1. Adversarial training：在模型训练的时候加入大量的adversarial samples去训练模型
2. Modifying the model：比如说隐藏模型梯度变化最快的方向
3. Detecting the adversarial samples：Feature squeezing
4. Feature set的处理方式：feature的随机取消使得模型变得non-determistic
5. 利用GAN对抗生成网络生成Adversarial samples
6. 思路：利用GAN生成对抗样本作为training set来构建Robust model
7. 主要包括两块内容，分别是G+D
8. 假设条件分别是：training set的feature和训练模型的参数
9. idea：
   1. G（优化目标）🡪D：feature已知，model已知，目的在于生成更好的Adversarial sample作为training set来构建model

做的过程中如何去扩大Adversarial sample的space，利用feature 消解方法，并且假设model的参数已知

* 1. G🡪D（优化目标）：feature已知，model已知

研究问题：

1. 如何在模型参数、feature等不可知的情况下，设计生成adversarial sample

在feature和模型参数不可知的情况下，但是app的malicious behaviors是已知的，因此使用这些malicious behaviors作为feature修改生成adversarial samples，然后构建一个substitute的DNN，猜测网络的参数和结构，通过比较substitute mode和real model的差异实现feature的猜测以及针对性的生成对抗样本

1. 如何去猜测gradient decrease的方向
2. 一定需要保证feature在gradient decrease的方向吗？修改其他的是否也是可以的？
3. 检测adversarial samples，抓住的点在于adversarial sample通常会修改feature导致在模型变化的gradient masking最关键的位置测试。假设的条件是：feature和model 的参数已知

在malware的adversarial sample detection中很容易做其他的假设，比如malware保持functionality不变等内容

1. 如何对feature进行适当的处理，在模型训练的过程中，使得model变得更加的complexity
2. 针对malware的检测中，多是假设malware的feature都是binary的形式，如何去突破这种形式，修改生成其他的内容
3. Malware adversarial samples 的特征：保持其main functionality不变，adversarial sample生成的部分是少数的，so detection is easy compare with image detection problems
4. 在adversarial malware sample generation中，利用softmax function生成reward function
5. 做graybox的attack，assumption: parameter is unknown, but the feature can be acquired and the output result of softmax layer
6. 做的工作包括：attack & defense，detection
7. Defense：利用生成的adversarial sample去做defense
8. Adversarial training
9. Adversarial samples生成的法则
10. 在Adversarial training中利用fuzzing测试的想法，生成大量高质量的Adversarial samples，如何去扩大sample的space is a problem should be tackled
11. 通过训练normal examples to craft adversarial samples
12. Adversarial sample生成的时候也使用random feature，不遵守相同的adversarial samples generation distribution
13. 基于梯度下降最快的防御方法，adversarial sample 生成的方法应该如何修改去evasion？
14. 在研究所谓的adversarial samples 生成的方法时关注，adversarial sample detection的方法有哪些？defense的方法？
15. 针对label目前尚未有工作进行相关的考虑
16. 在malware detection上面，分析normal example 的manifold feature，然后在保证malware的main functionality不变的情况下，通过在adversarial sample中添加perturbation，这些的perturbation符合normal samples的manifold feature。
17. 针对图片等，生成的时候符合manifold feature的，而不是添加perturbation等，将GAN添加进去
18. 在增加perturbation时，针对每个类别是否都有一个perturbation
19. 在Defense中一个科可行的思路：找出normal samples 和 adversarial samples之间能够区分的feature. Differentiate features
20. 在malware adversarial samples生成中必须要保证的是main functionality
21. Defense的三种思路：
    1. Enhancing model complexity
    2. Adversarial samples detection：including feature
    3. Adversarial training
22. Attack 针对defense而言