PhD school: Addressing Data Challenges in Edge AI Systems

Katarina Petrovic
PhD student of Telecommunication Engineering at Graz University of Technology
Graz, Austria
katarina.petrovic@tugraz.at

Abstract

As the volume of data generated from connected IoT devices continues to rise, concerns such as data poisoning have become increasingly significant. We examine the impact of data quality on machine learning models, with a particular focus on sensor and time-series data. We investigate the thresholds at which poisoned data can mislead these models, models transferability and explore techniques to enhance data efficiency. Participation in a PhD school will provide valuable feedback regarding some current challenges and enhance our research.

CCS Concepts

• **Computer systems organization** → Embedded software; • **Security and privacy** → **Systems security**; • **Computing methodologies** → *Machine learning algorithms.*

Keywords

IoT, data quality, data poisoning, model robustness, adversarial attacks, backdoor attacks

1 Introduction

Rapid advancements in hardware, software, and sensing technologies over the past decades have enabled the exponential growth of data available for machine learning. Parallel developments in communication technologies have further accelerated this expansion, particularly with the rise of the Internet of Things (IoT) and Internet-connected sensory devices that continuously gather observations and measurements from the physical-world. As of 2022, approximately 15 billion IoT devices are in use globally, with estimates predicting this number to rise to around 40 billion by 2033 [13]. While the IoT has significantly increased the volume of generated and shared data, the growing reliance on large datasets for high model performance presents both opportunities and challenges for machine learning. This surge in IoT-generated data raises critical questions: Does more data necessarily lead to better outcomes? Is it the quantity or the quality of data that ultimately matters for model performance? May the increasing volume of data compromise the integrity of the learning process? Researchers have begun to explore the potential drawbacks of large-scale datasets [2, 5, 6, 10], particularly in terms of time and cost efficiency, as well as issues related to data quality, reliability, and safety. This research will specifically examines these challenges in the domain of sensor and time series data generated by IoT systems, focusing on how data quality affects model safety and how different models vary in their robustness to data-related problems.

Our research focuses on the critical role of clean, high-quality data from IoT systems in training machine learning models. We will investigate the risks associated with using uncontrolled datasets, particularly those containing poisoned or noisy data. Specifically, we aim to determine the threshold at which poisoned data can mislead a model and explore whether these issues can be detected in advance. These challenges not only degrade model performance but also present significant risks in safety-critical applications, such as water distribution systems. Additionally, we will examine the potential of data compression techniques to reduce dataset size without compromising data quality. By compressing data effectively, we aim to maintain model accuracy while reducing computational costs, addressing both efficiency and data integrity.

Our research seeks to provide a systematic approach to mitigating the dangers posed by "bad" data in machine learning. With a focus on sensor and time series data produced by IoT devices, we address an important domain where these challenges are particularly prevalent. Ultimately, our goal is to make machine learning models not only more robust and reliable but also more efficient in handling large-scale datasets.

2 Methodology

Large datasets, whether sourced from the Internet, sensor readings, or user-generated content, often contain corrupted or manipulated data, which can compromise the integrity and performance of machine learning models. In this research, we focus on addressing the growing concern of data poisoning in modern machine learning. Our methodology is designed to systematically investigate these threats and develop strategies to mitigate the risks posed by poisoned data in machine learning models.

2.1 Adversarial attacks

Adversarial attacks present a significant threat to object detection algorithms, drawing substantial attention due to their big disruptive potential. Research in this area has focused on both developing efficient attack methods and designing defense strategies. In the context of object detection, adversarial attacks are commonly executed using adversarial patches. Initially introduced in 2018 [1], and further extended in 2019 [9], adversarial patches have since been adapted for physical-world scenarios [8, 12], highlighting the challenges of maintaining attack effectiveness across varying environmental conditions. Building on this body of work, our previous research [11] systematically evaluated the performance of physical adversarial patches in object detection tasks.

Looking ahead, we plan to extend our investigation into adversarial attacks on time series data [3, 7]. In particular, we aim to explore attack strategies where small perturbations in the training data can mislead deep learning models, as demonstrated in [14]. Our approach will leverage gradient information to generate these small perturbations, creating hard to detect adversarial attacks that could compromise the safety of machine learning models

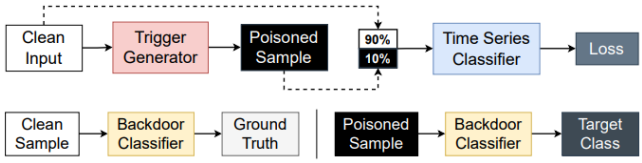

Figure 1: Overview of the proposed attack pipeline [4]. Top: training the attack trigger pattern generator; Bottom: inference with the backdoored model on clean vs poisoned samples.

in real-world applications. We will examine these attacks in systems where sensor readings are physically connected and cannot be easily disrupted. Since real-world systems often rely on distributed sensors, we also plan to investigate the transferability of models trained on individual sensor data and assess how effectively they can generalize across multiple sensors.

2.2 Backdoor attacks

Backdoor attacks [4] have emerged as a significant security threat to deep learning models, as they allow adversaries to manipulate the model's test-time predictions by injecting backdoor triggers during training. These attacks are particularly difficult to detect, making them a dangerous risk. Our research will focus on investigating backdoor attacks in the context of sensor and time series data, which pose greater challenges compared to image-based attacks due to the complex, temporal nature of the data.

Timeseries backdoor attacks involve subtly altering time series data during training so that the model performs normally on clean data but misclassifies inputs when specific patterns, referred to as triggers, are added. These attacks operate undetected by traditional defense mechanisms, making them especially dangerous. There are two primary types of time series backdoor attacks: (1) the attacker controls the entire training process and creates a backdoored model that appears normal but misclassifies inputs containing the trigger; (2) the victim unknowingly uses poisoned data that contains hidden triggers, leading the model to fail on specific adversary-selected inputs.

A key component of backdoor attacks (Figure 1), is the "trigger generator" network, which creates dynamic, sample-specific triggers. These triggers make poisoned data more difficult to detect and remove, as they are not static and are harder to identify. Our research focuses on thoroughly analyzing these attacks, with the aim of developing strategies to determine the threshold at which poisoned data can successfully mislead a model while remaining undetected. Additionally, we will investigate the transferability of these attacks in systems with distributed sensors, such as water distribution systems, assessing how the attacks propagate across different sensor networks.

3 Challenges

Finding the Dataset. One of the challenges in our research is sourcing a high-quality, representative dataset without the ability to independently collect data. Safety-critical real-world systems, such as water distribution systems, rarely release their data collections, and it is almost impossible to gather such data on our own due to strict regulations and access limitations. As a result, we often have to rely on artificially generated datasets that comply with all the physical laws of the system, but may not fully capture the complexities of real-world scenarios. This limitation can affect the validity and applicability of our findings.

Flexibility. Developing flexible methods in machine learning is a significant challenge. Many existing approaches are highly specialized and tailored to specific domains, making it difficult to adapt them to other areas. This lack of generalizability limits the applicability of these methods across different use cases, limiting the creation of versatile models that can effectively address a range of problems. Building adaptable solutions remains a complex task, requiring innovative approaches to bridge the gap between domain-specific and generalizable methodologies.

Networking and Collaboration. Attending conferences, workshops, and doctoral schools offers valuable opportunities to build connections within the academic and research community. Identifying potential collaborators with complementary skills and expertise is crucial, yet finding the right network to support such collaborations requires considerable effort and strategic engagement within the community.

Balancing Responsibilities. Maintaining a balance between research and project work can be challenging, as their core focuses often diverge. Project work sometimes tends to be more engineering-oriented, which can lead to deviations from the primary research topic. Suggestions on how to manage situations where project tasks drift away from the research focus and strategies for effectively navigating these circumstances would be greatly appreciated.

Inspiration. At times, researchers may experience a lack of inspiration in their work. Engaging with fellow PhD students can foster fresh perspectives that inspire creativity. Exchanging ideas within this community can be highly beneficial, providing new insights and rejuvenating one's approach to research. Such interactions create opportunities for collaboration and enhance the overall research experience.

The PhD school offers an excellent environment to tackle the identified challenges. Feedback from experts and fellow participants can be invaluable for guiding and enhancing research efforts. Furthermore, engaging with the experiences and insights shared by others can generate fresh ideas and perspectives.

References

[1] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. 2017. Adversarial Patch. *CoRR* abs/1712.09665 (2017). arXiv:1712.09665 http://arxiv.org/abs/1712.09665

[2] Lisa Ehrlinger, Verena Haunschmid, Davide Palazzini, and Christian Lettner. 2019. A DaQL to Monitor Data Quality in Machine Learning Applications. In *Database and Expert Systems Applications*, Sven Hartmann, Josef Küng, Sharma Chakravarthy, Gabriele Anderst-Kotsis, A Min Tjoa, and Ismail Khalil (Eds.). Springer International Publishing, Cham, 227–237.

[3] Samuel Harford, Fazle Karim, and Houshang Darabi. 2020. Adversarial Attacks on Multivariate Time Series. arXiv:2004.00410 [cs.LG] https://arxiv.org/abs/2004.00410

[4] Yujing Jiang, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. 2023. Backdoor Attacks on Time Series: A Generative Approach. arXiv:2211.07915 [cs.LG] https://arxiv.org/abs/2211.07915

[5] Siddharth Joshi and Baharan Mirzasoleiman. 2024. Foundations of Data-efficient Machine Learning. International Conference on Machine Learning.

[6] Ashish Juneja and Nripendra Narayan Das. 2019. Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. 559–563. https://doi.org/10.1109/COMITCon.2019.8862267

[7] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. 2021. Adversarial Attacks on Time Series. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2021), 3309–3320. https://doi.org/10.1109/TPAMI.2020.2986319

[8] M. Lee and J. Z. Kolter. 2019. On Physical Adversarial Patches for Object Detection. *CoRR* abs/1906.11897 (2019). arXiv:1906.11897 http://arxiv.org/abs/1906.11897

[9] X. Liu, H. Yang, L. Song, H. Li, and Y. Chen. 2018. DPatch: Attacking Object Detectors with Adversarial Patches. (2018). http://arxiv.org/abs/1806.02299

[10] Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. 2019. Data Validation for Machine Learning. In *Proceedings of Machine Learning and Systems*, A. Talwalkar, V. Smith, and M. Zaharia (Eds.), Vol. 1. 334–347. https://proceedings.mlsys.org/paper_files/paper/2019/file/928f1160e52192e3e0017fb63ab65391-Paper.pdf

[11] Jakob Schack, Katarina Petrovic, and Olga Saukh. 2024. Breaking the Illusion: Real-world Challenges for Adversarial Patches in Object Detection. In *The 21st International Conference on Embedded Wireless Systems and Networks (EWSN'24)*.

[12] S. Shrestha, S. Pathak, and E. Kugler Viegas. 2023. Towards a Robust Adversarial Patch Attack Against Unmanned Aerial Vehicles Object Detection.

[13] Statista. 2024. Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2033, by vertical. https://www.statista.com/statistics/1194682/iot-connected-devices-vertically/

[14] Tao Wu, Xuechun Wang, Shaojie Qiao, Xingping Xian, Yanbing Liu, and Liang Zhang. 2022. Small perturbations are enough: Adversarial attacks on time series prediction. *Information Sciences* 587 (2022), 794–812. https://doi.org/10.1016/j.ins.2021.11.007