EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models

Hyrum Anderson, Phil Roth
Endgame.

https://github.com/endgameinc/ember

Overview

EMBER is a benchmark dataset and model created to accelerate research in *static malware classification*:

- **dataset:** labels, sha256 hashes, and extracted features from 1.1M Windows executable (PE) files collected in 2017.
- **model**: LightGBM model trained with *default parameters*, achieving a ROC AUC of 0.999112 on the test set.
- **codebase:** allows researchers to calculate features for new PE files, classify them, or modify / append to the feature set.

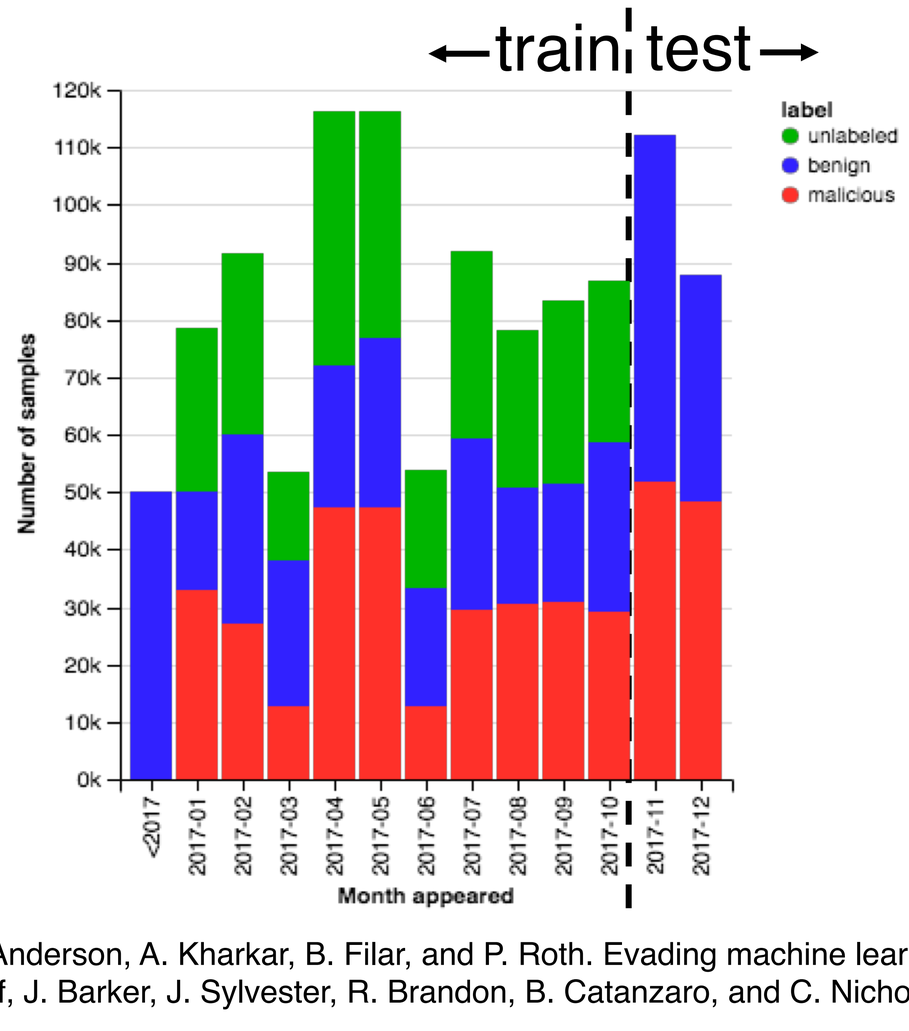
Feature Groups

Strings	*No PE file format knowledge required for calculation.*
Byte Histogram	
Byte Entropy Hist	
General/Header	*LIEF is required to parse the PE file format before calculating these features.*
Import/Export	
Section	

https://lief.quarkslab.com/

Temporal Split

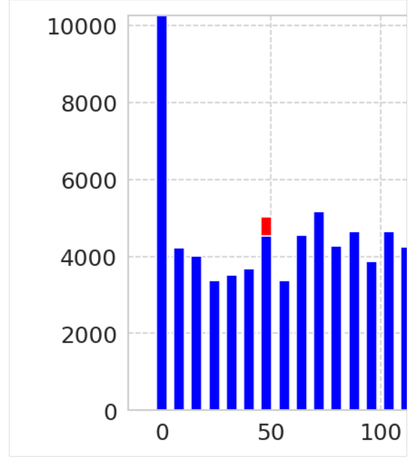
Training set appears before test set to reflect the evolving and adversarial nature of the problem.

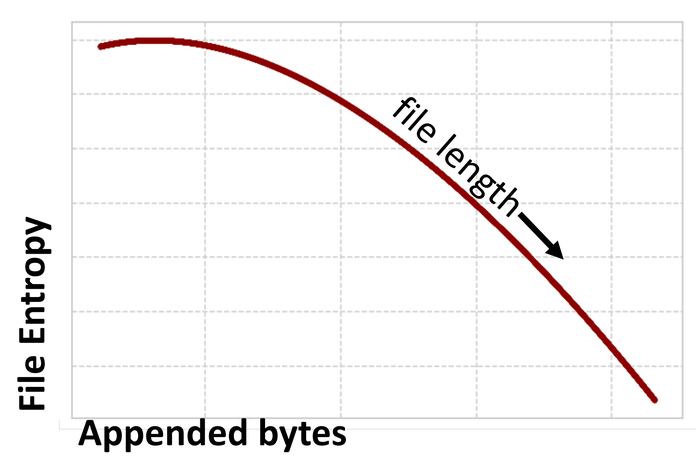


Example Research Challenges

Adversarial ML for Malware

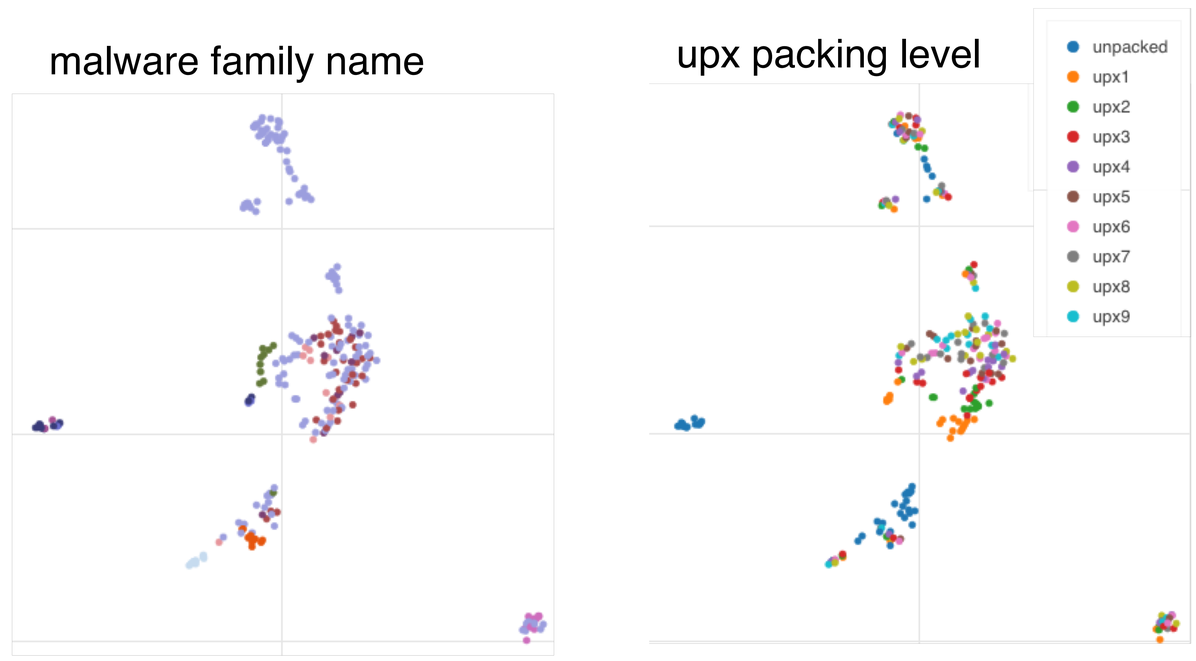
Unlike images, size of adversarial tweaks may be unbounded as long as they are functionality preserving [1]. L_p ball is not a realistic threat model in this domain. Here, adding a single byte (0x30) to the end of a file is shown.





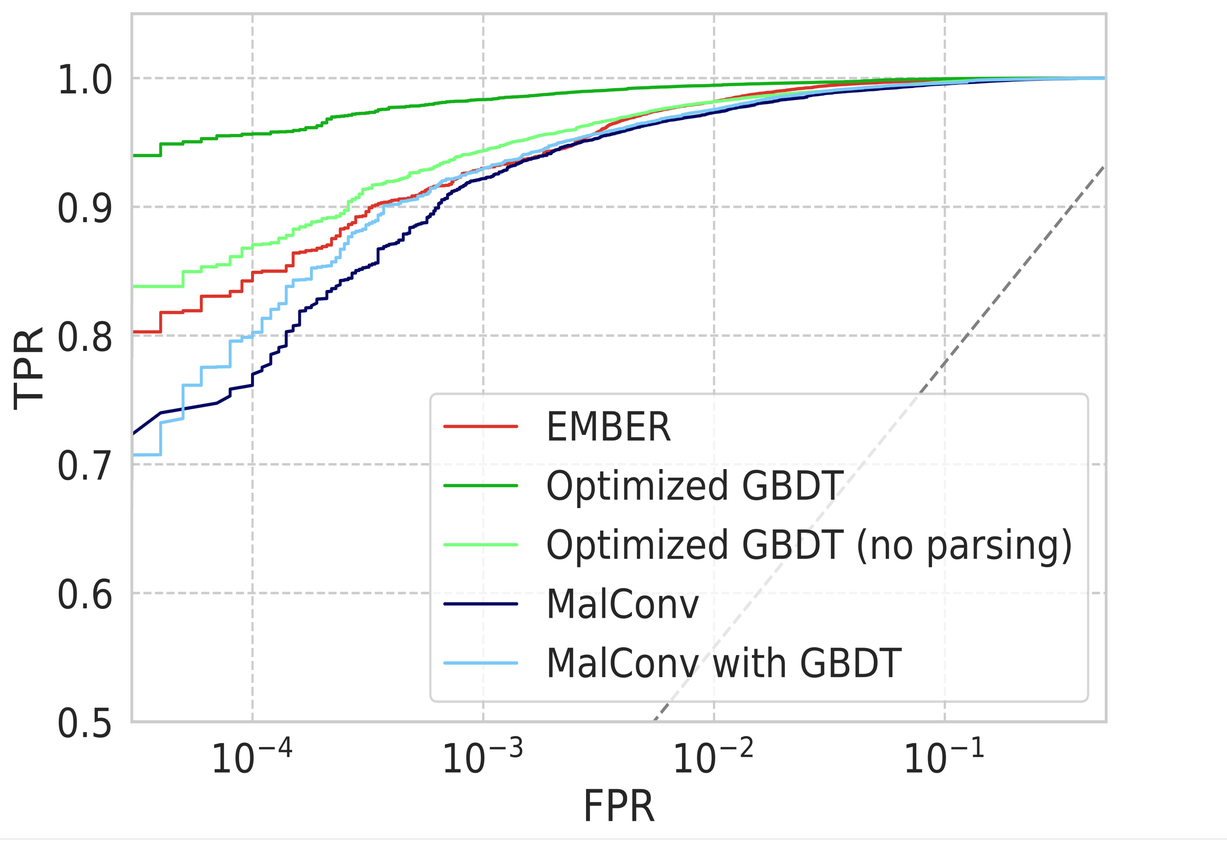
Representation Learning

Representations can be learned from ember feature vectors, and they can be used to distinguish between samples with features that relate to different things (ie. high entropy could indicate malicious or packed)



End-to-end Deep Learning

MalConv [2] is a featureless neural network framework for static malware classification. After optimizing LightGBM parameters, GBDT models beat MalConv performance even without using PE file format features. In this problem domain, feature engineering continues to beat featureless models.



[1] H. S. Anderson, A. Kharkar, B. Filar, and P. Roth. Evading machine learning malware detection. In *Black Hat*, 2017.
[2] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas. Malware detection by eating a whole exe. *arXiv preprint arXiv:1710.09435*, 2017.