基于生成对抗网络（GAN）的恶意软件变体生成与检测规避技术研究

郭鑫

浙江公路技师学院 交通服务学院，浙江省杭州市，310023；

摘要：本文是围绕着基于生成对抗网络的恶意软件变体生成与检测规避技术进行研究。文章阐述了相关的研究意义，具体指出，网络安全形势下的传输检测方法具有的局限性。开始是介绍了生成对抗网络的基础理论，包括原理架构、训练过程和变体应用。深入的剖析了基于生成对抗网络生成恶意软件变体的原理以及利用这个原理能够规避检测的机制。分析了检测规避技术面临的 GAN 训练稳定性变体质量评估和检测模型进化等问题，最终提出了改进 GAN 的训练、建立评估的体系和强化检测模型的应对策略，以此为网络的安全研究提供理论支撑。

关键词：生成对抗网络；恶意软件变体；检测规避技术

DOI：10.69979/3041-0673.26.01.006

在当今的数字化发展时代，网络安全问题越来越严峻。恶意的软件成为了主要的威胁，随着恶意软件与日俱增的数量层出不穷，传统的恶意软件检测方法虽然在一定期间能够发挥作用，但随着恶意软件技术的不断演进，出现了一些局限性。这些方法只依赖于已知的恶意软件特征库，对于新出现的具有未知特征的恶意软件变体检测效果欠佳。而生成对抗网络作为一种强大的深度学习模型，具有生成逼真数据的能力，能够将其应用到恶意软件分析领域中，为恶意软件的检测带来新的思路，也能够生成恶意软件的变体来规避检测。因此，深入研究基于GAN的恶意软件变体生成和检测规避技术最有重要的现实意义。

1 生成对抗网络（GAN）基础理论

1.1 GAN 的原理与架构

生成对抗网络作为深度学习领域的创新成果，为数据的生成和分析提供了强大的工具。在恶意软件研究等领域上展现出了独特的应用价值。生成对抗网络的核心思想是来源于博弈论中的零和博弈，是由生成器和判别器两个神经网络构成的对抗架构。生成器主要是负责接收随机的噪声作为输入，通过多层非线性的变换，能够生成最真实的数据和相似的样本，生成器的目标是要欺骗判别器，使生成的样本被误认为真实的数据；判别器要接收真实的数据和生成器生成的假数据，然后输入一个概率值，表示输入的样本是真实数据的概率，判别器的目标就是要区分真实和生成样本，这两者在对抗训练中通过不断优化生成器，要提升生成样本的真实性，判别器要提高鉴别能力，最终才能够达到均衡，生成器能生成以假乱真的样本。

1.2 GAN 的训练过程

生成对抗训练采用的是交替迭代的方式进行的，在每次迭代过程中，要固定好生成器的数据参数，使用真实的数据和生成器。当前应用的假数据训练判别器，判别器也要通过最大化对真实数据进行正确分类和最小化对生成数据的错误分类概率来更新自己的参数，提高区分能力。接着要固定判别器的参数训练生成器，生成器要通过最小化判别器将生成样本识别为假的概率来更新参数，这种方法能够提升生成样本的真实性，这种循环反复的交替训练，能够让生成器生成的样本逐渐接近真实数据分布，判别器也难以准确的分辨最终生成高质量的数据样本，这样能够为后续的恶意软件等领域应用奠定相应的基础。

1.3 GAN 的变体与应用领域

为了解决原始的生成对抗网络训练过程中不稳定的问题，研究人员提出了多种变体。使用卷积层和转置卷积层处理图像数据，能够更好的捕捉图像的特征，生成高质量的图像。WGAN 通过改变损失的函数使用距离，衡量真实的分布和生成分布的差异，改变了训练的稳定性和生成样本的质量，在具体的应用领域中生成对抗网络。除了在图像生成和风格迁移等方面表现比较出色，还能够用于数据增强解决数据稀缺的问题。在语音合成方面也能够生成自然流畅的语音。在恶意软件研究中，能够生成多样性的恶意软件变体用于检测模型的训练和评估，提升检测的能力。

2 基于 GAN 的恶意软件变体生成原理

2.1 恶意软件的特征表示

恶意软件的特征本质就是在数据层面的映射。从代码的结构来看，抽象的语法树能够将代码的语法结构通过树状的形式进行呈现，清晰的展示了函数的调用和变

量定义等关系，方便于捕捉代码的逻辑特征。在行为模式方面，系统的调用了序列记录软件与操作系统交互的操作顺序，比如打开文件和读取数据等，这些能够充分的反映出软件的实际运行行为。而且恶意的软件 API 调用和网络通信模式是重要的特征，通过将这些多维特征进行数码化的编码，能够转换为GAN可处理的向量形式，为后续的生成器生成变体提供了基础性的数据，使生成的变体能够在特征层面模拟最真实的恶意软件。

2.2 生成器生成恶意软件变体的过程

生成器通过随机噪声和恶意软件的特征表示作为输入。随机噪声表现为生成过程引入随机性特点，保证生成变体的多样性。生成器通过多层神经网络，如全连接层、卷积层等，对输入进行非线性变换。在训练过程中，生成器不断学习恶意软件特征的分布规律。它会根据判别器的反馈，逐步调整自身参数，从简单的代码片段生成开始，逐渐组合和优化，生成完整的恶意软件变体代码。生成的变体不仅在功能上模拟真实恶意软件，还在代码结构和行为模式上尽可能贴近，以增加其迷惑性和检测难度。

2.3 判别器对生成变体的评估与反馈

判别器接收真实恶意软件样本和生成器生成的变体样本。它通过内部的神经网络结构，对样本的特征进行提取和分析，输出一个概率值，表示样本为真实恶意软件的概率。若判别器判定生成变体为假的概率较高，说明生成器生成的变体与真实恶意软件存在差异。判别器会将这一评估结果反馈给生成器，指导生成器调整参数。生成器根据反馈，优化生成策略，使生成的变体在特征上更接近真实恶意软件。这种生成器与判别器的对抗博弈，不断推动生成器生成质量更高、更难被检测的恶意软件变体。

3 利用 GAN 生成恶意软件变体以规避检测的原理

3.1 传统恶意软件检测方法的局限性

传统恶意软件检测主要依赖特征码检测和行为检测两种方式。特征码检测基于已知恶意软件的特征代码库，通过比对样本代码与库中特征码来识别恶意软件。然而，随着恶意软件技术的快速发展，攻击者频繁对恶意软件进行混淆、加密等操作，改变其代码特征，使得基于固定特征码的检测方法难以应对新变种。行为检测则是通过监控软件运行时的行为，如系统调用、网络通信等，与预设的恶意行为模式进行匹配。但这种方法需要预先定义全面的恶意行为规则，对于一些具有隐蔽性、采用新攻击手段的恶意软件，其行为模式可能不在预设规则内，导致检测失败。而且，传统方法对未知恶意软件的检测能力较弱，无法及时应对不断涌现的新型威胁。

3.2 GAN 生成变体对检测方法的挑战

GAN 生成的恶意软件变体给传统检测方法带来了巨大挑战。由于 GAN 的生成器具有强大的学习能力，它可以生成大量多样化的恶意软件变体。这些变体在代码结构和行为模式上与已知恶意软件存在差异，使得基于特征码的检测方法难以准确匹配。同时，生成器能够学习到检测模型的特征提取和判断规律，有针对性地生成能绕过检测的变体。例如，生成器可以调整恶意软件的系统调用顺序、网络通信频率等行为特征，使其在行为检测中表现得与正常软件相似。GAN 生成的变体具有高度的随机性和不确定性，增加了检测的难度和复杂性，传统检测方法难以建立有效的模型来识别这些变体。

3.3 检测规避技术的具体实现方式

利用 GAN 实现检测规避有多种具体方式。一种是对恶意软件代码进行混淆处理，生成器通过插入冗余代码、重命名变量和函数等操作，改变代码的外观结构，同时保持其恶意功能不变，使特征码检测无法识别。另一种是模拟正常软件行为，生成器分析正常软件的行为模式，如系统资源使用情况、网络通信模式等，让生成的恶意软件变体在行为上尽可能接近正常软件，从而绕过行为检测。还可以采用对抗样本技术，对生成的恶意软件变体进行微调，使其在检测模型的输入空间中产生微小的扰动，导致检测模型输出错误的判断结果，成功规避检测，提高恶意软件的生存能力。

4 检测规避技术面临的挑战

4.1 GAN 训练的稳定性问题

GAN 训练过程稳定性极差。生成器和判别器相互对抗，若一方过于强大，另一方就难以有效学习。例如，当判别器性能远超生成器时，它会迅速准确判别生成样本，使生成器得不到有效反馈，无法提升生成质量；反之，生成器生成的高质量样本会让判别器难以区分，导致判别器参数更新混乱。而且，GAN 训练对超参数设置极为敏感，学习率、批次大小等微小变化都可能引发训练崩溃，如梯度消失或爆炸，使模型无法收敛到理想状态，难以持续稳定地生成能有效规避检测的恶意软件变体。

4.2 恶意软件变体的质量评估

准确评估恶意软件变体质量是检测规避技术的关键，却也困难重重。传统评估指标如生成样本的多样性、真实性等，难以全面衡量变体对检测的规避能力。一个看似真实多样的变体，可能在关键恶意特征上未做有效改变，仍易被检测模型识别。并且，不同检测模型对变体的敏感度不同，缺乏统一的评估标准。此外，恶意软

件变体的质量还受其功能完整性和隐蔽性影响，若变体在生成过程中破坏了原有恶意功能，或隐蔽性不足，就无法达到规避检测的目的，这使得质量评估变得复杂且具有挑战性。

4.3 对抗检测模型的不断进化

检测模型也在持续进化以应对恶意软件变体的威胁。安全研究人员不断改进检测算法，采用更先进的特征提取方法和机器学习模型，提高检测的准确性和泛化能力。例如，深度学习模型能够自动学习恶意软件的复杂特征，对新型变体有更强的识别能力。同时，检测模型还会结合多源数据进行综合分析，增加恶意软件变体规避检测的难度。面对检测模型的进化，利用 GAN 生成规避检测的变体需要不断调整策略，这要求 GAN 具备更强的适应性和学习能力，否则将难以跟上检测模型更新的步伐，导致检测规避技术失效。

5 应对基于 GAN 的恶意软件检测规避的策略

5.1 改进 GAN 训练方法

为提升 GAN 生成样本的质量与稳定性，可从多方面改进训练方法。可以采用更先进的优化算法，如 AdamW 优化器，它在 Adam 的基础上增加了权重衰减项，能有效防止过拟合，使生成器和判别器在训练过程中更稳定地更新参数，避免梯度消失或爆炸问题，从而生成更具多样性和真实性的恶意软件变体样本，为检测模型提供更丰富的训练数据。另一方面，引入条件 GAN 的思想，将恶意软件的类别、攻击类型等条件信息作为额外输入，指导生成器生成特定类型的恶意软件变体。这样生成的样本更具针对性，能帮助检测模型更好地学习不同类型恶意软件的特征，提高检测的准确性。采用渐进式训练策略，先训练生成器和判别器处理简单的恶意软件特征，随着训练的进行逐渐增加特征的复杂度，使模型能够逐步适应复杂的恶意软件生成任务，提升整体训练效果。

5.2 建立全面的恶意软件变体评估体系

一个全面的评估体系是准确衡量恶意软件变体质量的关键。这个体系应包含多个维度的评估指标，除了传统的生成样本多样性、真实性外，还需要重点关注对检测的规避能力。可以构建一个包含多种主流检测模型的测试环境，将生成的恶意软件变体在这些模型上进行检测，统计其被检测出的概率，以此评估变体的规避效果。同时，考虑变体的功能完整性，确保在生成过程中恶意软件的核心功能不受破坏。通过综合这些指标，对恶意软件变体进行全面、客观的评估，为改进生成策略和检测方法提供有力依据。

5.3 强化检测模型的鲁棒性

为应对不断进化的恶意软件变体，检测模型需具备更强的鲁棒性。可以采用集成学习的方法，将多个不同类型的检测模型进行集成，同时，引入对抗训练技术，在训练检测模型时，使用基于 GAN 生成的恶意软件变体作为对抗样本，让模型在训练过程中不断学习如何识别这些具有规避能力的变体，从而增强模型对新型恶意软件的抵御能力。持续更新检测模型的特征库和知识，及时纳入新发现的恶意软件特征和攻击模式，使检测模型能够跟上恶意软件技术的发展步伐，有效检测基于 GAN 生成的恶意软件变体。

6 结束语

综上所述，基于生成对抗网络的恶意软件变体生成与检测规避技术的研究，能够为网络的安全领域带来新的挑战和机遇。GAN 具有的强大生成能力能够让恶意软件变体生成变得具有多样性和隐蔽化，这个变化给传统的检测方法带来了巨大的冲击。通过对 GAN 的训练方法进行改进，建立全面的恶意软件变体评估体系和强大的检测模型等策略，能够有效的应对这些问题。在未来的发展中，随着技术的不断变化，我们能够持续的关注 GAN 在恶意软件领域中的应用动态，不断的优化和完善检测和防范技术，同时，要加强与国际间的合作和交流，共同应对全球性的网络安全威胁，构建安全、稳定的网络环境。

参考文献

[1]陈元昭,林良勋,王蕊,等.基于生成对抗网络GAN的人工智能临近预报方法研究[J].大气科学学报,2019,42(02):311-320.
[2]杨涛.基于生成对抗网络（GAN）自主海报生成的研究[D].湖北省:长江大学,2020.
[3]王良基.基于生成对抗网络（GAN）的地质统计学反演方法研究[D].四川省:电子科技大学,2021.
[4]郑旭廷,孙三祥,崔善坤.基于生成对抗网络(GANs)的隧道施工通风研究与应用[J].隧道建设(中英文),2025,45(S1):330-339.
[5]姚大斌.基于生成对抗网络(GANs)的绘本自动化创作研究与实现[J].艺术与设计（理论版）,2023,(12):95-97.

作者简介：郭鑫（1979.05—），男，汉，安徽安庆，本科，工程师，主要研究方向为职业院校基础学科在计算机应用教学领域的实践和理论创新，长期从事与职业院校信息化建设领域，在智慧校园建设、专业管理等方向有着丰富的现实案例和管理经验，主编出版《计算机常用工具软件应用》（天津科技出版社）、《计算机网络基础》（同济大学出版社）。