別紙２

令和４年度　防衛装備庁
安全保障技術研究推進制度

研究成果報告書
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

令和５年５月

情報セキュリティ大学院大学

本報告書は，防衛装備庁の安全保障技術研究推進制度による委託業務として，情報セキュリティ大学院大学が実施した令和４年度「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」の成果を取りまとめたものです．

１．委託業務の目的

１．１　研究課題の最終目標

本研究の目標は，深層強化学習を用いて，CTFを解く自律サイバー推論システムのプロトタイプを構築し，強化学習を用いたセキュリティシステムの有効性を示すことを目指す．SeqGANやAttentionといった技術を利用して，ソフトウェアを入力すると自動でエクスプロイト又はパッチを作成するシステムを構築できる可能性がある．CTFの大会等で優秀な成績を目指して自律サイバー推論システムの有用性等を評価し，サイバーセキュリティの発展に寄与することを目指す．なお，問題の本質を掴むために，本研究ではまず対象との入出力，すなわちシステムからのレスポンス列と，システムへの入力コマンド列を，キーボードから入力可能なテキスト文字列と仮定する．

１．２　最終目標を実現するために克服又は解明すべき要素課題

最終目標に対する要素課題には，以下の(1)～(3)が挙げられる．

(1) 大量パラメータの最適化と大量サンプルの取得

近年のニューラルネットワークの学習においては，膨大な数のパラメータの最適化が必要である．研究過程で生じるネットワーク変更のたびに，学習をスクラッチから繰り返す必要があり，大量パラメータの最適化を高速に実行できる実験環境を整えることが課題である．また，大量パラメータの最適化には大量のサンプルデータが必要であり，サンプルデータの収集も大きな課題である．これらの課題を解決するために，強力なGPUを搭載したサーバー機器やCTFサーバーによる実験環境を構築して高速化・効率化する必要がある．

(2) 深層強化学習手法の解明

囲碁，将棋等の完全情報ゲームと異なり，CTFやサイバー攻撃対処では環境に対するアクションの実行の結果得られるレスポンス情報を手掛かりにシステムの状態を推定し，次のアクションを推定する問題になっている．効率的に最善手を推定でいる深層学習ネットワークの構成方法や，熟練した人間の解法(サンプルデータ)から自然に知識を獲得する手法，サンプルデータなしで自律的に学習する手法などの未解決問題を解明する必要がある．

(3) 性能評価指標の確立

良い深層強化学習モデルに関して，識別問題や回帰問題などの一般的な問題については，多くの分野ですでに確立した指標があるが，サイバーセキュリティにおける強化学習で性能の概念は明確になっていない．セキュリティ性能概念の探索と，深層強化学習モデルとモデルが使用可能なツール群の組み合わせによる複合性能の概念を考案し，確立する必要がある．

１．３　要素課題に対する実施項目及び体制

(1) 深層学習実験環境の強化（要素課題(1)に対応）

深層学習の実験環境強化を初年度に完了する．Docker等のコンテナ技術を活用して複数の研究者による同時並行作業が可能な環境を整える．

(2) CTFサーバーの構築とCTF解法ログの収集（要素課題(1)，(2)に対応）

本研究で用いるデータを収集する．収集するデータは，情報システムに対する攻撃・防御データであり，具体的には，エクスプロイトコードやネットワークトラフィック，パッチコード，トリアージに用いるデータ，コマンドラインシーケンス，システムコールシーケンス，ファイルシステムに対する操作履歴等である．これらは，(N)IDS/IPS/EDRやシステムログ等からはもとより，OSカーネルに手を入れることでカーネル空間の情報も広く取得することを検討している．

(3) ハッキング手法の調査と体系化，ツール化（要素課題(1)，(2)，(3)に対応）
エクスプロイトコードやネットワークトラフィック，パッチコード，トリアージに用いるデータ，コマンドラインシーケンス，システムコールシーケンス，ファイルシステムに対する操作履歴等を分析整理し，深層強化学習で利用可能なツールとして体系化する．また，(N)IDS/IPS/EDRやシステムログ等からはもとより，OSカーネルに手を入れることでカーネル空間の情報も広く取得することを検討する．

(4) 深層強化学習モデルの構築（要素課題(2)に対応）
End-to-endアプローチ，SeqGANアプローチ，転移学習アプローチの3つを起点に，システム状態の理解と適切なコマンド列を生成するモデルを探索する．

(5) 敵対的学習による自律学習手法の検討（要素課題(2)に対応）
当初はサンプルデータ，CTF解法ログ等を大量に収集して学習を試みるが，平行してCTFサーバーやUnixシステムの応答を元に自律学習モデルからの転移学習や，敵対的学習による学習のアプローチを探索する．

(6) 評価手法の検討（要素課題(3)に対応）
自律サイバー推論システムの性能評価指標を検討する．CTF問題を解くシステムの場合には，CTFスコアに加えて，CTFの解に至る前段階での性能評価方法も検討する．さらに，人間の試行錯誤過程を参考にした際の学習時間や性能指標への影響から学習効果や汎化性能を評価する．

(7) プロジェクトの総合的推進
各要素課題に関する研究の進捗を管理し，本委託業務の実施により得られた成果について，国内外の学会等において積極的に発表し，本研究のさらなる進展に努める．

２．研究開始時に設定した研究目標の達成度

我々の知る限り，従来研究は全てサイバー攻撃対応の問題を仮想的なゲームでモデル化し，状態空間や行動空間を有限の狭い空間に限定するものであった．これに対し，本研究ではニューラル自然言語処理技術(GTrXL[1])を応用したPOMDP型[2]の深層強化学習アプローチを探索した．その結果，実際のUnix環境を対象にしたCTF問題において，攻撃機序の異なる6種類のUnixサーバーへの侵入問題を設定し，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．また，一般のCTF問題に対しても，新たに大規模言語モデルに基づく手法を検討し，上位のスコアを達成できることを確認した．これらにより研究開始時に設定した研究目標を達成した．

３．委託業務における研究の方法及び成果

3．1 深層学習実験環境の強化

機械学習計算サーバーの設置を令和3年1月29日に完了し，A100x4基の稼働を開始した．また，

[1] GTrXL(Gated Transformer XL)は，POMDP 型の深層強化学習向けに LSTM の代替として，2019 年に DeepMind の Parisotto らによって提案された Transformer の改良版である．
[2] POMDP (Patial Observation Markov Decision Process)は，部分観測情報に基づき状態を推定することで，次のアクションと遷移先状態を推定する強化学習モデルである．自然言語処理技術に基づいて部分観測情報から状態を推定する必要があるため，状態が環境から与えられる MDP よりも一般に実現が困難である．

AWS SageMakerのV100x8基を利用した．OSにUbuntu Linux 20.04.2 LTSを導入し，DockerでJupyter Labを立ち上げ，複数の独立した環境からGPUを利用できる環境を構築した．

3．2 CTFサーバーの構築とCTF解法ログの収集（要素課題(1)，(2)に対応）

Container 技術を用いて攻撃機序の異なる多数の標的サーバーを効率的に生成するCTFサーバーのプロトタイプCyExec*[関連発表 3件]を構築した．CyExec*は，攻撃シナリオ中の権限昇格などの中間目標となるマイルストーンをノードとし，SQLインジェクションなどの攻撃をエッジとする非巡回有向グラフで攻撃シナリオを表現することにより，最終目標のノードに至る複数のパスをランダムに選択することで，様々な攻撃機序に脆弱性を持つコンテナを生成する機能を有する．この結果，開始ノードから最終目標ノードに至るグラフ上のパスをランダムに選択することで，多様なサーバー侵入関連のCTF問題を自動生成できる．先行研究(SecGen)では，VM(仮想マシン)でCTF問題のランダム化を実現していたが，本研究ではコンテナで実現することにより，CTF問題生成の効率を大幅に改善した．

3．3 ハッキング手法の調査と体系化，ツール化

Metasploitを導入し，5000個超のモジュールを利用する深層強化学習の実験環境を構築した．MetasploitはCWEに登録された脆弱性およびそのエクスプロイトコードをモジュールの形で格納しており，Metasploit Shellを通じて制御できる機能を備えている．モジュールには主に次の4種類が存在し，CTFのペネトレーションテスト問題では，これらを適切に組み合わせて目標サーバーに侵入することが求められる．

(1) Exploitモジュール
既知の脆弱性を利用して侵入するためのモジュール

(2) Payloadモジュール
Exploitが成功した後に実行されるコードであり，標的システム内でのシェル（コマンドラインインターフェイス）の起動や特定のコマンドを実行するためのモジュール

(3) Auxiliaryモジュール
情報収集やスキャン，サービスの中断（DoS攻撃）などのタスクを実行するためのモジュール

(4) Post exploitationモジュール
侵入後に，データの収集または侵入を永続化するためのツール等を含むモジュール

本研究では，これらのモジュールを行動空間に持つ深層強化学習エージェントを試作した．

また，マルウェアの検出に関し，２つの手法を開発した．第一の手法は，バイナリを画像化し，全セクションを含む画像の分類モデルと識別性の高いセクションのみからなる画像の分類モデルを用意し，それらの予測結果を組み合わせることで，大局的特徴量と局所的特徴量の両者を考慮したアーキテクチャを提案した．BIG2015データセットを用いた検証にて，ベースラインと比較して精度向上が見られた．

第二の手法は，関数呼び出しの関係に注目し，分類理由を解釈可能な提供する初のマルウェア分類手法であるFCGATを提案した．FCGATは，自然言語処理技術を適用して関数に含まれる命令の分布に依存した特徴ベクトルを作成し，関数間の呼び出し関係を反映したグラフニューラルネットワークと注意機構を適用し，マルウェア分類性能の向上に重要な関数をAttention Weightで強調することで，識別性能の向上と分類の理由に解釈を与えることに成功した．分類性能はF1-Scoreで98%以上に達し，これまでの類似研究で世界最高の性能を示した．さらに，FCGATで新たに可能になった関数毎のAttention Weightを調べた結果，意外なことに，上位6個（サンプルあたりの平均）の関数に注目するだけで70%の精度をもたらすことが判明した．また，マルウェア分類に寄与するAttention Weightの高い関数の多くは直感的にもマルウェアの特徴を反映していることを確認した．以上から，FCGATは，少数の関数で信頼性の高い説明を解析者に提供することができ，マルウェア解析の効率化，傾向の分析などへの応用が期待できる．

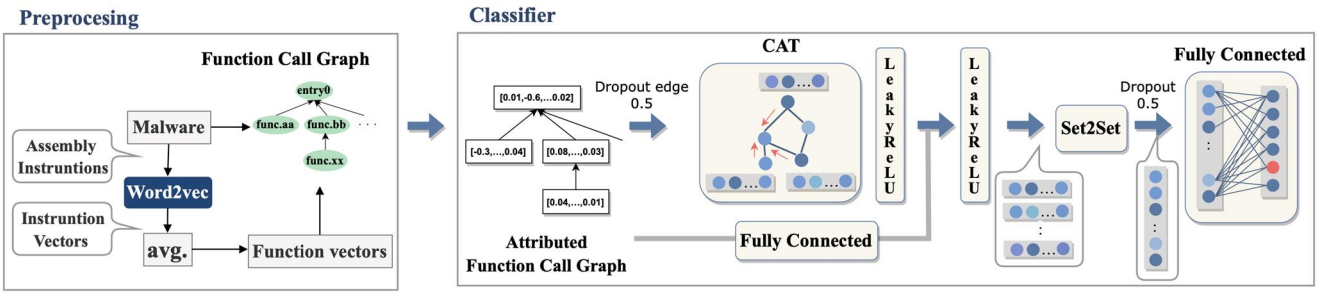

図 1　FCGAT (Function Call Graph and Attention Mechanism)の概要

3. 4 深層強化学習モデルの構築

本研究では，テキストで与えられる観測情報をニューラル自然言語処理(NLP[3])に基づいて状態推定を行うPOMDP型自律サイバー推論システムを構築し，標的Unixサーバー内に設置されたファイルに格納されたFlagを取得することを目標とする自律サイバー推論システムを2種類構築した．

まず，比較対象のベースラインシステム[4]の概要を図に示す．本システムでは部分観測情報(出力結果等)とアクションとして使用できる全コマンド情報をニューラルエージェントの入力として与える．それらの情報はNLPでベクトル空間に埋め込まれ，GRU(Gated Recurrent Unit)を介して，観測履歴を勘案した推定状態h_t^*を出力する．行動は，C_1から C_kまでのUnixコマンド文字列の埋込ベクトルと推定状態h_t^*から行動の選択に関する確率分布を出力し，その分布に従ってサンプリングして行動を選択する．行動として選択されたUnixコマンドを再び実システムに入力することで，そのコマンドに対するシステムの応答が，次の部分観測情報として再びエージェントに入力される．これをFlagが出力されるまで繰り返すことで，自律的にCTF問題を解くことが期待される．

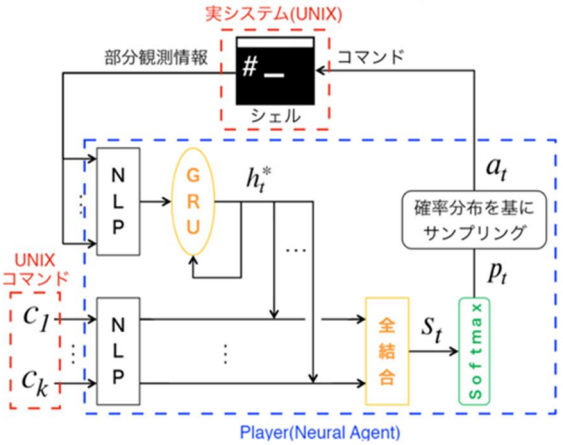

図 2 自律サイバー推論システムの概要

3.4.1 GTrXL の構造

これに対し，GTrXLに基づく自律サイバー推論システムは， NLP部をGTrXLに置き換えたアーキ テクチャを構築した．GTrXL のレイヤー構造を図３に示す．GTrXLは残差接続とゲーティング機構を備えることを特徴とし，状態再利用を目的とする再帰ループを持つニューラルネットワークである．

(1) 埋め込みと位置符号化

Input EmbeddingはText情報を単語毎に分割し，それぞれの単語の分散表現（埋込ベクトル）に変換する層である．単語の位置に応じて，以下の式で与えられる位置符号化(positional encoding)を用いて，それぞれの単語の近傍にある単語との間の相関を高め，文中の位置が離れた単語との相関を弱める働きを有する．

$$PE(\mathrm{pos}, 2i) = \sin\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right)$$

$$PE(\mathrm{pos}, 2i+1) = \cos\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right)$$

[3] NLP: Natural Langage Processing

[4] Adolphs, L. & Hofmann, T. LeDeepChef: Deep Reinforcement Learning Agent for Families of Text-Based Games. in 34th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI 2020) 5180 (2020).

(2) 全結合層

2つの線形変換とReLU活性化関数によって構成されている.

$$FFN(x) = \mathrm{Re}\,LU(xW_1 + b_1)W_2 + b_2$$

(3) 相対マルチヘッド注意機構

相対的マルチヘッド注意機構は，通常のマルチヘッド注意機構に相対的位置符号化を加えて，可変長のテキストに対応できるようにしたものである.相対位置エンコーディングを用いることで,長さの異なるシーケンスを扱う場合に，入力シーケンス内の要素間の関係性を考慮することができる．下記にその式を示す.

$$\tilde{E}^{(l-1)} = \left[M^{(l-1)}, E^{(l-1)}\right]$$

Eは位置符号が加えられた文章の分散表現である．$l \in [1, L]$ は層のインデックスを表している.

$$Q^{(l)}, K^{(l)}, V^{(l)} = W_Q^{(l)} E^{(l-1)}, W_K^{(l)} \tilde{E}^{(l-1)}, W_V^{(l)} \tilde{E}^{(l-1)}$$

また，正弦標準行列Φを学習パラメータで線形変換した

$R = W_R^{(l)}\Phi$を相対位置

Eを$W_Q^{(l)}, W_K^{(l)}, W_V^{(l)}$ で線形変換した$Q^{(l)}, K^{(l)}, V^{(l)}$を用いて,

次の添字dに関する内積$\alpha_{htm}^{(l)}$を計算し,

$$\alpha_{htm}^{(l)} = Q_{htd}K_{hmd} + Q_{htd}R_{hmd} + u_{h*d}K_{htm} + v_{h*d}R_{hmd}$$

添字mに関するSoftmaxにより注目すべき単語の確率分布を求める.

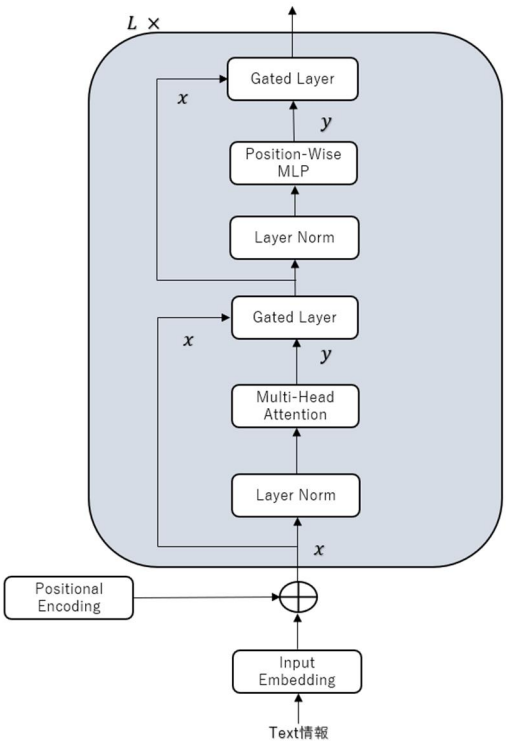

図 3　GTrXL のレイヤー構造

$$W_{htm}^{(l)} = \ \mathrm{MaskedSoftmax}\ \left(\alpha^{(l)},\ \mathrm{axis}\ = m\right)$$

さらに層正規化を加えた操作を再帰的にL回繰り返し，部分観測情報(Text情報)に対応したマルチヘッド数に応じたh個の潜在ベクトルを出力する.

(4) ゲート層

GTrXLではtransformerの残差接続をゲーティング機構に変更している．入力テキストの埋込ベクトルxと,相対マルチヘッド注意機構の出力yとして,学習パラメータ$W_g^{(l)}, U_g^{(l)}W_r^{(l)}, U_r^{(l)}$を用いて,

ゲート層の出力値$\hat{h}$を計算する．ここで，⊙はアダマール積を表す.

$$\hat{h} = \tanh\left(W_g^{(l)}y + U_g^{(l)}(\sigma\left(W_r^{(l)}y + U_r^{(l)}x\right) \odot x)\right)$$

この機構により，入力xと注意機構の出力yの値に依存して残差接続の重みを単語毎に制御することが可能になる．さらに，単語毎の内分係数zを以下の式で求め,

$$z = \sigma\left(W_z^{(l)}y + U_z^{(l)}x - b_g^{(l)}\right)$$

ゲート層の出力に含める入力xと，ゲート層の出力値$\hat{h}$を単語位置毎に選択する．

$$g^{(l)}(x,y) = (1-z) \odot x + z \odot \hat{h}$$

3.4.2 GTrXL に基づく自律サイバー推論システム

前述のGTrXLを用いて，本研究では以下のGRUレイヤーをGTrXLに変更し，GTrXLモデルの出力する512次元のベクトルを，次のGRUレイヤーに通し，次状態であるh_t^*を推定する自律サイバー推論システムを構成した．部分観測情報から状態を推定するネットワークを図4に示す．前述したGTrXLをcontext encodingの部分に組み込むことで，状態推定の精度が上昇することが期待される．本研究では，GTrXLに基づく自律サイバー推論システムがGRUに基づくシステムやDeepExploitと比べて桁違いに著しく少ないエピソード数で効率良く最適戦略を学習できることを確認した．有効性を確認した．以下に，実験の詳細を述べる．

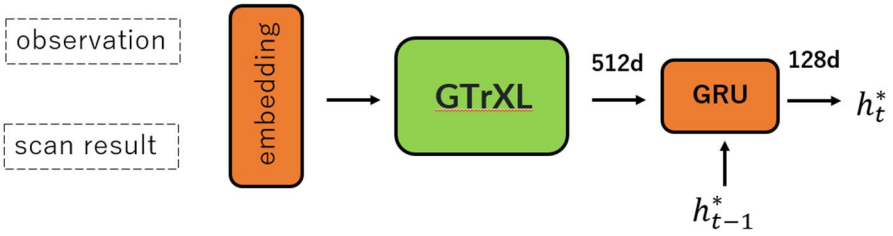

図 4　GTrXL に基づく部分観測情報からの状態推定

(1) 実験結果1

単一のサーバに対してポリシーの最適化を行い，GRU， GTrXLに基づく2つの自律サイバー推論システムが最適ポリシーを獲得するまえに要したエピソード数を比較した．また，比較のためにMetasploitを用いた自動ペネトレーションツールであるDeepExploit[5]の実験結果も示す．

実験1の設定

- Metasploitable2ベースの単一サーバー
- 部分観測情報に以下を補足情報(Text情報)として追加している．
  - 過去の行動履歴
  - 標的サーバーに対するnmapの出力（各ポートの状況）
- 行動空間
  - set rport 〈ポート番号〉　　ポート番号の数～10^1
  - set module 〈モジュール番号〉　　モジュール番号の数～10^3
  - set ip 〈ipアドレス〉　　ipアドレスの数～10^1
  - set payload 〈ペイロード番号〉　　ペイロードの数～10^2

  4つの情報が揃い，exploitを実行可能な状態になると標的サーバーに対して攻撃を実行する．

200ステップ以内にexploitに成功しなければ当該エピソードは失敗とする．学習パラメータの更新は10ステップ毎に行い，3500ステップまでのステップ毎の期待報酬を記録した．行動空間のサイズは10^7程度である．標的サーバの種類によってそれぞれの数は変動する．割引率γは0.9に設定している．また，報酬は標的exploitに成功すると，スコアとして100が与えられ，1エピソード内に標的サーバーのexploitに成功しなかった場合，負の報酬として-100を与える．また，1ステップごとに-1を与えている．最適ポリシーのステップ数は5であり，ステップ毎の期待報酬は19である．

[5] https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit

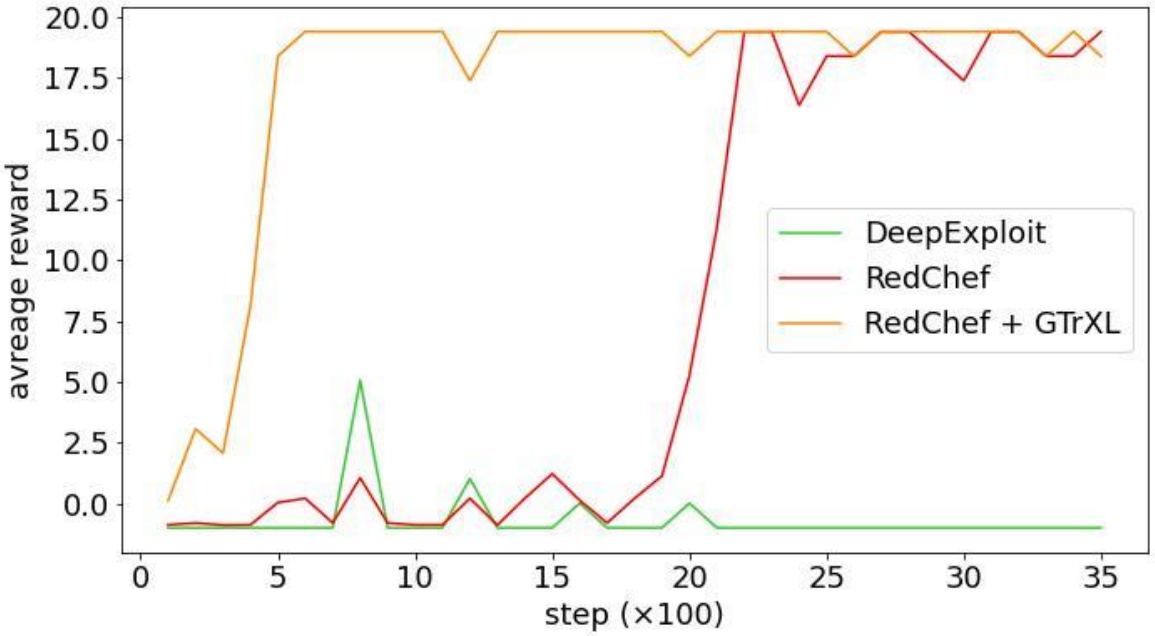

図 5 実験 1 の結果（単一サーバーへの Exploit 問題）

実験1の結果を図5に示す．この図ではステップ毎の期待報酬をプロットしている．GTrXLベースの手法(RedChef + GTrXL)は約500ステップ目で最適な平均報酬となっている．GRUベースの手法(RedChef)は約2000ステップ目に最適な平均報酬となっており,GTrXLベースの手法が1桁近く速く最適ポリシーに収束していることが分かる.比較対象のDeepExploitはモジュール番号やペイロード番号の選択をランダムに行っているため，ステップ毎の期待報酬の増加は見られない.

(2) 実験結果2

攻撃機序の異なる複数の標的サーバに対してポリシーの最適化を行い，GRUベースの手法とGTrXLベースの手法の比較実験を行った.標的サーバにはMetasploitable2を含む計6種類を用意した.割引率や報酬は実験1と同じ値とし,500ステップ以内にexploitに成功しなければ失敗とした.

実験結果を図6に示す． GTrXLベースの手法(RedChef + GTrXL)は約7500ステップ目で最適な平均報酬となっており，6種類の標的サーバ全てに対する最適ポリシーの獲得に成功している．GRUベースの手法 (Redchef) は約17500 ステップ目で最適戦略の獲得に成功しており，GTrXLベースの手法が効率的に最適ポリシーを獲得していることを確認した.

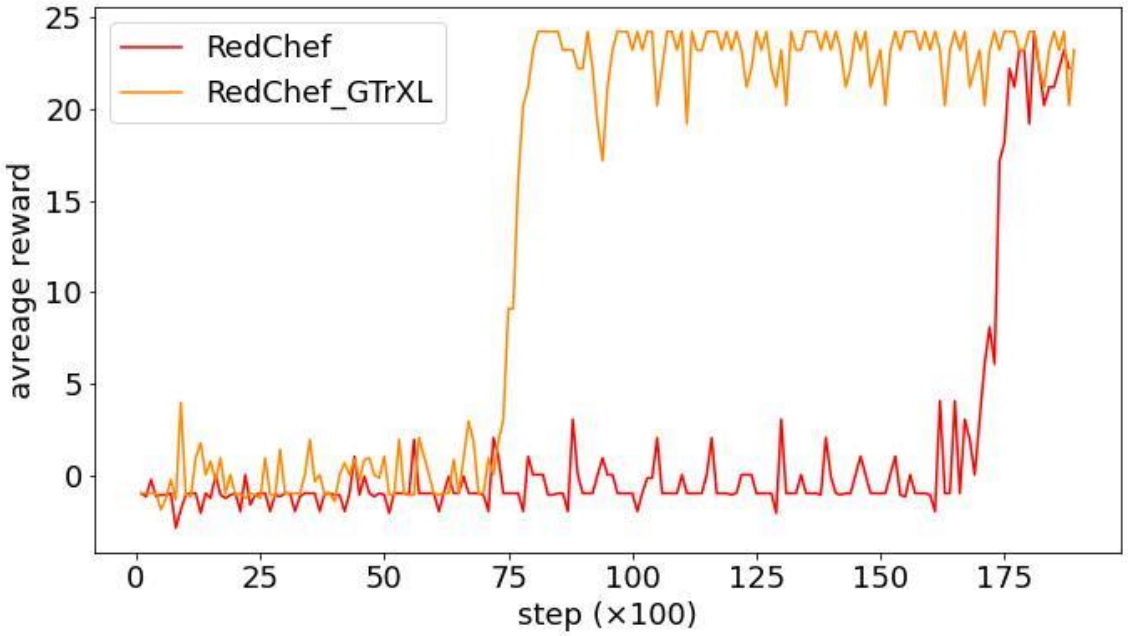

図 6　実験 2 の結果（複数サーバーへの Exploit 問題）

3.4.3 考察

今回の実験は,Metasploitシェルへのテキストデータによる入出力を直接POMDP型深層強化学習による自律サイバー推論システムで処理できることを示している．POMDP型の深層強化学習は，部分観測情報から状態を推定する必要があることから，MDP型のゲームを仮定している従来研究よりも遙かに実践的な環境でCTF問題に対応できることを示唆している．

しかし，標的サーバーがネットワーク構造を持つような場合や，行動の選択肢が桁違いに多い場合に対応するには，本研究を単純にスケールアップしただけでは難しい感触である．次章以降では，最近話題になっている大規模言語モデルを適用するアプローチを検討し，極めて良い結果を得たので報告する．

3. 5 敵対的学習による自律学習手法の検討

深層強化学習によるアプローチを実サーバーに適用する際の問題の一つに学習時に要する時間が挙げられる．前節で述べた実験では，観測-行動の間に環境であるMetasploitシェルと対話し，Metasploitシェルは標的システムに実際に攻撃を実施して結果を返す．このため，行動から観測までのタイムラグが無視できないほど大きい．そこで，敵対的学習等により，標的システムへのアクセスを減らし，効率的に学習するアプローチを模索していたところ，大規模言語モデルによるアプローチで極めて良好な結果が得られた．

3.5.1 大規模言語モデルによる CTF 求解

OpenAI社が開発したLLMの一つであるChatGPT[6] を用いて，エージェント（人間による仲介が必要）と環境がインタラクティブにやり取りを行う環境（Game）において，エージェントがFlagを見つけ出し,CTFサーバにFlagを提出することでスコアを獲得するという設定で実験を行った.Gameではまず初めに,Kali Linux環境にpicoCTF2022サーバーの問題をダウンロードし,問題から得られる部分観測情報O_tを基に人間がテンプレートに従ってプロンプトを作成し， ChatGPTエ

[6] https://chat.openai.com/

ージェントに入力する.ChatGPTエージェントは,プロンプトに従って行動A_tを生成しフラグ獲得を目指す.フラグ獲得のために生成される行動$A_t^{\star}$には,Kali linuxで実行可能なOSコマンドの他,Pythonプログラムなどがある.ここで,行動Aはトークンの集合, tはタイムステップ, $A^{\star}$はトークンの系列とする.

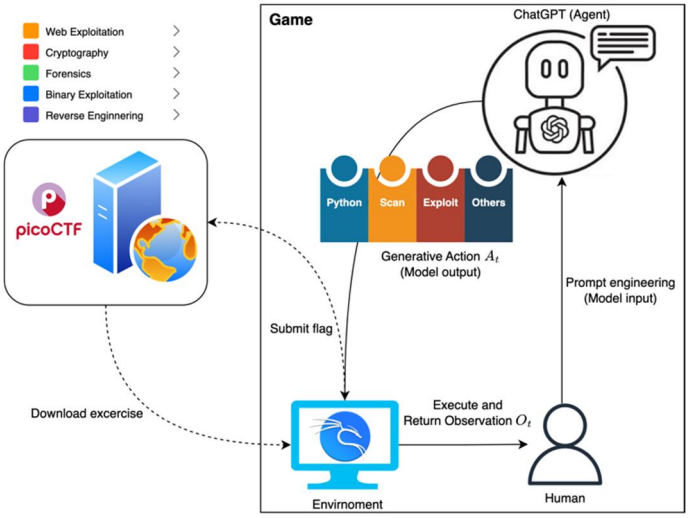

図 7 ChatGPT による CTF チャレンジ概要図

実験では, picoCTF2022[7]を使用した.picoCTFは,主に中高生や初学者向けに作られたオンラインのCTF競技大会で,Carnegie Mellon Universityが主催している.問題は全部で64問あり,それぞれ100,200,300,400,500のポイントが与えられる.一般的に問題の難易度に応じてポイントが高くなる.最終的なスコアが高かった上位チームには賞金が与えられる場合がある.picoCTF2022では5つの問題カテゴリがある.それぞれの説明と攻撃例を表1に示す.

表 1 CTF カテゴリと攻撃例

カテゴリ	説明	攻撃例
Web Exploitation	Web アプリケーションにおいて,脆弱性を突いた攻撃を行う.	SQL インジェクション, クロスサイトスクリプティング (XSS).
Cryptography	暗号理論を用いた問題が出題され,解読を試みる.	暗号文の解読, 暗号鍵の復号化
Forensics	決められたデータから情報を探索する.	メモリダンプからの情報抽出, ファイルからの隠された情報の発見
Binary Exploitation	プログラムのバイナリを解析し,攻撃手法を見つける.	バッファオーバーフロー, スタックの上書き
Reverse Engineering	プログラムのバイナリを解析し,仕様を特定する.	プログラムの逆アセンブル, 実行ファイルからの情報抽出

本研究では,ChatGPT の CTF における性能を評価するため,Zero-round,Few-rounds,Failure の 3 つの指標を独自に設定して,フラグ獲得までのプロンプトの情報量を測定した. Zero-round は,

[7] https://picoctf.org/competitions/2022-spring.html

プロンプト情報に人間のサポートなしでフラグ獲得に成功したことを示す.また,Few-rounds は,プロンプト情報に人間のサポートありでフラグ獲得に成功したことを示す.Failure は,フラグ獲得に失敗したことを示す.従って,ChatGPT の CTF における性能評価のために,Zero-round と Few-rounds 割合が多ければ多いほど良いという指標ができる.

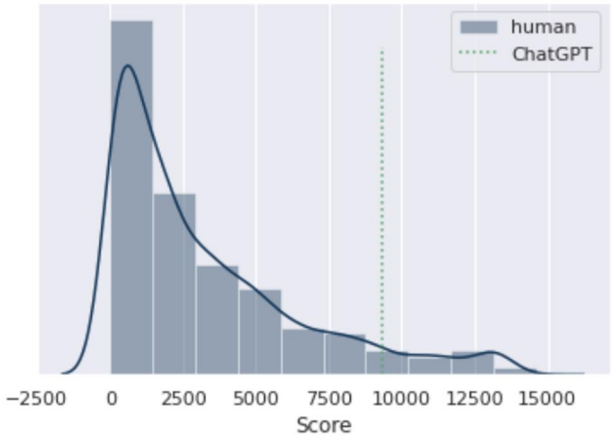

図 8　picoCTF のスコア分布

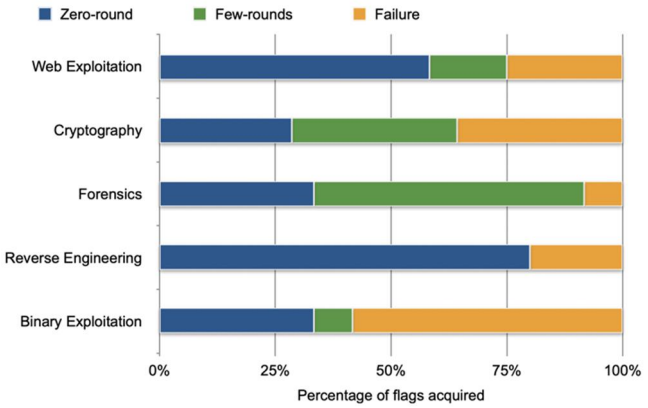

図 9 カテゴリ毎の chatGPT の正答率

先に示した検証方法を基に評価を行ったところ,図 7-10 の結果となった.総合的な結果として,64 問中 48 問のフラグを獲得することに成功し,全参加者 7794 人中 575 位となった.スコアに関する統計データを表 2 にまとめた.

表 2　スコア分布と chatGPT の得点

平均点	標準偏差	ChatGPT の得点	ChatGPT の偏差値
3244.4	3280.9	9300	68.5

図 9 は,カテゴリ毎にソートしてデータをまとめた結果である.Binary Exploitation 以外のカテゴリにおいて,60%以上の割合でフラグの獲得に成功した.この結果について,ChatGPT は,一般的な Q&A のような対話だけでなく,プログラムコード生成の性能が高いため,Reverse Engineering のようなプログラム解析の問題に Zero-round で解答できたと推測される.一方,Binary Exploitation のような参加者でも平均的に解くことが難しいカテゴリでは,Failure の割合が多い結果となった.

図 10 は,問題ポイント毎にまとめた結果である.この結果から,300 ポイント以下の比較的低いポイントの問題においてフラグ獲得に至っているケースが多いことが確認できる.一方,400 ポイント以上の問題ではフラグ獲得率が低く,その原因として大きく 2 つ考えられる. 一つ目は,参加者が Web で簡

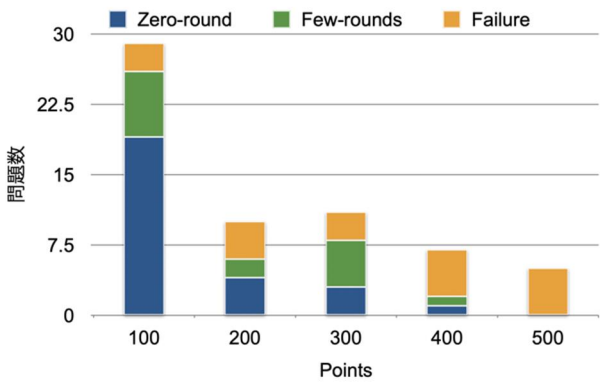

図 10　ポイント毎の chatGPT の正答数

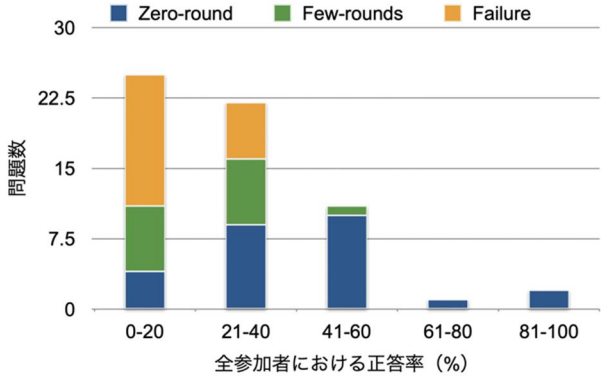

図 11　参加チームの正答率と chatGPT の正答数

単にアクセスできるような解析ツールを ChatGPT の場合利用できない点がある. ChatGPT は,学習時に高度な解析ツール生成について学習している可能性が低いため,モデルのパラメータだけで解析ツールを生成することは難しいと考えられる. 二つ目は,プロンプトの入力制限(4,000トークン).例えば,ファイル数が非常に多い Web アプリケーションの問題を解く場合, 全てのソースコードをプロンプトに入力することは難しい. 加えて,人間のように Web ブラウザにアクセスして探索することもできないため,問題のステートレスな状態を把握することができない.

図 11 は,正答率の分布毎にソートしてデータをまとめた結果である.正答率は,正解フラグを取得した数と picoCTF2022 に参加した累計数で算出した数値である.全体的な Failure の傾向として,前節述べた解釈と同じであると言えるが,Few-rounds において異なった解釈が得られる.Few-rounds の数は,全参加者における正答率が低いほど多くなっている.この分析結果から,Few-rounds のプロンプトサポートにより,ChatGPT が依然として高難易度の CTF 問題でフラグを獲得する可能性があることが示唆されている.

3. 6 評価手法の検討

自律サイバー推論システムの性能評価指標を検討した. CTF問題を解くシステムの場合には, CTFスコアに加えて, CTFの解に至る前段階での性能評価方法も検討する. 一般に, 強化学習のポリシーの評価は初期状態の価値関数で与えられることが多い. また, 最適ポリシーが獲得できる場合は, 最適ポリシー獲得までのエピソード数で評価することも考えられる. 本研究における深層強化学習の評価では最適ポリシー獲得までのエピソード数で学習効率を評価した.

いずれの方法も強化学習に固有の数値であり,人間の能力との比較が必要な場合には,picoCTFの実験で行ったようなzero-roundでの正答数や偏差値で比較せざるを得ない.

４.委託業務全体の成果

4.1 計画時に想定していなかった成果(副次的成果)や, 目標を超える成果

自然言語処理技術に注目し, 当初からPOMDP型深層強化学習にアプローチを絞って研究した結果, 従来の関連研究の結果を大幅に上回る結果が得られたと考えている. また, 令和４年12月のChatGPTの発表以降, 大規模言語モデルの研究が急速に注目を集めており, 本研究においても, CTFに適用した結果, 極めて汎用性の高い能力を有することが判明した.

4.2 研究課題の発展性(間接的成果を含む)

多数の大規模言語モデルが公開され, 研究に用いるための環境が整ってきている. 今後は, 大規模言語モデルを強化学習のモデルでサイバー攻撃対応に即したファインチューニングを行い, 強力な自律サイバー推論システムを構築する方向に大きな発展の可能性があると思われる.

4.3 研究成果の発表・発信に関する活動

これまでに国際ジャーナル1件, 国際会議3件, 国内学会6件の発表を行うなど, 積極的に研究成果を発表してきた. また, 人工知能学会「安全性とセキュリティ研究会」の設置に参画し, 国内のAIセキュリティ領域の研究コミュニティ形成に貢献した.

５. プロジェクトの総合的推進

５. １　研究実施体制とマネジメント

静岡大学を再委託先に加え, 研究推進体制を強化した. また, 研究協力者を追加し, CSS2022, 人工知能学会合同研究会, SCIS2023, NDSS Workshop BAR2023等で研究成果を発表した. さらに, 研究実施場所に一橋大学を追加し, 高速なネットワーク環境に機械学習サーバ

ーを接続し，研究を加速した．研究終了後も2023年6月に開催される人工知能学会全国大会でLLMの成果を発表する予定である．

５．２　経費の効率的執行

特になし．

６．まとめ，今後の予定

本研究では、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)レスポンスを部分観測情報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，複数の異なった脆弱性を有するサーバーへの侵入するCTF問題の求解に成功した．我々の知る限り同領域の既存研究は，単純なCTFゲームを対象とし，かつ状態が環境から与えられるMDP型の強化学習に関する研究に限られており，本研究は実Unixサーバーを対象とする現実的な設定でのCTF求解に成功するための基本原理を明らかにしたことの意義は大きいと考えている．

さらに，本研究では大規模言語モデル(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は特筆に値する．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の中核に成長する可能性がある．

今後は，大規模言語モデルを強化学習のモデルでサイバー攻撃対応に即したファインチューニングを行う学習フレームワークを開発し，強力な自律サイバー推論システムを構築する方向で研究を進める予定である．

７．研究発表，知的財産権等の状況

（１）研究発表等の状況

本研究では，積極的に学会発表を行い．研究論文1件，国際会議での発表3件，国内発表6件を行った．

種別	件数
学術論文	1
学会発表	9
展示・講演	該当なし
雑誌・図書	該当なし
プレス	該当なし
その他	該当なし

（２）知的財産権等の状況

該当無し

（３）その他特記事項

該当無し