

Harshita DIDDEE

SCAI Center Fellow @ Microsoft Research

@ harshitadd@gmail.com  github.com/harshitadd  Google Scholar  Website

EDUCATION

2017 B.Tech, COMPUTER SCIENCE AND ENGINEERING, Guru Gobind Singh Indraprastha University
2021 Major GPA - 9.5/10.0, CGPA - 9.46/10.0

Graduated as the Best Outgoing Student for the Class of 2021

EXPERIENCE

July 2021 | SCAI Centre Fellow, MICROSOFT RESEARCH INDIA, Bangalore, India
Present
> Developed Quantized and Distilled Machine Translation models (approximate size being <400MB) for extremely low-resource languages (data <25K sentences).
> Evaluating the efficacy of using such lightweight models to provide dynamic assistance to data providers via assistive-translation interfaces.
> **Mentored by Kalika Bali, Principal Researcher**

Neural Machine Translation Data Quality Estimation

June 2022 | Visting Pre-Doc for JSALT'22, JOHNS HOPKINS UNIVERSITY, Baltimore, USA
August 2022
> Evaluated the generalizability of speech and text cross-lingual models to extremely low-resource languages.
> Exploring the development of multilingual variant of speechT5.
> **Participated in the Speech Translation for Under-Resourced Languages Track**


Neural Speech Translation

October 2020 | Intern, AI4BHARAT, Remote
March 2021
> Implemented a tesseraact-based OCR pipeline;
> Assisted the mining of parallel sentences from a dense (12M+), monolingual embedding space (between a target and source language) using FAISS
> **Mentored by Dr.Mitesh M. Khapra, Associate Professor (CSE), IIT Madras**

Neural Machine Translation Optical Character Recognition

March 2020 | Research Intern, IIIT DELHI, Delhi
November 2020
> Validating a Cross-Silo Federated Learning (FL) hypothesis for heterogeneous clients having distributed and exclusive feature space.
> **Mentored by Dr.Koteswar Rao Jerripothula, Assistant Professor (CSE), IIIT Delhi**

Cross-Silo Federated Learning Dropout Regularization

May 2019 | Research Intern, CELESTINI PROGRAM INDIA, IIT Delhi
October 2019
> Developed a federated learning enabled custom deep learning model that powers VisionAir, an android application that predicts the Real-Time Air Quality Index of an image.
>  **Work published by TensorFlow**
> **Mentored by Dr.Aakanksha Chowdhery, Google Brain**

Cross-Device Federated Learning Computer Vision

SELECT PUBLICATIONS - GOOGLE SCHOLAR

TOO BRITTLE TO TOUCH : COMPARING THE STABILITY OF QUANTIZATION AND DISTILLATION TOWARDS DEVELOPING LIGHTWEIGHT LOW-RESOURCE MT MODELS OCTOBER 2022

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, Kalika Bali

 Code  Paper

To Appear in the Conference In Machine Translation, 2022 (WMT)

Harshita Diddee*, Kalika Bali*, Monojit Choudhury*, Namrata Mukhija*

[🔗 Paper](#)

ACM COMPASS, 2022

SAMANTAR : THE LARGEST PUBLICLY AVAILABLE PARALLEL CORPORA COLLECTION FOR 11 INDIC LANGUAGES

FEBRUARY 2022

Ramesh et. al.

[🔗 Paper](#)

Transactions of the Association for Computational Linguistics (TACL)

PSUEDOPROP AT SEMEVAL-2020 TASK 11: PROPAGANDA SPAN DETECTION USING BERT-CRF AND ENSEMBLE SENTENCE LEVEL CLASSIFIER

DECEMBER 2020

Annirudha Chauhan, Harshita Diddee

[🔗 Paper](#) [🔗 Code](#)

Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval)

 HONORS AND GRANTS

- September 2020 [ACM] Grant to present work at The 35th IEEE/ACM International Conference on ASE
- August 2020 [Google AI] Selected to attend the Google AI Summer School 2020
- March 2020 [Google LLC] Travel Grant to attend the TensorFlow Dev Summit at Sunnyvale, CA
- October 2019 [The Marconi Society, Google LLC] Runner's Up at the Celestini Prize India 2019
- October 2019 [Jointly by the Government of Singapore and India] Awarded 8000 SGD as 1st Runner's Up at the Singapore India Hackathon
- March 2019 [Government of India] Winner at Nationals for the e-Yantra Robotics Competition

 SELECT PROJECTS

INMT-LITE

JULY 2021 - PRESENT

[🔗 Codebase](#)

Advised by Tanuja Ganu, Dr. Monojit Choudhury, Sandipan Dandapat and Dr. Kalika Bali, Microsoft Research India

Acknowledging that there will always be languages which have little to no data : we are working with Gondi , a severely under-resourced language spoken by 2.4M across India, to identify (a) if intermediate models (trained with < 25000) samples can be utilized to generate recommendations that can accelerate the data provision yield of a crowdsourced system targetting such a language's collection and (b) what would be the best interface (dropdown lists, Bag of Words, gisting, etc) to present such low-quality assistance so as to not make it disruptive for the user's task to providing data.

Low-Resource Machine Translation Quantization Distillation

MODEL SELECTION FOR LOW-RESOURCE ASR

OCTOBER 2022 - PRESENT

[🔗 Codebase](#)

Advised by Dr. Sunayana Sitaram and Dr. Kalika Bali, Microsoft Research India

Evaluation metrics are often not accurate proxies of the performance of systems on low-resource languages so I'm exploring (a) under what conditions can these metrics be used optimal model selection and (b) are there more representative metrics that can be monitored during training to select the best model.

Automatic Speech Recognition

VISIONAIR

JUNE 2019 - FEBRUARY 2020

[🔗 Codebase](#) [🔗 Project Website](#) [🔗 Video](#)

Advised by Dr. Aakanksha Chowdhery, Google Brain

I worked on developing a cascaded CNN + random forest based classifier that utilized meteorological A privacy-preserving smartphone application that can be used to estimate the Air Quality Index from an image that the user takes. The application uses On-Device training to ensure complete isolation of the user's private data and also uses federated learning to diversify the response of the global deep learning model iteratively.

Federated Learning Image Processing

</> COMMUNITY OUTREACH, LEADERSHIP AND VOLUNTEERING

- Present Organizer and Reviewer for SLT'2022 Hackathon
- 2021 - 2022 Organizer for Technology for Empowerment Reading Group at Microsoft Research India
- 2019-2021 Volunteer and Speaker PyData, NumFOCUS: [🔗 Talk on Open Problems in Federated Learning](#)
- July 2020 Women In Datascience: [🔗 Talk on Emerging Setups in Federated Learning](#)