References

Alzantot, Moustafa, Sharma, Yash, Elgohary, Ahmed, Ho, Bo-Jhang, Srivastava, Mani, Chang, Kai-Wei, 2018. Generating natural language adversarial examples. arXiv preprint. arXiv:1804.07998.

Alzantot, Moustafa, Sharma, Yash, Chakraborty, Supriyo, Zhang, Huan, Hsieh, Cho-Jui, Srivastava, Mani B., 2019. Genattack: practical black-box attacks with gradient-free optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1111–1119.

Andriushchenko, Maksym, Hein, Matthias, 2019. Provably robust boosted decision stumps and trees against adversarial attacks. arXiv preprint. arXiv:1906.03526.

Aramoon, Omid, Chen, Pin-Yu, Qu, Gang, 2021. Don't forget to sign the gradients! Proceedings of Machine Learning and Systems 3.

Arya, Vijay, Bellamy, Rachel K.E., Chen, Pin-Yu, Dhurandhar, Amit, Hind, Michael, Hoffman, Samuel C., Houde, Stephanie, Liao, Q. Vera, Luss, Ronny, Mojsilović, Aleksandra, et al., 2019. One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. arXiv preprint. arXiv:1909.03012.

Athalye, Anish, Sutskever, Ilya, 2018. Synthesizing robust adversarial examples. In: International Conference on International Conference on Machine Learning.

Athalye, Anish, Carlini, Nicholas, Wagner, David, 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: International Conference on International Conference on Machine Learning.

Aurenhammer, Franz, Klein, Rolf, 1999. Voronoi diagrams. Handbook of computational geometry 5 (10), 201–290.

Bagdasaryan, Eugene, Veit, Andreas, Hua, Yiqing, Estrin, Deborah, Shmatikov, Vitaly, 2018. How to backdoor federated learning. arXiv preprint. arXiv:1807.00459.

Balaji, Yogesh, Goldstein, Tom, Hoffman, Judy, 2019. Instance adaptive adversarial training: improved accuracy tradeoffs in neural nets. arXiv preprint. arXiv:1910.08051.

Balın, Muhammed Fatih, Abid, Abubakar, Zou, James, 2019. Concrete autoencoders: differentiable feature selection and reconstruction. In: International Conference on Machine Learning, pp. 444–453.

Beck, Amir, Teboulle, Marc, 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2 (1), 183–202.

Belghazi, Mohamed Ishmael, Baratin, Aristide, Rajeshwar, Sai, Ozair, Sherjil, Bengio, Yoshua, Courville, Aaron, Hjelm, Devon, 2018. Mutual information neural estimation. In: International Conference on Machine Learning, pp. 531–540.

Bhagoji, Arjun Nitin, Chakraborty, Supriyo, Mittal, Prateek, Calo, Seraphin, 2019. Analyzing federated learning through an adversarial lens. In: International Conference on Machine Learning, pp. 634–643.

Bhattad, Anand, Chong, Min Jin, Liang, Kaizhao, Li, Bo, Forsyth, David A., 2019. Big but imperceptible adversarial perturbations via semantic manipulation. arXiv preprint. arXiv:1904.06347.

Bishop, Christopher M., 2006. Pattern recognition and machine learning. Machine Learning 128 (9).

Blum, Avrim, Dick, Travis, Manoj, Naren, Zhang, Hongyang, 2020. Random smoothing might be unable to certify linf robustness for high-dimensional images. Journal of Machine Learning Research 21, 211–1.

Bogdan, Małgorzata, van den Berg, Ewout, Su, Weijie, Candès, Emmanuel Jean, 2013. Statistical Estimation and Testing via the Ordered ℓ_1 Norm. Stanford University.

Boopathy, Akhilan, Weng, Tsui-Wei, Chen, Pin-Yu, Liu, Sijia, Daniel, Luca, 2019. CNN-cert: An efficient framework for certifying robustness of convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3240–3247.

Boopathy, Akhilan, Weng, Lily, Liu, Sijia, Chen, Pin-Yu, Zhang, Gaoyuan, Daniel, Luca, 2021. Fast training of provably robust neural networks by singleprop. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6803–6811.

Brendel, Wieland, Rauber, Jonas, Bethge, Matthias, 2018. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. In: International Conference on Learning Representations.

Brown, Tom B., Mané, Dandelion, Roy, Aurko, Abadi, Martín, Gilmer, Justin, 2017. Adversarial patch. arXiv preprint. arXiv:1712.09665.

Brown, Tom B., Carlini, Nicholas, Zhang, Chiyuan, Olsson, Catherine, Christiano, Paul, Goodfellow, Ian, 2018. Unrestricted adversarial examples. arXiv preprint. arXiv:1809.08352.

Brown, Tom B., et al., 2020a. Language models are few-shot learners. In: NeurIPS.

Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Pranav, Shyam, Sastry, Girish, Askell, Amanda, et al., 2020b. Language models are few-shot learners. arXiv preprint. arXiv:2005.14165.

Brunner, Thomas, Diehl, Frederik, Le Truong, Michael, Knoll, Alois, 2019. Guessing smart: biased sampling for efficient black-box adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4958–4966.

Bunel, Rudy R., Turkaslan, Ilker, Torr, Philip, Kohli, Pushmeet, Mudigonda, Pawan K., 2018. A unified view of piecewise linear neural network verification. Advances in Neural Information Processing Systems 31.

Bunel, Rudy, De Palma, Alessandro, Desmaison, Alban, Dvijotham, Krishnamurthy, Kohli, Pushmeet, Torr, Philip, Kumar, M. Pawan, 2020a. Lagrangian decomposition for neural network verification. In: Conference on Uncertainty in Artificial Intelligence. PMLR, pp. 370–379.

Bunel, Rudy, Lu, Jingyue, Turkaslan, Ilker, Kohli, P., Torr, P., Mudigonda, P., 2020b. Branch and bound for piecewise linear neural network verification. Journal of Machine Learning Research 21 (2020).

Candès, Emmanuel J., Wakin, Michael B., 2008. An introduction to compressive sampling. IEEE Signal Processing Magazine 25 (2), 21–30.

Carion, Nicolas, Massa, Francisco, Synnaeve, Gabriel, Usunier, Nicolas, Kirillov, Alexander, Zagoruyko, Sergey, 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

Carlini, Nicholas, Wagner, David, 2017a. Adversarial examples are not easily detected: bypassing ten detection methods. In: ACM Workshop on Artificial Intelligence and Security, pp. 3–14.

Carlini, Nicholas, Wagner, David, 2017b. Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy, pp. 39–57.

Carlini, Nicholas, Wagner, David, 2018. Audio adversarial examples: targeted attacks on speech-to-text. arXiv preprint. arXiv:1801.01944.

Carlini, Nicholas, Athalye, Anish, Papernot, Nicolas, Brendel, Wieland, Rauber, Jonas, Tsipras, Dimitris, Goodfellow, Ian, Madry, Aleksander, Kurakin, Alexey, 2019a. On evaluating adversarial robustness. arXiv preprint. arXiv:1902.06705.

Carlini, Nicholas, Liu, Chang, Erlingsson, Úlfar, Kos, Jernej, Song, Dawn, 2019b. The secret sharer: evaluating and testing unintended memorization in neural networks. In: 28th {USENIX} Security Symposium ({USENIX} Security 19), pp. 267–284.

Carlucci, Fabio M., D'Innocente, Antonio, Bucci, Silvia, Caputo, Barbara, Tommasi, Tatiana, 2019. Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2229–2238.

Carmon, Yair, Raghunathan, Aditi, Schmidt, Ludwig, Liang, Percy, Duchi, John C., 2019. Unlabeled data improves adversarial robustness. Neural Information Processing Systems.

Cavallari, Gabriel, Ribeiro, Leonardo, Ponti, Moacir, 2018. Unsupervised representation learning using convolutional and stacked auto-encoders: a domain and cross-domain feature space analysis. In: IEEE SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 440–446.

Chen, Pin-Yu, 2022. Model reprogramming: resource-efficient cross-domain machine learning. arXiv preprint. arXiv:2202.10629.

Chen, Jinghui, Gu, Quanquan, 2020. Rays: A ray searching method for hard-label adversarial attack. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1739–1747.

Chen, Xinlei, He, Kaiming, 2020. Exploring simple Siamese representation learning. arXiv preprint. arXiv:2011.10566.

Chen, Pin-Yu, Zhang, Huan, Sharma, Yash, Yi, Jinfeng, Hsieh, Cho-Jui, 2017a. 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.

Chen, Xinyun, Liu, Chang, Li, Bo, Lu, Kimberly, Song, Dawn, 2017b. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint. arXiv:1712.05526.

Chen, Hongge, Zhang, Huan, Chen, Pin-Yu, Yi, Jinfeng, Hsieh, Cho-Jui, 2018a. Attacking visual language grounding with adversarial examples: a case study on neural image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2587–2597.

Chen, Pin-Yu, Sharma, Yash, Zhang, Huan, Yi, Jinfeng, Hsieh, Cho-Jui, 2018b. EAD: elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Chen, Pin-Yu, Vinzamuri, Bhanukiran, Liu, Sijia, 2018c. Is ordered weighted ℓ_1 regularized regression robust to adversarial perturbation? a case study on OSCAR. In: IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1174–1178.

Chen, Tian Qi, Rubanova, Yulia, Bettencourt, Jesse, Duvenaud, David K., 2018d. Neural ordinary differential equations. In: Advances in Neural Information Processing Systems, pp. 6572–6583.

Chen, Ting, Kornblith, Simon, Norouzi, Mohammad, Hinton, Geoffrey, 2018e. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning.

Chen, Hongge, Zhang, Huan, Boning, Duane, Hsieh, Cho-Jui, 2019a. Robust decision trees against adversarial examples. In: International Conference on Machine Learning. PMLR, pp. 1122–1131.

Chen, Hongge, Zhang, Huan, Si, Si, Li, Yang, Boning, Duane, Hsieh, Cho-Jui, 2019b. Robustness verification of tree-based models. arXiv preprint. arXiv:1906.03849.

Chen, Jianbo, Jordan, Michael I., Wainwright, Martin J., 2020a. Hopskipjumpattack: a query-efficient decision-based attack. In: IEEE Symposium on Security and Privacy, pp. 1277–1294.

Chen, Mark, Radford, Alec, Child, Rewon, Wu, Jeffrey, Jun, Heewoo, Luan, David, Sutskever, Ilya, 2020b. Generative pretraining from pixels. In: International Conference on Machine Learning. PMLR, pp. 1691–1703.

Chen, Tianlong, Liu, Sijia, Chang, Shiyu, Cheng, Yu, Amini, Lisa, Wang, Zhangyang, 2020c. Adversarial robustness: from self-supervised pre-training to fine-tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 699–708.

Chen, Ting, Kornblith, Simon, Swersky, Kevin, Norouzi, Mohammad, Hinton, Geoffrey, 2020d. Big self-supervised models are strong semi-supervised learners. arXiv preprint. arXiv:2006.10029.

Chen, Xinlei, Fan, Haoqi, Girshick, Ross, He, Kaiming, 2020e. Improved baselines with momentum contrastive learning. arXiv preprint. arXiv:2003.04297.

Chen, Xiangning, Hsieh, Cho-Jui, Gong, Boqing, 2022. When vision transformers outperform ResNets without pre-training or strong data augmentations. In: International Conference on Learning Representations (ICLR).

Cheng, Chih-Hong, 2019. Towards robust direct perception networks for automated driving. arXiv preprint. arXiv:1909.13600.

Cheng, Minhao, Le, Thong, Chen, Pin-Yu, Yi, Jinfeng, Zhang, Huan, Hsieh, Cho-Jui, 2019a. Query-efficient hard-label black-box attack: an optimization-based approach. In: International Conference on Learning Representations.

Cheng, Shuyu, Dong, Yinpeng, Pang, Tianyu, et al., 2019b. Improving black-box adversarial attacks with a transfer-based prior. In: NeurIPS.

Cheng, Hao, Xu, Kaidi, Liu, Sijia, Chen, Pin-Yu, Zhao, Pu, Lin, Xue, 2020a. Defending against backdoor attack on deep neural networks. arXiv preprint. arXiv:2002.12162.

Cheng, Minhao, Lei, Qi, Chen, Pin-Yu, Dhillon, Inderjit, Hsie, Cho-Juih, 2020b. Cat: customized adversarial training for improved robustness. arXiv preprint. arXiv:2002.06789.

Cheng, Minhao, Singh, Simranjit, Chen, Patrick H., Chen, Pin-Yu, Liu, Sijia, Hsieh, Cho-Jui, 2020c. Sign-OPT: a query-efficient hard-label adversarial attack. In: International Conference on Learning Representations.

Cheng, Minhao, Yi, Jinfeng, Zhang, Huan, Chen, Pin-Yu, Hsieh, Cho-Jui, 2020d. Seq2sick: evaluating the robustness of sequence-to-sequence models with adversarial examples. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Cheng, Minhao, Chen, Pin-Yu, Liu, Sijia, Chang, Shiyu, Hsieh, Cho-Jui, Das, Payel, 2021. Self-progressing robust training. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Cohen, Jeremy M., Rosenfeld, Elan, Kolter, J. Zico, 2019. Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning.

Croce, Francesco, Hein, Matthias, 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning. PMLR, pp. 2206–2216.

Dai, Hanjun, Li, Hui, Tian, Tian, Huang, Xin, Wang, Lin, Zhu, Jun, Song, Le, 2018. Adversarial attack on graph structured data. In: International Conference on Machine Learning. PMLR, pp. 1115–1124.

Dau, Hoang Anh, Bagnall, Anthony, Kamgar, Kaveh, Yeh, Chin-Chia Michael, Zhu, Yan, Gharghabi, Shaghayegh, Ratanamahatana, Chotirat Ann, Keogh, Eamonn, 2019. The ucr time series archive. IEEE/CAA Journal of Automatica Sinica 6 (6), 1293–1305.

Davis, Luke M., 2013. Predictive modelling of bone ageing. PhD thesis. University of East Anglia.

Davis, Jason V., Kulis, Brian, Jain, Prateek, Sra, Suvrit, Dhillon, Inderjit S., 2007. Information-theoretic metric learning. In: International Conference on Machine Learning (ICML), pp. 209–216.

de Andrade, Douglas Coimbra, Leo, Sabato, Da Silva Viana, Martin Loesener, Bernkopf, Christoph, 2018. A neural attention model for speech command recognition. arXiv preprint. arXiv:1808.08929.

De Palma, Alessandro, Behl, Harkirat Singh, Bunel, Rudy, Torr, Philip H.S., Kumar, M. Pawan , 2021a. Scaling the convex barrier with active sets. In: International Conference on Learning Representations (ICLR).

De Palma, Alessandro, Bunel, Rudy, Desmaison, Alban, Dvijotham, Krishnamurthy, Kohli, Pushmeet, Philip Torr, H.S., Pawan Kumar, M., 2021b. Improved branch and bound for neural network verification via Lagrangian decomposition. arXiv preprint. arXiv: 2104.06718.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li, 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805.

Dhillon, Guneet S., Azizzadenesheli, Kamyar, Lipton, Zachary C., Bernstein, Jeremy, Kossaifi, Jean, Khanna, Aran, Anandkumar, Anima, 2018. Stochastic activation pruning for robust adversarial defense. In: International Conference on Learning Representations.

Dhurandhar, Amit, Chen, Pin-Yu, Luss, Ronny, Tu, Chun-Chen, Ting, Paishun, Shanmugam, Karthikeyan, Das, Payel, 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. Neural Information Processing Systems.

Dhurandhar, Amit, Pedapati, Tejaswini, Balakrishnan, Avinash, Chen, Pin-Yu, Shanmugam, Karthikeyan, Puri, Ruchir, 2019. Model agnostic contrastive explanations for structured data. arXiv preprint. arXiv:1906.00117.

Ding, Gavin Weiguang, Sharma, Yash, Chau Lui, Kry Yik, Huang, Ruitong, 2018. Mma training: direct input space margin maximization through adversarial training. arXiv preprint. arXiv:1812.02637.

Dong, Yinpeng, Liao, Fangzhou, Pang, Tianyu, Su, Hang, Zhu, Jun, Hu, Xiaolin, Li, Jianguo, 2018. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193.

Donsker, Monroe D., Varadhan, S.R. Srinivasa, 1983. Asymptotic evaluation of certain Markov process expectations for large time. iv. Communications on Pure and Applied Mathematics 36 (2), 183–212.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, et al., 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. arXiv:2010.11929.

Dubey, Abhimanyu, van der Maaten, Laurens, Yalniz, Zeki, Li, Yixuan, Mahajan, Dhruv, 2019. Defense against adversarial images using web-scale nearest-neighbor search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8767–8776.

Duchi, John, Hazan, Elad, Singer, Yoram, 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12 (7).

Duchi, John C., Jordan, Michael I., Wainwright, Martin J., Wibisono, Andre, 2015. Optimal rates for zero-order convex optimization: the power of two function evaluations. IEEE Transactions on Information Theory 61 (5), 2788–2806.

Dvijotham, Krishnamurthy Dj, Hayes, Jamie, Balle, Borja, Kolter, Zico, Qin, Chongli, Gyorgy, Andras, Xiao, Kai, Gowal, Sven, Kohli, Pushmeet, 2020. A framework for robustness certification of smoothed classifiers using f-divergences.

Ehlers, Ruediger, 2017. Formal verification of piece-wise linear feed-forward neural networks. In: International Symposium on Automated Technology for Verification and Analysis. Springer, pp. 269–286.

Elsayed, Gamaleldin F., Goodfellow, Ian, Sohl-Dickstein, Jascha, 2019. Adversarial reprogramming of neural networks. In: International Conference on Learning Representations.

Engstrom, Logan, Tran, Brandon, Tsipras, Dimitris, Schmidt, Ludwig, Madry, Aleksander, 2017. A rotation and a translation suffice: fooling cnns with simple transformations. arXiv preprint. arXiv:1712.02779.

Engstrom, Logan, Ilyas, Andrew, Santurkar, Shibani, Tsipras, Dimitris, Tran, Brandon, Madry, Aleksander, 2019a. Learning perceptually-aligned representations via adversarial robustness. arXiv preprint. arXiv:1906.00945.

Engstrom, Logan, Tran, Brandon, Tsipras, Dimitris, Schmidt, Ludwig, Madry, Aleksander, 2019b. Exploring the landscape of spatial robustness. In: International Conference on Machine Learning, pp. 1802–1811.

Eslami, Taban, Mirjalili, Vahid, Fong, Alvis, Laird, Angela R., Saeed, Fahad, 2019. Asd-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fmri data. Frontiers in Neuroinformatics 13.

Evtimov, Ivan, Eykholt, Kevin, Fernandes, Earlence, Kohno, Tadayoshi, Li, Bo, Prakash, Atul, Rahmati, Amir, Song, Dawn, 2017. Robust physical-world attacks on machine learning models. arXiv preprint. arXiv:1707.08945.

Eykholt, Kevin, Evtimov, Ivan, Fernandes, Earlence, Li, Bo, Rahmati, Amir, Xiao, Chaowei, Prakash, Atul, Kohno, Tadayoshi, Song, Dawn, 2018. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625–1634.

Fan, Lijie, Liu, Sijia, Chen, Pin-Yu, Zhang, Gaoyuan, Gan, Chuang, 2021. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? Advances in Neural Information Processing Systems 34.

Fawzi, Alhussein, Frossard, Pascal, 2015. Manitest: Are classifiers really invariant? In: BMVC.

Feinman, Reuben, Curtin, Ryan R., Shintre, Saurabh, Gardner, Andrew B., 2017. Detecting adversarial samples from artifacts. In: International Conference on Machine Learning.

Finn, Chelsea, Abbeel, Pieter, Levine, Sergey, 2017. Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint. arXiv:1703.03400.

Fong, Ruth, Patrick, Mandela, Vedaldi, Andrea, 2019. Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2950–2958.

Foret, Pierre, Kleiner, Ariel, Mobahi, Hossein, Neyshabur, Behnam, 2020. Sharpness-aware minimization for efficiently improving generalization. In: International Conference on Learning Representations (ICLR).

Freund, Yoav, Schapire, Robert E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

Gan, Chuang, Gong, Boqing, Liu, Kun, Su, Hao, Guibas, Leonidas J., 2018. Geometry guided convolutional neural networks for self-supervised video representation learning. In: CVPR, pp. 5589–5597.

Gao, Bolin, Pavel, Lacra, 2017. On the properties of the softmax function with application in game theory and reinforcement learning. arXiv preprint. arXiv:1704.00805.

Gao, X., Jiang, B., Zhang, S., 2014. On the information-adaptive variants of the admm: an iteration complexity perspective. Optimization Online 12.

Garcia, Washington, Chen, Pin-Yu, Jha, Somesh, Clouse, Scott, Butler, Kevin R.B., 2021. Hard-label manifolds: unexpected advantages of query efficiency for finding on-manifold adversarial examples. arXiv preprint. arXiv:2103.03325.

Geiping, Jonas, Fowl, Liam, Huang, W. Ronny, Czaja, Wojciech, Taylor, Gavin, Moeller, Michael, Goldstein, Tom, 2021. Witches' brew: industrial scale data poisoning via gradient matching. In: International Conference on Learning Representations.

Ghadimi, Saeed, Lan, Guanghui, 2013. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization 23 (4), 2341–2368.

Gidaris, Spyros, Singh, Praveer, Komodakis, Nikos, 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint. arXiv:1803.07728.

Goldberger, Jacob, Hinton, Geoffrey E., Roweis, Sam T., Salakhutdinov, Ruslan R., 2004. Neighbourhood components analysis. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 513–520.

Goldblum, Micah, Fowl, Liam, Goldstein, Tom, 2019. Adversarially robust few-shot learning: a meta-learning approach. ArXiv. ArXiv–1910.

Goldblum, Micah, Tsipras, Dimitris, Xie, Chulin, Chen, Xinyun, Schwarzschild, Avi, Song, Dawn, Madry, Aleksander, Li, Bo, Goldstein, Tom, 2022. Dataset security for machine learning: data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, Bengio, Yoshua, 2014. Generative adversarial nets. Advances in Neural Information Processing Systems 27.

Goodfellow, Ian J., Shlens, Jonathon, Szegedy, Christian, 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representation.

Gowal, Sven, Dvijotham, Krishnamurthy, Stanforth, Robert, Bunel, Rudy, Qin, Chongli, Uesato, Jonathan, Mann, Timothy, Kohli, Pushmeet, 2018. On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint. arXiv:1810.12715.

Gowal, Sven, Huang, Po-Sen, van den Oord, Aaron, Mann, Timothy, Kohli, Pushmeet, 2021. Self-supervised adversarial robustness for the low-label, high-data regime. In: International Conference on Learning Representations. https://openreview.net/forum?id=bgQek2O63w.

Grill, Jean-Bastien, Strub, Florian, Altché, Florent, Tallec, Corentin, Richemond, Pierre H., Buchatskaya, Elena, Doersch, Carl, Avila, Bernardo Pires, Guo, Zhaohan Daniel, Azar, Mohammad Gheshlaghi, et al., 2020. Bootstrap your own latent: a new approach to self-supervised learning. arXiv preprint. arXiv:2006.07733.

Gu, Tianyu, Dolan-Gavitt, Brendan, Garg, Siddharth, 2017. Badnets: identifying vulnerabilities in the machine learning model supply chain. arXiv preprint. arXiv:1708.06733.

Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S., 2019. BadNets: evaluating backdooring attacks on deep neural networks. IEEE Access 7, 47230–47244.

Guo, Chuan, Frank, Jared S., Kilian, Weinberger Q., 2018. Low frequency adversarial perturbation. arXiv preprint, arXiv:1809.08758.

Hambardzumyan, Karen, Khachatrian, Hrant, May, Jonathan, 2021. Warp: Word-level adversarial reprogramming. arXiv preprint. arXiv:2101.00121.

Han, Song, Pool, Jeff, Tran, John, Dally, William, 2015. Learning both weights and connections for efficient neural network. In: NeurIPS.

Hard, Andrew, Rao, Kanishka, Mathews, Rajiv, Beaufays, Françoise, Augenstein, Sean, Eichner, Hubert, Kiddon, Chloé, Ramage, Daniel, 2018. Federated learning for mobile keyboard prediction. arXiv preprint. arXiv:1811.03604.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, Kaiming, Fan, Haoqi, Wu, Yuxin, Xie, Saining, Girshick, Ross, 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738.

Heinsfeld, Anibal Sólon, Franco, Alexandre Rosa, Craddock, R. Cameron, Buchweitz, Augusto, Meneguzzi, Felipe, 2018. Identification of autism spectrum disorder using deep learning and the abide dataset. In: NeuroImage: Clinical.

Hendrycks, Dan, Zhao, Kevin, Basart, Steven, Steinhardt, Jacob, Song, Dawn, 2021. Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15262–15271.

Herman, Amy, 2016. Are you visually intelligent? What you don't see is as important as what you do see. Medical Daily. http://www.medicaldaily.com/are-you-visually-intelligent-what-you-dont-see-important-what-you-do-see-397963.

Hjelm, R. Devon, Fedorov, Alex, Lavoie-Marchildon, Samuel, Grewal, Karan, Bachman, Phil, Trischler, Adam, Bengio, Yoshua, 2019. Learning deep representations by mutual information estimation and maximization. In: International Conference on Learning Representations.

Ho, Chih-Hui, Vasconcelos, Nuno, 2020. Contrastive learning with adversarial examples. In: Advances in Neural Information Processing Systems.

Holland, J.K., Kemsley, E.K., Wilson, R.H., 1998. Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purees. Journal of the Science of Food and Agriculture 76 (2), 263–269.

Hosseini, Hossein, Poovendran, Radha, 2018. Semantic adversarial examples. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1614–1619.

Hsieh, Cheng-Yu, Yeh, Chih-Kuan, Liu, Xuanqing, Ravikumar, Pradeep, Kim, Seungyeon, Kumar, Sanjiv, Hsieh, Cho-Jui, 2020. Evaluations and methods for explanation through robustness analysis. arXiv preprint. arXiv:2006.00442.

Hsu, Chia-Yi, Chen, Pin-Yu, Lu, Songtao, Liu, Sijia, Yu, Chia-Mu, 2022. Adversarial examples can be effective data augmentation for unsupervised machine learning. In: AAAI.

Huang, Po-Sen, Wang, Chenglong, Singh, Rishabh, Yih, Wen-tau, He, Xiaodong, 2018. Natural language to structured query generation via meta-learning. arXiv preprint. arXiv:1803.02400.

Huang, Po-Sen, Stanforth, Robert, Welbl, Johannes, Dyer, Chris, Yogatama, Dani, Gowal, Sven, Dvijotham, Krishnamurthy, Kohli, Pushmeet, 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4074–4084.

Hubara, Itay, Courbariaux, Matthieu, Soudry, Daniel, El-Yaniv, Ran, Bengio, Yoshua, 2017. Quantized neural networks: training neural networks with low precision weights and activations. Journal of Machine Learning Research 18 (1), 6869–6898.

Ilyas, Andrew, Engstrom, Logan, Madry, Aleksander, 2019. Prior convictions: black-box adversarial attacks with bandits and priors. In: International Conference on Learning Representations.

Jagielski, Matthew, Oprea, Alina, Biggio, Battista, Liu, Chang, Nita-Rotaru, Cristina, Li, Bo, 2018. Manipulating machine learning: poisoning attacks and countermeasures for regression learning. In: IEEE Symposium on Security and Privacy, pp. 19–35.

Jia, Robin, Raghunathan, Aditi, Göksel, Kerem, Liang, Percy, 2019. Certified robustness to adversarial word substitutions. arXiv preprint. arXiv:1909.00986.

Jiang, Ziyu, Chen, Tianlong, Chen, Ting, Wang, Zhangyang, 2020. Robust pre-training by adversarial contrastive learning. arXiv preprint. arXiv:2010.13337.

Joshi, Ameya, Mukherjee, Amitangshu, Sarkar, Soumik, Hegde, Chinmay, 2019. Semantic adversarial attacks: parametric transformations that fool deep classifiers. arXiv preprint. arXiv:1904.08489.

Julian, Kyle D., Sharma, Shivam, Jeannin, Jean-Baptiste, Kochenderfer, Mykel J., 2019. Verifying aircraft collision avoidance neural networks through linear approximations of safe regions. arXiv preprint. arXiv:1903.00762.

Kantorovich, L.V., Rubinstein, G., 1958. On a space of completely additive functions. Vestnik Leningradskogo Universiteta 13 (7), 52–59.

Katz, Guy, Barrett, Clark, Dill, David L., Julian, Kyle, Kochenderfer, Mykel J., 2017. Reluplex: An efficient smt solver for verifying deep neural networks. In: International Conference on Computer Aided Verification. Springer, pp. 97–117.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, Tang, Ping Tak Peter, 2017. On large-batch training for deep learning: generalization gap and sharp minima. In: International Conference on Learning Representations.

Khatri, Devvrit, Cheng, Minhao, Hsieh, Cho-Jui, Dhillon, Inderjit, et al., 2020. Voting based ensemble improves robustness of defensive models. arXiv preprint. arXiv:2011.14031.

Khosla, Prannay, Teterwak, Piotr, Wang, Chen, Sarna, Aaron, Tian, Yonglong, Isola, Phillip, Maschinot, Aaron, Liu, Ce, Krishnan, Dilip, 2020. Supervised contrastive learning. arXiv preprint. arXiv:2004.11362.

Kim, Minseon, Tack, Jihoon, Hwang, Sung Ju, 2020. Adversarial self-supervised contrastive learning. arXiv preprint. arXiv:2006.07589.

Kingma, Diederik, Ba, Jimmy, 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint. arXiv:1609.02907.

Ko, Ching-Yun, Lyu, Zhaoyang, Weng, Lily, Daniel, Luca, Wong, Ngai, Lin, Dahua, 2019. POPQORN: Quantifying robustness of recurrent neural networks. In: International Conference on Machine Learning (ICML).

Koch, Gregory, Zemel, Richard, Salakhutdinov, Ruslan, 2015. Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop. Lille, vol. 2.

Kolouri, Soheil, Rohde, Gustavo K., Hoffmann, Heiko, 2018. Sliced Wasserstein distance for learning Gaussian mixture models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3427–3436.

Komkov, Stepan, Petiushko, Aleksandr, 2021. Advhat: Real-world adversarial attack on arcface face id system. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 819–826.

Kozlov, Mikhail K., Tarasov, Sergei P., Khachiyan, Leonid G., 1980. The polynomial solvability of convex quadratic programming. U.S.S.R. Computational Mathematics and Mathematical Physics 20 (5), 223–228.

Krizhevsky, Alex, Hinton, Geoffrey, et al., 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Krizhevsky, Alex, Sutskever, Ilya, Hinton, Geoffrey E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Kurakin, Alexey, Goodfellow, Ian, Bengio, Samy, 2016. Adversarial examples in the physical world. arXiv preprint. arXiv:1607.02533.

Kurakin, Alexey, Goodfellow, Ian, Bengio, Samy, 2017. Adversarial machine learning at scale. International Conference on Learning Representations.

Lapuschkin, Sebastian, Binder, Alexander, Montavon, Grégoire, Müller, Klaus-Robert, Samek, Wojciech, 2016. The lrp toolbox for artificial neural networks. Journal of Machine Learning Research 17 (114), 1–5. http://jmlr.org/papers/v17/15-618.html.

Lax, Peter D., Terrell, Maria Shea, 2014. Calculus with Applications. Springer.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, Haffner, Patrick, et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.

LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey, 2015. Deep learning. Nature 521 (7553), 436–444.

Lecuyer, Mathias, Atlidakis, Vaggelis, Geambasu, Roxana, Hsu, Daniel, Jana, Suman, 2019. Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 656–672.

Lee, Kimin, Lee, Kibok, Lee, Honglak, Shin, Jinwoo, 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems 31.

Lei, Qi, Wu, Lingfei, Chen, Pin-Yu, Dimakis, Alexandros G., Dhillon, Inderjit S., Witbrock, Michael, 2019. Discrete adversarial attacks and submodular optimization with applications to text classification. In: SysML.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics. Doklady 10 (1), 845–848.

Li, Jiwei, Monroe, Will, Jurafsky, Dan, 2016. Understanding neural networks through representation erasure. arXiv preprint. arXiv:1612.08220.

Li, Jinfeng, Ji, Shouling, Du, Tianyu, Li, Bo, Wang, Ting, 2018a. Textbugger: generating adversarial text against real-world applications. arXiv preprint. arXiv:1812.05271.

Li, Yao, Renqiang Min, Martin, Yu, Wenchao, Hsieh, Cho-Jui, Lee, Thomas C.M., Kruus, Erik, 2018b. Optimal transport classifier: defending against adversarial attacks by regularized deep embedding. arXiv preprint. arXiv:1811.07950.

Li, Bai, Chen, Changyou, Wang, Wenlin, Carin, Lawrence, 2019a. Certified adversarial robustness with additive noise. Neural Information Processing Systems.

Li, Juncheng, Schmidt, Frank, Kolter, Zico, 2019b. Adversarial camera stickers: a physical camera-based attack on deep learning systems. In: International Conference on Machine Learning. PMLR, pp. 3896–3904.

Li, Linyang, Ma, Ruotian, Guo, Qipeng, Xue, Xiangyang, Qiu, Xipeng, 2020a. Bert-attack: adversarial attack against bert using bert. arXiv preprint. arXiv:2004.09984.

Li, Qizhang, Guo, Yiwen, Chen, Hao, 2020b. Practical no-box adversarial attacks against dnns. In: Advances in Neural Information Processing Systems.

Lin, Chang-Sheng, Hsu, Chia-Yi, Chen, Pin-Yu, Yu, Chia-Mu, 2021. Real-world adversarial examples involving makeup application. arXiv preprint. arXiv:2109.03329.

Liu, Ziwei, Luo, Ping, Wang, Xiaogang, Tang, Xiaoou, 2015. Deep learning face attributes in the wild. In: ICCV 2015.

Liu, Yannan, Wei, Lingxiao, Luo, Bo, Xu, Qiang, 2017a. Fault injection attack on deep neural network. In: IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp. 131–138.

Liu, Yanpei, Chen, Xinyun, Liu, Chang, Song, Dawn, 2017b. Delving into transferable adversarial examples and black-box attacks. In: International Conference on Learning Representations.

Liu, Sijia, Chen, Jie, Chen, Pin-Yu, Hero, Alfred O., 2018a. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. AISTATS.

Liu, Sijia, Kailkhura, Bhavya, Chen, Pin-Yu, Ting, Paishun, Chang, Shiyu, Amini, Lisa, 2018b. Zeroth-order stochastic variance reduction for nonconvex optimization. In: Advances in Neural Information Processing Systems, pp. 3727–3737.

Liu, Xuanqing, Cheng, Minhao, Zhang, Huan, Hsieh, Cho-Jui, 2018c. Towards robust neural networks via random self-ensemble. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 369–385.

Liu, Changliu, Arnon, Tomer, Lazarus, Christopher, Strong, Christopher, Barrett, Clark, Kochenderfer, Mykel J., 2019a. Algorithms for verifying deep neural networks. arXiv preprint. arXiv:1903.06758.

Liu, Hsueh-Ti Derek, Tao, Michael, Li, Chun-Liang, Nowrouzezahrai, Derek, Jacobson, Alec, 2019b. Beyond pixel norm-balls: parametric adversaries using an analytically differentiable renderer. In: International Conference on Learning Representations.

Liu, Sijia, Chen, Pin-Yu, Chen, Xiangyi, Hong, Mingyi, 2019c. Signsgd via zeroth-order oracle. In: International Conference on Learning Representations.

Liu, Xuanqing, Li, Yao, Wu, Chongruo, Hsieh, Cho-Jui, 2019d. Adv-BNN: improved adversarial defense through robust Bayesian neural network. In: International Conference on Learning Representations.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Stoyanov, Veselin, 2019e. Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv:1907.11692.

Liu, Sijia, Chen, Pin-Yu, Kailkhura, Bhavya, Zhang, Gaoyuan, Hero, Alfred, Varshney, Pramod K., 2020a. A primer on zeroth-order optimization in signal processing and machine learning. IEEE Signal Processing Magazine.

Liu, Sijia, Lu, Songtao, Chen, Xiangyi, Feng, Yao, Xu, Kaidi, Al-Dujaili, Abdullah, Hong, Mingyi, O'Reilly, Una-May, 2020b. Min-max optimization without gradients: convergence and applications to black-box evasion and poisoning attacks. In: International Conference on Machine Learning, pp. 6282–6293.

Liu, Xuanqing, Si, Si, Cao, Qin, Kumar, Sanjiv, Hsieh, Cho-Jui, 2020c. How does noise help robustness? Explanation and exploration under the neural sde framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 282–290.

Liu, Yong, Siqi, Mai, Chen, Xiangning, Hsieh, Cho-Jui, You, Yang, 2022. Towards efficient and scalable sharpness-aware minimization. In: IEEE Conference on Computer Vision and Pattern Recognition, 2022. CVPR 2022.

Lu, Jingyue, Kumar, M. Pawan, 2020. Neural network branching for neural network verification. In: International Conference on Learning Representation (ICLR).

Luss, Ronny, Chen, Pin-Yu, Dhurandhar, Amit, Sattigeri, Prasanna, Zhang, Yunfeng, Shanmugam, Karthikeyan, Tu, Chun-Chen, 2021. Leveraging latent features for local explanations. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1139–1149.

Ma, Xingjun, Li, Bo, Wang, Yisen, Erfani, Sarah M., Wijewickrema, Sudanthi, Schoenebeck, Grant, Song, Dawn, Houle, Michael E., Bailey, James, 2018. Characterizing adversarial subspaces using local intrinsic dimensionality. In: International Conference on Learning Representations.

Madry, Aleksander, Makelov, Aleksandar, Schmidt, Ludwig, Tsipras, Dimitris, Vladu, Adrian, 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint. arXiv:1706.06083.

Madry, Aleksander, Makelov, Aleksandar, Schmidt, Ludwig, Tsipras, Dimitris, Vladu, Adrian, 2018. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations.

Maicas, Gabriel, Bradley, Andrew P., Nascimento, Jacinto C., Reid, Ian, Carneiro, Gustavo, 2018. Training medical image analysis systems like radiologists. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 546–554.

Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, Goodfellow, Ian, Frey, Brendan, 2016. Adversarial autoencoders. In: ICLR Workshop.

McMahan, Brendan, Moore, Eider, Ramage, Daniel, Hampson, Seth, Aguera y Arcas, Blaise, 2017. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research, vol. 54. PMLR, pp. 1273–1282.

Mehra, Akshay, Kailkhura, Bhavya, Chen, Pin-Yu, Hamm, Jihun, 2021a. How robust are randomized smoothing based defenses to data poisoning? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13244–13253.

Mehra, Akshay, Kailkhura, Bhavya, Chen, Pin-Yu, Hamm, Jihun, 2021b. Understanding the limits of unsupervised domain adaptation via data poisoning. In: Thirty-Fifth Conference on Neural Information Processing Systems.

Meng, Dongyu, Chen, Hao, 2017. Magnet: a two-pronged defense against adversarial examples. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., Dean, Jeff, 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26.

Miyato, Takeru, Maeda, Shin-ichi, Koyama, Masanori, Ishii, Shin, 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (8), 1979–1993.

Mohapatra, Jeet, Weng, Tsui-Wei, Chen, Pin-Yu, Liu, Sijia, Daniel, Luca, 2020. Towards verifying robustness of neural networks against a family of semantic perturbations. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–252.

Moosavi-Dezfooli, Seyed-Mohsen, Fawzi, Alhussein, Fawzi, Omar, Frossard, Pascal, 2017. Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1765–1773.

Munkhdalai, Tsendsuren, Yu, Hong, 2017. Meta networks. Proceedings of Machine Learning Research 70, 2554.

Neekhara, Paarth, Hussain, Shehzeen, Dubnov, Shlomo, Koushanfar, Farinaz, 2018. Adversarial reprogramming of text classification neural networks. arXiv preprint. arXiv: 1809.01829.

Nesterov, Yurii, Spokoiny, Vladimir, 2017. Random gradient-free minimization of convex functions. Foundations of Computational Mathematics 17 (2), 527–566.

Neyshabur, Behnam, Bhojanapalli, Srinadh, McAllester, David, Srebro, Nati, 2017. Exploring generalization in deep learning. In: Advances in Neural Information Processing Systems, pp. 5947–5956.

Nguyen, Anh, Tran, Anh, 2020. Input-aware dynamic backdoor attack. In: Neural Information Processing Systems.

Nichol, Alex, Achiam, Joshua, Schulman, John, 2018. On first-order meta-learning algorithms. arXiv preprint. arXiv:1803.02999.

Noroozi, Mehdi, Favaro, Paolo, 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. Springer, pp. 69–84.

Novak, Joseph D., Gowin, D. Bob, 1984. Learning How to Learn. Cambridge University Press.

Papernot, Nicolas, McDaniel, Patrick, 2018. Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. arXiv preprint. arXiv:1803.04765.

Papernot, Nicolas, McDaniel, Patrick, Goodfellow, Ian, 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint. arXiv:1605.07277.

Papernot, Nicolas, McDaniel, Patrick, Goodfellow, Ian, Jha, Somesh, Celik, Z. Berkay, Swami, Ananthram, 2017. Practical black-box attacks against machine learning. In: ACM Asia Conference on Computer and Communications Security, pp. 506–519.

Paul, Sayak, Chen, Pin-Yu, 2022. Vision transformers are robust learners. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Pennington, Jeffrey, Socher, Richard, Manning, Christopher D., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Peyré, G., Cuturi, M., 2018. Computational optimal transport. arXiv preprint. arXiv:1803.00567.

Purushwalkam, Senthil, Gupta, Abhinav, 2020. Demystifying contrastive self-supervised learning: invariances, augmentations and dataset biases. arXiv preprint. arXiv:2007.13916.

Qin, Yao, Carlini, Nicholas, Cottrell, Garrison, Goodfellow, Ian, Raffel, Colin, 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: International Conference on Machine Learning. PMLR, pp. 5231–5240.

Qin, Yunxiao, Xiong, Yuanhao, Yi, Jinfeng, Hsieh, Cho-Jui, 2021. Training meta-surrogate model for transferable adversarial attack. arXiv preprint. arXiv:2109.01983.

Raghu, Aniruddh, Raghu, Maithra, Bengio, Samy, Vinyals, Oriol, 2019. Rapid learning or feature reuse? Towards understanding the effectiveness of maml. arXiv preprint. arXiv: 1909.09157.

Raghunathan, Aditi, Steinhardt, Jacob, Liang, Percy S., 2018. Semidefinite relaxations for certifying robustness to adversarial examples. In: Advances in Neural Information Processing Systems, pp. 10877–10887.

Raghuram, Jayaram, Chandrasekaran, Varun, Jha, Somesh, Banerjee, Suman, 2020. Detecting anomalous inputs to DNN classifiers by joint statistical testing at the layers. arXiv preprint. arXiv:2007.15147.

Ranzato, Marc'Aurelio, Huang, Fu Jie, Boureau, Y-Lan, LeCun, Yann, 2007. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Rao, K. Ramamohan, Yip, Ping, 2014. Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press.

Ravi, Sachin, Larochelle, Hugo, 2016. Optimization as a model for few-shot learning.

Ribeiro, Marco, Singh, Sameer, Guestrin, Carlos, 2016. "why should I trust you?" explaining the predictions of any classifier. In: ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115 (3), 211–252.

Sablayrolles, Alexandre, Douze, Matthijs, Schmid, Cordelia, Jégou, Hervé, 2020. Radioactive data: tracing through training. In: International Conference on Machine Learning. PMLR, pp. 8326–8335.

Salman, Hadi, Yang, Greg, Li, Jerry, Zhang, Pengchuan, Zhang, Huan, Razenshteyn, Ilya, Bubeck, Sebastien, 2019a. Provably robust deep learning via adversarially trained smoothed classifiers. arXiv preprint. arXiv:1906.04584.

Salman, Hadi, Yang, Greg, Zhang, Huan, Hsieh, Cho-Jui, Zhang, Pengchuan, 2019b. A convex relaxation barrier to tight robustness verification of neural networks. Advances in Neural Information Processing Systems 32.

Salman, Hadi, Ilyas, Andrew, Engstrom, Logan, Vemprala, Sai, Madry, Aleksander, Kapoor, Ashish, 2020a. Unadversarial examples: designing objects for robust vision. arXiv preprint. arXiv:2012.12235.

Salman, Hadi, Sun, Mingjie, Yang, Greg, Kapoor, Ashish, Kolter, J. Zico, 2020b. Denoised smoothing: provable defense for pretrained classifiers. Advances in Neural Information Processing Systems 33, 21945–21957.

Samangouei, Pouya, Kabkab, Maya, Chellappa, Rama, 2018. Defense-gan: protecting classifiers against adversarial attacks using generative models. arXiv preprint. arXiv:1805.06605.

Santoro, Adam, Bartunov, Sergey, Botvinick, Matthew, Wierstra, Daan, Lillicrap, Timothy, 2016. Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning, pp. 1842–1850.

Shafahi, Ali, Huang, W. Ronny, Najibi, Mahyar, Suciu, Octavian, Studer, Christoph, Dumitras, Tudor, Goldstein, Tom, 2018. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: Advances in Neural Information Processing Systems, pp. 6103–6113.

Shafahi, Ali, Najibi, Mahyar, Ghiasi, Amin, Xu, Zheng, Dickerson, John, Studer, Christoph, Davis, Larry S., Taylor, Gavin, Goldstein, Tom, 2019. Adversarial training for free!. arXiv preprint. arXiv:1904.12843.

Shan, Shawn, Wenger, Emily, Zhang, Jiayun, Li, Huiying, Zheng, Haitao, Zhao, Ben Y., 2020. Fawkes: protecting privacy against unauthorized deep learning models. In: 29th {USENIX} Security Symposium ({USENIX} Security 20), pp. 1589–1604.

Shao, Rulin, Shi, Zhouxing, Yi, Jinfeng, Chen, Pin-Yu, Hsieh, Cho-Jui, 2021a. On the adversarial robustness of vision transformers. arXiv preprint. arXiv:2103.15670.

Shao, Rulin, Shi, Zhouxing, Yi, Jinfeng, Chen, Pin-Yu, Hsieh, Cho-Jui, 2021b. Robust text captchas using adversarial examples. arXiv preprint. arXiv:2101.02483.

Sharif, Mahmood, Bhagavatula, Sruti, Bauer, Lujo, Reiter, Michael K., 2016. 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, pp. 1528–1540.

Sharma, Yash, Ding, Gavin Weiguang, Brubaker, Marcus A., 2019. On the effectiveness of low frequency perturbations. In: AAAI.

Shi, Zhouxing, Zhang, Huan, Chang, Kai-Wei, Huang, Minlie, Hsieh, Cho-Jui, 2020. Robustness verification for transformers. In: International Conference on Learning Representations (ICLR).

Shi, Zhouxing, Wang, Yihan, Zhang, Huan, Yi, Jinfeng, Hsieh, Cho-Jui, 2021. Fast certified robust training via better initialization and shorter warmup. In: NeurIPS.

Simonyan, Karen, Zisserman, Andrew, 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556.

Singh, Gagandeep, Gehr, Timon, Mirman, Matthew, Püschel, Markus, Vechev, Martin, 2018a. Fast and effective robustness certification. In: Advances in Neural Information Processing Systems, pp. 10802–10813.

Singh, Gagandeep, Gehr, Timon, Püschel, Markus, Vechev, Martin, 2018b. Boosting robustness certification of neural networks. In: International Conference on Learning Representations.

Singh, Gagandeep, Ganvir, Rupanshu, Püschel, Markus, Vechev, Martin, 2019a. Beyond the single neuron convex barrier for neural network certification. In: Advances in Neural Information Processing Systems (NeurIPS).

Singh, Gagandeep, Gehr, Timon, Püschel, Markus, Vechev, Martin, 2019b. An abstract domain for certifying neural networks. Proceedings of the ACM on Programming Languages 3 (POPL), 41.

Sitawarin, Chawin, Wagner, David, 2019. Defending against adversarial examples with k-nearest neighbor. arXiv preprint. arXiv:1906.09525.

Smith, Virginia, Chiang, Chao-Kai, Sanjabi, Maziar, Talwalkar, Ameet S., 2017. Federated multi-task learning. In: Advances in Neural Information Processing Systems, pp. 4424–4434.

Snell, Jake, Swersky, Kevin, Zemel, Richard, 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, pp. 4077–4087.

Stallkamp, Johannes, Schlipsing, Marc, Salmen, Jan, Igel, Christian, 2012. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition. Neural Networks 32, 323–332.

Stanforth, Robert, Fawzi, Alhussein, Kohli, Pushmeet, et al., 2019. Are labels required for improving adversarial robustness? Neural Information Processing Systems.

Stutz, David, Hein, Matthias, Schiele, Bernt, 2019. Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6976–6987.

Stutz, David, Chandramoorthy, Nandhini, Hein, Matthias, Schiele, Bernt, 2020. Bit error robustness for energy-efficient dnn accelerators. arXiv preprint. arXiv:2006.13977.

Su, Dong, Zhang, Huan, Chen, Hongge, Yi, Jinfeng, Chen, Pin-Yu, Gao, Yupeng, 2018. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648.

Su, Jiawei, Vargas, Danilo Vasconcellos, Sakurai, Kouichi, 2019. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23 (5), 828–841.

Sugiyama, Masashi, 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. Journal of Machine Learning Research 8, 1027–1061.

Sun, Lichao, Dou, Yingtong, Yang, Carl, Wang, Ji, Yu, Philip S., He, Lifang, Li, Bo, 2018. Adversarial attack and defense on graph data: a survey. arXiv preprint. arXiv:1812.10528.

Sun, Xiaowu, Khedr, Haitham, Shoukry, Yasser, 2019a. Formal verification of neural network controlled autonomous systems. In: Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control, pp. 147–156.

Sun, Yu, Wang, Shuohuan, Li, Yukun, Feng, Shikun, Chen, Xuyi, Zhang, Han, Tian, Xin, Zhu, Danxiang, Tian, Hao, Wu, Hua, 2019b. Ernie: Enhanced representation through knowledge integration. arXiv preprint. arXiv:1904.09223.

Sutskever, Ilya, Vinyals, Oriol, Le, Quoc V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, Fergus, Rob, 2014. Intriguing properties of neural networks. In: International Conference on Learning Representations.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, Wojna, Zbigniew, 2016. Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826.

Thrun, Sebastian, Pratt, Lorien, 2012. Learning to Learn. Springer Science & Business Media.

Thys, Simen, Van Ranst, Wiebe, Goedemé, Toon, 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Tian, Yonglong, Sun, Chen, Poole, Ben, Krishnan, Dilip, Schmid, Cordelia, Isola, Phillip, 2020. What makes for good views for contrastive learning. arXiv preprint. arXiv:2005.10243.

Tibshirani, Robert, 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, Methodological, 267–288.

Tjandraatmadja, Christian, Anderson, Ross, Huchette, Joey, Ma, Will, Patel, Krunal Kishor, Vielma, Juan Pablo, 2020. The convex relaxation barrier, revisited: tightened single-neuron relaxations for neural network verification. Advances in Neural Information Processing Systems 33, 21675–21686.

Tramer, Florian, Boneh, Dan, 2019. Adversarial training and robustness for multiple perturbations. In: Advances in Neural Information Processing Systems.

Tramer, Florian, Carlini, Nicholas, Brendel, Wieland, Madry, Aleksander, 2020. On adaptive attacks to adversarial example defenses. arXiv preprint. arXiv:2002.08347.

Trinh, Trieu H., Luong, Minh-Thang, Le, Quoc V., 2019. Selfie: self-supervised pretraining for image embedding. arXiv preprint. arXiv:1906.02940.

Tsai, Yun-Yun, Chen, Pin-Yu, Ho, Tsung-Yi, 2020. Transfer learning without knowing: reprogramming black-box machine learning models with scarce data and limited resources. In: International Conference on Machine Learning, pp. 9614–9624.

Tsai, Yu-Lin, Hsu, Chia-Yi, Yu, Chia-Mu, Chen, Pin-Yu, 2021a. Formalizing generalization and adversarial robustness of neural networks to weight perturbations. Advances in Neural Information Processing Systems 34.

Tsai, Yu-Lin, Hsu, Chia-Yi, Yu, Chia-Mu, Chen, Pin-Yu, 2021b. Non-singular adversarial robustness of neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3840–3844.

Tsai, Yun-Yun, Hsiung, Lei, Chen, Pin-Yu, Ho, Tsung-Yi, 2022. Towards compositional adversarial robustness: generalizing adversarial training to composite semantic perturbations. arXiv preprint, arXiv:2202.04235.

Tu, Chun-Chen, Ting, Paishun, Chen, Pin-Yu, Liu, Sijia, Zhang, Huan, Yi, Jinfeng, Hsieh, Cho-Jui, Cheng, Shin-Ming, 2019. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 742–749.

van den Oord, Aaron, Li, Yazhe, Vinyals, Oriol, 2018. Representation learning with contrastive predictive coding. arXiv preprint. arXiv:1807.03748.

Van der Maaten, Laurens, Hinton, Geoffrey, 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9 (11).

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia, 2017. Attention is all you need. arXiv preprint. arXiv:1706.03762.

Vinod, Ria, Chen, Pin-Yu, Das, Payel, 2020. Reprogramming language models for molecular representation learning. arXiv preprint. arXiv:2012.03460.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, Erhan, Dumitru, 2015. Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164.

Wang, Tongzhou, Isola, Phillip, 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. arXiv preprint. arXiv:2005.10242.

Wang, Shiqi, Pei, Kexin, Whitehouse, Justin, Yang, Junfeng, Jana, Suman, 2018a. Efficient formal safety analysis of neural networks. In: Advances in Neural Information Processing Systems, pp. 6367–6377.

Wang, Yining, Du, Simon, Balakrishnan, Sivaraman, Singh, Aarti, 2018b. Stochastic zeroth-order optimization in high dimensions. In: AISTATS.

Wang, Bao, Shi, Zuoqiang, Osher, Stanley, 2019a. Resnets ensemble via the Feynman-Kac formalism to improve natural and robust accuracies. Advances in Neural Information Processing Systems 32.

Wang, Bolun, Yao, Yuanshun, Shan, Shawn, Li, Huiying, Viswanath, Bimal, Zheng, Haitao, Ben Zhao, Y., 2019b. Neural cleanse. Identifying and Mitigating Backdoor Attacks in Neural Networks.

Wang, Lu, Liu, Xuanqing, Yi, Jinfeng, Zhou, Zhi-Hua, Hsieh, Cho-Jui, 2019c. Evaluating the robustness of nearest neighbor classifiers: a primal-dual perspective. arXiv preprint. arXiv:1906.03972.

Wang, Yisen, Ma, Xingjun, Bailey, James, Yi, Jinfeng, Zhou, Bowen, Gu, Quanquan, 2019d. On the convergence and robustness of adversarial training. In: Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97. PMLR, pp. 6586–6595.

Wang, Xiao, Wang, Siyue, Chen, Pin-Yu, Wang, Yanzhi, Kulis, Brian, Lin, Xue, Chin, Sang, 2019e. Protecting neural networks with hierarchical random switching: towards better robustness-accuracy trade-off for stochastic defenses. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), pp. 6013–6019.

Wang, Guangting, Luo, Chong, Sun, Xiaoyan, Xiong, Zhiwei, Zeng, Wenjun, 2020a. Tracking by instance detection: a meta-learning approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6288–6297.

Wang, Haohan, Wu, Xindi, Huang, Zeyi, Xing, Eric P., 2020b. High-frequency component helps explain the generalization of convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8684–8694.

Wang, Lu, Liu, Xuanqing, Yi, Jinfeng, Jiang, Yuan, Hsieh, Cho-Jui, 2020c. Provably robust metric learning. arXiv preprint. arXiv:2006.07024.

Wang, Ren, Zhang, Gaoyuan, Liu, Sijia, Chen, Pin-Yu, Xiong, Jinjun, Wang, Meng, 2020d. Practical detection of trojan neural networks: data-limited and data-free cases. In: European Conference on Computer Vision, pp. 222–238.

Wang, Yihan, Zhang, Huan, Chen, Hongge, Boning, Duane, Hsieh, Cho-Jui, 2020e. On lp-norm robustness of ensemble decision stumps and trees. In: International Conference on Machine Learning. PMLR, pp. 10104–10114.

Wang, Zifan, Yang, Yilin, Shrivastava, Ankit, Rawal, Varun, Ding, Zihao, 2020f. Towards frequency-based explanation for robust cnn. arXiv preprint. arXiv:2005.03141.

Wang, Jingkang, Zhang, Tianyun, Liu, Sijia, Chen, Pin-Yu, Xu, Jiacen, Fardad, Makan, Li, Bo, 2021a. Adversarial attack generation empowered by min-max optimization. Advances in Neural Information Processing Systems 34.

Wang, Ren, Xu, Kaidi, Liu, Sijia, Chen, Pin-Yu, Weng, Tsui-Wei, Gan, Chuang, Wang, Meng, 2021b. On fast adversarial robustness adaptation in model-agnostic meta-learning. In: International Conference on Learning Representations.

Wang, Shiqi, Zhang, Huan, Xu, Kaidi, Lin, Xue, Jana, Suman, Hsieh, Cho-Jui, Kolter, J. Zico, 2021c. Beta-crown: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. arXiv preprint. arXiv:2103.06624.

Wang, Siyue, Wang, Xiao, Chen, Pin-Yu, Zhao, Pu, Lin, Xue, 2021d. Characteristic examples: high-robustness, low-transferability fingerprinting of neural networks. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI, pp. 575–582.

Weinberger, Kilian Q., Saul, Lawrence K., 2009. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10, 207–244.

Weng, Tsui-Wei, Zhang, Huan, Chen, Hongge, Song, Zhao, Hsieh, Cho-Jui, Boning, Duane, Dhillon, Inderjit S., Daniel, Luca, 2018a. Towards fast computation of certified robustness for relu networks. In: International Conference on International Conference on Machine Learning.

Weng, Tsui-Wei, Zhang, Huan, Chen, Pin-Yu, Yi, Jinfeng, Su, Dong, Gao, Yupeng, Hsieh, Cho-Jui, Daniel, Luca, 2018b. Evaluating the robustness of neural networks: an extreme value theory approach. In: International Conference on Learning Representations.

Weng, Lily, Chen, Pin-Yu, Nguyen, Lam, Squillante, Mark, Boopathy, Akhilan, Oseledets, Ivan, Daniel, Luca, 2019. PROVEN: Verifying robustness of neural networks with a probabilistic approach. In: International Conference on Machine Learning, PMLR, pp. 6727–6736.

Weng, Tsui-Wei, Zhao, Pu, Liu, Sijia, Chen, Pin-Yu, Lin, Xue, Daniel, Luca, 2020. Towards certificated model robustness against weight perturbations. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6356–6363.

Wong, Eric, Kolter, J. Zico, 2017. Provable defenses against adversarial examples via the convex outer adversarial polytope. arXiv preprint. arXiv:1711.00851.

Wong, Eric, Kolter, Zico, 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In: International Conference on Machine Learning, pp. 5286–5295.

Wong, Eric, Rice, Leslie, Kolter, J. Zico, 2020a. Fast is better than free: revisiting adversarial training. arXiv preprint. arXiv:2001.03994.

Wong, Eric, Schneider, Tim, Schmitt, Joerg, Schmidt, Frank R., Kolter, J. Zico, 2020b. Neural network virtual sensors for fuel injection quantities with provable performance specifications. arXiv preprint. arXiv:2007.00147.

Wu, Dongxian, Wang, Yisen, Xia, Shu-Tao, Bailey, James, Ma, Xingjun, 2020a. Skip connections matter: on the transferability of adversarial examples generated with resnets. In: International Conference on Learning Representations.

Wu, Dongxian, Xia, Shu-tao, Wang, Yisen, 2020b. Adversarial weight perturbation helps robust generalization. In: NeurIPS.

Xiao, Chaowei, Zhu, Jun-Yan, Li, Bo, He, Warren, Liu, Mingyan, Song, Dawn, 2018. Spatially transformed adversarial examples. In: International Conference on Learning Representations.

Xiao, Chang, Zhong, Peilin, Zheng, Changxi, 2019a. Enhancing adversarial defense by k-winners-take-all. arXiv preprint. arXiv:1905.10510.

Xiao, Kai Y., Tjeng, Vincent, Shafiullah, Nur Muhammad, Madry, Aleksander, 2019b. Training for faster adversarial robustness verification via inducing relu stability. In: ICLR.

Xie, Cihang, Wang, Jianyu, Zhang, Zhishuai, Ren, Zhou, Yuille, Alan, 2017. Mitigating adversarial effects through randomization. arXiv preprint. arXiv:1711.01991.

Xie, Cihang, Zhang, Zhishuai, Zhou, Yuyin, Bai, Song, Wang, Jianyu, Ren, Zhou, Yuille, Alan L., 2019. Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2730–2739.

Xie, Chulin, Huang, Keli, Chen, Pin-Yu, Li, Bo, 2020. DBA: Distributed backdoor attacks against federated learning. In: International Conference on Learning Representations.

Xu, Kaidi, Chen, Hongge, Liu, Sijia, Chen, Pin-Yu, Weng, Tsui-Wei, Hong, Mingyi, Lin, Xue, 2019a. Topology attack and defense for graph neural networks: an optimization perspective. In: IJCAI.

Xu, Kaidi, Liu, Sijia, Zhao, Pu, Chen, Pin-Yu, Zhang, Huan, Fan, Quanfu, Erdogmus, Deniz, Wang, Yanzhi, Lin, Xue, 2019b. Structured adversarial attack: towards general implementation and better interpretability. In: International Conference on Learning Representations.

Xu, Kaidi, Liu, Sijia, Chen, Pin-Yu, Sun, Mengshu, Ding, Caiwen, Kailkhura, Bhavya, Lin, Xue, 2020a. Towards an efficient and general framework of robust training for graph neural networks. In: ICASSP.

Xu, Kaidi, Shi, Zhouxing, Zhang, Huan, Wang, Yihan, Chang, Kai-Wei, Huang, Minlie, Kailkhura, Bhavya, Lin, Xue, Hsieh, Cho-Jui, 2020b. Automatic perturbation analysis for scalable certified robustness and beyond. Advances in Neural Information Processing Systems (NeurIPS).

Xu, Kaidi, Zhang, Gaoyuan, Liu, Sijia, Fan, Quanfu, Sun, Mengshu, Chen, Hongge, Chen, Pin-Yu, Wang, Yanzhi, Lin, Xue, 2020c. Adversarial t-shirt! Evading person detectors in a physical world. In: European Conference on Computer Vision, pp. 665–681.

Xu, Kaidi, Zhang, Huan, Wang, Shiqi, Wang, Yihan, Jana, Suman, Lin, Xue, Hsieh, Cho-Jui, 2021. Fast and complete: enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In: International Conference on Learning Representations (ICLR).

Yan, Xueting, Misra, Ishan, Gupta, Abhinav, Ghadiyaram, Deepti, Mahajan, Dhruv, 2020. Clusterfit: improving generalization of visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6509–6518.

Yang, Timothy, Andrew, Galen, Eichner, Hubert, Sun, Haicheng, Li, Wei, Kong, Nicholas, Ramage, Daniel, Beaufays, Françoise, 2018. Applied federated learning: improving Google keyboard query suggestions. arXiv preprint. arXiv:1812.02903.

Yang, Qiang, Liu, Yang, Chen, Tianjian, Tong, Yongxin, 2019a. Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2), 12.

Yang, Zhilin, Dai, Zihang, Yang, Yiming, Carbonell, Jaime, Salakhutdinov, Ruslan, Le, Quoc V., 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint. arXiv:1906.08237.

Yang, Zhuolin, Li, Bo, Chen, Pin-Yu, Song, Dawn, 2019c. Characterizing audio adversarial examples using temporal dependency. In: International Conference on Learning Representations.

Yang, Chao-Han, Qi, Jun, Chen, Pin-Yu, Ma, Xiaoli, Lee, Chin-Hui, 2020a. Characterizing speech adversarial examples using self-attention u-net enhancement. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3107–3111.

Yang, Chao-Han Huck, Qi, Jun, Chen, Pin-Yu, Ouyang, Yi, Hung, I., Danny, Te, Lee, Chin-Hui, Ma, Xiaoli, 2020b. Enhanced adversarial strategically-timed attacks against deep reinforcement learning. In: ICASSP.

Yang, Greg, Duan, Tony, Hu, J. Edward, Salman, Hadi, Razenshteyn, Ilya, Li, Jerry, 2020c. Randomized smoothing of all shapes and sizes. In: International Conference on Machine Learning. PMLR, pp. 10693–10705.

Yang, Puyudi, Chen, Jianbo, Hsieh, Cho-Jui, Wang, Jane-Ling, Jordan, Michael, 2020d. Ml-loo: Detecting adversarial examples with feature attribution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6639–6647.

Yang, Puyudi, Chen, Jianbo, Hsieh, Cho-Jui, Wang, Jane-Ling, Jordan, Michael I., 2020e. Greedy attack and Gumbel attack: generating adversarial examples for discrete data. Journal of Machine Learning Research 21 (43), 1–36.

Yang, Yao-Yuan, Rashtchian, Cyrus, Wang, Yizhen, Chaudhuri, Kamalika, 2020f. Robustness for non-parametric classification: a generic attack and defense. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 941–951.

Yang, Chao-Han Huck, Qi, Jun, Chen, Samuel Yen-Chi, Chen, Pin-Yu, Marco Siniscalchi, Sabato, Ma, Xiaoli, Lee, Chin-Hui, 2021a. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6523–6527.

Yang, Chao-Han Huck, Tsai, Yun-Yun, Chen, Pin-Yu, 2021b. Voice2series: reprogramming acoustic models for time series classification. In: International Conference on Machine Learning.

Yang, Chao-Han Huck, Hung, I., Danny, Te, Ouyang, Yi, Chen, Pin-Yu, 2022. Training a resilient Q-network against observational interference. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Yao, Yuanshun, Li, Huiying, Zheng, Haitao, Ben Zhao, Y., 2019. Latent backdoor attacks on deep neural networks. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 2041–2055.

Yen, Hao, Ku, Pin-Jui, Huck Yang, Chao-Han, Hu, Hu, Marco Siniscalchi, Sabato, Chen, Pin-Yu, Tsao, Yu, 2021. A study of low-resource speech commands recognition based on adversarial reprogramming. arXiv preprint. arXiv:2110.03894.

Yin, Chengxiang, Tang, Jian, Xu, Zhiyuan, Wang, Yanzhi, 2018. Adversarial meta-learning.

Yuan, Xuejing, Chen, Yuxuan, Zhao, Yue, Long, Yunhui, Liu, Xiaokang, Chen, Kai, Zhang, Shengzhi, Huang, Heqing, Wang, Xiaofeng, Gunter, Carl A., 2018. Commandersong: A systematic approach for practical adversarial voice recognition. arXiv preprint. arXiv:1801.08535.

Zantedeschi, Valentina, Nicolae, Maria-Irina, Rawat, Ambrish, 2017. Efficient defenses against adversarial attacks. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 39–49.

Zawad, Syed, Ali, Ahsan, Chen, Pin-Yu, Anwar, Ali, Zhou, Yi, Baracaldo, Nathalie, Tian, Yuan, Yan, Feng, 2021. Curse or redemption? How data heterogeneity affects the robustness of federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: ECCV. Springer, pp. 818–833.

Zeng, Xiangrong, Figueiredo, Mário AT, 2014a. Decreasing weighted sorted ℓ_1 regularization. IEEE Signal Processing Letters 21 (10), 1240–1244.

Zeng, Xiangrong, Figueiredo, Mario AT, 2014b. Solving oscar regularization problems by fast approximate proximal splitting algorithms. Digital Signal Processing 31, 124–135.

Zhai, Xiaohua, Oliver, Avital, Kolesnikov, Alexander, Beyer, Lucas, 2019. S4l: self-supervised semi-supervised learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1476–1485.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, Vinyals, Oriol, 2017. Understanding deep learning requires rethinking generalization. In: International Conference on Learning Representations.

Zhang, Huan, Weng, Tsui-Wei, Chen, Pin-Yu, Hsieh, Cho-Jui, Daniel, Luca, 2018. Efficient neural network robustness certification with general activation functions. In: Advances in Neural Information Processing Systems, pp. 4944–4953.

Zhang, Dinghuai, Zhang, Tianyuan, Lu, Yiping, Zhu, Zhanxing, Dong, Bin, 2019a. You only propagate once: accelerating adversarial training via maximal principle. arXiv preprint. arXiv:1905.00877.

Zhang, Hongyang, Yu, Yaodong, Jiao, Jiantao, Xing, Eric, El Ghaoui, Laurent, Jordan, Michael, 2019b. Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. PMLR, pp. 7472–7482.

Zhang, Huan, Chen, Hongge, Xiao, Chaowei, Li, Bo, Boning, Duane, Hsieh, Cho-Jui, 2020a. Towards stable and efficient training of verifiably robust neural networks. In: International Conference on Learning Representations (ICLR).

Zhang, Wei Emma, Sheng, Quan Z., Alhazmi, Ahoud, Li, Chenliang, 2020b. Adversarial attacks on deep-learning models in natural language processing: a survey. ACM Transactions on Intelligent Systems and Technology (TIST) 11 (3), 1–41.

Zhang, Gaoyuan, Lu, Songtao, Zhang, Yihua, Chen, Xiangyi, Chen, Pin-Yu, Fan, Quanfu, Martie, Lee, Horesh, Lior, Hong, Mingyi, Liu, Sijia, 2022. Distributed adversarial training to robustify deep neural networks at scale. In: The Conference on Uncertainty in Artificial Intelligence.

Zhao, Yue, Li, Meng, Lai, Liangzhen, Suda, Naveen, Civin, Damon, Chandra, Vikas, 2018. Federated learning with non-iid data. arXiv preprint. arXiv:1806.00582.

Zhao, Pu, Liu, Sijia, Chen, Pin-Yu, Hoang, Nghia, Xu, Kaidi, Kailkhura, Bhavya, Lin, Xue, 2019a. On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method. In: IEEE International Conference on Computer Vision, pp. 121–130.

Zhao, Pu, Wang, Siyue, Gongye, Cheng, Wang, Yanzhi, Fei, Yunsi, Lin, Xue, 2019b. Fault sneaking attack: a stealthy framework for misleading deep neural networks. In: ACM/IEEE Design Automation Conference (DAC), pp. 1–6.

Zhao, Pu, Chen, Pin-Yu, Das, Payel, Ramamurthy, Karthikeyan Natesan, Lin, Xue, 2020a. Bridging mode connectivity in loss landscapes and adversarial robustness. In: International Conference on Learning Representations.

Zhao, Pu, Chen, Pin-Yu, Wang, Siyue, Lin, Xue, 2020b. Towards query-efficient black-box adversary with zeroth-order natural gradient descent. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Zhu, Xiaojin, Goldberg, Andrew B., 2009. Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 3 (1), 1–130.

Zhu, Chen, Huang, W. Ronny, Li, Hengduo, Taylor, Gavin, Studer, Christoph, Goldstein, Tom, 2019a. Transferable clean-label poisoning attacks on deep neural nets. In: Proceedings of the 36th International Conference on Machine Learning, pp. 7614–7623.

Zhu, Chen, Huang, W. Ronny, Li, Hengduo, Taylor, Gavin, Studer, Christoph, Goldstein, Tom, 2019b. Transferable clean-label poisoning attacks on deep neural nets. In: International Conference on Machine Learning. PMLR, pp. 7614–7623.

Zhu, Sicheng, Zhang, Xiao, Evans, David, 2020a. Learning adversarially robust representations via worst-case mutual information maximization. In: International Conference on International Conference on Machine Learning.

Zhu, Xizhou, Su, Weijie, Lu, Lewei, Li, Bin, Wang, Xiaogang, Dai, Jifeng, 2020b. Deformable detr: deformable transformers for end-to-end object detection. arXiv preprint. arXiv:2010.04159.

Zhuang, Juntang, Gong, Boqing, Yuan, Liangzhe, Cui, Yin, Adam, Hartwig, Dvornek, Nicha, Tatikonda, Sekhar, Duncan, James, Liu, Ting, 2022. Surrogate gap minimization improves sharpness-aware training. In: International Conference on Learning Representations (ICLR).

Zou, Hui, Hastie, Trevor, 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, Statistical Methodology 67 (2), 301–320.

Zügner, Daniel, Günnemann, Stephan, 2019. Adversarial attacks on graph neural networks via meta learning. In: International Conference on Learning Representations.

Zügner, Daniel, Akbarnejad, Amir, Günnemann, Stephan, 2018. Adversarial attacks on neural networks for graph data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2847–2856.