Tai Vu

taivu@meta.com • linkedin.com/in/taiducvu
taivu1998.github.io • github.com/taivu1998

EDUCATION

Stanford University   Sep 2018 - Jun 2022
- B.S., Computer Science (AI Specialization) • GPA: 4.18 / 4.0 • Graduated with Distinction (Summa Cum Laude)
- [Endorsed](#) by Fields Medalist Prof. [Ngo Bao Chau](#) for exceptional mathematical talent and research potential.

AWARDS

International Mathematical Olympiad: Bronze Medal	American Mathematics Competition: Gold Medal
Vietnam Math Olympiad: Gold Medal (Rank #1)	International Math Local Tournament: Gold Medal
Hanoi Open Math Competition: Gold Medal	Regional Math Competitions: 5+ Gold Medals

WORK EXPERIENCE

Meta Platforms   Senior Machine Learning Engineer   July 2022 - Present
- **#1 impact driver** to organizational topline metrics, directly contributing to **90-120% of revenue and session targets** and **40-50% of DAU goals** through end-to-end ownership of high-leverage ML solutions. Recognized with Exceeded Expectations+ ratings over multiple halves.
- Spearheaded **large-scale, cross-organizational initiatives spanning 20+ engineers** across infrastructure, product, and modeling teams. Independently defined strategies, aligned stakeholders, and executed system-wide upgrades that materially advanced Meta's notifications recommendation system.
- Pushed the frontier of recommendation modeling by pioneering advanced architectures and optimization strategies. Developed novel techniques including **contextual and long-horizon representation learning**, **dynamic user interest modeling**, **causal inference**, **knowledge distillation**, **pairwise ranking**, and **value-aware multi-task learning** to capture latent interests and maximize user engagement.
- **Scaled team capability by 3x** by proactively scoping and launching new technical roadmaps, **mentoring and onboarding 10+ IC4/IC5 engineers**, and creating a culture of technical rigor that enabled sustained growth and research-grade innovation.

Meta Platforms   Software Engineering Intern   Jun 2021 - Sep 2021
- Developed large-scale ML models to generate contextual embeddings of users and ads, enhancing **semantic understanding** and **fine-grained personalization** in ad retrieval and ranking systems.
- Designed and deployed representation learning frameworks for user-ad matching across multiple surfaces, significantly improving relevance and driving **ad revenue growth**.

Meta Platforms   Software Engineering Intern   Jun 2020 - Sep 2020
- Enhanced latency profiling tools with module-level debugging. Achieved a **5x increase in runtime efficiency**.
- Accelerated Conv1D and channel shuffle operations through low-level optimizations, delivering up to **10x operator-level speedup** for on-device speech and NLP model inference.

RESEARCH EXPERIENCE

Stanford AI Lab   Undergraduate Researcher   Apr 2020 - Sep 2021
- Built large-scale data pipelines to ingest, align, and preprocess satelite imagery and global forest loss driver labels. Implemented advanced **data augmentation** and **stratified sampling** techniques to improve signal diversity and model robustness across diverse geographies.
- Designed and trained deep learning models, including **CNNs**, **LSTMs**, and **multimodal fusion architectures**, to classify forest loss drivers from multi-temporal satelite imagery, achieving **80% classification accuracy** and supporting scalable environment monitoring.

Stanford InfoLab   Undergraduate Researcher   Jan 2020 - Apr 2020
- Engineered a high-throughput input pipeline for loading and preprocessing underwater video data, including **image normalization**, **spatial augmentation**, and **temporal slicing**, to support robust model training.
- Developed **Mask R-CNN** and **U-Net** models to detect, localize, and temporally track coral structures in underwater environments, enabling fine-grained analysis and monitoring of reef health over time.

Computer Science Research Lab   Undergraduate Researcher   Sep 2019 - Dec 2019
- Researched model compression techniques, including **structured pruning**, **regularization-based sparsity**, and **weight quantization**, to reduce the computational footprint of deep convolutional networks.
- Applied regularization and pruning strategies to ResNet, achieving a **15% reduction in FLOPs** with a **2% improvement in accuracy**, demonstrating efficient compression without performance trade-offs.