テンソルデータ拡充を用いた組織内ネットワーク攻撃判定方式の回避攻撃に対するロバスト性の向上

宍戸 克成[1,a)] 森川 郁也[1] 及川 孝徳[1] 海野 由紀[1]

概要：サイバー攻撃の増加に伴い， AI を用いた攻撃検知の研究が活発に行われている． 2019 年に及川らはテンソルデータ拡充を用いたニューラルネットワークによる組織内ネットワーク攻撃判定方式を提案し，判定精度 95% を攻撃の見逃し無しで達成した．一方で， AI システムの潜在的な特性を突いた攻撃が報告されており，モデルを騙す攻撃 (回避攻撃) は AI を用いたセキュリティアプリケーションに対して大きな脅威となる．本研究の目的は及川らの方式に対する新たな回避攻撃の発見と，及川らの方式の精度を保ちながら 回避攻撃 に対する攻撃検知精度を高め，モデルのロバスト性を向上することである．本稿では適切にテンソルデータを拡充することで，目的を達成できることを報告する．

キーワード：AI, 回避攻撃, Adversarial Training, 標的型攻撃

1. はじめに

情報技術の発達と共に，特定機関や組織の機密情報の窃取やサービス妨害を目的とするサイバー攻撃が社会問題となっている．特に，機密情報の窃取・破壊・改ざんを目的とする標的型攻撃は年々活動の高度化・巧妙化，そして攻撃数が増加しており，攻撃者の組織内ネットワークへの侵入を完全に防ぐことが難しい．近年のサイバー攻撃対策は攻撃者に侵入された際に可能な限り早い対処が要求されている．そこで，近年脚光を浴びている人工知能 (AI) 技術，特に機械学習や深層学習を応用したセキュリティアプリケーションの研究開発が盛んに行われており，可能な限り早い対処を実現する技術が多く提案されている．

一方で，2015 年頃から AI システムに対する攻撃報告が増加しており，特に AI システムを騙す攻撃 (回避攻撃) が盛んに議論されている．特に，画像分類タスクに対する攻撃“Adversarial Example”はセキュリティや機械学習系の学会でも議論されている．

図 1 は元々 55.7%の確率でパンダと判定される画像に「人が認識できない小さなノイズ (摂動)」を加えた画像を攻撃者が画像分類器に入力すると，99.3%の確率でテナガザル (gibbon) とモデルが誤判定する [1] ことを示している．道路標識を誤判定させることに成功した事例 [2] も報告されており，Adversarial Example は自動運転車の課題の一つとして考えられている．セキュリティ領域ではマルウェア検知に対する回避攻撃 [3], [4], [5] が報告されている．AI システムに対する既存の回避攻撃 [1], [3], [4], [5] は攻撃の前提条件が必ずしも現実的でないこともある．しかし，回避攻撃は AI システムの潜在的な特性を利用するため，AI システムの分類精度の劣化等の副作用を最小限に抑えながら潜在的特性の影響を抑えることが重要である．

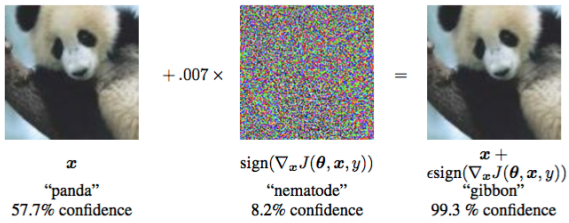

図 1 Adversarial Example の例 [1]

本稿では 2019 年に及川らが提案したニューラルネットワークを用いた組織内ネットワーク攻撃判定方式 [6] の回避攻撃とその対策を述べる．我々の目的は [6] の方式の精度を劣化させずに回避攻撃に対するモデルのロバスト性を向上することである．[6] は通信路上のパケットをテンソルデータに変換して業務操作と攻撃操作を学習させることで，近年見つけにくくなっている標的型攻撃と業務データを高精度に分類することができる．

本稿は攻撃者が意図的に攻撃通信を業務通信と判定させる回避攻撃，そして分類精度を劣化させずに回避攻撃に対してモデルをロバストにする防御手法を示す．我々はマルウェア検知に対する回避攻撃である Append attack[5] を

[1] 株式会社富士通研究所，神奈川県川崎市中原区上小田中 4-1-1
FUJITSU LABORATORIES LTD., 4-1-1, Kamiodanaka, Nakahara-ku Kawasaki, Kanagawa 211–0053, Japan
[a)] k.shishido@fujitsu.com

ネットワークセキュリティの攻撃検知用 AI に応用し，初めて [6] の回避攻撃を実現した．我々は回避攻撃の対策として，画像等の metric なデータに対する回避攻撃の有効な防御手段である Adversarial Training[1] を non-metric なデータに適用し，non-metric なデータに Adversarial Training を適用した際に生じる新たな課題とその解決手法を初めて示した．提案した解決手法は従来と同等の性能を有するロバストなモデルの学習を実現できる．

本稿の構成は 2 章で，回避攻撃と攻撃の前提条件を述べ，3 章で関連研究を紹介する．4 章で，我々が提案する回避攻撃と防御手法を述べ，5 章で実験と結果を示す．最後に 6 章で提案手法の有効性と回避攻撃の要因について述べる．

2. 準備

本章では回避攻撃と攻撃の前提条件について説明する．

2.1 回避攻撃

定義 2.1 (回避攻撃) あるデータとラベルの組を $(\mathbf{x}, y)$，攻撃対象の分類器を f とする．データ $\mathbf{x}$ を改ざんしたデータ $\mathbf{x}'$ が「$f(\mathbf{x}') \neq y$」かつ「分類器のタスクに関する特定の条件 P を満たす」とき，回避攻撃に成功したという．

回避攻撃は攻撃対象の分類器のタスクに依存して攻撃成功条件が変わることに注意されたい．例えば，1 章で紹介した Adversarial Example は画像分類タスクに対する回避攻撃であり，条件 P は「人間がデータ $\mathbf{x}$ と改ざんデータ $\mathbf{x}'$ の違いに気が付かないこと」である．この条件 P はある小さな実数 $\epsilon > 0$ に対して，$|\mathbf{x} - \mathbf{x}'|_\infty < \epsilon$ と定式化されることが多い．本稿で扱う攻撃判定器に対する回避攻撃の条件 P は「攻撃データ $\mathbf{x}$ の改ざんデータ $\mathbf{x}'$ の命令が攻撃機能を保持すること」である．

2.2 攻撃者の能力

攻撃者の能力について説明する．はじめに，Black-box attack と White-box attack を定義する．

定義 2.2 (Black-box attack) 攻撃者が攻撃対象の分類器 f に関する任意の入出力ペア $(\mathbf{x}, f(\mathbf{x}))$ を得ることができる条件の下で行う攻撃を Black-box attack という．

定義 2.3 (White-box attack) 攻撃者は Black-box attack の条件に加え，攻撃対象の分類器 f の内部状態を得られる条件の下で行う攻撃を White-box attack という．

攻撃者の能力に加え，攻撃者の前提知識も重要になる．例えば，攻撃対象の分類器 f の訓練データや評価データに関する知識は攻撃者にとって非常に有用である．本稿では学習・評価データに関する知識を持つ攻撃者を「知識のある攻撃者」と呼び，知識を持たない攻撃者を「知識がない攻撃者」と呼ぶ．

3. 関連研究

本章ではニューラルネットワークによる組織内ネットワーク攻撃判定方式 [6] の概要と Adversarial Example に対して，現状最も効果的な防御手法である Adversarial Training[1], [7] について説明する．

3.1 組織内ネットワーク攻撃判定方式

2019 年に及川らが提案した組織内ネットワーク攻撃判定方式は海野らの高速フォレンジック技術 [8] と多次元の配列データ (テンソルデータ) の学習に特化したニューラルネットワーク [9] を組み合わせた方式である．一般的に，我々は観測できるネットワークから業務データを大量に収集できる．しかし，安定した教師あり学習を行うために必要な攻撃データを収集することは難しい．攻撃データ収集の困難性から，攻撃検知システムは攻撃データが業務データから外れる値であることを前提として，業務データから外れ値を検知するアノマリ検知を採用しているものが多い．一方で，及川らは「ATA Suspicious Activity Playbook」[10] 等を参考に作成した攻撃データと観測した業務データ (以下，元データ) を用いて学習したモデルに，深層学習の説明機能のひとつである LIME[11] を適用して学習網羅性を満たす攻撃データを生成するデータ拡充技術を提案した．本稿では「データ拡充技術で生成した攻撃データ」を攻撃亜種データと呼ぶ．海野らの高速フォレンジック技術はネットワークを流れる通信から通信プロトコルを分析し，実行されたリモート操作コマンドの種類を特定してコマンドの再構成を行う．及川らは海野らの高速フォレンジック技術を用いて，ネットワーク通信から抽出した「コマンド名・オプション・使用アカウント名・ファイル拡張子名等」の情報を通信データの特徴量として利用し，[9] の手法で学習・推論を実施した．そして，生成した攻撃亜種データを元データにに加えて再学習を行う．表 1 に示すように，[6] の方式は攻撃亜種データを追加することにより，精度 0.945 を攻撃の見逃しなしで達成した．

表 1 元データ vs. 元データ+攻撃亜種データの精度比較

	Accuracy	Precision	Recall	F-measure
元データ	0.890	–	0.556	–
元データ+亜種	0.945	0.386	1.000	0.557

3.2 Adversarial Training

画像分類タスクに対する回避攻撃である Adversarial Example に対して，最も有望な防御手法である Adversarial Training について説明する．n 枚のデータとラベルの組 $(\mathbf{x}_i, y_i)_{i\in[n]}$ を訓練データとし，得られた学習済みモデルを F とする．はじめに，Adversarial Example について

説明する．ある入力画像とそのラベルの組 $(\mathbf{x}, y)$ に対して，$F(\hat{\mathbf{x}}) \neq y$ を満たす改ざんした入力 $\hat{\mathbf{x}} = (= \mathbf{x} + \delta)$ を Adversarial Example という．ここで，δ は摂動と呼ばれ，攻撃者は人が認識できない大きさの任意の値を選択できる．

Adversarial Training のアイデアは非常にシンプルである．モデル学習者はすべての訓練データ $(\mathbf{x}_i, y_i)_{i\in[n]}$ に対し，学習済みモデル F に対する Adversarial Example $\hat{\mathbf{x}}_i$ を生成し，データとラベルの組 $(\hat{\mathbf{x}}_i, y_i)$ を訓練データに加えて再学習を行う．直感的に，各訓練データの Adversarial Example を学習することによって，新たなモデルは Adversarial Example が入力されても，誤判定を起こす可能性が低くなる．これまでに提案された Adversarial Training は訓練データのラベルを用いて Adversarial Example の再学習を実施する手法 [1], [12] と事後確率分布 (確信度の分布) を滑らかにする手法 [7] に分けられる．

Goodfellow らが提案した手法 [1] は先に述べたように，各訓練データの Adversarial Example を訓練データに加えて再学習することで，ロバストなモデルを得るものである．Madry らは各訓練データとラベルの組 $(\mathbf{x}_i, y_i)$ に対して，損失を最大化する摂動 $\delta \in \Delta$ を用いた Adversarial Example を訓練データに加えて再学習する手法 [12] を提案している．[12] では次式の最適化問題を解く．θ はモデル F の重み，ϵ は 0 以上の実数である．

$$\min_{\theta} \sum_{i} \max_{\delta\in\Delta} \ell(f_\theta(\mathbf{x}_i + \delta), y_i), \text{where } \Delta = \{|\delta|_\infty \leq \epsilon\}.$$

Miyato らが提案した手法 [7] は事後確率分布を滑らかにすることで，Adversarial Examples に対してモデルをロバストにする．ここで，モデル F の重みを θ，事後確率分布を $p(y|\mathbf{x}, \theta)$ とすると，訓練データ $\mathbf{x}$ 近傍のデータ $\mathbf{x} + \delta$ に関する事後確率分布は $p(y|\mathbf{x} + \delta, \theta)$ となる．Adversarial Example はデータ x 近傍のデータとなる条件から，Miyato らは事後確率分布 $p(y|\mathbf{x}, \theta)$ と $p(y|\mathbf{x} + \delta, \theta)$ が近い値になるように制約を加えることで，データ x 近傍のデータの推論結果が急に変わらないようにしてモデルをロバストにした．具体的に，次式に示す KL ダイバージェンスを正則化項として採用する学習方法となっている．

$$\Delta_{\mathrm{KL}}(\delta, \mathbf{x}, \theta) = \mathrm{KL}[p(y|\mathbf{x}, \theta)||p(y|\mathbf{x} + \delta, \theta)].$$

4. 提案手法

はじめに，標的型攻撃判定方式 [6] のデータ形式を説明し，我々が提案する回避攻撃と防御手法を説明する．

4.1 データ形式

3.1 節で述べたように，海野らの高速フォレンジック技術を用いて通信データからコマンドを再構成する．通信データからコマンドを再構成した後，標的型攻撃において攻撃者が常套的に使用するリモート操作コマンドを含めてセッション単位で抽出し，それらをテンソルに変換する．セッションとは再構成した一連のコマンド列で，セッション 1 個が 1 つのテンソルとなり，業務通信，または，攻撃通信を表す入力になっている．図 2 に示すように，各行が 1 コマンドで各列が特徴量を表している．本研究では，[6] と同様に「コマンド名・コマンドオプション・アカウント名・共有フォルダ名・ファイル拡張子名・フォルダ名」の 6 種の特徴を使用する．

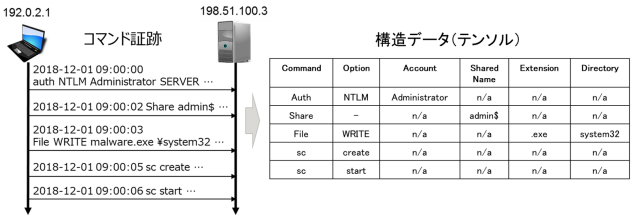

図 2 テンソル化のイメージ

4.2 標的型攻撃判定器に対する回避攻撃

画像分類器に対する攻撃である Adversarial Example は攻撃対象のモデルの内部状態 (損失関数の勾配) を観察して条件を満たす摂動 δ を求める．しかし，セキュリティアプリケーションの多くは入力が離散値であるため，Adversarial Example の生成方法を素直に適用することができない．定義 2.1 に示したように，標的型攻撃判定器 F に対する回避攻撃は改ざんしたデータの命令が元の攻撃データの命令を保持する必要があるため，入力の変更や削除といった改ざんは困難である．さらに，現実のネットワーク環境で実現ができない通信は AI システムに入力されることがないため，改ざん後のデータが表す通信は現実的に実現できる必要がある．これらの理由から，標的型攻撃判定器に対する回避攻撃が満たす条件を以下で定義する．

定義 4.1 (Strict con.) 標的型攻撃の回避攻撃が満たす以下の条件を strict con. と呼ぶ．

・攻撃機能の保持，

・改ざんデータが表す通信は現実的に実現可能．

我々は回避攻撃の条件を満たすために，「攻撃データの前方，または後方に業務データを追加するデータ改ざん法」を提案する．この手法は Suciu らのマルウェア検知に対する Append attack[5] を応用している．Append attack は攻撃対象モデルの推論の結果，高確率で異常なしと判定されるバイナリファイルの一部をマルウェアに追加することで攻撃判定を回避する手法である．攻撃者は攻撃判定器 F に対して，以下の条件を満たすものを攻撃データ $\mathbf{x}$ に追加するデータ $\mathbf{x}_{add}$ として選択する．ここで，$F(\mathbf{x}) \in \{0, 1\}$ で 0 は業務通信，1 は攻撃通信を表すラベルである．

(i) $F(\mathbf{x}_{add}) = 0$ を満たす，
(ii) 改ざんデータが表す通信は現実的に実現可能．
この攻撃は $\mathbf{x}_{add}$ を探索するため，組織内を流れる業務データの傾向を知る必要がある．よって，知識のある攻撃者の Black-box attack で実現が可能である．

4.3 回避攻撃に対する防御手法

我々が提案する防御手法について説明する．提案手法の基本的なアイデアは 3.2 節で説明した Adversarial Training に類似する．以下に，[1], [12] の手法との相違点を挙げる．

(1) Adversarial Training: 攻撃判定器に有効な全ての回避攻撃データ $\mathbf{x}'$ と攻撃ラベル $(y = 1)$ の組 $(\mathbf{x}', 1)$ を訓練データに追加
(2) Balanced data: 業務データを増強し，攻撃データ数と業務データ数を統一

(1) は既存手法 [1], [12] と同様に，訓練データに攻撃データを加える．既存手法は誤判定を発生させるデータのうち，特定の改ざんデータのみ加える．一方で，我々の手法は誤判定を発生させるすべての改ざんデータを追加する．Adversarial Example は攻撃データが元データ近傍に存在することが保証されているため，最も損失を大きくする改ざんデータのみで十分と考えられる．しかし，我々の攻撃手法は攻撃データが元データ近傍に存在するとは限らないため，すべての改ざんデータを訓練データに追加している．
(2) はデータ不均衡対策である．[1], [12] では訓練データに追加するデータ数が元データに対して少ないため，各クラスのデータ数の偏りが小さくモデルの精度に与える影響が弱い．しかし，提案手法は訓練データに追加するデータ数が多く，クラス間のデータ数の偏りがモデルの精度に大きな影響を与える．本研究ではデータ不均衡対策として up sampling を採用する．Up sampling は回避攻撃のデータ生成法と同じ条件で，業務データの後方にデータ $\mathbf{x}_{add}$ を追加してデータ数を増強する．

5. 評価実験

我々は「4.2 節で提案した攻撃手法が攻撃判定方式 [6] に対して有効か否か」と「4.3 節で提案した防御手法が [6] の方式を精度劣化させず，ロバスト性を向上できるか」の 2 点について評価を行った．

5.1 実験データ

本研究で使用した攻撃データと業務データは [6] で使われたデータである．本稿では概要のみ説明する．

攻撃データのうち，訓練データは「ATA Suspicious Activity Playbook」[10] 等を参考に作成し，実際に模擬環境において得られた通信観測データを使用した．

攻撃データのうち，評価データには動的活動観測 BOS(Behavior Observable System) の研究用データセット (以下，BOS Dataset) とサイバー攻撃誘引基盤 STARDUST[13] の通信観測データが利用されている．なお、BOS Dataset はマルウェア対策のための研究用データセット (MWS Datasets)[14] に含まれている。STARDUST の通信観測データのうち，日本や台湾の組織を標的とした DragonOK[15] と PowerShell Empire[16] による攻撃の通信観測データを使用した．

業務データは 2018 年 3 月から 8 月までの 6 ヶ月間に富士通株式会社が管理するネットワークから収集したデータである．そのうち，前半 3 ヶ月分を訓練データ，後半 3 ヶ月分を評価データとして使用した．

5.2 学習

実験に用いた攻撃データと業務データのデータ数は表 2 の通りである．[6] の実験では業務データの数が攻撃データよりも多い不均衡状態を解消するために，各パターン，業務データの選択を変えて 10 個のデータセットを作成し，各データセットを学習した 10 個の分類器を作成する．攻撃判定は 10 個の分類器の平均で算出している．

表 2 訓練データと評価データのデータ数

	訓練データ	評価データ
攻撃データ	171	27
業務データ	171	748

5.3 実験方法

5.3.1 攻撃実験

回避攻撃用の改ざんデータは評価データの攻撃データと業務データを用いて生成する．4.2 節で述べたように，evasion dataset(回避攻撃用の改ざんデータセット) は攻撃機能を保持し，通信が現実のネットワークで実現できなければいけない．しかし，すべての evasion dataset が strict con. を満たすことを確認するには専門家によるチェックが必要で，時間的コストの観点から非現実的である．本研究では回避攻撃の条件を緩和し，高確率で条件を満たすことが期待できる evasion dataset を生成する．ここで，strict con. を緩和した relax con. を定義する．

定義 5.1 (Relax con.) 本研究が生成した回避攻撃が満たす以下の条件を relax con. と呼ぶ．
・攻撃機能の保持，
・改ざんデータが表す通信は高確率で現実的に実現可能．

Relax con. を満たすために，我々は以下の条件を満たす業務データを追加データ $\mathbf{x}_{add}$ として選択した．
(i') $F(\mathbf{x}_{add}) = 0$ を満たす，
(ii') $\mathbf{x}_{add}$ のコマンド名に特定の値を含まない．

現実のネットワークで実現できる通信を表す改ざんデータを生成するために，我々は少なくとも 1 セッションで複数発生しえないコマンドを表す特定の値を含まない追加デー

タ $\mathbf{x}_{add}$ を選択し，現実的なデータを生成した．非現実的な通信を表すデータを網羅的に除去できると限らないが，生成したデータは高確率で現実的に実現可能な通信であることが期待できる．生成した evasion dataset を [6] で提案された攻撃分類器に入力し，攻撃検出率を算出した．

5.3.2 防御実験

5.2 節で述べたように，[6] では 10 個のデータセットを作成して各データセットを学習した 10 個の分類器を利用して攻撃分類を行っている．我々は各データセットに対して，提案した防御手法を適用し，増強した訓練データを用いて再学習を行った．学習と評価に使用したデータ数を表 3 に示す．新たに生成した分類器が「standard dataset([6] の評価データセット) に対して精度劣化していないか」,「evasion dataset を正しく分類できるか」の 2 点を評価するために，standard dataset と evasion dataset を再学習して得られたモデルに入力し，精度と攻撃検出率を算出した．また，本研究では relax con. の下で，攻撃改ざんデータを生成しているため，攻撃現実性の観点の評価が弱い．そこで，strict con. を満たす攻撃改ざんデータを手作業で作成し，evasion dataset に加えることで現実性の観点の評価を補った．なお，手作業で作成したデータのうち，[6] で検知されない改ざんデータを実験で利用した．

表 3 提案手法の訓練データと評価データ数

	訓練データ	評価データ
攻撃データ (relax con.)	3678.2(平均)	5319
攻撃データ (strict con.)	—	9
業務データ	3678.2(平均)	748

5.4 実験結果

我々の目的に加え，データ不均衡状態 (inbalanced data) が精度に与える影響を観察するために，inbalanced data と balanced data のモデルの評価も行った．攻撃実験と防御実験の結果を表 4 に示す．

6. 議論

6.1 標的型攻撃判定器に対する回避攻撃

我々が提案した回避攻撃について実験結果から考察する．実験の過程で専門家が回避攻撃の条件の下で改ざんデータを手作業で生成し，[6] の方式に検知されない攻撃データの生成が可能であることを確認した．知識のある攻撃者は Black-box attack で我々の攻撃手法を適用して意図的に攻撃判定器の推論結果を騙すデータを生成できることになる．さらに，回避攻撃の緩和条件の下で作成した改ざんデータも 24%しか攻撃判定されなかった．これらの結果から，我々の攻撃手法は有効で，知識のある攻撃者が Black-box attack で回避攻撃を実現できると思われる．

6.2 回避攻撃に対する防御手法の効果

本研究では Adversarial Example に対して最も効果的である Adversarial Training を [6] の方式に適用した．既存研究は Adversarial Training を適用すると，攻撃全体の約 50%を正しく分類できている傾向がある．

Inbalanced data では relax con. を満たす evasion dataset に対して，49.7%の回避攻撃の検知を実現し，従来の Adversarial Training と同様の結果が得られた．しかし，「standard dataset に対して，[6] から精度劣化をさせないこと」と「strict con. を満たす evasion dataset のうち，11.1%の回避攻撃の検知に留まったこと」より，inbalanced data では我々の目標達成ができなかった．Inbalanced data は Precision が 0.250 であることから攻撃の過剰検知が顕著に表れている．過剰検知の原因は訓練データに含まれる攻撃データが業務データよりも多く，当てずっぽうで攻撃判定しても正解する確率が高くなるためである．また，strict con. を満たす evasion dataset のうち，11.1%しか検知ができなかったことから，手作業で作成した回避攻撃はかなり業務データに近いデータになっていることが予想される．一方で，balanced data では relax con. を満たす evasion dataset に対して，inbalanced data よりも低い 39.0%の回避攻撃の検知となった．しかし，inbalanced data で達成できなかった「standard dataset に対する精度劣化をさせないこと」と「strict con. を満たす evasion dataset に対して，55.6%の回避攻撃を検知したこと」から，目標を達成できた．攻撃判定器は二値分類であるため，画像分類器のような多クラス分類よりもデータ不均衡状態がモデルの精度に与える影響が大きいと考えられる．

6.3 根本的な原因について

回避攻撃が成功する原因について考察する．訓練データは各データ毎にコマンド数が異なる．図 3 は訓練データの

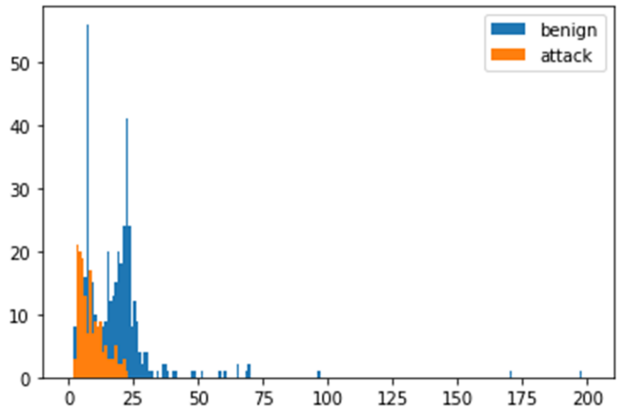

図 3 訓練データのコマンド数に関するヒストグラム

コマンド数に関するヒストグラムである．横軸がコマンド数，縦軸が度数を表す．業務データ (benign) と攻撃デー

表 4　Standard dataset に対する精度と evasion dataset の検知率

Model	Standard dataset				Evasion dataset	
					relax con.	strict con.
	Accuracy	Precision	Recall	F-measure	Accuracy	Accuracy
[6]'s model	0.945	0.386	1.000	0.557	0.240	—
Adversarial Training model on inbalanced data	0.895	0.250	1.000	0.400	0.497	0.111
Adversarial Training model on balanced data	0.947	0.397	1.000	0.568	0.390	0.556

タ (attack) の間でコマンド数という観点で分布に偏りがある．特に，通常データのコマンド数が攻撃データよりも長い傾向がある．この傾向が攻撃の判定基準になっている仮説を立てると，コマンド数が長い攻撃データを作ることで攻撃判定を回避できている説明することができる．実際に，我々が提案した回避攻撃は攻撃データに条件を満たす業務データを追加するため，コマンド数が長くなる傾向がある．訓練データのクラス間分布の偏りは推論結果に大きな影響を及ぼすため，影響を最小化する工夫が必要である．

6.4　回避攻撃の実現可能性

現実における回避攻撃の実現可能性について考察する．提案した回避攻撃を成功させるために，攻撃者は「訓練データや評価データの知識を持つこと」が必要不可欠と考えられる．攻撃者がこれらの知識を持たない，つまり，組織内ネットワークを常套的に流れるデータの知識を持たない場合，組織内であまり使われていない通信を発生させて検知される可能性が高い．また，機械学習や深層学習に対する攻撃論文の多くは White-box attack よりも条件が厳しい Black-box attack で回避攻撃を実現することで攻撃の有用性や現実性を主張している．しかし，実際の AI システムに対する攻撃は攻撃者が標的型攻撃判定器に任意のデータを入力できたとしても，組織が分類器の判定結果を外部に公開しないため，攻撃者は入力に対応する出力を得ることができない．したがって，実際の AI システムの攻撃条件は Black-box attack よりも厳しい条件となる．

上記を踏まえ，判定機の学習に用いたデータの公開は回避攻撃の高いリスクが伴うため，望ましくない．そして知識がない攻撃者は回避攻撃の実現が困難と考えられる．

7.　まとめ

本稿は標的型攻撃判定器 [6] に対する回避攻撃を提案し，その攻撃に対するモデルのロバスト性を改善した．画像分類タスクに対する回避攻撃対策である Adversarial Training を non-metric なテンソルデータに適用した．それに伴い発生するデータ不均衡状態を解消することで，標的型攻撃判定器の精度劣化を抑えながらモデルのロバスト性が向上することを明らかにした．我々の提案手法は従来手法と同じく recall = 1.0 を達成し，24%しか検知できなかった回避攻撃を約 40-50%検知することに成功した．今後は根本的な原因を調査して，モデルの直接的な改良やデータ形式の検討を行い，モデルのロバスト性のさらなる向上を目指す．

謝辞　本研究を進めるにあたり，貴重な攻撃データの提供，及び，サイバー攻撃誘引基盤“STARDUST”を開発した独立行政法人情報通信研究機構（NICT）に感謝いたします．

参考文献

[1] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[2] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D.: Robust Physical-World Attacks on Deep Learning Visual Classification, *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, IEEE Computer Society, pp. 1625–1634 (2018).

[3] Chen, L., Ye, Y. and Bourlai, T.: Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense, *European Intelligence and Security Informatics Conference, EISIC 2017, Athens, Greece, September 11-13, 2017* (Brynielsson, J., ed.), IEEE Computer Society, pp. 99–106 (2017).

[4] Huang, Y., Verma, U., Fralick, C., Infante-Lopez, G., Kumar, B. and Woodward, C.: Malware Evasion Attack and Defense, *49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, DSN Workshops 2019, Portland, OR, USA, June 24-27, 2019*, IEEE, pp. 34–38 (2019).

[5] Suciu, O., Coull, S. E. and Johns, J.: Exploring Adversarial Examples in Malware Detection, *2019 IEEE Security and Privacy Workshops, SP Workshops 2019, San Francisco, CA, USA, May 19-23, 2019*, IEEE, pp. 8–14 (2019).

[6] 及川孝徳，西野琢也，矢野翔太郎，海野由紀，古川知快，鳥居　悟，伊豆哲也，金谷延幸，津田　侑，井上大介：テンソルデータ拡充を用いた組織内ネットワーク攻撃判定方式，暗号と情報セキュリティシンポジウム (SCIS) (2019).

[7] Miyato, T., Maeda, S., Koyama, M. and Ishii, S.: Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 41, No. 8, pp. 1979–1993 (2019).

[8] 海野由紀，森永正信，及川孝徳，古川和快，金谷延幸，津田　侑，遠峰隆史，井上大介，鳥居　悟，伊豆哲也：標的型攻撃の被害範囲を迅速に分析するネットワークフォレンジック手法の改良，コンピュータセキュリティシンポジウム (CSS) (2018).

[9] Maruhashi, K., Todoriki, M., Ohwa, T., Goto, K., Hasegawa, Y., Inakoshi, H. and Anai, H.: Learning Multi-Way Relations via Tensor Decomposition With Neural Networks, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Press, pp. 3770–3777 (2018).

[10] Harris, A. and Levitz, G.: *ATA Suspicious Activity Playbook*.

[11] Ribeiro, M. T., Singh, S. and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D. and Rastogi, R., eds.), ACM, pp. 1135–1144 (2016).

[12] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks, *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net (2018).

[13] 津田　侑，遠峰隆史，金谷延幸，牧田大佑，丑丸逸人，丑丸逸人，神宮真人，高野祐輝，安田真悟，三浦良介，太田悟史，宮地利幸，神薗雅紀，衛藤将史，井上大介，中尾康二：サイバー攻撃誘引基盤 STARDUST，コンピュータセキュリティシンポジウム (CSS) (2017).

[14] 荒木粧子，笠間貴弘，千葉 大紀充弘，寺田真敏：マルウェア対策のための研究用データセット～ MWS Datasets 2019 ～，情報処理学会, Vol.2019-CSEC-86, No.8, 2019 年 7 月 (2019).

[15] Haq, T., Moran, N., Vashisht, S. and Scott, M.: *Operation Quantum Entanglement* (2014).

[16] 石川芳浩：PowerShell Empire を利用した標的型攻撃 (2017).