Interpretability in Artificial Intelligence

SOUL course

Course Description

As artificial intelligence models like chatGPT become increasingly capable and ubiquitous, the need to understand their inner workings intensifies. Imagine an autonomous vehicle taking an unexpected turn or a medical AI diagnosing a life-altering condition; the importance cannot be overstated. Despite their widespread applications, our grasp of these models remains alarmingly limited. This course is designed to survey this gap. A special emphasis is placed on *mechanistic interpretability*, a subfield that rigorously investigates AI networks at the neuronal level.
The course will engage students through rigorous reading assignments, interactive discussions, and a hands-on project. Papers from researchers/groups in this area, such as AnthropicAI, OpenAI, and MIT's Tegmark Group, form the cornerstone of our course.

Course Topics**

This was the original one, and maybe completely modified accordingly to student desire. Various research papers and articles to be distributed during the course.

1. Introduction to Interpretability in AI [1 week] and basics of transformer language models. [1 Week]

2. Mechanistic Interpretability [4 weeks], example papers include, [11], and a good collection of articles is collected here. One over arching goal is to reproduce the paper [10], which shows the algorithms in the phenomena of gorkking uses discrete Fourier transforms and trigonometric identities to convert addition to rotation about a circle.

3. Concept-based Interpretability [1 week] example papers include, [3].

Required Background

Students are expected to have a basic understanding of calculus, linear algebra, probability theory, and coding. A reading list will be provided to be completed before the commencement of the course.

Assessment

Student assessment will be based on both weekly reading assignments and a course-long project. Students can opt for either a coding project or a written project,

aimed at deeply exploring a sub-field of interpretability. The grade distribution will be as follows: Weekly Readings: 60%, Course-long Project: 40%.

1 Timeline

Week 1: introduction to Alignment

Introduction to the topics and scope of the course. The goal is to give a sense of the topics involved in the AI Safety community where the topics of the course fit in this picture. As examples, we discuss the problems of alignment and robustness in a nontechnical way. Some of these concerns are near-term: how do we prevent driverless cars from misidentifying a stop sign in a blizzard? Others are more long-term: if general AI systems are built, how do we make sure these systems pursue safe goals and benefit humanity?

Textbook and references: The field is extremely young, with no complete surveys about this. However, there are a number of useful survey article on this subject.

- The main article we will be looking at is the survey *Unsolved Problems in ML safety*, [5]. This is by Dan Hendricks at the center of AI safety.
- For broader perspective, there are articles by Rohin Shah [1] and Nanda [9].

Homework: this week's homework will require student to read over certain articles that examinines the socio-technical complexities of ensuring that AI and humans share compatible goals.

Week 2: The mathematical prerequisites

The language of matrices will form the basis of our discussion, we will go through a few basic matrix manipulation. The background is minimal, we do not require full Linear Algebra course background equivalent to that of 110.201, but similar content to the first 2 weeks.
Textbook and references:

- Linear Algebra with Applications, 5th Edition, [2], Otto Bretscher, Prentice Hall, December 2012, ISBN-13: 978-0321796974. This is the course text book for 110.201.
- A concise and sufficient introduction is in Stanford CS229, see [8, Sec. 1-3].
- Section 1-3 of Strictland's notes lectures [13, p1-3]

Week 3: Learning Methods and Robustness

The goal of this week is to give an overview of what is meant by *AI*, in the context of this course. This involves introducing the the first of the three most common machine learning techniques;

1. Supervised learning.
2. Reinforcement learning.
3. Unsupervised learning.

We will discuss the second technique the following week, and not really go in detail of the third example.

The examples is used to understand the *capabilities* of ML method, and highlight examples of failure modes, particularly in the context of adversarial examples, [6], see **??** for a more detailed collection of references.

One important aspect of this week is that we will phrase the machine learning paradigm in terms of *function learning*. This is particularly simplistic but is sufficient for our purpose.
References and texts:

- This lecture will be a summary of the field of ML, similar to Ngo's post.
- The requisite is minimal and we will be going through the most basic supervised learning example.

Week 4: Homework discussion and buffer

Week 5 - 6: Reinforcement Learning and Goal Misgeneralization

We will discuss one-two basic examples of each, and finish with how this leads to the problem of outer and inner alignment. The main reference for this week is by Rohin Shah et al. on *Goal Misgenralizations*, [12]. As in previous topics we will go through the main example in the paper, and discuss current methods of approaching it.
References:

- We will cover section 1 of the reinforcement learning textbook. We will use the classic reference by Sutton and Barto [14].

Weeks 7: Oversight of AI Systems

Exploration of the challenges in AI system oversight. Topics include reinforcement learning from human feedback [4]. Introductory talk [7].

Week 8: Homework discussion and buffer

References

[1] *AI Alignment 2018-19 Review — AI Alignment Forum — alignmentforum.org.* [Accessed 28-09-2023].

[2] Otto Bretscher, *Linear algebra with applications, 5th edition*, 2012.

[3] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt, *Discovering latent knowledge in language models without supervision*, 2022.

[4] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei, *Deep reinforcement learning from human preferences*, 2023.

[5] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt, *Unsolved problems in ml safety*, 2022.

[6] Robin Jia and Percy Liang, *Adversarial examples for evaluating reading comprehension systems*, 2017.

[7] Jared Kaplan, *AI Safety, RLHF, and Self-Supervision - Jared Kaplan | Stanford MLSys, youtube.com.* [Accessed 28-09-2023].

[8] Zico Kolter and Chuong Do, 2015.

[9] Neel Nanda, *My Overview of the AI Alignment Landscape: Full Sequence — docs.google.com.*

[10] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt, *Progress measures for grokking via mechanistic interpretability*, 2023.

[11] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter, *Zoom in: An introduction to circuits*, Distill (2020). https://distill.pub/2020/circuits/zoom-in.

[12] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton, *Goal misgeneralization: Why correct specifications aren't enough for correct goals*, 2022.

[13] Neil Strickland, *Linear mathematics for applications*, 2020.

[14] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: An introduction*, IEEE Transactions on Neural Networks **16** (2005), 285–286.