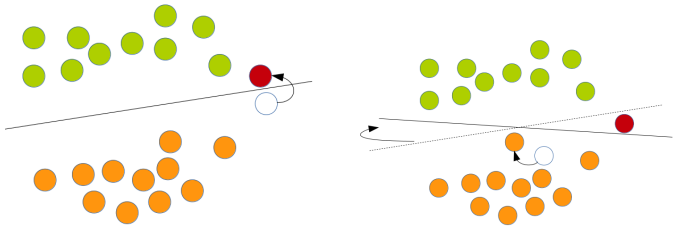
Adversarial poisoning against deep networks

Adrien CHAN-HON-TONG

Despite adversarial example is now a well known phenomenons [3], data poisoning has received much less attention. In data poisoning, the hacker does not modify training image but testing image (see figure 1).



classical adversarial attack data poisoning

Figure 1: Illustrations of classical adversarial attack and poisoning attack: in both case the red point is a testing sample that should be classifier as orange, and, the goal of attacks is to have it classified as green either by modifying testing or training data.

Pioneer works like [2] has offered poisoning attack against deep network. But these attack has focused on large data modification (including label modification). Recently, [1] offers an invisible poisoning based on adversarial example phenomenon. At the end of the attack, nobody can detect it even with a careful review of the training data. Yet, [1] only tackles deep feature + SVM.
In this short paper, we show that [1] attack can be extended to deep network. Two extensions are offered: based on GAN like method or based on better energetic landscape perturbation than in [1] (extended paper in hal.archives-ouvertes.fr/hal-02139074v2).

References

[1] Adrien CHAN-HON-TONG. An algorithm for generating invisible data poisoning using adversarial noise that breaks image classification deep learning. *MAKE*, 2018.

[2] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Artificial Intelligence and Security*, 2017.

[3] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, 2016.