Boyao Wang

Champaign, Illinois | Email: boyaow2@illinois.edu | Tel: +1 2179740197

beryex.github.io | github.com/Beryex

EDUCATION BACKGROUND

University of Illinois Urbana-Champaign (UIUC) **Urbana, US**

Bachelor of Science in **Computer Engineering**, GPA: **4.00/4.00**, Transcript *Sep 2021 - Jun 2025*

- Core Courses: Artificial Intelligence (A+), Machine Learning (A+), Computer Systems & Programming (A), Computer Systems Engineering (A), Intro to Algorithms & Models of Computation (A+), Data Structures (A+), Game Development (A+), Analog Signal Processing (A+), Digital Signal Processing (A+), Probability with Engineering Applications (A+)
- Awards: Dean's List (Fall 2023 & Spring 2024)

Zhejiang University (ZJU) **Hangzhou, China**

Bachelor of Engineering in **Electronic and Computer Engineering**, GPA: **3.98/4.00**, Transcript *Sep 2021 - Jun 2025*

- Core Courses: Discrete Mathematics (A+), Linear Algebra (A), Calculus (A), Differential Equations (A)
- Awards: ZJU-UIUC Institute First-Class Academic Excellence Award (Oct 2023), ZJU First-Class Scholarship (Nov 2024)

Stanford University **Stanford, US**

Undergraduate **Summer Visitor**, GPA: **4.075/4.30**, Transcript *Jun 2024 - Aug 2024*

- Core Courses: Data Mining & Analysis (A), Convex Optimization (A)

RESEARCH EXPERIENCE & PUBLICATIONS

Adapt-Pruner: Adaptive and Structured Pruning for Efficient LLMs (Co-First Author) *Jun 2024 - Current*

UIUC | Supervisor: **Prof. Tong Zhang** | In Preparation for ICML 2025 **Preprint**

- Proposed Adapt-Pruner, a method for structured pruning of LLMs that adaptively evaluates the importance of each decoder layer and assigns a corresponding sparsity level by measuring the distance between each layer's input and output tensors
- Evaluated performance on LLaMA-3.1-8B, Qwen2.5-7B, and Gemma-2-9B, demonstrated 5% improvement in common sense benchmark scores over state-of-the-art methods including SliceGPT at 50% sparsity
- Compressed LLaMA-3.2-3B to 1.3B and post-trained on 0.05B tokens, outperforming TinyLlama-1.1 pretrained on 10B tokens, demonstrating structured pruning efficiently generates smaller models with far less computational resources

RL-Pruner: Structured Pruning Using Reinforcement Learning for CNNs (First Author) *Jan 2024 - Aug 2024*

UIUC | Supervisor: **Prof. Volodymyr Kindratenko** | Submitted to CVPR 2025 **Preprint** | **Source Code**

- Developed an end-to-end approach to compress CNNs using structured pruning and Q-learning, with the accuracy of compressed models as the reward, which learned the optimal layer-wise pruning distribution to minimize performance loss
- Applied response-based Knowledge Distillation to post-train the compressed models, using the original uncompressed model as the teacher to transfer learned representations and efficiently recover accuracy lost during compression
- Independently implemented the entire framework in about 2,000 lines of PyTorch code and experimented on ResNet, GoogLeNet and MobileNet, achieving 81% parameter reduction on VGG-19 (CIFAR-100) within 1% performance drop

TEACHING EXPERIENCE

Teaching Assistant, ECE 374: Algs & Models (ZJU) | Instructor: **Prof. Pavel Loskot** *Sep 2024 - Current*

- Hosted a weekly 2-hour lab session for 20 students to review and reinforce class content

Course Assistant, CS 415: Game Development (UIUC) | Instructor: **Prof. Eric Shaffer** *Jan 2024 - May 2024*

- Held office hours and mentored two project teams of 4 students each through their final projects

PROJECTS

ECE 391 POSIX-compliant Unix-like Operating System **Source Code**

- Implemented a POSIX-compliant Unix-like operating system, featuring a terminal driver, real-time clock driver, and system calls, along with advanced capabilities like signal handling and dynamic memory allocation

CS 415 Game Development: The Final Boss **Source Code**

- Implemented an advanced action system for the main character featuring dynamic combo mechanics and environmental interactions, plus a context-aware NPC dialogue system that evolves based on game progression and player choices

STATS 202 URL Relevance Prediction **Source Code**

- Applied comprehensive feature engineering for data preprocessing, outlier removal, and valuable feature extraction, then leveraged various classification algorithms, including boosting techniques, to optimize URL relevance prediction

RESEARCH INTERESTS

My research focuses on developing efficient neural network **compression** and **distillation** techniques for **CNNs** and **LLMs**. I aim to significantly reduce model size and inference latency while preserving model capabilities, enabling secure deployment on edge devices and local hardware for privacy-preserving, application-specific optimizations.

TECHNICAL SKILLS

Programming Languages & Tools: Python, C/C++, PyTorch, CUDA, SLURM, x86 Assembly, Linux, Git, MATLAB