Supplementary Material for Adversarial Deep Learning with Stackelberg Games

No Author Given

No Institute Given

1 Related Work

In this section we provide additional comparison of our game theoretical adversarial learning model with existing deep generative models in literature.

1.1 Adversarial Learning Benchmarks for Misleading Classifiers

In trying to interpret the solutions in deep neural networks, Szegedy et al. [9] introduced adversarial examples as discontinuities in input-output mappings learnt by deep neural networks. Goodfellow et al. [4] propose a training regime called Fast Gradient Sign Method (FGSM) for generating adversarial examples as analytical perturbations to a linear model that can be computed efficiently using backpropagation. Papernot et al. [8] introduce a blackbox attack strategy where adversarial examples are generated without knowledge of target deep neural networks internals and inputs. Baluja et al. [1] proposes a targeted attack where feed-forward neural networks called Adversarial Transformation Networks (ATNs) are trained to generate adversarial examples against adversarial targets in deep neural network. ATNs generate adversarial examples that minimally modify classifier's outputs given original input. By contrast, Moosav et al. [7] construct an untargeted attack technique DeepFool which is optimized by distance metrics between adversarial examples and normal examples.

1.2 Generative Adversarial Networks for Adversarial Learning

Generative Adversarial Networks (GANs) [3] estimate data likelihood with a adversarial framework involving a two-player game between a generator network G and a discriminator network D. IWGAN [5] improves GAN with regularization that does not introduce correlations between generated examples. InfoGAN [2] proposes an information-regularized generator to disentangle interpretable representations from generated data. The distribution of adversarial perturbations have been modelled with AdvGAN [11] in whitebox attacks as well as blackbox attacks. By using adversarial generator, Adversarial Autoencoders [10] impose a prior distribution to the output of encoder network learning latent representations of training data. In this architecture, a new autoencoder network is trained to discriminatively predict whether a sample comes from latent space of autoencoder or from prior distribution determined by the user. Adversarial examples have been defined for deep generative models [6]. The objective of our game formulation is not to improve classification accuracy by augmenting the original data training autoencoders. We solve a supervised learning problem while deep generative models generally solve either unsupervised learning or semi-supervised learning problems.

References

1. Baluja, S., Fischer, I.: Learning to attack: Adversarial transformation networks. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
2. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems 29 (2016)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (NIPS) (2014)
4. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of International Conference on Learning Representations (2015)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems 30 (2017)
6. Kos, J., Fischer, I., Song, D.: Adversarial examples for generative models. In: Proceedings of 2018 IEEE Security and Privacy Workshops (SPW) (2018)
7. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition CVPR (2016)
8. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (2017)
9. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of International Conference on Learning Representations (2014)
10. Tran, N.T., Bui, T.A., Cheung, N.M.: Dist-gan: An improved gan using distance constraints. In: Proceedings of European Conference on Computer Vision (ECCV) (2018)
11. Xiao, C., Li, B., Zhu, J., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI (2018)