REN YUXIN

☎ (1) 520 230 6130 | ✉ yuxinr@arizona.edu

EDUCATION

University of Arizona | Electrical and Computer Engineering 08/2024-Present
Ph.D. in Progress

EIT Digital Master School | Autonomous System
- **France – EURECOM |** Sensing, Communicating and Big Data 09/2021-01/2023
- **Sweden - KTH Royal Institute of Technology |** School of EECS 09/2020-05/2023

China - Xi'an Jiaotong University
- **Bachelor of Engineering in Automation |** School of ECE 08/2016-07/2020
- **Honors Youth Program |** Qian Xuesen College **Outstanding Graduate** 08/2014-07/2016
  Aim to cultivate multifaceted qualities of those most talented middle school students

RESEARCH EXPERIENCES

Function-preserving Attention Replacement | ***University of Arizona, Tucson*** 08/2024- Present
- Proposed and developed distillation methods to replace attention modules in Transformers with simpler alternative recurrent or structured architectures (e.g., Bi-LSTM, Mamba), aiming to reduce inference cost while preserving functionality.
- Designed and conducted comprehensive experiments on token compression and alternative sequence modeling to rigorously test the core hypothesis: attention-based pretraining decompose complicated knowledge into simpler mappings which can be learned by efficient substitutes in inference stage.
- Implemented block-wise distillation, pruning, and efficiency evaluation across multiple architectures, demonstrating the feasibility of efficient inference under constrained hardware settings.
- Published preliminary results as a preprint: Is Attention Required for Transformer Inference? Explore Function-preserving Attention Replacement (arXiv:2505.21535).

Scalable Network-Aware Recommendation System | ***EURECOM*** 10/2021- 01/2022
- Built a deep reinforcement learning algorithm using Markov Decision Process and aimed on designing and improving a caching-friendly user recommendation algorithm to save both network cost and user experience.
- Compared the proposed SARSA Learned the mainstream network recommendation algorithms and understood the meaning and usage of sequential neural networks.

A New Method of Convolutional Neural Network Pruning | ***Bachelor Graduation Design, Xi'an Jiaotong University*** 12/2019-06/2020
- Transformed CNNs Pruning problem into a nonlinear constraint problem by combining Gradual Global Pruning with ThiNet structure and completed a filter level convolutional layers pruning method by evaluating similarity among input channels of convolutional kernels.
- Tested the proposed pruning method on VGG-16 and ResNet structure to optimize the computational amount based on Caffe and CIFAR-10.
- Implemented several widely used CNNs pruning methods of different granularity, tested and compared their performance on existing CNN structures, and wrote a literature review.

PUBLICATIONS

Preprints

1. **Y. Ren**, M. D. Collins, M. Hu, and H. Yang. *Is Attention Required for Transformer Inference? Explore Function-preserving Attention Replacement*. arXiv preprint arXiv:2505.21535, 2025.
2. D. Wang, Z. Liu, S. Wang, **Y. Ren**, J. Deng, J. Hu, T. Chen, and H. Yang. *FIER: Fine-Grained and Efficient KV Cache Retrieval for Long-context LLM Inference*. arXiv preprint arXiv:2508.08256, 2025.

WORKING EXPERIENCES

AI Intern: Efficient Deployment of Function-preserving Attention Replacement (FAR) Models on IMC Hardware | *TetraMem Inc., Fremont, CA* 05/2025-08/2025

- Designed and implemented the end-to-end deployment pipeline for FAR models, converting training-oriented architectures (e.g., BiLSTM) into optimized ONNX subgraphs and preparing them for quantization and inference on TetraMem's MX100 in-memory computing hardware.
- Developed static quantization and graph decomposition strategies to enable efficient execution of complex recurrent structures on resource-constrained hardware accelerators.
- Contributed to extending the FAR framework beyond LSTM, including experiments with alternative structures (e.g., Mamba) and token compression methods, validating broader applicability of the approach.

Algorithm Intern: Traffic Flow Simulation Algorithm Developing for OASIS Autonomous Driving Simulation Platform | *Synkrotron Ltd., China* 05/2023-12/2023

- Designed and implemented a method of generating initial traffic flow state in simulation environment, which provides authentic and diverse micro perspective traffic status covering multiple scenarios by applying optional statistical algorithm or conditional GAN-based method.
- Participated in implementing the company's self-developed OASIS traffic flow simulation control model, in which conducted improvements to Behavior Cloning and GAIL controlling algorithms on simulation platforms.
- Performed data cleaning and feature extraction on vehicle trajectory datasets using Python and simulation APIs; conducted road map editing and acquisition and statistics of simulated traffic flow information based on SUMO simulation platform.

Graduation Intern: Deployment and Improving Strategies on a CNN-Based CIR Fingerprinting Method for UMB Indoor Localization | *Siemens AG, Munich Germany* 07/2022-12/2022

- Implemented an indoor localization system which use UMB channel impulse response as a fingerprint to give locating predictions by deep CNNs, deployed and tested in real official and industrial scenarios based on Pytorch, and achieved a leading accuracy among similar methods.
- Applied transfer learning methods into the built indoor localization architecture and proposed a universal pre-trained model on large-scale dataset which successfully increased locating accuracy and training stability on small datasets.
- Utilized data augmentation method to further improve the performance of deployed system on damaged and uneven datasets. Combined the CNN model with WGAN into a semi-supervised learning model by generating realistic fake data and reach an up to 25% positioning improvement.

ACTIVITIES

Vice Captain of School Debate Team 05/2018-05/2020

- Took charge of making plans and leading the implementation of 28-weeks daily training and 1-month high-intensity training plan of World Championship, did lectures in logic and debate skills training.
- Won the Champion of the 1st Silk Road International Film Festival Debate (1/24), and Runner-up of the 12th China Cup (2/24), making the team move up to 9th place in world ranking.
- As the organizer of New Urbanization Debate Competition, planned and organized the competition, docked with local government and school leaders, coordinated 100+ staff, and completed the reception of 24 teams from all over the country and material deployment of 31 debate competitions.

OTHER INFORMATION

Skills Python, C++, MATLAB

Language IELTS 8.0, GRE: 321/340