Aditya Shah

+1 (540) 824 9021 aditya.shahh3@gmail.com aditya-shahh Google Scholar

Summary

- I am closely working with the leadership and research teams to build large-scale enterprise GenAI solutions.
- My core responsibilities include:
  - Implementing Large Language Model architectures (Llama2, Mixtral, etc) and performing domain adaptive pre-training on multi-GPU sytems using DeepSpeed / Megatron.
  - Synthesizing domain specific enterprise data into required prompt-instruction pairs and performing instruction fine-tuning using DPO / RLHF.
  - Developing a solid understanding of novel architectural innovations (KV cache, flash attention, MQA, etc), optimization techniques (quantization, FSDP, etc), and implementing newer research methods.
- Key Expertise:
  - **Tools/Tech**: Python, PyTorch, DeepSpeed, Megatron-LM, Accelerate, GPU, CUDA, Docker, NumPy, Pandas, SQL.
  - **Deep Learning**: Pretraining & Finetuning Large Language Models (LLMs), RLHF, DPO, FSDP, RAG, Vector DB.

Work Experience

- **Capital One** — Mclean, USA
  Machine Learning Scientist - AI Foundations — *Jun 2023 - Present*
  - Built in-house models with **Llama2** and **Mixtral**, loaded checkpoint weights, and conducted further **pre-training** on enterprise data (casual language modelling) using **Megatron**, **FSDP** on multiple GPUs.
  - Developed prompt-instruction dataset from enterprise data and **fine-tuned** these base models using **DPO/RLHF** to align them for chat-based use cases.
  - Implemented various **optimization techniques** like KV cache, reduced precision, Multi Query Attention, Rotary Position Embeddings, etc to optimize **fine-tuning** and **inference** pipelines.
  - Building domain specific **LLM agents** using **RAG** and **VectorDB** to provide AI based virtual assistance with different financial legalities and preventing risks.
  - Delivered various **keynotes** and **training sessions** on Generative AI, and NLP, highlighting personal expertise and **leadership** in **upskilling teams**.

- **Google** — Seattle, USA
  Research Scientist Intern - LLMs — *Sep 2022 - Dec 2022*
  - Worked with DeepMind to integrate **soft prompt parameters** and **adapters** in a Multimodal Large Language Model (MLLM) for Document Extraction.
  - Developed an **efficient optimization** pipeline and performed **parameter-efficient prefix fine-tuning** on **TPUs** to extract data from invoice documents.
  - Enhanced model's adaptability and robustness in sequential uptraining, which **reduced catastrophic forgetting** by **14%.**

- **Capital One** — Mclean, USA
  Data Science Intern - NLP — *Jun 2022 - Aug 2022*
  - Fine-tuned transformer-based language models (RoBERTa, XLNet, T5) on enterprise wide call transcript data to extract relevant knowledge, identify entities and summarize the transcript.
  - Improved customer request fulfillment and agent performance through **co-reference resolution** and eliminated **70% of false positives** with **94% accuracy.**

- **Indian Institute of Technology (IIT)** — Indore, India
  Research Scientist - Machine Learning — *Sep 2020 - Aug 2021*
  - Developed a novel multimodal neural network architecture for sarcasm detection which **outperformed** existing benchmarks by **6.14% F1** score. *Research Paper accepted in ICONIP 2021*
  - Proposed an efficient self-attention based model to capture incongruity for code-mixed sarcasm detection. Achieved competitive F1 score as compared to multilingual models while **training 10x faster** and using **lower memory footprint**. *Research Paper accepted in ICON — (ACL 2021)*

- **Saarthi.ai** — Bangalore, India
  Machine Learning Engineer — *Jul 2020 - Oct 2020*
  - Conducted applied research on ASR and developed a deep learning model based on BiLSTM and 1-D CNN for gender identification from audio data.
  - Achieved a **test accuracy of 96%** with **15% improvement** over previously designed approach. Further, worked on age identification and specific keyword detection from real-time audio input.

Selected Publications

- **A. Shah**, A. Jain, S. Thapa, and L. Huang, "ADEPT: Adapter-based Efficient Prompt Tuning Approach for Language Models", *The 61st Annual Meeting of the Association for Computational Linguistics* (ACL), 2023 *Paper*
- B. Yao*, **A. Shah***, L. Sun, and L. Huang, "End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models", *International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), 2023 *Paper* (*Best Paper Honorable Mention*)
- S. Thapa*, **A. Shah***, F. Jafri, U. Naseem, and I. Razzak, "A Multi-Modal Dataset for Hate Speech Detection on Social Media:Case-study of Russia-Ukraine Conflict", *Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2022. *Paper*
- **A. Shah** and C. Maurya, "How effective is incongruity? Implications for code-mixed sarcasm detection", *Proceedings of the* 18^{th} *International Conference on Natural Language Processing* — (ACL), 2021. *Code Paper*
- S. Gupta, **A. Shah**, M. Shah, L. Syiemlieh, and C. Maurya, "FiLMing Multimodal Sarcasm Detection with Attention", *Proceedings of the* 28^{th} *International Conference on Neural Information Processing* (ICONIP), 2021. *Code Paper*
- L. Kurup, M. Narvekar, R. Sarvaiya, and **A. Shah**, "Evolution of Neural Text Generation: A Comparative Analysis", *Advances in Intelligent Systems and Computing*, Springer (IC4S), 2020. *Paper*

Skills Summary

- **Libraries and Technologies**: PyTorch, DeepSpeed, Megatron-LM, Accelerate, GPU, CUDA, NumPy, Pandas, SpaCy
- **Languages & Frameworks**: Python, C++, SQL, MongoDB, Flask, Docker, Kubernetes, Spark

Education

- **Virginia Tech** — Blacksburg, USA
  Masters of Science in Computer Science - Research — *2021 - 2023*
  Thesis: NLP based Episodic Future Thinking (EFT). (Funded by NIH)
- **Dwarkadas J. Sanghvi College of Engineering** — Mumbai, India
  Bachelors of Science in Computer Science — *2016 - 2020*

Academic Projects

- **Code Interpretability on transformer models using SHAP**: Conducted an exclusive research study on analysing code interpretability using SHAP values and Logit manipulation for Codebert and Graph Codebert models. Code
- **Adaptive pooling based Electra model for Multi Label Relation Classification**: Proposed an Adaptive pooling based method on top of Electra model — *AdaElectra* for multilabel relation classification achieving F1 score of 0.88 on the NYT29 dataset. Code
- **Weighted Contextual N-gram method for evaluation of Text Summarization**: Finetuned T5 model on Extreme Summarization (XSum) Dataset and proposed the use of *Weighted Contextual N-gram (WCN)* method – an alternative metric for evaluation of text generation. Code
- **Supervised Text Generation using GPT2 model, BiLSTM, and GloVe Embedding**: Fined tuned GPT2 model on wikisent data for generating context-dependent text samples. Developed a BiLSTM with GloVe embedding and N-gram model to generate text with 90% test accuracy. Code
- **Food-101 Challenge by ETH Zurich**: Designed a Neural Network model on top of the Xception network and fine-tuned it to achieve State-of-the-Art result on the challenging Food 101 Dataset with a test accuracy of 87%. Code

Honors and Awards

- Received "Best Paper Honorable Mention" for the work on Multimodal Fact Checking at 'SIGIR', 2023.
- Served as a Reviewer for: NAACL SRW 2023, ICON 2023, EMNLP 2022, COLING 2022, ICON - ACL 2021.
- Selected for AI fellowship program, Fellowship.ai, May 2020.
- Awarded "Best Research Project" at HaXplore, IIT BHU Machine Learning hackathon, 2019.
- Received "Innovative Research Project" award in 'CodeShastra Intercollege Hackathon', 2019.
- Served as "Co-Technical Head" for ACM, 2017-18. Mentored a team of 10 students for Software Development and ML.
- Awarded "Google India Scholarship", 2017. in Android Application Development.