CS 6220 - Big Data Systems and Analytics

Saurab Sirpurkar
Assignment 2

24 September 2021

A critique on
Adversarial Deception in Deep Learning: Analysis and Mitigation

The authors of the paper argue the importance of studying adversarial attacks on deep learning models, and give us an exhaustive characterisation of adversarial attacks. Based on the study of instance-level divergence of these attacks, they propose the *Strategic Input Teaming Defense* (SITD) technique that combats the problem of adversarial attacks, on par with the existing defense paradigms. They also holistically evaluate 15 adversarial attacks and perform 3 case studies on instance-level divergence of these attacks. On top of designing a state-of-the-art mitigation strategy, the paper has a significant contribution in attack characterization through introduction of terminology to quantify both the adverse effect of an attack and its perturbation cost (such as DistACOC and DistPercept-SSIM respectively), which help us talk numerically about these attacks. For each of the 15 considered attacks, they calculate these effect and cost metrics and arrive at meaningful observations about the attacks and the metrics.

The proposed adversary mitigation strategy, *SITD*, is centered around instance-level divergence. Consider the following image:

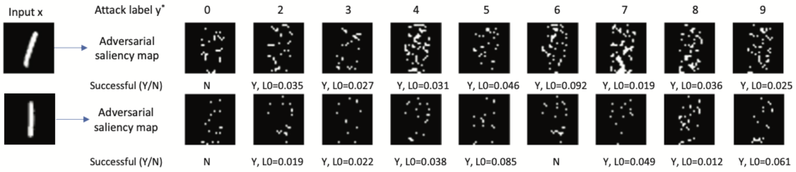

Fig. 4: Illustration of targeted instance-level diversity by visualizing Adversarial Saliency Map-based Noise Injection in JSMA.

Figure 1: Showing the Saliency Map for both instances of 1. Source: The current paper

A simple rotation of 1 made it closer (considering L_0) to perturb towards 5 from the original 6. Furthermore, the authors observed that the perturbed examples of a class are often misclassified into different destination classes during untargeted attacks. In case of a targeted attack on two images of the same class with a small perturbation (or a transformation) separating them, they are often classified as two different classes or need different orders of perturbations (or transformations).

Based on these observations, the authors inferred that transforming the input multiple times individually, using a team of transformers, would deter the adversary, as it needs to trick at least half of them as they apply different transformations. Based on the above saliency example, we see that adversary no longer can force the prediction towards a specific class due to high instance-level divergence among predictions of these transformed inputs. In the coordinate space, each transformation could move the input coordinate in a different direction, therefore making a planned attack less likely as the model could give different predictions for each resulting point (Fig 2). The team of transformers arrive at independent decisions using the model before ensembling them into a final decision - either through consensus or using confidence metrics. The teaming is logical - consider good transformers that are proven to mitigate adversarial noises of various kinds, without affecting the classification of benign examples. In cases when the adversary manages to trick the transformers team, the authors mention that selecting random sets of transformers would make number of teamings exponentially large, thereby avoiding any attacks in reasonable time. But this comes with a tradeoff with the performance. Overall, across many attacks, the authors observed more than 70% defense accuracy, often exceeding 85%, thus making it a competitive one among its class.

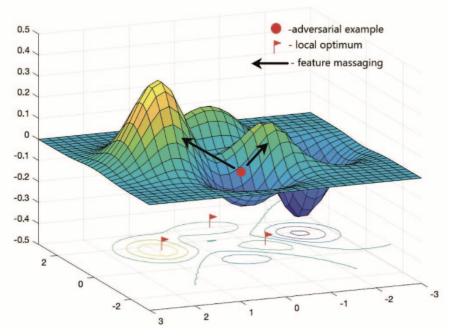

Fig. 7: Illustration of multiple input transformation teaming defense. Different input transformation methods could push the adversarial example to different sub-optimums in the decision space.

Figure 2: Showing the transformations spatially

Although this paper presents an elegant way to mitigate the adversarial effects, it does not explore if the technique *SITD* could be used alongside other defense mechanisms, perhaps as a first step. Also, despite a good selection of input transformations (smoothing, rotation, depth-reduction), the authors could drive the point further home by including more complex transformations. Finally, when comparing to AdvTraining which outperforms *SITD* in some scenarios by taking longer time to train, the authors could expand upon *SITD*'s exponential selection possibilities and show the tradeoff between runtime vs mitigation accuracy more explicitly.