PPR Seminar

Advances in Perception, Prediction, and Reasoning

Changhoon Kim

Postdoctoral Scientist
Amazon

https://www.changhoonkim.com/

amazon

ASU Arizona State University

Strengthening Image Generative AI: Integrating Fingerprinting and Revision Methods for Enhanced Safety and Control

Nov 25, 2024; 4:00 – 5:15 PM; Math&Psych 106 / Webex

Abstract: In the rapidly evolving field of Generative Artificial Intelligence (Gen-AI) for imaging, models such as DALL·E3 and Stable Diffusion have transitioned from theoretical concepts to practical tools with significant impact across various sectors including entertainment, art, journalism, and education. These advancements represent a substantial technological evolution, enhancing creative and professional practices. However, the widespread accessibility of Gen-AI also facilitates misuse by malicious actors who create deepfakes and spread misinformation, posing serious risks to societal well-being and privacy. This talk will address these critical challenges by focusing on enhancing the reliability of Image Gen-AI models through the identification and mitigation of inherent vulnerabilities and the development of computational tools and frameworks for enabling better community oversight. The talk will describe the development of innovative fingerprinting techniques that trace malicious Gen-AI outputs back to their sources, and the implementation of strategies to prevent the generation of unauthorized content. These efforts collectively strengthen the robustness and accountability of Gen-AI technologies, particularly in sensitive applications..

About the Speaker: Dr. Changhoon Kim is a Postdoctoral Scientist in the Bedrock Team at Amazon. He completed his Ph.D. in Computer Engineering at Arizona State University. His primary research focuses on the creation of secure machine learning systems. He has dedicated his efforts to developing user-attribution methods for generative models, a critical area of research in the age of AI-generated hyper-realistic content for tracing malicious usage, and machine unlearning for removing private or harmful content from AI models. Kim's research has been recognized at prestigious conferences such as ICLR, ICML, ECCV, and CVPR, and a U.S. patent for user-attribution in generative models. To further contribute to the community, he has organized tutorials and workshops at leading conferences to emphasize the importance of secure generative AI.

UMBC

PPR Seminar: https://www.tejasgokhale.com/seminar.html
Hosted by Dr. Tejas Gokhale gokhale@umbc.edu