arXiv:2302.03768v1 [cs.AI] 7 Feb 2023

猫鼠游戏：运用Q学习算法提升网络安全领域中的对手能力

Arti Bandhana[1], Ondˇrej Lukáˇs[1] b,Sebastian Garcia[2] and Tomáˇs Kroupa[2] 捷克理工大学 *bandhart@fel.cvut.cz, lukasond@fel.cvut.cz, sebastian.garcia@agents.fel.cvut.cz, tomas.kroupa@fel.cvut.cz*

关键词: Q学习、强化学习、马尔可夫决策过程、网络安全、学习智能体、高级持续性威胁。

摘要: 网络攻击持续增加，而网络安全领域缺乏能够应对这些攻击的专业人才，这表明我们需要性能良好的自动化工具来检测攻击。攻击者会伪装其动作，并发动由多个动作构成、难以检测的攻击。因此，要改进防御工具，就需要使用训练有素的攻击者对其进行校准。本文提出了一种攻击智能体与环境模型，并使用基础Q学习、朴素Q学习和双重Q学习（均为Q学习的变体）评估其性能。攻击智能体接受训练，目标是窃取数据；与此同时，网络中的所有主机都具有非零的检测概率。结果表明，双重Q学习智能体的总体性能最佳，在70%的交互中成功实现了目标。

1 引言

网络攻击的风险正在不断增加。攻击者日益变得复杂老练，并设法发现新的漏洞加以利用，使网络防御者的角色变得更加偏斜和不对称。大多数攻击技术几乎不涉及攻击者与防御者之间的直接交互。在勒索软件（ENISA，2022年）、端口扫描或加密货币挖矿等攻击中，交互可能少到攻击者仅需执行一次操作。在网银木马或高级持续性威胁（APT）攻击（Draˇsar 等，2020年）等更复杂的攻击中，攻击者必须在网络或目标设备内执行一系列步骤，才能在不被发现的情况下取得成功。此类攻击极难检测，但其影响也最为严重。高级持续性威胁攻击通常持续时间较长，其中许多决策通常由人类作出，人类会调整其战术和技术以避免被检测；而在大多数情况下，防御机制并不够灵活，无法适应攻击者的行为。

APT 攻击者可以建模为在与环境交互的同时追求自身目标的智能体（目标设备或网络）。这些交互大多通过博弈论或强化学习（RL）模型进行刻画，旨在提升网络防御能力。博弈论框架用于为最优防御提供解决方案（例如蜜罐分配），而强化学习模型主要用于改进渗透测试攻击（Durkota 等人，2016年；Mitchell 和 Healy，2018年）。LSTM 网络和 Q学习技术也被用于预测 APT 数据集中的攻击者动作（Dehghan 等人，2022年）。然而，对现实防御进行建模不可避免地需要学习攻击者的近似最优决策。据我们所知，目前尚无研究以改进攻击者决策为目标，对 APT 攻击者的行为进行建模。为攻击者创建现实的推理模型，需要考虑攻击者的意图、能力、目标、机会和可用资源等因素（Moskal 等人，2018年；Liu 等人，2005年）。由于这些属性十分复杂，开发通用框架变得极具挑战性。为应对这些挑战，强化学习模型通常从防御者的角度进行训练，以求解最优策略；然而，据我们所知，目前尚无用于优化 APT 攻击者动作的强化学习模型。

本文对 APT 攻击者和网络环境进行建模，以训练能够优化攻击的强化学习智能体。攻击者的目标是窃取

a https://orcid.org/0000-0002-3711-3645 b
https://orcid.org/0000-0002-7922-8301 c
https://orcid.org/0000-0001-6238-9910 d
https://orcid.org/0000-0003-1531-2990

数据，将其从本地网络中的特定服务器外传至互联网上的命令与控制（C&C）服务器。为找到攻击者的最优策略，我们训练了三种离策略强化学习算法：Q学习、朴素Q学习和双重Q学习。

我们的结果表明，基于双重Q学习的攻击者智能体能够在近70%的交互中窃取数据。此外，我们还表明，该智能体能够学会规划并执行多阶段数据窃取攻击，而这种攻击被检测到的比例不到40%。从网络安全的角度来看，这意味着可以学习并改进攻击者模型，因此，未来的研究可以进一步学习更好的防御者模型。

本文的主要贡献如下：

- 在对抗性环境中，提出了一种决策实体（APT 攻击者）的新模型；
- 在定制环境中，为攻击智能体实现强化学习算法；以及
- 分析 APT 攻击者模型对网络安全领域的影响。

本文结构如下。第2节介绍研究动机和前人工作。第3节描述强化学习环境。第4节介绍强化学习算法；第5节介绍实验设置；第6节展示结果并讨论其影响。结论和未来工作见第7节。

2 动机与相关工作

研究本地网络中高级持续性威胁攻击的攻击者行为模型，主要有两个动机。首先，基于过去攻击的知识改进防御机制（算法、防病毒系统等），凸显了深入了解真实网络中近乎最优攻击行为特征的必要性。其次，通过创建并训练攻击者行为的强化学习模型，可以优化未来的防御机制以及此类系统的动态特性。

多年来，博弈论和强化学习（Shiva等人，2010年）在建模众多领域中的攻击与防御机制方面日益受到重视，其中包括网络安全。网络安全问题通常十分复杂，需要理性决策。博弈论为多个决策者（玩家或智能体）之间的策略互动提供数学模型，并提供用于求解此类场景中解（均衡）的算法。将博弈论应用于网络安全的潜在益处在于自动化网络管理员所执行的全面威胁检测流程。然而，现实世界中的网络安全模型在玩家所能观测到的信息方面可能存在局限。通常，防御者对攻击者策略和决策的了解有限（Patil等人，2018年）。这会导致出现部分观测或信息不完备的博弈，而这类博弈极难扩展到问题所需的规模。

在安全博弈论领域，蜜罐技术方面已有颇具前景的研究（Anwar和Kamhoua，2022年）。作者通过构建防御者与攻击者之间的双人零和博弈，设计了一种蜜罐分配的最优方法，该博弈以攻击图为基础展开。防御者在机器上部署蜜罐，而攻击者则选择一条穿过攻击图的攻击路径，试图在不被发现的情况下抵达目标机器。除了为网络中的蜜罐部署制定有效策略外，作者还对多种蜜罐配置进行了实验。采用多样化的蜜罐配置，可以确保即使一个蜜罐遭到攻陷，也不会导致所有蜜罐都被发现；但这会增加运营成本。为实现对网络攻击的自动响应，Hammar和Stadler（2020年）研究了无需人工干预、也不需要领域知识的策略演化方法。作者将网络交互建模为马尔可夫博弈，并通过自博弈模拟，让智能体彼此交互，根据先前对局中的经验更新策略。

另一项很有前景的研究方向采用近端策略优化（PPO）结合自博弈，来解决防御者与攻击者轮流行动的随机（马尔可夫）双人博弈（Du 等，2022）。该博弈在攻击图上进行，作者表明，PPO 策略的性能优于启发式策略。初步结果颇具前景，但作者所采用的场景仅限于包含五个节点和四条边的攻击图。相比之下，我们的工作只涉及单智能体环境。

攻击图很有帮助，因为它们能够根据网络中存在的漏洞预测攻击者的行动路径。同时，防御者可以利用攻击图找到有效的防御策略。具体而言，Guo 等，2021 通过在活动目录中构建的攻击图进行边阻断，提供了防御方案。另一类研究关注如何借助攻击工具改进渗透测试或网络安全训练，例如使用深度 Q 学习（Niculae 等，2020）。作者比较了 Q学习，

扩展分类器系统（XCS）和深度 Q 网络（DQN），以寻找攻击者策略。为确定对网络中可疑用户的最佳响应，Chung 等，2016 将不同版本的 Q学习与随机博弈进行了比较。

3 环境模型

Q学习是应用最广泛的无模型离策略强化学习算法之一（Jang 等人，2019a）。它使智能体能够在具有马尔可夫性质的领域中学习，因此可将这类领域建模为马尔可夫决策过程（MDP）。通过 ε- 贪婪策略对环境进行充分探索。 ε- 贪婪方法以概率 ε 选择均匀随机操作，以概率 1 – ε选择贪婪操作。超参数 ε 用于平衡探索与利用，旨在最大化累积奖励。

由于重点是训练单个攻击智能体，因此采用MDP作为底层模型（Sutton 和 Barto，2018年）。这种方法使防御者成为环境的一部分。在现实场景中，成功检测需要多个步骤，从部署防御措施、检测并生成警报，到评估和应对威胁。在本研究中，防御者被建模为环境中随机且全局的一部分。

3.1 网络

用于定义环境的计算机网络代表一个小型组织，其中有五台客户端、五台服务器和一台提供互联网连接的路由器；见图1。网络中的每台主机都可以访问互联网。该路由器同时也是防火墙，用于控制子网 2 中的哪些客户端可以访问子网 1 中的服务器（对应图1中的虚线）。如果计算机位于同一子网中，就可以相互连接。

在该环境中，我们假设攻击者已经获得了网络中某台客户端的访问权限。此外，攻击者知道互联网上某台外部命令与控制服务器的地址。攻击者的目标是找到并窃取位于子网 1 中某台服务器上的数据。

3.2 防御者

在我们的模型中，防御者是一个同时存在于所有客户端/服务器上的实体，并被分配了一个检测攻击者操作的概率。一旦

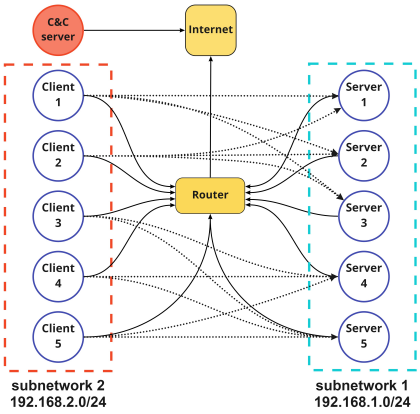

图1：包含两个本地子网和一台位于互联网上的命令与控制服务器的网络拓扑。黑色实线表示直接的网络连接（例如以太网线缆）。虚线表示防火墙允许客户端与服务器之间建立的逻辑连接。在非随机化实验中，攻击者从客户端 1 开始。在随机起始位置的实验中，攻击者从子网 2 中的一台客户端开始。攻击者始终知道命令与控制服务器的 IP 地址。

攻击者被检测到，回合便会结束，环境也会重置为初始状态。这在环境中通过一个终止状态表示。鉴于防御者拥有完整的网络可见性，所有客户端和服务器上的每项操作都对应一个检测概率。

3.3 攻击者

攻击者通常不了解网络情况，因此必须通过反复试错来弥补知识不足。根据我们的假设，我们模拟一名已在子网 1（图1）站稳脚跟的攻击者。这符合现实场景，因为网络上有许多联网设备，初始入侵可以通过多种方式实现，而从某些方面来说，要阻止初始入侵极其困难。因此，我们当前的设置中忽略了对攻击者入侵的建模。攻击者的目标是找到通往子网 2 中服务器的最优路径，该服务器包含敏感数据；然后找到并窃取这些数据，并使其可在网络上访问。可执行的动作是完成目标所需的最少操作：查找主机、查找服务、获取访问权限、查找数据和窃取数据。我们将攻击者建模为采取最优行为的理性攻击者。

3.4 状态

状态是从攻击者视角对环境进行的抽象表示。它包含若干资产，攻击者可以利用这些资产，或通过先前的动作发现它们。

因此，环境的状态会根据攻击者的动作和当前状态而变化。概率$p(a|s)$表示攻击者在状态s中执行a操作成功的概率，而$p(\text{检测}|s,a)$表示在状态s下执行a操作时被检测到的概率。这些成功概率和检测概率（表 1）是根据渗透测试专业人员的专家评估设定的：将领域知识与各种检测工具对恶意行为的发现能力评估进行比较并匹配，相关评估见（Huang 等，2022）。

表 1：攻击者在网络中执行的每项操作的成功概率和检测概率。

操作	成功概率	检测概率
扫描网络	0.9	0.2
查找服务	0.9	0.3
在服务中执行代码	0.7	0.4
FindData	0.8	0.1
数据外泄	0.8	0.1

成功概率基于已知的工具和技术。虽然网络问题是大多数动作失败的原因，但对于执行代码-在服务中，还必须考虑服务版本和漏洞利用程序质量等其他问题。检测概率考虑了人工操作者在真实网络的良性流量中发现的误报。有些动作（例如使用 ARP 扫描的扫描网络）成功率很高，且几乎不会被识别（现成的最先进的入侵检测系统无法检测到它（Hou 等，2010））。通常，即使检测到这些扫描，为了限制误报，这些警报也会被忽略。查找数据也是如此，因为它是在本地执行的，因而几乎无法检测；而数据外泄只要操作得当，就极难与良性流量区分。

在每个时间步，以下信息都属于状态：

- 攻击者已发现的网络集合；
- 攻击者已发现的主机集合；
- 攻击者已控制的主机集合；
- 攻击者在每台主机上已发现的服务集合；以及
- 攻击者在主机上已发现的数据集合。

由于状态由资产构成，我们可以采用最初为规划设计的著名 STRIPS 表示法（Fikes 和 Nilsson，1971年）。STRIPS 将系统中的转移描述为操作符，当前置条件 满足时，操作符即可应用。最初，每个操作符都可以指定添加 和删除 效果。不过，在我们的方法中，我们完全省略了删除效果，从而得到松弛 的问题表示（Bonet 和 Geffner，2001）。问题松弛是人工智能各个领域中常用的方法。这种方法简化了状态空间的遍历问题。

3.5 动作

攻击者的动作遵循 Mitre ATT&CK[1] 中列出的对手行为技术子集。由于我们在此模型中仅表示一种目标——数据窃取，因此仅使用与数据窃取相关的 Mitre 动作子集：

1. 主动扫描：
   (a) 查找网络中的计算机 (b) 查找网络中主机上运行的服务 (c) 查找计算机中的数据
2. 攻击服务以执行代码；以及
3. 将数据外泄至互联网。

我们模型中的攻击者遵循表 2 所示的五步操作，以达成其目标。

表 2：导致状态变化的网络动作及其效果列表

操作	描述	前置条件	效果
扫描网络	扫描完成 给定网络的范围	网络 +掩码扩展'已知主机'	-
查找服务	扫描给定主机 针对运行中的服务	主机 IP	扩展"已知服务" 包含主机-服务对
ExecudeCode InService	在服务中运行漏洞利用 以控制一台主机	主机:服务	扩展"受控主机" -
FindData	运行代码以发现 受控主机中的数据	主机 IP	扩展"已知数据"，加入 主机:数据对
数据外泄	将数据从一个 从一个受控主机到另一个	主机:数据:主机	扩展"已知数据" 使用"目标:数据"

3.6 奖励

奖励是智能体针对状态-动作对获得的一种激励。在我们的模型中，智能体的奖励设定如下：-1每执行一次操作，-50如果操作被检测到，以及$+100$如果达到目标状态。

每次操作给予少量负奖励，旨在促使智能体寻找通往目标的最短路径。$+100$实现目标所获得的奖励，使攻击者能够采取预期检测概率较高的动作，前提是这些动作能带来更高的预期奖励。

[1] https://attack.mitre.org/

3.7 实现

如第 3.4 节所述，状态的表示方式使我们无需从头重新训练智能体即可修改环境，而不必从头重新训练。这使我们的环境模型有别于其他模型，并为各种网络安全场景提供了更高程度的模块化。我们的智能体不会在训练前预先分配完整的 Q表，而是在训练过程中动态创建 Q 值，从而节省内存和时间。

4 学习智能体

为了训练并评估攻击者的性能，我们使用 Q学习（Jang 等人，2019a）及其变体：朴素Q学习和双重Q学习。Q学习是一种强化学习算法，无论所遵循的策略如何，都能近似最优状态-动作价值函数。它是一种离策略算法，通过更新 Q 值 $Q(s,\ a)$，将学习过程与当前执行策略分离；该值表示某个状态-动作对的优劣。Q 值更新公式如下：

$$Q(s,a) \coloneqq Q(s,a) + \alpha(R_{t+1} + \gamma V^t(s')), \quad (1)$$

其中，$\alpha \in [0,1]$ 是学习率，$\gamma \in [0,1]$ 是折扣因子，体现了折旧的概念。数值越接近 0，表示越倾向于当前奖励而非未来奖励。

在朴素Q学习中，部分学习率会分配给先前结果，以便在学习过程中结合过往历史知识、当前迭代中的实际即时奖励以及预期的未来奖励（ Chung 等人，2016年）。由此得到方程（1）的以下变体：

$$Q(s,a) \coloneqq \alpha Q(s,a) + (1-\alpha)(R_{t+1} + \gamma V^t(s')) \quad (2)$$

双重Q学习（Hasselt，2010年；Jang 等人，2019b年）提出学习两个 Q 函数，而不是一个。每个 Q 函数都使用另一个 Q 函数对下一个状态进行更新。这两个 Q 函数对动作价值的估计都是无偏的。随后，对每个动作的两个 Q 值求平均或相加，再根据得到的 Q 值执行ε-贪心动作选择。本文在执行 ε-贪心策略之前，先将两个 Q 值相加。

$$Q^A(s,a) \coloneqq Q^A(s,a) + \alpha(R + \gamma Q^B(s',a') - Q^A(s,a)) \quad (3)$$

$$Q^B(s,a) \coloneqq Q^B(s,a) + \alpha(R + \gamma Q^A(s',a') - Q^B(s,a)) \quad (4)$$

另外两个学习智能体也按照原始论文的做法，使用 ε-贪心作为动作选择准则。

5 实验设置

使用了三种不同的场景来训练学习智能体：攻击者位置固定、攻击者位置随机，以及随机选择要攻击的目标服务器。

在第一个场景中，攻击者被放置在子网 2 的客户端 1 上（图1）。我们将客户端定义为网络上用于工作的正式设备，将服务器定义为存储数据并提供客户端可访问服务的设备。攻击者的目标是到达目标服务器（指定为子网 1 中的服务器 3），并将目标服务器上的数据窃取到本地网络之外的命令与控制服务器。

子网 2 中有五个客户端；实际上，网络中的任何联网设备都容易受到攻击。因此，在第二个场景中，我们随机分配攻击者的起始位置。这样做是为了比较学习智能体的性能，并观察它们如何适应起始位置的随机性。除了随机化起始位置，我们还随机选择了用于数据窃取的目标服务器，这构成了第三个场景。

要成功达到目标状态，在全部 3 个场景中都必须至少执行 5 次成功操作；但是，如果智能体在每个回合中超过 25 次操作的上限，交互就会终止。

在全部 3 个场景中，防御者都是一个拥有无限可见性的实体，并且存在于所有主机上；也就是说，每次操作都可能以预先设定的概率被检测到。此外，我们假设主机上运行的所有服务都可被利用，并且所有主机都可以连接互联网。

各算法的学习参数如表 3 所示。实验从一个随机攻击者开始，该攻击者会随机选择一项操作。Q学习智能体和双重Q学习智能体的训练学习率为 0.3，而朴素Q学习智能体的训练学习率为 0.8。所有智能体的 epsilon 控制的动作选择参数均保持为 0.2；不过，双重Q学习使用了从 0.2 线性递减至 0.05 的 ε 。

在所有实验中，我们测量了回合的胜率、检测率和平均回报。胜率表示攻击者成功完成交互的百分比，即

在 10000 个回合中，攻击者成功到达目标状态并窃取数据的次数所占的比例。检测率表示被检测到的交互所占的比例，这些交互会导致攻击者获得 −50奖励。

表 3：训练参数：Q学习和 DoubleQ 学习智能体经过 10000 个回合的训练，而 NaiveQ-Learning 经过 5000 个回合的训练。所有学习智能体的折扣因子 γ 均设为 0.9。

算法	α	ε	γ	回合数个
随机	-	-	-	-
Q学习	0.3	0.2	0.9	10 000
朴素Q学习	0.8	0.2	0.9	5 000
双重Q学习	0.3	0.2	0.9	10 000

6 实验结果

以下结果是在实验的第一种场景中获得的，此时攻击者在网络中的位置已指定。表4总结了不同学习智能体的性能。随机攻击者既不了解网络，也没有任何策略，其检测率为99.58%；而双重Q学习攻击者的检测率为33%。Q学习和朴素Q学习智能体具有相近的检测率。

表4：攻击者从客户端*1*开始且位置固定时学习智能体的性能比较。Q学习和双重Q学习训练了10000个回合，而朴素Q学习训练了5000个回合。

算法	获胜率（%）	检测比率（%）	Mean回报
随机	0.48	99.58	53.03
Q学习	66.4	40.4	43.94
朴素Q学习	66.91	40.19	43.94
双重Q学习	74.0	33.0	54.61

表5：在攻击者随机化起始位置的场景中，学习智能体性能的比较。Q学习和双重Q学习训练了 10 000 回合，而朴素Q学习训练了 5 000 回合。

算法	获胜率 (%)	检测率 (%)	Mean返回
随机	0.34	99.48	-54.04
Q学习	65.4	39.27	41.97
朴素Q学习	54.27	50.78	25.59
双重Q学习	68.9	36.8	47.58

表6：攻击者的起始点和目标服务器均随机化场景下学习智能体的比较。所有算法均经过10000个回合的训练。

算法	获胜率 (%)	检测率 (%)	Mean返回
Q学习	53.3	53	23.45
朴素Q学习	61.8	44.1	36.8
双重Q学习	64.9	41.7	41.2

如表5所示，随机化起始位置会降低所有学习智能体的胜率，并提高检测率。在所有学习智能体中，由于起始位置的随机性，朴素Q学习智能体的性能受到的影响最大。检测率从 40.4% 上升至 50.78%

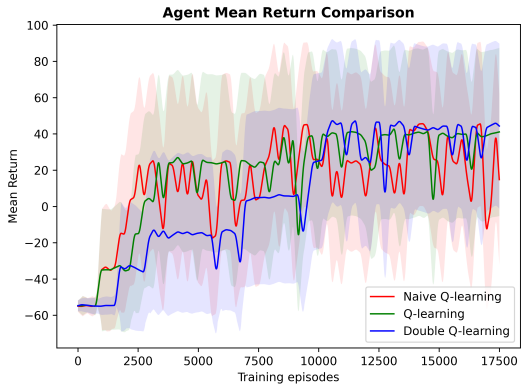

图2：在存在防御者且攻击者起始位置随机化的场景中，智能体在学习过程中的平均累积奖励比较。

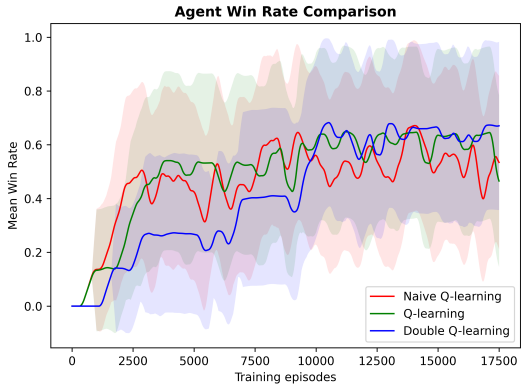

图3：在存在防御者且起始位置随机化的场景中，智能体在学习过程中的获胜率比较。

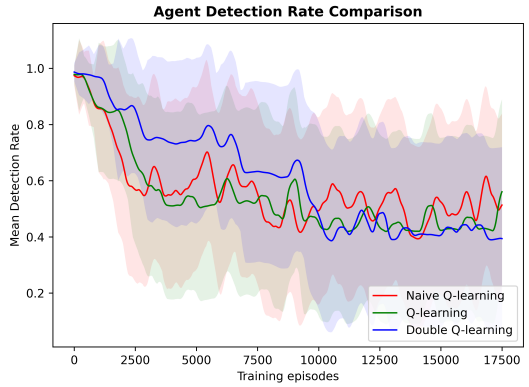

图 4：在防御者在场且起始位置随机的场景中，智能体在学习过程中的检测率对比。

6.1 结果分析

我们比较了在防御者具有完全可见性的情况下，不同参数的智能体如何在拥有十台主机的网络中学习策略。所有客户端和服务器上的动作，其被检测的概率都不为零。在比较所有学习智能体的胜率和检测率时，显然，双重Q学习在所有场景中的表现都优于其他智能体。此类智能体使用两个Q函数进行训练，但它们来自不同的回合，因此训练更加稳健。在推理过程中，使用两个Q函数的总和。这避免了Q学习的过高估计偏差，并在噪声环境中也能提高训练稳定性。在起始点已指定的第一个场景中，Q学习攻击者与朴素Q学习攻击者的性能相同。这是因为学习率按照公式（1）和（2）进行分配。朴素Q学习使用的学习率为 0.8；相比之下，Q学习使用 0.2 的学习率，得到的结果与之相似。然而，起始位置随机化后，朴素Q学习的性能下降。这归因于Q值更新规则的权重，如公式（2）所示。考虑负奖励时，由于采用了拆分更新 α，这种更新对Q值的影响大于标准Q学习。尽管在正奖励较高的情况下这可能有益，但结果表明，该方法在随机环境中缺乏适应性。

图3和图4显示，在获胜率和检测率方面，双重Q学习优于另外两个智能体。图2所示的平均回报方差较大，这是由第3.6节所述的随机环境和奖励分布造成的。图表还显示，尽管双重Q学习在初期表现不佳，但随着回合数增加、状态-动作值不断更新，其表现最终优于另外两个学习智能体。尤其是，即使智能体的策略达到最优，它也无法影响 −50是否被检测到以及随后获得的奖励。因此，三个智能体的平均回报都呈现类似的高方差，但在侧重于达成目标的指标上差异显著，其中双重Q学习展现出最令人期待的结果。

尽管三个基于 Q学习的智能体都采用了随机探索 ε，但第一和第二个场景的结果表明，对于纯粹执行随机动作的智能体而言，环境和目标并非简单易解；这类智能体成功达到目标的情况不到 1%。因此，在图 2、3 和 4 以及第三个场景中，比较时排除了随机智能体。

实验结果表明，尽管防御者对网络具有完全可见性，理性攻击者仍然能够到达目标并窃取数据。从安全角度来看，这表明需要改进网络中的防御工具，以阻止攻击者在系统中横向移动。

7 结论

本文提出了一种基于 Q学习、能够执行数据窃取的攻击智能体。

我们的结果表明，尽管这三个学习型智能体都能找到有意义的策略，但双重 Q 学习的表现优于其他方法，且训练过程最为稳定。它在 70% 的交互中达成了目标，其中 37% 未被检测到。这表明，即使防御者始终存在，理性的攻击者仍可能达到目标。

初始的成功概率和检测概率是根据专家知识设定的，但我们的结果清楚地表明，防御者的检测能力仍有提升空间。攻击者操作具有较高的成功概率，凸显了建立强健防御机制的必要性，这种机制应能够检测到任何隐蔽攻击者。这为研究和改进攻击者技术、增强网络防御能力奠定了基础。

目前，该方法仅适用于小型或中型网络。尽管交互和世界表征模型可以轻松扩展，以适应规模更大、动作空间更复杂的网络环境，但此类扩展的可扩展性和计算可行性仍有待评估。

因此，未来研究的自然方向

是将我们的方法扩展到更大的环境中。由于这些设置较为复杂，还需要进行后续的可扩展性测试。我们还计划将其他类型的网络攻击纳入 Mitre 分类体系，并将防御者建模为在交互中拥有自己一套动作的理性实体。此外，我们计划在模拟环境中测试智能体的性能。

除了增加环境的复杂性之外，为攻击者设定更复杂的目标也已提上日程，这将导致智能体需要进行更多侦察。

致谢

作者感谢信息学研究中心（CZ.02.1.01/0.0/0.0/16 019/0000765）以及捷克共和国安全研究发展战略支持2019–2025（IMPAKT 1）项目的支持。该项目由捷克共和国内政部资助，项目编号为VJ02010020——AI-Dojo：用于研究和测试人工智能驱动网络安全技术的多智能体测试平台。

参考文献

Anwar, A. H. 和 Kamhoua, C. A.（2022年）。通过蜜罐分配与多样性实施网络欺骗：一种博弈论方法。载于*2022*年*IEEE*第*19*届年度消费者通信与网络会议（*CCNC*）。Bonet, B. 和 Geffner, H.（2001年）。规划即启发式搜索。人工智能，129(1):5–33。Chung, K.、Kamhoua, C. A.、Kwiat, K. A.、Kalbarczyk, Z. T. 和 Iyer, R. K.（2016年）。用于网络安全监控的学习型博弈论。载于*2016*年*IEEE*第*17*届高可信系统工程国际研讨会（*HASE*），第1–8页。Dehghan, M.、Sadeghiyan, B.、Khosravian, E.、Moghaddam, A. S. 和 Nooshi, F.（2022年）。ProAPT：利用深度强化学习预测高级持续性威胁。arXiv:2209.07215 [cs]。Draˇsar, M.、Moskal, S.、Yang, S. 和 Zat'ko, P.（2020年）。由会话级攻击者意图驱动的网络攻击模拟器。载于*2020*年*IEEE/ACM*第*24*届分布式仿真与实时应用国际研讨会（*DS-RT*），第1–9页。Du, Y.、Song, Z.、Milani, S.、Gonzales, C. 和 Fang, F.（2022年）。学习如何参与自适应网络欺骗博弈。载于自主智能体与多智能体系统国际会议（*AAMAS*）第*13*届多智能体系统中的优化与学习研讨会。Durkota, K.、Lisy, V.、Kiekintveld, C.、Bosansky, B. 和 Pechoucek, M.（2016年）。利用攻击图博弈进行网络防御的案例研究。*IEEE*智能系统。ENISA（2022年）。ENISA勒索软件攻击威胁态势。技术报告，ENISA，卢森堡。Fikes, R. E. 和 Nilsson, N. J.（1971年）。STRIPS：将定理证明应用于问题求解的新方法。人工智能，2(3):189–208。Guo, M.、Li, J.、Neumann, A.、Neumann, F. 和 Nguyen, H.（2021年）。用于防御活动目录风格攻击图的实用固定参数算法。Hammar, K. 和 Stadler, R.（2020年）。通过强化学习和自博弈寻找有效的安全策略。载于*2020*年第*16*届网络与服务管理国际会议（*CNSM*）。IEEE。Hasselt, H.（2010年）。双重Q学习。神经信息处理系统进展，23。Hou, X.、Jiang, Z. 和 Tian, X.（2010年）。基于Snort的ARP欺骗检测与防范。载于*2010*年国际计算机应用与系统建模会议（*ICCASM 2010*），第5卷，第V5–137–V5–139页。Huang, Y.-T.、Lin, C. Y.、Guo, Y.-R.、Lo, K.-C.、Sun, Y. S. 和 Chen, M. C.（2022 年）。利用开源情报发现并解读恶意行为。*IEEE*可靠与安全计算汇刊。Jang, B.、Kim, M.、Harerimana, G. 和 Kim, J. W.（2019a）。Q学习算法：全面分类与应用。*IEEEAccess*。Jang, B.、Kim, M.、Harerimana, G. 和 Kim, J. W.（2019b年）。Q学习算法：全面分类与应用。*IEEE Access*。Liu, P.、Zang, W. 和 Yu, M.（2005年）。基于激励的攻击者意图、目标与策略建模和推理。*ACM*信息与系统安全汇刊（*TISSEC*），8(1):78–118。Mitchell, R. 和 Healy, B.（2018年）。计算机网络利用活动的博弈论模型。载于*2018*年*IEEE*第*8*届年度计算与通信研讨会暨会议（*CCWC*）。Moskal, S.、Yang, S. J. 和 Kuhl, M. E.（2018年）。通过攻击场景模拟评估网络威胁：一种集成攻击者与网络建模的方法。《国防建模与仿真杂志》。Niculae, S.、Dichiu, D.、Yang, K. 和 B¨ack, T.（2020年）。利用强化学习实现渗透测试自动化。Patil, A.、Bharath, S. 和 Annigeri, N.（2018年）。博弈论在网络安全系统中的应用：综述。国际应用工程研究杂志，13(17):12987–12990。Shiva, S.、Roy, S. 和 Dasgupta, D.（2010年）。网络安全博弈论。载于第六届网络安全与信息智能研究年度研讨会论文集，第1–4页。Sutton, R. S. 和 Barto, A. G.（2018年）。强化学习：导论。麻省理工学院出版社。