Sabbir Ahmed

Website: sabbiracoustic1006.github.io
Email: sahmed9@binghamton.edu
LinkedIn | GitHub | Google Scholar
Phone: 9518239360

Education

State University of New York at Binghamton — Binghamton, New york
PhD in Computer Science (GPA: 4.0/4.0) — Jan 2023–Present

University of California Riverside — Riverside, California
MS in Electrical Engineering (GPA: 3.96/4.0) — Sep 2021–Dec 2022

Bangladesh University of Engineering and Technology — Dhaka, Bangladesh
B.Sc. in Electrical and Electronic Engineering (GPA: 3.51/4.0) — Feb 2015–Apr 2019

Work Experience

SONY AI — Binghamton, New York (Remote)
Research Intern, Efficient Vision Transformers — Aug 2024–Current

SUNY Binghamton — Binghamton, New York
Graduate Research Assistant, ML Security Research Lab — Jan 2023–Current

REVE Systems — Dhaka, Bangladesh
Machine Learning Engineer — Feb 2020–Aug 2021

Research Experience

Developing Novel Compression Techniques for Memory Efficiency in LLMs — Jan 2025–Current
PhD Research

- Developing novel weight compression to improve throughput in memory-bound LLM decoding.
- Improving memory efficiency on top of Post-Training Quantization (e.g., GPTQ). Currently achieved ∼25% memory reduction on top of GPTQ in OPT models with a slight decrease in perplexity.

Developed Novel Attention to improve Efficiency of ViTs — Aug 2024–Nov 2024
Internship work at SONY AI

- Developed MixA, a novel attention mechanism that improves efficiency of ViTs.
- Maintains performance comparable to softmax attention while improving inference speed by 15-30% at edge.

Developed Novel Compression Method for ViT — Jan 2024–Present
PhD Research

- Developed DeepCompress-ViT, an encoder-decoder based weight compression strategy for 15-20× ViT compression with high performance.
- Introduced Optimized-Test Time Decoding to mitigate weight decoding overhead.
- Achieved up to ∼1500× energy reduction and ∼70× latency reduction on edge platform.

Improving Safety of Source-Free Domain Adaptation — Jan 2023–Current
PhD Research

- Developed SSDA, the first secure SFDA framework against backdoor attacks.
- Proposed SSDA can successfully defend attack ($< 5\%$ ASR) without degrading SFDA performance.

Projects

- **Participated in The eBay 2024 University ML Competition to build a model that extracts vehicle parts compatibility (fitment) data from eBay listings.**
  - Converted the raw HTML data to text and denoised using Qwen2.5-32B-Instruct to remove irrelevant information.
  - Used two open-source LLMs (Mistral-Nemo, Qwen2.5-14B) and fine-tuned them using k-fold LoRA on labeled training data to handle the fitment prediction task.
  - Ensembled predictions using majority voting strategy to generate the final fitment prediction.
  - Achieved 79.6% F0.2 score, securing second runner-up position in the challenge.
- **Participated in the eBay 2023 University ML Competition to build a model that extracts and labels named entities from eBay item titles.**
  - Implemented a teacher-student training framework with XLM-ROBERTA and ELECTRA models for robust NER by generating pseudo labels on unlabeled data using teacher model and filtering out uncertain samples to train student model.
  - Used k-fold training and ensemble strategies, combining predictions from both XLM-ROBERTA and ELECTRA models to improve performance.
  - Achieved an F1 score of 94.37%, securing second runner-up position in the challenge.

Publications

1. **Sabbir Ahmed**, Jingtao Li, Weiming Zhuang, Chen Chen, Lingjuan Lyu, "MixA: A Mixed Attention approach with Stable Lightweight Linear Attention to enhance Efficiency of Vision Transformers at the Edge" (submitted to ICCV).
2. **Sabbir Ahmed**, Abdullah Al Arafat, Deniz Najafi, Akhlak Mahmood, Mamshad Nayeem Rizve, Mohaiminul Al Nahian, Ranyang Zhou, Shaahin Angizi, Adnan Siraj Rakin, "DeepCompress-ViT: Rethinking Model Compression to Enhance Efficiency of Vision Transformers at the Edge" (accepted at CVPR 2025).
3. **Sabbir Ahmed**, Ranyang Zhou, Shaahin Angizi, Adnan Siraj Rakin, "Deep-TROJ: An Inference Stage Trojan Insertion Algorithm through Efficient Weight Replacement Attack" (CVPR 2024).
4. **Sabbir Ahmed**, Abdullah Al Arafat, Mamshad Nayeem Rizvee, Rahim Hossain, Zishan Guo, Adnan Siraj Rakin, "SSDA: Secure Source-Free Domain Adaptation", 2023 International Conference of Computer Vision (ICCV 2023).
5. **Sabbir Ahmed**, Uday Kamal, Md. Kamrul Hasan, "DFR-TSD: A Deep Learning Based Framework for Robust Traffic Sign Detection Under Challenging Weather Conditions", IEEE Transactions on Intelligent Transportation Systems.

Skills

- **Programming Languages:** Python, MATLAB, C, C++, Intel-8086 Assembly
- **Machine Learning Libraries:** PyTorch, vllm, triton, mmdetection, Scikit-Learn

Awards and Honors

- **Clog Loss: Advance Alzheimer's Research with Stall Catchers**, Team leader of team *"acoustic_user"* that won 6th place among 922 teams from the whole world.
- **Bengali Handwritten Digit Recognition Competition**, Won $5th$ position among 57 teams from the whole country.
- **Kaggle APTOS 2019 Blindness Detection**, Team leader of team *"cholo model re shikhai"* that won 38th place among $2,943$ teams from the whole world.
- **Kaggle Human Protein Atlas Image Classification**, Member of team *"The Unseens"* that won $98th$ place among $2,169$ teams from the whole world.
- **IEEE Signal Processing Cup 2019**, Member of team *"Maverick"* that won $6th$ place among 24 teams from the whole world.
- **Served as Reviewer**, at CVPR 2025 and IEEE Transactions on Intelligent Transportation Systems.