Jinqiang Yu

PhD Candidate at Monash University

Wellington Rd, Clayton VIC 3800, Australia

jinqiang.yuuu@gmail.com | linkedin.com/in/jinqiang-yu-404bb0187 | Jinqiang Yu

Summary

I am passionate about machine learning and AI areas, with research experience in the fields of machine/deep learning and explainable AI, which is a division of responsible AI. Explainable AI can be applied in machine learning area, such as debugging and monitoring machine learning models/ datasets, helping ML system deplyment, and improving LLMs to generate more understandable and trustworthy outputs, etc. As an enthusiast of cutting-edge AI techniques, I particularly has a strong desire to explore them, such as generative AI and large language models. With the enthusiasm and experience in machine learning and AI areas, I am expecting to utilise my strong technical skills to contribute to real-world projects.

Education

Monash University *Melbourne, Australia*

PhD in Data Science & AI *Feb 2021 - Aug 2024*

- **Thesis Topic**: Explainable AI with the Use of Formal Reasoning
- **Supervisors**: Prof. Peter J. Stuckey, Dr. Alexey Ignatiev
- **Thesis Description**: In this PhD project, we aim to deal with explainable AI problems, including computing interpretable machine learning (ML) models, generating accurate and concise explanations to explain ML models in various domains like NLP, CV and general classification tasks, and applying explainable AI in machine learning, such debugging ML models and improving LLMs.

Monash University *Melbourne, Australia*

Master of Information Technology *Mar 2019 - Dec 2020*

- Graduated with H1
- **Core units**: Master Minor Thesis, Applied Data Science, Machine Learning, Data Processing for Big Data.
- **Minor Thesis Topic**: Computing optimal interpretable machine learning models.
- **Thesis Description**: The thesis focuses on interpretable models, aiming at developing advanced approaches to computing machine learning models that are both accurate and interpretable.

Publications

From Formal Boosted Tree Explanations to Interpretable Rule Sets
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey
29th International Conference on Principles and Practice of Constraint Programming (CP), 2023

On Formal Feature Attribution and Its Approximation
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey
arXiv preprint arXiv:2307.03380, 2023

Eliminating the Impossible, Whatever Remains Must Be True: On Extracting and Applying Background Knowledge in the Context of Formal Explanations
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, Joao Marques-Silva
37th AAAI Conference on Artificial Intelligence (AAAI), 2023

Learning Optimal Decision Sets and Lists with SAT
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Pierre Le Bodic
Journal of Artificial Intelligence Research (JAIR) 72 (2021) pp. 1251–1279. 2021

Computing Optimal Decision Sets with SAT
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Pierre Le Bodic
26th International Conference on Principles and Practice of Constraint Programming (CP), 2020

Experience

Optima *Melbourne, Australia*

PhD Researcher *Apr 2021 - Present*

- Under the supervision of Prof. Peter J. Stuckey from Optima, I am engaged in a PhD project at Monash University. My research focuses on explainable AI, including generating interpretable ML models and computing accurate and concise explanations, aiming at developing methods to help users understand and explain ML model inferences and predictions, and applying explainable AI to improve ML performance.

Monash University *Melbourne, Australia*

Teaching Associate *Jun 2021 - Nov 2021*

- Tutoring and grading students in tutorials, assignments, and final exams.
- **Unit**: FIT5220 - Solving discrete optimisation problems.
- **Topics**: Constraint Programming, Mixed Integer Programming, Boolean Satisfiability (SAT) Solving, Large Neighbourhood Search.

Research Projects

Applying Explainable AI in Machine Learning and LLMs *Oct 2023 - present*

- Explainable AI has been widely used in ML. With the use of explainable AI, we can debug and improve machine/ deep learning model performance, help others understand the models' behavior, and build trust of the models. For example, explanation regularization (ER) methods aim to improve LLM generalization by aligning the model's machine rationales (which tokens it focuses on) with human rationales. These automated rationales provide feedback to LLMs during training. This approach can improve the accuracy of LLMs by 10-25% for various tasks, even when human rationale is lacking. Another recent work leverages richer signals beyond just input-output pairs to teach smaller models to mimic the reasoning process of large foundation models like GPT-4. Specifically, the authors collect training data consisting of prompts and detailed explanatory responses from GPT-4. To allow GPT-4 to generate explanations, system instructions such as "You are a helpful assistant, who always provides explanation. Think like you are answering to a five-year-old." are utilized. These models are demonstrated to outperform models trained using conventional instruction tuning in complex zero-shot reasoning benchmarks. In this project, we aim at connecting explainable AI and machine learning and such that improving ML/ DL/ LLM models.

Explainability in NLP and Image Classification *Dec 2022 - present*

- As ML models for NLP/image classification problems are black-box models, users cannot understand the prediction made by the models and thus it is hard to trust the predictions. Although existing model-agnostic approaches are able to provide explainability for predictions, these approaches are known to suffer from fundamental explanation issues. Inspired by the limitation, in this project, we target developing the approach to providing trustable explanations for NLP/image predictions in machine/deep learning models.

Computing Succinct and Accurate Explanations *Feb 2021 - present*

- In recent years the growing practical AI and ML applications have given the rise to Explainable AI (XAI). One of the major approaches to XAI is to compute explanations to ML predictions on demand, including post-hoc (*abductive*) explanations answering a "*why?*" question and (*contrastive*) explanations targeting a "*why not?*" question. In this project, we focus on developing the approach to computing both abductive and contrastive formal explanations making use of background knowledge, which can positively affect the quality of both kinds of explanations. In addition, we generate accurate feature attribution which indicates the contribution of a feature.

Learning Optimal Interpretable Machine Learning Models *Feb 2020 - present*

- In order to make explanations easy for humans to understand the interpretable models, e.g. decision trees, lists and sets, they should be as concise as possible. In addition, such models should provide accurate predictions such that humans can make proper decisions based on the predictions. Therefore, this project focuses on devising approaches to computing interpretable ML models that are both small in size and accurate, making use of modern formal reasoning.

Skills

Proficient in Python, Java, R, SQL.

Familiar with C/C++, Spark, MongoDB, MATLAB, ML systems, machine/deep learning models (neural networks, transformers, LLMs, GANs, and gradient boosting trees, etc.), LLM framework (LangChain), and prompt engineering (Chain of Thought Prompting, Tree of Thoughts, Satisfiability-Aided Prompting).

Experience with explainable AI, data mining, NLP, CV, and diverse libraries such as pandas, numpy, scikit-learn, TensorFlow, PyTorch, and PySAT.

Awards and Scholarships

2021-2024	**Monash Graduate Scholarship**	Scholarship covers living expenses and tuition
2020	**Best Paper Award**	Our paper "Computing Optimal Decision Sets with SAT" has been selected for the Best Paper Award for the CP/ML Track of CP 2020.

Scientific Activities

2024	PC member of the AAAI Conference on Artificial Intelligence(AAAI-2024)

References available upon request.