Malware Antivirus Scan Pattern Mining via Tensor Decomposition

Prajna Bhandary*[1], Iman Adetunji[1], Auguste Kiendrebeogo[1], Colin Vieson[1], Robert J. Joyce[1], Maksim E. Eren†[1], and Charles Nicholas‡[1]

[1]Department of Computer Science and Electrical Engineering, UMBC
Baltimore, MD, USA 21045

Mar 4, 2022

1 Abstract

Accurate labeling is important for detecting malware and building reference datasets which can be used for evaluating machine learning (ML) based malware classification and clustering approaches. Labels obtained from Anti-Virus (AV) vendors (such as *Kaspersky*, *Malwarebytes*, and *McAfee*) are one source of information; however, despite ongoing research efforts there is still inconsistency with the labeling across AV vendors [7, 5, 9]. AV vendors use differing formats and naming conventions when reporting labels of malware samples, and the reported labels between any two vendors can disagree. We address this problem in our work utilizing CP-APR, a powerful tensor decomposition method for unsupervised ML, to discover the hidden patterns across AV vendors in the way they report the malware labels. In comparison to the traditional ML methods, tensor decomposition models the multi-dimensional properties of the data and produces interpretable results [6]. The higher-dimensional representation of the AV scans enables the discovery of multi-faceted and complex details of those scans.

A subset of the collection of +25 million VirusTotal reports for malware from the VirusShare corpus (which we call the VirusShare-VT dataset) is used for this research [1, 8]. The VirusShare-VT dataset consists of malware scan reports from AV vendors, where each report has aliases for the malware if detected. An example of an alias can look like *"Trojan.Win32.Backdoor"*. In this example, the tokens from the scan result would be *Trojan*, *Win32*, and *Backdoor* where the token *Trojan* is located at position 0, *Win32* at position 1, and *Backdoor* at position 2. The AV scan data is naturally multi-dimensional; therefore, we can represent it using tensors. We build a 3-dimensional

*prajnab1@umbc.edu

†meren1@umbc.edu

‡nicholas@umbc.edu

Abstract Submission for the MTEM 2022 Poster Session.

Expected Level of Materials and Discussion: Unclassified Unlimited Release

Share Presentation Materials: Opt-In

count tensor $\mathcal{X} \in \mathbb{R}^{\text{AV x Tokens x Location}}$, where the dimension *AV* represents each AV vendor, the *Tokens* dimension represents each individual token in the alias, and the *Location* dimension represents the position of the token within the alias. An entry $\mathcal{X}_{a,t,l}$ represents the number of times AV vendor a used the token t at location l across each scanned malware in the VirusShare-VT dataset. We model the tensor $\mathcal{X}$ using Canonical Polyadic Decomposition (CPD) where $\mathcal{X}$ is approximated with a sum of R rank-1 tensors, also called components, such that $\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(AV)} \circ \mathbf{a}_r^{(Tokens)} \circ \mathbf{a}_r^{(Location)}$. Here, $\circ$ is the outer-product and $\mathbf{a}_r^{(AV)}$, $\mathbf{a}_r^{(Tokens)}$, and $\mathbf{a}_r^{(Location)}$ are the latent factors for each dimension. Because tensor decomposition produces interpretable results, we can observe the patterns via visual inspection of each R latent factor. We decompose the tensor $\mathcal{X}$ with CP-APR, a non-negative CPD method [2], specifically the recently published Python implementation with GPU capability [4, 3]. The non-negativity constraint in CP-APR further improves the interpretability.

Our preliminary results allowed us to identify groups of AV vendors with similar labeling patterns. In tensor decomposition, the information across each dimension is analyzed simultaneously, enabling the discovery of complex latent patterns. To be specific, we observed which vendors produced similar results based on the particular tokens, the token family, and the token location. We also found that tensor decomposition, in an unsupervised manner, was able to identify tokens with the same meaning but different reporting format. Our current experiments indicate that utilization of the multi-dimensional patterns can provide more effective and detailed observations as compared to the traditional clustering methods used in the analysis of the AV scan reports [10]. Our future work includes expanding the experiments to the entire VirusShare-VT dataset, using an improved token extraction method, and exploring different tensor configurations.

References

[1] Virusshare.com - because sharing is caring. `https://virusshare.com/`, Last accessed on 2022-3-3.

[2] Eric C. Chi and Tamara G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, December 2012.

[3] M. E. Eren, J. S. Moore, E. Skau, M. Bhattarai, G. Chennupati, and B. S. Alexandrov. pycp_apr. `https://github.com/lanl/pyCP_APR`, 2021.

[4] M. E. Eren, J. S. Moore, E.W. Skau, M. Bhattarai, E. A. Moore, and B. S. Alexandrov. General-purpose unsupervised cyber anomaly detection via nonnegative tensor factorization. *Digital Threats: Research and Practice*, 2022.

[5] Robert J. Joyce, Dev Amlani, Charles Nicholas, and Edward Raff. Motif: A large malware reference dataset with ground truth family labels, 2021.

[6] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[7] Marcos Sebastián, Richard Rivera, Platon Kotzias, and Juan Caballero. Avclass: A tool for massive malware labeling. In *RAID*, 2016.

[8] John Seymour. Labeling the virusshare corpus- lessons learned. 2016.

[9] Yanxin Zhang, Yulei Sui, Shirui Pan, Zheng Zheng, Baodi Ning, Ivor Tsang, and Wanlei Zhou. Familial clustering for weakly-labeled android malware using hybrid representation learning. *IEEE Transactions on Information Forensics and Security*, 15:3401–3414, 2020.

[10] Shuofei Zhu, Jianjun Shi, Limin Yang, Boqin Qin, Ziyi Zhang, Linhai Song, and Gang Wang. Measuring and modeling the label dynamics of online Anti-Malware engines. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2361–2378. USENIX Association, August 2020.