Automatic Mapping of Vulnerability Information to Adversary Techniques

Otgonpurev Mendsaikhan, Hirokazu Hasegawa, Yukiko Yamaguchi, Hajime Shimada

Presenter: Otgonpurev Mendsaikhan

Nagoya University

ogo@net.itc.nagoya-u.ac.jp

NAGOYA UNIVERSITY

Presenter resume

OTGONPUREV MENDSAIKHAN (Student Member, IEEE) received the M.S. degree in information security policy and management from Carnegie Mellon University, Pittsburgh, PA, USA.

He is currently pursuing the Ph.D. degree with the Graduate School of Informatics, Nagoya University, Japan. His main research interests include cybersecurity situational awareness and cyber threat intelligence.

Outline

- Motivation
- Background
  - Vulnerability Modeling
  - MITRE ATT&CK
  - Multi-label classification
  - Evaluation measures
- Experiment
  - Dataset
  - Text representation
  - Model selection
  - Model evaluation
  - Model analysis
- Conclusion

Motivation

Mapping vulnerability to adversarial techniques

- Software vulnerabilities are increasing rapidly
- Impossible to keep track of all the reported vulnerabilities
- Threat models have been developed to generalize the threat landscape
  - e.g., MITRE ATT&CK
- **Analogy:** MITRE ATT&CK is the playbook of steps that house robber would take to rob a house (e.g., find open access) and software security vulnerability is the weaknesses of the house security (e.g., unlocked door or broken window)

BACKGROUND

Vulnerability Modeling

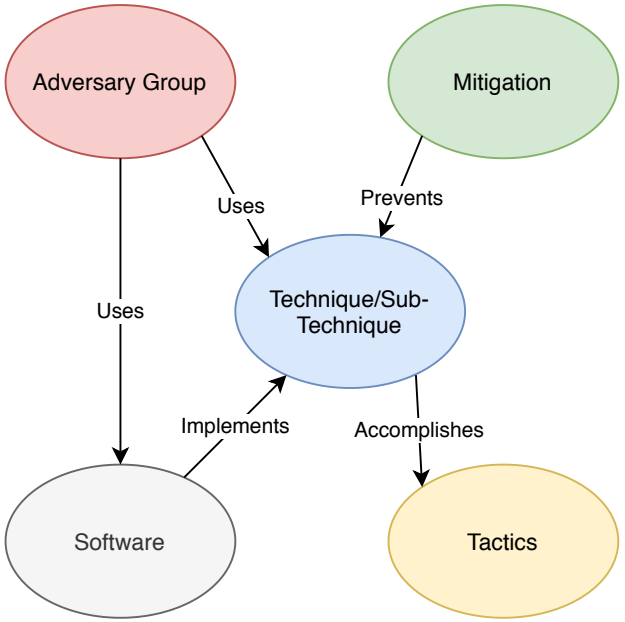
- **Common Vulnerabilities and Exposures (CVE):** is a list of entries, each containing an identification number, description, and at least one public reference for publicly known cybersecurity vulnerabilities
- **Common Attack Pattern Enumeration and Classification (CAPEC):** efforts provide a publicly available catalog of common attack patterns that helps users understand how adversaries exploit weaknesses in applications and other cyber-enabled capabilities.
- **MITRE's Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) framework:** Curated knowledge base and threat model for adversarial tactics and techniques.

MITRE ATT&CK

- Started in 2013 at MITRE corporation to systemically categorize adversary behavior
- Apart from ATT&CK Enterprise there are complementary models such as PRE-ATT&CK, ATT&CK for Mobile, and ATT&CK for ICS
- Constantly enriched with techniques and sub-techniques.
- As of June 2020:
  - 266 techniques/sub-techniques of 12 tactics in the Enterprise model
  - 174 techniques of 15 tactics in the PRE-ATT&CK
  - 79 techniques of 13 tactics in ATT&CK for Mobile

MITRE ATT&CK Components

- **Adversary group:** Known adversaries that are tracked and reported in threat intelligence reports
- **Tactics:** The adversary's tactical objective: the reason for performing an action.
- **Technique/Sub-Technique:** "How" an adversary achieves its tactic, whereas Sub-technique further breaks down techniques into more specific descriptions of actions to reach the goal
- **Software:** An instantiation of a technique or sub-technique at the software level
- **Mitigation:** Security concepts and technologies to prevent a technique or sub-technique from being successfully executed



Multi-label Classification

- Classification is the task of learning to classify the set of examples that are from a set of disjoint labels L, $|L| > 1$
- If $|L| = 2$, binary or single-label classification and if $|L| > 2$, multi-class classification
- Multi-class classification -> single example has single class or label

  Multi-label classification -> single example has more than one labels $Y \subseteq L$

  where: Y - set of labels per example

Multi-label classification categories

- **Algorithm adaptation methods:** The existing machine learning algorithms that are adapted, extended, and customized for multi-label classification problem.

  Eg: boosting, k-nearest neighbors, decision trees, neural networks.
- **Problem transformation methods:** Transforms the multi-label classification into one or more single-label classification or regression problems. Further divided into categories as binary relevance, label power-set, and pair-wise methods.
- **Ensemble classification:** Developed on top of existing problem transformation or algorithm adaptation methods.

  eg: Random k-label sets (RAkEL) and ensembles of pruned sets (EPS) etc.

Evaluation measures

- Evaluation metrics to measure the performance of the multi-label classification is different
- It falls into the following categories:
  - Example based measures
  - Label based measures
  - Ranking based measures

Evaluation measures used in this study

- **Subset Accuracy:** Most strict metric, indicating the percentage of samples that have all their labels classified correctly
- **Micro Averaged F1 Score:** Harmonic mean of micro-precision and micro-recall, the measures averaged over all the example/label pair
- **Macro Averaged F1 Score:** Harmonic mean between precision and recall where the average is calculated per label and then averaged across all labels
- **Hamming Loss:** Evaluates how many times an example-label pair is misclassified
- **Ranking Loss:** Evaluates the average fraction of label pairs that are reversely ordered for the particular example

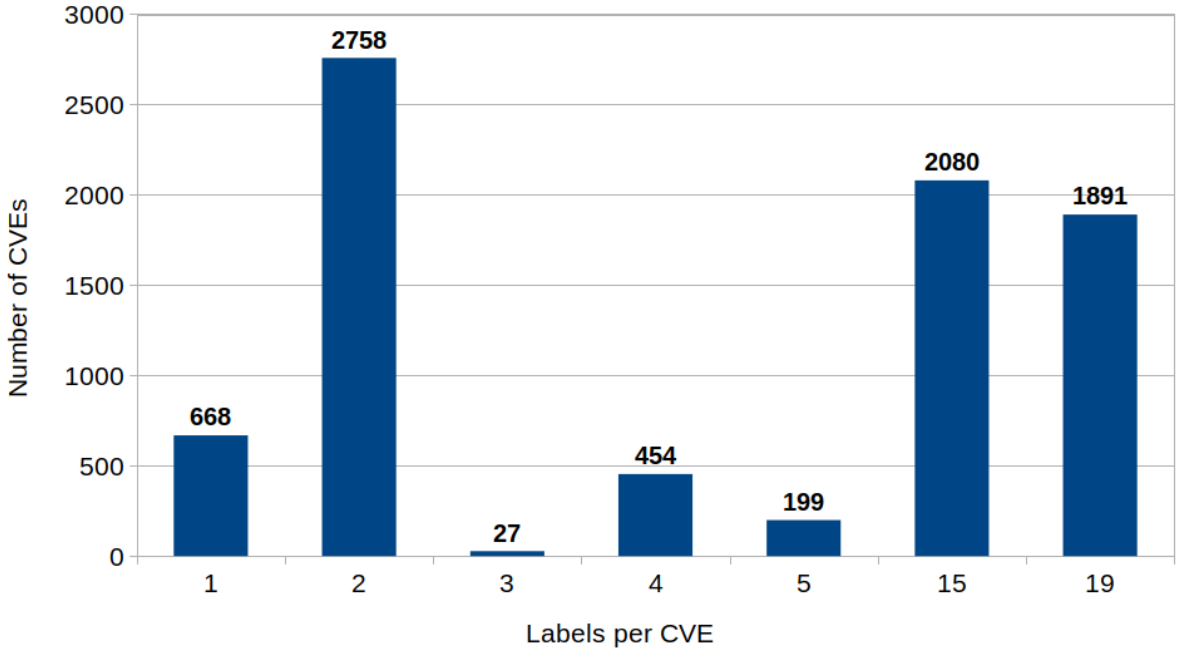
EXPERIMENT

Dataset

- The European Union Agency for Cybersecurity (ENISA) released report “State of Vulnerabilities 2018/2019”
- 27,471 vulnerability information during 1st January 2018 to 30th September 2019 has been collected as part of the report
- ENISA did analysis on the vulnerability information and mapped CVEs to MITRE ATT&CK Technique using CAPEC information

Dataset

- 8,077 CVEs mapped to 52 Techniques
- Dataset cardinality: 9.43
- Dataset density: 0.18
- Distributed into 7 discrete buckets of Technique combinations as shown



Dataset

- CVE descriptions of the dataset:
  - Min length: 40 char
  - Max length: 3,655 char
  - Oldest CVE: CVE-2007-6763
  - Newest CVE: CVE-2019-9975
- Out of total 8,077 examples:
  - 7,877 for train and evaluate the model
  - 200 for validate and analyze the model

Text representation

- Need to convert CVE descriptions into numerical vectors
- Universal Sentence Encoder (USE) from Google Research is used
- USE has good task performance with little task specific training data
- Deep Averaging Network (DAN) based USE model has been chosen
- CVE description converted into fixed 512 dimensional vector representation

Model selection (traditional)

- Algorithm adaptation methods:
  - Multi-label k-nearest neighbors (MlkNN)
- Problem transformation:
  - LabelPowerset
  - ClassifierChain
  - BinaryRelevance
- Ensemble:
  - Random k-laELsets multi-label classifier (RAkELd)

Model selection (neural)

- Since neural networks has been proven to be superior in almost every task neural approaches have been tested as well
- Multi-label classification task doesn't require the sequential input or memory state of the input, a simple Multilayer Perceptron (MLP) neural models have been tested
  - LabelPowerset(neural) MLP as base classifier, 2 hidden layers and softmax activation
  - BinaryRelevance(neural) MLP as base classifier, 2 hidden layers and sigmoid activation

Model evaluation

- 10 fold cross validation result

Model	Accuracy score	Micro Average F1 score	Macro Average F1 score	Hamming loss	Ranking loss
MlkNN	0.6138	0.6740	0.5576	0.1079	0.3595
LabelPowerset	0.6133	0.6369	0.5174	0.1157	0.3654
ClassifierChain	0.5036	0.6089	0.4243	0.1427	0.4298
BinaryRelevance	0.3744	0.6158	0.4907	0.1471	0.3263
RakelD	0.4237	0.6230	0.5024	0.1340	0.3411
LabelPowerset (neural)	**0.7432**	**0.7452**	**0.6396**	0.0911	**0.2448**
BinaryRelevance (neural)	0.5538	0.7426	0.6279	**0.0883**	0.2885

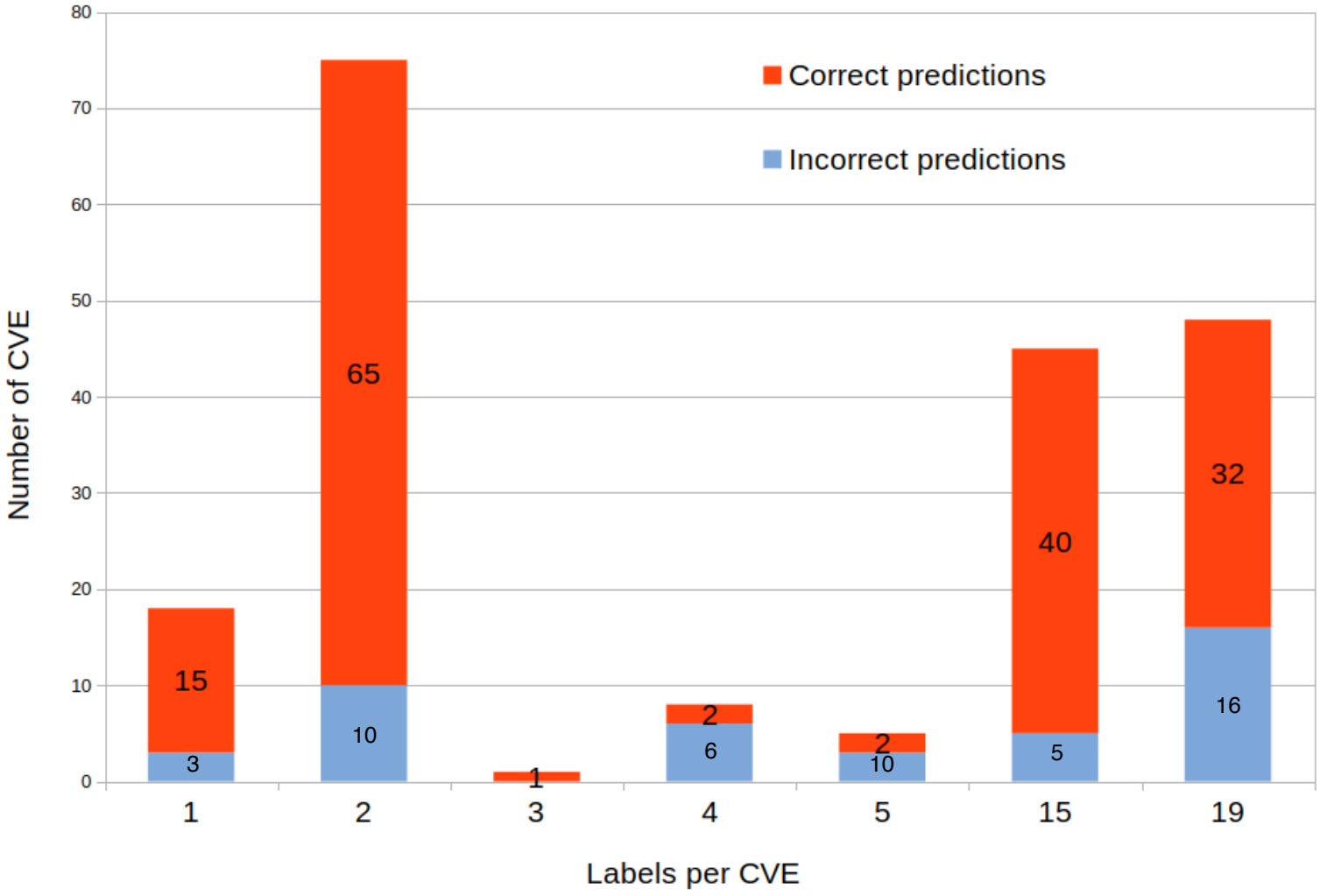
Model evaluation

- Neural models have better performance than traditional models
  - Neural LabelPowerset model has best scores except in Hamming loss
  - Neural BinaryRelevance model has best score only in Hamming loss
- The results fall within acceptable range when compared to other benchmarking study results
- Thus Neural LabelPowerset model has been chosen as best performing model

Model analysis

- Best performing neural LabelPowerset model is trained with 7,877 examples
- Model analysis by predicting the validation dataset of previously unseen 200 examples

Prediction results



Model analysis

- More labels, more incorrect predictions
- Probably due to the skewed training data
  - The dataset consists of 8,077 examples
  - Mapped to only 52 adversarial techniques of 266 techniques
  - Distributed to only 7 different buckets

CONCLUSION

Conclusion

- We proposed a novel approach to map the vulnerability information to adversary techniques
- We converted vulnerability description into vector space and applied various multi-label classification methods
- We found that neural LabelPowerset method performs best in our experiment
- Due to the partial nature of the dataset the experimental result could not be fully tested
- However the chosen methods show good performance, indicating comprehensive dataset may yield production-ready system
- In the future, we would like to build comprehensive dataset by correlating CAPEC information

THANK YOU FOR YOUR ATTENTION