JUNYANG CAI

272 · 788 · 0465 ◇ caijunya@usc.edu
https://caidog1129.github.io/

EDUCATION

University of Southern California — Los Angeles, CA
Ph.D. in Computer Science — *Aug 2023 - Present*
- Advisor: Bistra Dilkina
- Research direction: Mixed Integer Programming, Learning + Optimization

Bucknell University — Lewisburg, PA
B.S. in Computer Science & Engineering; B.A. in Mathematics — *Aug 2019 - May 2023*
- GPA: 3.98 / 4.0, Summa Cum Laude

WORK EXPERIENCE

Fidelity Investments — Boston, MA
AI Center of Excellence — *June 2025 - Aug 2025*
- Advisor: Serdar Kadioglu
- Project: Agentic Framework of LLM modeling Constraint Optimization Problem

Fidelity Investments — Boston, MA
AI Center of Excellence — *June 2024 - Aug 2024*
- Advisor: Serdar Kadioglu
- Project: BALANS: Multi-Armed Bandits-based Adaptive Large Neighborhood Search for Mixed-Integer Programming Problems

PUBLICATIONS

1. **Junyang Cai**, Serdar Kadioglu, Bistra Dilkina. Global Constraint LLM Agents for Text-to-Model Translation. *Under Review*
2. Alican Yilmaz, **Junyang Cai**, Serdar Kadioglu, Bistra Dilkina. ParBalans: Parallel Multi-Armed Bandits-based Adaptive Large Neighborhood Search. *Under Review*
3. **Junyang Cai**, Weimin Huang, Jyotirmoy V. Deshmukh, Lars Lindemann, Bistra Dilkina. Neuro-Symbolic Acceleration of MILP Motion Planning with Temporal Logic and Chance Constraints. *Under Review*
4. **Junyang Cai**, Taoan Huang, Bistra Dilkina. Multi-task Representation Learning for Mixed Integer Linear Programming. *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research (CPAIOR)* 2025.
5. **Junyang Cai**, Serdar Kadioglu, Bistra Dilkina. BALANS: Multi-Armed Bandits-based Adaptive Large Neighborhood Search for Mixed-Integer Programming Problems. *International Joint Conference on Artificial Intelligence (IJCAI)* 2025.
6. **Junyang Cai**, Taoan Huang, Bistra Dilkina. Learning Backdoors for Mixed Integer Linear Programs with Contrastive Learning. *European Conference on Artificial Intelligence (ECAI)* 2024.
7. Jiatai Tong, **Junyang Cai**, Thiago Serra. Optimization Over Trained Neural Networks: Taking a Relaxing Walk. *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research (CPAIOR)* 2024.
8. **Junyang Cai**, Christopher M. Haggerty, Joshua V. Stough. Co-Unet-GAN: a Co-Learning Domain Adaptation Model on Echocardiography Segmentation. *Proceedings Volume 12464, Medical Imaging 2023: Image Processing; 305-311.*

9. **Junyang Cai***, Nguyen Nguyen*, Nishant Shrestha, Aidan Good, Ruisen Tu, Xin Yu, Thiago Serra. Getting away with more network pruning: From sparsity to geometry and linear regions. *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research (CPAIOR)* 2023.

RESEARCH EXPERIENCE

Agentic LLM Modeling for Constraint Optimization Problems *June 2025 – Present*
Supervised by Serdar Kadioglu, Fidelity Investments

- Proposed an agentic LLM pipeline for text-to-MiniZinc model translation, where specialized LLM agents detect and generate global constraints which are then integrated by an assembler agent.
- Developed the Global Constraint Agents framework, improving executability and solve rates over baseline prompting and chain-of-thought methods.
- Current research focuses on constraint-type fine-tuning, prompt optimization, and scaling evaluations with stronger open- and closed-weight LLMs.

Large Neighborhood Search for Mixed-Integer Programming *June 2024 – Aug 2025*
Supervised by Serdar Kadioglu, Fidelity Investments

- Designed BALANS, a multi-armed bandits–based adaptive large neighborhood search framework for MIP, enabling automated selection of heuristics during search.
- Extended to ParBALANS, a parallelized variant exploring multiple adaptive configurations concurrently for improved scalability.
- Released BALANS as an open-source meta-solver applicable across diverse MIP problems.

Machine Learning for Improving Mixed-Integer Programming Solvers *Aug 2023 – Present*
Supervised by Bistra Dilkina, University of Southern California

- Developed graph attention network (GAT) and contrastive learning–based methods to guide key decisions in Branch-and-Bound search, achieving over 15-20% performance gains compared to Gurobi.
- Proposed multi-task representation learning to jointly optimize solver policies and improve generalization across problem classes.
- Currently exploring chaining of solver decisions through multitask learning for further solver efficiency improvements.

Pruning Neural Networks Based on Linear Regions *Jan 2022 – May 2024*
Supervised by Thiago Serra, Bucknell University

- Investigated geometry- and linear region–based approaches to neural network pruning, extending beyond traditional sparsity-driven methods.
- Derived upper bounds on linear regions for pruned networks and designed pruning strategies aligned with these bounds, improving accuracy in fully connected and convolutional architectures.
- Contributed to optimization-based analysis of trained neural networks, applying mathematical optimization to understand and reason about network behavior.

AWARDS AND FELLOWSHIPS

2025 CPAIOR 2025 Best Paper Award
2025 3rd AI4OPT Student Workshop 3-Minute Research Pitch Competition First Place
2023 INFORMS 2023 Undergraduate Operations Research Prize Finalist
2023 Viterbi School of Engineering/Graduate School Fellowship
2023 The Bucknell Prize in Computer Science and Engineering
2022 Ciffolillo Healthcare Technology Inventors Program