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.
- **Finetuned and productionized a compact LLM** to enhance content generation and candidate ranking.
- Pushed the frontier of recommendation modeling by pioneering advanced architectures and optimization strategies. Developed novel techniques including **reinforcement learning**, **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.
- 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.
- **Scaled team capability by 3x** by proactively scoping and launching new technical roadmaps, **mentoring and onboarding 10+ IC4/IC5 engineers**, and enabling sustained growth and research-grade innovation.

Meta Platforms — Software Engineering Intern — Jun 2021 - Sep 2021
- Developed large-scale ML models in PyTorch 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**. Used Python, SQL, C++, PHP, and Spark.

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 via low-level kernel optimizations for the **PyTorch** framework, delivering up to **10x operator-level speedup** for on-device speech and NLP models. Used Python and C++.

Research Experience

Stanford AI Lab — Undergraduate Researcher — Apr 2020 - Sep 2021
- Built large-scale data pipelines in Python, NumPy, and Pandas 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 in PyTorch, 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 in TensorFlow to detect, localize, and temporally track coral structures in underwater environments, enabling fine-grained analysis and monitoring of reef health over time.

Computer Science 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.

Publications

FlapAI Bird: Training an Agent to Play Flappy Bird Using Reinforcement Learning Techniques
Tai Vu, Leon Tran [paper] [github]

How Not to Give a FLOP: Combining Regularization and Pruning for Efficient Inference
Tai Vu, Emily Wen, Roy Nehoran [paper] [github]

Privacy Preserving Inference of Personalized Content for Out of Matrix Users
Michael Sun, Tai Vu, Andrew Wang [paper] [github]

GANime: Generating Anime and Manga Character Drawings from Sketches with Deep Learning
Tai Vu, Robert Yang [paper] [github]

BERT-VQA: Visual Question Answering on Plots
Tai Vu, Robert Yang [paper] [github]

Pixel-Perfect Piloting: Superhuman Control of Pixelcopter via Reinforcement Learning
Tai Vu, Brad Nikkel, Jenny Yang [paper] [github]

Beyond the Panels: A Deep Neural Network Approach for Manga Object Detection
Tai Vu, Robert Yang [paper] [github]

Amplifying Emotional Signals: Data-Efficient Deep Learning for Robust Speech Emotion Recognition
Tai Vu [paper] [github]

From Bayes to BERT: A Comprehensive Benchmark for State-of-the-Art Intent Detection
Tai Vu, Robert Yang [paper] [github]

Skills

Languages: Python • C • C++ • SQL • JavaScript • TypeScript • Java • R • PHP
Technologies: PyTorch • TensorFlow • HuggingFace • NumPy • Pandas • vLLM • Accelerate • FSDP • TRL OpenRLHF • Scikit-learn • AWS • Google Cloud • Azure • Spark • HTML • CSS • React • Node.js • MongoDB Express.js • Next.js • React Native • Django Heroku • Sass • Git • Linux
Coursework: Machine Learning • Deep Learning • LLMs • NLP • Reinforcement Learning • Computer Vision Information Retrieval • Spoken Language Processing • Graph ML • Parallel Computing • Data Structures & Algorithms • Systems • Web Applications • Statistics • Linear Algebra • Convex Optimization • Game Theory

Projects

FlapAI Bird: Training AI Agents to Play Flappy Bird
- Designed a Flappy Bird agent in PyTorch and NumPy that achieved superhuman performance with **2,000+ scores** by implementing and fine-tuning **reinforcement learning** algorithms, including **SARSA**, **Q-learning**, **function approximation**, and **deep Q networks**. Recognized by Dong Nguyen, creator of Flappy Bird.

Privacy Preserving Inference of Personalized Content for Out of Matrix Users
- Architected a novel graph-based recommender system using **BERT-powered embeddings** and **graph neutral networks** to deliver personalized, privacy-preserving recommendations for cold-start users, outperforming WMF by **7x** and DropoutNet by **1.5x** in user recall. Used PyTorch, HuggingFace, NumPy, Pandas, and Azure.

Pixelcopter-RL: Building Reinforcement Learning Algorithms for Playing Pixelcopter
- Developed an AI agent in PyTorch and NumPy to master Pixelcopter using **reinforcement learning** techniques (**Q-learning**, **SARSA**, **function approximation**, and **policy gradients**), achieving a **peak score of 379**.

ConnAIsseur: An AI-driven Recipe Recommendation Website
- Engineered a **web application** with a **BERT-based recommendation engine** that infers latent taste profiles from unstructured user preference data to provide highly personalized recipe suggestions. Used Python, Javascript, PyTorch, HuggingFace, NumPy, Pandas, Django, React, Node.js, HTML, CSS, and AWS.

GANime: Generating Anime Character Drawings from Sketches
- Developed generative models in TensorFlow and AWS, including **neural style transfer**, **Pix2Pix**, and **CycleGAN**, to automate the colorization of anime sketches, attaining **state-of-the-art scores of 220.5 FID and 0.76 SSIM**.

MangaNet: Building an Object Detection System for Mangas
- Designed advanced object detection model architectures, including **Faster R-CNN**, **RetinaNet**, and **YOLOv3**, for the manga domain, achieving a **state-of-the-art mAP of 71.0**. Used PyTorch, NumPy, and Pandas.