A New Algorithm for Crafting Adversarial Examples

Shang Da
Iowa State University
Computer Science Department
Email: dashang@iastate.edu

***Abstract*—Deep neural networks(DNNs) achieve high success rate in a large variety of tasks such as image classification, natural language processing, reinforcement learning. Recent research shows that DNNs are vulnerable to adversarial attacks, where adversarial examples are carefully crafted by adding imperceptible perturbation to natural samples. This technique report introduce a new algorithm for crafting adversarial examples. It uses information provided by decision boundaries obtained from a SVM that trained on output of last hidden layer to decided direction of perturbation. In an application to computer vision, the algorithm reliably produce untargeted adversarial examples that are close to natural samples but misclassified by a DNN with 99.9% success rate. Improving algorithm for targeted attack and study of attack generalization is the next goal.**

References

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

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[3] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Classification regions of deep neural networks," *arXiv preprint arXiv:1705.09552*, 2017.

[4] Y. Li, L. Ding, and X. Gao, "On the decision boundary of deep neural networks," *CoRR*, vol. abs/1808.05385, 2018. [Online]. Available: http://arxiv.org/abs/1808.05385

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

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

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

[8] A. Sinha, H. Namkoong, and J. Duchi, "Certifiable distributional robustness with principled adversarial training," *stat*, vol. 1050, p. 29, 2017.