2022

Robustness and Vulnerability Measures of Deep Learning Methods for Cyber Defense

Martinsen, Thor; Kang, Wei

Monterey, California: Naval Postgraduate School

ROBUSTNESS AND VULNERABILITY MEASUREMENT OF DEEP LEARNING METHODS FOR CYBER DEFENSE

EXECUTIVE SUMMARY

Principal Investigator (PI): CAPT Thor Martinsen, USN, PhD, Applied Mathematics

Additional Researcher(s): Dr. Wei Kang, Applied Mathematics

Student Participation: ENS Elana Kozak, USN, Applied Mathematics and ENS Philip Smith, USN, Applied Mathematics

Prepared for:
Topic Sponsor Lead Organization: N2/N6 - Information Warfare
Topic Sponsor Organization(s): Navy Cyber Defense Operations Command
Topic Sponsor Name(s): Executive Officer, CDR Matt Caylor, USN
Topic Sponsor Contact Information: mcaylor@ncdoc.navy.mil, 757-203-1054

This research is supported by funding from the
Naval Postgraduate School, Naval Research Program (PE 0605853N/2098).

Project Summary

Navy networks and infrastructure are under frequent cyberattack. One developing area of application for machine learning (ML) is cybersecurity. However, the susceptibility of ML to adversarial data is an important issue that must be studied and addressed before these systems can safely be incorporated into US Navy systems and operations. The robustness of deep learning (DL) techniques used in computer vision and language processing have been extensively studied. Less is, however, currently known about the vulnerabilities and robustness of DL methods suitable in cybersecurity applications. The goal of this study is to investigate mathematical concepts of robustness and vulnerability of ML systems that are subjected to data poisoning attacks. The first phase of the project includes a thorough literature review. The second phase of research focuses on robustness analysis of infrastructure cybersecurity. Using a microgrid power system model and learning-based fault detection as our testbed, we investigate the robustness of neural networks subjected to noisy or poisoned data. In the third and final phase of the project, we explore distributional robustness of neural networks. The findings and conclusions from our study include the following: The concept of robustness is not uniquely defined in the existing body of adversarial machine learning research literature. When incorporating ML technology into US Navy systems, it is important that deep neural networks (DNN) be purpose-built and that they are trained with a slightly higher level of noise than what is expected in their normal operating environments. Whenever possible, we also recommend incorporating ML-redundancy and simultaneously operating several DNNs with different network topologies. Research indicates that although they may perform similarly under normal conditions, their performance can differ in the presence of noise. Such disparities can be used to detect the presence of noise and inform operators that a data poisoning attack may be taking place.

Keywords: *machine learning, deep learning, adversarial machine learning, data poisoning, robustness, cybersecurity, infrastructure security, fault detection*

Background

Sophisticated cyber actors and nation-states are developing capabilities to disrupt, destroy, or threaten the delivery of essential services. According to the Cybersecurity & Infrastructure Security Agency (n.d.), "As information technology becomes increasingly integrated with physical infrastructure operations, there is increased risk for wide scale or high-consequence events that could cause harm or disrupt services upon which our economy and the daily lives of millions of Americans depend." Many of the United States Navy's operations depend upon physical infrastructure such as the power grid and computer networks. These days, cybersecurity plays an essential role in protecting a wide spectrum of critical systems and infrastructures. This research will provide the Navy Cyber Defense Operations Command with information regarding potential vulnerabilities of DL systems as well as propose some computational methods to aid the command with determining the robustness of DL embedded systems should they be employed in future Navy networks.

The use of AI and ML in cybersecurity settings is an area that is attracting increased attention. For instance, supervised learning is studied by researchers to classify particular security problems such as

denial-of-service attacks or to identify different classes of network attacks such as scanning and spoofing. There have been many approaches to network intrusion detection using DL such as deep belief networks and restricted Boltzmann machines. Deep belief networks can also be applied to detect malware attacks. Studies show that the combination of feature selection and deep neural networks is capable of performing malware classification tasks. The stacked auto-encoder, a special kind of DNN, can be used for network traffic identification and protocol classification. Despite promising results from existing DL research, neural nets have been shown to be susceptible to adversarial data poisoning (i.e., small perturbations of input data designed to cause output variations that lead to errors such as mislabeling or misclassification). The robustness of DL techniques used in computer vision and language processing have been extensively studied. However, less is currently known about the vulnerabilities and robustness of DL methods suitable in cybersecurity applications. The problem of quantitatively measuring the robustness and vulnerability of DL for cybersecurity applications is the focal area of this project.

The first phase of the project includes a thorough literature review. The second phase of research focuses on robustness analysis of infrastructure cybersecurity. Using a microgrid power system model and learning-based fault detection as our testbed, we investigate the robustness of neural networks subjected to noisy or poisoned data. In the third and final phase of the project, we explore distributional robustness of neural networks.

Findings and Conclusions

Our research shows that the concept of robustness is not uniquely defined in the existing body of ML research literature. There are a variety of different aspects of robustness that may affect the performance of an ML model. The study and evaluation of robustness should, therefore, be tailored with the specific application in mind. Most machine learning robustness studies have thus far focused on traditional machine learning applications such as computer vision, image classification, and language processing. However, some techniques found in the existing literature are applicable to a wider spectrum of application scenarios.

This study focused on robustness analysis of infrastructure cyber security. Using a microgrid power system model and learning-based fault detection as the testbed, we investigated the robustness of DNNs under noisy or adversarial data. From the study, we conclude that noise drawn from a variety of different distributions with the same mean and standard deviation produce the same or very similar results when it comes to the DNN's ability to detect faults. However, adding adversarial noise in the direction of the network gradient produced substantially different effects from that of random noise distributions. In addition, computation shows that more complex networks do not necessarily result in more robust networks. During the project, we added uniform random noise to the training data. Subsequent testing showed that the lowest error rates were achieved when a network was trained with slightly higher noise levels than those present in the testing data. This finding is important in real-world applications where noise is expected in the input data. In such cases, the ML model should be trained with additional noise added to the training data set.

Distributional robustness was also investigated. DNNs may sometimes be used outside of the environment in which they were trained. If the input data's distribution is significantly different from that of the training data, it could negatively impact the performance of the network. The dynamic behavior of the input data is critical to distributional robustness of the system. For dynamical systems, the initial distributions for trajectories of the system can be vastly different. However, the dynamic nature causes the data points along trajectories to converge to a similar pattern, regardless of the initial state distributions.

The findings in this study are intended to inform the technical staff and leadership at the Navy Cyber Defense Operations Command. When incorporating ML technology into Navy Networks and Defenses, we recommend ensuring the DNNs in question are purpose-built for the intended application and are trained with a slightly higher level of noise than what is normally found in the testing data. Whenever possible, we also recommend that ML-redundancy be built into the system. Although DNNs with different network topologies may perform similarly under normal conditions, our research indicates that their performance can differ in the presence of noise. This disparity can be leveraged to detect the presence of noise in the system and inform operators of aberrant operational conditions or the potential of a data poisoning attack is taking place.

Recommendations for Further Research

One of the conclusions of this research project is that the evaluation of machine learning (ML) robustness should be tailored with the specific application in mind. Future study of ML robustness in support of the Navy Cyber Defense Operations Command should focus on applying the testing methods developed in this project to real-world Navy network data. This includes selecting specific machine learning tools to evaluate, collect, and format real-world network data, and conducting off-line computations and simulations to find the effects of added random noise and adversarial perturbations. If possible, we recommend incorporating and testing several neural networks with different network topologies that simultaneously process incoming data. Results from our study indicate that performance and predictions disparities among different neural networks could potentially be used to indicate the presence of noise or that adversarial attacks are occurring within a network. Finally, our research also indicates that the dynamic behavior of data can be critical to the overall robustness of a system. Identifying and studying the dynamic nature of data resident within operational US Navy networks should consequently be considered when carrying out future ML robustness research.

References

Cybersecurity & Infrastructure Security Agency. (n.d.). *CISA's Role in Cybersecurity*. Retrieved December 6, 2022, from *https://www.cisa.gov/cybersecurity*

Acronyms

AI artificial intelligence
DL deep learning

DNN	deep neural network
ML	machine learning