UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

Qualifying Exam

Thursday February 29, 2024 at 2PM, online via Zoom (Remote)

Mohammad Maghsoudimehrabani

Towards Attack-Resilient Natural Language Processing Systems: A Framework for Enhancing Security and Intellectual Property Protection

Chair: Dr. Stacey Scott
Advisor: Dr. Ali Dehghantanha
Co-Advisor: Dr. Gautam Srivastava (external – Brandon University)
Non-Advisory: Dr. Neil Bruce
Non-Advisory: Dr. Fattane Zarrinkalam (SoE)

Abstract:

The escalating adoption of NLP in various digital applications, such as chatbots and automated content generation, underscores an urgent need for robust security measures. As these NLP systems become integral components of digital infrastructures, they are increasingly vulnerable to adversarial attacks and intellectual property violations. This research presents a comprehensive framework comprising three sophisticated models designed to bolster the security and integrity of NLP systems. The first model within this framework is the "Stateful Query Analysis (SQA)," model, a standalone query filtering mechanism. Designed to operate in conjunction with existing NLP models, the SQA model proactively identifies and mitigates query-based black-box adversarial attacks. By scrutinizing sequences of queries for suspicious patterns, the SQA model enhances the overall security resilience of NLP systems. The second model, the "Anti-Distillation Backdoor Attack Testing (ADBAT)," model, serves as both a vulnerability assessment and an audit tool for encoder-only transformers in text processing. This model introduces a novel method to craft anti-distillation backdoors hidden in the teacher model that survive in Knowledge Distillation (KD), highlighting the unreliability of models trained by KD. The ADBAT model underscores the need for rigorous security checks in NLP models, especially in the context of distilled models, which have not been extensively explored for such vulnerabilities in transformers. The third model, "Strategic Watermark Injection for Encoder-Only Transformer Security (EOTShield)," is a strategic watermarking scheme for pre-trained encoder-only transformers, termed the EOTShield model. Given a clean pre-trained encoder, EOTShield injects a watermark into it and outputs a watermarked version.

This model is anchored in the principles established by the ADBAT model and aims to safeguard intellectual property rights while preserving model utility during potential model stealing processes such as distillation. Collectively, these models form a cohesive framework that addresses the dual challenges of adversarial robustness and intellectual property protection in NLP systems. This framework is characterized by its strategic layering of security measures, where each model serves a distinct purpose but complements the others, ensuring that NLP systems remain resilient to both adversarial threats and intellectual property infringements. In summary, this research provides a holistic approach to fortifying NLP systems, ensuring they remain secure and trustworthy in an increasingly interconnected digital landscape.