# 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.**

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

## References

[1] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[2] F. Berkhahn, R. Keys, W. Ouertani, N. Shetty, and D. Geißler, "One model to rule them all," *arXiv preprint arXiv:1908.03015*, 2019.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] Y. Cheng, "Semi-supervised learning for neural machine translation," in *Joint Training for Neural Machine Translation*. Springer, 2019, pp. 25–40.

[7] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6940–6944.

[8] R. Fergus, Y. Weiss, and A. Torralba, "Semi-supervised learning in gigantic image collections," in *Advances in neural information processing systems*, 2009, pp. 522–530.

[9] M. Shi and B. Zhang, "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," *Bioinformatics*, vol. 27, no. 21, pp. 3017–3023, 2011.

[10] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Advances in neural information processing systems*, 2015, pp. 919–927.