Lilly Kumari

☎ (+1) 206-765-0813 | ✉ lkumari@uw.edu | lillykumari8.github.io | lillykumari8 | lillykumari

Education

University of Washington *Seattle, WA*

PhD Candidate in Electrical and Computer Engineering *Oct 2018 - Mar 2025 (expected)*

- MELODI Lab, Advisor: Prof. Jeff Bilmes, GPA: 3.95/4.0

Indian Institute of Technology, Roorkee *Roorkee, India*

Bachelors in Electrical Engineering *2012 - 2016*

- GPA: 9.5/10.0, Silver Medalist

Research Experience

UW MELODI Lab *Seattle, WA*

Research Assistant (Advisor: Jeff Bilmes) *Oct 2018 - present*

- Working on **memory selection in large multimodal models** for long-range video understanding.
- Working on a **submodular data selection framework for targeted instruction tuning** of Large Language Models (LLMs).
- Worked on **targeted data selection** for **few-shot fine-tuning of Vision Language Models (VLMs)** in a parameter-efficient manner.
- Proposed a two-stage submodular optimization framework for **data-efficient in-context learning using LLMs** and achieved state-of-the-art results for few-shot learning on several natural language processing tasks.
- Proposed a novel retrospective **adversarial data augmentation algorithm for mitigating catastrophic forgetting in continual learning** of deep models and showed state-of-the-art results on incremental learning benchmarks.
- Proposed a two-stage submodular span **query-focused summarization** framework utilizing a single submodular function to capture **query relevance and representativeness** and achieved state-of-the-art **document, video, and image summarization** results. Used dense representations from **fine-tuned BERT, GAN, and CLIP** for instantiating the **submodular functions for multimodal data**.

Google *Seattle, U.S.A*

Student Researcher *Jan 2024 - Aug 2024*

- Designed a novel framework for **long-context KV cache summarization for efficient inference** using transformers-based Large Language Models (LLMs). Achieved state-of-the-art results on various **long-context natural language processing** tasks from LongBench using LLaMA and LongChat models.

Google *Mountain View, U.S.A*

Research Intern *Jun 2023 - Sep 2023*

- Worked on **Retrieval Augmented Generation (RAG)** for long-range **persona-grounded and knowledge-grounded dialog modeling**.

Google Research *NYC, U.S.A*

Research Intern (Mentors: Srikumar Ramalingam, Ayan Chakrabarti) *Jun 2022 - Sep 2022*

- Worked on data re-weighting and **curriculum-guided replay** strategies for faster convergence of deep neural networks training.
- Explored a novel **knowledge distillation**-based loss objective for performing **efficient importance sampling** using lightweight models.

Adobe Inc. *Bengaluru, India*

Member of Technical Staff *Jun 2016 - Jul 2018*

- Worked on fashion attribute prediction for **large-scale visual search and ranking** in the E-Commerce domain, involving millions of product images.
- Collaborated with Big Data Experience Lab and proposed modifications to existing **positive-unlabeled learning** techniques for **non-human traffic detection** in analytics data.
- Designed an **LSTM-based deep learning model** that predicts **personalized skill sets** required for completing Adobe Photoshop tutorials.
- Developed and optimized machine learning models for **Click-Through Rate (CTR) prediction on highly imbalanced datasets**.

Adobe Research *Bengaluru, India*

Research Intern (Mentors: Ritwik Sinha, Atanu Sinha) *May 2015 - Jul 2015*

- Developed a novel machine learning pipeline for **dynamic audience segmentation** and predicting the **segment membership** of a new reader, enhancing **content recommendation and personalization**.

Publications

BumbleBee: Dynamic KV-Cache Streaming Submodular Summarization for Infinite-Context Transformers

Lilly Kumari, Shengjie Wang, Tianyi Zhou, Nikhil Sarda, Anthony Rowe, Jeff Bilmes. *COLM, 2024.* [paper]

An End-to-End Submodular Framework for Data-Efficient In-Context Learning

Lilly Kumari, Shengjie Wang, Arnav Das, Tianyi Zhou, Jeff Bilmes. *NAACL Findings 2024.* [paper | code]

High resolution point clouds from mmwave radar

Akarsh Prabhakara, Tao Jin, Arnav Das*, Gantavya Bhatt*, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, Anthony Rowe. *ICRA, 2023.* [paper | code]

Retrospective Adversarial Replay for Continual Learning

Lilly Kumari, Shengjie Wang, Tianyi Zhou, Jeff Bilmes. *NeurIPS (NIPS), 2022.* [paper | code]

Submodular Span, with Applications to Conditional Data Summarization

Lilly Kumari, Jeff Bilmes. *AAAI, 2021.* [paper]

Audience Prism: Segmentation and Early Classification of Visitors Based on Reading Interests

Lilly Kumari, Sunny Dhamnani, Akshat Bhatnagar, Atanu R Sinha, Ritwik Sinha. *India-KDD-CoDS, 2016.* [paper]

Workshop Publications

COBRA: COmBinatorial Retrieval Augmentation for Few-Shot Learning

Arnav Das*, Gantavya Bhatt*, Lilly Kumari, Sahil Verma, Jeff Bilmes. *ICML Workshop on Data-Centric Machine Learning Research, 2024.* [paper]

Retrieval Augmented Generation for Dialog Modeling

Lilly Kumari, Usama Shafqat, Nikhil Sarda. *NeurIPS Workshop on Efficient Natural Language and Speech Processing, 2023.* [paper]

Botcha: Detecting malicious non-human traffic in the wild

Sunny Dhamnani, Ritwik Sinha, Vishwa Vinay, Lilly Kumari, Margarita Savova. *RecSys Workshop on Online Misinformation- and Harm-Aware Recommender Systems, 2020.* [paper]

Patents

Classification of website sessions using one-class labeling techniques	*US-PTO 10785318*
Detecting robotic internet activity across domains utilizing one-class and domain adaptation machine-learning models	*US-PTO 15982393*
Makeup identification using deep learning	*US-PTO 10755447*

Teaching Experience

Signals, Systems, and Data (EE 242)

UW, Seattle

Teaching Assistant (Instructor: Nathan Kutz) — *Autumn 2024*

TinyML (EEP 595)

UW, Seattle

Teaching Assistant (Instructor: Dinuka Sahanabandu) — *Spring 2024*

Introduction to Statistical Learning (EE 511)

UW, Seattle

Teaching Assistant (Instructor: Jeff Bilmes) — *Winter 2024*

Information Theory (EE 514)

UW, Seattle

Teaching Assistant (Instructor: Jeff Bilmes) — *Autumn 2021*

Deep Learning (EEP 596)

UW, Seattle

Teaching Assistant (Instructor: Jeff Bilmes) — *Spring 2021*

Advanced Introduction to Machine Learning (EEP 596) *UW, Seattle*

TEACHING ASSISTANT (INSTRUCTOR: JEFF BILMES) *Winter 2021*

Honors & Awards

2022	**NeurIPS Scholar Award**
2022	**NVIDIA Academic Hardware Grant Award**
2020	**Azure Compute Grant Award for $40k**
2016	**Institute Silver Medal - Department Rank 1 at IIT, Roorkee**

Services

Program Committee (Reviewer)	ICML 2022-2023, SubsetML@ICML 2021, ICLR 2023, NeurIPS 2022-2024, ENLSP@NeurIPS 2023-2024, ARR 2023-Present (NAACL 2024, ACL 2024, EMNLP 2024, NAACL 2025), AISTATS 2025

Interests and Skills

Interests	Generative AI, Large Language Models, Multimodal LLMs, NLP, Efficient Deep Learning
Languages	Python, R, C++, LaTeX
Packages	PyTorch, TensorFlow, Hugging Face, spaCy, NLTK