Janvijay Singh
Champaign, Illinois

🌐 https://iamjanvijay.github.io/
✉ jvsingh2@illinois.edu ▪ +1 (470) 753-5241

EDUCATION

University of Illinois Urbana-Champaign — Urbana, IL
Ph.D. in Computer Science · Focus: Reasoning in LLMs · Advisors: Profs. Dilek Hakkani-Tur & Gokhan Tur — Aug 2024 - Present

Georgia Institute of Technology — Atlanta, GA
M.S. in Computer Science · Specialization: Machine Learning · GPA: 3.91/4.00 — Aug 2021 - May 2023

Indian Institute of Technology Varanasi — Varanasi, India
B.Tech. in Computer Science and Engineering · GPA: 9.62/10.00 · Class Rank: 2/65 — July 2014 - May 2018

INTERESTS & SKILLS

- **Interests**: Large Language Models (LLMs), Natural Language Processing, Multimodal Learning, Speech Recognition & Synthesis, Machine/Deep Learning, AI Applications for Societal Impact.
- **Languages**: Python, Bash, C++, C, Java, Objective-C.
- **Technologies**: PyTorch, TensorFlow, MXNet, JAX, CUDA, HuggingFace, DeepSpeed, Megatron-LM, vLLM, TGI, NVIDIA Triton, TensorBoard, Numpy, Pandas, Scikit-Learn, Matplotlib, MATLAB, Docker, Kubernetes, Helm, Terraform, GCP, AWS, FastAPI, gRPC, Django, Flask, Git, LaTeX.

PUBLICATIONS

- **JANVIJAY SINGH**, VILEM ZOUHAR, MRINMAYA SACHAN. *"Enhancing Textbooks with Visuals from the Web for Improved Learning"*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023. https://aclanthology.org/2023.emnlp-main.731.pdf

- **JANVIJAY SINGH**=, MUKUND RUNGTA=, DIYI YANG, SAIF M. MOHAMMAD. *"Forgotten Knowledge: Examining the Citational Amnesia in NLP"*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) 2023. ***Best Paper Honourable Mention***. https://aclanthology.org/2023.acl-long.341.pdf

- **JANVIJAY SINGH**, FAN BAI, ZHEN WANG. *"Entity Tracking via Effective Use of Multi-Task Learning Model and Mention-guided Decoding"*. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL) 2023. https://aclanthology.org/2023.eacl-main.90.pdf

- MUKUND RUNGTA=, **JANVIJAY SINGH**=, SAIF M. MOHAMMAD, DIYI YANG. *"Geographic Citation Gaps in NLP Research"*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022. https://aclanthology.org/2022.emnlp-main.89.pdf

  = *denotes equal contribution.*

EXPERIENCE

TikTok — San Jose, CA
Social Recommendation – Machine Learning Engineer 2 — Jan 2024 - July 2024
- Performed feature engineering and value tree tuning for a set of neural recommendation models, significantly improving various engagement and followership metrics through A/B testing on live traffic.

Verneek AI — Manhattan, NY
Personalised Search for E-Commerce – Machine Learning Researcher/Engineer — June 2023 - Jan 2024
- Designed a personalized multimodal retrieval system using SOTA embedding models and developed its evaluation framework, enabling text and interaction-based personalization. Fine-tuned embedding models using adversarially generated e-commerece data from LLMs (LLAMA 2 70B, GPT-4), improving personalization results.

Walmart Group — Bangalore, India
Voice-Assistant for E-Commerce – Applied Scientist 2 — Aug 2018 - July 2021
- **Speech Recognition for Indic Languages**: Improved speech recognition for Indic languages using hierarchical CTC-based neural models. Achieved a ∼38% relative WER improvement via joint-fusion of neural language models and Transducer loss. Improved decoder efficiency by introducing algorithmic approximations.
- **Speech Synthesis for Indic Languages**: Benchmarked SOTA architectures, including Tacotron2, WaveGlow, WaveNet, and ClariNet. Innovated a multilingual training framework using automatic transliteration and learnable sentence-style embeddings to enhance conversational prosody. Outperformed Google and Amazon APIs in code-mixed and vernacular domains.
- Used multi-node training with Torchrun & DeepSpeed; developed 2x fast production-ready CUDA/cuDNN code.

- **Microsoft**   Hyderabad, India
  MS Excel – Sofware Engineer Intern   *May 2017 - Aug 2017*
  - Developed UI dialogs and back-end routines for pivot-table functionality in Microsoft Excel for MacOS. Offered a full-time role for outstanding contributions.

PROJECTS

- **RNN-Transducer Prefix Beam Search**: Optimized prefix-beam search for speech recognition model with caching, batching, and 2-D beam pruning. >10x speed-up. Studied the increment in error-rates caused by these approximations. Open-sourced CUDA implementation. `https://github.com/iamjanvijay/rnnt_decoder_cuda` (66 stars)
- **RNN-Transducer Loss Function**: Devised diagonal parallelism to reduce time complexity from $\mathcal{O}(T * U)$ to $\mathcal{O}(T + U)$. Open-sourced TensorFlow implementation as a Python package. `https://github.com/iamjanvijay/rnnt` (45 stars)

ACHIEVEMENTS

- Honorable Mention for Best Paper Award, ACL conference, 2023.
- ETH Zurich Summer Research Fellowship (<0.8% acceptance), 2022.
- Best Team Award at Flipkart for Text-to-Speech excellence in vernacular domain, 2021.
- Winner, FinSBD-2 Task at FinNLP@IJCAI (Prize: USD 1000), 2020.
- Runner-Up, Walmart Data Science Hackathon (Prize: INR 20,000), 2019.
- Candidate Master on Codeforces (Max Rating: 1920), 2017.
- Ranked Top 0.82% globally in Algorithms on Hackerrank, 2017.
- 67th in ACM ICPC India Regionals (out of 402 teams), 2016.

RELEVANT COURSEWORK

- **Graduate Courses**: Deep Learning*, Machine Learning*, Natural Language Processing*, Computational Data Analysis*, Machine Learning Theory, Advanced Algorithms and Uncertainty, Computational Social Science, Languages and Computers, Language Interfaces and Communication.
- **Undergraduate Courses**: Artificial Intelligence, Computer Vision, Intelligent Computing, Theory of Computation, Optimisation Techniques, Probability and Statistics, Mathematical Methods, Operation Research, Discrete Mathematics, Computer Programming and Linux*.

  * indicates teaching assistantship.