Suvodeep Majumder

Email: smajumd3@ncsu.edu
Linkedin: https://www.linkedin.com/in/suvodeep-majumder/
Mobile: +1-(774)-208-663
Website: https://www.smajumder.us

Career Objective

As an Applied Scientist at Amazon AWS with a Ph.D. in Computer Science from North Carolina State University, I specialize at the intersection of Machine Learning (ML), Software Engineering (SE), and data science, with a current focus on developing generative AI models for understanding multimodal aspects of data. My expertise is grounded in extensive work with large language models, including semi-supervised fine-tuning strategies for BERT and context identification for machine translation, complemented by my Columbia University research in vision models for image segmentation and explainable tumor identification frameworks. This unique combination positions me to innovate in the rapidly evolving field of multimodal AI that bridges visual and textual understanding. Drawing on my experience in developing LLM evaluation frameworks, including "LLM as a Judge" for complex numerical reasoning, I aim to push the boundaries of AI systems that can seamlessly integrate and comprehend diverse data modalities. My goal is to apply this specialized background to create more intuitive, context-aware multimodal solutions that address complex, real-world challenges while maintaining my commitment to fairness and explainability in AI systems.

Education

- **North Carolina State University** — Raleigh, NC
  Ph.D. in Computer Science; GPA: 4.02/4.0 — *Aug 2019 - Aug 2023*
- **North Carolina State University** — Raleigh, NC
  MS in Computer Science; GPA: 4.1/4.0 — *Aug 2017 - May 2019*
- **West Bengal University of Technology** — Kolkata, India
  Bachelor of Computer Science and Engineering; GPA: 8.21/10.0 — *Aug 2008 - May 2012*

Experience

- **AMAZON AWS** — New York, NY
  Applied Scientist — *June 2024 - Present*
  - **Multimodal Interaction**: Developed bidirectional multimodal interaction framework for LLMs.
  - **Multimodal Model Based Evaluation**: Developed model based evaluation for evaluating multimodal data including documents, html, images, audio/videos.
  - **Visual Language Models**: Developed features that helps large language models to answer visual questions from visually rich documents.
  - **LLM as a Judge**: Developed frameworks for LLM as a judge for complex numerical reasoning type question answers.
- **Columbia University** — New York,NY
  Post Doctoral Research Scientist — *Aug 2023 - June 2024*
  - **Image Segmentation**: Created a model agnostic framework to extract different image, model and tumor properties to identify a probable tumor segmentation failure. This process includes first extracting different statistical features and analyzing their contribution towards segmentation failure and then use those features to build an explainable framework then can accurately identify possible segmentation failures with reasoning for failures.
- **North Carolina State University** — Raleigh, NC
  Research Assistant — *Aug 2019 - Aug 2023*
  - **Transfer Learning**: Developed a highly scalable hierarchical transfer learner to find exemplary source projects using bellwether method.
  - **AI Fairness**: Developed algorithms to build fair models by using causal inference to select data instances which are less likely to introduce bias. Also developed a minority oversampling technique which not only oversamples the minority class, also takes the protected attribute into consideration.
  - **Semi-supervised fine-tuning of Large Language Models**: Used semi-supervised strategies in large language models such as BERT for fine-tuning using a very small labeled corpus and large unlabeled corpus on classification and generative task.
  - **Large Scale Socio-Technical Graph Mining**: Developed a large-scale system for mining software engineering project repository for analyzing impact of developer code and social interaction on code quality using Social Network Analysis.
- **AMAZON AWS** — New York, NY
  Applied Scientist II Intern — *May 2022 - Aug 2022*
  - **Select Appropriate Context for NMT**: Researched on identification of relevant context and developed a system to identify context during machine translation by defining new down stream tasks for Large Language Models such as BERT to find context and fine tune the model to achieve competitive performance when compared to state of the art models.

- **AMAZON AWS** New York, NY
  Applied Scientist II Intern *May 2021 - Aug 2021*
  - **Effect of Model Architecture on Contextual NMT**: Researched on effect of model size on contextual neural machine translation and showed larger models understands context better then smaller models. Between models with similar capacity deeper models understands context better then wider models.
  - **Knowledge Distillation for Contextual Translation**: Researched on developing a system to contextual knowledge transfer between models using knowledge distillation by showing shallow decoder student models can achieve comparable performance to much deeper teacher models in contextual translations.
- **AMAZON AWS** New York, NY
  Applied Scientist II Intern *May 2020 - Aug 2020*
  - **Improving Pivot NMT Performance**: Developed a process for improving pivot Neural Machine Translation (NMT) performance by supplying linguistic information as source and target context.
  - **Automated Test Set Creation Process For Pivot NMT**: Developed a process for to automatically create contrastive and matching test suite for measuring the Neural Machine Translation (NMT) performance by disambiguation of word senses and a way of measuring the error propagation in pivot NMT translations.
- **IBM** Raleigh, NC
  Data Scientist Intern *June 2018 - Aug 2018*
  - **ML Based Insights For Developers**: Created ML based insights for dev teams by analyzing GitHub projects as part of DevOps insight team.
  - **Large Scale Data Processing**: Modeled large scale data processing system using spark cluster and python to analyze terabytes of software project raw data and create meaningful usable form to be used in machine learning algorithms.
- **Infosys Ltd.** Bhubneswar, India
  Test Analyst and Test Engineer *March 2013 - June 2017*
  - **Test Design and Execution**: Created Test Scenarios and Automation Scripts for Automation, Manual, Performance testing scenarios for web-based portal, Mobile Application and IVR systems.

Publications

- J Chakraborty, Suvodeep Majumder, and T Menzies, Bias in Machine Learning Software: Why? How? What to do?, **FSE, 2021. [ACM SIGSOFT Distinguished Paper Award Winner]**
- Suvodeep Majumder, J Chakraborty, T Menzies, When Less is More: On the Value of "Co-training" for Semi-Supervised Software Defect Predictors,**EMSE, 2023**
- Suvodeep Majumder, J Chakraborty, G R. Bai, K T. Stolee, and T Menzies, Fair Enough: Searching for Sufficient Measures of Fairness", **TOSEM, 2023**
- Suvodeep Majumder, P Mody, T Menzies, Revisiting process versus product metrics: a large scale analysis, **EMSE, 2022**
- Suvodeep Majumder, T Xia, R Krishna, T Menzies, Methods for stabilizing models across large samples of projects (with case studies on predicting defect and project health), **MSR, 2022**
- J Chakraborty, Suvodeep Majumder, H Tu, T Menzies Fair-SSL: Building fair ML Software with less data, **FSE, 2022**
- NC Shrikanth, Suvodeep Majumder, T Menzies, Early life cycle software defect prediction, why? how? **ICSE, 2021**
- J Chakraborty, Suvodeep Majumder, Z Yu, and T Menzies, Fairway: A way to build fair ML Software, **FSE, 2020**
- Suvodeep Majumder, N Balaji, K Brey, W Fu, T Menzies, 500+ times faster than deep learning:(a case study exploring faster methods for text mining StackOverflow), **MSR, 2018**
- Suvodeep Majumde, S Lauly, M Nadejde, M Federico, G Dinu, A baseline revisited: Pushing the limits of multi-segment models for context-aware translation, arXiv
- Suvodeep Majumder, J Chakraborty, and T Menzies, Communication and Code Dependency Effects on Software Code Quality (an Empirical Analysis of Herbsleb Hypothesis), arXiv
- J Chakraborty, Suvodeep Majumder, and T Menzies, Causality-Based Testing for Software Fairness, arXiv
- J Chakraborty, Suvodeep Majumder, H Tu, T Menzies, Can we achieve fairness using semi-supervised learning?, arXiv

Skills Summary

- **Languages**: Python, R, Java
- **Tools**: sklearn, Tensorflow, Keras, Sockeye, mxNet, PyTorch, pyradiomics, Torchvision

Honors and Awards

- ACM SIGSOFT Distinguished Paper Award, ESEC/FSE, 2021
- Making the Difference Award, Infosys, Jan 2017
- Certificate of Recognition, Infosys, Jan 2016
- Milestone Award for domain Champion, Infosys, Aug 2015
- Laurel Best Debutant Award, Infosys, Jul 2014
- INSTAINSTA, Infosys, Oct 2013