Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

基于GAN的网络流量对抗样本生成技术

张荣倩

导师：张笈

2021年1月10日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1.了解入侵检测面临的安全问题
  - 2.了解基于GAN的网络流量对抗样本生成方法
  - 3.了解生成对抗网络在网络安全领域内的应用

背景介绍

- 入侵
  - 由内部或外部入侵者执行的一系列恶意行为，旨在破坏目标系统
- 入侵检测
  - 监视计算机系统和网络流量，并分析活动以检测针对系统的可能入侵

背景介绍

- 入侵检测系统(IDS)
  - 用于检测以系统为目标的内部入侵和外部入侵，识别潜在入侵和可疑活动异常
  - 检测方法
    - 误用检测：预定义恶意活动检测恶意行为（仅适用于检测已知攻击）
    - 异常检测：定义正常模式并根据与正常模式的偏差识别恶意活动
    - 混合技术：混合利用误用检测和异常检测

- 传统的入侵检测系统(IDS)使用误用检测和异常检测技术检测异常行为
- 随着机器学习广泛应用于入侵检测，IDS的检测速度和准确性得到进一步提升

背景介绍

- 安全问题
  - 基于机器学习的检测技术存在安全隐患，对抗攻击严重威胁计算机系统和网络的安全
  - 基于网络的入侵检测系统缺少可靠的训练数据集

- 解决方案
  - 构造强鲁棒性的入侵检测系统
  - 生成可靠的网络流量数据，用于训练提升IDS的鲁棒性
  - 生成对抗IDS的恶意网络流量数据，用于评估IDS的性能

基本概念

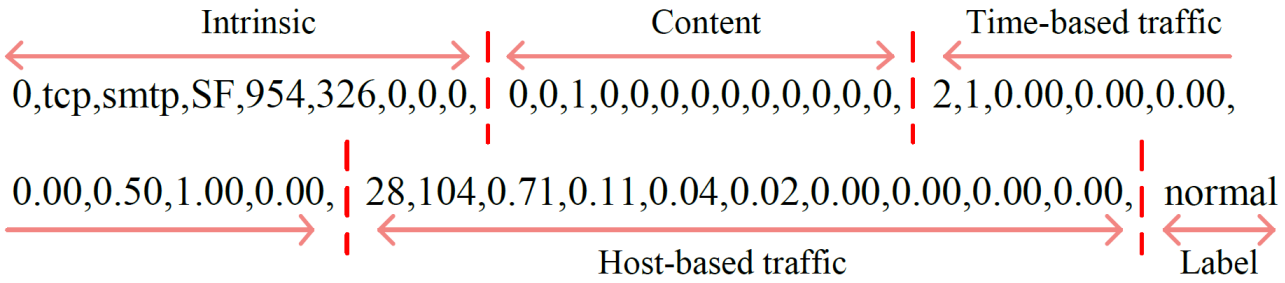
基本概念

- 网络流量数据
  - 分类
    - 基于包（包含完整的有效负载的信息）
      - 格式：pcap（数据报存储格式）
      - 数据集：CICIDS2017
    - 基于流（只包含网络连接的元数据）
      - 五元组：源IP地址、源端口、目标IP地址、目标端口、传输协议
      - 单向流、双向流
      - 数据集：CIDDS-001
    - 混合数据
      - 既不是纯基于包也不是纯基于流的数据集
      - 数据集：KDD CUP 1999

#	Attribute
1	Date first seen
2	Duration
3	Transport protocol
4	Source IP address
5	Source port
6	Destination IP address
7	Destination port
8	Number of transmitted bytes
9	Number of transmitted packets
10	TCP flags

基本概念

- 网络流量数据
  - 来源
    - 在真实或模拟环境生成
      - 手动标记实际网络流量困难且耗时
      - 隐私问题阻碍了实际网络流量的共享
    - 使用现有基准数据集
      - 不支持实时处理
      - 现有数据集较为老旧
      - 数据集
        - » KDDCUP99
        - » NSL-KDD



基本概念

- 生成式对抗网络（GAN）
  - 通过对抗训练的方式来使得生成网络产生的样本服从真实数据分布
    a. 判别网络(D) :尽量准确地判断一个样本是来自于真实数据还是生成网络产生
    b. 生成网络(G) :尽量生成判别网络无法区分来源的样本
    c. 判别网络和生成网络不断交替训练直至收敛

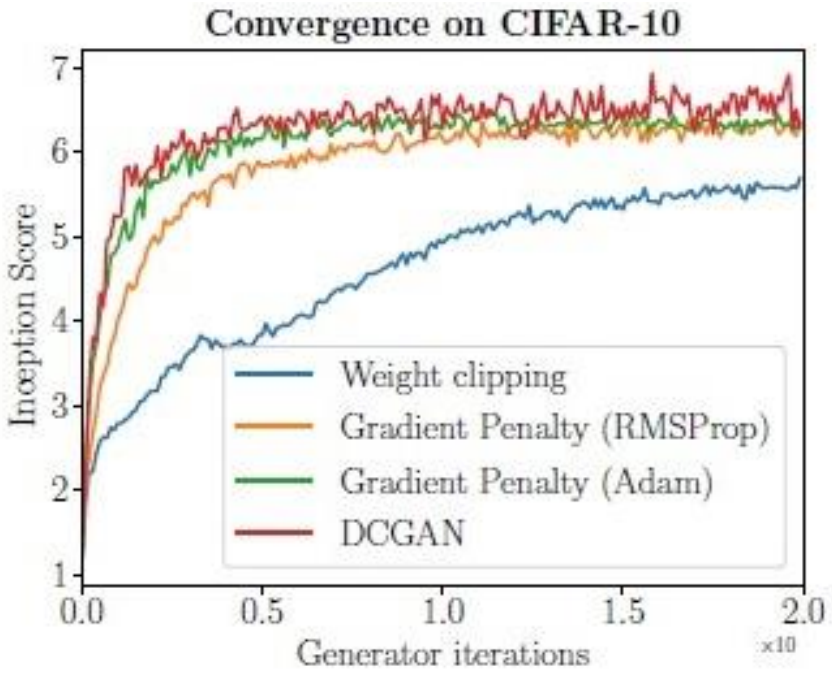
基本概念

- 生成式对抗网络（GAN）
  - 缺陷
    - 难训练，不稳定
    - Mode collapse
  - 变体
    - DCGAN
    - CGAN
    - WGAN
    - WGAN-GP
    - PROGAN
    - SAGAN
    - BIGGAN

基本概念

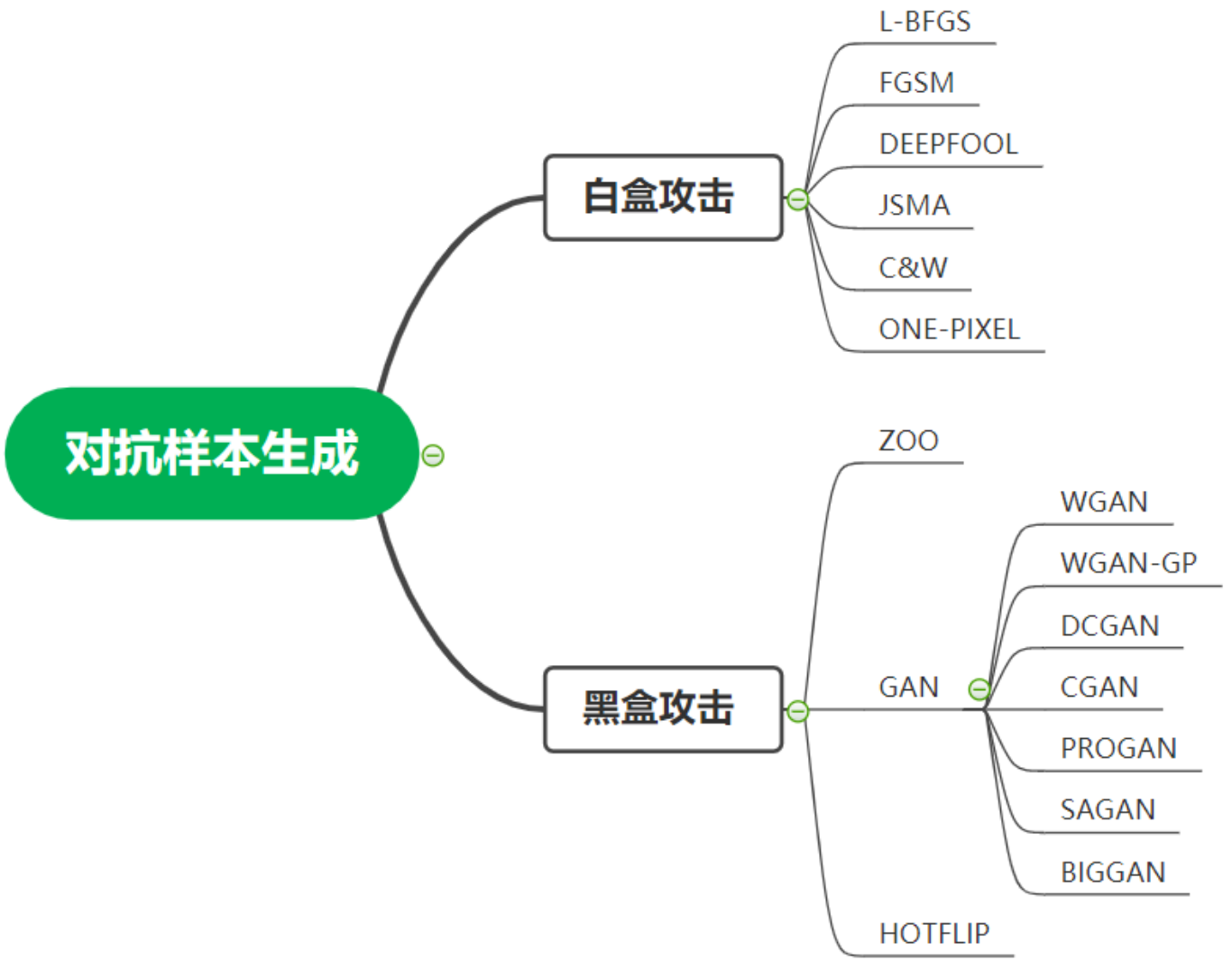
- WGAN-GP
  - WGAN遗留问题
    - WGAN采用权重裁剪(weight clipping)策略，即权重w不超过范围[-c,c]，强行满足鉴别器的Lipschitz连续条件，较难收敛
  - 改进：提出梯度惩罚(gradient penalty)策略

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}\left[D(\tilde{\boldsymbol{x}})\right] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}\left[D(\boldsymbol{x})\right]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}\left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right]}_{\text{Our gradient penalty}}.$$



基本概念

- 对抗样本生成方法



算法原理

算法原理

- 基于GAN的恶意网络流量对抗样本生成方法

T	伪装DOS攻击流量为正常网络流量，逃逸NIDS的检测
I	DOS攻击网络流量
P	1.根据恶意功能标记“已更改”和“未更改”特征 2.输入正常流量样本、恶意流量样本和噪声，根据标记特征训练WGAN和WGAN-GP模型 3.从转换器C获取恶意流量样本
O	能欺骗NIDS的DOS恶意网络流量样本
P	在保留DOS流量攻击能力的前提下，生成能够成功欺骗NIDS的恶意网络流量样本
C	攻击者能够操纵被目标分类器分类为恶意的恶意样本，但不了解NIDS检测器的详细结构
D	在转化器C中结合对抗扰动G(z)和恶意流量X生成对抗样本
L	IJMLC2019（2区）

算法原理

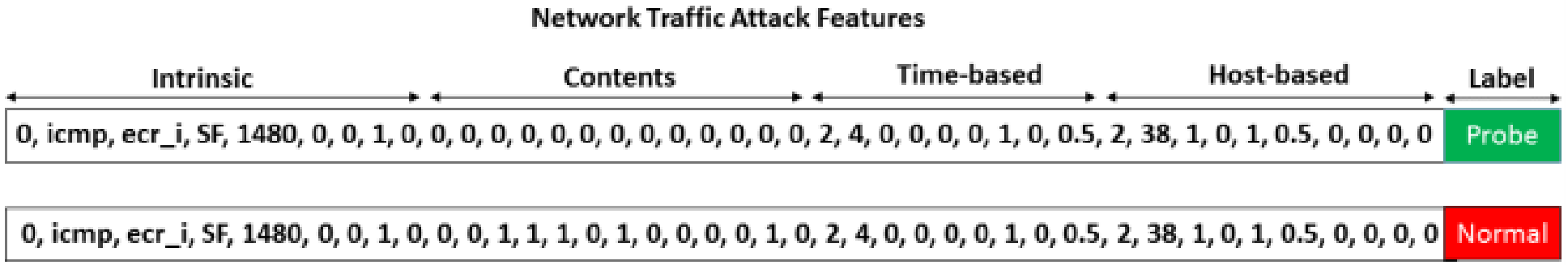
- 总体思想
  - 基于全部特征训练生成的恶意流量对抗样本不具备攻击功能，通过标记“未修改”和“已修改”特征在转换器中生成具备攻击能力的对抗样本
- 实现方案
  - 标记无法修改的特征为“未更改”，标记可以修改的特征为“已修改”
  - 在生成器G中利用噪声生成41维的对抗扰动G(z)
  - 在转换器C中将原始DOS攻击样本中标记为“未更改”的特征与G(z)中标记为“已修改”的特征相结合，生成恶意网络流量样本C(G(z),X)
  - 在鉴别器D中对C(G(z),X)和正常网络流量样本分类
  - 训练模型WGAN和WGAN-GP模型，生成具备攻击能力的恶意流量对抗样本

算法原理 数据集

- NSL-KDD
  - 从模拟的美国空军局域网上采集的9个星期的网络连接数据
  - 基于KDDCUP99改进
    - 不包含冗余和重复的记录，分类器结果更为准确
    - 训练集和测试集的记录数量合理
  - 异常类型
    - Probe：端口监视或扫描
    - R2L(Remote-to-Local)：远程到本地的入侵过程
    - U2R(User-to-Root)：本地提权，未授权的本地超级用户特权访问
    - DOS：拒绝服务攻击，短时间内发送大量无效数据包，以消耗资源

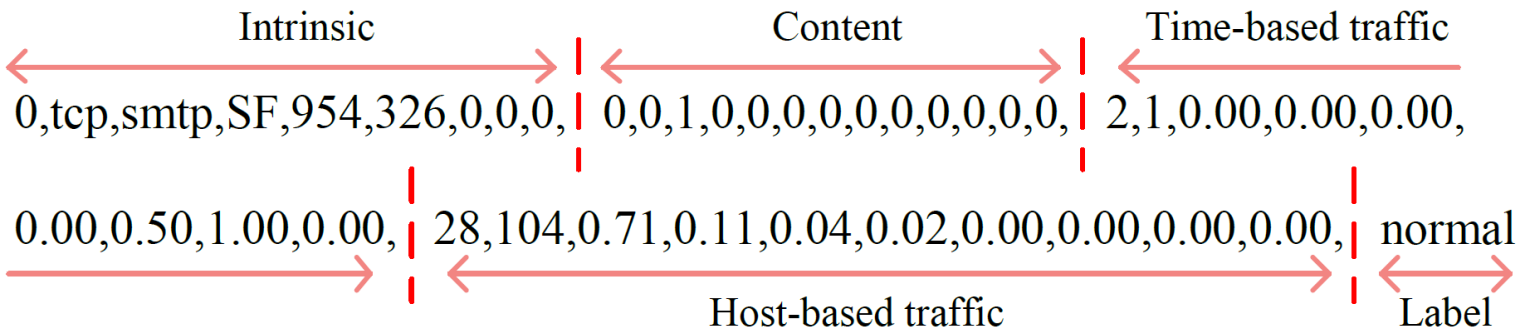
算法原理 数据集

- NSL-KDD
  - 特征类型（共41维）
    - Intrinsic：反映常规网络分析中连接的固有特征。” duration”” protocol_type”
    - Content：标记连接的内容。“num_failed_logins”” num_root”
    - Time-based traffic features：检查在过去的两秒钟内，与当前连接具有相同目标主机或相同服务的连接。“count”” srv_count”
    - Host-based traffic features：检查在过去100个连接中，与当前连接具有相同目标主机或相同服务的连接。” dst_host_count” “dst_host_srv_count”



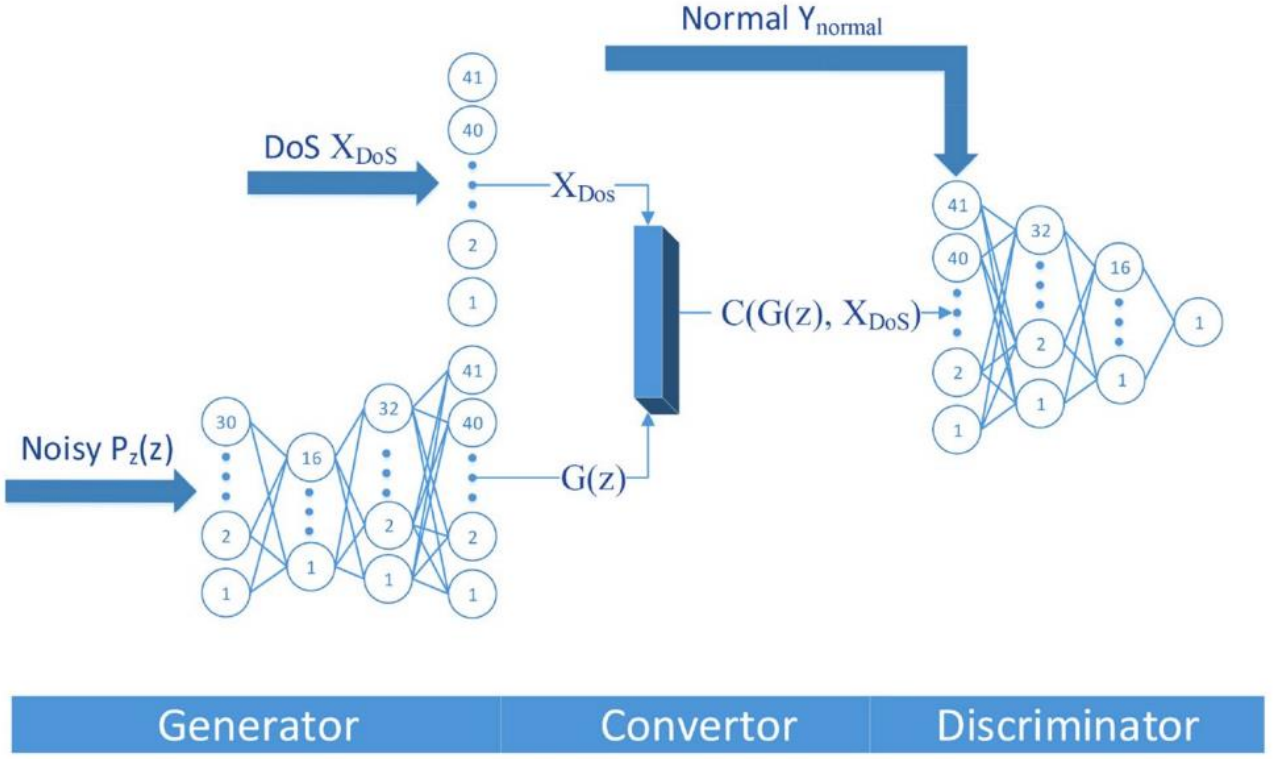
算法原理 特征标记

- DOS攻击流量数据特征标记
  - “未更改”
    - Intrinsic特征集：“duration”，“protocol_type”，“service”，“flag”
    - Time-based traffic特征集：在时间上具有很强的相关性
  - “已更改”
    - Intrinsic特征集：其他特征
    - Content特征集
    - Host-based traffic特征集



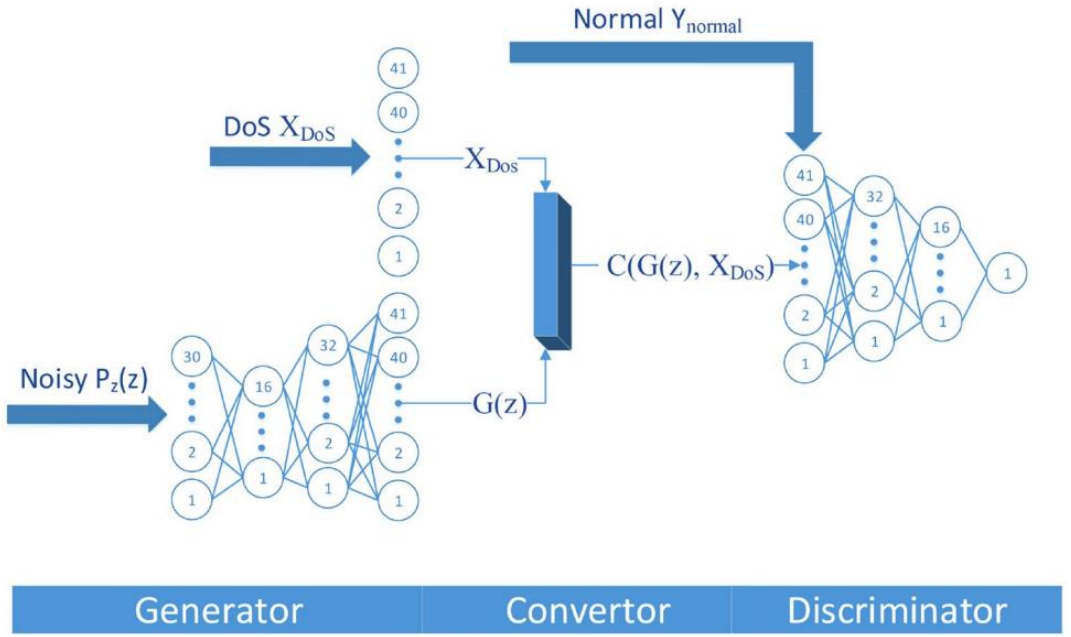
算法原理 WGAN

- WGAN
  - 生成器G和转换器C尽可能生成愚弄鉴别器D的恶意流量对抗样本
  - 鉴别器D接收伪造的网络流量和正常网络流量，并尽可能区分两者
  - 生成器G
    - 目标：生成对抗扰动
    - 输入：均匀分布的[0,1)噪声
    - 输出：41维的对抗扰动 G(z)



算法原理 WGAN

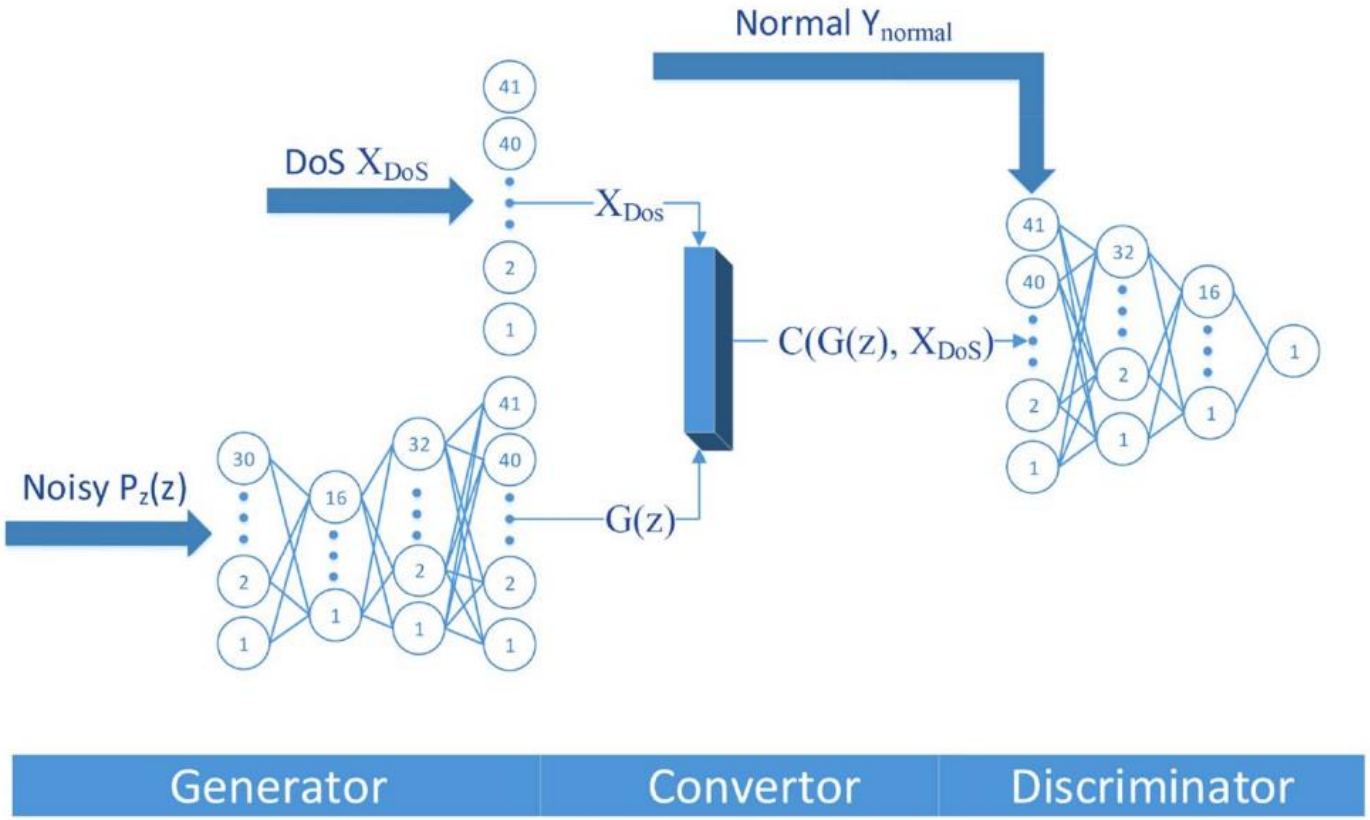
- WGAN
  - 转换器C
    - 目标：生成愚弄鉴别器D的恶意流量对抗样本
    - 输入
      - 对抗扰动 G(z)
      - DOS攻击流量样本 X
    - 操作
      - 特征标记
      - 将X中标记为 “未更改”的特征与G(z)中标记为“已修改”的特征相结合
    - 输出：DOS攻击网络流量对抗样本 C(G(z),X)



算法原理 WGAN

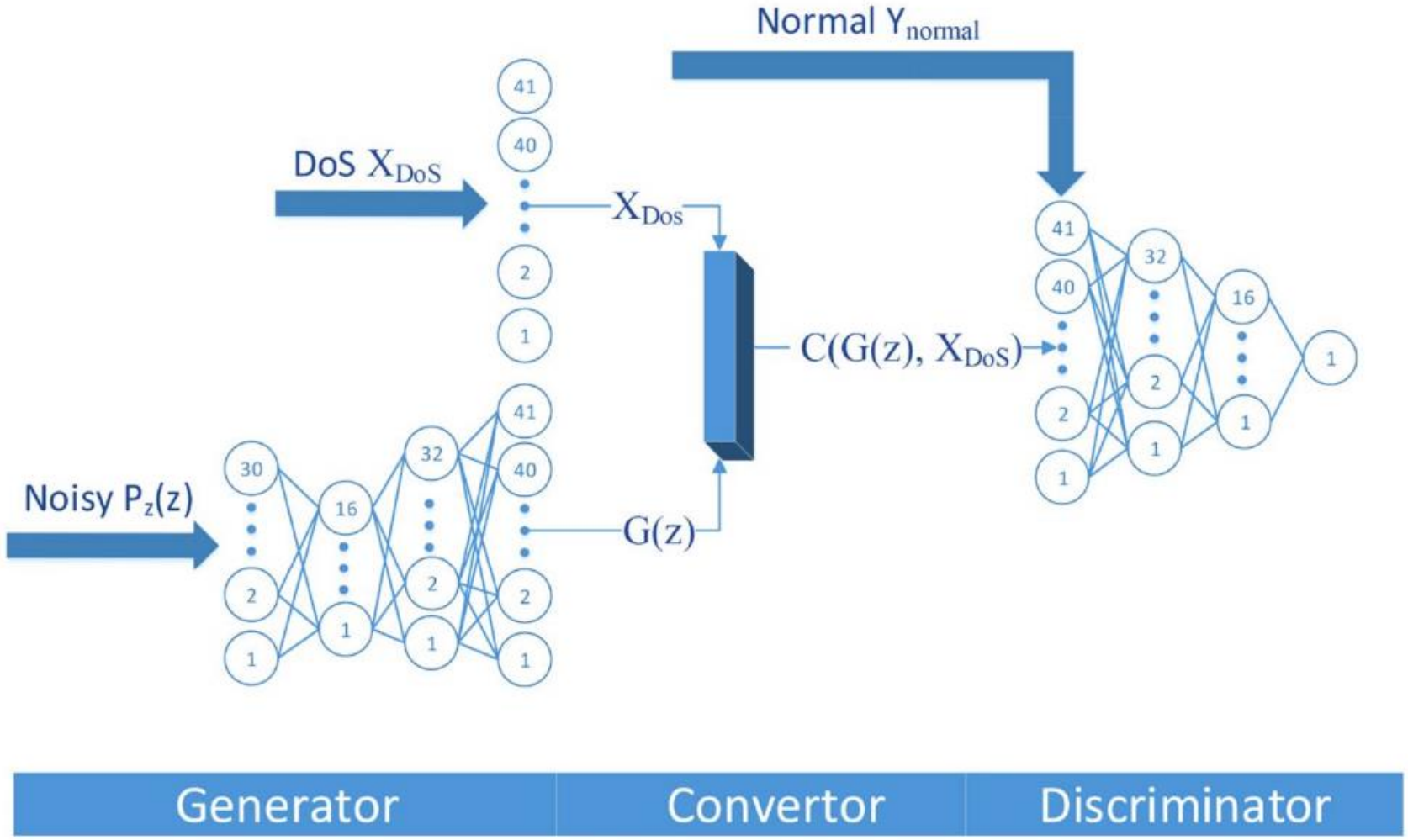
- WGAN
  - 鉴别器D
    - 目标：尝试分辨出正常数据和对抗样本
    - 输入
      - 恶意流量对抗样本 C(G(z),X)
      - 正常流量样本
    - 输出：对输入样本的预测标签
    - 模型：MLP
  - 损失函数：

$$L = \mathbb{E}_{x_n \sim \mathbb{P}_{normal}(x)}[f_\omega(x_n)]$$
$$- \mathbb{E}_{z \sim P_z(z), x_D \sim \mathbb{P}_{DoS}(x)}[f_\omega(C(g_\theta(z), x_D)]$$
$$- \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\| - 1)^2]$$



算法原理

- 算法原理图


Normal Y_{normal}
DoS X_{DoS}
X_{Dos}
$C(G(z), X_{DoS})$
Noisy $P_z(z)$
$G(z)$
41
40
2
1
30
16
32
Generator
Convertor
Discriminator

算法原理 评估方法

- 评估方法
  - 欧几里得距离(Euclidean distance)
    - 度量不同数据的相似性
    - 改进：标准化欧几里得距离(Standardized Euclidean distance)
      - 公式：$$d = \sqrt{\sum_{k=1}^{41}(\frac{x_{1k} - x_{2k}}{S_k})^2}.$$
  - 信息熵(Information entropy)
    - 衡量无序程度或分布分散程度
    - 信息的熵越大，表示的数据类型就越多样化
    - 一段时间内信息熵变化小表明训练稳定
    - 公式：$$H(X) = -\sum_{x\sim\chi} p(x)\log p(x)$$

算法原理 评估方法

- 评估方法
  - 性能评估
    - 对抗样本同原始样本的相似性
    - 对抗样本的多样性
    - 训练过程的稳定性
  - 具体操作
    - 计算网络流量与对抗样本之间的标准化欧几里得距离，具有最小标准欧几里德距离的网络流量是对抗性样本到原始数据集分布的映射
    - 计算对抗样本的分布的信息熵，信息熵越大对抗样本越多样化
    - 观察GAN训练期间信息熵的变化，如果信息熵急剧变化，则训练不稳定

实验步骤

- 实验步骤
  - 预处理
    - 字符串特征转化为数字特征
      - “protocol”特征：用1,2,3代替tcp,udp,icmp
    - 数据归一化
      - 目的：加快学习速度，收敛加快
      - 方法：
        - » 最小-最大归一化 $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
        - » z分数归一化
        - » 十进制缩放归一化
  - WGAN/WGAN-GP模型训练，生成恶意流量对抗样本
  - 训练黑盒基于CNN的NIDS检测器
  - 评估恶意流量对抗样本性能

实验设置

- 实验设置
  - 训练集
    - GAN、WGAN-GLIP、WGAN-GP和NIDS使用相同训练集
    - GAN、WGAN-GLIP、WGAN-GP使用相同训练集
  - 攻击模型
    - GAN(DoS-WGAN)
    - WGAN-GLIP(具有削波功能的DoS-WGAN)
    - WGAN-GP(具有梯度损失的DoS-WGAN)
  - 被攻击对象：基于CNN的NIDS
  - 数据集：KDD99
  - 评价指标：真实阳性率（TPR）

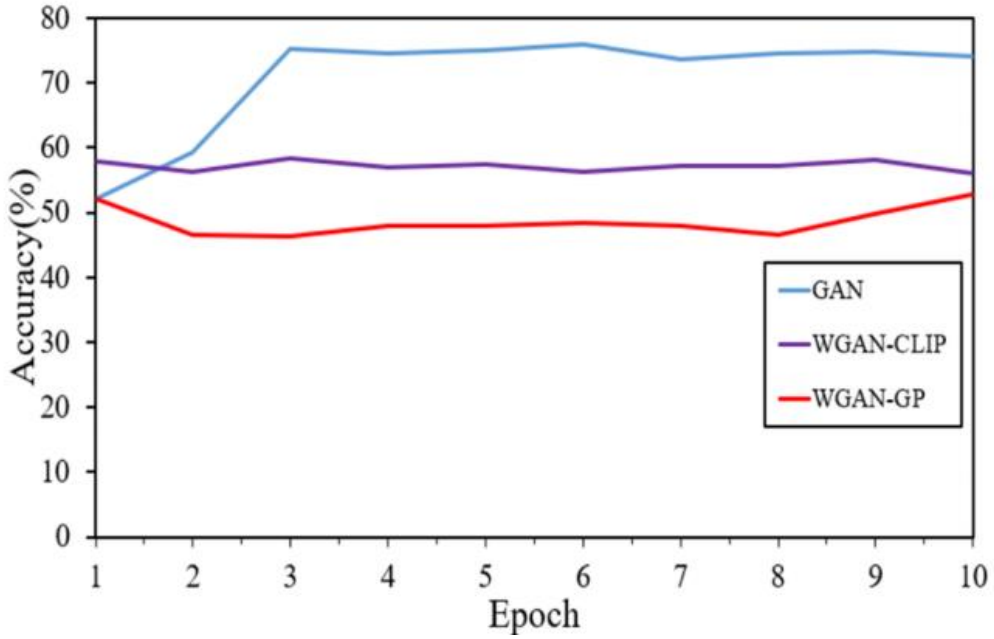
实验结果

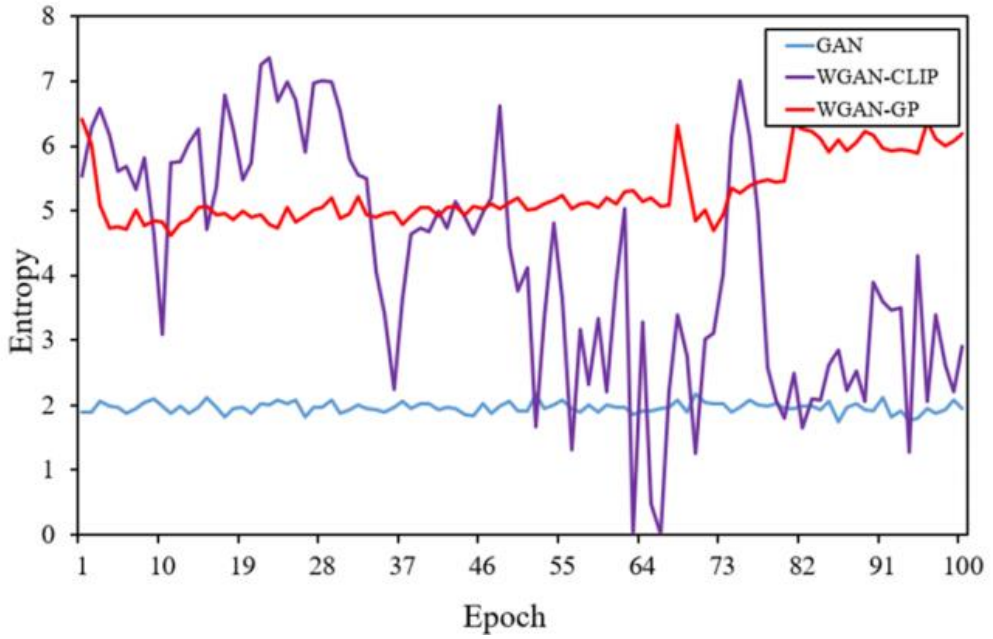
- 实验结果
  - TPR
    - WGAN-GP模型生成的对抗样本具有更好的性能，可以将TPR从97.3461降低到47.627%
    - 对于这三个模型，对抗性示例的TPR均低于原始样本
    - WGAN-GP模型是KDDCup99数据集上最好的模型

Models	Accuracy
Models without GAN	97.3461%
Models with GANs	
GAN	$69.722 \pm 1.0848\%$
WGAN-CLIP	$57.153 \pm 0.5615\%$
WGAN-GP	$\mathbf{47.627 \pm 1.1903\%}$

实验结果

- 实验结果
  - 信息熵
    - GAN模型的熵低于其他模型，这表明GAN模型可能会陷入崩溃模式
    - WGAN-CLIP模型的曲线剧烈抖动且不平滑
    - WGAN-GP模型的曲线更平滑，WGAN-GP模型的熵比其他模型高，是最稳定的





结果分析

- 结果分析
  - 针对基于CNN的网络入侵检测系统，WGAN-GP模型生成的对抗样本比WGAN-CLIP和GAN的对抗样本具备更好的逃逸性能，证实了GAN模型容易产生一些重复且惩罚低的样本
  - WGAN-GP模型的训练过程比WGAN-CLIP的训练过程更稳定，证实了梯度惩罚能很好的解决WGAN难以收敛，训练不稳定的问题

优劣分析

- DOS-WGAN
  - 优点：
    - 生成了保留DOS攻击能力的恶意流量数据，并能降低基于CNN的NIDS的检测率到47%
  - 缺点：
    - 每次训练只能生成DOS攻击类型的对抗恶意流量数据
    - 尽管大幅度降低了NIDS的TPR，逃逸率仍较低

应用总结

应用总结

- 算法应用领域
  - 恶意软件检测
  - 网络入侵检测
  - 恶意软件对抗攻击
  - 恶意流量对抗攻击

- 发展方向
  - 提高恶意流量数据的逃逸率
  - 提升入侵检测系统的鲁棒性
  - 提升网络流量数据的多样性和准确率

参考文献

[1] YAN Q, WANG M, HUANG W, et al. Automatically synthesizing DoS attack traces using generative adversarial networks[J]. International Journal of Machine Learning and Cybernetics,2019,10(12):3387-3396.

[2] RING M, SCHL O R D, LANDES D, et al. Flow-based network traffic generation using generative adversarial networks[J]. Computers \& Security,2019,82:156-172.

道德经

上善若水。水善利万物而不争，处众人之所恶，故几於道。居善地，心善渊与善仁，言善信，正善治，事善能，动善时。夫唯不争，故无尤。