安全保障技術研究推進制度　令和４年度終了課題　終了評価結果

1. 評価対象研究課題

（１）研究課題名：深層強化学習を用いた自律サイバー推論システムの研究

（２）研究代表者：学校法人岩崎学園　情報セキュリティ大学院大学　大塚 玲

（３）研究期間　：令和２年度～令和４年度

2. 終了評価の実施概要

日時　　：令和５年１１月９日

場所　　：ＴＫＰ秋葉原カンファレンスセンター

評価委員：未来工学研究所　理事長、上席研究員／東京大学　名誉教授
平澤　泠（委員長）

元 三菱ケミカルホールディングス 顧問
岩野 和生

理化学研究所　革新知能統合研究センター　副センター長
上田 修功

玉川大学　脳科学研究所　特別研究員
大森 隆司

兵庫県立大学　大学院情報科学研究科　教授
田中 俊昭

千葉商科大学　総合教育センター長、
東工大　名誉教授、筑波大　名誉教授
寺野 隆雄

産業技術総合研究所　人間拡張研究センター・主任研究員
長谷川 良平

（委員長以外は五十音順・敬称略）

3. 研究と成果の概要

研究の概要

本研究は、CTF*を解くための自律サイバー推論システムのプロトタイプを、深層強化学習を用いて構築し、強化学習を用いたセキュリティシステムの有効性を示し、サイバーセキュリティの発展に寄与することを目指した。

特に、より現実的な環境に対応するために、自然言語処理(NLP)技術をシェル(Metasploit Shell)レスポンス文字列の認識に利用することで、人間の利用者と同じインターフェイスで、AI が自律的に CTF 求解過程の状態を認識し、求解のためのコマンド列を生成する深層強化学習モデルの構築を目指した。

*CTF: Capture The Flag の略。コンテスト形式で情報セキュリティ技術者の訓練に一般的に用いられる問題。

成果の概要

深層強化学習に基づく自律サイバー推論システムについて、自然言語処理(NLP)技術をシェル(Metasploit Shell) レスポンスの認識に利用した深層学習により現在の状態を推定し、CTF 問題のフラグ獲得を報酬とする POMDP（部分観測マルコフ決定過程）型の深層強化学習に適用することで、攻撃機序の異なる 6 種類の Unix サーバーへ侵入する CTF 問題に必要な最適戦略を、従来よりもはるかに効率的(少ないエピソード数)に獲得する技術の開発に成功した。

さらに、研究開始時に想定していなかった大規模言語モデル(LLM)の革命的進化が生じたため、大規模言語モデル(LLM)をサイバーセキュリティ分野に応用し、CTF において LLM の活用可能性を探究した。人間と LLM の協調作業での実験ではあったが、ファインチューニングなしで、 PicoCTF2022 の問題において 64 問中 48 問のフラグを獲得するなど、予想外に良い結果が得られたことから、LLM と深層強化学習を組み合わせたサイバーセキュリティ技術は AI for Security の中核技術として、今後、飛躍的に発展する可能性が示唆された。

4. 終了評価の評点

A　十分な成果をあげた。

5. 総合コメント

最近の大きな技術変革（LLM の出現）の流れを取り入れ、情報セキュリティの手段にも活用できることを明らかにしたこと、CTF 求解を関して目標をほぼ達成したことは評価できる。

サイバーセキュリティにかかる具体的なデータを持たない大学の環境では、深層強化学習や LLM の活用に留まらず、独創的に掘り下げた研究など、更なる工夫が望まれた。

LLM の活用に関して、課題を整理・解決するなど、将来の業界の発展に資するデータやアプローチに関するさらなる検討を期待する。

6. 主な個別コメント

- 当初計画していた深層強化学習に基づく CTF 求解技術を開発し、既存のベンチマークテストで有用性を実証している。
- 大規模言語モデルが情報セキュリティシステムの強化に貢献することを示した点が評価できる。
- 複数の自律サイバー推論システムによる敵対的学習への発展なども見込んでお

り、当該分野の発展に貢献し得る成果と言える。

- プロジェクトの途中で大きな技術変革（大規模言語モデル）があり、その流れを取り入れたところは評価できる。一方で、時間的な制約があり、大規模言語モデルを用いた分析について十分な知見が得られなかった。
- 一つの方向性を見出してはいるが、まだまだ突き抜けた研究レベルになっているとはいいがたい。
- LLM の出現という想定外の状況ではあるが、成果としてはやや深みに欠けている。なぜ、という問いにもう少し踏み込んでほしかった。
- 幅広い領域への LLM 適用の可能性の一例と考えられるが、結果は試行にとどまっており、不十分である。
- なぜうまくいくのか、なぜうまくいかないのかなどの掘り下げが行われることで、更なる発展が期待できそうである。
- 内容的には研究成果がまだ主要な業績として確認できておらず、国際的な評価が不明瞭であるので、成果発表を着実なものにしていただきたい。