Project Prosal: Robust Machine Learning and Adversarial Attacks

AM 221 Advanced Optimization

Vishnu Rudrasamudram

February 27, 2019

For the project, the topic I have chosen is "Robust Machine Learning and Adversarial Attacks". I will conduct a literature survey for the existing attacking methods and the relevant optimization techniques for computing the adversarial noise. Then, I will proceed to implementing and comparing different methods for attacking that I found as part of literature survey. Next steps would be, after implementation, to empirically asses and compare their performances. If time permits, I will study the effect of adverse noise in reinforcement learning setting, where deep learning models are used to learn control policies [1].

1 Motivation

With significant progress in a wide spectrum applications, deep learning is being employed in various applications. Some of them, like autonomous vehicles where they require to perceive the environment using cameras and lidars [2], are safety critical. The same techniques are used to train agents to play atari games and achieve super-human level performance [1]. However, a recent study has found the deep neural networks vulnerable to well-designed input samples, called adversarial examples. Adversarial examples look quite normal to a human but can easily fool deep neural networks in various stages. This problem arises mainly because of the non-convex nature of neural network optimization which makes it very difficult to understand the failure cases of these systems.

As the systems are becoming increasingly safety-critical, studying such attacks to find ways of defending against them is an extremely important task.

2 Problem Statement

Given a trained deep learning model f and an original input data sample $\mathbf{x}$, generating an adversarial example $\mathbf{y}$ can be formulated as a box-constrained optimization problem [3]:

$$
\begin{aligned}
\operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d} \ & \|\mathbf{y}-\mathbf{x}\| \\
\text{s.t. } & f(\mathbf{y}) = l', \\
& f(\mathbf{x}) = l, \\
& l \neq l', \\
& \mathbf{y} \in [0,1],
\end{aligned}
$$

where l and l' denote the output labels of $\mathbf{x}$ and $\mathbf{y}$, $\|\cdot\|$ denotes the distance between two data samples. Let us define $\mathbf{r} = \mathbf{y}-\mathbf{x}$, the perturbation added on $\mathbf{x}$. This optimization problem minimizes the perturbation while mis-classifying the prediction with a constraint of input data. This optimization problem formulation captures various methods used to generate adversarial examples.

3 Algorithms

Depending upon the objectives of the adversaries, adversarial examples can, generally, be put under two categories: mis-classification attacks and targeted attacks. To generate such adversarial, a number of algorithms have been proposed, such as the Fast Gradient Sign Method (FGSM) by Goodfellow et. al. [4], and the Jacobian Saliency Map Algorithm (JSMA) approach by Papernot et. al., [5] . Most of the algorithms assume that the details of the neural network to be attacked are available. A black-box approach to generating adversarial examples is also proposed by Papernot et. al., [6].

The list of algorithms mentioned above is not exhaustive. There are only few algorithms that are suitable for reinforcement learning setting. So, I intend to finalize the algorithms to be considered for implementation and evaluation after the literature survey.

4 Data and Experiment Setup

I am planning to primarily work with MNIST [7] and CIFAR-10 [8] datasets with PyTorch [9] as a learning framework. If time permits I will be working on studying the effect of adversarial noise on performance of DQN for which I will use OpenAI gym (Atari Games) [10].

5 Deliverables

I will implement the approaches, with and without the knowledge of the target network, that I found through literature survey and make an attempt to study the algorithms on DQN. The deliverable will be a survey of recent literature and report along with the code.

6 Next Steps

I will start with reviewing the current literature and get familiarize with state-of-the-art algorithms for adversarial attacks.

Below is the tentative plan for milestone 2

- Done with literature survey and finalize the algorithms to implement
- Gather data and set up the environment for experiments
- Implement few of the finalized algorithms
- Prepare status report

References

[1] Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves et al. "Human-level control through deep reinforcement learning." Nature 518, no. 7540 (2015): 529.

[2] Bojarski, Mariusz, et al. "End to end learning for self-driving cars." arXiv preprint arXiv:1604.07316 (2016).

[3] Yuan, Xiaoyong, et al. "Adversarial examples: Attacks and defenses for deep learning." IEEE transactions on neural networks and learning systems (201

[4] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[5] Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." 2016 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2016.

[6] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM, 2017.

[7] MNIST Dataset `http://yann.lecun.com/exdb/mnist/`

[8] CIFAR-10 Dataset `https://www.cs.toronto.edu/~kriz/cifar.html`

[9] PyTorch `pytorch.org`

[10] OpenAI Gym `gym.openai.com`