FEDERATION UNIVERSITY

TABLE I: List of data poisoning attack driven by mathematical perturbation function

No	Attack Name	Mathematical Function	Defence
1	DPA-M-PGD	PGD[29][35][36]	Certified Robust[32]
2	DPA-M-Auto-PGD	Auto-PGD[15][14]	WSNNS[20]
3	DPA-M-LL-PGD	LL-PGD[31]	WSNNS[20]
4	DPA-M-PGD Iterative	PGD Iterative[45]	Vector Defence[28]
5	DPA-M-PGD-Single Shot	PGD-Single Shot[27]	Vector Defence[28]
6	DPA-M-MT-Linf/MT-L2	MT-Linf/MT-L2[25]	Adversarial Training[8]
7	DPA-M-L-BFGS	BFGS[21]	APE-GAN[44]
9	DPA-M-FGSM	FGSM[2]	FGSM Counter
10	DPA-M-LL-FGSM	LL-FGSM(Step-LL)[47]	Prakash et al.[40]
11	DPA-M-ADA-FGSM	ADA-FGSM[45]	Carrara et al.[7]
12	DPA-M-IFGSM(MI-Linf/MI-L2)	IFGSM(MI-Linf/MI-L2)[15]	Prakash et al.[40]
13	DPA-M-MI	MI[15]	Adversarial Training[8]
14	DPA-M-MI-FGSM	MI-FGSM(Momentum Iterative)[42]	Mustafa et al.[39]
15	DPA-M-TGSM	TGSM[41]	Feature Distillation*[34]
16	DPA-M-IFGSM	IFGSM	SAP[17]
17	DPA-M-ZOO	ZOO[11]	Hybrid Random Forest[18]
18	DPA-M-cADV	cADV Colorisation attack[4]	JPEG defence[16]
19	DPA-M-tAdv	tADV texture transfer attack	JPEG defence[16]
20	DPA-M-StAdv	Spatial Transformation	Adversarial Training[8]
21	DPA-M-BIM	BIM(Iterative FGSM)[29]	Progressive Defence [48]
22	DPA-M-BIM-A	BIM-A[29]	Vector Defence[28]
23	DPA-M-BIM-B	BIM-B[29]	Vector Defence[28]
24	DPA-M-FFF	Fast Feature Fool [38]	Adversarial Training[8]
25	DPA-M-ILCM	Iterative Least-likely class method[29]	Adversarial Training[8]
26	DPA-M-BIM	Momentum BIM[39]	Mustafa[39]
27	DPA-M-Shadow Attack	Semantic spoofed certificates[22]	Mustafa [39]
28	DPA-M-JSMA	Gradient Based[24]	Vector Defence[28]
29	DPA-M-NTM	Metamorphic Relation Based [9]	AT [30]
30	DPA-M-MGA	Momentum Gradient Based[10]	Vector Defence[28]
31	DPA-M-WitchCraft	Gaussian Noise[12]	Certified Robustness [32]
32	DPA-M-QL Attack	Gradient Estimation[26]	Adversarial Training[8]
33	DPA-M-Basic	Least-Likely-Class Iterative Methods[2]	Adversarial Training[8]
34	DPA-M-One Pixel	One Pixel[46]	Pixel Defend [43]
35	DPA-M-Momentum Iterative	Momentum Iterative[19]	Super resolution [39]
36	DPA-M-JigSaw Attack	UAP[37]	Adversarial Training[8]
37	DPA-M-UPSET and ANGRI	UPSET and ANGRI[1]	Adversarial Training[8]
38	DPA-M-Houdini	Houdini[13]	Adversarial Training[8]
39	DPA-M-ATN	AAE-ATN[3]	Adversarial Training[8]
40	DPA-M-SimBA	SimBA[23]	Randomisation[14]
41	DPA-M-SimBA-DCT	SimBA-DCT[26]	Randomisation[14]
42	DPA-M-Patch Attack	Generated Patch[33]	Pixel Defend [43]
43	DPA-M-Adversarial Patch	Adversarial Patch[15]	Pixel Defend [43]
44	DPA-M-DPatch	DPatch[23]	Pixel Defend [43]
45	DPA-M-Carlini & Wagner	C&W [6]	Stochastic Elements[5]
46	DPA-M-IFS	IFS[23]	Adversarial Training[8]
47	DPA-M-QL Attack	QL[26]	Adversarial Training[8]
48	DPA-M-QeBB	QeBB[29]	Adversarial Training[8]

REFERENCES

[1] Naveed Akhtar, Jian Liu, and Ajmal Mian. "Defense against universal adversarial perturbations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3389–3398.

[2] Naveed Akhtar and Ajmal Mian. *Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey*. 2018.

[3] Shumeet Baluja and Ian Fischer. *Adversarial Transformation Networks: Learning to Generate Adversarial Examples*. 2017. arXiv: 1703.09387 [cs.NE].

[4] Anand Bhattad et al. *Unrestricted Adversarial Examples via Semantic Manipulation*. 2020. arXiv: 1904.06347 [cs.CV].

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. *Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models*. 2018. arXiv: 1712.04248 [stat.ML].

[6] Nicholas Carlini and David Wagner. *Towards Evaluating the Robustness of Neural Networks*. 2017. arXiv: 1608.04644 [cs.CR].

[7] Fabio Carrara et al. "Adversarial image detection in deep neural networks". In: *Multimedia Tools and Applications* 78.3 (2019), pp. 2815–2835.

[8] Anirban Chakraborty et al. *Adversarial Attacks and Defences: A Survey*. 2018. arXiv: 1810.00069 [cs.LG].

[9] Alvin Chan et al. *Metamorphic Relation Based Adversarial Attacks on Differentiable Neural Computer*. 2018. DOI: 10.48550/ARXIV.1809.02444. URL: https://arxiv.org/abs/1809.02444.

[10] Jinyin Chen et al. *MGA: Momentum Gradient Attack on Network*. 2020. DOI: 10.48550/ARXIV.2002.11320. URL: https://arxiv.org/abs/2002.11320.

[11] Pin-Yu Chen 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, pp. 15–26.

[12] Ping-Yeh Chiang et al. *WITCHcraft: Efficient PGD attacks with random step size*. 2019. DOI: 10.48550/ARXIV.1911.07989. URL: https://arxiv.org/abs/1911.07989.

[13] Moustapha Cisse et al. *Houdini: Fooling Deep Structured Prediction Models*. 2017. arXiv: 1707.05373 [stat.ML].

[14] Francesco Croce and Matthias Hein. *Mind the box: l_1-APGD for sparse adversarial attacks on image classifiers*. 2021. arXiv: 2103.01208 [cs.LG].

[15] Francesco Croce and Matthias Hein. *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks*. 2020. arXiv: 2003.01690 [cs.LG].

[16] Nilaksh Das et al. "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression". In: *arXiv preprint arXiv:1705.02900* (2017).

[17] Guneet S. Dhillon et al. *Stochastic Activation Pruning for Robust Adversarial Defense*. 2018. arXiv: 1803.01442 [cs.LG].

[18] Yifan Ding et al. "Defending against adversarial attacks using random forest". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.

[19] Yinpeng Dong et al. *Boosting Adversarial Attacks with Momentum*. 2017. DOI: 10.48550/ARXIV.1710.06081. URL: https://arxiv.org/abs/1710.06081.

[20] Abhimanyu Dubey et al. *Defense Against Adversarial Images using Web-Scale Nearest-Neighbor Search*. 2019. arXiv: 1903.01612 [cs.CV].

[21] Ibrahim Gashaw and H L Shashirekha. "Machine Learning Approaches for Amharic Parts-of-speech Tagging". In: *arXiv preprint arXiv:2001.03324* (2020).

[22] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. *Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates*. 2020. DOI: 10.48550/ARXIV.2003.08937. URL: https://arxiv.org/abs/2003.08937.

[23] Anteneh Girma, Mosses Garuba, and Rajini Goel. "Advanced Machine Language Approach to Detect DDoS Attack Using DBSCAN Clustering Technology with Entropy". In: *Information Technology - New Generations*. Ed. by Shahram Latifi. Cham: Springer International Publishing, 2018, pp. 125–131. ISBN: 978-3-319-54978-1.

[24] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. "Adversarial and clean data are not twins". In: *arXiv preprint arXiv:1704.04960* (2017).

[25] Sven Gowal et al. *On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models*. 2019. arXiv: 1810.12715 [cs.LG].

[26] Chuan Guo et al. “Simple black-box adversarial attacks”. In: *arXiv preprint arXiv:1905.07121* (2019).

[27] Yunseok Jang et al. “Adversarial defense via learning to generate diverse attacks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 2740–2749.

[28] Vishaal Munusamy Kabilan, Brandon Morris, and Anh Nguyen. *VectorDefense: Vectorization as a Defense to Adversarial Examples*. 2018. arXiv: 1804.08529 [cs.CV].

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. “Adversarial machine learning at scale”. In: *arXiv preprint arXiv:1611.01236* (2016).

[30] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. *Perceptual Adversarial Robustness: Defense Against Unseen Threat Models*. 2021. arXiv: 2006.12655 [cs.LG].

[31] Alfred Laugros, Alice Caplier, and Matthieu Ospici. *Are Adversarial Robustness and Common Perturbation Robustness Independent Attributes ?* 2019. arXiv: 1909.02436 [cs.LG].

[32] Mathias Lecuyer et al. *Certified Robustness to Adversarial Examples with Differential Privacy*. 2019. arXiv: 1802.03471 [stat.ML].

[33] Xiang Li and Shihao Ji. *Generative Dynamic Patch Attack*. 2021. DOI: 10.48550/ARXIV.2111.04266. URL: https://arxiv.org/abs/2111.04266.

[34] Zihao Liu et al. “Feature distillation: DNN-oriented JPEG compression against adversarial examples”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2019, pp. 860–868.

[35] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML].

[36] Aleksander Madry et al. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017).

[37] Seyed-Mohsen Moosavi-Dezfooli et al. *Universal adversarial perturbations*. 2017. arXiv: 1610.08401 [cs.CV].

[38] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. “Fast feature fool: A data independent approach to universal adversarial perturbations”. In: *arXiv preprint arXiv:1707.05572* (2017).

[39] Aamir Mustafa et al. “Image Super-Resolution as a Defense Against Adversarial Attacks”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 1711–1724. ISSN: 1941-0042. DOI: 10.1109/tip.2019.2940533.

[40] Aaditya Prakash et al. *Deflecting Adversarial Attacks with Pixel Deflection*. 2018. arXiv: 1801.08926 [cs.CV].

[41] Andrew Slavin Ross and Finale Doshi-Velez. *Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients*. 2017. arXiv: 1711.09404 [cs.LG].

[42] Vivek B. S. and R. Venkatesh Babu. *Single-step Adversarial training with Dropout Scheduling*. 2020. arXiv: 2004.08628 [cs.LG].

[43] Ali Shafahi et al. *Label Smoothing and Logit Squeezing: A Replacement for Adversarial Training?* 2019. arXiv: 1910.11585 [cs.LG].

[44] Shiwei Shen et al. "Ape-gan: Adversarial perturbation elimination with gan". In: *arXiv preprint arXiv:1707.05474* (2017).

[45] Yucheng Shi et al. "Adaptive iterative attack towards explainable adversarial robustness". In: *Pattern Recognition* 105 (2020), p. 107309. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2020.107309. URL: http://www.sciencedirect.com/science/article/pii/S0031320320301138.

[46] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One Pixel Attack for Fooling Deep Neural Networks". In: *IEEE Transactions on Evolutionary Computation* 23.5 (Oct. 2019), pp. 828–841. ISSN: 1941-0026. DOI: 10.1109/tevc.2019.2890858. URL: http://dx.doi.org/10.1109/TEVC.2019.2890858.

[47] Florian Tramèr et al. *Ensemble Adversarial Training: Attacks and Defenses*. 2020. arXiv: 1705.07204 [stat.ML].

[48] Ling Wang et al. *Progressive Defense Against Adversarial Attacks for Deep Learning as a Service in Internet of Things*. 2020. arXiv: 2010.11143 [cs.CR].