Jay Gala

Research Associate @ MBZUAI

🌐 Website   @ Email   GitHub   Google Scholar   Semantic Scholar   in LinkedIn

Education

Dwarkadas J. Sanghvi College of Engineering (University of Mumbai)   Aug 2017 - Jul 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

MBZUAI   May 2024 - Present
Research Associate   Advisors: Yova Kementchedjhieva and Alham Fikri Aji

> Exploring how to efficiently retrofit visual modality knowledge into pre-trained LLMs without explicit vision encoder.

AI4Bharat (IIT Madras)   Aug 2022 - Apr 2024
AI Resident   Advisors: Mitesh Khapra, Anoop Kunchukuttan and 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).

> Developed efficient Indic-Indic (non-English) translation models by repurposing components from independently pretrained English-centric translation models. Distilled IndicTrans2 translation models with a **~5x reduction in model size** and **~36% reduced inference time**. Checkout the blog for more details.

> Study various aspects influencing ICL abilities of LLMs like BLOOM (Scao et al., 2022) and Llama 2 (Touvron et al., 2023) for MT task to ascertain if ICL is example-drive or instruction-driven.

Cohere For AI Research Collaboration   Jun 2023 - Feb 2024
Independent Researcher (Remote)   Advisors: Sara Hooker, Julia Kreutzer and Bruce Bassett

> Working on understanding the effective ways of data pruning for MT by leveraging Checkpoints Across Time (**CAT**).

> Experimental results demonstrate superior performance using perplexity from early model checkpoints compared to sentence embedding models for high-resource pairs (En-De, En-Fr) and vice-versa for low-resource pairs (En-Sw).

MBZUAI Research Collaboration   Sep 2021 - Dec 2022
Independent Researcher (Remote)   Advisor: 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: 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.

Publications   Complete List at Google Scholar and Semantic Scholar (∗ = equal contribution)

[10]   **MMTEB: Massive Multilingual Text Embedding Benchmark** [Code]
Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, et al.
In Submission to International Conference on Learning Representations   **[Under Review]**

[9]   **Leverage Class-Specific Accuracy to Guide Data Generation for Improving Image Classification** [Paper]
Jay Gala, Pengtao Xie
International Conference on Machine Learning   **[ICML 2024]**

[8] **Critical Learning Periods: Leveraging Early Training Dynamics for Efficient Data Pruning in MT** [Paper]
Everlyn Chimoto, <u>Jay Gala</u>, Orevaoghene Ahia, Julia Kreutzer, Bruce Bassett, Sara Hooker
Findings of the Annual Meeting of the Association for Computational Linguistics   **[Findings - ACL 2024]**

[7] **An Empirical Study of In-context Learning in LLMs for Machine Translation** [Paper | Code]
Pranjal A. Chitale*, <u>Jay Gala</u>*, Raj Dabre
Findings of the Annual Meeting of the Association for Computational Linguistics   **[Findings - ACL 2024]**

[6] **RomanSetu: Efficiently unlocking multilingual capabilities of LLMs via Romanization** [Paper]
Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, <u>Jay Gala</u>, Thanmay Jayakumar, Ratish Puduppully, Anoop Kunchukuttan
The Annual Meeting of the Association for Computational Linguistics (🏆 **Senior Area Chair Award**)   **[ACL 2024]**

[5] **CVQA - Culturally-diverse Multilingual Visual Question Answering Benchmark** [Paper | Website]
David Romero, . . . , <u>Jay Gala</u>, . . . , Alham Fikri Aji
Conference on Neural Information Processing Systems Datasets & Benchmark track   **[NeurIPS 2024]**

[4] **Airavata: Introducing Hindi Instruction-tuned LLM** [Paper | Code]
<u>Jay Gala</u>, Thanmay Jayakumar, et al.
ArXiv Preprint (Technical Report)   **[arXiv 2024]**

[3] **IndicTrans2: Towards High-Quality and Accessible MT Models for Indian Languages** [Paper | Code]
<u>Jay Gala</u>*, Pranjal A. Chitale*, et al.
Transactions on Machine Learning Research   **[TMLR 2023]**

[2] **NICT-AI4B's Submission to the Indic MT Shared Task in WMT 2023** [Paper]
Raj Dabre, <u>Jay Gala</u> and Pranjal Chitale
Conference on Machine Translation   **[WMT - EMNLP 2023]**

[1] **A Federated Approach for Hate Speech Detection** [Paper | Code]
<u>Jay Gala</u>*, Deep Gandhi*, Jash Mehta*, Zeerak Talat
European Chapter of the Association for Computational Linguistics   **[EACL 2023]**

Projects

Ocubot - Image-based Dialog [Report | 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).

Pothole Detection and Depth Estimation [Report | Code]   Advisor: Prof. Ruhina Karani
> Developed an autonomous surveillance system for road safety to identify potholes using YOLOv4 and estimate the depth and dimensions of the pothole using triangular similarity.
> Collected and released a dataset of 1.2K pothole images annotated as per the YOLO labeling format.

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%.

Skills

Languages	Python, C, Java, JavaScript, SQL, HTML5, LaTeX
Libraries	PyTorch, Keras, Fairseq, Transformers, Scikit-learn, NumPy, Pandas, OpenCV, SpaCy, NLTK, Flask, FastAPI
Others	Git, Jupyter, Docker, Raspberry Pi, LaTeX

Academic Service

Volunteer	EACL 2023, ACL 2024
Reviewer	EACL 2024, ACL Rolling Review

Co-Curricular Activities

> Gave a talk on in-context learning capabilities of LLMs for MT at the SNLP Reading Group, Microsoft Research India.
> **Presented Tutorial on Developing SOTA MNMT Systems for Related Languages at AACL-IJCNLP 2023.**
> **Teaching Assistant** for Summer Machine Learning Course, UMLSC 2021, supported by **Google Research India**.
> Collaborated with SimPPL to develop Parrot, a tool for auditing online disinformation on social media, in partnership with The Sunday Times and Ippen Digital.