8 REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[3] Nguyen, Anh & Yosinski, Jason & Clune, Jeff. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 427-436. 10.1109/CVPR.2015.7298640.

[4] Papernot, Nicolas & McDaniel, Patrick & Jha, Somesh & Fredrikson, Matt & Celik, Z. Berkay & Swami, Ananthram. (2016). The Limitations of Deep Learning in Adversarial Settings. 372-387. 10.1109/EuroSP.2016.36.

[5] Y. LeCun, L. Bottou, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[6] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," 2016 EEE Symposium on Security and Privacy (SP), San Jose, CA, 2016, pp. 582-597, doi: 10.1109/SP.2016.41.

[7] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in Deep Learning and Representation Learning Workshop at NIPS 2014. arXiv preprint arXiv:1503.02531, 2014.

[8] Goodfellow, Ian & Shlens, Jonathon & Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.

[9] Kurakin, Alexey & Goodfellow, Ian & Bengio, Samy. (2016). Adversarial examples in the physical world.

[10] S. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2574-2582, doi: 10.1109/CVPR.2016.282.

[11] Moosavi-Dezfooli, Seyed-Mohsen & Fawzi, Alhussein & Frossard, Pascal. (2016). DeepFool: a simple and accurate method to fool deep neural networks. CVPR.

[12] Gu, Shixiang & Rigazio, Luca. (2014). Towards Deep Neural Network Architectures Robust to Adversarial Examples.

[13] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In Joint European conference on machine learning and knowledge discovery in databases, pp. 387–402. Springer, 2013.

[14] Szegedy, Christian & Zaremba, Wojciech & Sutskever, Ilya & Bruna, Joan & Erhan, Dumitru & Goodfellow, Ian & Fergus, Rob. (2013). Intriguing properties of neural networks.

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

[16] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 39-57, doi: 10.1109/SP.2017.49.

[17] Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612, 2018b.

[18] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. 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.

[19] Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017.

[20] Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. 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.

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

[22] Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1885–1894. JMLR. org, 2017.

[23] Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean label poisoning attacks on neural networks. In Advances in Neural Information Processing Systems, pp. 6103–6113, 2018a.

[24] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945, 2017.

[25] Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=S18Su--CW.

[26] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[27] Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068, 2014.

[28] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[29] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. ArXiv preprint arXiv:1705.07204, 2017.

[30] Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! ArXiv preprint arXiv:1904.12843, 2019.

[31] Carlini, N., Katz, G., Barrett, C., and Dill, D. L. Provabl minimally distorted adversarial examples. arXiv preprint arXiv:1709.10207, 2017.

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

[33] Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344, 2018a.

[34] Carlini, N. and Wagner, D. 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, 2017a.

[35] Grosse, K., Manoharan, P., Papernot, N., Backes, M., and Mc- Daniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280, 2017

[36] Hendrycks, D. and Gimpel, K. Early methods for detecting adversarial images. arXiv preprint arXiv:1608.00530, 2016.

[37] Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155, 2017.

[38] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. 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. ACM, 2016.

[39] Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., and Yuille, A. Adversarial examples for semantic segmentation and object detection. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1369–1378, 2017b.

[40] Xie, Meiyan & Xue, Yunzhe & Roshan, Usman. (2019). Stochastic Coordinate Descent for 01 Loss and Its Sensitivity to Adversarial Attacks. 299-304. 10.1109/ICMLA.2019.00056.