Meng et al. MagNet: a Two-Pronged Defense against Adversarial Examples. (CCS 2017)

1. Summary

This paper proposed to use separate networks from the target classifier models to defend adversarial attacks. MagNet, the proposed system, is consisted of one or more detectors and a reformer. The detectors are trained to distinguish adversarial examples and normal examples, and the reformer is trained to normalize adversarial examples. MagNet does not modify the target networks while keeping the model safe from adversarial attacks.

2. Pros

- MagNet does not modify the target classifier so there’s no trade-off between perecision and adversarial robustness.
- MagNet is able to defend Carlini's attack well, which was SOTA at that time.
- The paper defined and evaluated the gray-box and black-box settings well.

3. Cons

- Separate models require additional resources to execute, for example, the detectors have to process every input prior to reformer or the target classifier.
- The paper evaluated accuracy of detectors but didn't explain false positives of detectors and how the reformer or overall system handle them.

4. Meaning of the paper

- This paper proposed detectors and reformer to prevent adversarial attacks which works independently from the target classifier

5. Discussion and Questions

- In Generative Adversarial Neworks(GAN), the generator always deceives the discriminator and the accuracy of the discriminator decreases as the training progresses. Given this, it is questionable whether training independent detectors can continue to defend adversarial examples.