Hao Bai

+86 198 8327 0881 • Haob2@illinois.edu

EDUCATION

University of Illinois at Urbana-Champaign Champaign, IL, USA
M.S. in Computer Science 08/2023- 06/2025

University of Illinois at Urbana-Champaign Champaign, IL, USA
B.S. in Computer Engineering **GPA: 3.7/4.0** 08/2019- 06/2023

Zhejiang University Hangzhou, ZJ, China
B.E. in Electronic & Computer Engineering **GPA: 3.8/4.0** 08/2019- 06/2023

SELECTED PUBLICATIONS

Bai, Hao. "GoAutoBash: Golang-based multi-thread automatic pull-execute framework with GitHub webhooks and queuing strategy." International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2022). Vol. 12253. SPIE, 2022.

Bai, Hao. "Modern Distributed Data-Parallel Large-Scale Pre-training Strategies for NLP models." 2022 6th International Conference on High Performance Compilation, Computing and Communications (HP3C 2022). ACM, 2022.

Bai, Hao. "A Practical Three-phase Approach To Fully Automated Programming Using System Decomposition And Coding Copilots." 2022 5th International Conference on Machine Learning and Machine Intelligence (MLMI 2022). ACM, 2022.

PROFESSIONAL EXPERIENCE

Book Author, Tsinghua University Press Beijing, China
Contracted Treatise Author, Editor In Charge: Meiying Shen, TUP CS Department 08/2022- Present

- Contracted with TUP as the author of the book *Modern Software Frameworks for Deep Learning*.
- Surveyed the mainstream architecture of deep learning frameworks and taxonomized popular existing ones.
- Implemented the *MedoFlow* framework as an educational software framework for deep learning.

Research Intern, Microsoft Research Beijing, China
Research Intern, Mentor: Shilin He, MSRA DKI Group 11/2022- 05/2023

- Worked as a research intern on facilitating the cloud incident management system using language models.
- Implemented an online version of BERT for online root cause category classification and solved the time drifting problem, increasing the accuracy from 0.47 to 0.63.
- Experimented and evaluated a variety of prompting techniques to solve root cause category classification problems, like In-context Learning (ICL), Chain-of-Thought prompting (CoT) and Role-playing prompting on their performance of matching-based classification by first retrieving similar documents.

CORE RESEARCH EXPERIENCE

Empirical Study on Codex-based Prompt Tuning For Moral Value Prediction Champaign, IL
Research project, Advisor: Prof. Heng Ji, UIUC CS Department 11/2022 -02/2023

- Predicted the moral values of a wide range of corpus on social media and evaluated the accuracy.
- Used the K-fold stratified strategy to split the moral datasets MFTC (Twitter) and MFRC (Reddit) for training and testing.
- Inferenced and evaluated the moral values by BERT and Codex respectively under normal and low-resource scenarios.
- Achieved approximately the same performance on MFRC using Codex with 50 samples when BERT used over 10,000 samples, and achieved a much higher performance using Codex than BERT in low-resource scenarios.

Instance Segmentation with Rendering 3D Objects Champaign, IL
Summer Research Project, Advisor: Prof. Jooyoung Seo, NCSA SPIN Program 06/2022 -08/2022

- Built an image segmentation dataset for segmenting from scratch and classifying each brick in a Lego stack.
- Completed a script for rendering images of 3D Lego brick objects from different directions using Maya to construct an artifact dataset in addition to the real dataset.
- Implemented a three-phase pipeline for instance segmentation, including background elimination (image segmentation with U-Net), instance segmentation (with Mask R-CNN), and image segmentation (with GoogLeNet).

COMPLETE PUBLICATIONS

Bai, Hao. “A Training Method For VideoPose3D with Ideology of Action Recognition.” 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML 2021). IEEE, 2021.
Bai, Hao. “VSC-WebGPU: A Selenium-based VS Code Extension For Local Edit And Cloud Compilation on WebGPU.” 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC 2021). IEEE, 2021.
Bai, Hao. “GoAutoBash: Golang-based multi-thread automatic pull-execute framework with GitHub webhooks and queuing strategy.” International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2022). Vol. 12253. SPIE, 2022.
Bai, Hao. “Modern Distributed Data-Parallel Large-Scale Pre-training Strategies For NLP models.” 2022 6th International Conference on High Performance Compilation, Computing and Communications (HP3C 2022). ACM, 2022.
Bai, Hao. “A Practical Three-phase Approach To Fully Automated Programming Using System Decomposition And Coding Copilots.” 2022 5th International Conference on Machine Learning and Machine Intelligence (MLMI 2022). ACM, 2022.
Bai, Hao. “ICP Algorithm: Theory, Practice And Its SLAM-oriented Taxonomy.” 2022 4th International Conference on Computing and Data Science (CONF-CDS 2022). EWA, 2022.
Bai, Hao. “Statistical and Geometric Views of Linear Algebra.” *Preprint.* Accepted by 2022 International Conference on Modeling, Algorithm and Artificial Intelligence (MAAI 2022). IEEE, 2022.
Bai, Hao. “Modern Software Frameworks for Deep Learning.” Book contracted with Tsinghua University Press (TUP), to be published by the end of 2023.

ADDITIONAL RESEARCH EXPERIENCE

Attributed Inline Citation Generation using LLM Champaign, IL
Summer Research project, Advisor: Prof. Heng Ji, UIUC CS Department 05/2023 -Present

- Surveyed state-of-the-art inline citation generation approaches using LM and LLM.
- Surveyed LM attribution methods and approaches to evaluate them.

Amazon Alexa Social Chatbot Grand Challenge (Research Group) Champaign, IL
Competition project, Advisor: Prof. Chengxiang Zhai, UIUC CS Department 10/2022 -Present

- Work as a part-time machine learning R&D, mainly in charge of prompt engineering, model fine-tuning and evaluation.
- Built a synthesized dataset from WoW and WoI using ChatGPT to fine-tune Flan-T5 for the entity tracker and query generator.
- Finetuned a search decision model based on heuristics of the output confidence and integrated it into the dialog system.
- Implemented several human evaluation interfaces, and reproduced several automatic evaluation metrics like DynaEval.

MedoFlow: An Educational Software Framework for Deep Learning Champaign, IL
Thesis project, Advisor: Prof. Volodymyr Kindratenko, UIUC ECE Department 08/2022 -05/2023

- Built a deep learning framework from scratch, which achieves state-of-the-art accuracy and decent training efficiency.
- Implemented a variety of operators including matrix multiplication, convolution, batch normalization, and utilized the framework to reproduce famous networks like MLP, LeNet-5, RNN, etc.
- Conducted experiments to show that MedoFlow works as accurate as PyTorch and TensorFlow, and can easily improve its computational efficiency using accelerations provided by TVM and hardware devices.

Self-supervised Open Domain Question Answering Chatbots As Teaching Assistants Champaign, IL
Research project, Advisor: Prof. Volodymyr Kindratenko, UIUC ECE Department 08/2022 -05/2023

- Bootstrapped the idea of utilizing self-supervised learning for developing open domain question answering systems.
- Generated fine-tuning datasets for the retriever, reader and generator models using GPT-3 based on the existing knowledge sources.
- Developed the Entity Tracker algorithm and integrated to the chatbot system, which enables the system to capture contexts.
- Implemented the front-end of the chatbot and deployed the chatbot on various platforms like Discord, Slack and WeCom.

Distributed Data Parallel Training for Large-scale Deep Learning Models Champaign, IL
Research project, Advisor: Prof. Kindratenko Volodymyr, UIUC ECE Department 12/2021 -05/2022

- Studied the distributed parallel training methods to speed up very large deep learning models.
- Reproduced the GPT-2 and RoBERTa models on the mini-supercomputer HAL at NCSA.
- Applied different data-parallel training strategies on the PyTorch version of GPT-2, like Single Parameter Server (DP), Distributed Parameter Server (DDP), Horovod and Apex.
- Profiled, visualized and compared the performance of each strategy in detail and completed a detailed technical report, which was accepted by the conference HP3C.

GoAutoExecuter: Golang-based Multi-Thread Automatic Pull-Execute Framework Hangzhou, China
Research project, Advisor: Prof. Steve S. Lumetta, UIUC ECE Department 10/2021 -02/2022

- Implemented the *GoAutoExecuter* framework inherited from Wenqing Luo's *GoAutoGrader* framework, which utilizes the GitHub WebHook to download students' repos, run the grading script, and pushes the results back to GitHub.
- Utilized Golang for the producer-consumer model and concurrent handling to improve the server throughput.

A Survey of The ICP and SLAM Algorithms Champaign, IL
Research project, Advisor: Prof. Hao Li, University of California, Berkeley 08/2021 -10/2021

- Surveyed different implementations of the ICP algorithm and its applications in SLAM, and taxonomized the ICP algorithm according to its applications in SLAM.
- Implemented the ICP algorithm utilizing point-point metric, point-line metric and line-line metric, compared and visualized the different implementations on the Stanford bunny using C++ and CMake.

Optimized Transformer SoC Design based on the Gemmini Architecture on Chipyard Framework Hangzhou, China
Summer Research Project, Advisor: Prof. Kejie Huang, Institute of VLSI Design, ZJU 06/2021 -10/2021

- Used CHISEL language to conduct FPGA system programming and improved Gemmini architecture to realize performance optimization module for the transformer.
- Adopted systolic array and PE weight to optimize some layers of the transformer and accordingly advanced the design of the chip to reduce power consumption and area.
- Deployed Docker as the working environment and mounted the working directories in the host machine with Git as the version control system.

Co-planar Data Enhancement of Human3.6M Hangzhou, China
Research Project, Advisor: Prof. Gaoang Wang, ZJU-UIUC ECE Department 03/2021 -02/2022

- Built a new dataset based on Human3.6M by transforming the dataset from single-object to multi-object.
- Designed, implemented and optimized an algorithm to eliminate the collisions of multiple co-planar objects.
- Visualized the experimental results of the objects and completed the experiments part of the paper.

INDIVIDUAL PROJECTS

Fully Automated Programming Using System Decomposition And Coding Copilots Champaign, IL
Individual research project 07/2022 -08/2022

- Proposed a simple neuro-symbolic approach that integrates the techniques of programming like system decomposition and the existing very large-scale language models to generate code fully automatically.
- Concluded several empirical prompt templates for generating better code using Codex and GPT-3.
- Profiled the performance of GitHub Copilot and GPT-3 on different tasks, and summarized their various capabilities.

VS Code Extension for AT&T i386/IA32 Assembly Language Support Champaign, IL
Individual Research project, Advisor: Prof. Steve S. Lumetta, UIUC ECE Department 01/2022 -03/2022

- Implemented the language server and client of AT&T x86 Assembly language as a VS Code extension with help from Prof. Lumetta and Qi Li, which has been officially approved to be an auxiliary tool to use for students taking ECE 391.
- Developed keyword highlighting, snippet auto completion, and code linting (static syntax and semantic check) for the language server.

VSC-WebGPU: VS Code Extension for Local Edit & Online Submission Tasks Champaign, IL
Individual research project, Advisor: Prof. Volodymyr Kindratenko, UIUC ECE Department 09/2021 -10/2021

- Developed VS Code extension which utilizes Node JS and Selenium to help students write code locally and push the code to the course website with the extension, which helps students write better code on the course website.
- Utilized technics like blocking and waiting using Node JS to ameliorate user's experience when uploading the code.

IntLife: Wechat applet development based on Vue.js and UniDB Hangzhou, China
Software Development Engineer, ZJU-UIUC Residential College 10/2020 -07/2021

- Developed an official efficient communication applet for the residential college that enables identity authentication, encryption communication, event announcement, lost & found, secondary market, grouping, etc.
- In charge of developing message list and message interchanging. Operated back-end systems.
- Constituted the front-end module using Vue.js. Used cloud functions provided by UniApp and cloud database by UniDB as an end-to-end communication method.

OTHER EXPERIENCE

Undergraduate Teaching Assistant of ECE220 Hangzhou, China

Teaching Assistant, Advisor: Prof. Steve S. Lumetta, ZJU-UIUC ECE Department 09/2021 -01/2022

- Assisted Prof. Lumetta in teaching students in the course ECE220: Computer Systems & Programming.
- Held weekly labs for international students to discuss about lecture contents, programming labs and homework, and arranged weekly office hour to interact with students.
- Developed and maintained the GitLab platform to release assignments for the course staff and to submit programming assignments for students.

Founder of Hepta Lab Hangzhou, China

Founder, Consultant: Prof. Zhaozhao Shao, ZJU-UIUC ECE Department 07/2020 -09/2022

- Established Hepta Lab with several divisions including UI group, main developer group, network safety group, product manager group, propagandizing group and research group.
- Published technical blogs and columns on a wide range of topics about deep learning and AI on Zhihu, including translations of trending papers and interpretation of popular technologies and algorithms.
- Completed a web app that forwards updates from course websites to domestic SMS apps for Chinese students.
- Constructed the portal website and implemented an email server for the lab to improve its publicity.

SKILLS & LANGUAGES

Computer: Python, C/C++, Golang, JavaScript, SQL, Assembly (AT&T), TeX

Language: Chinese (Native), English (Proficient), Japanese & Spanish (Rudimental)