SURABHI BHARGAVA

surabhi9b@gmail.com | in/surabhibhargava

EDUCATION

Columbia University — New York, NY
M.S. in Computer Science, Specialization: Machine Learning — *Dec 2018*

Indian Institute of Technology, Guwahati — Guwahati, India
Bachelor of Technology in Electronics and Electrical Engineering — *Jul 2016*

INDUSTRY EXPERIENCE

Adobe Inc. — San Jose, CA
Machine Learning Tech Lead | Document Cloud — *Feb 2019 - Present*

Adobe Acrobat AI Assistant

- Leading development of the AI Assistant in Acrobat, introducing capabilities like multi-document Q&A and generative summaries.
- Developed a hybrid solution for multilingual document question answering, integrating Retrieval-Augmented Generation (RAG) and MapReduce-based methods. Experimented with LLMs like GPT, Llama, and Claude, performing prompt tuning to optimize text generation quality and style. Created an intent-based router to route user requests to appropriate services (e.g., document Q&A, Acrobat Help, or action-related Q&A).
- Designed automated and human evaluation systems, encompassing metrics used for evaluations, human study protocols, and a framework for automated assessments.

  Received TIME best invention award in 2024.

Liquid Mode and Adobe Extract

- **Document Layout Analysis**
  - Developed a framework for document layout analysis using object detection, evaluating both two-stage (e.g., Faster R-CNN) and single-stage (e.g., SSD) detectors along with feature extractors like UNet, ResNet, and VGG. Designed a single-stage document layout detector and created an evaluation metric that more accurately reflects user experience than mAP.
- **Table Structure Extraction**
  - Developed deep learning models for table structure detection in PDFs, utilizing both semantic segmentation and object detection techniques to decompose table structures.
  - Led data curation efforts by defining a data labeling taxonomy, collaborating with annotators to ensure high-quality annotations, and implementing human-in-the-loop strategies to enhance efficiency and minimize annotator fatigue.
  - Built a synthetic data generation pipeline to overcome data scarcity challenges and designed custom evaluation metrics aligned with end-user experience.

  Received TIME best invention award in 2023.

MLOps & Monitoring: Developed tools to monitor deployed models and gather insights from production data in a privacy-preserving manner, enabling the identification of data and model drift and facilitating timely data and model adjustments.

Few-shot learning and Active learning in NLP: Developed a framework for few-shot and active learning in NLP to efficiently train models with limited domain-specific data, targeting enterprise applications such as legal contracts. Leveraged pre-trained language models (e.g., BERT, RoBERTa) to enhance feature quality in low-data settings.

PayPal — Chennai, India
Software Engineer — *2016-2017*

- Hadoop GetWell Project: Developed a dashboard to monitor MapReduce jobs and task attempts, optimizing resource utilization and identifying faulty jobs, tasks, hosts, or nodes.
- Automatic Feeder Restart and Maintenance: Created a tool to continuously monitor and automatically restart over 50 log feeders, ensuring consistent availability.
- Configuration Management System: Established rules to verify attribute consistency across different data servers for Oracle and Couchbase databases.

ACADEMIC RESEARCH

Image-phrase grounding
New York, NY
Research Project, Advisor: Prof. Shih-Fu Chang & Dr. Svebor Karaman, Columbia University *Sep - Dec 2018*

- Performed image-phrase grounding by learning a multi-level common semantic space shared by textual and visual modalities. Showed 20-60% relative gains over existing state of the art for certain data sets. (Published in CVPR 2019)

Image Generation from Sketches
New York, NY
Research Project, Advisor: Prof. Carl Vondrick, Columbia University *Sep - Dec 2018*

- Performed **image-to-image translation** to generate photo realistic images from sketches using a **conditional GANs**.

Multimodal social media analysis for gang violence prevention
New York, NY
Research Project, Advisor: Prof. Shih-Fu Chang & Dr. Svebor Karaman, Columbia University *Jan - Jun 2018*

- Built a system to identify **social media content which may lead to violence**. Used a **Faster R-CNN** to identify objects of interest (related to loss, aggression or substance use) in social media images. These detections were further combined with text features to determine whether the content should be flagged. (Published in ICWSM 2019)

Generalized Nuclear Segmentation in Computational Pathology
Guwahati, India
Bachelor Thesis Project, Advisor: Prof. Amit Sethi, IIT Guwahati *Jul 2015 - Dec 2016*

- Developed a generalized technique to segment nuclei in histological images of different organs using CNNs. As a part of this work, we also released a large publicly available data set of annotated H&E stained tissue images. (Published in IEEE TMI 2017)

PATENTS

- **Fine-grained Attribution for Document Question Answering** Patent pending at USPTO, 2024
- **Processing Tables in Documents for Prompt Answering** Patent pending at USPTO, 2024

PUBLICATIONS

Total Citations: 1096 (updated 11/2024)

- **Challenges, Solutions, and Best Practices in Post-Deployment Monitoring of Machine Learning Models**
  Under submission at IJCTT Journal 2024
  S. Bhargava, S. Singhal
- **A Recipe For Taking Better Interviews** Medium article, 2022
  S. Bhargava
- **Multi-Level Multimodal Common Semantic Space for Image-Phrase Grounding** CVPR 2019
  H. Akbari, S. Karaman, **S. Bhargava**, B. Chen, C. Vondrick, S.F. Chang
- **Multimodal Social Media Analysis for Gang Violence Prevention** ICWSM 2019
  P. Blandfort, D.U. Patton, W.R. Frey, S. Karaman, **S. Bhargava**, F.T. Lee, S. Varia, C. Kedzie, M.B. Gaskell, R. Schifanella and S.F. Chang, K. McKeown
- **Data set & technique for generalized nuclear segmentation in computational pathology** TMI 2017
  N. Kumar, R. Verma, **S. Bhargava**, S. Sharma, A. Vahadane, A. Sethi
- **Timing model for two stage buffer and its application in ECSM characterization** VDAT 2015
  Y. Chaurasiya, **S. Bhargava**, A. Sharma, B. Kaur and B. Anand

INTERNATIONAL TALKS AND INTERVIEWS

- Live Interview on **Responsible development of GenAI applications** with *Ticker News, Australia*, 2024
- Featured in **The Future of LLMs** article by *UseTech*, 2024
- Featured in **Big Tech Offers to Watermark AI Content — Can AI-Generated Misinformation Be Stopped?** by *Techopedia*, 2024
- **You Deployed Your Machine Learning Model, What Could Possibly Go Wrong?**
  Re-Work Enterprise AI summit , 2022
- **Workshop on Building a Powerful Resume & SOP**
  IIITD, 2022

- **Panel Discussion - Computer Vision In Industry: Use Cases, Challenges, & Roadmap**
  Ai4 Conference, 2022
- **Active Few Shot Learning: The Future of Training Machine Learning Models**
  Codemotion Online Conference (English & Italian Edition), 2021

AWARDS

- Finalist in Venturebeat Women in AI Award, 2024
- TIME best invention award for AI Assistant in Acrobat, 2024
- TIME best invention award for Liquid Mode in Acrobat, 2023
- Two times Meta GHC Scholarship recepient – awarded to only 50 women all over the world, 2017 & 2018

TEACHING & SERVICE

- **Reviewer** at ACM Multimedia 2022, NeurIPS TRL Workshop 2022, AISTATS 2023, Elsevier Journal 2023 & 2024, NeurIPS Math AI Workshop 2024
- **IEEE Senior Member** 2024 onwards
- **Workshop on Resume and SOP Building**, IIIT Delhi 2022
- **All About Embeddings** mini-summit at Adobe 2022
- **Teaching**
  - **Course Assistant**, *Natural Language Processing* Spring 2018
  - **Course Assistant**, *Deep Learning* Fall 2017