Jay Gala

AI Resident, AI4Bharat (IIT Madras)

Website · GitHub · Google Scholar · @ Email · in LinkedIn

Education

Dwarkadas J. Sanghvi College of Engineering (University of Mumbai) 2017 - 2021
Bachelor of Engineering (B.E.) in Computer Engineering Overall GPA: **9.86/10**
Applied Math, Discrete Math, Algorithms, Machine Learning, Artificial Intelligence, Natural Language Processing.

Experience

AI4Bharat (IIT Madras) Aug 2022 - Present
AI Resident *Advisors: Prof. Mitesh Khapra, Dr. Anoop Kunchukuttan and Dr. Raj Dabre*

- Mined 5M high-quality bitext pairs from the web (ebooks, lecture transcripts, etc) using LaBSE and margin score.
- Developed SOTA IndicTrans2 translation models and created a challenging IN22 benchmark for 22 Indian languages. Notably, these models are used by the **Supreme Court of India** to translate legal proceedings and **Wikimedia Foundation** to translate Wikipedia content (Coverage).
- Developing multilingual text generation models – IndicBART v2 supporting 22 Indian languages and also exploring multilingual instruction-tuning on BLOOM (Scao et al., 2022) and LLaMa 2 (Touvron et al., 2023).
- Analyze the impact of directionality of test-sets on task-specific NMT models and in In-Context settings with multilingual LLMs such as BLOOM (Scao et al., 2022).

Research Collaboration June 2023 - Present
Independent Researcher (Remote) *Advisor: Dr. Sara Hooker, Prof. Bruce Bassett, Orevaoghene Ahia*

- Working on understanding the effective ways of data pruning for MT by leveraging checkpoints across time (CAT).
- Experimental results demonstrate superior performance using entropy measure from early model checkpoints compared to sentence embedding models for English-German (high-resource) and vice-versa for English-Swahili (low-resource).

Research Collaboration Sep 2021 - Dec 2022
Independent Researcher (Remote) *Advisor: Dr. Zeerak Talat*

- Proposed cross-dataset generalization for hate speech detection using Federated Learning extending Fortuna et al. (2021).
- Experiments show around 10% improvement in f1-score with relatively less data compared to centralized training.

University of California San Diego Jun 2021 - Jun 2022
Research Intern (Remote) *Advisor: Prof. Pengtao Xie*

- Implementation of Learning from Mistakes for Neural Architecture Search (Garg et al., 2021) in PyTorch [Code].
- Proposed an efficient multi-level optimization algorithm as an extension to Garg et al. (2021) for improving NAS by conducting performance-aware data generation using class-wise evaluation during the architecture search.
- Model-agnostic framework that can be coupled with any gradient-based (differentiable) search approaches.

Tata Consultancy Services Dec 2019 - Feb 2020
Machine Learning Intern

- Developed models using VAEs and K-means clustering for customer behavior analysis to prevent customer churn.
- Prepared a custom dataset by developing surveys to handle open-ended and closed-ended questions.
- Extracted feedback responses from handwritten survey forms using OCR achieving 12% CER and 18% WER.

Unicode Research Aug 2020 - Dec 2022
Research Student *Advisor: Swapneel Mehta*

- Worked on SimPPL to develop tools for policymakers and journalists to audit online disinformation on social media (currently supported by NYC Media Lab, Wikimedia Foundation, and AI4ABM).
- Collaborated with The Sunday Times and Ippen Digital to develop parrot.report, part of SimPPL.
- **Teaching Assistant:** Summer Machine Learning Course, UMLSC 2021, supported by **Google Research India**.

Publications

Complete List at Google Scholar ($*$ = equal contribution)

IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages [🔗] [Code]
Jay Gala*, Pranjal A. Chitale*, *et al.*
Transactions on Machine Learning Research **[In Submission to TMLR]**

NICT-AI4B's Submission to the Indic MT Shared Task in WMT 2023
Raj Dabre, Jay Gala and Pranjal Chitale
Proceedings of the 8th Conference on Machine Translation **[WMT - EMNLP 2023]**

Learning from Mistakes based on Class Weighting with Application to Neural Architecture Search [🔗]
Jay Gala, Pengtao Xie
e-Print (ArXiv) **ArXiv**

A Federated Approach for Hate Speech Detection [🔗] [Code]
Jay Gala*, Deep Gandhi*, Jash Mehta*, Zeerak Talat
17th Conference of the European Chapter of the Association for Computational Linguistics **[EACL 2023]**

Expanding Access to ML Research through Student-led Collaboratives [🔗]
Deep Gandhi, Raghav Jain, Jay Gala, Jhagrut Lalwani, Swapneel S Mehta
NeurIPS Workshop on Broadening Research Collaborations 2022 **[WBRC - NeurIPS 2022]**

Improving Image-Based Dialog by Reducing Modality Biases [🔗] [Code]
Jay Gala, Hrishikesh Shenai, Pranjal Chitale, Kaustubh Kekre, Pratik Kanani
5th International Conference on Advances in Computing and Data Sciences **[ICACDS 2021]**

Pothole Detection and Dimension Estimation System using Deep Learning (YOLO) and Image Processing [🔗] [Code]
Pranjal A. Chitale, Kaustubh Y. Kekre, Hrishikesh R. Shenai, Ruhina Karani, Jay P. Gala
35th International Conference on Image and Vision Computing New Zealand **[IVCNZ 2020]**

Projects

Ocubot - Image-based Dialog [Code] *Advisor: Prof. Pratik Kanani*

- Bachelor's project which focused on improving performance on the multimodal task of Visual Dialog.
- Adversarial analysis of existing systems to identify modality biases towards historical context and salient visual features.
- Reduced modality biases by improving visual context with dense captions and attention over these captions.
- Achieved competitive performance to the baseline with around 70% training data (85K images out of 120K images).

Anomaly Detection in ECG Signals *Advisor: Prof. Pratik Kanani*

- Industry collaboration to develop neural models for detecting anomalies in processed ECG signals from IoT devices with a human-in-the-loop approach to semi-automate the process while ensuring the safety of human lives.
- Applied distributed computing algorithms for speed improvements during inference and load balancing by 60%.

Annotated PyTorch Paper Implementations [Code]

- Annotated PyTorch implementations of deep learning papers as interactive jupyter notebooks.
- Includes papers such as Word2Vec, GloVe, KimCNN, Bahdanau Attention, Transformer, Neural Style Transfer, etc.

C Programming Exam Portal [📹]

- A paperless solution for conducting C programming exam for over 500 students at D. J. Sanghvi institution.
- Generated data-driven detailed reports for students and instructors to enhance the overall learning experience.

Skills

Languages	Python, C, Java, JavaScript, SQL, HTML5
Databases	MySQL, SQLite, PostgreSQL, MongoDB
Libraries	PyTorch, Keras, Transformers, Scikit-learn, NumPy, Pandas, OpenCV, Gensim, SpaCy, NLTK, Flask, FastAPI, Streamlit, Gradio, ReactJs, NodeJs
Others	Git, Jupyter, Docker, Raspberry Pi, LaTeX

Academic Service

Volunteer EACL 2023

Co-Curricular Activities

- Former Member of Shalizi–Stats reading group which focuses on the stats book Advanced Data Analysis from an Elementary Point of View by Cosma Shalizi and Bayesian Statistics.
- Attended the Eastern European Machine Learning Summer School (EEML) 2022.
- Former ML Collective NLP Reading Group Moderator.
- Cohere for AI Interactive Reading Group Organizer.
- **Presented Tutorial on Developing SOTA MNMT Systems for Related Languages at AACL-IJCNLP 2023.**