**摘要**

近年来，随着勒索病毒和恶意PE可执行文件木马程序的流行，目前存在的一些反病毒软件的查杀引擎，例如360杀毒的QVM2反病毒引擎，对于恶意PE可执行文件的检测仍然有较多的漏报、误报问题，仍需要用户主动向反病毒软件厂商提交样本漏报和误报的样本。同时，部分反病毒软件过度依赖动态检测来对抗恶意软件，静态检测能力仍然有不足。本文提出了一种基于强化学习的Sarsa算法的静态对抗型样本生成框架，输入已有的PE可执行程序恶意样本来生成能够规避部分反病毒软件静态检测的对抗性样本。本实验的行为表定义了对PE文件执行扰动的操作，这些操作基于Python的lief库和pefile库，样本集基于国内反病毒软件的官方论坛提供的样本以及Virus Share的样本， 某个样本集经处理前后，Clam AV对其查杀率从27.8%下降到了21.9% 360杀毒对其的查杀率从99.3%下降到了66.5%。本实验中强化学习模型生成的对抗性样本能有一定程度的静态检测对抗性，虽然对于火绒杀毒的查杀率下降影响并不显著。总而言之，使用本文中提出的方法生成的对抗样本集对杀毒软件进行训练，可以训练基于深度学习和机器学习的反病毒引擎和反病毒软件的对于恶意软件的检测能力，对恶意PE可执行文件的对抗性样本生成领域有一定的贡献。

**Abstract**

In recent years, with the prevalence of ransomware and malicious PE executable trojans, existing antivirus detection engines — such as 360 Antivirus's QVM2 engine — still exhibit significant false negatives and false positives in detecting malicious PE executables. This necessitates users to proactively submit samples of missed or misclassified threats to antivirus vendors. Furthermore, some antivirus solutions over-rely on dynamic analysis to counter malware, revealing deficiencies in static detection capabilities.

This paper proposes a static adversarial sample generation framework based on the reinforcement learning SARSA algorithm. By inputting existing malicious PE executable samples, the framework generates adversarial variants capable of evading static detection by certain antivirus software. The experiment defines perturbation operations for PE files using Python's *lief* and *pefile* libraries. The dataset combines samples from Chinese antivirus vendors' official forums and Virus Share. After processing, the detection rates declined as follows: Clam AV dropped from 27.8% to 21.9%, while 360 Antivirus fell sharply from 99.3% to 66.5%. Although the adversarial samples showed limited impact on Huo rong Antivirus's detection rate, the reinforcement learning model demonstrated measurable success in evading static analysis.

In conclusion, training antivirus engines (particularly those leveraging deep learning or machine learning) with adversarial samples generated by this method can enhance malware detection capabilities. This work contributes to the domain of adversarial sample generation for malicious PE executables.

**关键词**

静态分析规避，PE恶意可执行文件，强化学习，Sarsa算法

**引言**

恶意软件是经过精心设计，被用来攻击计算机系统或计算机网络并且造成损害的软件，危害不仅限于感染型病毒和蠕虫的自身复制耗尽系统资源，破坏操作系统、后门软件开放系统端口供黑客连接从而形成僵尸网络对服务器发起分布式拒绝服务攻击、木马伪装成正常程序，实则窃取用户的敏感数据和破坏系统，为黑客提供后门、勒索病毒加密文件要求用户支付赎金等。根据统计数据，仅2022年在全球大约发生了55亿次恶意软件攻击事件。

SonicWall

Sonicwall cyber threat report

<https://www.sonicwall.com/medialibrary/en/white-paper/2023-cyber-threat-report.pdf> (2023)

在早期，恶意软件的代码较为简单，容易被反病毒软件检测到特征值从而处理清除。但是多年以来，恶意软件的复杂性不断发展，传统的基于黑名单哈希值的恶意软件检测技术难以应对当今恶意软件复杂的混淆策略，尽管这种方法速度很快，但是难以识别新一代的恶意软件以及一些0day恶意软件。

[1] Gopinath M., Sibi Chakkaravarthy Sethuraman Ph.D., A comprehensive survey on deep learning based malware detection techniques

目前，为了应对恶意软件带来的威胁，许多开源以及商业杀毒软件厂商不断升级病毒库，更替杀毒软件版本。目前针对恶意软件的识别主要分为静态分析、动态分析以及混合分析（同时结合了静态分析和动态分析）这三种。而本实验研究的强化学习模型，是生成针对反病毒软件的静态特征分析的PE可执行程序对抗性样本，目的是用于反病毒软件厂商的机器学习和深度学习反病毒引擎模型训练和逆向分析工作者们学习研究，本实验未研究动态特征分析对抗性样本生成。静态特征分析包括文件Hash（如MD5 SHA256）匹配，这在很多病毒样本分析网站，例如VirusTotal，VirSCAN等被使用，如果上传的文件和病毒库里面存在的样本的Hash值匹配，则反病毒软件判断该文件是病毒。此外，静态分析还包括资源节分析，时间戳检查，数字签名检查，函数导入表检查，特征字符串匹配，DEBUG信息检查等。

在过去的十几年中，学术界出现了大量利用机器学习和强化学习的恶意软件查杀模型判断恶意软件的研究成果，甚至可以运用启发式杀毒规则，检测分析软件的代码行为，来判断从未出现过的新型恶意软件。

[5] Hyrum S Anderson and Phil Roth. 2018. Ember: an open dataset for training static PE malware machine learning models. arXiv preprint arXiv:1804.04637 (2018).

[15] George E Dahl, Jack W Stokes, Li Deng, and Dong Yu. 2013. Large-scale malware classification using random projections and neural networks. In 2013 IEEE Inter national Conference on Acoustics, Speech and Signal Processing. IEEE, 3422–3426.

[55] Edward Raff, Jon Barker, Jared Sylvester, Robert Brandon, Bryan Catanzaro, and Charles K Nicholas. 2018. Malware detection by eating a whole exe. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

[56] Konrad Rieck, Philipp Trinius, Carsten Willems, and Thorsten Holz. 2011. Auto matic analysis of malware behavior using machine learning. Journal of Computer Security 19, 4 (2011), 639–668.

[60] Joshua Saxe and Konstantin Berlin. 2015. Deep neural network based malware de tection using two dimensional binary program features. In 2015 10th International Conference on Malicious and Unwanted Software (MALWARE). IEEE, 11–20.

[61]J.X. Geng, J.F. Wang, Z.Y. Fang, Y.J. Zhou, D. Wu, W.H. Ge

A survey of strategy-driven evasion methods for PE malware: transformation, concealment, and attack

Comput. Secur., 137 (2024), Article 103595, [10.1016/j.cose.2023.103595](https://doi.org/10.1016/j.cose.2023.103595)

但不幸的是，某些基于静态特征分析的反病毒软件在某些情况下有着很高的误报率，可能因为机器学习模型自身的问题，误判一些正常的软件的行为，认为这些正常软件是恶意软件，例如QVM反病毒引擎误报Microsoft Visual Studio Complier、Clang等C/C++语言编译器。此外，基于机器学习的反病毒软件也存在严重的漏洞，攻击者只需要对恶意软件进行修改，甚至有时只需要增加一个资源文件改变恶意软件的Hash值，就能绕过反病毒软件的检测，恶意软件的编写者的规避技术对于反病毒软件的识别带来了极大的挑战，这促使攻防对抗的双方不断采取更先进的措施。

在过往的PE对抗性样本生成实验中，许多已有的对抗性模型生成，有些采用值函数的强化学习算法，例如使用深度学习神经网络的DQN。

[4] HyrumSAnderson,AnantKharkar,BobbyFilar,DavidEvans,andPhilRoth.2018. Learning to evade static PE machine learning malware models via reinforcement learning. arXiv preprint arXiv:1801.08917 (2018).

有些使用遗传编程算法，基于适应度进行选择，交叉，编译。

[18] Luca Demetrio, B. Biggio, Giovanni Lagorio, F. Roli, and A. Armando. 2020. Functionality-preserving Black-box Optimization of Adversarial Windows Mal ware. arXiv: Cryptography and Security (2020).

有些采用蒙特卡洛搜索树方法，将对抗样本转化为路径搜索问题。

[54] Erwin Quiring, Alwin Maier, and Konrad Rieck. 2019. Misleading authorship attribution of source code using adversarial learning. In 28th {USENIX} Security Symposium ({USENIX} Security 19). 479–496.

这些现有的方法已经展现了一定有效性，但时序差分算法，例如Sarsa Q-learning的研究却很稀少，因此本实验的主要研究目标是使用时序差分算法Sarsa构建强化学习对抗性样本生成模型，来生成能够规避部分静态检测的PE可执行程序对抗样本，使部分样本逃逸反病毒软件的查杀，本文的主要贡献如下：

（1）研究了时序差分强化学习算法Sarsa对于对抗性样本的生成。

（2）能够生成高概率逃逸360杀毒云查杀的对抗性样本并且保证恶意程序的功能，生成的恶意程序样本即使开启了允许上传可疑文件，仍然需要经过多次扫描后上传分析才得以查杀。

（3）预测了未来可能出现的高危害性感染型病毒变种和结合了静态分析的挂马网站，以及部分木马下载器变种，并且建议反病毒软件厂商们加以防范，尽管动态分析技术仍然可以针对感染型病毒变种，但该类型病毒仍具有一定威胁性。这种感染型病毒变种会在感染新文件时对于新产生的病毒自带进行自动地静态检测规避处理，同时被感染的文件相较于传统的感染型病毒难以恢复。

（4）action更加倾向于模块化，减少了模块之间的耦合程度，便于后续的研究人员从项目中直接提取函数，而无需修改大量内容，对于某些扰动函数只需要传入原恶意程序文件的绝对路径，和用户期望生成的文件的绝对路径，就能实现对抗型扰动操作。相比原有框架PSP-Mal

[70]Dazhi Zhan, Wei Bai, Xin Liu, Yue Hu, Lei Zhang, Shize Guo, and Zhisong Pan. 2023. PSP-Mal: Evading malware detection via prioritized experience-based reinforcement learning with Shapley prior. In *Proceedings of the 39th Annual Computer Security Applications Conference*. 580–593.

，本实验项目中的扰动函数与行为（Action）相互分离，并且额外加入日志系统，扰动函数自身不需要传递过多的参数。而原框架PSP-Mal对于恶意软件的操作相对混乱，难以单独从一个行为（Action中）抽象出修改函数，对恶意软件的扰动行为和智能体之间耦合度过高，导致修改较为困难。

**关于过往的对抗性样本生成的工作：**

基于强化学习的方法是通过构建动作（Action），状态（State），奖励（reward），来使智能体与环境交互，通过多次操作获取经验，挑选一个相对更好的策略来修改已有的PE恶意程序来达到逃逸反病毒软件检测。

在2017年，基于强化学习生成静态对抗性样本绕过黑盒的反病毒软件的GYM-Malware框架被提出。Anderson等人定义了10种不会影响恶意软件和功能的扰动操作，例如利用签名漏洞来更改恶意软件的签名，修改恶意软件的调试（Debug）信息，修改可选首部检验码（checksum），修改现有的节的名称。这些操作与DQN强化学习算法结合和病毒检测软件交互以指导选择的扰动操作。

[1]H.S. Anderson, A. Kharkar, B. Filar, P. Roth

Evading Machine Learning Malware Detection

Black Hat (2017)

[2] H.S. Anderson, A. Kharkar, B. Filar, D. Evans, P. Roth

Learning to evade static pe machine learning malware models via reinforcement learning

ArXiv preprint

[arXiv:1801.08917](http://arxiv.org/abs/1801.08917) (2018)

但遗憾的是，一部分使用强化学习模型生成的对抗性样本无法在虚拟机中正常执行，有一部分恶意软件经过某些操作后被破坏了。尽管某些操作修改的部分似乎与恶意软件的代码部分无关，但还是影响到了恶意程序的功能。

[61]J.X. Geng, J.F. Wang, Z.Y. Fang, Y.J. Zhou, D. Wu, W.H. Ge

A survey of strategy-driven evasion methods for PE malware: transformation, concealment, and attack

但GYM-Malware模型仍然对后续的研究有着很大的作用，许多后续的研究基于其工作来进行。其中封装的一部分动作（Action），被拆成了函数用于其他的研究中，例如Mab-malware。

[1]W. Song, X. Li, S. Afroz, D. Garg, D. Kuznetsov, H. Yin

Mab-malware: a reinforcement learning framework for blackbox generation of adversarial malware

Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security (2022), pp. 990-1003

有一些研究，关注了GYM-Malware中的一些造成恶意软件功能损坏无法正常执行的操作。Mab-Malware认为是Python的LIEF库导致的，并且对其进行修复，将使用LIEF库的一些对PE可执行程序的操作改为使用Python的Pefile库，以减少损坏的恶意程序数量。也有一些研究，直接删除了可能导致恶意程序遭到破坏的操作，引入随机化操作来缩小动作空间，限制强化学习智能体可执行的动作次数，鼓励强化学习智能体寻找更优秀的Action集合。

[1] Z. Fang, J. Wang, B. Li, S. Wu, Y. Zhou, H. Huang

Evading anti-malware engines with deep reinforcement learning

IEEE Access, 7 (2019), pp. 48867-48879

而由Labaca-Castro等人开展的研究，则考虑修正奖励函数，对奖励函数添加惩罚因子，为了进一步优化强化学习智能体的操作。对奖励函数的修正鼓励智能体尽可能用更少的步骤对恶意PE程序进行修改以逃逸反病毒软件的检测。

[2]R. Labaca-Castro, S. Franz, G.D. Rodosek

Aimed-rl: exploring adversarial malware examples with reinforcement learning

Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML, Proceedings, Part IV 21, PKDD 2021, Bilbao, Spain, September 13–17, 2021, Springer (2021), pp. 37-52

类似的，Gibert et al.等人使用空操作，即插入大量NOP指令来修改恶意软件，表明使用插入无意义空操作的方法对于绕过MalConv等反病毒软件一样是有效的。

[3]D. Gibert, M. Fredrikson, C. Mateu, J. Planes, Q. Le

Enhancing the insertion of nop instructions to obfuscate malware via deep reinforcement learning

Comput. Secur., 113 (2022), Article 102543

同时，目前也存在基于梯度的对抗性样本生成，例如FGSM、 Carlini和Wagner创建的C&W、以及deepfool模型。

[1]I.J. Goodfellow, J. Shlens, C. Szegedy

Explaining and harnessing adversarial examples

3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015)

[2]N. Carlini, D. Wagner

Towards evaluating the robustness of neural networks

2017 IEEE Symposium on Security and Privacy (SP), IEEE (2017), pp. 39-57

[3] S.M. Moosavi-Dezfooli, A. Fawzi, P. Frossard

Deepfool: a simple and accurate method to fool deep neural networks

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 2574-2582

且存在一些研究能够在已知梯度信息的情况下，通过基于梯度的方法对恶意软件的字节或者外观表现形式进行修改来规避静态特征检测。

[1] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, F. Roli

Adversarial malware binaries: evading deep learning for malware detection in executables

2018 26th European Signal Processing Conference (EUSIPCO), IEEE (2018), pp. 533-

[2] A. Al-Dujaili, A. Huang, E. Hemberg, U.M. O'Reilly

Adversarial deep learning for robust detection of binary encoded malware

2018 IEEE Security and Privacy Workshops (SPW), IEEE (2018), pp. 76-

[3] F. Kreuk, A. Barak, S. Aviv-Reuven, M. Baruch, B. Pinkas, J. Keshet

Deceiving end-to-end deep learning malware detectors using adversarial examples

ArXiv preprint

[4] O. Suciu, S.E. Coull, J. Johns

Exploring adversarial examples in malware detection

2019 IEEE Security and Privacy Workshops (SPW) (2019), pp. 8-14

[5] X. Liu, J. Zhang, Y. Lin, H. Li

Atmpa: attacking machine learning-based malware visualization detection methods via adversarial examples

Proceedings of the International Symposium on Quality of Service (2019), pp. 1-10

[6] D. Park, H. Khan, B. Yener

Generation & evaluation of adversarial examples for malware obfuscation

2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE (2019), pp. 1283-1290

相比已有的工作，本实验额外考虑了UPX加壳操作和sigthief造成的假数字签名，resourceHacker调用导致的图标类资源添加等行为，对静态检测的对抗性样本生成可能带来的影响。同时，本实验生成的部分样本能逃逸已有的反病毒软件，为相关研究提供了参考，为基于机器学习的反病毒引擎提供了可选的训练样本生成器，本实验采用的强化学习算法是基于时序查分算法的Sarsa算法。关注PE恶意程序的原因是Windows操作系统在个人电脑（PC），服务器等操作系统中占有72%左右的份额，因此超过70%的恶意软件将Windows操作系统作为攻击目标。

Statista

Global market share held by operating systems for desktop PCs, from january 2013 to february 2024

(2024)

<https://www.statista.com/statistics/218089/global-market-share-of-windows-7/>. [Online; Accessed 2024]

AV-TEST

Statistics

(2023)

<https://portal.av-atlas.org/malware/statistics>. [Online; Accessed 2023]

**环境配置：**

本实验使用如下开发环境：

实体机操作系统及版本号：Windows10 22H2

虚拟机软件，虚拟机操作系统：VMWare Workstation 16.1.2 build-17966106

Windows7 旗舰版

集成开发环境：PyCharm 2024.3.5 (Professional Edition)

Anaconda 24.11.3 AnacondaNavigator2.6.5

对抗性样本测试：主机使用Clam AV，并且手动更新病毒库到最新版本。

虚拟机采用火绒杀毒和360杀毒。

项目需要的依赖已经导出在项目根目录下的requirements.txt文件中，读者可以使用anaconda直接导入requirements.txt文件来创建Python虚拟环境。

实体机存储：

内存空间 32GB

磁盘空间 512GB

**预备知识：**

**关于PE文件：**

PE文件是Windows操作系统下的可执行文件形式，包括EXE（可执行文件），DLL（动态链接库），SYS（系统文件）等类型。

而本实验的对抗性样本生成操作，需要修改PE文件某些节的信息。

PE文件中包含PE文件头（IMAGE\_NT\_HEADERS），其中的COFF文件头（IMAGE\_FILE\_HEADER）和可选首部（IMAGE\_OPTIONAL\_HEADER）中的有些部分是我们需要修改的目标。

COFF文件头（IMAGE\_FILE\_HEADER）包含如下关键字段：

Machine（目标CPU架构，指明了能运行这个程序的机器码，可以指明支持程序运行的机器架构是x86、x64、PowerPC、ARM等）

NumberOfSections：指明所有节区的数量。

TimeDateStamp：时间戳，指明了这个文件被编译生成的时间。

SizeOfOptionalHeader：可选首部的大小。

Characteristics：文件的类型，是动态链接库，还是可执行文件等类型。

可选首部（IMAGE\_OPTIONAL\_HEADER）包含如下字段，可选首部指明了当PE程序被载入内存后的一些情况：

Magic：魔法位，包含PE信息，一定要和COFF文件头中的Machine对应，否则就会报错导致程序无法启动。。

AddressOfEntryPoint：程序入口点，即相对虚拟地址。

ImageBase：加载机制。

SectionAlignment：内存中节区的对齐粒度，不建议修改，否则程序可能无法启动。

FileAlignment：文件中节区的对齐粒度，不建议修改，否则程序将无法启动

SizeOfImage：加载到内存后的总大小。

SubSystem：子系统类型。

DataDirectory：数据目录表，记录某些数据的位置及其大小。

PE文件中的节表（Section table）描述了每个节的属性，由多个IMAGE\_SECTION\_HEADER组成，每个条目对应一个节区，包括下列关键属性：

Name：节区的名字。

VirtualAddress：虚拟地址中的起始相对位置。

SizeOfRawData：节区中数据的大小。

PointerToRawData：节区偏移量。

Characteristics：节区属性。

节区数据正常情况下应包含如下内容：

.text：代码段，存放该PE程序执行的指令。

.data：已经完成初始化的某些数据。

.rdata：只读（Read Only）的数据。

.rsrc：资源节，存放PE文件的图标等信息。

.reloc：重定位信息，可以用于动态链接库的装载过程。

.idata：import data，即导入函数信息。

数据目录表中包含的内容如下：

Import Table：导入表，用于存放该PE文件依赖的动态链接库和调用的某些函数。

Export Table：导出表，可以用于存放这个PE文件封装好的函数，多半运用于动态链接库封装函数供其余PE程序调用使用。

Relocation Table：重定位表，修复地址偏移相关问题。

TLS：线程存储表，与多线程程序有关，存储线程初始化数据。

Debug Directory：存放该PE程序的调试信息。

本实验对抗性模型生成只使用了PE可执行文件（EXE），并未使用动态链接库（DLL）和系统文件（SYS）。

**强化学习相关知识：**

**基本概念相关：**

**1.智能体：**在强化学习中决策，行动，学习。智能体是一个感知者，能感知并且理解当前的状态，智能体是一个决策者，能够知道在一个状态下应该采取什么行动，智能体是一个执行者，通过改变状态从而获取奖励。

**2.状态：**描述了智能体与环境的相对状况。

**3.状态空间：**所有状态的集合。

**4.动作：**智能体在某一状态下能选择的操作。

**5.动作空间：**所有动作的集合。

**6.状态转移：**当执行一个动作时，智能体可能从一个状态转移到另一个状态的过程。

**7.策略：**智能体在每一个状态下应该采取什么样的动作，允许分为确定性策略和随机性策略。**8.奖励：**作为人机交互的一个重要手段，可以设置合适的奖励来引导智能体按照我们的预期选择正确的决策，正数奖励表明我们鼓励智能体执行该行动，负数奖励表明我们不鼓励智能体执行该行动。

**9.回合/尝试：**智能体执行一个策略与环境交互的过程中，智能体从开始状态到终止状态停止的过程被称为一个回合或尝试，一般用英文episode来表示

**10.折扣因子：**用于调整智能体对于近期奖励和远期奖励的重视程度，可以记作折扣因子为γ，γ在(0,1)的范围，且折扣因子的引入允许了无限长的轨迹

**11.状态值：**

说明智能体在一个状态之下，最终能获取到的回报。

首先，需要了解基于时序差分策略的方法可被用于估计状态值：

时序差分方法的表达式如下：

(

代表t时刻对于的估计，是t状态下对于状态的学习率。

在t时刻，只有当时正在被访问的状态的估计值会被更新。

公式中：

作为时序差分方法的目标。

作为时序差分方法的误差。

为在t时刻对于状态的学习率。

为新的t+1时刻对于状态值估计值。

为t时刻对于状态的状态值的估计值。

Sarsa（state action reward action state action）是基于时序查分方法的强化学习算法，但是该方法不是估计状态值而是估计动作值。

[35]G. A Rummery and M. Niranjan, Online Q-learning using connection systems.

Technical Report, Cambridge University, 1994.

[3]R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction(2nd Edition).MIT Press, 2018

首先要引入动作值的概念：

对于一个状态-动作配对，动作值定义为=表示在一个状态采取一个动作之后获得回报的期望值。

这里将的估计值记作。

给定一个策略π，需要估计其动作值，可以从π的经验样本中，使用Sarsa算法来估计动作

值，其表达式如下：

其中，学习率为

在t时刻，只有的动作值被更新，其它的动作值保持不变。

Sarsa算法主要是用于求解Bellman方程近似算法，近似算法的表达式如下所示：



这是一个基于动作值而不是状态值的Bellman方程。

运用Sarsa算法学习最优策略的伪代码如下：

目标：学习最优策略，使智能体能从给定状态到达目标状态。

初始化函数：

*Init():*

*for all values and all t values:*

*select =*

*var*

*set init value as*

训练学习函数：

*Train():*

*for episode in episodes:*

*if s == :*

*from let a =*

*if !=target\_state:*

*collect experience samples**(, , , )*

*run action*

*generate ,*

*generate from*

*update value:*

*update strategy of :*

*if a ==* *:*

*else if a!= :*

*let =*

*let =*

Sarsa算法描述如下：

算法初始化：

对于所有数值和所有t数值，选取学习率因子 *= ，*贪婪因子，并且设定所有的初始值为，从获取初始贪婪策略。

算法目标：学习最优策略，使智能体能从给定的状态出发，到达目标终止状态。

对于每个回合：

如果当前状态是根据选取策略 a =

在时刻t，如果不是目标状态

收集经验样本(, , , )：在通过与环境交互生成，。

再根据生成

随后更新(, 的值

Sarsa算法也存在一些推广，例如Expected Sarsa算法、n-step Sarsa算法等，尽管本实验中并未涉及到这些推广算法，但是读者在构建强化学习对抗性样本生成模型上也可以尝试使用它们。

其中Expected Sarsa算法类似Sarsa算法，但是它们的时序差分方法目标上不同。具体实现流程上二者相似。

已知策略π，则Expected Sarsa算法对于该策略π的动作值估计公式如下：

其中：

是在策略下的期望值。

不同于Sarsa算法的目标是求出，

Expected Sarsa算法的目标是求出。

引入期望值会增加计算复杂度，但是会减小所求目标的方差，Expected Sarsa的涉及的随机变量是{}，相比Sarsa算法的{}，减少了一项，因此对于减少估计方差是有正面作用的。

[3]R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction (2nd Edition). MIT Press, 2018.

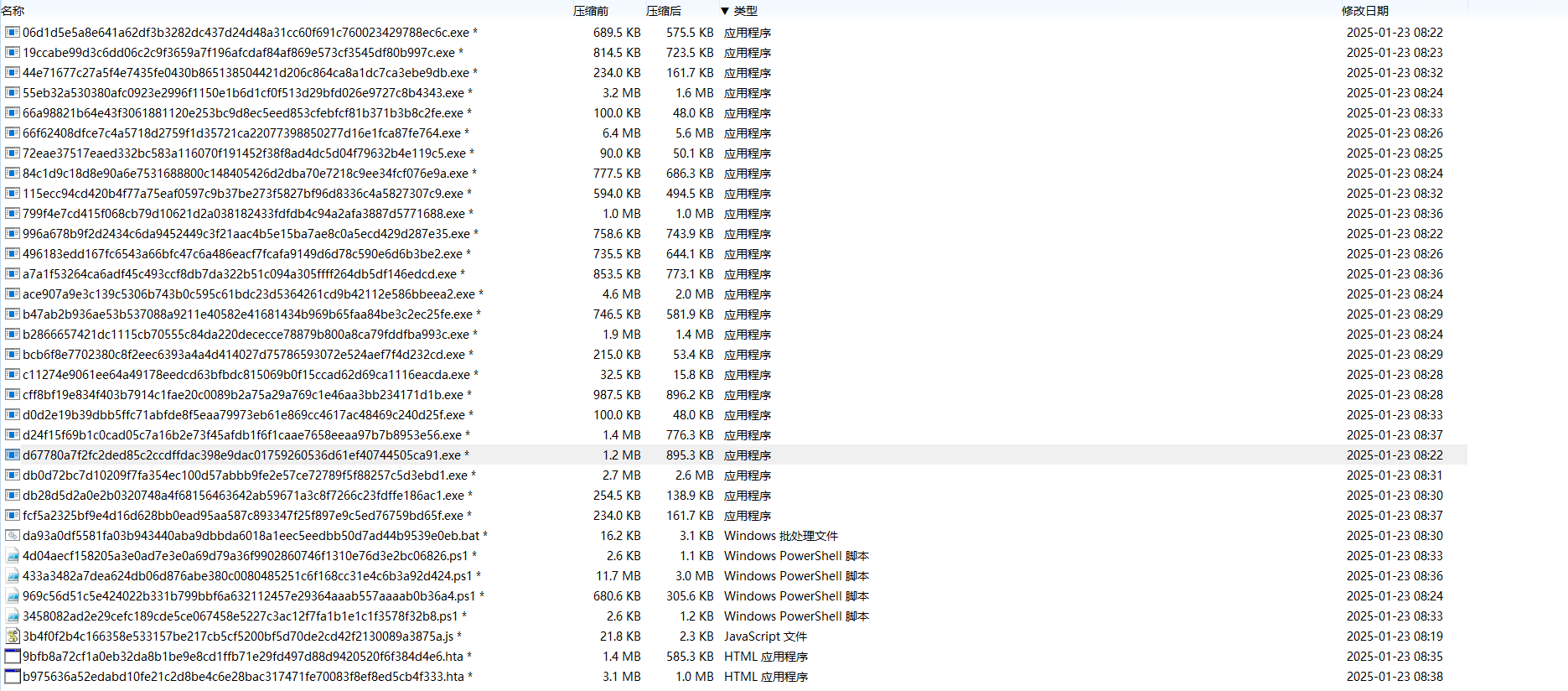
[36]H. Van Seijen, H. Van Hassekt, S. Whiteson, and M. Wiering, “A theoretical and empirical analysis of Expected Sarsa,” in IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, 2009, pp. 177-184

[37]M. Ganger, E. Duryea, and W. Hu,“Double Sarsa and double expected Sarsa with shallow and deep learning,” Journal of Data Analysis and Information Processing, vol.4, no.4,2016, pp. 159-176

**实验方案以及过程：**

本方案基于强化学习模型开展，首先，需要收集足够量的恶意样本以供强化学习模型使用，因此，笔者从国外Virus Share的开放恶意样本库，以及部分国内安全软件论坛，例如火绒杀毒，360杀毒等官方论坛，收集并且下载了大约3000个恶意样本。

但需要注意的是，有些样本可能不是PE文件，如图所示，笔者在手动筛选样本时，发现了其中的部分样本包包含其他类型的恶意程序，如下图所示。

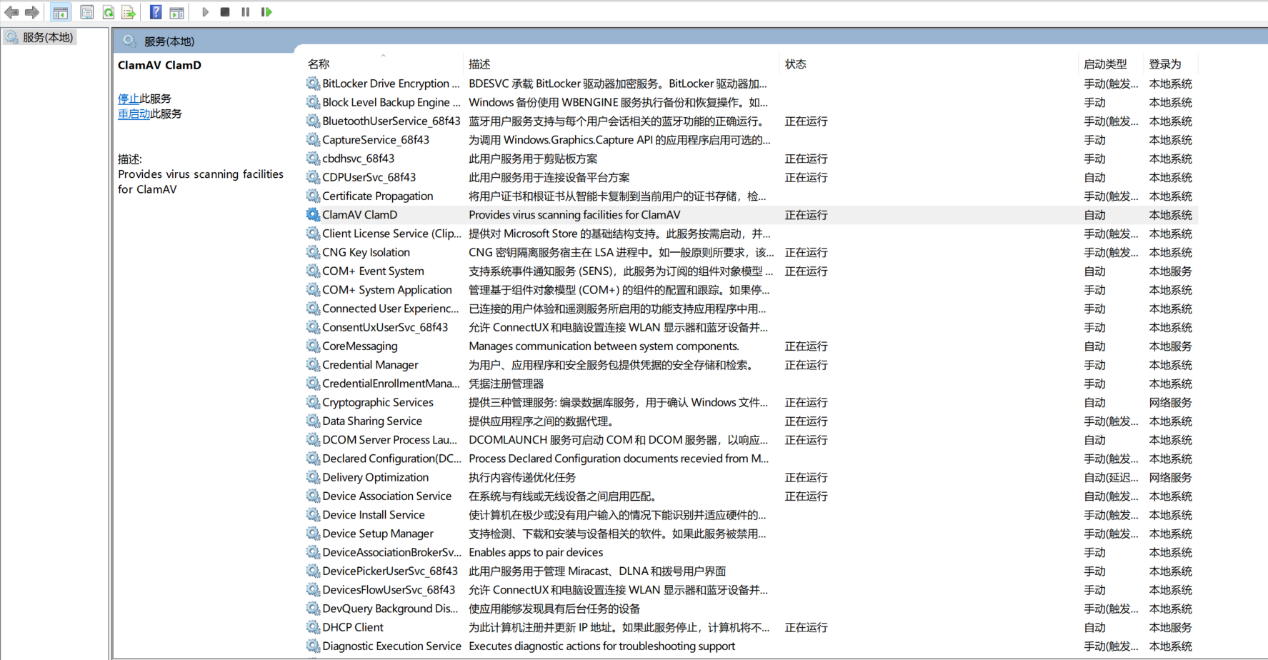


这是其中一个样本包，不难看出里面包含了恶意Windows Power Shell脚本和HTML恶意应用程序，恶意JavaScript脚本，恶意Visual Basic Script等非PE恶意程序。而非PE恶意程序的对抗性样本生成不是本研究的目标，所以对于样本的处理，需要删除掉这些非PE文件。其次，本实验并没有对恶意动态链接库（DLL）和SYS文件进行对抗性样本生成，只是对恶意PE可执行文件（EXE）进行了对抗性样本生成，是因为反病毒软件厂商提供的样本不包含动态链接库和SYS文件，且检验动态链接库在对抗性样本生成后是否因为对抗性扰动导致损坏，相比检验PE可执行文件是否损坏更困难。

本实验在筛选样本这一步操作时，前四组样本集，笔者选择了在Windows7虚拟机中手动筛选，利用Windows文件系统的相关功能，设置类型排序，然后删除后缀文件名非exe的文件，只保留PE可执行程序样本。然而，这有着比较大的操作隐患，很容易误操作，意外执行恶意程序导致可能遭受到恶意软件攻击。在后续的第五组样本集中，笔者将筛选操作自动化封装成了一个Python函数，只需要传入样本集所在的绝对路径即可自动化完成操作，同时也规避掉了手动筛选误操作的风险。

然而，上述的筛选看似合理，但是在实验中会存在一些比较严重的问题，例如有些是伪装的PE可执行文件，这些文件使用了结构化欺骗，实际上这些文件自身并不含有COFF文件头等内容，如果封装的对抗性扰动函数认为它们是PE可执行文件并且对其操作，就会因为PE格式解析异常报错，进而Python主程序将会抛出异常，没有被处理的异常会导致整个智能体的训练被迫终止，幸运的是，这可以通过在do\_action( )函数中设置异常处理来解决这个问题，即使遇到了不是正常的PE可执行文件，程序仍然因为不会抛出异常，导致强制终止整个程序。另外还存在一些PE可执行文件恶意软件，自身大小非常大，大约64MB到数百MB，可能是这些文件本身已经经过了一些对抗性操作导致的，笔者在对对抗性扰动的函数进行黑盒测试中发现，生成对抗性样本的扰动函数在处理它们时极其缓慢，有些甚至需要数分钟的时间，这无疑严重拖累了强化学习智能体自身运行的速度，因为每个回合（episode）是不能多线程并发执行的，必须要按照单线程，进行顺序执行。其次，这些较大的PE可执行程序读取和写入操作对磁盘的损害也是巨大的，极有可能在运行中发生系统蓝屏死机的问题，甚至可能会导致数据丢失。笔者的计算机在运行第五组样本时已经发生过一次因为硬盘故障导致的Windows蓝屏问题，幸运的是，重启后固态硬盘自检后成功修复。因此，我们还需要对样本进行第二次筛选，在本实验中，样本中大多数大于4096字节的PE可执行程序在挑选样本时需要被删除，以避免智能体出现故障导致程序强制终止。

当筛选出大约400~600个PE可执行程序恶意软件时，将这些样本归为一个样本集，以备后续使用。同时，我们需要对原始样本使用反病毒软件扫描。这一步笔者是在Windows10的实体机上配置了Clam AV，使用最新的病毒库，并且在系统服务中需要手动开启Clam AV的服务。对于Clam AV的病毒扫描服务，笔者选用的是3310端口号，需要在clamd.conf中修改TCPSocket属性的值来指定服务端口号，读者在这一步环境配置需要注意端口号冲突问题，否则服务将无法启动，同时这个端口号也是Python调用Clam AV服务所需要的。下图中所示是Clam AV服务成功启动的状况。



但是在使用360杀毒，火绒杀毒等具有云查杀功能的反病毒软件对原始样本进行扫描时，需要注意样本污染问题，对于360杀毒，需要关闭自动上传发现的可疑文件，如下图所示，否则可能导致有些样本被360杀毒上传到云查杀服务器数据库中，导致后续结果查杀率偏高的问题。



**实验中所需的恶意软件检测器配置：**

为了测试智能体的操作是否有效，本实验使用了Python的pyclamed库的0.4.0版本搭建了一个恶意软件检测器，并且封装成了一个函数，如图所示，函数需要传入带扫描文件的绝对路径。

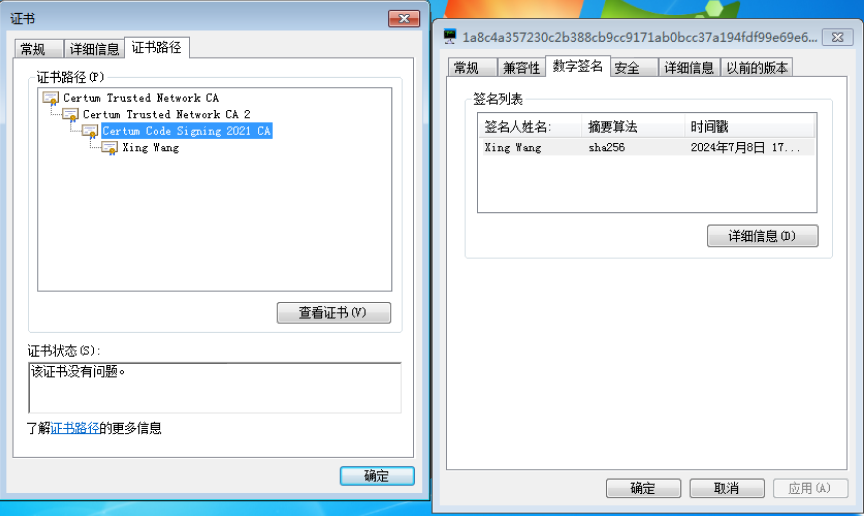


首先需判断Clam AV的服务是否处于活动状态，如果不处于活动状态，就抛出一个网络连接异常，同时，需要注意参数file\_path对应的文件是否存在，如果文件不存在，也需要抛出异常。当文件存在时，调用scan\_file接口对该绝对路径下的文件进行扫描，如果扫描结果是空，说明文件是良性的，如果扫描结果非空（None），说明改文件是恶意软件。这个函数最终返回一个bool值，为True，说明函数参数中传入的文件是恶意软件，为False，说明函数参数中传入的文件是良性的。不返回具体恶意软件类型名称是因为智能体不需要用到恶意软件类型名称，只需要检测Clam AV扫描传入的文件是否是恶意软件即可。

**实验所需的PE文件操作模块配置：**

本实验中针对PE的对抗性操作包括如下内容：

1.使用sigthief，将一个从良性软件中提取出的数字签名附加在恶意软件上，如图所示。



2.在PE程序的尾部添加无意义的字节:

添加的字节使用随机生成的字符串。

3.修改校验和（checksum）:

该操作需要传递一个checksum参数，用于指定修改后的checksum内容，如果缺少这个参数，则默认指定为0

4.修改机器码（因为会导致修改后的PE可执行文件无法正确运行，因此已经被本实验后续弃用）:

修改PE可执行文件的Machine和Magic内容。但如果一旦修改后的机器码和软件运行平台的机器码不同，则程序会无法启动。此操作被笔者认为是对恶意程序有破坏性的操作，会使恶意程序丧失原本的功能。

5.修改时间戳：

此功能本实验用pefile库和lief库分别实现了该功能，函数会随机生成一个过去一年之内的时间对应的时间戳，并且将生成的时间戳替换掉恶意软件原有的时间戳

6.UPX加壳操作：

使用The Ultimate Packer for executables加壳工具，是一个GitHub网站上的开源壳项目。

[2]<https://github.com/upx/upx>

同时，它也能做到压缩原PE可执行程序的大小。

读者在运行UPX加壳操作时需要注意UPX加壳程序的路径，防止因为路径问题导致抛出异常加壳失败。

7.向节之间的空洞中添加无意义的字节：

添加的字节使用随机生成的字符串。

同时，也需要一个search\_cave()函数用于寻找节之间的空洞。

8.修改导入表，增加PE可执行程序导入的函数：

从small\_dll\_imports.json中选取导入的函数来修改导入表，会添加引用的动态链接库或函数。

9.在PE可执行文件后添加一个节，节中包含无意义的字符串数据：

此操作使用LIEF库实现，添加的字节使用随机生成的字符串

10.修改PE可执行程序首部校验和：

将PE可执行程序的首部校验和修改成0

11.删除PE可执行程序的调试信息：

可能因为没有debug信息，导致操作无意义，但不会抛出异常。

12.重命名原PE可执行程序中的某个节，备选命名名称会在

.abitcsa  
.bbitcsb  
.cbitcsc  
.dbitcsd  
.ebitcse  
.fbitscf  
.gbitcsg  
.hbitcsh  
.ibitcsi  
.jbitcsj  
.kbitcsk

中任选一个。

13.向原PE可执行程序中增加资源：

增加的资源为多个ICO文件，是从备选的ICO文件中随机选取，然后截取随机大小，再调用resourcehacker将处理后的ICO文件加入到资源节中。

这些操作需要被封装成函数，传入参数需要包含待修改的可执行文件，经过对抗性干扰操作后，将修改后的可执行文件需要移动到原路径。

读者应注意此操作的resourcehacker路径和备选ico路径，防止Python抛出找不到文件的异常。

但上述的这些操作不适合直接作为动作（action），需要使用logging模块再进行一次封装，记录修改前和修改后文件的SHA256哈希值和使用的操作名称在一个日志文件中。经过这样的操作有利于中途程序进行黑盒测试时排查错误，以及根据文件的哈希值判断这些对抗性干扰操作是否有效。修改成功的一个必要条件是修改后文件的哈希值与修改前文件的哈希值不同。

如图所示：



**关于查杀率的定义：**

对于某个恶意软件集合，反病毒软件的查杀率被定义为经反病毒软件扫描该恶意软件样本集合的检出恶意软件数量除以恶意软件集合中总恶意软件数量。

**实验中生成PE对抗性恶意样本的算法：**

输入样本集，行为集，初始化后的Sarsa模型，日志是否被允许记录。

伪代码如下：

void alogrithm\_generate\_controversy\_samples(sample\_directory,actions\_list,Sarsa\_model,enable\_log):

    Sarsa\_model.init(sample\_directory,actions\_list,gamma,alpha,epsilon,test\_mode,enable\_log);

    train(Sarsa\_model);

int train(Sarsa\_model):

    episodes = get\_episodes\_for\_training(Sarsa\_model.sample\_directory);

    evasion\_count = 0;

    for episode in episodes:

        count = 0;

        state = 0;

        total\_reward = 0;

        status\_done = False;

        reward = 0;

        action = choose\_action(state);

        do\_action(action,episode,enable\_log);

        while not status\_done:

            judgement = can\_evasion\_clamAV(episode);

            if judgement == True :

                evasion\_count++;

                next\_state = None;

                reward = SUCCESS\_REWARD;

                status\_done = True;

            else:

                count++;

                if count >= Max\_Count:

                   reward = FAIL\_REWARD;

                   next\_state = None;

                   status\_done = True;

                else:

                   next\_state = count;

            total\_reward+=reward;

            if status\_done:

                Sarsa\_model.update(state,action,reward,next\_state=None,next\_action=None)

            else:

                next\_action = choose\_action(next\_state);

                Sarsa\_model.update(state,action,reward,next\_state,next\_action);

                action = next\_action

                do\_action(action,state,enable\_log);

            write\_total\_reward\_to\_log(enable\_log);

    return evasion\_count;

总体生成流程如下：

Step1：传入恶意样本集的目录和其它参数，先将恶意样本的绝对目录存储，以备后续Sarsa智能体模型使用，同时预加载Sarsa模型（Sarsa\_model）。

Step2：对于样本集中的每个样本episode，先初始化奖励。

Step3：根据状态当前选择一个行为，episode经由处理后，经由前文中封装的clam AV反病毒软件扫描工具，检测是否成功逃逸反病毒软件的扫描。

Step4：如果成功，则直接进入终止态，获取奖励SUCCESS\_REWARD并且更新Sarsa模型的Qtable，否则进入下一个状态。

Step5：如果失败，但是次数没有达到Max\_Count，则更新Sarsa模型的Qtable，并且重复Step3的过程。如果失败且次数达到了Max\_Count，进入终止状态，并获取奖励FAIL\_REWARD（可以设置为负数）和更新Sarsa模型的Qtable。

这个算法在实际运行时，会被略微修改，因为伪造数字签名的行为完成后，几乎不能使用其它行为，例如新增包含无害字节段，可执行的程序尾部新增直接的操作，否则会导致恶意软件损坏。本实验中一旦完成了伪造数字签名操作，将会修改count时完成本次操作后就会转变到终止状态。

此外，本实验引入了惩罚因子，它与恶意PE可执行文件处理前后的大小有一定关系，这里定义改变比：

令惩罚因子

在因为成功达到最终状态获取的奖励reward会被再次修正为reward=reward\*。

看似能尽量促进智能体选择尽量不过分扩大原恶意PE可执行文件的大小，然而实行运行时很小概率下会发生问题，这表现在使用UPX加壳后，过于小，导致惩罚因子变成负数，理论上本该因为达成了逃逸效果，被Clam AV检测为良性程序，应该获取的是正数的奖励，但实际上获取的是负值奖励，这与达成逃逸获取正数奖励的设定矛盾。

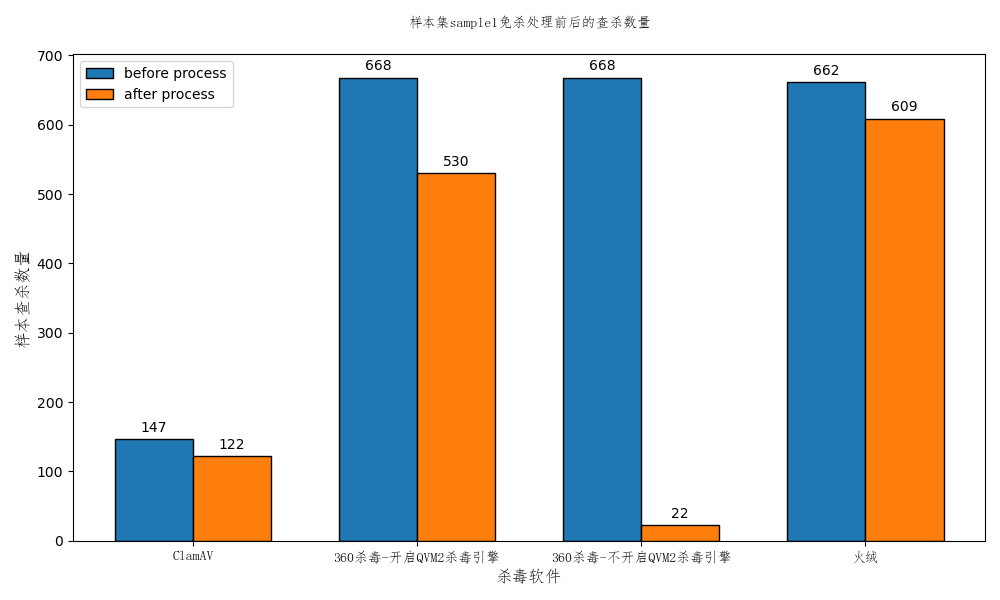
因此，最好修正惩罚因子，修正后的惩罚因子为.

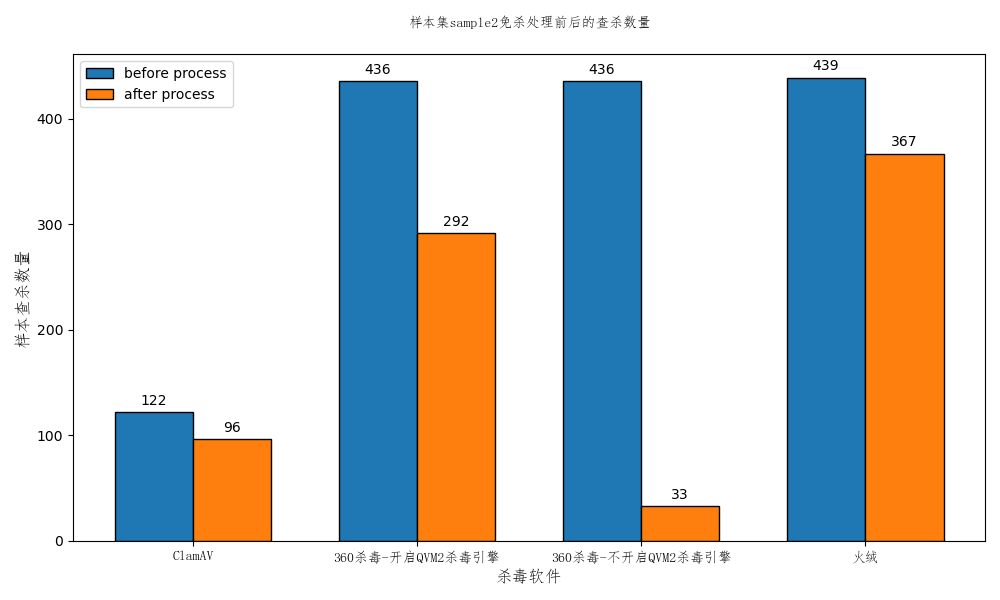
令惩罚因子

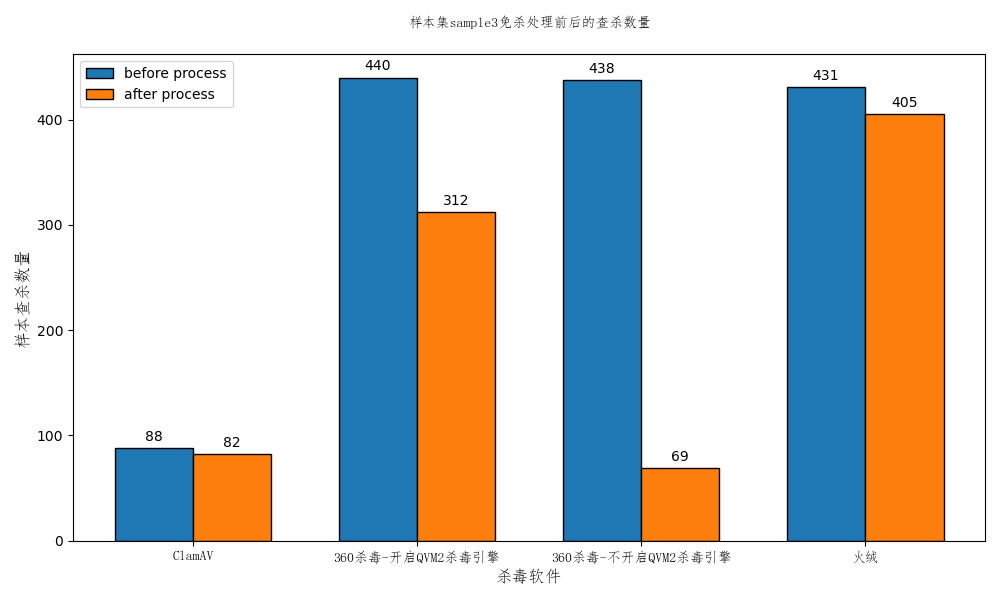
这样可以避免使惩罚因子变为负数，从而影响智能体做出更优的决策的意外。

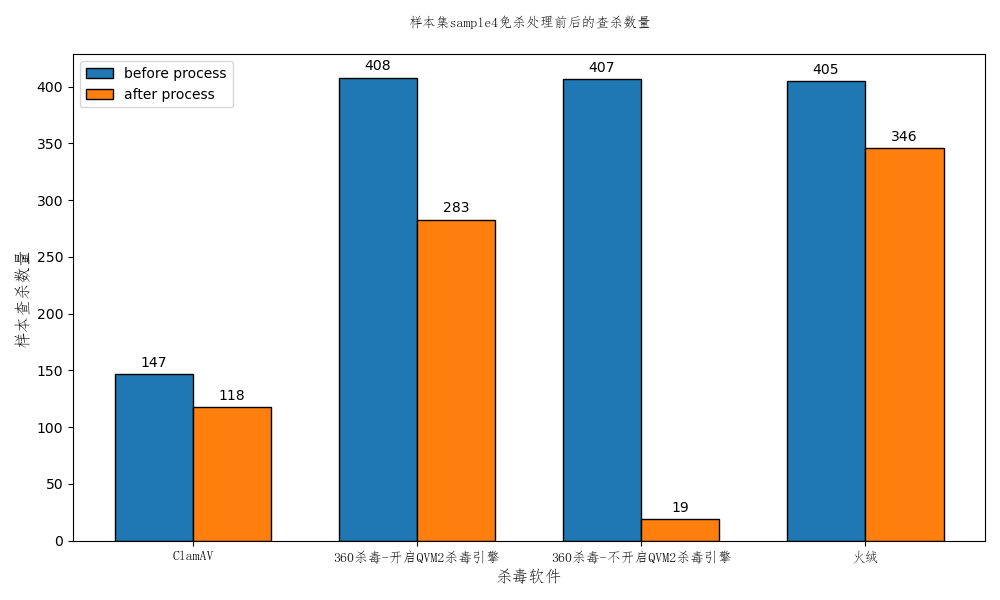
**关于实验结果的讨论：**

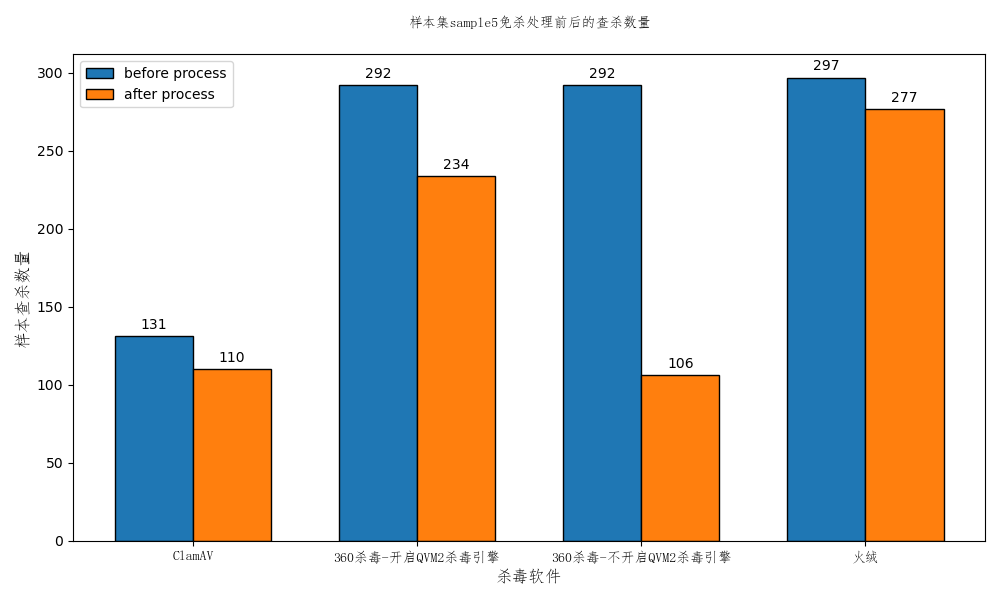
在本次实验中，笔者一共使用了五组数据，合计大约2000个样本来进行测试，测试结果如下所示。值得注意的是，在虚拟机中使用360杀毒和火绒对样本扫描之前，需要打开虚拟机的FTP连接，并且关闭虚拟机中的反病毒软件，再在主机上使用XFTP等软件向虚拟机中发送样本集，防止样本集发送过程中受到反病毒软件的干扰和潜在的数据污染问题，例如360杀毒会默认禁止存在风险的允许远程主机向本主机的FTP协议，在本次实验中对于虚拟机我们不需要这个防御功能。且在主机和虚拟机的FTP协议完成数据传输后，虚拟机接收到文件后，360杀毒会自动扫描文件，这是不应该被允许的。因为会导致后续检出率偏高以及360杀毒等反病毒软件会直接隔离虚拟机接收到的样本集中的一些样本，对后续统计查杀数量带来了不便。



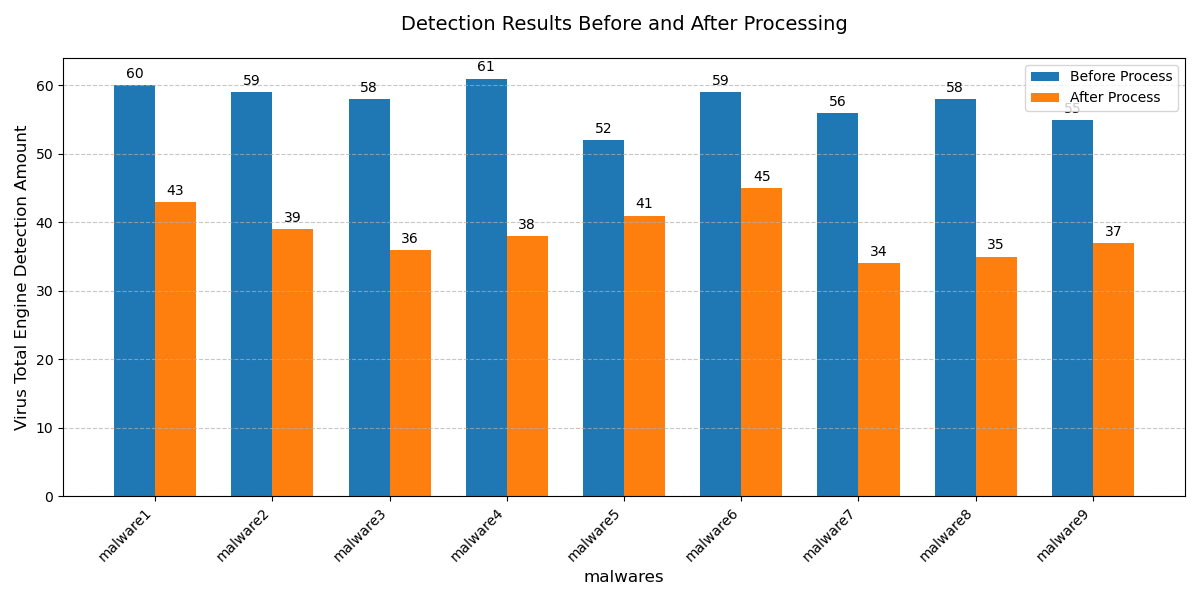








此外，对于某组样本中的部分程序，笔者使用了Virus Total来进行上传扫描，结果如下。



经过计算，可以发现：

对于样本集sample1：

Clam AV对强化学习模型处理前后的样本检出率从22.0%下降到了18.3%。

开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从100.0%下降到了79.3%。

只开启360云查杀引擎，不开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从100.0%下降到了3.3%。

火绒杀毒对强化学习模型处理前后的样本检出率从99.1%下降到了91.2%。

对于样本集sample2：

Clam AV对强化学习模型处理前后的样本检出率从27.8%下降到了21.9%。

开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从99.3%下降到了66.5%。

只开启360云查杀引擎，不开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从99.3%下降到了7.5%。

火绒杀毒对强化学习模型处理前后的样本检出率从100%下降到了83.6%。

对于样本集sample3：

Clam AV对强化学习模型处理前后的样本检出率从20.0%下降到了18.6%。

开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从100%下降到了70.9%。

只开启360云查杀引擎，不开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从99.5%下降到了15.7%。

火绒杀毒对强化学习模型处理前后的样本检出率从98.0%下降到了92.0%。

对于样本集sample4

Clam AV对强化学习模型处理前后的样本检出率从36.0%下降到了29.0%。

开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从100%下降到了69.4%。

只开启360云查杀引擎，不开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从99.8%下降到了4.7%。

火绒杀毒对强化学习模型处理前后的样本检出率从99.3%下降到了84.8%。

对于样本集sample5

Clam AV对强化学习模型处理前后的样本检出率从43.7%下降到了37.7%。

开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从97.3%下降到了78.0%。

只开启360云查杀引擎，不开启QVM2杀毒引擎的360杀毒对强化学习模型处理前后的样本检出率从97.3%下降到了35.3%。

火绒杀毒对强化学习模型处理前后的样本检出率从99.0%下降到了92.3%。

为了更好计算样本集的逃逸反病毒软件查杀效果，且要求生成对抗性样本前后使用同一款反病毒软件，则查杀下降率定义如下：

根据查杀下降率的定义，对于五个样本集合的查杀下降率计算结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sample1 | Sample2 | Sample3 | Sample4 | Sample5 |
| Clam AV | 17.0% | 14.3% | 6.8% | 19.7% | 16.0% |
| 360杀毒开启云查杀引擎和QVM2引擎 | 20.7% | 33.0% | 29.1% | 30.6% | 19.9% |
| 360杀毒开启云查杀引擎不开启QVM2引擎 | 96.7% | 92.4% | 84.2% | 95.3% | 63.7% |
| 火绒 | 8.0% | 16.4% | 6.0% | 14.6% | 6.7% |

计算平均值如下：

|  |  |
| --- | --- |
| Clam AV | 14.76% |
| 360杀毒开启云查杀引擎和QVM2引擎 | 26.66% |
| 360杀毒开启云查杀引擎不开启QVM2引擎 | 86.46% |
| 火绒 | 10.34% |

从结果中看出，对抗性样本对于不同反病毒软件的有效性，对360云查杀引擎最有效，对火绒的有效性最差。

前文中使用Virus Total扫描上传从某个样本集合内随机抽样的PE恶意程序，在云端总共使用了70余个反病毒软件进行分析。

Virus Total检出率的定义为：

Virus Total检出下降率定义如下

处理前后检出率变化以及检出下降率变化如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 处理前检出率 | 处理后检出率 | 检出下降率 |
| malware1 | 60/72=83.3% | 43/72=59.7% | 28.3% |
| malware2 | 59/73=80.8% | 39/72=54.2% | 33.0% |
| malware3 | 58/73=79.5% | 36/72=50% | 37.1% |
| malware4 | 61/73=83.6% | 38/71=53.5% | 36.0% |
| malware5 | 52/73=71.2% | 41/72=56.9% | 20.0% |
| malware6 | 59/72=81.9% | 45/73=61.6% | 24.8% |
| malware7 | 56/72=77.8% | 34/72=47.2% | 39.3% |
| malware8 | 58/72=80.6% | 35/72=48.6% | 39.7% |
| malware9 | 55/73=75.3% | 37/72=51.4% | 31.8% |

这9个随机抽样的样本，处理前，平均有57.6个反病毒引擎可以检测出样本是恶意软件。

处理后，平均有38.7个反病毒引擎能检测出。处理前检出率平均值为79.3%，处理后检出率平均值为53.7%，平均检出下降率为32.2%，说明了本实验提出的强化学习模型对抗样本生成对于逃逸反病毒软件查杀有一定的作用。

以下是笔者在结果研究中遇到的一些问题的思考以及一些建议。

**问题1：为什么火绒和Clam AV对于免杀处理后的样本和处理前的样本检出差距相比360杀毒比较低？**

笔者发现，处理后的样本对于火绒杀毒来说，逃逸概率低的原因是火绒有一定的反扰动能力，处理后的样本很多都会判断为如下的病毒类型：



说明有一部分反病毒软件的静态特征分析能力，能识别出对抗样本中的扰动。这也是为什么采用Sarsa算法不宜设置过大的state数值的原因，一方面是处理前后的PE程序反复读写，当程序大小因为新增节，新增无意义尾部内容，用无意义字节填充空洞等操作多次使用后变的过于庞大（在某次样本处理时，笔者观察到了一个PE程序处理前后，变为原大小大约6.6倍，从2.67 MB变成了17.6MB），如果不限制state值，程序中途会变得很大，不但导致后续的对抗性样本生成的干扰操作极其缓慢，因为要先读取原文件，然后再加以修改，最后写入磁盘中，Python对于文件的操作是很缓慢的。而且会导致根据查杀结果获取result的速度相当慢，因为Clam AV反病毒软件扫描一个样本的时间也和样本大小正相关，当样本很大的时候，判断查杀结果会相当的慢。另一方面，当无意义尾部内容过多，会被反病毒软件认为是疑似恶意程序刻意扰动，从而直接判定此程序为为了绕过反病毒软件检查的恶意程序。而从火绒杀毒对于样本的描述也可以看到这一点，火绒杀毒认为经过处理后的某些恶意程序是代码混淆器。

同样对于Clam AV，可能它的部分检测也能对抗样本中的扰动，尽管经过处理后的恶意程序中有大量良性内容，但Clam AV仍然会查杀这些恶意程序。

**问题2：为什么需要研究QVM引擎开启先后的情况？**

这是因为QVM作为一个自学习的人工智能引擎，这个病毒检测引擎是基于机器学习的模型，通过检测文件的某些特征，来判断该文件是否是恶意程序，如果某程序被QVM引擎检测出是恶意程序，那么360杀毒在查杀结果中会报HEUR/QVM.xx.Malware.Gen字样。这里使用QVM引擎作为检测标准是为了研究论文中的强化学习对抗性样本生成模型产生的处理后静态检测规避对抗性样本在基于不同模型的反病毒软件之间的逃逸可迁移性，以及验证机器学习的病毒检测引擎对于静态检测规避对抗性样本的抗干扰能力。但遗憾的是，QVM引擎对于静态检测规避对抗性样本的抗干扰能力比较差，有时甚至仅仅增加了无意义的图标ICO资源，QVM引擎就认为样本是良性的。

**问题3：为什么不测试360杀毒的鲲鹏引擎？**

这是因为大多数反病毒软件，倾向于防御现有的威胁和威胁变种，而不是没有广泛传播的对抗性样本，鲲鹏引擎也如此，它多被360杀毒用于对抗目前广泛流行的恶意样本（例如银狐和某些流行的勒索病毒），而非对抗性样本或是离目前时间较久远的样本。这是因为有时需要反病毒软件对未知程序的判断更加精准以及判断速度更快，也就是尽量降低误报的可能性以及判断时间。为了减少误报率，很多反病毒软件厂商可能会使用白名单数字签名放行，或是使用文件哈希白名单来进行放行。因为例如Intel，AMD的驱动程序，它们会加载驱动，但这些硬件驱动程序跟某些病毒的行为很相似，某些病毒也会使用驱动来提升自己的操作权限，使自己具有更强的破坏力以及更好地针对和破坏反病毒软件，如果这些硬件外设驱动程序因为释放了一些驱动文件被反病毒软件误报没有放行，将会导致某些外设驱动不能正常安装，对用户来说后果无疑是灾难性的，极有可能导致操作系统或是计算机崩溃。

**问题4：对抗性样本攻击对于使用反病毒软件的用户是否有很大的危害性？**

这里笔者是在VMware创建的Windows7虚拟机中使用360杀毒进行测试静态检测规避对抗性样本的查杀情况，并且在360杀毒的设置中打开了自动上传发现的可疑文件，尽管部分处理后的样本逃逸了360杀毒的自定义查杀的检出，但是360杀毒仍然将这些逃逸样本归类为可疑文件。虽然这些逃逸样本不会报为高危风险项，但是360杀毒仍然会将这些样本上传到360云安全中心来进行检验，在可疑样本被360杀毒上传之后的一段时间，再次启动自定义查杀扫描剩余的逃逸样本，有部分样本被360杀毒检出判定为高危风险项。说明对于开启未知程序上报的云查杀（笔者后续得知360云安全中心是有动态行为检测的，并不完全只有静态检测）的安全软件来说，对抗性样本攻击虽然有一定危害，但大多数仍然最终会被检出是恶意程序。同时，360杀毒也具有动态检测功能，例如Behavioral脚本引擎、主动防御模块这些基于动态行为分析的检测，如果用户执行了恶意程序，360杀毒会拦截并且隔离恶意程序。当反病毒软件具有动态检测功能，则静态对抗性样本攻击的效果是有限的，但仍然不可以忽视。究其原因，是因为某些病毒具有反虚拟化策略、反沙箱、识别监控进程的功能。

Yokoyama A., Ishii K., Tanabe R., Papa Y., Yoshioka K., Matsumoto T., Kasama T., Inoue D., Brengel M., Backes M., *et al.*

Sandprint: Fingerprinting malware sandboxes to provide intelligence for sandbox evasion

International Symposium on Research in Attacks, Intrusions, and Defenses, Springer (2016), pp. 165-187

Rudd E., Rozsa A., Gunther M., Boult T.

A survey of stealth malware: Attacks, mitigation measures, and steps toward autonomous open world solutions

IEEE Commun. Surv. Tutor., 19 (2) (2017), pp. 1145-1172

Bulazel A., Yener B.

A survey on automated dynamic malware analysis evasion and counter-evasion: PC, mobile, and web

Proceedings of the 1st Reversing and Offensive-Oriented Trends Symposium, ACM, ACM, New York, NY, USA (2017), p. 2

Apostolopoulos T., Katos V., Choo K.-K.R., Patsakis C.

Resurrecting anti-virtualization and anti-debugging: Unhooking your hooks

Future Gener. Comput. Syst., 116 (2021), pp. 393-405

云查杀技术对这些病毒来说未必有效，因为在被上传到云端沙箱（例如Cuckoo Sandbox）中之后，恶意软件不会表现出预期的恶意行为，从而阻碍了恶意软件行为分类和恶意行为的自动识别检测。过度依赖云查杀仍然可能会存在潜在的威胁，同时，云查杀需要联网访问云端恶意软件样本数据库的问题也是一大安全隐患，对抗性恶意程序可能在自己被运行时，因为反病毒软件因为对抗性恶意程序自身的扰动导致判断该恶意程序是未知安全性，判断后需要上传到云端服务器，来利用云查杀技术进行分析，而这时的恶意程序自身可能会通过修改网络设置导致计算机无法连接到互联网，或利用Windows防火墙规则，本地Hosts文件篡改反病毒软件请求服务器地址到本地环回地址127.0.0.1等方式，使反病毒软件无法连接到互联网，从而阻止了反病毒软件从云端服务器获取云端沙箱检测等传回来的信息，这就导致反病毒软件对于对抗性恶意程序无法进行精确地判断是否具有恶意性。更糟的情况是当云查杀技术面对未来可能出现的感染型病毒变种对抗性样本和蠕虫对抗性样本。

传统的感染型病毒，例如文件型病毒，只感染固定的目标文件，例如PE可执行文件（.exe），动态链接库文件（.dll），html文件（.html/.htm），以及目前发现的一些宏病毒，只感染office文件，例如后缀为docx，xlsx，pptx，xls，doc，ppt的文件等。

"Parasitic malware: The resurgence of an old threat", Network Security, vol. 3, pp. 15-18, 2008.

其中的一部分传统感染型病毒，通过将自身的代码附加在正常程序后来感染正常程序。当程序被运行时，其中的病毒代码也随之运行。然而，传统的感染型病毒很容易被反病毒软件查杀，因为反病毒软件只需要进行全盘扫描，扫描文件系统中的所有文件来寻找已被病毒感染的文件，通过静态分析，检查其尾部内容是否有病毒库中黑名单的恶意代码字段，即可判断这个文件是否已经被感染型病毒所感染。而且反病毒软件对被感染的文件可以通过匹配相应二进制字符串，如通过正则表达式匹配，从而直接去除截断掉尾部的病毒代码进行文件修复。因为病毒代码和正常程序代码是物理分离的，所以反病毒软件的操作可以将受感染型病毒感染的文件恢复到原始文件状态，即对被感染的文件实行修复而不是直接删除或者隔离，可以使被感染的文件保留其原始正常文件的功能。

然而，通过本实验，可以说明，将来很可能存在某种感染型病毒变种对抗性样本。不同于传统的感染型病毒，它的危害相对较大，本身具有感染新文件后，自身也会突变不断制造新变种的新型攻击方式。这种病毒除了自身的恶意代码，也携带了对抗性扰动操作的相关模块，例如产生一个内部包含被pyinstaller打包的附带Python编译的PE可执行文件，用于产生静态检测对抗性扰动来针对基于机器学习和深度学习的反病毒软件以实现逃逸静态检测的查杀。在病毒因为用户运行导致入侵未被病毒感染的系统后，该病毒会在磁盘中通过遍历文件系统的目录来寻找新的可被该病毒感染的目标文件，将自身代码附加在被感染程序后，并判断产生对抗性扰动操作的程序对应的Environment\_Variable\_α环境变量是否存在，若存在则调用产生对抗性扰动的可执行程序对被新感染的程序进行处理，否则释放可产生对抗性扰动的可执行程序，随后将其移动到安装有操作系统的磁盘的某个位置（可以为系统根目录）下，随后设置文件为隐藏属性，添加该可执行程序到系统的环境变量中，最后调用产生对抗性扰动的可执行程序对被新感染的程序进行静态免杀处理。··

该变种对抗性样本的部分行为可用伪代码和程序流程图表述如下：

start();

scan\_all\_files\_in\_system\_to\_get\_can\_infected();

function1:

if System\_Environment\_Variables.contains(Environment\_Variable\_α):

    for file in files\_virus\_can\_infected:

        append\_malicious\_code(file);

        run\_evasion\_module(Environment\_Variable\_α);

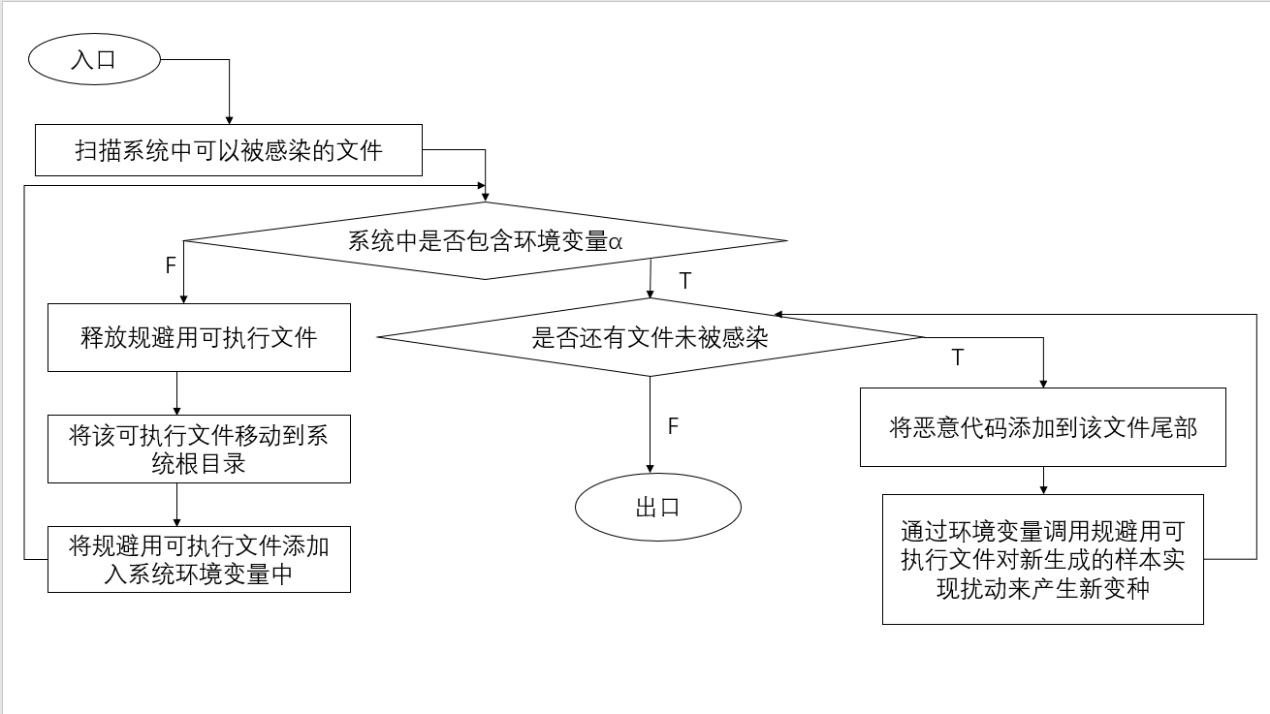
else:

    release\_executable\_file\_for\_evasion();

    move\_evasion\_module\_file\_to\_system\_root\_folder();

    add\_environment\_variable\_to\_system\_environment\_variables(Environment\_Variable\_α);

    go to function1;



这种感染型病毒变种每次新感染的程序的尾部病毒代码将互不相同，因为受到了对抗性扰动可执行程序的修改。基于静态分析的反病毒软件很难查杀这种变种对抗性样本，很容易面临能杀毒，但是查杀不完全的问题，每次杀毒如果不能检测出感染型病毒变种对抗性样本感染的全部文件，遗留下来的受感染文件一旦被用户无意间通过启动，例如通过某些Windows应用程序的开机自启动，用户打开浏览器访问网页但是浏览器主程序已被感染，仍然会导致其他文件被感染。更危险的是，被感染的文件难以像应对传统文件型病毒的修复措施，即精准去掉尾部的病毒代码来恢复原始文件。因为需要处理对抗性扰动新增的节和无效字节，甚至有可能是导入表被篡改，节表被修改的行为，一旦去除有偏差，则会导致受感染文件无法正常运行。因此反病毒软件极有可能无法通过黑名单中的病毒代码来计算线性偏移量，进而直接截断固定长度的尾部从而修复直接受感染文件，只能选择删除，这对用户来说是危害巨大的，因为很多的Windows软件包含了大量的动态链接库（.dll）文件和PE可执行文件（.exe），一旦某些动态链接库被反病毒软件删除后，这些软件极有可能无法正常运行。即使反病毒软件拥有云查杀功能，频繁上传大量变种样本不但消耗了用户带宽资源，而且变种样本因为每个被感染的实例的尾部特征值唯一，难以在第一次扫描中就被反病毒软件检测出是恶意程序，只会被定性为不流行的可疑文件，这导致对感染型病毒的响应延迟，很有可能会面对云查杀查杀病毒并不彻底的问题。这将是一种可能存在的高级持续性威胁（APT）。

幸运的是，即使面对这种将来可能存在的感染型病毒变种对抗性样本，因为它只有不断产生静态规避变种的功能，而没有不断产生动态规避变种的功能，反病毒软件的动态行为检测仍能在病毒程序未执行感染操作时触发，识别出恶意程序并警告用户。同时，PE文件的动态规避目前尚未出现规模化的攻击方法，尽管已有研究证明可通过代码分片来攻击动态行为检测。

[33]2012 IEEE Symposium on Security and Privacy

Smashing the Gadgets: Hindering Return-Oriented Programming Using In-Place Code Randomization

Vasilis Pappas, Michalis Polychronakis, and Angelos D. Keromytis

Columbia University

[50] Jithin Pavithran, Milan Patnaik, and Chester Rebeiro. 2019. D-TIME: Distributed Threadless Independent Malware Execution for Runtime Obfuscation. In 13th USENIX Workshop on Offensive Technologies (WOOT 19). USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/woot19/presentation/ pavithran

[29] Kyriakos K. Ispoglou and Mathias Payer. 2016. malWASH: Washing Malware to Evade Dynamic Analysis. In 10th USENIX Workshop on Offensive Technolo gies (WOOT 16). USENIX Association, Austin, TX. https://www.usenix.org/ conference/woot16/workshop-program/presentation/ispoglou

同时，基于静态规避对抗性生成模型，未来也有可能会出现新的木马下载器变种，以及基于Web请求拦截器和对抗性生成模型挂马网站变种，这值得反病毒软件厂商警惕。许多网页框架，例如Java的SpringBoot CGI服务器，具有Interceptor接口用于拦截器操作，然而，这可以被黑客用于木马下载器来下载其他木马的静态规避变种，黑客可以使用SpringBoot的Interceptor接口确保每次访问挂马网站下载到的木马都是不同的。当木马下载器试图使用http的get请求访问黑客的挂马服务器来将木马下载到本地，这里假设为<https://www.malicous.net/trojan_file>，在黑客对挂马服务器业务逻辑处理进行修改后，当服务器后端接收到对于trojan\_file子网页的请求，首先会被拦截器拦截，拦截器含有的业务逻辑会先将原始木马文件生成对抗性样本，然后，拦截器的业务逻辑会删除掉上一次生成的对抗性样本，使用新生成的对抗性样本对其覆盖，当对抗性操作结束后，拦截器放行请求，相对应的trojan\_file\_servlet对http请求进行处理，返回301状态码进行重定向操作，定向到挂马服务器上已经经过对抗性模型处理后的对抗性样本，在这里假设为<https://www.malicous.net/trojan1_modified.exe> ，服务器返回静态对抗性木马trojan1\_modified.exe给木马下载器，随后，木马下载器将该文件保存到受害者的计算机上以备黑客后续入侵使用。尽管反病毒软件可能识别到了下载行为，但是下载的文件因为自身是静态分析对抗性样本，这会导致反病毒软件可能认为下载的文件是无恶意的文件。且木马下载器自身几乎不带有主动破坏计算机的恶意行为，只作为下载和启动其它木马的桥梁。此时，反病毒软件更多情况是认为木马下载器类似Motrix等的小型下载工具或者是正常程序的补丁更新，同时因为下载产生的数据流量不会过大，难以被判断为网络攻击。此时，相当于木马下载器完成了植入阶段的过程，在受害者的计算机上从黑客的挂马服务器上成功完成了下载木马程序。随后，将可能进入孵化期，上线期，攻击期。

[2] J. Xie, S. Li, X. chun Yun, Y. Zhang, P. Chang

Hstf-model: an http-based Trojan detection model via the hierarchical spatio-temporal features of traffics

Comput. Secur., 96 (2020), Article 101923

幸运的是，反病毒软件可以根据网站安全可信度，并将某个IP或某个域名及其子域名下的所有内容划分为黑名单，禁止用户和应用程序访问。但不容忽视的是，域名和IP安全性鉴定仍然需要时间。虽然开始时木马下载器变种在植入阶段选择下载能逃逸反病毒软件静态分析的对抗型木马，避免自身因为下载恶意程序被反病毒软件直接判断为恶意软件，由此逃逸反病毒软件的检测，但反病毒软件随后仍然可以通过判断高风险http请求拦截木马下载器的危险请求，相比之下，基于静态分析对抗性的木马下载器变种危害性较小且反制难度小于感染型病毒变种对抗性样本。但不可忽视的是，黑客可以通过上述的变种挂马服务器，导致每个被木马感染的计算机的木马都是具有不同的Hash值且产生变种木马的原始木马样本被经过修改的操作组合不同，仅仅对单个原始木马程序经过前文中10种行为（Action）和16次step就可能存在10^10以上量级的变种，原始木马程序庞大的变种可以达到干扰甚至摧毁云查杀威胁情报的共享机制。

因此，在特定的条件下，如过度依赖本地静态检测忽视动态上传上报可疑文件、过度依赖云查杀的动态检测沙箱和云端病毒库遭遇网络中断、反病毒软件无条件直接信任某些公司签发的安全证书，很大可能会导致静态规避攻击绕过杀毒软件的防御，构成实际威胁。

**问题五：本实验的对抗型样本生成和传统人工对抗型样本生成的区别：**

传统的对抗性样本需要恶意攻击者手动修改自身的恶意程序二进制代码，繁琐低效，而且可能因为修改失误导致恶意程序受损，产生对抗型样本的数量少，基于人工手动修改的对抗性样本难以对抗反病毒软件的云查杀，人工手动修改的变种终究只有少量。而本实验基于强化学习的Sarsa模型自动生成对抗型样本，对于单个恶意程序，能产生的对抗性样本数量较大，当变种数量足够多时，在有限的时间内，反病毒软件的云查杀系统未必能全部识别出。大量的对抗型样本对云查杀的压力也是较大的。

**对研究人员的建议：**

本研究使用对抗样本对采用机器学习技术的病毒查杀引擎进行黑盒测试，但需要注意，对某些特征的修改和删除，是否是修改容易被规避的脆弱特征、容易损坏原有恶意文件恶意行为的特征，例如机器码，本研究中尽管代码中保留了修改机器码的函数，但最终并未将其封装为操作函数，而是将其弃用，因为修改机器码在测试中并未通过，测试时使用了一个输出HelloWorld的C语言机器码为X86-64 PE可执行程序，在修改机器码后，导致代码段重定位失败，程序无法执行、以及一些容易导致程序被直接查杀的特征，例如UPX加壳，在测试中，只采取UPX加壳，QVM引擎会直接判定该程序是恶意程序，甚至被UPX加壳的文件只是毫无危害的HelloWorld C语言程序。而采用UPX加壳以及多种方式混合，QVM引擎认为该程序是恶意程序的概率会大幅下降。同时，证明修改后的样本对计算机仍然有危害是至关重要的，但需要注意测试的安全问题，因为几乎没有一种杀毒软件能100%的概率查杀静态处理后的样本，即使在测试中使用虚拟机，仍然可能会存在虚拟机逃逸问题，例如VMware WorkStation的漏洞CVE-2023-34048和CVE-2023-34056，测试前尽量对重要数据进行备份以及设置虚拟机还原点，防止运行恶意程序危害到实体机的意外发生。

**对反病毒软件厂商的建议：**

本研究表明静态分析器容易被攻破，虽然目前没有出现前文中提及到的感染型病毒变种对抗性样本这种对静态分析针对性逃逸的流行性恶意软件，但规避静态分析的研究仍然具有价值，这也是为什么EMBER，MalConv等静态分析模型日益被关注的原因，因为静态检测属于对恶意软件的第一道防线，保证能在用户执行恶意软件前就识别出恶意软件并且清除。

MAB-Malware: A Reinforcement Learning Framework for Attacking Static Malware Classifiers

arXiv:2003.03100v3 [cs.CR] 29 Apr 2021

此外尽量不要过度依赖于本地病毒库的Hash值匹配病毒扫描策略，因为静态规避样本相比原样本Hash值不再相同，且达成静态检测规避的处理步骤不同，也会使不同的静态规避样本的Hash值不相同。同时，对某些异常PE可执行文件，例如尾部具有过长的字符串，可以考虑将它们判断为代码混淆器，疑似风险、不受欢迎文件提示给用户，并且上传到反病毒软件的云安全中心进行分析，而不是直接认为是无危害文件，尽管这样来说会有误报的可能，但是可以对抗一部分的规避静态分析的对抗性样本。例如Clam AV反病毒软件中允许开启的检测潜在不受欢迎程序（Potentially Unwanted Applications PUA）选项，虽然会误报一些用于计算机管理员监控计算机状态的网络嗅探、筛选工具和产生大量流量使用P2P网络的下载器程序，以及一些远程登录程序，JavaScript、ActiveX等类似语言编写的问题脚本、PC Hunter这种加载了驱动用于杀死进程的通用系统工具，但能检测出更多潜在恶意程序。此外，需要注意加强反病毒软件的自我保护功能和动态行为监测机制。同时，反病毒软件厂商也需要注意静态病毒库和动态病毒库的存放问题，本地可以尽量少存放静态病毒库，只存放一些近期流行的恶意程序样本，并且一段时间进行本地病毒库更新，这样可以为用户节省磁盘空间；而动态高风险动态行为检测库最好存放在本地，这样即使在网络环境不佳或者是离线状态，仍然能拦截病毒的运行，防止病毒危害计算机和造成用户损失。同时，云端病毒库可以对某些Hash值或者携带某些有效签名的程序的动态高风险行为设置白名单，例如一些签名注册工具。上述操作在网络正常的情况下，可以在经过本地高风险行为检测后，再进行云端查询判断程序是否是恶意程序，尽管会有离线模式下无法连接到云端病毒库导致误报某些非恶意软件的敏感操作的问题，但是可以很大程度上规避恶意程序运行造成的损失。在本实验中使用的Clam AV反病毒软件是基于本地离线病毒库，需要手动更新，大约有630MB大小。

**动态检测对于静态对抗性样本的缓解作用：**

尽管本实验的方案无法绕过动态检测，但静态规避研究仍具有重要价值，可以为用户执行潜在的恶意软件提供预判防御。动态规避属于另一个研究方向，目前shellcode的逃逸检测上已经出现了动态规避，通过异或加密shellcode后封装在C/Python代码的数组中，在执行时解密即可实现动态规避。然而动态检测相比静态检测，对某些恶意软件的检测效果并不佳，因为有些恶意软件会在运行时试图破坏反病毒软件，例如关闭相应的防御服务，加载驱动让自己常驻内存来对抗反病毒软件等操作，恶意软件针对反病毒软件的破坏行为需要反病毒软件的一些自我保护机制来应对，例如采用Intel-VT虚拟化技术（目前已被360安全卫士采用），HVCI内存防护技术。但遗憾的是，反病毒软件使用Intel-VT虚拟化技术会对用户运行VMware workstation、Virtual Box等虚拟机软件的流畅造成影响。同时，反病毒软件仍然需要考虑到许多服务器和个人电脑使用的老旧计算机硬件和操作系统官方宣布已经过时不再支持维护的操作系统，例如Windows2000、Windows XP操作系统，以及过时的Windows server操作系统Windows Server 2003，这些CPU和操作系统，由于发行时间较早，当时没有考虑到后续2005年之后Intel和AMD硬件厂商推出的硬件虚拟化技术，因此反病毒软件通过虚拟化技术的自我保护在这些过时系统上无法成功实现。尽管这些过时不再被微软支持维护的Windows操作系统的使用者在逐渐减少，然而，对于恶意攻击者而言，攻击这些微软不再提供安全更新和漏洞修复的系统相对于攻击Windows10和Windows11这种仍然被长期支持的操作系统而言更容易，也是部分恶意攻击者更加青睐的目标，因为目前几乎没有第三方机构愿意为这些过时的操作系统提供安全更新服务，这些过时的操作系统一旦被恶意攻击者发现漏洞并且针对漏洞编写木马程序，后果相比仍然被支持提供安全更新的系统是灾难性的。不能按照常理认为更多用户选用被支持提供安全更新的系统，因此恶意攻击者更倾向于攻击用户数更多的操作系统，事实上，仍有大量计算机由于数据迁移困难，硬件不合格，软件和驱动的支持问题，没有淘汰掉过时的操作系统，而这些操作系统有可能是木马和勒索病毒利用Windows通用漏洞攻击下发生的重灾区，在过去几年前的永恒之蓝漏洞就被勒索病毒WannaCry利用445端口针对几乎所有Windows操作系统的攻击。

因此，尽管动态检测对于静态规避类对抗型恶意程序样本对于计算机的攻击确实具有一定的阻止作用，且在阻止对抗恶意程序的运行时的恶意行为起到了重大作用，但仍然可能因为恶意程序对反病毒软件的破坏导致反病毒软件的动态检测丧失功能。且因为防御生态的割裂性，杀毒软件的自我保护功能在某些过时操作系统能否有效仍然是不确定的。

**本实验中一些存在的不足：**

由于本实验使用的机器性能问题，为避免程序崩溃，Sarsa算法允许的最大步数上限不能调的很高，导致可能存在一些样本，这些样本再经过一些操作也能达到规避静态检测，但智能体在指定步数之内会判定扰动产生的对抗性样本无效。

其次，笔者在寻找加壳工具时，发现很多加壳工具难以实现自动化操作，许多闭源的加壳工具作者都会使用C语言和汇编来编写加壳工具，而本项目使用的是Python语言，导致调用函数来使用这些闭源加壳工具对样本进行处理几乎不可能，尤其是对于具有图形化界面的闭源加壳工具，难以自动化调用。

本实验中除了Clam AV能使用Python封装成检测恶意程序，其他反病毒软件并未向用户公布其可查杀接口，例如360杀毒，只有360杀毒企业版才可以通过提供的查杀接口来判断单个文件是否是恶意程序。

本实验只针对反病毒软件只采用了黑盒攻击，因为未知反病毒软件模型的内部信息以及内部扫描机制，只能通过反病毒软件的扫描结果获取恶意程序是否规避了反病毒软件的检测。

最后，本实验尚未考虑到Windows下其它恶意程序的对抗性样本生成，例如恶意shellcode，Java恶意程序，ActiveX恶意脚本程序，恶意JavaScript代码等。大多数恶意软件的传播是通过网络，对于含有JavaScript恶意代码的挂马网站等自动下载木马程序。目前通过移动设备，例如U盘，光盘等传播的恶意软件已经相比过去减少。但是黑客基于网络发动的攻击仍然频发，因此在未来，研究网页内恶意JavaScript的对抗性样本生成也是具有必要的，这可以帮助浏览器提醒用户更谨慎地访问这些网页，或是使用一个浏览器沙箱以规避恶意JavaScript对用户的计算机造成损害。

**总结：**

在本实验中，笔者使用了Sarsa算法构建了一个强化学习的静态查杀规避处理框架，用于生成对一些黑盒的反病毒软件和反病毒引擎的对抗性扰动静态规避攻击样本，动作选择可以认为是多臂老虎机问题，在有限的尝试次数下探索较好的行为，尽量规避无效动作，降低修改总次数来实现规避概率最大化。实验表明，处理前后在Clam AV和火绒杀毒下检出率变化较小，这可能是因为这两款杀毒软件自带一定的抗扰动能力，能对抗本实验中对于恶意软件的针对静态检测的扰动处理，然而360杀毒的云查杀引擎在没有云上传功能下，对处理后的恶意软件样本查杀率很低。实验表明了，在采用相似检测机制的反病毒软件之间，该模型具有可迁移攻击特性，且能推测出QVM反病毒引擎的抗扰动能力弱于Clam AV，说明部分基于机器学习的反病毒引擎仍然有待升级。本实验的对抗性样本生成强化学习框架在基于原有的GYM-Malware，PSP-Malware，Mab-Malware基础上，尝试了新的强化学习算法，并且降低了模块之间的耦合性，便于后续的研究人员利用已有的研究成果创建新的强化学习框架。