

Sandeep Kumar Routray

sroutra2@cs.cmu.edu | (412) 450-2068 | [linkedin.com/in/sandeep-routray](https://www.linkedin.com/in/sandeep-routray) | sroutray.github.io

EDUCATION

Carnegie Mellon University

Master of Science in Machine Learning

Relevant Coursework: Deep Reinforcement Learning, Probabilistic Graphical Models, Multimodal Learning

Dec. 2025

GPA: 4.00 / 4

Indian Institute of Technology Kanpur

Bachelor of Technology in Electrical Engineering | Department Rank: 2

Relevant Coursework: Data Structures, Algorithms, NLP, Digital Signal Processing, Optimization Algorithms

Honors: Dean's List (all semesters), Summa Cum Laude, Gold Medal for Undergrad Project

May 2021

GPA: 4.00 / 4

PUBLICATIONS

[1] S. R. Dash*, S. Routray*, P. Varshney* and A. Modi, “CS-NET at SemEval-2020 Task 4: Siamese BERT for ComVE”, in Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics (COLING), Dec 2020

EXPERIENCE

Carnegie Mellon University

Graduate Research Assistant, Prof. Deepak Pathak's Lab

- Research on adapting **multimodal LLMs** and **video diffusion models** to predict actions from internet-scale human and robot video datasets enabling few-shot task, environment and embodiment generalization

Oct. 2024 - Present

Pittsburgh, PA

Samsung Research | [\[CES 2024\]](#)

Machine Learning Engineer, SmartThings Team

- Spearheaded a project to convert home layouts to 3D models. Showcased at **CES 2024** and deployed across **1 million homes** globally
- Trained a **ConvNext** model with **focal loss** to identify rooms, walls, doors and junctions and designed a custom raster to vector pipeline
- Performed integer quantization for mobile deployment with TF Lite C API obtaining **4x** reduction in size and **3x** increase in inference speed
- Achieved 3D reconstruction from single image by training **neural radiance field (NeRF)** on multi-views generated from a diffusion model

Sep. 2021 - June 2024

Seoul, South Korea

Vector Institute for Artificial Intelligence

Research Fellow, Prof. Sanja Fidler's Lab

- Leveraged inter-image relationships in a **Slot Attention** framework to learn object-centric features with self-supervised learning (SSL)
- Created an image context aware score function to mine positives and negative slots for **contrastive loss** to improve feature consistency
- Ablated **vision transformers** training with SSL losses on multi-GPU clusters, obtained **2 % mIoU** improvements over existing baselines

Oct. 2020 - July 2022

Toronto, Canada

Samsung Research | [\[Report\]](#)

Software Engineer Intern, 6G Research Team

- Implemented a reinforcement learning based scheduler for LTE system with **Deep Deterministic Policy Gradient (DDPG)** algorithm
- Devised two reward mechanisms to maximize throughput while maintaining QoS requirements of delay and fair allocation among users
- Obtained **80% lower delay** and better **user scalability** than prevalent Proportional Fair scheduler without compromising data rates

May 2020 - July 2020

Seoul, South Korea

PROJECTS

Simulator-based Scaling of Inference Time Compute for Robotics

- Scaling inference-time compute for robotics by combining **Chain-of-Thought** reasoning with diffusion transformer **world model** rollouts
- Attained **30%** improvement with **model-based RL** and **reward modeling** to optimize trajectory search and policy performance at test time

Jan. 2024 - Present

Common Sense Validation And Explanation | [\[Paper\]](#)

- Proposed a **Siamese architecture** and **Mixture-of-Experts** with encoder based **LLMs** for efficient inter-relational information extraction
- Coupled with cross attention, achieved **94.8%** accuracy for Validation task and **89%** for Explanation task. Results published in **COLING '20**

June 2020 - Dec. 2020

Minimax Optimization in Non-Euclidean Space Using Bregman Divergences | [\[Slides\]](#)

- Designed a novel restarting algorithm to minimize smooth, strongly convex functions in non-Euclidean space based on Nesterov's AGD
- Proposed a new algorithm for smooth minimax optimization using above result. Improved convergence rate by **order of 2** in both cases

May 2020 - Nov. 2020

SKILLS

Languages: C, C++, Python, MATLAB, SQL

Technologies: Docker, Git, PyTorch, JAX, CUDA, ONNX, TF Lite, NetworkX, OpenCV, Fast APIs