

Ayush Kumar Tarun

📞 (+91)-704-827-5592
✉ ayushtarun210@gmail.com
in ayushkumartarun
🌐 ayushkumartarun

Education

2018 - 2022 **B.E. Computer Science with Minor in Data Science**, Birla Institute of Technology and Science, Pilani, India, CGPA - 9.30/10

Experience

- Jan 2022 - **ML Engineer**, Pepper Content, India
- Present
- Automated the entire business's creator allocation pipeline by developing a recommender system using state-of-the-art NLP models. Used embeddings of a creator's past work to calculate topic relevance for an assignment and added business metrics for the recommender to optimize on. *Reduced the mean allocation time from 29 hours to 1.5 hours.*
 - Developed the in-house NMT (neural machine translation) models for indic languages *to kick-start the company's localization services with a deal worth 1 million USD.*
 - Created the data-based project completion probability predictor by modeling creator working speeds as a probability distribution and forecasting their available bandwidth using LSTMs. The tool is internally used for deadline suggestions, real-time ETA for clients, and measurement of supply depth. *Reduced deadline breaches by 52.1%.*
 - Reduced the churn rate by 34.8%* by designing and implementing a client sentiment prediction model. Used my business insights to assign importance scores to different stages in a project, added them, and defined cut-offs for sentiment categories on the basis of historical client behavior.
- Oct 2022 - **Researcher**, N-CRiPT, National University of Singapore
- Present
- Working on more efficient and generalizable techniques to assign price tags to data points in a free market.
 - Working on privacy attacks and defenses on models obtained using machine unlearning.
 - Working on machine unlearning in Large Language Models.
- May 2022 - **Researcher**, N-CRiPT, National University of Singapore
- Oct 2022
- Pioneered the first machine unlearning technique specifically tailored for neural networks utilized in regression tasks.
 - Prepared a blindspot model with partial training on the data to retain and used it to transfer wrong knowledge about the data to forget. The original model is used to distill and retain knowledge about the retain dataset.
 - The method is 20x faster than the available baselines.
- Aug 2021 - **Thesis**, School of Computing, National University of Singapore
- May 2022
- Developed UNSIR, a class-level machine unlearning technique which uses loss maximization for the data to forget, and loss minimization on a subset of the rest of the data. *The method is up to 154x faster than existing methods.*
 - Introduced the concept of zero-shot machine unlearning. Also proposed two techniques, one based on conditional zero-shot knowledge distillation, and the other based on model inversion to obtain samples with minimal loss on retain data, and maximum loss on forget data, and subsequently training a new model on these samples.
 - Developed a domain-agnostic machine unlearning method *which is 70x faster than competing methods* and works for all modes of unlearning.
 - Supervisor:** Dr. Mohan Kankanhalli and Dr. Murari Mandal
- June 2021 - **SDE Intern**, Salesforce, India
- July 2021
- Developed a VS Code extension and an extension pack given a list of desired functionality and extensions used extensively by technical writers.
 - Designed and developed [SalesforceDocs Markdown Assistant](#) [Download count: 2300+] and [SFDocs Extension Pack](#) [Download count: 1400+] using VS Code Extension API and TypeScript.
 - Resulted in an increase in writer efficiency by 19.8%.
- Jan 2021 - **Data Science Intern**, LTI (Larsen & Toubro Infotech), India
- Feb 2021
- Upgraded the implementation of the Resolution Notes Classifier to support multiple languages by fine-tuning multilingual-T5 on the available data and *achieved an accuracy of 92.9%.*
 - Added the ability to produce more information about the note if it is classified as a bad note by modeling this as a subsequent multi-class classification problem. *Achieved an accuracy of 73.3%.*
- May 2020 - **Research Intern**, CSIR-CEERI (Central Electrical Engineering Research Institute), India
- June 2020
- Built a solution to classify chest X-Ray images into healthy, pneumonia, and covid-19.
 - Trained segmentation models like UNet to separate out the lungs, and subsequently applied classification models like ResNet-50, VGG-19, and DenseNet-121. Used GradCAM to explain the results.
 - Achieved a classification accuracy of 86.6%.

Publications

May 2023 **Deep Regression Unlearning**, Ayush K Tarun*, Vikram S Chundawat*, Murari Mandal, Mohan Kankanhalli, ICML 2023

- April 2023 **Fast Yet Effective Machine Unlearning**, Ayush K Tarun*, Vikram S Chundawat*, Murari Mandal, Mohan Kankanhalli, *IEEE Transactions on Neural Networks and Learning Systems*
- March 2023 **Zero-Shot Machine Unlearning**, Vikram S Chundawat*, Ayush K Tarun*, Murari Mandal, Mohan Kankanhalli, *IEEE Transactions on Information Forensics and Security*
- Dec 2022 **TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data**, Vikram S Chundawat*, Ayush K Tarun*, Murari Mandal, Mohan Kankanhalli, *IEEE Transactions on Artificial Intelligence*
- Nov 2022 **Can Bad Teaching Induce Forgetting? Unlearning in Deep Networks using an Incompetent Teacher**, Vikram S Chundawat*, Ayush K Tarun*, Murari Mandal, Mohan Kankanhalli, *AAAI-2023*
- *→ equal contribution

Projects

- Aug 2021 - **Nuclei Instance Segmentation**, BITS Pilani, Pilani Campus, India
- Dec 2021
- Built image segmentation pipeline to segment out nuclei from cell patches.
 - Used image sharpening and a novel self-supervised SuperResGAN trained on down-scaled cell patch images with original images as the target. Subsequently, the images are divided into a 2x2 grid, and this SRGAN is used to upscale each patch in this grid.
 - Trained UNet and HRNet on the obtained models and achieved a dice coefficient of 0.80.
- Jan 2021 - **Unsupervised Text Based Fake News Detection**, BITS Pilani, Pilani Campus, India
- May 2021
- Implemented an unsupervised approach to detect fake news.
 - The method involved live scraping of news articles, fact-checking websites, and other relevant websites related to the claims text, and subsequently state-of-the-art NLP models are used for semantic text similarity. A similarity of <0.5 is marked as fake news.
 - Achieved an accuracy of 95.9% on Liar Dataset and 81.6% on Covid-19 Fake News Dataset.
- Jan 2021 - **Augmenting Tabular Data using Deep Learning Techniques**, BITS Pilani, Pilani Campus, India
- May 2021
- Designed new architectures to produce synthetic tabular data.
 - Developed a new tabular GAN architecture by encoding the data using a variational Gaussian mixture model and the Encoder of a Transformer model. Developed another diffusion-based model in which predictors were trained for each feature and starting from random data at each step, these predictors are used to modify the features in a random permutation. *Achieved a similarity score of 68% with real data.*
 - Devised a new evaluation criterion that encapsulates all the desired features, i.e. basic statistical adherence, machine learning utility, rare-class(es) coverage, and distribution comparison for good-quality synthetic data. *Achieved a similarity coefficient of 0.97 vs 0.55 of that of competing methods within real data.*
- Oct 2020 - **Earthquake Nowcasting for Indonesian cities**, BITS Pilani, Pilani Campus, India
- Nov 2020
- Built a solution to analyze the current state of earthquake hazard in the region of Indonesia (-10°-10°N and 90°-110°E) by predicting the probability and magnitude of 'large' (magnitude ≥ 6.5) earthquakes.
 - We fit various statistical distributions on the number of small (magnitude < 6.5) earthquakes occurring between large earthquakes and use the obtained cumulative distribution function for the probability. An LSTM is used to predict the magnitude.
 - Achieved a MAPE of 4.0%.
- Sep 2020 - **Manipulating latent space of GANs to achieve high-resolution zoom synthesis**, BITS Pilani, Pilani Campus, India, Details
- Nov 2020
- Built an approach to achieve high-resolution zoom-in and zoom-out of images.
 - Trained SVMs on zoomed-in and zoomed-out versions of sampled images from BigGAN to obtain the latent direction to zoom-in. The images to zoom-in are then inverted to the GAN's latent space and the latent vector is moved in or against the obtained direction of zooming-in to zoom-in or zoom-out respectively.

Coursework

- Computer Science Data Structures and Algorithms, Object Oriented Programming (OOP), Database Management Systems, Operating Systems, Computer Networks, Probability and Statistics, Differential Equations, Linear Algebra
- Data Science Machine Learning, Foundations of Data Science, Applied Statistical Methods, Optimization

Skills

- Languages *Proficient:* Python, MySQL | *Familiar:* C++, R, JavaScript, TypeScript, C, Java
- Frameworks *Proficient:* PyTorch, TensorFlow | *Familiar:* NLTK, OpenCV, spaCy, Keras, Apache Spark
- Tools AWS, GitHub, Unix

Academic Honors

- Scholarships **Merit Scholarship**
- Received 100% tuition fee waiver for ranking in the top 1% of the entire batch of 1000+ students