SAND2022-9781D

The Center for Cyber Defenders
Expanding computer security knowledge

EMLAT: Explainable Machine Learning for Alert Triage

Alycia N. Carey – University of Arkansas, Ph.D. Fair Machine Learning, Dec. 2024
Akul A. Goyal – UIUC, Ph.D. Cybersecurity, May 2024
Thomas R. Quig – UIUC, M.CS. Computer Science, Dec. 2022
Manager: Tiawna L. Cayton 056831, Mentor: Eric L. Goodman 05555

Problem Statement

- Alert triage plays an important role in the protection of computing systems by determining the severity of threats –costly in time and resources required
- Machine learning (ML) has been shown to be viable in classifying alerts as true positive (TP) or false positive (FP) – no intuition behind classification
- Goal 1: Use explainable ML techniques to classify alerts from SCOT (Sandia Cyber Omni Tracker) as TP or FP and provide reasoning for classification
- Goal 2: Use explainable ML to determine if a specific packet is malicious or not (based on raw packet bytestreams) and provide reasoning for classification

Objectives and Approach

- Use transformer based ML architectures and explainable ML techniques to analyze the alerts/packets
  - Transformers are deep learning models that use self-attention to assign significance to input data
  - Can determine alert type, if it should be promoted (TP) or closed (FP), and what features of the alert contributed most to the classification
  - Gradient-based explanation techniques – simplest and most popular
- Create alert data through SNORT on the ISOT Cloud Intrusion Detection dataset

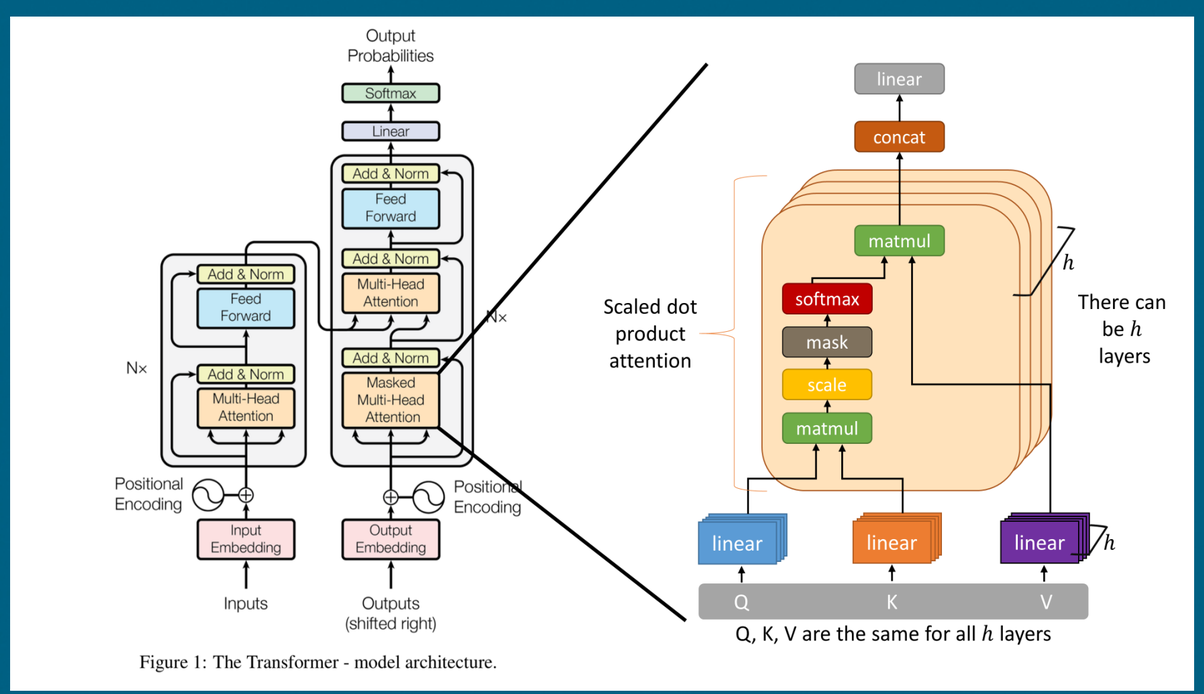


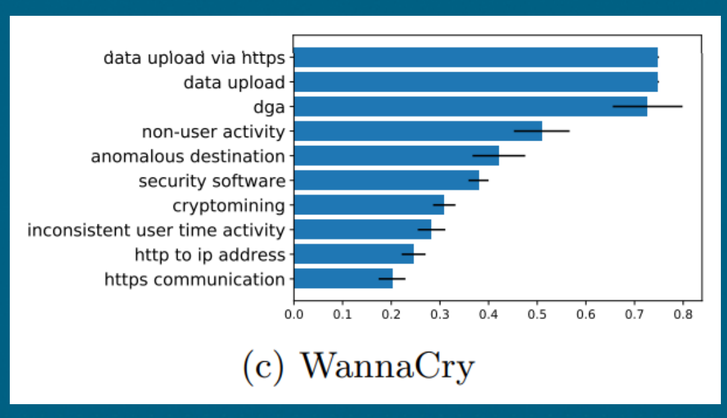
Figure 1: The Transformer - model architecture.

[1] Learning Explainable Representations of Malware Behavior – Prasse et al.

(Expected) Results

- **Alert-Based Analysis:**
  - Effectively filter out FP alerts from SCOT
  - Correctly group alerts that are part of the same attack
  - Predict future alerts based on the attack type by using the model trained on previous alerts

Example outputs for alert-based explanation techniques provided by [1]



(c) WannaCry

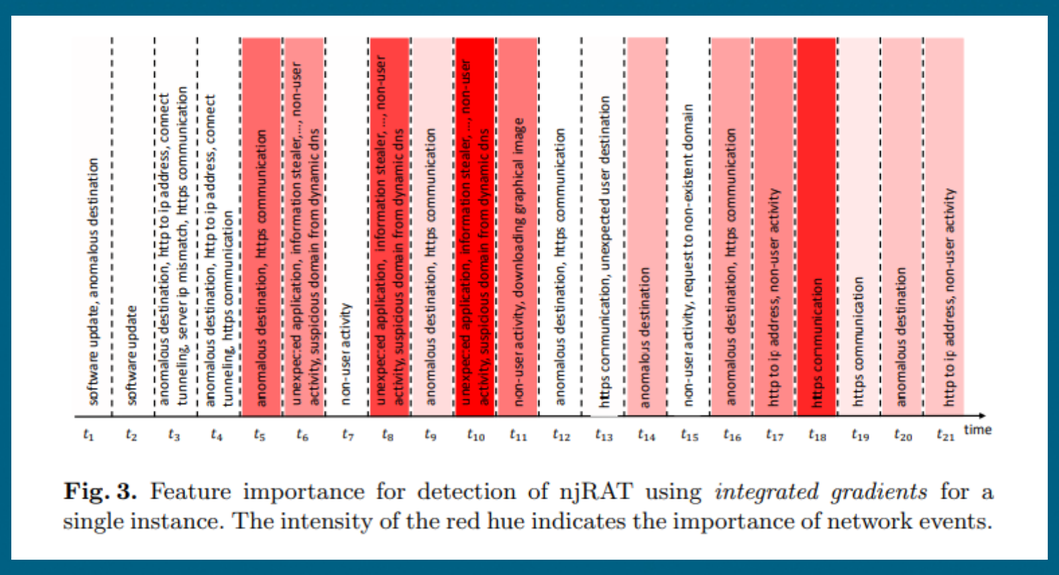


Fig. 3. Feature importance for detection of njRAT using *integrated gradients* for a single instance. The intensity of the red hue indicates the importance of network events.

- **Raw Packet-Based Analysis:**
  - Expect the model to operate on general packets and return a classification for maliciousness
  - Expect to find higher accuracy for malicious packet detection on the ISOT dataset than our previous naïve approach "packet2vec"
  - Expect that there will still be overfitting issues on metadata such as timestamps
  - Further optimization is required for generalization
  - Future work: will integrate additional explainability features into the model

Impact and Benefits

- Saves time and resources – analyst can focus on TP alerts rather than FP
- Provides an explanation into why an alert was fired, giving the analyst a place to start digging

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2017-XXXXC.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.