Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, 2014.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine, pp. 82–97, 2012.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, 2014.

[7] J. Fauw, J. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nature Medicine, 2018.

[8] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in Proceedings of the 10th ACM Conference on Recommender Systems, p. 191–198, 2016.

[9] S. Tobiyama, Y. Yamaguchi, H. Shimada, T. Ikuse, and T. Yagi, "Malware detection with deep neural network using process behavior," in 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), 2016.

[10] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," in NIPS Deep Learning Symposium, 2016.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in International Conference on Learning Representations, 2014.

[12] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," arXiv preprint arXiv:1605.07277, 2016.

[13] M. Naseer, S. Khan, M. H. Khan, F. Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," in Advances in Neural Information Processing Systems, 2019.

[14] L. Batina, S. Bhasin, D. Jap, and S. Picek, "Csi nn: Reverse engineering of neural network architectures through electromagnetic side channel," in 28th USENIX Security Symposium (USENIX Security 19), 2019.

[15] R. Joud, P.-A. Moellic, R. Bernhard, and J.-B. Rigaud, "A review of confidentiality threats against embedded neural network models," arXiv preprint arXiv:2105.01401, 2021.

[16] M. Dumont, P.-A. Moellic, R. Viera, J.-M. Dutertre, and R. Bernhard, "An overview of laser injection against embedded neural network models," arXiv preprint arXiv:2105.01403, 2021.

[17] B. Brik, A. Ksentini, and M. Bouaziz, "Federated learning for uavs-enabled wireless networks: Use cases, challenges, and open problems," IEEE Access, vol. 8, pp. 53841–53849, 2020.

[18] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, and et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3.

[19] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018.

[20] C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in Proceedings of the 36th International Conference on Machine Learning, 2019.

[21] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 2012.

[22] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 27–38, ACM, 2017.

[23] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

[24] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., University of Toronto, 2009.

[25] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[27] D. Sgandurra, L. Muñoz-González, R. Mohsen, and E. C. Lupu, "Automated dynamic analysis of ransomware: Benefits, limitations and use for detection," arXiv preprint arXiv:1609.03020, 2016.

[28] C. L. Blake and C. J. Merz, "Uci repository of machine learning databases," 1998.

[29] W. Guo, B. Tondi, and M. Barni, "A master key backdoor for universal impersonation attack against dnn-based face verification," Pattern Recognition Letters.

[30] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," in NIPS Workshop on Mach. Learn. and Comp. Sec, 2017.

[31] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," arXiv preprint arXiv:1912.02771, 2019.

[32] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in European Conference on Computer Vision, 2020.

[33] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in Proceedings of the 23rd USENIX Conference on Security Symposium, 2014.

[34] S. Mehnaz, N. Li, and E. Bertino, "Black-box model inversion attribute inference attacks on classification models," arXiv preprint arXiv:2012.03404, 2020.

[35] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[36] X. Zhao, W. Zhang, X. Xiao, and B. Y. Lim, "Exploiting explanations for model inversion attacks," arXiv preprint arXiv:2104.12669, 2021.

[37] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.

[38] M. Wu, X. Zhang, J. Ding, H. Nguyen, R. Yu, M. Pan, and S. T. Wong, "Evaluation of inference attack models for deep learning on medical data," arXiv preprint arXiv:2011.00177, 2020.

[39] "The general social survey." `https://gss.norc.org/`.

[40] V. Feldman, "Does learning require memorization? a short tale about a long tail," in Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, 2020.

[41] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in 28th USENIX Security Symposium (USENIX Security 19), 2019.

[42] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al., "Extracting training data from large language models," arXiv preprint arXiv:2012.07805, 2020.

[43] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in Proceedings of the 15th European Conference on Machine Learning, 2004.

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., "Language models are unsupervised multitask learners," OpenAI blog, 2019.

[45] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High accuracy and high fidelity extraction of neural networks," in 29th USENIX Security Symposium (USENIX Security 20), 2020.

[46] N. Carlini, M. Jagielski, and I. Mironov, "Cryptanalytic extraction of neural network models," arXiv:2003.04884 [cs], 2020.

[47] S. J. Oh, B. Schiele, and M. Fritz, "Towards reverse-engineering black-box neural networks," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019.

[48] S. Maji, U. Banerjee, and A. P. Chandrakasan, "Leaky nets: Recovering embedded neural network models and inputs through simple power and timing side-channels–attacks and defenses," IEEE Internet of Things Journal, 2021.

[49] W. Hua, Z. Zhang, and G. E. Suh, "Reverse engineering convolutional neural networks through side-channel information leaks," in 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC), 2018.

[50] U. Gupta, D. Stripelis, P. K. Lam, P. Thompson, J. L. Ambite, and G. V. Steeg, "Membership inference attacks on deep regression models for neuroimaging," in Medical Imaging with Deep Learning, 2021.

[51] S. P. Liew and T. Takahashi, "Faceleaks: Inference attacks against transfer learning models via black-box queries," arXiv preprint arXiv:2010.14023, 2020.

[52] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," arXiv preprint arXiv:1709.01604, 2017.

[53] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Security and Privacy (SP), 2017 IEEE Symposium on, pp. 3–18, IEEE, 2017.

[54] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated white-box membership inference," in 29th USENIX Security Symposium (USENIX Security 20), 2020.

[55] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jegou, "White-box vs black-box: Bayes optimal strategies for membership inference," in Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, pp. 5558–5567, 2019.

[56] C. A. C. Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," arXiv preprint arXiv:2007.14321, 2020.

[57] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in Advances in Neural Information Processing Systems, pp. 125–136, 2019.

[58] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.

[59] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[60] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling cnns with simple transformations," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, pp. 1802–1811, 2019.

[61] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "Colorfool: Semantic adversarial colorization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[62] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, 2019.

[63] M. Cheng, S. Singh, P. Chen, P.-Y. Chen, S. Liu, and C.-J. Hsieh, "Sign-opt: A query-efficient hard-label adversarial attack," in International Conference on Learning Representations, 2020.

[64] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519, ACM, 2017.

[65] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," arXiv preprint arXiv:1611.03814, 2016.

[66] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge examples: Energy-latency attacks on neural networks," 2020.

[67] I. Shumailov, Z. Shumaylov, D. Kazhdan, Y. Zhao, N. Papernot, M. A. Erdogdu, and R. Anderson, "Manipulating sgd with data ordering attacks," arXiv preprint arXiv:2104.09667, 2021.

[68] K. Grosse, T. A. Trost, M. Mosbach, M. Backes, and D. Klakow, "On the security relevance of initial weights in deep neural networks," in International Conference on Artificial Neural Networks, pp. 3–14, Springer, 2020.

[69] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011.

[70] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[71] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9, 2015.

[72] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III, (Berlin, Heidelberg), pp. 387–402, Springer-Verlag, 2013.

[73] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in International Conference on Learning Representations, Wiley Online Library, 2015.

[74] "A complete list of all (arxiv) adversarial example papers." https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html.

[75] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in International Conference on Learning Representations, 2018.

[76] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14, ACM, 2017.

[77] N. Carlini and D. Wagner, "Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples," arXiv preprint arXiv:1711.08478, 2017.

[78] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proceedings of the 35th International Conference on Machine Learning, ICML 2018, 2018.

[79] J. Uesato, B. O'Donoghue, A. v. d. Oord, and P. Kohli, "Adversarial risk and the dangers of evaluating against weak attacks," in Proceedings of the 35th International Conference on Machine Learning, ICML 2018, pp. 5025–5034, 2018.

[80] A. Athalye and N. Carlini, "On the robustness of the cvpr 2018 white-box adversarial example defenses," arXiv preprint arXiv:1804.03286, 2018.

[81] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," arXiv preprint arXiv:1902.06705, 2019.

[82] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in Advances in Neural Information Processing Systems, 2020.

[83] E. Wong, F. R. Schmidt, and J. Z. Kolter, "Wasserstein adversarial examples via projected sinkhorn iterations," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, pp. 6808–6817, 2019.

[84] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," in Advances in Neural Information Processing Systems 31, 2017.

[85] F. Karim, S. Majumdar, and H. Darabi, "Adversarial attacks on time series," IEEE transactions on pattern analysis and machine intelligence, 2020.

[86] Y. Zang, F. Qi, C. Yang, Z. Liu, M. Zhang, Q. Liu, and M. Sun, "Word-level textual adversarial attacking as combinatorial optimization," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[87] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," 2020.

[88] S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," arXiv preprint arXiv:2010.03593, 2020.

[89] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, pp. 1310–1320, 2019.

[90] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," in Advances in Neural Information Processing Systems 32, 2019.

[91] J. Z. Kolter and E. Wong, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 2019.

[92] S. Singla and S. Feizi, "Second-order provable defenses against adversarial attacks," in Proceedings of the 37th International Conference on Machine Learning, 2020.

[93] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Adversarial attacks on deep neural networks for time series classification," 2019 International Joint Conference on Neural Networks (IJCNN).

[94] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[95] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in Proceedings of the 36th International Conference on Machine Learning, 2019.

[96] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "Waveguard: Understanding and mitigating audio adversarial examples," in 30th USENIX Security Symposium (USENIX Security 21), 2021.

[97] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," Workshop of International Conference on Learning Representations (ICLR), 2017.

[98] H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh, "Robust deep reinforcement learning against adversarial perturbations on state observations," in Advances in Neural Information Processing Systems, 2020.

[99] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in Proceedings of the 35th International Conference on Machine Learning, 2018.

[100] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in International Conference on Learning Representations, 2016.

[101] C. Sitawarin, A. N. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang, "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos," 1st Deep Learning and Security Workshop (IEEE S&P 2018), 2018.

[102] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Simple physical adversarial examples against end-to-end autonomous driving models," in 2019 IEEE International Conference on Embedded Software and Systems (ICESS), 2019.

[103] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," arXiv preprint arXiv:1803.04683, 2018.

[104] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[105] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57, IEEE, 2017.

[106] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: elastic-net attacks to deep neural networks via adversarial examples," in The Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[107] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in The IEEE International Conference on Computer Vision (ICCV), 2019.

[108] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in 2020 IEEE Symposium on Security and Privacy (SP), 2020.

[109] T. Tanay and L. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," arXiv preprint arXiv:1608.07690, 2016.

[110] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147, ACM, 2017.

[111] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," in International Conference on Learning Representations, 2018.

[112] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in International Conference on Learning Representations, 2018.

[113] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in International Conference on Learning Representations, 2019.

[114] T. Zhang and Z. Zhu, "Interpreting adversarially trained convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning, 2019.

[115] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[116] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[117] C. Etmann, S. Lunz, P. Maass, and C. Schoenlieb, "On the connection between adversarial robustness and saliency map interpretability," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, pp. 1823–1832, 2019.

[118] S. Kaur, J. Cohen, and Z. C. Lipton, "Are perceptually-aligned gradients a general property of robust classifiers?," in Advances in Neural Information Processing Systems, 2019.

[119] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," arXiv preprint arXiv:1906.00945, 2019.

[120] F. Tramèr and D. Boneh, "Adversarial training and robustness for multiple perturbations," in Advances in Neural Information Processing Systems 33, 2019.

[121] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Security and Privacy (EuroS&P), 2016 IEEE European Symposium on, pp. 372–387, IEEE, 2016.

[122] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[123] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582, 2016.

[124] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," arXiv preprint arXiv:1801.02610, 2018.

[125] S. Wang, Y. Chen, A. Abdou, and S. Jana, "Enhancing gradient-based attacks with symbolic intervals," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 2019.

[126] Y. Tashiro, Y. Song, and S. Ermon, "Diversity can be transferred: Output diversification for white- and black-box attacks," in Advances in Neural Information Processing Systems 33, 2020.

[127] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "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, pp. 15–26, ACM, 2017.

[128] J. C. Spall et al., "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," IEEE transactions on automatic control, vol. 37, no. 3, pp. 332–341, 1992.

[129] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 2019.

[130] 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.

[131] T. Maho, T. Furon, and E. Le Merrer, "Surfree: A fast surrogate-free black-box attack," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[132] M. Cheng, T. Le, P.-Y. Chen, H. Zhang, J. Yi, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in International Conference on Learning Representations, 2019.

[133] S. Moon, G. An, and H. O. Song, "Parsimonious black-box adversarial attacks via efficient combinatorial optimization," in Proceedings of the 36th International Conference on Machine Learning, pp. 4636–4645, 2019.

[134] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," in Optimization Techniques, 1978.

[135] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in Proceedings of the 35th International Conference on Machine Learning, ICML 2018, 2018.

[136] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao, "Blacklight: Defending black-box adversarial attacks on deep neural networks," arXiv preprint arXiv:2006.14042, 2020.

[137] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2018.

[138] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[139] C. Xie, Z. Zhang, J. Wang, Y. Zhou, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[140] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," in The IEEE International Conference on Computer Vision (ICCV), 2019.

[141] W. Wu, Y. Su, X. Chen, S. Zhao, I. King, M. R. Lyu, and Y.-W. Tai, "Boosting the transferability of adversarial samples via attention," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[142] T. Huang, V. Menkovski, Y. Pei, Y. Wang, and M. Pechenizkiy, "Direction-aggregated attack for transferable adversarial examples," arXiv preprint arXiv:2104.09172, 2021.

[143] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, "Transferable adversarial perturbations," in Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[144] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," in International Conference on Learning Representations, 2020.

[145] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen, "Feature space perturbations yield more transferable adversarial examples," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[146] J. M. Springer, M. Mitchell, and G. T. Kenyon, "Uncovering universal features: How adversarial training improves adversarial transferability," 2021.

[147] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 2019.

[148] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," arXiv preprint arXiv:1703.00410, 2017.

[149] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," arXiv preprint arXiv:1702.06280, 2017.

[150] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1787, 2018.

[151] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[152] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in International Conference on Computer Aided Verification, 2017.

[153] V. Tjeng, K. Y. Xiao, and R. Tedrake, "Evaluating robustness of neural networks with mixed integer programming," in International Conference on Learning Representations, 2019.

[154] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, T. Mann, and P. Kohli, "On the effectiveness of interval bound propagation for training verifiably robust models," arXiv preprint arXiv:1810.12715, 2018.

[155] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in Advances in Neural Information Processing Systems, pp. 2266–2276, 2017.

[156] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," 2018.

[157] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 854–863, JMLR. org, 2017.

[158] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Fixing data augmentation to improve adversarial robustness," arXiv preprint arXiv:2103.01946, 2021.

[159] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," in Advances in Neural Information Processing Systems 33, 2019.

[160] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in International Conference on Learning Representations, 2020.

[161] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 2020.

[162] P. Maini, E. Wong, and J. Z. Kolter, "Adversarial robustness against the union of multiple perturbation models," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 2020.

[163] F. Croce and M. Hein, "Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model," arXiv preprint arXiv:2105.12508, 2021.

[164] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in International Conference on Learning Representations, 2020.

[165] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in International Conference on Learning Representations, 2021.

[166] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "Stylized adversarial defense," arXiv preprint arXiv:2007.14672, 2020.

[167] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein, "Adversarially robust distillation," Proceedings of the AAAI Conference on Artificial Intelligence, p. 3996–4003, 2020.

[168] F. Liu, R. Zhao, and L. Shi, "Adversarial feature stacking for accurate and robust predictions," arXiv preprint arXiv:2103.13124, 2021.

[169] C. Song, K. He, J. Lin, L. Wang, and J. E. Hopcroft, "Robust local features for improving the generalization of adversarial training," in International Conference on Learning Representations, 2020.

[170] D. Stutz, M. Hein, and B. Schiele, "Confidence-calibrated adversarial training: Generalizing to unseen attacks," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 2020.

[171] A. Chan, Y. Tay, Y. S. Ong, and J. Fu, "Jacobian adversarially regularized networks for robustness," in International Conference on Learning Representations, 2020.

[172] S. Addepalli, B. Vivek, A. Baburaj, G. Sriramanan, and R. Venkatesh Babu, "Towards achieving adversarial robustness by enforcing feature consistency across bit planes," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[173] F. Lin, R. Mittapali, P. Chattopadhyay, D. Bolya, and J. Hoffman, "Likelihood landscapes: A unifying principle behind many adversarial defenses," in Adversarial Robustness in the Real World (AROW), ECCV, 2020.

[174] R. Bernhard, P.-A. Moellic, M. Mermillod, Y. Bourrier, R. Cohendet, M. Solinas, and M. Reyboz, "Impact of spatial frequency based constraints on adversarial robustness," in 2021 International Joint Conference on Neural Networks (IJCNN), 2021.

[175] J. Jo and Y. Bengio, "Measuring the tendency of cnns to learn surface statistical regularities," arXiv preprint arXiv:1711.11561, 2017.

[176] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in International Conference on Learning Representations, 2019.

[177] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition," Psychological science, vol. 5, no. 4, pp. 195–200, 1994.

[178] M. Mermillod, P. Bonin, L. Mondillon, D. Alleysson, and N. Vermeulen, "Coarse scales are sufficient for efficient categorization of emotional facial expressions: Evidence from neural computation," Neurocomputing, vol. 73, no. 13-15, pp. 2522–2531, 2010.

[179] R. M. French, M. Mermillod, A. Chauvin, P. C. Quinn, and D. Mareschal, "The importance of starting blurry: Simulating improved basic-level category learning in infants due to weak visual acuity," in Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 24, 2002.

[180] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau, "Shield: Fast, practical defense and vaccination for deep learning using jpeg compression," Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018.

[181] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: Dnn-oriented jpeg compression against adversarial examples," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[182] Z. Zhang, C. Jung, and X. Liang, "Adversarial defense by suppressing high-frequency components," arXiv preprint arXiv:1908.06566, 2019.

[183] Z. Wang, Y. Yang, A. Shrivastava, V. Rawal, and Z. Ding, "Towards frequency-based explanation for robust cnn," arXiv preprint arXiv:2005.03141, 2020.

[184] Y. Sharma, G. W. Ding, and M. A. Brubaker, "On the effectiveness of low frequency perturbations," Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

[185] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[186] S. Zagoruyko and N. Komodakis, "Wide residual networks," in British Machine Vision Conference (BVCM), 2016.

[187] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[188] R. Bernhard, P.-A. Moellic, and J.-M. Dutertre, "Impact of low-bitwidth quantization on the adversarial robustness for embedded neural networks," in 2019 International Conference on Cyberworlds (CW), 2019.

[189] M. Denil, B. Shakibi, L. Dinh, N. De Freitas, et al., "Predicting parameters in deep learning," in Advances in neural information processing systems, pp. 2148–2156, 2013.

[190] G. B. Hacene, V. Gripon, M. Arzel, N. Farrugia, and Y. Bengio, "Quantized guided pruning for efficient hardware implementations of convolutional neural networks," arXiv preprint arXiv:1812.11337, 2018.

[191] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," arXiv preprint arXiv:1412.6115, 2014.

[192] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.

[193] Y. Choi, M. El-Khamy, and J. Lee, "Towards the limit of network quantization," arXiv preprint arXiv:1612.01543, 2016.

[194] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in Advances in neural information processing systems, pp. 3123–3131, 2015.

[195] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv preprint arXiv:1602.02830, 2016.

[196] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations.," Journal of Machine Learning Research, vol. 18, no. 187, pp. 1–30, 2017.

[197] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in European Conference on Computer Vision, pp. 525–542, Springer, 2016.

[198] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," 2016.

[199] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," arXiv preprint arXiv:1612.01064, 2016.

[200] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in International Conference on Machine Learning, pp. 1737–1746, 2015.

[201] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," arXiv preprint arXiv:1606.06160, 2016.

[202] R. Ding, Z. Liu, R. Shi, D. Marculescu, and R. Blanton, "Lightnn: Filling the gap between conventional deep neural networks and binarized networks," in Proceedings of the on Great Lakes Symposium on VLSI 2017, pp. 35–40, ACM, 2017.

[203] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," arXiv preprint arXiv:1802.05668, 2018.

[204] S. Darabi, M. Belbahri, M. Courbariaux, and V. P. Nia, "Bnn+: Improved binary network training," arXiv preprint arXiv:1812.11800, 2018.

[205] A. Galloway, G. W. Taylor, and M. Moussa, "Attacking binarized neural networks," in International Conference on Learning Representations, 2018.

[206] J. Lin, C. Gan, and S. Han, "Defensive quantization: When efficiency meets robustness," in International Conference on Learning Representations, 2019.

[207] Y. Zhao, I. Shumailov, R. Mullins, and R. Anderson, "To compress or not to compress: Understanding the interactions between adversarial attacks and neural network compression," Conference on Systems and Machine Learning (SysML), 2019.

[208] A. S. Rakin, J. Yi, B. Gong, and D. Fan, "Defend deep neural networks against adversarial examples via fixed anddynamic quantized activation functions," arXiv preprint arXiv:1807.06714, 2018.

[209] E. B. Khalil, A. Gupta, and B. Dilkina, "Combinatorial attacks on binarized neural networks," arXiv preprint arXiv:1810.03538, 2018.

[210] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 2017.

[211] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[212] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, "Technical report on the cleverhans v2.1.0 adversarial examples library," arXiv preprint arXiv:1610.00768, 2018.

[213] L. Wu, Z. Zhu, C. Tai, et al., "Towards understanding and improving the transferability of adversarial examples in deep neural networks," in Proceedings of The 12th Asian Conference on Machine Learning, 2020.

[214] R. Bernhard, P.-A. Moellic, and J.-M. Dutertre, "Luring of transferable adversarial perturbations in the black-box paradigm," in 2021 International Joint Conference on Neural Networks (IJCNN), 2021.

[215] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 2019.

[216] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi, "Unlabeled data improves adversarial robustness," in Advances in Neural Information Processing Systems, 2019.

[217] U. Hwang, J. Park, H. Jang, S. Yoon, and N. I. Cho, "Puvae: A variational autoencoder to purify adversarial examples," IEEE Access, vol. 7, pp. 126582–126593, 2019.

[218] S. Shan, E. Wenger, B. Wang, B. Li, H. Zheng, and B. Y. Zhao, "Using honeypots to catch adversarial attacks on neural networks," in Proceedings of ACM Conference on Computer and Communications Security (CCS), 2019.

[219] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High accuracy and high fidelity extraction of neural networks," in 29th {USENIX} Security Symposium ({USENIX} Security 20), pp. 1345–1362, 2020.

[220] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 6106–6116, 2018.

[221] A. S. Rakin, Z. He, and D. Fan, "Bit-flip attack: Crushing neural network with progressive bit search," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.