DPA DRIVEN BY ALGORITHM TABLE

REFERENCES

[1] Maksym Andriushchenko et al. "Square attack: a query-efficient black-box adversarial attack via random search". In: *European Conference on Computer Vision*. Springer. 2020, pp. 484–501.

[2] Battista Biggio et al. "Evasion Attacks against Machine Learning at Test Time". In: *Lecture Notes in Computer Science* (2013), pp. 387–402. ISSN: 1611-3349. DOI: 10.1007/978-3-642-40994-3_25. URL: http://dx.doi.org/10.1007/978-3-642-40994-3_25.

[3] Franziska Boenisch, Philip Sperl, and Konstantin Böttinger. *Gradient Masking and the Underestimated Robustness Threats of Differential Privacy in Deep Learning*. 2021. arXiv: 2105.07985 [cs.CR].

[4] 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].

[5] Jacob Buckman et al. "Thermometer encoding: One hot way to resist adversarial examples". In: *International Conference on Learning Representations*. 2018.

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

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

[8] 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.

[9] Gabriela F. Cretu et al. "Casting out Demons: Sanitizing Training Data for Anomaly Sensors". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008, pp. 81–95. DOI: 10.1109/SP.2008.11.

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

TABLE I: List of data poisoning attack driven by algorithm

No	Algorithm Name	Algorithm	Defence
1	DPA-A-APGD	APGD[12][10]	Differential Approximation[10]
2	DPA-A-PPGD	PPGD[28]	PAT[28]
3	DPA-A-Cassidi	Cassidi[28]	PAT[28]
4	DPA-A-Deepfool	Deelfool[37][38]	Divide - Denoise[39]
5	DPA-A-LPA	LPA[27]	Trades[27]
6	DPA-A-Fast-LPA	Fast-LPA[27]	Trades[27]
7	DPA-A-Square Attack	Square Attack[1][21]	Bandlimiting * [30]
8	DPA-A-AutoAttack	Auto Attack[12]	Stochastic Elements [4]
9	DPA-A-NewtonFool	NewtonFool[42][45][41]	Adversarial Training[7]
10	DPA-A-R-FGSM	Rand-FGSM[52]	Adversarial Training[7]
11	DPA-A-N-FGSM	N-FGSM[47]	Adversarial Training[7]
12	DPA-A-Fast-FGSM	FAST-FGSM[52]	Adversarial Training[7]
13	DPA-A-Rapid-FGSM	Rapid-FGSM[47]	Adversarial Training[7]
14	DPA-A-Robust-FGSM	Robust$_F GSM[47]$	JPEG Compression[32]
15	DPA-A-UAP	UAP Universal Adversarial Perturbation [26]	Shardped Edges [14]
16	DPA-A-TUAP	Targeted Universal Adversarial Perturbation	Adversarial Training [7] [40]
17	DPA-A-TUAP-DeepFool	TUAP - DeepFool	Adversarial Retraining [40]
18	DPA-A-TUAP-CW	TUAP-CW	Adversarial Training[7]
19	DPA-A-DFO	Stochastic Derivative Free Optimization[35]	Adversarial Retraining [40]
20	DPA-A-CW	CW-$L0$[6]	Vectro Defence[22] PixelDefend[51]
21	DPA-A-CW	-$L2$[6]	Vectro Defence[22] PixelDefend[51]
22	DPA-A-CW	CW-$L\infty$[6]	Vectro Defence[22] PixelDefend[51]
23	DPA-A-AdvPreprocessing	Image Scaling [16][44]	Robust scaling algorithm and Image reconstruction [44]
24	DPA-ShadowAttack	Shadow Attack [17]	Random Smoothing Certified Defence* [17]
25	DPA-A-Biggio	Biggio Poisonning [2]	Adversarial Training[7]
26	DPA-A-FrogsAttack	Frogs Poisonning [49]	Data Sanitizing* [9]
27	DPA-A-Salt-Pepper	Salt and Pepper [33]	Adversarial Training[7]
28	DPA-A-SignHunter	Momentum Gradient Based [15]	Randomisation [30]
29	DPA-A-FastMN	Fast Minimum-norm (FMN) Attack[43]	Adversarial Training[7]
30	DPA-A-FAB	Minimally distorted with a Fast Adaptive[11]	Adversarial Training[7]
31	DPA-A-BB	Minimally distorted with a Fast Adaptive[11]	Adversarial Training[7]
32	DPA-A-KKT Based	KKT[25]	Adversarial Training[7]
33	DPA-A-Square Attack	L1$-APGDAnd$L1$-AutoAttack(APGD-AT)$[1][21]	Logit Squeezing* [48],Pixel Defend [48]
34	PIA (partial Information Attack)	(QLA variation)[20]	Logit pairing [23]
35	DPA-A-JSMA-F	JSMA-F[6]	Vector Defence[22]
36	DPA-JSMA-Z	JSMA[6]	Vectro Defence[22]
37	DPA-A-JPEG-Linf	JPEG-L_p	JPEG Compression* [13]
38	DPA-A-ReColorAdv	$ReColorAdv$[27]	PAT [28]
39	DPA-A-SimBA (simple black box attack)	$L1$-APGD And $L1$-AutoAttack(APGD-AT)[18]	Pixel Defend [18]
40	DPA-A-SimBA-DCT (simple black box attack)	(SimBA variation)[18]	Pixel Defend [48]
41	DPA-A-Parsimonious(Efficient Combinatorial Optimization)	$L1$-APGD And $L1$-AutoAttack (APGD-AT), Single and Multi APGD[36]	Randomisation[10]
42	DPA-A-DFO -(1+1)-ES	DFO variation-(1+1)-ES[35]	Adversarial Retraining [40]
43	DPA-A-DFO-CMA-ES	DFO variation CMA-ES[35]	Adversarial Retraining [40]
44	DPA-A-Bandits	Bandits [19]	Logit Squeezing* [48]
45	DPA-A-Bandits$_T$	Bandits$_T$ [19]	Logit Squeezing* [48]
46	DPA-A-Bandits$_T D$	Bandits$_T D$ [19]	Logit Squeezing* [48]
47	DPA-A-NES	NES[53]	Augmented Adv Training [4]
48	NES-GE	NES-GE[20]	Augmented Adv Training [4]
49	NES-PIA	NES-PIA[20]	Augmented Adv Training [4]
50	DPA-A-ZOO Attack [31]	ZOO Attack [31]	Shardped Edges[14]
51	DPA-A-ZOO-SGD	ZOO-SGD[31]	Stochastic Element [14]
52	DPA-A-ZOO-SignSGD	ZOO-SignSGD[31]	Stochastic Element [14]
53	DPA-A-ZOO-M-signSGD	ZO-M-signSGD[31]	Stochastic Element [14]
54	DPA-A-ZOO-NES	ZOO-NES[31]	Stochastic Element [14]
55	DPA-A-ZOO-SCD	ZOO-SCD[31]	Stochastic Element [14]
56	DPA-A-FMN	FMN[43]	Adversarial Training[7]
57	DPA-A-Semantic Attack	Semantic[17] [34]	Adversarial Training[7]
58	DPA-A-Discretized Inputs	Discrete Gradient Ascent PGD / PGA[29]	One Hot [5]
59	DPA-A-CROWN-IBP	Shadow-Penalties[17]	Random Smoothing Certified Defence* [17]
60	DPA-A-BPDA	BPDA (Gradient Free) [55]	Adversarial Training[7]
61	DPA-A-BNN-GA	BNN-GA(Gradient Free) [55]	Adversarial Training[7]
62	BNN-ZOO	BNN-ZOO (Gradient Free) [55]	Stochastic Element [14]
63	DPA-A-Koh-Liang attack	Koh-Liang[24]	Adversarial Training[7]
64	DPA-A-ZOO-ADAM	ZOO-ADAM[8]	Gradient Masking [3]
65	DPA-A-ZOO-Newton	ZOO-Newton[8]	Gradient Masking [3]
66	DPA-A-SADS	Saddle Point[46]	Byzantine-Robust Distribution[54]
67	DPA-A-FMN	Fast Minimum-norm[43]	Adversarial Training[7]
68	DPA-A-Physical Attack	Recusrsive Impersonation[50]	Adversarial Training[7]

[11] Francesco Croce and Matthias Hein. "Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack". In: Proceedings of Machine Learning Research 119 (13–18 Jul 2020). Ed. by Hal Daumé III and Aarti Singh, pp. 2196–2205. URL: https://proceedings.mlr.press/v119/croce20a.html.

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

[13] Nilaksh Das et al. "SHIELD: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression". In: (2018).

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

[15] Abdullah Al-Dujaili and Una-May O'Reilly. "Sign Bits Are All You Need for Black-Box Attacks". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=SygW0TEFwH.

[16] Yue Gao and Kassem Fawaz. *Scale-Adv: A Joint Attack on Image-Scaling and Machine Learning Classifiers*. 2021. arXiv: 2104.08690 [cs.LG].

[17] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. *Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates*. 2020. arXiv: 2003.08937 [cs.LG].

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

[19] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. *Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors*. 2019. arXiv: 1807.07978 [stat.ML].

[20] Andrew Ilyas et al. *Black-box Adversarial Attacks with Limited Queries and Information*. 2018. arXiv: 1804.08598 [cs.CV].

[21] Matthew Jagielski et al. "High Accuracy and High Fidelity Extraction of Neural Networks". In: *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2020.

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

[23] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. *Adversarial Logit Pairing*. 2018. arXiv: 1803.06373 [cs.LG].

[24] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: *International conference on machine learning*. PMLR. 2017, pp. 1885–1894.

[25] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. *Stronger Data Poisoning Attacks Break Data Sanitization Defenses*. 2018. arXiv: 1811.00741 [stat.ML].

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

[27] Cassidy Laidlaw and Soheil Feizi. *Functional Adversarial Attacks*. 2019. arXiv: 1906.00001 [cs.LG].

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

[29] Qi Lei et al. *Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification*. 2018. DOI: 10.48550/ARXIV.1812.00151. URL: https://arxiv.org/abs/1812.00151.

[30] Yuping Lin, Kasra Ahmadi K. A., and Hui Jiang. *Bandlimiting Neural Networks Against Adversarial Attacks*. 2019. arXiv: 1905.12797 [cs.LG].

[31] Sijia Liu et al. "signSGD via Zeroth-Order Oracle". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=BJe-DsC5Fm.

[32] 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.

[33] Pratyush Maini, Eric Wong, and Zico Kolter. "Adversarial robustness against the union of multiple perturbation models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6640–6650.

[34] Jan Hendrik Metzen et al. "On detecting adversarial perturbations". In: *arXiv preprint arXiv:1702.04267* (2017).

[35] Laurent Meunier, Jamal Atif, and Olivier Teytaud. *Yet another but more efficient black-box adversarial attack: tiling and evolution strategies*. 2019. arXiv: 1910.02244 [cs.LG].

[36] Seungyong Moon, Gaon An, and Hyun Oh Song. *Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization*. 2019. arXiv: 1905.06635 [cs.LG].

[37] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. *DeepFool: a simple and accurate method to fool deep neural networks*. 2016. arXiv: 1511.04599 [cs.LG].

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2574–2582.

[39] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel. *Divide, Denoise, and Defend against Adversarial Attacks*. 2019. arXiv: 1802.06806 [cs.CV].

[40] Elior Nehemya et al. "Taking Over the Stock Market: Adversarial Perturbations Against Algorithmic Traders". In: ().

[41] Maria-Irina Nicolae et al. "Adversarial Robustness Toolbox v1. 0.0". In: *arXiv preprint arXiv:1807.01069* (2018).

[42] Anay Pattanaik et al. "Robust deep reinforcement learning with adversarial attacks". In: *arXiv preprint arXiv:1712.03632* (2017).

[43] Maura Pintor et al. *Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints*. 2021. arXiv: 2102.12827 [cs.LG].

[44] Erwin Quiring et al. "Adversarial Preprocessing: Understanding and Preventing Image-Scaling Attacks in Machine Learning". In: *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2020.

[45] Rajeev Ranjan et al. "Improving network robustness against adversarial attacks with compact convolution". In: *arXiv preprint arXiv:1712.00699* (2017).

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

[47] Leo Schwinn, René Raab, and Björn Eskofier. *Towards Rapid and Robust Adversarial Training with One-Step Attacks*. 2020. arXiv: 2002.10097 [cs.LG].

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

[49] Ali Shafahi et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6103–6113.

[50] Mahmood Sharif et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition". In: *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*. 2016.

[51] Yang Song et al. *PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples*. 2018. arXiv: 1710.10766 [cs.LG].

[52] Florian Tramèr et al. "Stealing machine learning models via prediction apis". In: *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 2016, pp. 601–618.

[53] Daan Wierstra et al. "Natural Evolution Strategies". In: (2014).

[54] Dong Yin et al. "Defending Against Saddle Point Attack in Byzantine-Robust Distributed Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 7074–7084.

[55] Matthew Yuan, Matthew Wicker, and Luca Laurenti. *Gradient-Free Adversarial Attacks for Bayesian Neural Networks*. 2020. arXiv: 2012.12640 [cs.LG].