Chengze Li

EDUCATION

Columbia University - *Fu Foundation School of Engineering* New York, NY
- *Master of Science, Applied Mathematics, GPA: 3.9162/4.0* 09/2023-12/2024

University of North Carolina Wilmington - *College of Arts and Sciences* Wilmington, NC
- *Bachelor of Science, Mathematics, GPA: 3.867/4.0, Dean's List* 08/2022-05/2023

Chongqing University of Arts and Sciences - *College of Mathematics and Artificial Intelligence* China
- *Bachelor of Science, Mathematics and Applied Mathematics, Ranked 1/80* 09/2019-07/2023

PUBLICATIONS

- Zhenjiang Li, Wei Zhang, Baosheng Li, Jian Zhu, Yinglin Peng, **Chengze Li**, Jennifer Zhu, Qichao zhou, Yong Yin, "Patient-specific Daily Updated Deep Learning Auto-Segmentation for MRI-Guided Adaptive Radiotherapy", Radiotherapy and Oncology, Dec. 2022, Vol.177, P222-230
- Ying Li, Xiaoyan Yin, **Chengze Li**, Zhenjiang Li, Zixuan Leng, Qichao zhou, Zirong Li, "Automated Framework for Optimized Intensity-Modulated Radiation Therapy Planning", Physics in Medicine and Biology. (In Progress)
- Tianyang Xu, Haojie Zheng, **Chengze Li,** Haoxiang Chen, Yixin Liu, Lichao Sun, "NodeRAG: Structuring Graph-based RAG with Heterogeneous Nodes", Conference on Empirical Methods in Natural Language Processing. (In Progress)

ACADEMIC EXPERIENCE

Columbia University New York, NY
BERT-based Semantic Role Labeling System using OntoNotes 5.0 Data 12/2024
- Preprocessed PropBank-style SRL data from OntoNotes 5.0, handling 4-line annotations per predicate and tokenized each sentence with a BERT tokenizer to align B-I-O labels. Implemented a custom Dataset class to manage [CLS], [SEP], and [PAD]. Managed unseen or rare roles by mapping them to 'O,' ensuring stable label distributions and reliable training input.
- Fine-tuned a transformer-based BERT-base-uncased model by integrating a linear classification layer for precise argument identification. Leveraged segment embeddings to emphasize predicate locations and utilized CrossEntropyLoss with ignore_index=-100. Adjusted learning rates and iterated over multiple epochs until achieving 96% token-level accuracy, refining model representations and convergence quality.
- Conducted evaluations on a development dataset using span-based metrics, calculating micro-averaged precision, recall, and F1 scores. Attained an F1 score of 0.82, outperforming baseline methods. Implemented an advanced decoding strategy to maintain coherent argument spans, enabling effective benchmarking against prior state-of-the-art models and showcasing enhanced model performance.

Columbia University New York, NY
Neural Networks as Universal Function Approximators in Image Processing 05/2024
- Engineered a neural network featuring 10 input nodes for extended spatial and temporal coordinates, 6 hidden layers, and 9 output neurons for multi-channel image reconstruction. Integrated positional encoding via Fourier Features to bolster the network's capacity to capture high-frequency image details.
- Developed custom forward and backward propagation algorithms, deriving and vectorizing analytical gradients for each layer. Employed the Adam optimizer with learning rate scheduling and gradient clipping to promote convergence and avert overfitting. Incorporated batch normalization after each hidden layer to expedite training and augment model stability.
- Refined the network architecture by integrating residual connections and fine-tuning hyperparameters through Bayesian optimization. Measured performance using Mean Squared Error and Structural Similarity Index Measure, and visualized reconstructed images by reshaping network outputs to their original dimensions, ensuring accurate and high-quality function approximation.

RESEARCH EXPERIENCE

Manteia Data Technology Co., Ltd. Milwaukee, WI
Research Scientist 01/2025-06/2025

- Engineered a multi-objective optimization framework integrating dose prediction and personalized evaluation. Established patient-specific dose constraints by leveraging GPU-accelerated single-objective optimizations for each organ at risk, enabling precise determination of optimal dose distributions.
- Implemented a customized nnU-Net architecture for dose prediction, enhancing feature extraction through advanced concatenation of non-overlapping Regions of Interest assigned numerical values. Employed a quantile loss function to predict dose intervals at different quantile values, facilitating flexible and precise radiotherapy planning.
- Automated treatment plan generation and evaluation for 45 patients using stochastic optimization algorithms to balance strict and lenient OAR constraints. Achieved superior plan quality in 73% of cases with an average dose difference of 1.18% in remaining patients.

Lamont-Doherty Earth Observatory New York, NY
Advanced Methodologies for Enhanced Tropical Cyclone 08/2024-12/2024

- Customized the TempestExtremes by setting shell script criteria (e.g., Great Circle Distance radius, Max offset for geopotential height) to filter variables across ERA5, MERRA and IBTrACS datasets from NASA. Performed reanalysis on variables of the same type, ensuring accurate cross-dataset comparisons.
- Enhanced NeuralGCM pretrained model by constructing fully connected neural networks with skip connections to process atmospheric inputs like total incident solar radiation and sea surface temperature. Implemented dropout regularization and batch normalization to improve generalization and training stability, achieving a 7.9% increase in prediction accuracy over traditional models.
- Integrated Finite Difference Methods with Physics-Informed Neural Networks to solve atmospheric flow equations under hydrostatic balance and shallow atmosphere assumptions, providing strong theoretical support from an inverse problem perspective for regions with sparse storm data. Embedding these solutions within PINNs led to a 6% increase in forecasting reliability for tropical storms.

The University of Virginia & Microsoft New York, NY
Application and Development of the Large Language Model Framework SGLang 07/2024-09/2024

- Executed the complete setup and deployment of the SGLang framework, systematically configuring its runtime across multiple environments to ensure optimal performance and compatibility. Conducted extensive testing to resolve dependencies, including handling FlashInfer CUDA kernels and managing GPU resource allocation for tensor parallelism.
- Implement the LRU caching algorithm through the eviction strategy in the cache module of SGLang, by setting a comparison condition between the future time and the current time, and reading the eviction log to remove contents that will no longer be used within a given time in the future.
- Optimize the new strategy, Lookahead, compared to SGLang by introducing a linear combination of arrival rate, prompt length, and generation length in the original code to enhance the LRU performance. For the first test of LRU, record all access visits to get the future information for Lookahead. Then, use this information for the Lookahead as the second test. Based on the test results, explore further improvements to the LRU algorithm and identify any shortcomings in the process.

CONTESTS AND AWARDS

ICM/MCM Math Modeling Competition 02/2022
Asia and Pacific Mathematical Contest in Modeling (APMCM) 11/2021
Contemporary Undergraduate Mathematical Contest in Modeling (CUMCM) 09/2021
The 13th China National Undergraduate Mathematics Competition (Third Prize) 11/2021
The 12th China National Undergraduate Mathematics Competition (Second Prize) 11/2020
National Scholarship 10/2021
Merit Student of Chongqing City 05/2022
Distinguished Graduate of Chongqing City 06/2023
Outstanding Undergraduate Thesis of Chongqing City 05/2023