Shiji (Neo) Xin

(857) 200-7727
shijixin@g.harvard.edu
Linkedin: www.linkedin.com/in/shijixin
Github: github.com/zinccat
Site: https://www.shijixin.me/

Education

Harvard University, M.S. in Data Science *May 2025 (Expected)*

Peking University, B.S. in Intelligence Science and Technology, Turing Class (Summa Cum Laude) *July 2023*

Experiences

Algorithm Developer Intern (Machine Learning Engineer Intern), Applied Materials *May. 2024 - August. 2024*

- **Image Generation**: Built an end-to-end image generation system with highly customizable user input and feedback (patent pending). Deployed a fully functional demo within two weeks using Gradio and achieved a 30% acceleration with TensorRT.
- **Image Segmentation**: Fine-tuned the Mask2Former and Segment Anything models for images with significant image-to-image variation, improving accuracy by 10%.
- **Pretraining**: Established self-supervising framework with DinoV2 and benchmarked on common downstream tasks.

Researcher, Harvard University *Sept. 2023 -*

- **Accelerating Agents**: Working on improving the throughput of current LLM based tool-use agent systems through context switch and scheduling by tweaking the vLLM system.

Research Intern, Peking University (CoRe Lab) *Nov. 2021 - Aug. 2023*

- **Physics Informed Deep Learning for Seismic Wave Simulation**: Developed various datasets to simulate seismic waves and constructed physics-informed neural networks. Effectively addressed time domain extrapolation by applying physical constraints. The outcomes demonstrated an 75% reduction in relative error when extrapolating to previously unseen time points.
- **Causal Reasoning**: Optimized a neuro-symbolic causal discovery approach and implemented an object-centric model. Designed a structured neural network for causal discovery, achieving the performance reported in the paper using only 1/15 of the dataset, resulting in a 93% reduction in data usage.
- **Word Learning**: Built a dataset for machine word learning and evaluated vision-language baselines. Compared to human study results we conducted, state-of-the-art models showed deficiencies in word learning.

Undergraduate Visiting Researcher, Stanford University (Stanford Vision and Learning Lab) *June 2022 - Aug. 2022*

- Created a framework for visual relationship understanding using CLIP. Identified CLIP's deficiency in relationship understanding and proposed hard negative mining strategy that resulted in over a 40% accuracy improvement on the VRD dataset.

Research Intern, Peking University (ZERO Lab) *June 2021 - Oct. 2021*

- Designed an adversarial training approach for domain generalization by leveraging the mathematical relationship between Invariant Risk Minimization and DAT objectives.

Skills

Languages: Python, C/C++, Rust, SQL, CUDA, Bash, R, JavaScript, HTML, CSS, MATLAB, Stata

Tools and Frameworks: Google, PyTorch, TensorFlow, Jax, Triton, Pandas, OpenCV, Kubernetes, GCP, AWS, HuggingFace, React, LangChain, Docker, Kubernetes, Git, Selenium, Flask, FastAPI, Streamlit, Gradio, NVIDIA Modulus, vLLM, DeepSpeed, Neo4J, Accelerate, CI/CD, Slurm, MLFlow

Publications

FastAgent (first author): Enhanced Scheduling Strategies for Efficient Tool-Integrated Large Language Model Serving. **Compound AI Systems Workshop 2024** Paper

GlobalTomo: A global dataset for physics-ML seismic wavefield modeling and FWI. **In Submission** Project Page

MEWL: Few-shot multimodal word learning with referential uncertainty. **ICML 2023** Project Page

On the Connection between Invariant Learning and Adversarial Training for Out-of-Distribution Generalization. (first author) **AAAI 2023 (Oral)** Paper