# 基础知识

## Attack

在攻击中主要包括

## Defense

### Adversarial detection

目标就是为了找出normal data和adversarial examples之间的区别

### Adversarial training

尽可能的覆盖adversarial sample space

### Enhancing model complexity

尽可能的隐藏model parameters，通过增加model，使得难以attack

## 其他

Gray-box是指除了model的parameter未知，其他的内容例如feature、training set等都是已知的。

# 想法

## Attack

### Adversarial examples based on manifold features

针对CCS2017的MagNet中提出的关于normal example和adversarial samples的manifold feature不同，构建了一个自动编码器来学习normal examples来区分normal and adversarial samples。

学习normal examples的manifold feature，按照normal examples的manifold feature去构造生成adversarial samples，具体的做法：通过设置initial parameters来生成多个manifold features，将多个manifold feature进行训练得出general 的feature，然后根据这个feature来生成adversarial samples

假设条件：model is unknown、training set is known、feature is also known

测试的对象设置为image，由于training set is easy to retrieve，在image生成的时候直接就符合manifold feature，而不是通过增加perturbation来实现。

idea：attack the new build model with the insight that adversarial sample follows the manifold feature of normal samples

### Adversarial examples based on reinforcement learning

针对CCS2017的EvadeHC中提出的general attack方法，利用malware构建一个attack方法。在black-box中，malware的feature是binary的形式，根据malware的domain knowledge有针对性的启发式修改生成adversarial samples，前提是保证其main functionality unchanged。在启发过程中尝试使用reinforcement learning进行不断的试探，在DNN的输出层存在softmax function，根据softmax function确定进展情况，同时可以根据normal samples的manifold feature。

在EvadeHC中benign和malicious的界限是否是非常的清晰和明显，implicit assumption is not realistic in real scenario.

测试的对象为malware，其中feature is easy to get

论文启发：CCS2017的论文EvadeHC

## Defense

### Label random nullification using GAN

根据feature random nullification的思想，在GAN训练的过程中将label进行random处理，能否提高model defense的能力。在image中label进行随机处理，例如label 1，在训练的时候随机的划分为其他的label等内容。GAN

在malware中训练一个opposite DNN model，benign labeled as malicious，malicious labeled as benign，通过判断两个model的输出来判断是否是一个adversarial samples。生成对抗样本的目标就是为了将malware输出为benign处理。

理论上证明难以找到比较折中的交叉点。

论文启发：KDD2017

### Feature

### 其他

1. 通过训练model去猜测real model的参数等，例如使用GAN等方法。
2. 在构建adversarial samples一定是需要利用gradient变化最快的方向吗？
3. 在adversarial training中如何利用fuzzing的思想，生成大量的adversarial samples
4. 在adversarial sample生成的时候使用random feature，刻意的保留malicious 内容，不易被发现检测

# 基础知识储备

学习关于DNN模型的训练，manifold的知识、manifold feature的内容、auto encoder的内容

下一步：做好文献综述的工作，目前常见的attack and defense的方法有哪些？