Exploiting the In-Distribution Embedding Space with Deep Learning and Bayesian inference for Detection and Classification of an Out-of-Distribution Malware (Extended Abstract)

1st Tosin Ige
Dept. of Computer Science
The University of Texas at El Paso
Texas, USA
toige@miners.utep.edu

2nd Christopher Kiekintveld
Dept. of Computer Science
The University of Texas at El Paso
Texas, USA
cdkiekintveld@utep.edu

3rd Aritran Piplai
Dept. of Computer Science
The University of Texas at El Paso
Texas, USA
apiplai@utep.edu

***Abstract*—Current state-of-the-art out-of-distribution algorithm does not address the variation in dynamic and static behavior between malware variants from the same family as evidence in their poor performance against an out-of-distribution malware attack. We aims to address this limitation by: 1) exploitation of the in-dimensional embedding space between variants from the same malware family to account for all variations 2) exploitation of the inter-dimensional space between different malware family 3) building a deep learning-based model with a shallow neural network with maximum of two connected layers to overcome overfitting from the scratch 4) building a Bayesian inference based computation algorithm that intertwine with connected network and is able to create new and adjust existing data points in response to an exposure to new out-of-distribution variants of existing or new malware family which determines the extent at which model weight is adjusted which in turn triggers update on the gradient. Preliminary result of our proposed framework gave an accuracy of 81% in the successful classification of a novel out-of-distribution malware attack, something that could not be achieve by any of the state-of-the-art algorithms on novel malware classification.**

The potency of malware to successfully infiltrate any system no matter how sophisticated made it an indispensable tool available to cybercriminals today [3], [4], as malware had proven to be highly successful in the extraction of sensitive data which could be used by cybercriminals against their victim. Several approaches had been widely proposed and adopted to combat the rampant threat of malware attack among which machine learning (ML) and Deep Learning (DL) had been the most promising but the out-of-distribution (OOD) problems had lead to vulnerabilities of machine and deep learning based approaches against novel previously unseen malware family or new variant of an existing family. This is due to the fact that current state-of-the-art approaches are based on the assumption that identically and independently distributed (IID) data will be available in test time which are unfortunately not true in new world scenarios [1], [2]. Hence, the close-world assumption of identically and independently distributed are violated whenever state-of-the-art machine or deep learning based model are deploy in real-world scenarios in the presence of previously unseen out-of-distribution malware family or a novel variant of an existing family, the high failure rate of state-of-the-art approaches to previously unseen OOD malware is cyberattack.

While several SOTA algorithm had been proposed to address the out-of-distribution problems on several benchmark datasets, none of the SOTA classifier had been applied to OOD malware dataset thereby leaving a gap to filled. We started by applying and training SOTA models on 4 different benchmark malware dataset (Sorel, Malevis, Malimg, and Avast) in an out-of-distribution settings, after seeing their poor performance on previously unseen OOD malware, we proceeded to investigate the possible cause by converting each variants in each malware family to bytes and calculating the mean square error (MSE) and vector space representation of each family member, same procedure was also repeated for other dataset on which SOTA models have good performance. Our result shows wide variation between variants of the same malware family in form of wide embedding spaces while other datasets shows little to no variation embedding spaces between samples from same class. It becomes obvious that the failure of the state-of-the-art algorithms to address the in-distributional embedding spaces which is unique to malware is the cause of their poor performance on OOD Malware attack classification. We propose a deep learning and Bayesian inference framework that will address this limitation by: 1) exploitation of the in-dimensional space between variants from the same malware family to account for all variations 2)exploitation of the inter-dimensional space from different malware family 3)building a deep learning model having a shallow neural network containing maximum of two connected layers to overcome overfitting from the scratch 4) building a Bayesian inference based computation algorithm that intertwine with connected network and is able to create new and adjust existing data point in response to an exposure to an out-of-distribution variants.

I. Methodology

A. Dataset

We use all three of MaleVis, Malimg, and Avast CTU benchmark malware dataset. Malevis dataset contains a total of 14,226 malware samples spanning 26 families of malware, and out of which 9100 are training samples while are 5126 validation samples in 3 channels format, Malimg datase contains a total number of 9435 executable malwares take from 25 malware families which were previously disarme before being converted to 32 by 32 images based on the neares neighbor interpolation, and Avast CTU Dataset which contain archived CAPEv2 reports of 48,976 malicious files with eacl files containing sha256, classification to malware family, typ of the malware, and date of detection under the 6 malwar family type namely (”banker”, ”trojan”, ”pws”, ”coinminer” ”rat”, ”keylogger”) and 10 malware families in the dataset.

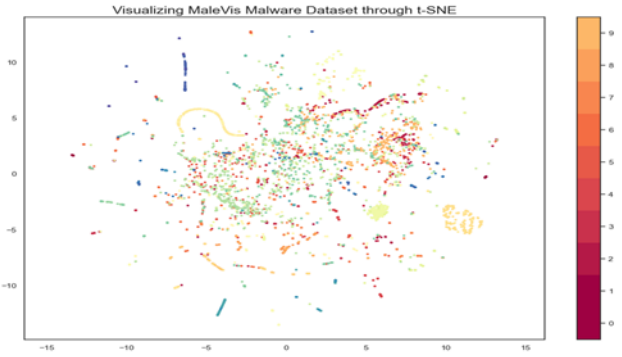

Fig. 1. Visualization of In-Distributional embedding spaces between samples from same malware family causing poor performance of current state-of-the-art algorithm on novel out-of-distribution malware classification

B. Experimental Set-up and Result

The proposed framework had 3 major stages =¿ exploitation of the in-dimensional embedding spaces between variants from the same malware family and their corresponding vector representation as well as the inter-dimensional space from different malware family takes place at stage 1. see (fig 1). At stage 2, a shallow neural network with two connected layer was built from the scratch, our network was restricted to maximum of 2 connected network layer to avoid over-fitting considering that novel out-of-distribution malware samples which doesn't exist or scanty are being targeted for classification, this significantly prevent the model from memorizing the data, the final stage involves the building and interconnection of Bayesian inference based computation algorithm that intertwine with the connected network network and is able to adjust the represented embedding space data point in response to an exposure to any previously unseen or novel out-of-distribution sample. It is this adjustment in data points that determines how and to what extent the weight of the network will be adjusted. Changes in the adjustment of the data points triggers update on the weight of the model which in turn trigger the overall gradient thereby enabling the proposed model to adjust to any unseen sample. Our proposed model shows gave preliminary of 81% accuracy in the successful classification of previously unseen out-of-distribution malware variants and family, a feat that could not be achieved by any of the state-of-the-art algorithm on OOD malware classification. see (fig 2)

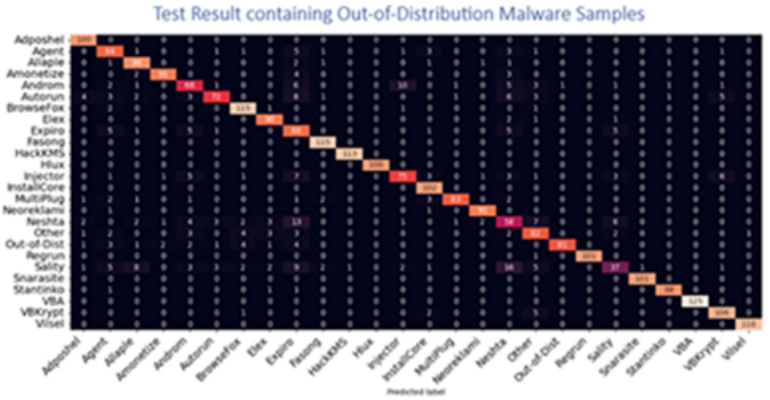

Fig. 2. Pre-liminary result of the proposed model showing 81 successful classification of unseen novel out-of-distribution malware samples

II. Conclusion

In this research, we proposed a deep learning framework with a shallow neural network that uses computations from Bayesian inference to trigger both the model weight and gradient of the network, So far, we had been able to successfully exploit some dimensional spaces between variants of same malware family with promising preliminary result **??** which current state-of-art OOD models couldn't achieve on an OOD malware classification. We are currently working on the parameter optimization setting to ensure the accuracy increases from 81% to state-of-the-art accuracy, Our focus is on the Bayesian inference computation algorithm which determine when and how the model weight should be regulated which in turns trigger update on the gradient. By integrating these technical elements into a cohesive framework, our novel framework provides a robust model that is intelligent enough to successfully classify both current out-of-distribution and future novel out-of-distribution malware correctly, something that is lacking in the current state-of-the-art algorithm.

References

[1] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R Bharat Rao. Learning classifiers when the training data is not iid. In *IJCAI*, volume 2007, pages 756–61. Citeseer, 2007.

[2] Navid Ghassemi and Ehsan Fazl-Ersi. A comprehensive review of trends, applications and challenges in out-of-distribution detection. *arXiv preprint arXiv:2209.12935*, 2022.

[3] Tosin Ige, Christopher Kiekintveld, and Aritran Piplai. Deep learning-based speech and vision synthesis to improve phishing attack detection through a multi-layer adaptive framework. *arXiv preprint arXiv:2402.17249*, 2024.

[4] Tosin Ige, Christopher Kiekintveld, and Aritran Piplai. An investigation into the performances of the state-of-the-art machine learning approaches for various cyber-attack detection: A survey. *arXiv preprint arXiv:2402.17045*, 2024.