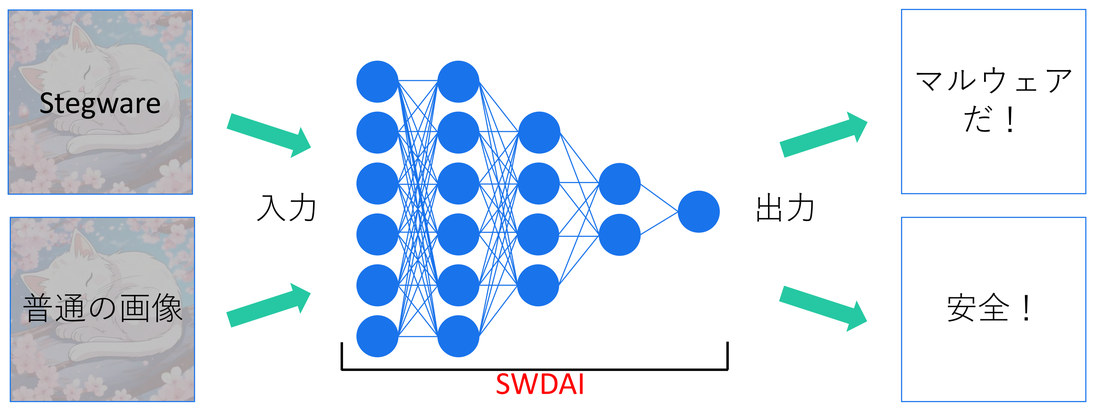
画像に埋め込まれたマルウェアを検知するAI「SWDAI」

学習駆動コース　坂井ゼミ　有賀菜摘

SWDAIとは

Stegwareを画像形式のまま検知する学習済みモデルのこと。Stegwareとは、ステガノグラフィーを用いて画像にマルウェアを埋め込むという難読化を施したマルウェアのこと。SWDAIは、StegWare Detection Artificial Intelligenceの略。普通の何の変哲も無い画像や悪意のないステガノグラフィーが施された画像等と区別してマルウェアが埋め込まれた画像のみを検知することができる。学習に利用したネットワークモデルはDNN(Deep Neural Network)、学習方法は教師あり学習である。学習・検証データとして利用したマルウェアのソースはVirusTotal、画像のソースは、自身で撮影・収集したもの。

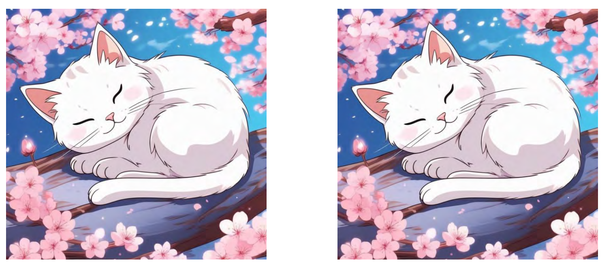
SWDAI概要図



学習・検証データの詳細

・学習データ　計91,000個
　Embedded Data : malicious file・・・45,000個
　　: benign file・・・・45,000個
　Cover Data : 90,000個
　Normal Image Data : 1000個
・検証データ　計11,000個
　Embedded Data : malicious file・・・5,000個
　　: benign file・・・・5,000個
　Cover Data : 10,000個
　Normal Image Data : 1000個
* Cover DataとNormal Image Dataはデータ拡張により数量を確保している

画像の比較



normal　1204KB　　steganography　1206KB

見た目の違いはほぼ無し

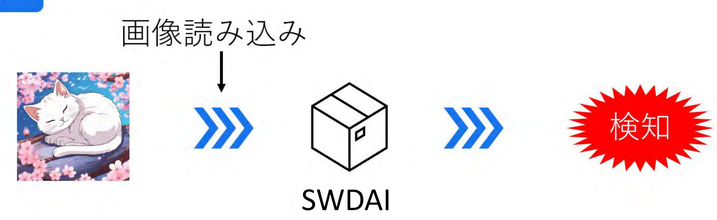
SWDAIの詳細

・入力層 : 1層・・・ニューロン数 : 1024個
・中間層 : 3層・・・1層目ニューロン数 : 1024個
　　・・・2層目ニューロン数 : 512個
　　・・・3層目ニューロン数 : 64個
・出力層 : 1層・・・ニューロン数 : 1個
・利用した活性化関数 : 中間層・・・ReLU関数
　　出力層・・・sigmoid関数
　　* 中間層では、各層で正規化を行っている
・エポック数 : 10個
・バッチサイズ : 2048個
・データセット : 50,000個

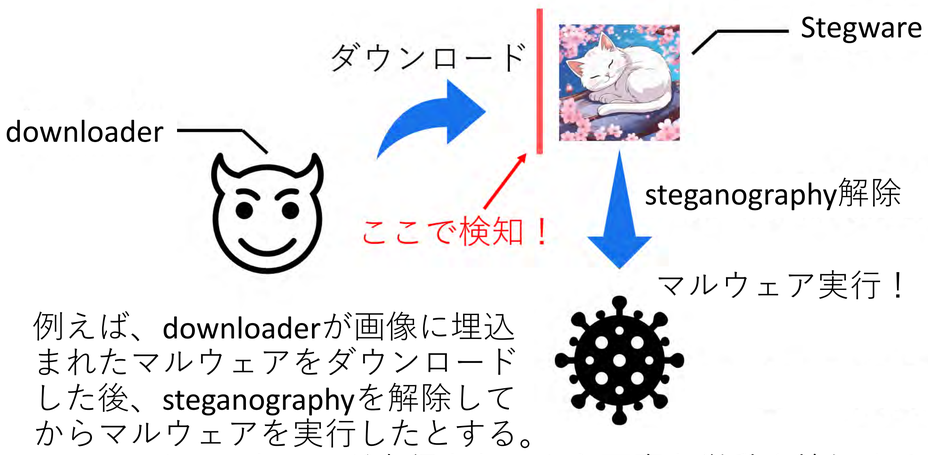
SWDAIはここが違う！

既存　Monika, A. & R. Eswari. *An Ensemble-Based Stegware Detection System for Information Hiding Malware Attacks.* *Journal of Ambient Intelligence and Humanized Computing*, 24 February 2023. etc.

SWDAI



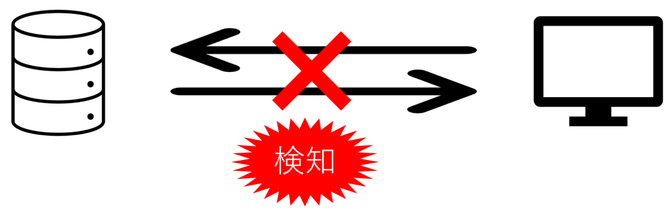
既存研究では、steganalysisとsteganography(難読化)解除ツールを組み合わせ、隠蔽部分を抽出した上でマルウェアを検知する。SWDAIは、steganalysisも難読化解除ツールも使用せず、画像形式のままマルウェアを検知可能である。



例えば、downloaderが画像に埋込まれたマルウェアをダウンロードした後、steganographyを解除してからマルウェアを実行したとする。antivirusはマルウェアが実行されてから不審な挙動を検知するが、SWDAIは画像形式のままマルウェアを検知可能なため、downloaderが画像をダウンロードした直後に検知できる。

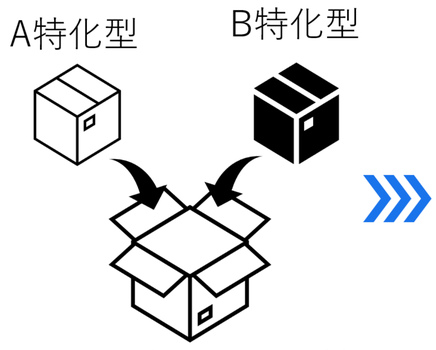
SWDAIの将来性

パケットキャプチャ × SWDAI



パケットキャプチャとSWDAIを組み合わせることで、送受信中にマルウェア検知が可能となり、実際にダウンロードや送受信が完了する前にマルウェアを検知して止めることができると考えられる。送受信完了前に止めることができれば、ゾンビコンピュータ攻撃によるマルウェア拡散を未然に防いだり、マルウェア感染によるリスクや被害を最小限に抑えることが可能になる。

特化型SWDAI × 特化型SWDAI



AとBの**圧倒的**な**検知精度の高さ**を**実現**！

汎用性・実用性UP！

ステガノグラフィーの様々な手法によって難読化された様々なマルウェアに特化したSWDAIを作成し、組み合わせることで、汎用性が高く、実用的なマルウェア検知AIを実現できると考えられる。1つのSWDAIでは、特定の範囲における検知精度は圧倒的であり、それぞれ特定の範囲に特化したSWDAIを組み合わせることで、汎用性を向上させつつ、圧倒的な検知精度を保つことができる。

SWDAIはこう使える！

1. 画像ファイルを1つ1つ診断する
   画像ファイルを入力すると、その画像ファイルがStegwareであるか否かについての判断と、判断結果の確度を出力する
2. ディレクトリを全探索して悪意のある画像ファイルを検知する
   環境内のディレクトリを全て探索して、全ての画像ファイルを診断し、Stegwareとして判断された画像ファイルのパスと判断結果の確度を出力する
3. 指定ディレクトリを定期診断
   プログラム内でディレクトリを指定して実行すると、任意の頻度で指定したディレクトリ内に含まれる画像ファイルがStegwareであるかを診断し、その判断結果と確度を出力する
4. 画像ファイルダウンロード時に診断
   Downloadフォルダを指定してプログラムを実行すると、そのフォルダに画像ファイルがダウンロードされたとき、即座にStegwareであるかを診断し、Stegwareであればアラートウィンドウを表示する

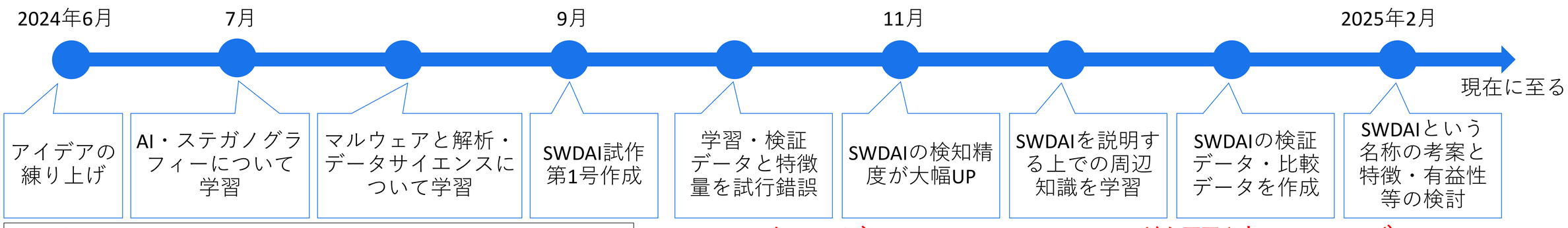
SWDAIのすごい所

Validation AUCは約0.945！
マルウェアの検知精度は96％超！

Stegwareの脅威と検知困難性について

Stegwareは、マルウェアを画像や音声ファイル等のメディアファイルに隠蔽したマルウェアであるため、従来のantivirusでは検知困難である。検知困難な理由として、データが隠蔽されているためコードを検出しにくい、メディアファイルに埋め込まれるためファイル形式が一見無害であり圧縮ファイルであれば解読されない、動的挙動が見えにくい等がある。

SWDAI作成の過程




```
model.fit_generator(
    validation_data=validation_data, generator=training_generator,
    steps_per_epoch=num_obs_per_epoch / batch_size, epochs=10,
    callbacks=[callbacks.ModelCheckpoint(
        save_best_only=True,filepath='results/best_model.h5',monitor='val_loss')],
    verbose=1)
```

ソースコードの一部

ステガノグラフィー

Cover Dataの中にEmbedded Dataを埋め込むことで情報を隠蔽する技術のこと。Cover Dataは画像・音声・動画・テキストファイル等のメディアが使われ、様々な方法で情報が埋込まれる。著作権保護や不正コピー抑止に役立つデジタルウォーターマークなどに利用されている。

学習済みモデル

既にデータを使って、特定のタスクを実行できる状態になっている、AIのネットワークモデルのこと。ネットワークモデルとは、データを学習するためのアルゴリズムや数学的構造のこと。学習済みモデルは、他のタスクに再利用したり、効率的に学習するために利用したりできる。