Caleb Kaiji Lu

Email : kaijil@andrew.cmu.edu

Education

Carnegie Mellon University Pittsburgh, Pennsylvania
PhD in Electrical and Computer Engineering *August 2017–May 2022*

Rice University Houston, Texas
Bachelor of Engineering in Electrical and Computer Engineering; Major GPA: 4.02 *August 2013–May 2017*
Bachelor of Arts in Cognitive Science; Overall GPA: 3.75 Cum Laude

Experience

CMU Accountable System Lab Pittsburgh, Pennsylvania
Research Assistant (Advisor: Prof. Anupam Datta) *Aug 2017 - Present*

Explainable NLP

- Create original gradient-based/internal/game-theoretic explanation techniques to systematically diagnose weaknesses and evaluate reliability of RNN and Transformer models for text and sequential data
- Invent *influence path*, a causal explainability method to understand how an RNN language model represents syntax such as subject-verb agreement, and compare the representation with actual grammatical rules
- Invent *influence patterns* to abstract computational graphs of Transformer models such as BERT and design algorithm to efficiently search from exponentially large number of possible abstractions
- Improve the faithfulness of explanations compared with prior methods and publish first-author papers in major AI/ML venues such as NeurIPS, ACL and KDD

Fairness in AI

- Discover substantial gender bias in various NLP tasks such as language modeling and coreference resolution
- Design algorithm to counteract bias while maintaining model utility using counterfactual data augmentation
- Quantify bias with new metrics and mitigate the bias in state-of-the-art conference resolution model by 83%
- Serve on programme committee of ACL workshops, Gender Bias in Natural Language Processing (GeBNLP)

Truera Inc. Redwood City, CA
Machine Learning Engineer Intern *May-Aug 2019, 2020, 2021*

NLP Explainability and Debugging Platform

- Build interactive platform from scratch for enterprise use with capabilities such as feature importance and interactions
- Visualize local and global token-level feature importance for Transformer-based sentiment analysis models such as BERT and RoBERTa and test on benchmarks datasets such as Covid-19 tweets and Yelp reviews
- Design debugging tool to stress test models by automatically generating adversarial examples

TruLens

- Build open-source, cross-framework library unifying a broad range of approaches for deep learning explainability
- Integrate different neural network frameworks, including TensorFlow, Pytorch, and Keras
- Present the library in multiple conference tutorials and demo including AAAI, KDD and NeurIPS

Time Series Data Explainability Platform

- Build platform for integrating, explaining, improving and monitoring production-grade RNNs model for financial corporation client, and TCNN models for insurance company client
- Implement explanation methods such as Integrated Gradients and evaluate clients' models and data
- Successfully improve the accuracy and reliability of models and deliver the POC and pilot products

Programming Skills

Languages: Python, Matlab, R, Java, SQL, C **Technologies**: TensorFlow, PyTorch, Keras, Spark, Unix

Publications

Google Scholar Link: https://tinyurl.com/4aby4jtt