TRANG NGUYEN ANH THUAN

Ho Chi Minh City, Vietnam

trangnguyenanhthuan@gmail.com ◊ www.linkedin.com/in/tranganhthuan/◊ +84352767692

OVERALL STATEMENT

AI Researcher and Engineer with over four years of experience in deep learning, 3D computer vision, and multimodal AI systems. Skilled in both research and deployment, with publications at top venues like AISTATS and TMLR. Experienced in building scalable AI pipelines, fine-tuning LLMs, and developing real-time applications. Currently pursuing a PhD in Computer Science, with a strong interest in advancing methods at the intersection of vision, language, and graph learning. Actively seeking new job opportunities to apply my skills and contribute to innovative AI projects.

EDUCATION

RMIT University, Vietnam May 2025 - **Current**
PhD Student in Computer Science

RMIT University, Vietnam (GPA 3.3**)** Oct 2018 - Jun 2021
Bachelor of Information Technology

PROFESSIONAL EXPERIENCE

Vulcan Labs, Ho Chi Minh City Dec 2024 - **Current**
AI Engineer
- Research state-of-the-art (SOTA) reasoning models.
- Deploy AI services via APIs and queue-based architectures.
- Collaborate with backend and mobile teams to develop a math-solving chatbot.
- Design and implement pipelines for efficient evaluation of newly published LLMs.

RMIT University, Ho Chi Minh City Jun 2024 - Jun 2025
AI Research Assistant
- Conduct research on Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), and Vision-Language Models (VLMs).
- Design and run experiments on Parameter-Efficient Fine-Tuning (PEFT).
- Develop a custom RAG system tailored for educational applications.

Aimesoft, Ho Chi Minh City Aug 2024 - Dec 2024
AI Engineer
- Researched existing methods by conducting literature reviews to identify the most suitable approach for assigned tasks based on accuracy, efficiency, and computational cost.
- Applied computer vision techniques such as object detection, recognition, tracking, segmentation, and re-identification to deliver solutions tailored to customer needs.
- Developed generative network pipelines for creating images and videos based on specific task requirements.
- Visualized results and built demos to showcase the performance of developed methods.

HySonLab, Remote Aug 2023 - Jun 2024
AI Researcher
- Adjusted model architecture and conducted experiments using proposed methods on additional datasets.
- Submitted the paper and engaged in the rebuttal process.
- Cleaned code, published, and developed a demo for the accepted paper.

FPT AI Center, Ho Chi Minh City May 2021 - Aug 2023
AI Resident - Batch 1
FPT Software AI Residency Program [page]

- FPT Software AI Residency is a two-year program focused on nurturing young talent in AI and machine learning. Participants receive guidance from world-class mentors and scientists in a creative environment.
- Having the privilege to work with MILA, a top AI research institute founded by Yoshua Bengio, and supervised by their professors on two projects.
- Conducting research on point clouds and videos under the mentorship of Prof. Truong Son Hy, Dr. Thieu N. Vo, and Prof. Siamak Ravanbakhsh.
- Working with Efficient Self-attention Modules and Group Equivariant Neural Networks on 3D objects such as point clouds and meshes.
- Involved in projects requiring the handling of long-range sequences and graphs.
- Assisting with a project on open-vocabulary semantic segmentation in video, involving a vision-language model.

AWARDS

PhD Scholarship at A2I2
- Covers full tuition fees.
- Provides a stipend of A$34,400 per year for the first three years.

RMIT Vietnam 2020 Academic Achievement Scholarship for current students
- Awarded to the top 5 students with the highest GPA at the time.
- The scholarship covers 50% of the remaining program

RESEARCH EXPERIENCE

E(3)-Equivariant Mesh Neural Networks [paper][abstract][code][poster] 2023 - 2024
AISTAT 2024 (CORE A)
- Thuan N.A. Trang*, Khang Nhat Ngo*, Daniel Levy*, Thieu N. Vo, Siamak Ravanbakhsh, Truong Son Hy
- Developed Equivariant Mesh Neural Networks (EMNN) for 3D mesh data, outperforming complex equivariant methods with an efficient architecture and fast runtime.
- Extended E(n)-Equivariant Graph Neural Networks (EGNNs) by integrating mesh face, edge, and node aggregation, preserving equivariance and enabling richer geometric representations.
- Achieved state-of-the-art (SOTA) performance, surpassing previous equivariant models with lower memory and computational costs, while maintaining excellent performance compared to non-equivariant methods.
- Built the full pipeline using PyTorch Geometric and reran other models in the benchmarks to validate their invariance and robustness.

Hierarchical Self-Attention with Learnable Hierarchy for Long-Range Interactions [paper][abstract] [code][video] 2023 – 2024
Transactions of Machine Learning Research (TMLR) 2024
Thuan N.A. Trang*, Khang Nhat Ngo*, Hugo Sonnery*, Thieu N. Vo, Siamak Ravanbakhsh, Truong Son Hy
- Developed an end-to-end mechanism for constructing data-dependent hierarchies to guide the self-attention mechanism.
- Reduced self-attention complexity from quadratic to log-linear while maintaining or surpassing the ability to model long-range dependencies.
- Achieved state-of-the-art performance on graph datasets with reduced computational cost and demonstrated competitive results on long-range graph benchmarks and point cloud tasks while maintaining efficiency.
- Utilized PyTorch and PyTorch Geometric for model development and data processing, and employed Accelerate for distributed training.

Sequoia: Hierarchical Self-Attention Layer with Sparse Updates for Point Clouds and Long Sequences [paper][poster] 2022 - 2023
ICLR Workshop 2023
Hugo Sonnery*, Thuan N.A. Trang*, Thieu N. Vo, Siamak Ravanbakhsh, Truong Son Hy

- Addressed the inefficiency of self-attention for long inputs by introducing a hierarchical approach that divides and groups long sequences into meaningful subgroups.
- Developed a novel attention module designed to work with hierarchical structures, reducing self-attention complexity to O(nlog(n)).
- Achieved promising results on point cloud tasks and sequence classification benchmarks.
- Built the model and created custom data loaders using PyTorch, along with preprocessing pipelines in NumPy using vectorized functions to enable parallel computing.

Design equivariant neural networks for 3D point cloud [paper][abstract] 2021 - 2022
Arxiv 2022
Thuan N.A. Trang, Thieu N. Vo, Khuong D. Nguyen
- Designed a plug-in equivariant module for point cloud processing models, enabling MLPs, CNNs, and Transformers to achieve equivariance to specific angles, enhancing model robustness.
- Achieved notable improvements in both complexity and performance over previous equivariant models.
- Modified PyTorch's gradient checkpointing to optimize the balance between GPU memory usage and processing time.

SIDE PROJECTS

Fine-tuning LLaMA on ChartQA with LoRA[code] 2024-2025
- Implemented Low-Rank Adaptation (LoRA) to fine-tune LLaMA models for a Chart QA application using multiple approaches.
- Designed a modular pipeline for dataset preprocessing, model training, and evaluation, enabling easy customization for new experiment settings.
- Supported both Hugging Face's PEFT and Unsloth—Unsloth optimizes memory usage and GPU efficiency, while PEFT offers broader model compatibility and can be extended to multi-GPU training in the future.

Minimal Retrieval-Augmented Generation (RAG) Pipeline[code] 2024-2025
- Built a custom RAG pipeline supporting custom datasets with flexibility to use ChatGPT API, Gemini, or self-hosted models.
- Developed a full-stack solution with a React library for the frontend and FastAPI for serving backend and AI services.
- Implemented both semantic search (vector embedding search) and BM25 keyword search to enhance retrieval accuracy.
- Integrated LangChain and Chroma Vector Store for document retrieval, with Llama Cloud for cloud-based PDF parsing, MinerU for local conversion, and JWT for authorization.
- Utilized Redis and Redis Queue to efficiently manage PDF parsing tasks and prevent processing overhead.

Video Call Web App with Sign Language Detection[code] 2024
- Implemented real-time video streaming and sign language recognition using WebRTC and WebSocket, enabling both client-to-client and client-to-server communication.
- Developed an intuitive interface to collect data via webcam and manage it directly within the browser.
- Leveraged Mediapipe and Keras for human pose detection and sign language gesture classification.

OpenVocab for Video Instance Segmentation 2023
- Adapt open-vocabulary segmentation from images to video.
- Utilize CLIP to map between texts and frames in a video.
- Implement techniques to aggregate temporal features in the video.
- Use PyTorch and frameworks like MMCV for building the model and distributed training.

SKILLS AND LANGUAGES

Technical skills:
- **Modeling and Training:** Proficient in developing, training, and fine-tuning models using **PyTorch**, **NumPy**, **PyTorch Geometric**, **Scikit-learn, PEFT,** and **Unsloth**.
- **Agentic Development:** Experience with building AI agents using **LangGraph**, **LangChain**, and **Chroma Vector Store**.

- **Service Deployment:** Familiar with serving AI services through **FastAPI**, **PubSub**, **Redis**, and **Redis Queue**.
- **Data Visualization:** Experienced in visualizing data with **Matplotlib** and **Plotly**.
- **Environment Setup:** Skilled in setting up environments using **Linux** and **Docker**.
- **Building Demos:** Competent in creating demos using **Streamlit**, **Django**, and **ReactJS**.
- **Real-time Communication:** Familiar with **WebSocket** and **WebRTC** for real-time data transmission and video streaming.
- **Generative Model Workflow:** Basic experience with **ComfyUI** for creating and managing visual workflows in AI models and pipelines.
- **Team Collaboration & Agile Development:** Familiar with the **Scrum** process and working with teammates using **Git** for version control and collaboration.

Communication skills:
- **Technical Writing:** Skilled in writing clear and concise research papers, reports, and documentation, with hands-on experience in the paper submission process.
- **Presentation Skills:** Good at presenting ideas and research findings to both technical and non-technical audiences during group meetings, with additional experience in designing posters and recording videos for accepted papers.
- **Collaboration:** Effective in working within teams, including strong communication with peers, mentors, and collaborators.

Languages:
- **Vietnamese:** Native
- **English:** Intermediate (IELTS Overall 7.5 - Listening: 8.0, Read: 8.5, Writing: 6.5, Speaking: 6.5)

CERTIFICATES

Large Language Model Agents - Mastery Tier [link]
By UC Berkeley - 2024

Natural Language Processing Specialization [link]
By Coursera - 2021

TensorFlow Developer Certificate [link]
By Google Developers Certification - 2020

Machine Learning Scientist Track [link]
By DataCamp - 2020

Deep Learning Specialization [link]
By Coursera - 2019

REFERENCES

Prof. Truong Son Hy [website] - thy@uab.edu
Assistant Professor, Department of Computer Science, University of Alabama at Birmingham

Dr. Thieu N. Vo [website] - thieuvo@nus.edu.sg
Lecturer & Researcher at TDTU, Research Fellow at NUS

Prof. Siamak Ravanbakhsh [website] - siamak@cs.mcgill.ca
Assistant Professor, McGill University, Canada CIFAR AI Chair