EX19502（名古屋大学推薦課題）

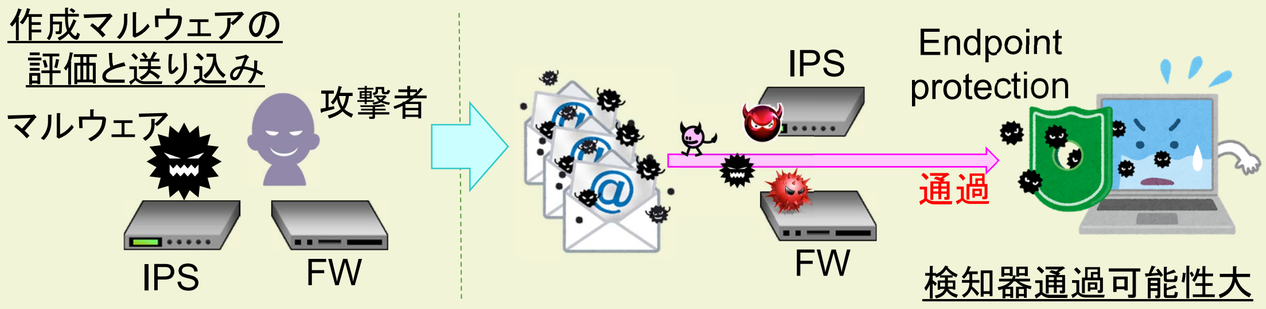
嶋田 創（名古屋大学）

特定マルウェアのみの検知逃れを実現する敵対的学習とその対抗手法

背景

- 新規に作られるマルウェアは増えている
- 攻撃者は事前にセキュリティ機器でマルウェアの検知可否を確認することも可能

→全世界共通なマルウェア検知器の有効性低下



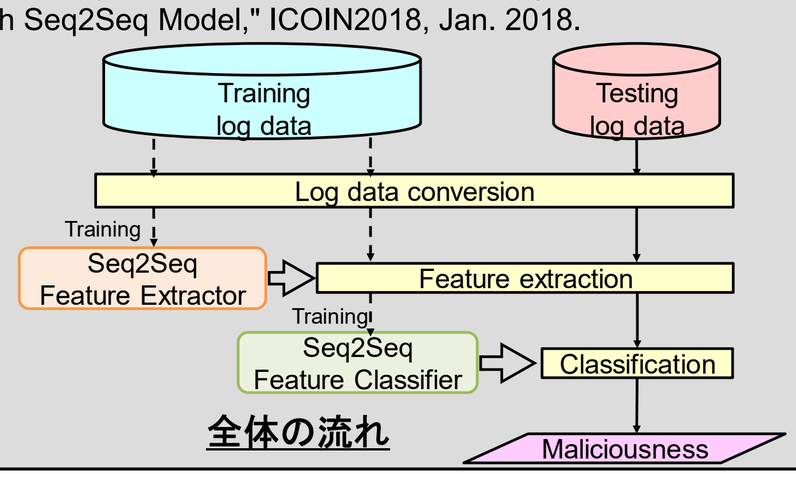
過去の研究

ローカルで学習を行った検知器(攻撃者が事前評価を行えない)を利用して、全世界共通のマルウェア検知器を補完

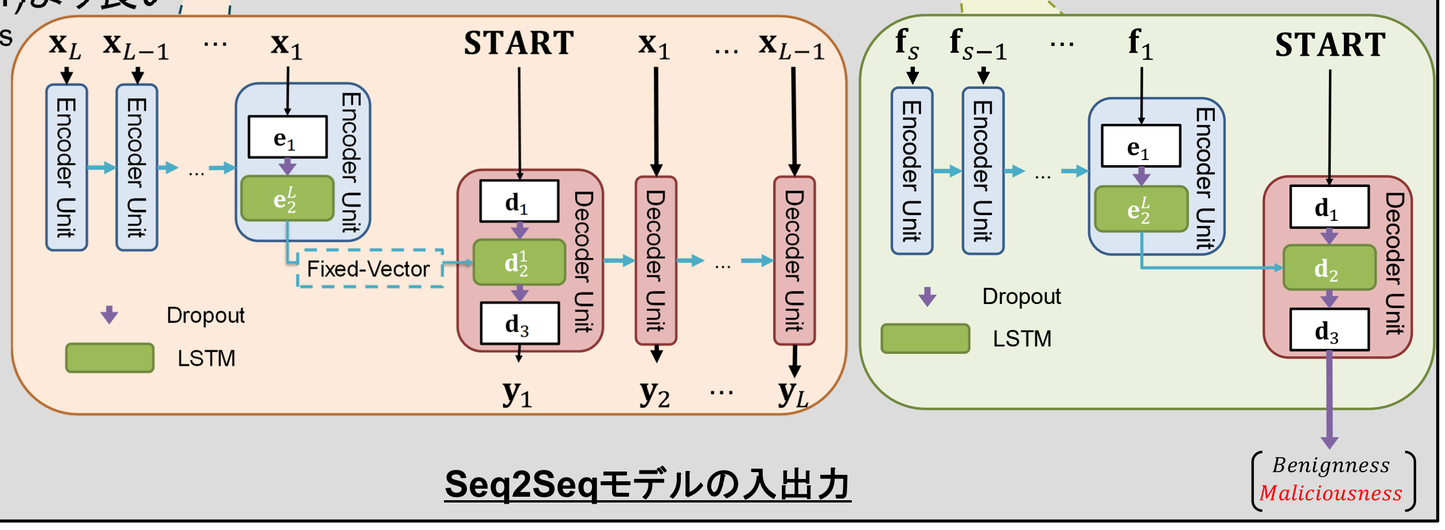
→正常指定したアプリとマルウェアデータから検知器作成

- 過去研究例: 2段Seq2Seqモデルによるマルウェア検知[1]
- 1段目のSeq2Seqモデルでプロセスの APIコールを要約
- 2段目のSeq2Seqモデルが要約した学習用データの学習および評価用データの識別
- 結果
  - FPR/TPRをxy軸としたときのAUCで0.979
  - 同一データでTF(AUC=0.936)やBoW(同0.974)より良い

[1] S. Tobiyamaら, "A Method for Estimating Process Maliciousness with Seq2Seq Model," ICOIN2018, Jan. 2018.



全体の流れ



Seq2Seqモデルの入出力

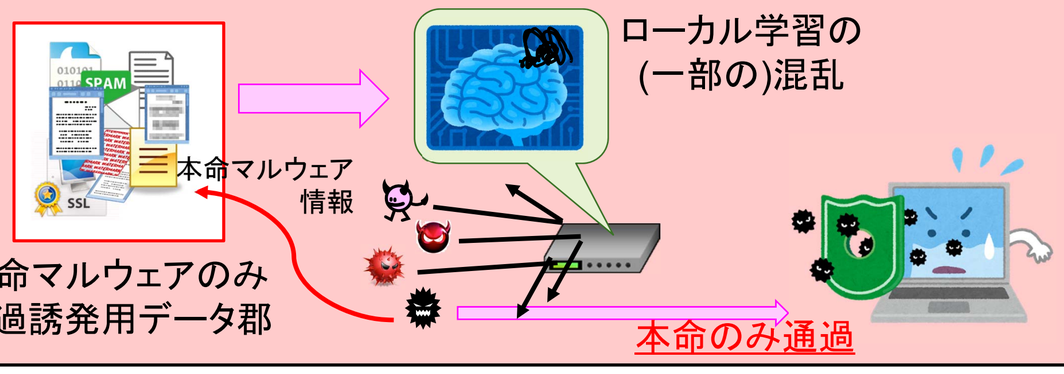
想定する新たな脅威

ローカル学習に対して、攻撃をしかけてくることが想定される

- すでに機械学習を混乱させる敵対的学習は広く行われている(例: 画像認識で誤認識を誘発させる)
- すでに機械学習のマルウェア検知を混乱させる試みも実施されている[1]
  - まだマルウェアの検知を等しく混乱させる段階

[1] H. S. Andersonら, "Evading Machine Learning Malware Detection," Black Hat USA 2017, Jul. 2017.

●将来的に、特定マルウェア(攻撃用)のみの検知を混乱させる攻撃が出てくるのでは?



採択課題の目標

- そもそも、特定マルウェアのみの検知逃れがまだ実現されていないので、その可能性の確認研究の実施
- その上で、特定マルウェアのみの検知逃れを誘発する敵対的学習への対抗研究を実施

→多種多様な機械学習の利用を試みるため、JHPCNにおける大規模計算機の機械学習の良い事例になるのでは?