

EDUCATION

State University of New York at Binghamton PhD in Computer Science (GPA: 4.0/4.0)	Binghamton, New York Jan 2023–Present
University of California Riverside MS in Electrical Engineering (GPA: 3.96/4.0)	Riverside, California Sep 2021–Dec 2022
Bangladesh University of Engineering and Technology B.Sc. in Electrical and Electronic Engineering (GPA: 3.51/4.0)	Dhaka, Bangladesh Feb 2015–Apr 2019

WORK EXPERIENCE

SONY AI Research Intern, Efficient Vision Transformers	Binghamton, New York (Remote) Aug 2024–Nov 2024
SUNY Binghamton Graduate Research Assistant, ML Security Research Lab	Binghamton, New York Jan 2023–Current
REVE Systems Machine Learning Engineer	Dhaka, Bangladesh Feb 2020–Aug 2021

RESEARCH EXPERIENCE

Developing Novel Compression Techniques for Memory Efficiency in LLMs PhD Research	Jan 2025–Current
<ul style="list-style-type: none">– 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 Internship work at SONY AI	Aug 2024–Nov 2024
<ul style="list-style-type: none">– 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 PhD Research	Jan 2024–Nov 2024
<ul style="list-style-type: none">– 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.	
Improved Safety of Source-Free Domain Adaptation PhD Research	Jan 2023–May 2023
<ul style="list-style-type: none">– Developed SSSDA, the first secure SFDA framework against backdoor attacks.– Proposed SSSDA 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.