Soham Dinesh Tiwari

soham.tiwari800@gmail.com | LinkedIn: soham-tiwari | Google Scholar | https://sohamdtiwari.com | +1 412-909-7089

EDUCATION

Carnegie Mellon University (CMU) Pittsburgh, PA

Master of Science in Intelligent Information Systems (MIIS) | CQPA 4.08 Dec 2023

Courses: *Advanced NLP, Question Answering, Multimodal ML, Artificial Social Intelligence, Speech Processing*

Manipal Institute of Technology (MIT) Manipal, India

Bachelor of Technology Honours Computer Science & Engineering, Computational Mathematics | 9.75/10.0 Jul 2022

SKILLS

Programming Languages: Python, Java, C/C++, JavaScript, Swift

Tools & Frameworks: Jax, PyTorch, TensorFlow, HuggingFace, AllenNLP, AWS, Azure, spaCy, Git, Jupyter Notebooks

WORK EXPERIENCE

Apple Seattle, WA

AIML Natural Language Intern May 2023 - August 2023

- Identified and mitigated weaknesses of TinyBERT style distillation for a decoder-only LLM, obtaining a better student model than training a small model without distillation. This process helped me understand the details of LLM training.
- Developed online LLM compression framework on distributed GPUs, removed bottleneck of fitting two decoder-only teacher and student models into GPU memory using model sharding and removing redundant objects in training.

Nanyang Technological University Singapore, Singapore | Remote

Research Intern Aug 2021 – Jul 2022

- Improved performance of pre-trained audio neural network (PANN) using frequency dynamic convolutions to mitigate translation invariance along frequency axis of log-Mel spectrograms, tuned Fmax, to improve F1 from 75% to 78%.
- Accepted paper at 🔗APSIPA 2022 on new curriculum-learning approach for the Automated Audio Captioning system.
- Published First-author 🔗poster at NeurIPS ENLSP Workshop 2021 and 🔗paper in IJACSA on related work.

Forty4Hz INSIA Bangalore, India | Remote

ReactJS and Data Science Intern Sep 2020 – Jun 2021, Jan 2022 – Jun 2022

- Led development of two production ReactJS platforms; integrated Bi-LSTMs in Python backend to model client data.

Gravitas AI London, UK | Remote

Natural Language Processing Engineering Intern Aug 2021 – Oct 2021

- Created a prototypical medical text processing pipeline - named entity recognition, co-reference resolution, and relationship extraction using SciSpacy and CoreNLP.
- Constructed a Neo4j knowledge graph containing subject-predicate-object triplets and a medical ontology utilizing Stanford's Protege Ontology Builder for use by senior researchers in the company.

University of British Columbia Vancouver, Canada | Remote

MITACS Globalink Research Intern May 2021 – Aug 2021

- Decoded EEG signals of infants using ML and signal processing and determined whether infants understood animacy.

RESEARCH PROJECTS

Language-Agnostism, ChatGPT (LLM) Query Rewriting for Multilingual Document QA | CMU Jan 2023 - May 2023

- Increased Vietnamese and French document grounded question answering Recall@1 scores by 22% over baseline.
- Investigated query rewriting performance using ChatGPT to achieve better retriever question understanding and improve detection of context switches in the last conversation turn, by employing LLM prompt engineering.
- Accepted paper at ACL 2023 🔗Third DialDoc Workshop. *https://aclanthology.org/2023.dialdoc-1.11.pdf*

On-Device Interactive Multimodal Educational Virtual Assistant | CMU Jan 2023 - May 2023

- Built a multimodal agent that runs on 4GB RAM Jetson Nano. Leveraged videos from depth cameras for pose estimation, direction of arrival of sound, to locate the user, on device. Used Azure Speech AI for speech recognition.
- Demo at 🔗ISLS 2023. *https://par.nsf.gov/biblio/10437737-traveling-bazaar-portable-support-face-face-collaboration*

Cascaded Code-switched Speech to Monolingual Speech Translation | CMU Jan 2023 - May 2023

- Developed a cascaded, speech to speech translation 🔗 system for code-switched Indic languages, using 🔗 Branchformer and 🔗 Conformer architectures and TTS. First such effort on the 🔗 Prabhupadavani dataset.

Optimal Resource Allocation for Multilingual Finetuning | CMU Aug 2022 - Dec 2022

- Used active learning and sample uncertainty using mBERT and LaBSE to identify a minimal set of source language texts that yield the best performance on unseen languages from the Multilingual Amazon Review Corpus.

Embodied Vision and Language Navigation | CMU Aug 2022 - Dec 2022

- Improved agent's collision recovery ability in the Room-Across-Room dataset's simulated house environments using self-orienting heuristics, curiosity, and alignment-based reinforcement learning reward functions.