DATA SELECTION FOR EFFICIENT BACKDOOR POISONING ATTACKS

Author's: R. Deepavathi[1], S. Rajeshkumar[2], K. Sujith[3], N. Vanjulavalli[4]

Abstract

Backdoor attacks pose a significant security threat to machine learning systems by embedding hidden triggers into training data such that the trained model behaves normally on clean inputs while misclassifying triggered inputs into an attacker-specified target class. A critical but underexplored factor in the effectiveness of such attacks is data selection strategy specifically, which training samples are chosen for poisoning. This study systematically investigates how different data selection strategies influence poison efficiency, defined as the ability to achieve high attack success rates (ASR) with minimal poisoning while maintaining acceptable clean accuracy (CA).Using a controlled synthetic multi-class classification environment, we evaluate four poisoning strategies: random selection, near-decision-boundary selection, prototype-based selection, and outlier-based selection. A trigger-based backdoor is injected into selected samples by modifying feature representations and relabeling them to a predefined target class. We measure performance using Clean Accuracy (CA), Attack Success Rate (ASR), and poison fraction sensitivity.Experimental results demonstrate that data selection significantly impacts poison efficiency. Near-boundary and prototype-based selection strategies typically require fewer poisoned samples to achieve high ASR compared to random or outlier selection. These findings highlight that not all poisoned samples contribute equally to backdoor formation; rather, strategically selected samples can amplify model vulnerability.This research provides insights into the mechanics of backdoor formation and contributes to the development of more robust training pipelines and defense mechanisms against data poisoning attacks.