Mayank Mishra

Research Engineer
IBM Research

Website: https://mayank31398.github.io/
Email: mayank31398@gmail.com
GitHub: https://github.com/mayank31398
LinkedIn: https://www.linkedin.com/in/mayank31398
Google Scholar: https://scholar.google.com/citations?user=YsbtW6cAAAAJ&hl=en

ACADEMIC DETAILS

- **Bachelor of Technology**, Electrical Engineering, Indian Institute of Technology Delhi (GPA: 8.17/10)
- Secured All India Rank 921 in JEE Advanced 2016
- Secured All India Rank 682 in JEE Mains 2016

WORK EXPERIENCE

IBM Research *Aug 2020 - present*

- Open-sourced code for benchmarking and deploying inference servers for BLOOM-176B to Megatron-DeepSpeed
- Trained a GPT2 model on 16 GPUs using Megatron-DeepSpeed with different parallel configurations on large text corpus
- Published paper titled "Variational Learning for Unsupervised Knowledge Grounded Dialogs" in IJCAI 2022 Main Track
- Worked on COVID-ASSIST (a conversational platform) centered around Watson in response to the COVID pandemic
- Worked as a Watson Assistant developer by contributing on aspects of planning, automation and customer care
- Participated in the DSTC9 challenge organized by Amazon Alexa to support conversations grounded in documents

SAMSUNG R&D Institute, Noida *May 2019 - Jul 2019*

- Implemented a new user authentication system using smartphone sensors for real-time authentication
- Used LSTM based Variational Autoencoders for projecting the obtained time-series on a lower dimensional manifold
- Implemented few-shot learning to learn user profiles with minimal data points in an online learning environment
- Created and deployed an android application on several devices and optimized the battery consumption of the same

PUBLICATIONS

Variational Learning for Unsupervised Knowledge Grounded Dialogs

- Proposed a model to generate responses for dialogs grounded on information present in external knowledge sources
- Retrieved the relevant textual documents from a large indexed collection of documents in an unsupervised fashion
- Used a variational framework to take advantage of the posterior distribution to retrieve better documents during training
- Also showed the efficacy of the proposed model on other tasks like question answering, classification etc
- Published in International Joint Conferences on Artificial Intelligence (IJCAI 2022)

Adversarial Approximate Inference for Speech to Electroglottograph Conversion

- Optimized the Speech to Laryngograph encoder using adversarial training for the network using informative priors
- Created a cosine based loss function for enforcing amplitude invariance between ground truth and network output
- Used a variational inference approach for learning optimal representations for speech signal to infer the EGG signal
- Demonstrated the advantages of using informative priors over Gaussian priors in the variational autoencoder setting
- Utilized continuous wavelet transforms using Ricker wavelets for robust peak picking
- Published in IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)

Variational Inference with Latent Space Quantization for Adversarial Resilience

- Implemented a defense mechanism capitalizing on the expressive power of regularized latent space generative models
- Trained Variational Autoencoders with a K-Lipschitz encoder to ensure closeness of similar images in the latent space
- Proposed a mechanism for defending neural networks against adversarial examples using latent space quantization
- Demonstrated the efficacy of the proposed mechanism against multiple attack types (black and white box) and methods
- Submitted to Association for the Advancement of Artificial Intelligence (AAAI 2020) - Arxiv preprint

PROJECTS

BLOOM-176B Inference Open-Source *IBM Research, Jul 2022 - present*

- Created an easy-to-use framework for deploying BLOOM-176B via a REST API or CLI with for inference
- Experimented with approaches like HuggingFace Accelerate, DeepSpeed-Inference and DeepSpeed ZeRO for inference
- Benchmarked throughput and latency for the 176 Billion parameter model on a single node with 8 A100 80GB GPUs
- Experimented with quantization approaches (LLM.int() and ZeroQuant) for reducing the memory footprint of the model
- Contributed the source code to Megatron-DeepSpeed for benchmarking and serving BLOOM-176B with ease. Also added support for fp16, bf16 and quantized BLOOM model using both LLM.int8() and ZeroQuant quantization approaches

Distributed Pretraining for GPT2 *IBM Research, Apr 2022 - Jun 2022*

- Pretrained a GPT2-like decoder model on 16 A100 80GB GPUs on a large text corpus using Megatron-DeepSpeed
- Experimented with different parallel configurations like Tensor Parallel, Pipeline Parallel & Fully Sharded Data Parallel
- Optimized the model (3.55 B parameters) for high training throughput with the different parallel configurations
- Found the best parallel configuration for taking advantage of both inter-node and intra-node GPU interconnects
- Experimented with both fp32 and fp16 training for better GPU utilization

COVID-ASSIST *IBM Research, Jul 2021 - Sep 2021*

- Worked on authoring a Watson Assistant skill to help out fellow IBMers in India during the COVID pandemic
- Authored the skill such that the users can request for medicines, emergency supplies, vaccination, information about COVID, doctor's appointment etc
- The work involved external collaborations including organizations such as Indian Council of Medical Research (ICMR), Department of Health Research, Ministry of Health and Family Welfare, Government of India

Watson Assistant Dialog Runtime *IBM Research, Feb 2021 - Jul 2021*

- Worked on improving customer experience, fixing customer issues, PII leaks, providing new features for easier authoring of skills, catching unexpected exceptions that sometimes led to the Assistant getting stuck in unforeseen states
- Also worked on upgrading dependencies like Google's gson project and the spring expressions project, to reduce vulnerabilities, which serve as the backbone of the dialog runtime
- Removed the cloned repos for these open-source projects and modified the entire codebase of the dialog runtime to source the jars (of these projects) from org.apache.maven's jar repository rather than building their entire repos from scratch
- Doing so reduced the dialog runtime's build times from 10-11 mins to 4-5 mins and also reduced the code size by 30,000 lines (approx.) making the code easier to understand for new team members. This also makes future upgrades a lot easier

DSTC9 Track 1 Challenge *IBM Research, Aug 2020 - Oct 2020*

- Participated in the DSTC9 challenge organized by Amazon Alexa
- Created models for generating responses to task-oriented dialogs where the required knowledge lies in external documents
- Worked on retrieving the relevant knowledge in both supervised and unsupervised settings

Real-time Visual Respiration Rate Estimation with Dynamic Scene Adaptation *IIT Delhi, Feb 2019 - May 2019*

- Used Computer Vision based techniques for estimating the respiration rate from the video footage of an individual
- Used the proposed algorithm to correctly identify the patients suffering from pneumonia (fast breathing)
- Implemented and optimized the algorithm to run on Raspberry Pi for detection in real-time in hospitals

Resource and profit optimization in electricity market *IIT Delhi, Sep 2019 - Dec 2019*

- Developed new models for evaluating flexible resources in two-settlement electricity markets (day-ahead and real-time)
- Worked on achieving equilibrium in two settlement electricity markets using ADMM

Bias Correction in Deep Neural Networks *IIT Delhi, Aug 2018 - Nov 2018*

- Worked on reducing dataset bias in neural networks for better generalization without training on multiple datasets
- Trained an Auxilliary Classifier GAN (ACGAN) to generate images conditionally given the class from MNIST dataset
- Used the original MNIST images and the conditionally generated images from the ACGAN to train a CNN classifier
- Tested this classifier on a hand-written digits dataset collected in classroom and achieved state of the art performance

Lecture Summarization using Deep Learning *SAMSUNG Research, IIT Delhi, Feb 2019 - May 2019*

- Trained Convolutional LSTMs for summarizing video lectures of various online courses
- Used Computer Vision techniques to find edge maps, optical flows and difference of consecutive frames of the videos
- Used the engineered features for increased accuracy over conventional recurrent networks trained using raw frames
- Implemented a WPF software in C# to summarize the video lectures and generate lecture notes in PDF format

Touch-Point Prediction using Deep Learning *IIT Delhi, May 2018 - Dec 2018*

- Worked on improving touch-screen latency for the SAMSUNG Flip device without explicitely changing the hardware
- Trained and benchmarked Fully Connected, RNNs and LSTMs and analyzed the performance of the said algorithms
- Implemented the said algorithms on the device yielding a low error rate with no significant impact on performance

Braille Tutoring Application *IIT Delhi, Jan 2018 - May 2018*

- Implemented tutorials and games using Python for comprehensive learning of Braille by visually challenged students
- Created a Linux based secondary software for the tutor to add customized exercises or games in the application
- Deployed the application on a Beaglebone-based Refreshable Braille Device
- Worked on providing tactile output, sound and an external Arduino based serial LCD display through the Braille device
- Tested the application with visually challenged students in the National Association for Blind

Identifying the Diabetic Neuropathic Patients using Machine Learning *IIT Delhi, Sep 2017 - Dec 2017*

- Trained bi-directional LSTMs for the identification of Diabetic Neuropathic patients using foot pressure data
- Implemented a WPF software in C# to record data using an Arduino based pressure mat

Crystal Ball Interface to view 3D Objects *IIT Delhi, Aug 2017 - Dec 2017*

- Implemented a crystal ball interface using OpenGL in C++ for viewing 3D objects saved in .obj file format

COURSES UNDERTAKEN

Machine Learning, Advanced Machine Learning, Information Theory, Data Structures, Computer Architecture, Embedded Systems Design Project, Probability and Stochastic Processes, Linear Algebra and Differential Equations, Calculus, Micro Economics, Signals and Systems, Control Engineering, Digital Electronics, Analog Electronics

TECHNICAL SKILLS

- **Programming Languages**: Python, Java, C++, C#
- **Machine Learning Frameworks**: TensorFlow, PyTorch, Keras, sklearn
- **Softwares**: Visual Studio, Android Studio, Eclipse, Vivado, Linux, MATLAB, Simulink, Unreal Engine 4, 3ds Max
- **Interests**: Deep Learning, Quantum Computing, Information Theory