ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

4 Conclusions

We proposed an iterative algorithm to find small adversarial perturbations that fool a given set of models simultaneously in a given pattern. This problem formulation has several applications including the generation of transferable adversarial examples, as well as *non-transferable* examples that target only a specific model and ensure that the other models are safe.

The algorithm applies the first-order approximation of the decision boundaries used in the DeepFool method. We evaluated the algorithm on a number of model sets over MNIST and CIFAR-10. We found that the algorithm consistently produces small perturbations in all the cases we examined. Perhaps the most interesting result is that small adversarial perturbations are present even when a non-transferable adversarial example was generated for the most robust model in the set, despite the fact that the models differed only in the regularization coefficient. The generalization of the method and making improvements to its convergence speed are under way.

References

[1] Ian J. Goodfellow and Jonathon Shlens Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd Intl. Conf. on Learning Representations (ICLR)*, 2015.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd Intl. Conf. on Learning Representations (ICLR)*, 2014.

[3] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, June 2016.

[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th Intl. Conf. on Learning Representations (ICLR)*, 2018.

[6] Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th Intl. Conf. on Learning Representations (ICLR)*, 2018.

[7] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Networks and Learning Syst.*, 30(9):2805–2824, Sep. 2019.

[8] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. 5th Intl. Conf. on Learning Representations (ICLR)*, 2017.

[9] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *Proc. of the 36th Intl. Conf. on Machine Learning, (ICML)*, pages 4970–4979, 2019.

[10] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *Proc. 6th Intl. Conf. on Learning Representations (ICLR)*, 2018.

[11] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, November 1998.

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

[13] Jimmy Ba and Diederik Kingma. Adam: A method for stochastic optimization. In *3rd Intl. Conf. on Learning Representations (ICLR)*, 2015.