Thesis Defense

Computer Science Master's Program

"Signature Based Model Attribution of Images Generated by Stable Diffusion"

By **Bella White**

Abstract:

Generative AI is a powerful yet controversial technology capable of spreading misinformation, non-consensual imagery, and offensive materials. This technology is being abused at an alarming rate to create realistic child sexual assault materials (CSAM) from just 3-5 images of an individual for the purposes of financial extortion, grooming, and re-victimization. The ability to not only distinguish between real images and generated images but also attribute an image to a generative model of origin allows the world of digital forensics to adapt to the ever-evolving threat landscape. Extensive research into AI image detection has been performed alongside initial forays into generator-level model attribution and hyperparameter detection. This work applies signature-based model attribution, an existing digital forensic technique, to Stable Diffusion images to attribute generated images to an individual finetuned SD model, model family, or LoRA. Model or model group signatures are created from and evaluated on a benign dataset of images from various finetuned Stable Diffusion models and associated LoRAs. Unseen image residuals are compared to model fingerprints through the use of comparison windows, $m\times m$ sections of an image or fingerprint, and a distance metric is calculated to attribute a given image to the model fingerprint it is most similar to. This thesis provides insight into the most effective size and location of a comparison window and compares performance across different signature creation methods in binary and multiclass classification. The results of this thesis indicate that signature-based attribution can be applied at a higher level of granularity than previously explored through the use of comparison windows.

Date: Monday, December 8th, 2025
Time: 4:00 PM – 6:00 PM
Location: 14-238b
Committee: Dr. Dekhtyar, Dr. Ventura, and Dr. Wright