Akshay Goindani

Available to work immediately

agoindan@andrew.cmu.edu | 412-696-2588 | LinkedIn | Google Scholar | Personal Website

Education

Carnegie Mellon University, School of Computer Science December 2024
Master of Science in Artificial Intelligent Information Systems, GPA: 4.17/4 *Pittsburgh, PA*

International Institute of Information Technology July 2022 (Graduated)
B.Tech (Honours) & M.S. (Research) in Computer Science & Engineering, GPA: 8.99/10 *Hyderabad, India*

Work Experience

Voyage AI February 2025 - Present
Founding Research Engineer *Palo Alto, CA*

- Led the development of instruction following capabilities for rerankers, resulting in the release of voyage-rerank-2.5 models, first in the industry that follow instance-level instructions, with improved performance on all benchmarks.
- Driving the development of the next-generation of multimodal embedding models, further enhancing the capabilities of voyage-multimodal-3, a state-of-the-art embedding model.
- Optimized multimodal training infrastructure with improved load-balancing and batch scheduling, boosting throughput by 40–50%.
- Advancing model adaptability through continual learning and model-merging approaches, ensuring long-term scalability and consistent performance across evolving tasks.

Machine Learning Intern *June 2024 - September 2024*

- Developed large-scale data curation, de-duplication, and quality filtering for token-efficient training. Improved data quality leads to the release of state-of-the-art voyage-3 text-embedding models
- Synthetic multimodal data generation for training embedding and reranking models. Generated high-quality synthetic data improves Vision-Language embedding models, and led to the release of voyage-multimodal-3.

ExaWizards | AI platform July 2022 - June 2023
Associate Machine Learning Engineer *Hyderabad, India*

- Deployed a pipeline for Temporal Activity Localization in videos using Natural Language Description.
- Enhanced the efficiency of the object recognition module, improving the inference speed of the platform by 30%

Amazon | International Machine Learning Team May 2022 – July 2022
Applied Scientist Intern *Bangalore, India*

- Built Reinforcement Learning based Attribute Extraction Model to predict missing values in product descriptions
- Leveraged trajectories of trained online RL agents to design Decision Transformer, improving recall by 8.5 points

ExaWizards | AI platform Jun 2021 - Aug 2021
AI Engineering Intern *Hyderabad, India*

- Designed Deep Learning techniques to retrieve body poses from images and videos in real-time, by utilizing probabilistic view-invariant pose embeddings to compute K-Nearest Neighbors of a query image

Selected Publications

- Liang, Paul Pu*, Akshay Goindani*, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. ”Hemm: Holistic evaluation of multimodal foundation models.” Advances in Neural Information Processing Systems 37 (2024): 42899-42940. NeurIPS 2024(* Equal Contribution)
- **Goindani, A**. and Shrivastava, M., 2021. A Dynamic Head Importance Computation Mechanism for Neural Machine Translation. International Conference on Recent Advances in Natural Language Processing.
- Sivaprasad, S.*, **Goindani, A.***, Fritz, M. and Gandhi, V., Class-wise Domain Generalization: A Novel Framework for Evaluating Distributional Shift. NeurIPS 2022 Workshop on Distribution Shifts.(* Equal Contribution)

Research Experience

Language Technologies Institute, Carnegie Mellon University August 2023 – May 2024
Research Student | Advisor: Prof. Ruslan Salakhutdinov, Prof. LP Morency, Paul Pu Liang *Pittsburgh, PA*

- Worked on evaluating and improving the performance of Generative Vision-Language Models (e.g., GPT4V, InstructBLIP). Proposed a diverse and comprehensive benchmark for evaluation, published at Neurips 2024

Language Technologies Research Centre March 2019 – July 2022
Research Assistant | Advisor: Professor Manish Shrivastava *Hyderabad, India*

- Developed a dynamic head importance computation mechanism for LLM based Neural Machine Translation.
- Augmented Transformer architecture that outperforms baseline Transformer model by a large margin, especially in low resource conditions, and learns better word alignment. Published in *RANLP 2021* [Paper].

Centre for Visual Information Technology June 2020 – Dec 2022
Research Assistant | Advisor: Professor Vineet Gandhi *Hyderabad, India*

- Proposed a novel class-wise domain generalization framework for evaluating distributional shift for image classification. Developed an Iterative Domain Feature Masking method that achieves SOTA performance [Paper]

PreCog Research Group Aug 2021 – Mar 2023
Research Assistant | Advisors: Prof. Ponnurangam Kumaraguru, Prof. Jisun An *Hyderabad, India*

- Analyzed hate speech on Twitter, and the impact of offline events during COVID-19 on online user activity
- Built BERT-based classifiers to detect religious hate speech in tweets, and predict user behavior

University of Calgary Jun 2021 – Aug 2021
Research Intern | Advisor: Professor Hadi Hemmati *Calgary, Canada*

- Developed Explainable AI model to generate interpretations for complex deep learning models' (e.g., CodeBERT) predictions, on sequence-to-sequence tasks such as method name prediction, code documentation generation

Sungkyunkwan University Dec 2020 – May 2021
Research Intern | Advisor: Professor Hogun Park *Seoul, South Korea*

- Proposed an approach for Augmenting Knowledge Graphs to Question-Answering Systems, using Graph Neural Networks (GNN), to impart commonsense knowledge to QA models, for Open-Domain Question-Answering task

Key Projects

Needle: A PyTorch-like training framework | LLM Training

- Built a deep learning framework in C++ and Python, supporting tensor operations (convolution, broadcasting, transpose) for training neural networks.
- Optimized training on NVIDIA GPUs with custom CUDA kernels for reduction, matrix multiplication, and other critical operations.
- Designed computation graph execution with backpropagation via topological sort, implementing efficient forward and backward passes for numerous operations.
- Extended capabilities with gradient checkpointing and FP16 mixed-precision training, significantly improving memory efficiency.

Synthetic Data Generation for Off-Policy Preference Optimization | Reasoning

- Generated a synthetic preference dataset using a model pool for fine-tuning vision language models for better abductive reasoning. [Dataset]
- Used CLIP for scoring the model generations, the scores are then used to determine the preference order.
- Fine-tuned PaliGemma-3B on the preference data using Direct Preference Optimization [Code].

MinBERT Classifier | Large Language Models (LLMs)

- Implemented the BERT architecture from scratch with essential components such as Positional Embeddings, Multi-Head attention, etc., and trained it for sentiment classification. Also implemented the AdamW optimizer for training/fine-tuning the LLM [PyTorch, Huggingface, GitHub]

Learning Bilingual Word Embeddings with Minimal Bilingual Data | Natural Language Processing

- Implemented unsupervised method to learn bilingual word embeddings for English & Italian, using a common embedding space via parameterized linear transformation
- Incorporated supervised learning with few known translation pairs to bilingual dictionary iteratively [GitHub]

Hybrid Machine Translation

- Proposed a Machine Translation approach that utilizes a combination of phrase tables extracted with Statistical Machine Translation methods, and a Sequence-to-Sequence architecture for Neural Machine Translation
- The proposed approach outperforms Bi-LSTM with attention mechanism by 1 - 1.3 BLEU points for low resource languages [GitHub]

Teaching Experience

- Tutored over 150 graduate students, on Transformers and LLaMA, designed and evaluated assignments for the course: Advanced Natural Language Processing, under Professor Graham Neubig.
- Mentored over 200 undergraduate students, developed assignments on ML algorithms and techniques such as Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Regularization, for course Statistical Methods in AI.

Skills

Languages: Python, C/C++, Matlab, Bash, HTML/CSS, JavaScript

Libraries: PyTorch, TensorFlow, Pandas, LIME, Fairseq, Flask, Hugging Face

ML/AI Techniques: LLMs, GPT, Transformers, LLaMA, BERT, RNN, CNN, GRU, LSTM, VisualBERT, GNN

Honors & Awards

- **Dean's List Award** for Academic Excellence (Top 5% of all students), 2018 - 2021
- ACM ICPC 2019 (Challenging Programming Contest, Worldwide) - Secured rank in Top 100 at Regional Level
- Recipient of competitive MITACS Globalink Graduate Fellowship for $15,000

Courses

Artificial Intelligence	Advanced Natural Language Processing, Multimodal Machine Learning, Question Answering, Machine Learning, Computer Vision, Vision Learning and Reasoning, Artificial Intelligence
Computer Systems	Database Systems, Operating Systems, Software Engineering, Digital Signal Analysis and Applications, Computer Graphics
Mathematics	Discrete Maths, Linear Algebra, Probability & Statistics, Complex Analysis, Multivariate Analysis, Formal Methods
Security & Networks	Advanced Computer Networks, Principles of Information Security
Algorithm & Programming	Data Structures and Algorithms, Computer Programming