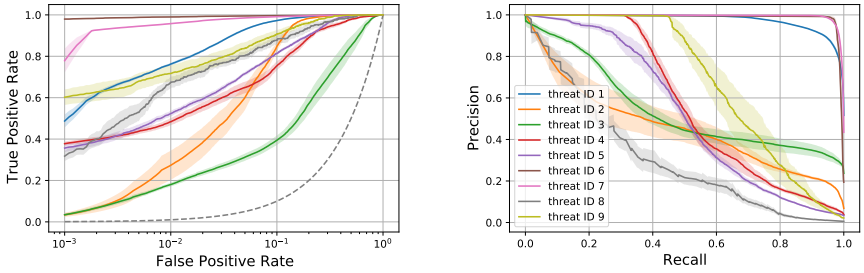
Appendix for Learning Explainable Representations of Malware Behavior

Paul Prasse[1], Jan Brabec[2], Jan Kohout[2],
Martin Kopp[2], Lukas Bajer[2], and Tobias Scheffer[1]

[1] University of Potsdam, Department of Computer Science, Germany
{prasse,scheffer}@uni-potsdam.de
[2] Cisco Systems, Cognitive Intelligence, Prague, Czech Republic
{janbrabe,jkohout,markopp,lubajer}@cisco.com

A Detailed Anaysis of Malware-Classification Performance

This section reports on detailed analysis results by measuring precision-recall curves for the detection of threat IDs, malware families, and malware categories. Precision—the fraction of alarms that are not false alarms—directly measures the amount of unnecessary workload imposed on security analysts, while recall quantifies the detection rate. We also compare the models in terms of ROC cuves because these curves are invariant to class ratios.



(a) ROC curves (log-scale for FPR) (b) Precision-recall curves

Fig. 1. ROC and precision-recall curves for the detection of threat IDs on the evaluation data. The dashed line shows the performance for random guessing and the colored bands the standard error.

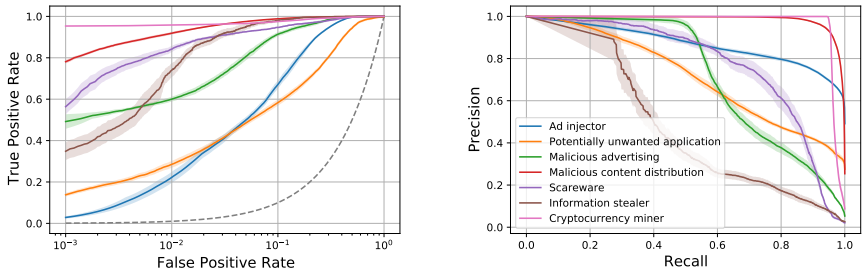
A.1 Threat ID Evaluation

Figure 1 shows the ROC and the precision-recall curves for detecting different threats using the transformer model. As Figure 1a shows, threat IDs that have

a one-to-one relationship with a malware category (threat IDs 6, 7, 8, 9) can be detected more easily than threat IDs that share the same malware category with several other threat IDs. We see that the threat IDs 2, 3, and 4 that belong to the category of *potentially unwanted application* are much harder to detect than other threat IDs. We can be explained by the similar behavior of multiple threat IDs within a category. In total, the transformer is able to detect 6 out of 9 threat IDs with a precision of 80% and a recall of at approximately 40%.

A.2 Malware-Category Evaluation

Figure 2 shows the ROC and the precision-recall curves for detecting different malware categories using the transformer model. The high-prevalence malware categories potentially *unwanted application* and *ad injector* are much harder to detect than the rare malware categories *cryptocurrency miner* and *malicious content distribution*. This can be explained by the larger behavioral variations in the frequent categories.Regarding the precision-recall curves, we can conclude that the transformer is able to detect 6 out of 7 malware categories with a precision of 80% with a recall higher than 40%.



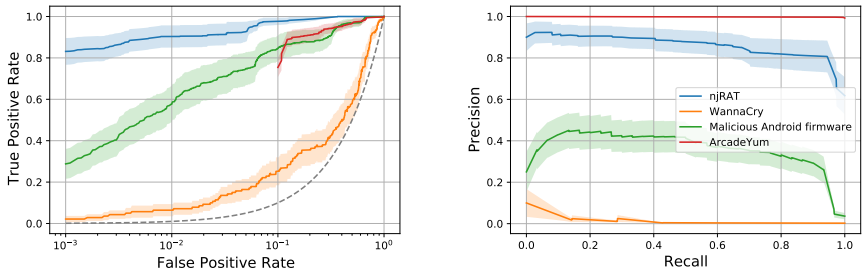
(a) ROC curves (log-scale for FPR) (b) Precision-recall curves

Fig. 2. ROC and precision-recall curves for the detection of malware categories on the evaluation data. The dashed line shows the performance for random guessing and the colored bands the standard error.

A.3 Malware Family Evaluation

Figure 3 shows the ROC and the precision-recall curves for detecting different malware families using the transformer model. Because of the highly unbalanced class ratios, attention should be given to the ROC curves of Figure 3a; the precision-recall curves of Figure 3b are less informative, but are included for completeness. We can conclude that the transformer performs best on the two information stealers. Because we only observe under 100 instances not belonging

to the malware family ArcadeYum, we can only draw the ROC curve up to a FPR of 0.1. In general, we see that the transformer is able to distinguish between malware families. Only the detection of WannaCry is significantly worse; this finding is plausible because especially newer versions of WannaCry create minimal network traffic.



(a) ROC curves (log-scale for FPR) (b) Precision-recall curves

Fig. 3. Malware Family Evaluation. Performance for models on test set. The dashed line shows the performance for random guessing and the colored bands the standard error.