Adversarial Learning Games with Deep Learning Models

Aneesh Chivukula and Wei Liu
Advanced Analytics Institute, University of Technology Sydney, Australia

Abstract and Motivation

- We design an adversarial learning algorithm for supervised learning and deep learning.
- Adversarial examples are generated by a game theoretic formulation on the performance of deep learning.
- The interaction between an intelligent adversary and deep learning model is a two-person sequential noncooperative Stackelberg game with stochastic payoff functions.
- The Stackelberg game is solved by the Nash equilibrium which is a pair of strategies (learner weights and genetic operations) from which there is no incentive for either learner or adversary to deviate.

Related Work and Proposed Algorithm

- Designing robust computing systems and machine learning algorithms for non-stationary data is the goal of adversarial learning.
- Adversarial learning is simulated by training a learning algorithm under various attack scenarios formulated by an intelligent adversary Huang et al. (2011).
- Adversarial learning is simulated by training a learning algorithm under various attack scenarios formulated by an intelligent adversary.
- The optimal attack policy for the adversary is defined in terms of the solution to an objective function.
- Adversarial examples can be crafted by prior knowledge, observation, and experimentation on the network layers and loss functions in the deep learning model.
- The existing adversarial learning algorithms are summarized in Table 1, Table 2.

Adversarial algorithm	**Attack strategy**	**Search algorithm**
Classifier ensembles Biggio et al. (2010)	Reorder features by importance for discriminant function	Randomized sampling
Feature weighting Kołcz and Teo (2009)	Addition/deletion of binary features	Feature bagging
SVM : inputs Biggio et al. (2012)	Train noise injection	Gradient ascent
SVM : labels Xiao et al. (2015)	Label noise injection	Gradient ascent
Deep learning Goodfellow et al. (2014)	Linear perturbation on x	Backpropagation with L-BFGS
Adversarial networks : DNN Papernot et al. (2016)	Observe DNN outputs given inputs chosen by the adversary	Jacobian-based dataset augmentation
Adversarial networks : DAE Gu and Rigazio (2014)	Gaussian additive noise	Stacking DAEs into a feed forward neural network
Game theory : support vector machines Globerson and Roweis (2006)	Delete different features from different data points	Quadratic programming
Game theory : deep learning (Our method)	Move positive samples towards negative samples	Genetic algorithm

Table 1: Adversarial Algorithms Comparision

Adversarial algorithm	**Adversary's knowledge**	**Learning games**
Classifier ensembles Biggio et al. (2010)	Train features	None
Feature weighting Kołcz and Teo (2009)	Train features	None
SVM : inputs Biggio et al. (2012)	Gradient of loss	None
SVM : labels Xiao et al. (2015)	Train labels	None
Deep learning Goodfellow et al. (2014)	Train and test data	None
Adversarial networks : DNN Papernot et al. (2016)	Test data	None
Adversarial networks : DAE Gu and Rigazio (2014)	Test data	None
Game theory : support vector machines Globerson and Roweis (2006)	Train features	Non-zero sum game
Game theory : deep learning (Our method)	Train and test data	Constant sum game

Table 2: Adversarial Algorithms Comparison

- We formulate the problem of finding defence mechanisms in adversarial learning as a maxmin optimization problem in learning-theoretic game theory.

$$Maxmin : (\alpha^*, w^*) = argmax_{\alpha \in A} J_L(\alpha, argmin_{w \in W} J_L(\alpha, w)) \quad (1)$$

- The attack processes specify the adversary's constraints and optimal attack policy. The learning processes specify the learner's gain and adversary's gain under the optimal policy.

$$J_L(\alpha, w) = 1 + \lambda * error(w) - cost(\alpha) \quad (2)$$

- The optimal attack policy is formulated in terms of stochastic optimization and evolutionary computing.

$$error(w) = 1 - recall(w) \quad (3)$$

$$cost(\alpha) = \sqrt{\Sigma \alpha^2 / (32 * 32 * 3)} / 255 \quad (4)$$

- Our algorithm can adapt to continuous adversarial data manipulations unlike most of the existing adversarial learning algorithms.
- We do not assume the adversary knows anything about the deep network structure which is close to real life settings.
- The adversarial data is constructed by the mutation, crossover, selection genetic operators defined on images.

Experiments and Analysis

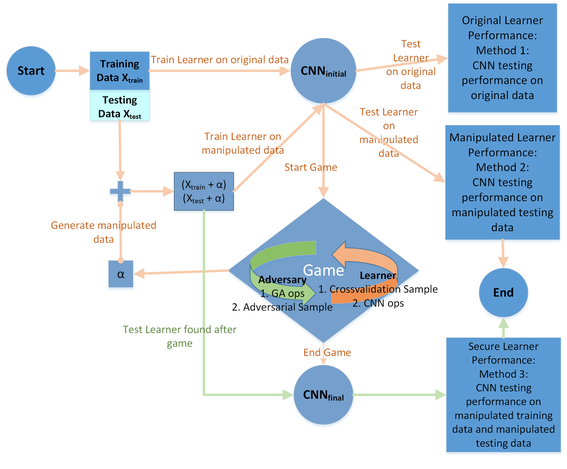

Figure 1: A flow chart illustrating the benefits of a game theoretic learner. The game has Adversary and Learner as the players. The game produces a final deep learning network CNN_{final} that is better equipped to deal with the adversarial manipulations than the initial deep learning network $CNN_{initial}$.

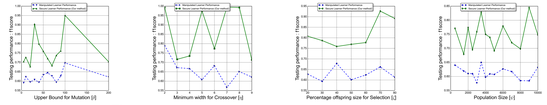

Figure 2: Performance Evaluation - f1scores before and after game.

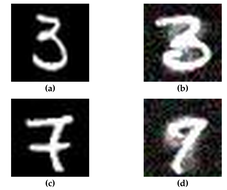

Figure 3: Examples of transformed images found at Nash equilibrium in a Stackelberg game. To avoid detection, the adversary adds pixels in (b), deletes pixels in (d) and changes shape in both (b) and (d)

Conclusion

- We have presented a maxmin problem for adversarial learning in deep learning networks.
- We propose a secure learner that is immune to the adversarial attacks on deep learning.

Acknowledgements

This research is funded by the Capital Markets Co-operative Research Centre.

References

L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 43–58.

B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems for robust classifier design in adversarial environments," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 27–41, 2010.

A. Kołcz and C. H. Teo, "Feature weighting for improved classifier robustness," in *CEAS09: sixth conference on email and anti-spam*, 2009.

B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *29th International Conference on Machine Learning (ICML)*, pp. 1807–1814, Jun. 2012.

H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, "Support vector machines under adversarial label contamination," *Neurocomputing*, vol. 160, pp. 53–62, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems (NIPS)*, 2014, pp. 2672–2680.

N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *2017 ACM Asia Conference on Computer and Communications Security*, 2016.

S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *CoRR*, vol. abs/1412.5068, 2014.

A. Globerson and S. Roweis, "Nightmare at test time: robust learning by feature deletion," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 353–360.