Course: 10-417/617 Intermediate Deep Learning

Instructor: Professor Ruslan Salakhutdinov

" The goal of this course is to introduce students to both the foundational ideas and the recent advances in deep learning." (Course Website: https://andrejristeski.github.io/10417-20/)

Overall Course Objectives:

- Understand and utilize concepts in supervised learning. (ex. neural networks, CNN, RNN).
- Understand and utilize concepts in unsupervised learning. (ex. VAE, Sparse-Coding, Boltzmann machines, and GANs)
- Build and gain familiarity of applications of Deep Learning in Image Recognition, Video Analysis, and Language Modelling.

https://andrejristeski.github.io/10417-20/

Constraints

- Class size: ~60 students
- Class Workload: 12 - 15 hours/week
- Course faculty not confirmed
- Constraints have not changed since previous pitch.

Learning Objectives

At the end of the module, students should be able to:

- Recognize ethical considerations and bias introduced in sequence-to-sequence models, particularly language models.
- Analyze ethical and psychological aspects of deep reinforcement learning with a focus on machine ethics.
- Identify ethical consequences of students' own deep learning application as an add-on component to the course project.

Topics Deep Dive

Language Models Ethics Reading, Journal, & Lecture

Article:

“We read the paper that forced Timnit Gebru out of Google. Here’s what it says.” (Hao)

Summary:

- 4 Dangers of Large Language Models
- Google AI ethics researcher departs from Google after her paper is disproved by Google for publication.

Motivation: Recognize and discuss ethical consequences of language models:

1. Academic Freedom and Agency
2. Privacy Concerns
3. Environmental Impacts

Hao, Karen. “We Read the Paper That Forced Timnit Gebru out of Google. Here's What It Says.” *MIT* Technology Review, MIT Technology Review, 4 Dec. 2020, www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

Deep Reinforcement Learning Reading/Journal and In-Class Activity

Article: We read, “Machines That Don’t Kill: How Reinforcement Learning Can Solve Moral Uncertainties”

Link to the article: https://analyticsindiamag.com/reinforcemenet-learning-moral-dilemma-ethics/

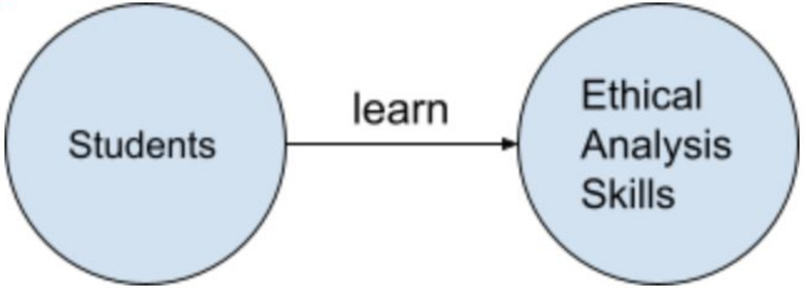
In-Class Student Lead Discussion:

- Students discuss about the design of reinforcement learning agents to handle certain moral dilemma such as the famous trolley problem in ethics and psychology.

- This is followed by a class discussion involving ideas about how moral philosophy and machine ethics can be be bridged.

Project Reflection

- Students reflect on the ethical considerations of their project and build a concept diagram as well as answer questions regarding the moral concerns of their work when deployed at scale.
- Students also have an opportunity to programmatically identify bias in their application as an optional extra credit question.



Implementation Overview

- **Part I: Language Models**
  - Pre-work: Read article and complete journal (40 minutes)
  - In-class: Slides developed for a 30 minute lecture + 15 minute instructor-led discussion
- **Part II: Deep Reinforcement Learning**
  - Pre-work: Read article and complete journal (40 minutes)
  - In-class: 20 minute student-led small group discussion discussion + 20 minute class-wide discussion
- **Part III: Project Reflection**
  - Project reflection (50 minutes total + 90 minute optional extra credit)
    - Reflection Question: Reflect on the ethical considerations of the application's use case when deployed at scale and check if students integrated ethics in the system design phase.
    - Concept Diagram: Construct a concept diagram that illustrates the relationship between the deep learning project and other entities that are in some way impacted by the application.
    - Extra Credit: Programmatically develop test cases to identify bias in application.

Instructor/Peer Feedback

- The module and supporting slides were presented to Professor Ruslan Salakhutdinov, and we received positive feedback.
- Professor particularly felt the language model and project material could be a great fit for the course.
- We added another article (total of 2) in response to peer feedback.
- We also addressed the issue Professors Victoria and Illah brought up about how in-class discussion will be formatted given the large class size.

Thank You