Refining Adversarial Training Methods Using Game Theory
15-300, Fall 2020

Akhil Nadigatla
http://akhilnadigatla.com/15400-proposal.pdf

November 4, 2020

1 Project Description

I will be working with Professor Pradeep Ravikumar and one of his graduate students, Arun Sai Suggala, in the Machine Learning Department on trying to improve and/or find new adversarial training methods using game theory principles.

Adversarial examples represent small perturbations to the input with respect to some distance measure that change the decision of the classifier. As expected, these compromise the safety and robustness of machine learning models. One defensive approach against these is to conduct adversarial training. Most of the algorithms proposed for adversarial training – including Projected Gradient Descent (PGD), Fast Gradient Sign Method (FGSM), or Randomized Smoothing – rely on heuristics. This means that they are not guaranteed to provide a model that exhibits the greatest robustness against these adversarial samples. However, studies conducted towards evaluating the adversarial robustness portrayed by existing training techniques against a (parameter-free, computationally affordable, and user-independent) ensemble of attacks showed almost all of the models achieved lower robust test accuracy than expected and/or reported in their respective papers, often by more than 10%.

Our direction for this project is to understand what is wrong with existing algorithms and see if we can bolster them using game theoretic methods. This boils down to viewing adversarial risk as a zero-sum game between a defender (the user building the model) and attacker. Specifically, the game is also likely to be Bayesian given that the defender might not know the exact cost of generating adversarial data and the attacker might not know the exact classification cost to the defender. Adversarial training techniques built on a game theoretic framework may be of significance as it helps represent the behaviors of the defender and attacker as it includes the benefit for the adversary to attack and the cost of generating the adversarial data to be fed to the model during the training phase and, from the other side, the cost to the learner to update the model.

If we are successful in this endeavor, we will be one step closer towards more robust machine learning systems, a crucial requirement for safety-critical applications. Moreover, we would have provided an alternative perspective for adversarial risk and training that may be built upon in the future.

A majority of the research work conducted in this project will involve testing the adversarial accuracy of proposed adversarial training algorithms on models built on the MNIST and CIFAR-10 image datasets. The reason for why these two have been chosen is because all landmark papers that introduced methods like PGD and FGSM presented their findings using models trained on these particular datasets, and the use of these will provides more direct points for comparison for the progress made by our research.

The major challenge going into this project will be consolidate the research done in the field of adversarial training so far to try and pinpoint the causes behind the shortcomings of current methods. While there is a lot of speculations behind each training algorithm's reason for lower-than-expected adversarial robustness, there is little to no facts regarding these. While we are not aiming to mathematically prove these pitfalls across all the training algorithms, it will be in the scope of this project to understand the common denominator behind the observations.

2 Project Goals

2.1 75% Project Goal

- Determine the issues in terms of adversarial accuracy for popular training algorithms including (but not limited to) PGD and its variants, FGSM, and randomized smoothing-based methods, using parameter-free attacks.
- Use the above findings (or otherwise) propose a few applications of existing game theory algorithms towards adversarial risk in the game setting mentioned above.

2.2 100% Project Goal

- Determine the issues in terms of adversarial accuracy for popular training algorithms including (but not limited to) PGD and its variants, FGSM, and randomized smoothing-based methods, using parameter-free attacks.
- Use the above findings (or otherwise) propose a few applications of existing game theory algorithms towards adversarial risk in the game setting mentioned above.
- Refine the above proposed algorithms by providing guarantees on which kinds of adversarial attacks they are able to resist.
- Provide empirical evidence – using models trained with MNIST and CIFAR-10 – that show how the game-theoretic training algorithms fare in terms of adversarial accuracy.

2.3 125% Project Goal

- Determine the issues in terms of adversarial accuracy for popular training algorithms including (but not limited to) PGD and its variants, FGSM, and randomized smoothing-based methods, using parameter-free attacks.
- Use the above findings (or otherwise) propose a few applications of existing game theory algorithms towards adversarial risk in the game setting mentioned above.
- Refine the above proposed algorithms by providing guarantees on which kinds of adversarial attacks they are able to resist.
- Provide empirical evidence – using models trained with MNIST and CIFAR-10 – that show how the game-theoretic training algorithms fare in terms of adversarial accuracy.
- Compare the above figures to claims made about the different conventional training algorithms in their respective papers and show any improvements/deficiencies in our methods.
- Explain the reasons behind why our methods perform they way they do in comparison to existing algorithms.
- Attempt to scale up our techniques to larger datasets like ImageNet.

3 Project Milestones

3.1 First Technical Milestone:

Get a better understanding of the workings behind the training algorithms I will investigate in the research project, which includes the motivations behinds them, how they are defined, and their objective functions. This may even involve training models using the MNIST and CIFAR-10 datasets and observing their adversarial accuracy, comparing them to provided estimates for this in the literature.

3.2 First Biweekly Milestone: February 15th

I hope to have run several tests on the aforementioned models. By doing so, I hope to better recognize some of the limitations of these algorithms, comparing this to what has already been said about them in the literature.

3.3 Second Biweekly Milestone: March 1st

I intend to identify some common causes for the observations made with respect to adversarial robustness for the common training algorithms. This will motivate what are the specific improvements required to better adversarial training.

3.4 Third Biweekly Milestone: March 15th

I hope to have begun exploring adversarial risk from a game theory perspective, understanding how the field applies to the problem at hand. I may also begin to explore potential algorithms for the above proposed game.

3.5 Fourth Biweekly Milestone: March 29th

I hope to have picked out one or more new training algorithms based on game theory, and detailed how they theoretically provide adversarial robustness. I may have also begun training models based on adversarial examples generated using these methods.

3.6 Fifth Biweekly Milestone: April 12th

I intend to have publishable results on the performance of the game-theoretic training methods in terms of adversarial robustness, and will have compared them to existing algorithms. I will also have explained why the proposed algorithms perform they way they do to some extent.

3.7 Sixth Biweekly Milestone: April 26th

I hope to have integrated all the work we have done so far into a draft paper. If done early, we may also begin to explore some of the ideas beyond the 100% project goal to bolster our results.

3.8 Seventh Biweekly Milestone: May 10th

I intend to have my project complete as proposed and will have tried to establish future directions of research that build on our findings. These may even form the backbone for my undergraduate senior thesis.

4 Literature Search

While viewing machine learning from a game theory perspective is not novel by any means, it is definitely a budding area, especially when it comes to adversarial robustness and training. There have been a couple of interesting discussions about game theory applied to general adversaries and the security of machine learning models [1] [2] [3]. There is a lot of material with regard to adversarial training methods (and adversarial risk in general), which I have begun to look deeper into as draw closer towards the first milestone [4] [5] [6] [7] [8] [9] [10] [11] [12].

5 Resources Needed

Our main resources will be Python and the deep learning framework PyTorch, all of which can be setup on Anaconda. We will make use of the MNIST and CIFAR-10 image datasets to train the models. In addition, if time permits, we may also attempt to train models using the CIFAR-100 and ImageNet datasets which have defined classes, which typically require a lot more resources in terms of time and computing power. We also intend to utilize the Machine Learning Department's GPU resources for this project, given the large number of models (of high volume) that we hope to train and evaluate in this research.

References

[1] Avrim Blum. Machine learning, game theory, and mechanism design for a networked world. 2006.

[2] Prithviraj Dasgupta and Joseph Collins. A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. 40(2), 2019.

[3] Wei Liu and Sanjay Chawla. A game theoretical model for adversarial learning. 2009.

[4] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020.

[6] Lemonia Dritsoula, Patrick Loiseau, and John Musacchio. A game-theoretic analysis of adversarial classification, 2017.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[8] David Rios Insua, Roi Naveiro, Victor Gallego, and Jason Poulos. Adversarial machine learning: Perspectives from adversarial risk analysis, 2020.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[10] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning, 2017.

[11] Arun Sai Suggala, Adarsh Prasad, Vaishnavh Nagarajan, and Pradeep Ravikumar. Revisiting adversarial risk, 2019.

[12] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. 2020.