Mayank Mishra

Research Software 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

BigCode *Oct 2022 - present*

- Working on the training and inference team for reducing training time and inference latency for large language models
- Working on implementing Multi-Query Attention and Flash Attention in Megatron-LM for distributed training

BigScience *Jun 2022 - present*

- Contributed to the training codebase for BLOOM-176B allowing the training of first-ever open-access large transformer
- Open-sourced code for optimized inference for large transformers in huggingface/transformers-bloom-inference
- Submitted paper titled 'BLOOM: A 176B-Parameter Open-Access Multilingual Language Model' to JMLR

IBM Research *Aug 2020 - present*

- Published paper titled 'Variational Learning for Unsupervised Knowledge Grounded Dialogs' in IJCAI 2022 Main Track
- Submitted paper titled 'Joint Reasoning on Hybrid-knowledge sources for Task-Oriented Dialog' to EACL 2023
- Created an optimized serving method for serving large transformers for inference to researchers within IBM
- Used prompt tuning for BLOOM-176B in a distributed environment for state-of-the-art performance on various datasets
- Used Megatron-DeepSpeed for distributed training of a GPT-2 model on 16 A100 GPUs on a large text corpus
- Created COVID-ASSIST using Watson Assistant for help with appointments, vaccines etc in response to COVID
- 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
- Used few-shot learning to reason over 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)

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

- Contributed to the training codebase of BLOOM, a 176-Billion parameter multilingual large language model
- Implemented state-of-the-art AliBi positional encodings to attend over long sequences unseen during training
- Also implemented a novel checkpoint reshaping strategy to change the distributed configuration of a large model
- Worked in open collaboration with researchers all over the world and gained experienced on large scale training
- Submitted to Journal of Machine Learning Research (JMLR) - Arxiv

Joint Reasoning on Hybrid-knowledge sources for Task-Oriented Dialog

- Worked on generating agent responses to conversations requiring reasoning over both structured databases and unstructured text documents
- Created a modified version of the MultiWOZ dataset and showed that existing methods failed on the created dataset
- Proposed a baseline model trained with Prompt+LM tuning to retrieve the relevant information (from both structured and unstructured sources) and generate the response
- Submitted to European Chapter for the Association for Computational Linguistics (EACL 2023) - Arxiv

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
- Demonstrated that the proposed approach for training was robust to change in knoledge modality and had comparable performance on modality unseen during training
- Submitted to Association for the Advancement of Artificial Intelligence (AAAI 2020) - Arxiv

PATENTS

EdgeEGG - A system and method for hand-held electrode free elctroglottograph using neural networks on programmable controllers

- Proposed a safe contact-free ElectroGlottoGraph which provides an accurate estimate of EGG signal
- Proposed a cost-effective and efficient mechanism with integrated speech sensors to allow edge computation of EGG
- Designed a resource efficient hardware device optimizing both energy consumption and prediction latency, by performing computations on very low power micro-controllers
- Indian Patent application, No. 201911036593

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 for inference purposes
- 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.int8() and ZeroQuant) to reduce 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
- Official code maintainer for huggingface/transformers-bloom-inference repository

BLOOM-176B prompt tuning *IBM Research, Jul 2022 - present*

- Used prompt tuning to improve the performance of BLOOM-176B on a variety of tasks and datasets
- Used DeepSpeed ZeRO stage 3 for prompt tuning in a distributed environment with 2 nodes with 8 GPUs each
- Achieved state-of-the-art performance by training only 100k ($6 \times 10^{-5}\%$) parameters of the BLOOM-176B model
- Also created a REST API interface for serving prompt-tuned large language models like BLOOM

BLOOM-176B large-scale serving *IBM Research, Jul 2022 - present*

- Created a framework for large-scale models built upon BLOOM-176B open source contribution for over 200 researchers
- Provided a variety of options (just like GPT-3) which included a UI, API calls for generated outputs, embeddings etc.
- Also implemented a novel continuous batching scheme to increase server throughput to serve multiple people concurrently

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 32-bit vs 16-bit for increasing training speed. Also experimented with fp16 and bf16 for stability

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 to allow the users to 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 and 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 Alternating Direction Method of Multipliers

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
- Released the 'Nearly MNIST' dataset for future research

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 Networks, RNNs and LSTMs and analyzed their runtime performance
- 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 device running a Refreshable Braille Display
- Provided tactile output and sound using an external Arduino based device connected to the Refreshable Braille Display
- 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, Probability and Stochastic Processes, Linear Algebra and Differential Equations, Calculus, Data Structures, Computer Architecture, Embedded Systems, Signals and Systems, Control Theory, Digital Electronics, Analog Electronics

TECHNICAL SKILLS

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