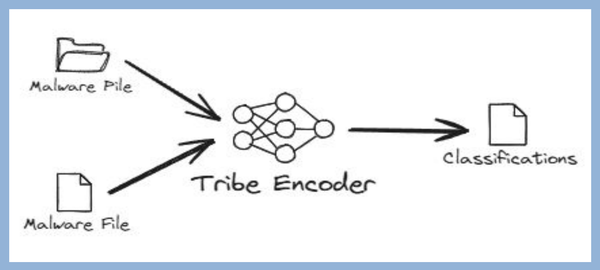
A Quest Called **TRIBE**: Clustering Malware Families for Enhanced Triage and Analysis

MIDN 1/C: Michael Chen, John Jenness, Chris Kim, Justin Liaw
Dr. Dane Brown, Cyber Operations Department
CDR. Edgar Jatho, Computer Science

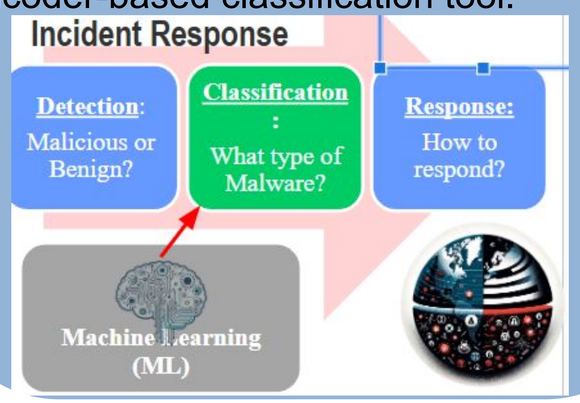
Abstract

We propose using a transformer-based sequence-to-sequence autoencoder to classify malware binaries.This novel approach will classify malware into Tribal Relation Inferential Binary Encoder (TRIBE) clusters which will reduce the total number malware family classifications.



Importance

Polymorphic and mutational malware has led to the rise in number of extraneous antivirus labels[11]. Our goal is to reduce the time it takes for Security Operation Centers (SOCs) to classify malware by optimizing labeling and produce an encoder-based classification tool.

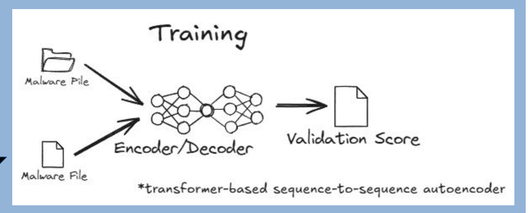


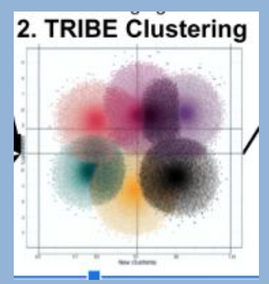
Methodology

1. **Obtain large set of labeled executable malware binaries**
2. **Create autoencoder (sequence-to-sequence and transformer based)**
3. **Determine the most optimal groupings of TRIBES**
4. **Build a transformer-based malware classifier using the TRIBES**
5. **Productizing the Classifier**

ClamAV

ABUSE.ch





Product Details

Our end deliverable will be a Dockerized command line interface tool that can take either a binary file or a folder of binary files and returns the classification of each binary. This tool will arrive pre-trained.

docker

Conclusions

We hope to find that by utilizing an unsupervised autoencoder we can “see” around the obfuscation presented by polymorphism and mutation in order to reduce the total number of classification malware can be grouped into and create an efficient classification tool.

Acknowledgement

Trident Research Program:
Provided $8600 GPU for malware.

GPU

Future Work

- Integration of recommended course of action based on classification
- User based modification of autoencoder and end-product trainability.

References

Literary Review: