Research Statement

Jinen Setpal

Department of Computer Science, Purdue University, West Lafayette, USA.

Updated: November 4, 2023

My time as an undergraduate at Purdue granted me exposure to a vast variety of research disciplines. It put me in a unique position to experimentally narrow down my interests and identify what I wanted to pursue as a doctoral candidate. I look forward to making that future the present, as I pursue opportunities in graduate studies with a Summer/Fall 2024 intake.

When I began venturing into machine learning, better data was stressed as the most straightforward way to improve model performance. Even though this is true – given the hypothesis space of the MLP encapsulates that of all modern feedforward architectures – it is impractical to expect CNNs or the attention mechanism to be intermedially obtained by backpropagation. By leveraging intuitive biases, researchers have repeatedly allowed us to develop architectures that establish state-of-the-art solutions which aren't practically backpropagatable even with perfectly engineered data, and it is this aspect of research I'm particularly excited to explore through my doctorate program.

My primary interest therefore lies in **Deep Learning Theory**; specifically, using theoretically-backed approaches to embed distributional robustness. I'm currently exploring *intrinsic interpretabitily*; the objective of which is to develop neural networks that are interpretable by default, with intermediate representations being backpropagatable checkpoints. I want to use this to establish a curriculum-style[1] approach towards learning. My hypothesis is that by setting up even single targets as multi-levelled optimization problems, it will allow for a more humanlike interpretation of a given challenge, preventing shortcut learning and promoting emergent behaviors such as converging loss to a convex optimization space, and subsequently rudimentary distribution generalization. The final project of my deep learning class – Interpretable Risk Minimization[2] – was an effort towards demonstrating the feasibility of my proposed approach.

Alongside my education, I also work as a Machine Learning Engineer at DagsHub: it's an MLOps startup binding popular open-source tools to a single remote host. Working at a company whose business model revolves around effective development practices has allowed me to refine my approach towards machine learning projects. I have been able to establish a systematic module-based method towards writing code to evaluate a given hypothesis. Additionally, it taught me best practices for data versioning, model lineage and experiment tracking to maximize the reproducibility of my research. In addition, it also lent me valuable insight into what working professionally in a non-research role entails.

Inspired by CGP Grey[3], my long term goal - intentionally ambiguous - is to explore every exciting avenue within my chosen field of study; and to pursue absolute commitment in every responsibility I undertake.

[1] model-based, rather than a data-based curriculum.

[2] old preprint.

[3] my favorite YouTube channel, has awesome videos.

Updated: November 4, 2023

As part of my earlier revisions, I used to storyboard research projects that I contributed to semester-wise. Since that list grew, I severed it from the original statement to the journal-esque format below. It's unabridged, so gets less relevant as you traverse it. (It's basically an appendix.)

Fall 2023 This semester, I'll once again teach **CS39000-WAP**, this time with just one co-instructor. I have a lot of takeaways from having taught it last year, and look forward to implementing the improvements we've set in place. Besides that, I'm also taking two graduate courses this semester: **Advanced Topics with Large Language Models** and **Applied Regression Analysis**. The TaskBot challenge enabled my first hands-on experience with Large Language Modelling, and I look forward to working on a project leveraging intrinsic interpretability to combat hallucinated outputs under the guidance of **Prof. Dan Goldwasser**. Unfortunately, we did not make it to the TaskBot Finals, and accordingly have wrapped up our pending experiments and prepared the camera-ready paper for publication. Finally, I flew to Monterey, California to present a talk on **The Machine Learning Angle for Open Source Science** @ LFMS 2023.

Summer 2023 We're in the **TaskBot Semi-Finals**! As part of the competition, Amazon provided us AWS credits towards pursuing our research objectives. Besides the engineering challenges, it gave me the opportunity to fine-tune Large Language Models; I have never appreciated HuggingFace's frameworks so much before. At DagsHub, I went back to pursuing full-time work. I developed **DPT**, a conversational agent using OpenAI's GPT-3.5 to answer questions about DagsHub Documentation, and completed reproducing **Panoptic Deeplab**. We aim to submit this towards the next edition of the reproducibility challenge. Finally, we also launched our flagship service for the quarter: the **Data Engine**. I made open-source contributions to the engine's client, developing custom Dataset and DataLoader classes for both PyTorch and TensorFlow.

Spring 2023 I took another graduate course! This time, I took **Deep Learning** taught by **Prof. Bruno Ribeiro**. This is definitely my favourite course to date. It gave me a lot of perspective on the statistical foundations that underpin machine learning and are abstracted by modern deep learning frameworks. The knowledge I've gained from this course has helped me identify foundational improvements in models I developed that I would previously attempt to hotfix with data. Advised by **Prof. Rajiv Khanna**, I developed **Interpretable Risk Minimization**: an approach that leverages Class Activation Mappings to build an intrinsically interpretable neural network for out-of-distribution generalization. This also doubled as my final project for the Deep Learning class. That aside, I also began working towards the TaskBot Challenge. Our objective is to build a task-oriented conversational agent that guides users through recipes and DIY (do-it-yourself) tasks like building a birdhouse, or painting a fence.

Fall 2022 This semester, I got the opportunity to teach a self-developed course end-to-end. Mentored by **Prof. Buster Dunsmore**, four student instructors developed the course syllabus, assignments, slides and schedule for **CS-39000-WAP**: Web Application Programming. In addition to instructing the course, we also evaluated assignments and submitted grades. Finally, a huge first: I got my **first grant proposal accepted**! Under the mentorship of **Prof. Julia Rayz**, Purdue

Updated: November 4, 2023

will participate in the **2nd Alexa Prize TaskBot Competition** as Team BoilerBot. I also flew to Toronto, to present a talk on **Interpretable Model Optimization** @ TMLS 2022! Over the winter break, I also contributed to HuggingFace's Transformers library, developing model trainer integrations for their framework.

Summer 2022 This summer was the first time that I worked full-time! I worked as a Machine Learning Engineer at DagsHub: it's an MLOps startup binding popular open-source MLOps tools to a single remote host. I worked on reproducing CheXNet, which is now-used as the flagship onboarding project for users new to the platform.

Spring 2022 I participated in my first graduate-level course! I took **Computer Security**, covering binary exploitation and reverse engineering, under **Prof. Antonio Bianchi**. While continuing my work at the Purdue Applied Cryptography Lab, I also joined the Q-Learning and Vision Lab under **Prof. Qiang Qiu** and **Prof. Wei Zakharov**, developing systems for Drone-Mounted Video Object Tracking. Additionally, I began serving as an Undergraduate Teaching Assistant for the first time under Purdue's Data Mine, mentoring students working to create industry solutions for corporate partner MISO. I summarize and grade bi-weekly agile reports, and hold office hours and lab sessions. Lastly: I lead Purdue's TE AI Cup submission for this year. Our team earned the **Best Innovation Award** and a scholarship worth $6,000 USD. We are currently in the process of filing two patents for developments made through the duration of the competition.

Fall 2021 I began pursuing research at the Applied Cryptography Research Lab at Purdue, under **Prof. Christina Garman**. We are working to establish baseline methods to identify cryptographic algorithms embedded within binaries. My focus within the project was centered around evaluating past efforts such as CryptoKnight, CryptoHunt and using primitive NLP techniques to develop a target baseline. I am also leading a publication-track project, reproducing **Panoptic-Deeplab** with Tensorflow 2.x, as part of **Papers with Code's Reproducibility Challenge**. (Note: we couldn't train the model owing to high compute requirements, and our paper was ultimately rejected).

Research Prior to University I collaborated with faculty at CERN, developing CutLang: a domain-specific analysis description language for particle physicists conducting Large Hadron Collider (LHC) experiments. CutLang eases programming proficiency required to process raw data from experimental apparatus, allowing researchers to focus on their experiments and not have to invest time learning the ROOT Framework over C++. CutLang was presented at the **8th Annual Conference on Large Hadron Collider Physics**, and has been peer-reviewed and published to **Frontiers in Big Data** in June 2021. I also published and presented first-author research within Natural Language Processing at **EVALITA** in Italy. We developed ArchiMeDe: a model for meme classification implementing an ensemble of pre-trained backbones utilizing multimodal inputs to label images obtained from Italian social media meme accounts. ArchiMeDe was presented at the Final Workshop of the EVALITA conference, and was peer-reviewed and published to **CEUR-WS** in December 2020.