ECE 590: Generative AI: Foundations, Applications, and Safety

Neil Gong

Instructor

- Neil Gong
- neil.gong@duke.edu
- Research area
  - AI and security
- Office hour
  - Time: Wed. 11:00am – 11:50am
  - Location: 413 Wilkinson Building
- Teaching assistant
  - Reachal Wang, reachal.wang@duke.edu
    Jason Wang, xilong.wang@duke.edu

Course overview

- Foundations of GenAI
  - Transformers, pre-training, post-training, LLM agent, diffusion model, etc.
- Applications and Security/Safety of GenAI
  - Prompt injection, detecting AI-generated content, safety and jailbreaking, data-use auditing, etc.

- Course webpage:

https://ece590-genai.github.io/

Goal of this class

- State-of-the-art literature on GenAI foundations and security/safety
- Get prepared to apply and research GenAI

Class format

- Read papers
  - Write comments and send to ecegenerativeai@gmail.com
  - Deadline: Sunday and Tuesday 11:59pm
  - Send your comments to all papers in a single email thread
  - Comment
    - One paragraph of summary of each assigned paper
    - Three or more strengths
    - Three or more weaknesses
    - Disclosure of AI use
- Lead a lecture
  - Forming a group of at most 4 students
  - A group sends three preferred dates to ecegenerativeai@gmail.com by 11:59pm, 01/25
- Participate in class
- One class project
  - Can be a group of at most 4 students
  - Your research project can be class project
  - 02/01: project proposal due.
  - 03/15: milestone report due.
  - 04/13, 04/15: project presentation.
  - 04/26: final project report due.

Lead a lecture

- Why lead a lecture
  - Understanding a topic better after teaching others about it
- Like how I give a lecture
- May read multiple papers on the selected topic
  - E.g., each group member leads discussion on one paper
- 75 mins for a lecture!
- Use whiteboard/blackboard if possible
- Be interactive

An example class project

- Problem: Detecting AI-generated images

- Solution: watermark
    - E.g., start from a watermarking method and optimize it to enhance its robustness, efficiency, and/or image quality

- Proposal abstract: one paragraph to describe the problem and potential solution.

Project report template

- Abstract
- Introduction
- Related work (can also be moved to be after empirical evaluation)
- Problem definition
- Method
- Theoretical evaluation (if any)
- Empirical evaluation
- Conclusion

Grading policy

- 50% project

- 25% reading assignment

- 10% class participation

- 15% class presentation

Course structure

- Part I: LLM
  - Foundations: transformers, pre-training, post-training, prompt engineering, LLM agent
  - Security/Safety: prompt injection, jailbreak and safety guardrails, detecting AI-generated text and robustness, and data-use auditing

- Part II: Image generation
  - Foundations: representation learning, diffusion model
  - Security/Safety: guardrails, jailbreak attacks, detecting AI-generated images, and robustness of detectors

- Part III: other modalities
  - Audio and video generation and security/safety issues