Session I Abstracts

Exploring Lightweight Debiasing and Enhancing Robustness Through Data Augmentation in Pretrained Language Models
Pengrui Han
Mentors: Anima Anandkumar and Rafal Kocielnik

Large-scale pretrained language models, while revolutionary, can inadvertently incorporate biases from their training data, leading to potential unintended outcomes. Traditional debiasing methods, although effective, are resource-intensive and may adversely affect a model’s linguistic capabilities. In contrast, lightweight debiasing offers a more efficient approach, but its effectiveness is often constrained and challenges in generalization arise. In response to these challenges, our research adopted a dual-pronged approach. We first investigated the Mass Memory Editing Technique, a novel lightweight method that modifies the original model by adjusting weights over specific critical layers. This was contrasted with the prevalent adapter tuning, an approach that retains and freezes the original model, introducing subsequent adapter layers for fine-tuning. Despite the Memory Editing Technique’s prowess in targeted debiasing, its generalizability was limited. This realization transitioned our focus to the second phase: fortifying the robustness of lightweight debiasing through data augmentation tailored for adapter tuning. This involved the exploration of term-based, sentence-based, synonym-based, and task-based augmentations. Preliminary findings highlight that these augmentations not only enhance debiasing efficacy but also preserve the model’s inherent linguistic strengths.

Developing End-user AI Inspection Tools for Large Language and Vision-Text Models
Roy Jiang
Mentors: Anima Anandkumar and Rafal Kocielnik

Modern language and vision models learn implicit social biases such as gendered and racial prejudices due to the media and texts they are trained on, which may propagate discriminatory effects where AI is employed. As such, methods for bias detection and quantification understandable to non-technical social experts is imperative. Existing methods for auditing models prove inadequate, relying on the use of fixed datasets and being severely unusable for non-technical experts. This project aimed to address these issues through the design, development, and iteration of end-user tools for the inspection of stereotypical bias in state-of-the-art large language (LLM) and text-vision AI models. These tools were developed through a process of UI/UX prototyping, technical feasibility evaluations, as well as stress testing with novice users and both technical and non-technical experts. Currently, we have a working version of the LLM testing tool and an initial prototype of the text-vision tool on the HuggingFace platform. We conclude that such a task is accompanied by challenges in balancing, efficiency, clarity and faithfulness, and that these tools are a necessary step in the development of fairer AI. Lastly, we recommend more principled evaluation of the tools with the target audience of social scientists and ethics experts.

Tipping Point Forecasting in Non-Stationary Dynamics on Function Spaces
Miguel Liu-Schiaffini
Mentor: Anima Anandkumar

Tipping points are abrupt, drastic, and often irreversible changes in the evolution of non-stationary and chaotic dynamical systems. For instance, increased greenhouse gas concentrations are predicted to lead to drastic decreases in low cloud cover, referred to as a climatological tipping point. In this paper, we learn the evolution of such non-stationary dynamical systems using a novel recurrent neural operator (RNO), which learns mappings between function spaces. After training RNO on only the pre-tipping dynamics, we employ it to detect future tipping points using an uncertainty-based approach. In particular, we propose a conformal prediction framework to forecast tipping points by monitoring deviations from physics constraints (such as conserved quantities and partial differential equations), enabling forecasting of these abrupt changes along with a rigorous measure of uncertainty. We illustrate our proposed methodology on non-stationary ordinary and partial differential equations, such as the Lorenz-63 and Kuramoto-Sivashinsky equations. We also apply our methods to forecast a climate system tipping point in stratocumulus cloud cover.

MineDojo 2: A Minecraft Environment With a Unified Observation and Action Space to Build More Powerful Embodied Agents
Bilal Khan
Mentors: Anima Anandkumar and Guanzhi Wang

Fan et al. 2022 introduced MineDojo: an environment built in Minecraft for training reinforcement learning agents on open-ended tasks. Recently, Wang et al. 2022 introduced Voyager, an LLM-powered agent that interacts with Minecraft through text observations and high-level actions using a text-only environment to solve complex tasks. To combine the strengths of MineDojo and Voyager and solve the most difficult open-ended tasks in Minecraft, we develop MineDojo 2: a Minecraft environment with a unified observation and action space that provides access to
both visual and text-based observation APIs and both high-level and low-level action APIs to build more powerful embodied agents.

**Joint Reconstruction-Segmentation Using Graph-Based Methods**  
Ashug Gurijala  
*Mentors: Franca Hoffmann and Jeremy Budd*

This research is focused on tackling the task of segmenting imperfect images. Prior implementations that sequentially reconstruct and segment images are limited because the reconstruction task is unaware that segmentation will follow. Another approach involves the use of neural networks, which although powerful, lack explainability. This research explores the novel approach of simultaneously reconstructing and segmenting the image. We allow for the reconstruction task to aid segmentation process, thereby overcoming the limitations of previous methods.

We use a graph-based approach, representing the image as a graph with pixels as vertices and weighted edges according to pixel similarity. The approach is integrable in a variational framework where the goal is to minimize functionals across functions defined on a graph utilizing gradient flows. We formulate the reconstruction-segmentation process as the minimization of Tikhonov reconstruction and Ginzburg-Landau segmentation energies.

To facilitate wider adoption and practical usage of the proposed approach, particularly in fields like medical imaging, we introduce a Python pipeline. Developing the Python pipeline required implementing various optimization strategies to align the principles of the proposed approach with the practical constraints of coding.

**Utilizing General Vision Language Models for Animal Action Recognition**  
Jonathan Lin  
*Mentors: Pietro Perona and Markus Marks*

Accurate animal action recognition holds significant implications across diverse fields such as bioengineering, nature conservation, and pharmaceutical experimentation. A multi-species action recognition model offers distinct advantages over single-species models due to its versatile applicability and out-of-the-box capabilities. The current challenge with multi-species action recognition is the variety of ways different species of animals can accomplish the same action, such as climbing. We present a method for multi-species action recognition by utilizing Contrastive Language-Image Pre-Training (Radford et al., 2021) and X-CLIP (Ma et al., 2022). CLIP encodes a latent space between an image and a set of captions, while X-CLIP encodes a latent space between several images, across the temporal setting, and a set of captions. In this project, we present the results of action recognition by utilizing CLIP and X-CLIP with the general pre-trained weights. Our findings underscore the insufficiency of generalized pre-trained CLIP and X-CLIP models for precise atomic action recognition in the animal domain, emphasizing the necessity of custom pre-training to achieve heightened accuracy.

**Improved Deep Learning-Based Reconstruction for 2D Phase Contrast MRI**  
Ishaan Mantripragada  
*Mentors: Shreyas S. Vasanawala, Matthew J. Middione, Julio A. Osceanoa, Daniel B. Ennis, and Lior S. Pachter*

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique to produce detailed soft tissue anatomical images without exposing the body to ionizing radiation. The measured MRI signal can be represented as complex numbers which generate magnitude (real component) and phase images (imaginary component). Typically, MRI images are presented as magnitude images, which generally offer a qualitative view of the patient's anatomy and structure. On the other hand, PC-MRI stores valuable information about the velocity of flowing blood. 2D PC-MRI oftentimes requires patients to hold their breath for up to 20 seconds during a scan, to eliminate the effects of respiratory motion. However, this poses challenges for both sick and older patients. To reduce breath hold durations, we propose a solution that involves undersampling the data collection, which shortens the scan time but produces inaccurate image reconstructions with conventional reconstruction approaches. We plan to use fully sampled 2D PC-MRI images and a Convolutional Neural Network to generate a deep learning-based reconstruction that learns how to reconstruct accurate PC-MRI images from undersampled images. Similar techniques have already been demonstrated in the literature and offer ≤12x reduction in scan time for magnitude images and ≤ 8x reduction in scan time for 2D PC-MRI.

**Implementing Property-Based Testing in NetworkX**  
Nilo Rivera  
*Mentors: Chris Umans and Ross Barnowski*

Currently, most software is tested with a set of example inputs and expected outputs. A newer approach is property-based testing, in which many example inputs are generated, and the outputs are tested for specific properties they are expected to have, such as commutativity. NetworkX, a Python package for graph data structures and algorithms, could benefit from this form of testing. However, as property-based testing is not yet
prevalent, I had to develop an algorithm for generating random graphs to use as test cases. This algorithm had to be capable of ensuring specified characteristics in its output, such as connectedness, and finding the simplest graph that fails a test. I will continue this work for the remainder of the summer by determining which properties each algorithm in NetworkX should have and writing tests for each of these properties, improving the efficacy, development time, and readability of the NetworkX test suite. Accompanied by an analysis and research paper, this work will contribute to the adoption of property-based testing in software used across the many industries where graphs play a vital role.