H, Yuan K, Huang Y, et al. Feedback is all you need: from ChatGPT to autonomous driving. Sci China Inf Sci, 2023, 66: 166201
2 Wu X, Tao R, Hong D, et al. The FrFT convolutional face: toward robust face recognition using the fractional Fourier transform and convolutional neural networks. Sci China Inf Sci, 2020, 63: 119103
3 Wang Z, Du Y, Wei K, et al. Vision, application scenarios, and key technology trends for 6G mobile communications. Sci China Inf Sci, 2022, 65: 151301
4 Akhtar N, Mian A, Kardan N, et al. Advances in adversarial attacks and defenses in computer vision: A survey. IEEE Access, 2021, 9: 155161–155196
5 Tao G, Ma S, Liu Y, et al. Attacks meet interpretability: Attribute-steered detection of adversarial examples. Adv Neural Inf Process Syst, 2018, 31
6 Li S, Zhu S, Paul S, et al. Connecting the dots: Detecting adversarial perturbations using context inconsistency. In: Proceedings of the European Conference on Computer Vision (ECCV), 2020, Glasgow, UK, 2020. 396–413
7 Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations. 2017. ArXiv:1711.00117
8 Raff E, Sylvester J, Forsyth S, et al. Barrage of random transforms for adversarially robust defense. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 6528–6537
9 Zhai R, Dan C, He D, et al. Macer: Attack-free and scalable robust training via maximizing certified radius. 2020. ArXiv:2001.02378
10 Jia J, Cao X, Wang B, et al. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. 2019. ArXiv:1912.09899
11 Cemgil T, Ghaisas S, Dvijotham K, et al. Adversarially robust representations with smooth encoders. In: Proceedings of the International Conference on Learning Representations (ICLR), 2019
12 He Z, Rakin AS, Li J, et al. Defending and harnessing the bit-flip based adversarial weight attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 14095–14103
13 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. 2017. ArXiv:1706.06083
14 Pang T, Yang X, Dong Y, et al. Bag of tricks for adversarial training. 2020. ArXiv:2010.00467
15 Peng SY, Xu W, Cornelius C, et al. Robust principles: Architectural design principles for adversarially robust cnns. 2023. ArXiv:2308.16258
16 Tang S, Gong R, Wang Y, et al. Robustart: Benchmarking robustness on architecture design and training techniques. 2021. ArXiv:2109.05211
17 Liu C, Dong Y, Xiang W, et al. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. 2023. ArXiv:2302.14301
18 Liu G, Zhang W, Li X, Fan K, Yu S. VulnerGAN: a backdoor attack through vulnerability amplification against machine learning-based network intrusion detection systems. Sci China Inf Sci, 2022, 65: 170303.
19 LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86(11): 2278–2324
20 Yang L, Jiang H, Cai R, et al. CondenseNetV2: sparse feature reactivation for deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3569–3578
21 Ye M, Huang W, Shi Z, Li H, Du B. Revisiting federated learning with label skew: an over-confidence perspective. Sci China Inf Sci, 2025, 68(9): 192102
22 Wong E, Rice L, Kolter JZ. Fast is better than free: Revisiting adversarial training. 2020. ArXiv:2001.03994
23 Shafahi A, Najibi M, Ghiasi MA, et al. Adversarial training for free! Adv Neural Inf Process Syst, 2019, 32
24 Zheng H, Zhang Z, Gu J, et al. Efficient adversarial training with transferable adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1181–1190
25 Zhang D, Zhang T, Lu Y, et al. You only propagate once: Accelerating adversarial training via maximal principle. Adv Neural Inf Process Syst, 2019, 32
26 Farnia F, Zhang JM, Tse D. Generalizable adversarial training via spectral normalization. 2018. ArXiv:1811.07457
27 Grabinski J, Jung S, Keuper J, et al. FrequencyLowCut Pooling–Plug & Play against Catastrophic Overfitting. In: Proceedings of the European Conference on Computer Vision (ECCV), 2022. 36-57
28 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the International Conference on Machine Learning (ICML), 2017. 1126–1135
29 Goldblum M, Fowl L, Goldstein T. Adversarially robust few-shot learning: A meta-learning approach. Adv Neural Inf Process Syst, 2020, 33: 17886–17895
30 Liu F, Zhao S, Dai X, et al. Long-term cross adversarial training: A robust meta-learning method for few-shot classification tasks. 2021. ArXiv:2106.12900
31 Qi J, Zhang R, Li C, et al. Cross domain few-shot learning via meta adversarial training. 2022. ArXiv:2202.05713
32 Ma C, Zhao C, Shi H, et al. Metaadvdet: Towards robust detection of evolving adversarial attacks. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019. 692–701
33 Metzen JH, Finnie N, Hutmacher R. Meta adversarial training against universal patches. 2021. ArXiv:2101.11453
34 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
35 Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 6105–6114
36 Nicolae MI, Sinn M, Tran MN, et al. Adversarial Robustness Toolbox v1.0.0. 2018. ArXiv:1807.01069
37 Ding GW, Wang L, Jin X. AdverTorch v0.1: An Adversarial Robustness Toolbox based on PyTorch. 2019. ArXiv:1902.07623
38 Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of the International Conference on Machine Learning (ICML), 2020. 2206–2216
39 Rice L, Wong E, Kolter JZ. Overfitting in adversarially robust deep learning. In: Proceedings of the International Conference on Machine Learning (ICML), 2020. 8093–8104
40 Rusu AA, Rabinowitz NC, Desjardins G, et al. Progressive neural networks. 2016. ArXiv:1606.04671
41 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014. ArXiv:1412.6572
42 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of

the 25th International Conference on Neural Information Processing Systems (NeurIPS), 2012. 1097–1105
43 Miyato T, Maeda S, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2018, 41(8): 1979–1993
44 Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 7472–7482
45 Jiang H, He P, Chen W, et al. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. 2019. ArXiv:1911.03437
46 Xie G, Shangguan A, Fei R, et al. Motion trajectory prediction based on a CNN-LSTM sequential model. Sci China Inf Sci, 2020, 63: 1–21
47 Qiaoben Y, Ying C, Zhou X, et al. Understanding adversarial attacks on observations in deep reinforcement learning. Sci China Inf Sci, 2024, 67(5): 1–15
48 Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness. 2019. ArXiv:1902.06705
49 Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization. 2017. ArXiv:1710.09412
50 DeVries T, Taylor G W. Improved regularization of convolutional neural networks with cutout. 2017. ArXiv:1708.04552
51 Zha J, Fan Y, Li K, et al. Decoupled multi-hierarchy Kalman filter for 3D object tracking. 2025. ArXiv:2505.12340
52 Yang L, Zheng Z, Chen B, et al. Nullu: mitigating object hallucinations in large vision–language models via HalluSpace projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. 14635–14645
53 Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 2016. 372–387
54 Moosavi-Dezfooli S-M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2574–2582
55 Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1765–1773
56 Jang U, Wu X, Jha S. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In: Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC), 2017. 262–277
57 Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. 2017. ArXiv:1712.04248
58 Chen P, Sharma Y, Zhang H, et al. Ead: elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018. 32(1)
59 Chen P, Zhang H, Sharma Y, et al. 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, 2017. 15–26
60 Engstrom L, Tran B, Tsipras D, et al. Exploring the landscape of spatial robustness. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 1802–1811
61 Chen J, Jordan MI, Wainwright MJ. Hopskipjumpattack: A query-efficient decision-based attack. In: Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP), 2020. 1277–1294
62 Guo C, Gardner J, You Y, et al. Simple black-box adversarial attacks. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 2484–2493
63 Ghiasi A, Shafahi A, Goldstein T. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. 2020. ArXiv:2003.08937
64 Rahmati A, Moosavi-Dezfooli S-M, Frossard P, et al. Geoda: a geometric framework for black-box adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 8446–8455
65 Wong E, Schmidt F, Kolter Z. Wasserstein adversarial examples via projected sinkhorn iterations. In: Proceedings of the International Conference on Machine Learning (ICML), 2019. 6808–6817
66 Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018. 99–112
67 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), 2017. 39–57
68 Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 9185–9193
69 Dong Y, Pang T, Su H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 4312–4321
70 Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses. 2017. ArXiv:1705.07204
71 Andriushchenko M, Croce F, Flammarion N, et al. Square attack: a query-efficient black-box adversarial attack via random search. In: Proceedings of the 16th European Conference on Computer Vision (ECCV), 2020. 484–501
72 Liu X, Li Y, Wu C, et al. Adv-bnn: Improved adversarial defense through robust bayesian neural network. 2018. ArXiv:1810.01279
73 Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput, 2019, 23(5): 828–841
74 Croce F, Hein M. Minimally distorted adversarial examples with a fast adaptive boundary attack. In: Proceedings of the International Conference on Machine Learning (ICML), 2020. 2196–2205