Thesis Defense

Computer Science Master's Program

"Investigating and Evaluating Poison Unlearning under Imperfect Detection"

By **Mitashi Parikh**

Abstract:

Machine unlearning is an emerging field focused on removing the influence of specific data points from trained machine learning models. Early work in this field has primarily focused on privacy-oriented unlearning, often motivated by data protection regulations such as the GDPR. More recently, security-oriented unlearning has emerged as a response to data poisoning attacks, where malicious inputs are injected into training data to manipulate model behavior. In these security-oriented settings, unlearning methods aim to reverse the influence of malicious training data after a model has already been deployed. However, existing methods assume access to an accurately identified portion of poisoned data. This thesis investigates the performance of state-of-the-art poison unlearning techniques under imperfect detection conditions, where the forget set is not only a portion of poisoned data but also includes false positives. An evaluation simulating these more realistic detection scenarios is designed. State-of-the-art unlearning method, Potion, is evaluated using this framework across standard image classification benchmarks. The results demonstrate that false positives in the forget set can impact model damage and unlearning effectiveness. More broadly, these findings highlight the need for robust and comprehensive evaluation frameworks that reflect real-world conditions.

Date: Thursday, June 5th, 2025
Time: 5:30 PM – 7:30 PM
Location: 14-232b
Zoom: https://calpoly.zoom.us/j/7024764308?omn=82201196936
Committee: Dr. Fang, Dr. Ventura, Dr. Sisodia, Dr. Hasan