Van Minh **Nguyen**

Gainesville, FL USA

vmnguyen251@gmail.com | n0k0m3.github.io | GitHub: n0k0m3 | LinkedIn: minhnguyen251

Work Experience

University of Florida
Gainesville, FL

AI Support Engineer — *Aug 2024 - Present*

- Provide technical support for researchers and PIs using HiPerGator supercomputer, specializing in multi-node/multi-GPU training
- Serve as NVIDIA Deep Learning Institute (DLI) Ambassador and Certified Instructor, leading training sessions on campus
- Teaching Assistant for DLI workshops, supporting students in hands-on deep learning applications
- Troubleshoot complex computational issues and optimize research workflows on HPC infrastructureteam
- Guide researchers in implementing efficient parallel computing strategies and best practices

Technetium Engineering
Remote

AI & Platform Engineer — *June 2024 - Present*

- Architected foundation for multi-node distributed computing systems for space applications
- Developed and implemented deployment inference and communication protocols integrating NVIDIA AI platform SDK
- Established and managed company's digital infrastructure including static website with email functionality
- Led organization setup using Microsoft Startup programs, implementing Microsoft Entra ID and Office for Business across the team
- Implementing computer vision algorithms for space object detection and 3D reconstruction

Florida Tech
Melbourne, FL

MLOps Technician - Neural Transmissions Lab — *Jan 2022 - July 2024*

- As the sole architect and manager, initiated and implemented an on-premise server cluster for the research lab, using Kubernetes on Ubuntu Server.
- Set up multi-user research environments with GPU support and role-based access control using JupyterHub, Kubernetes, and Keycloak
- Secured deployments using HTTPS, DNS configuration, short-lived SSH, and VNC over HTTPS

Teaching Assistant - Department of Mathematical Sciences — *Aug 2018 - May 2024*

- Taught and graded exams for Probability & Statistics, Neural Networks, Calculus I, II, III
- Aided students studying Stochastic Modeling and Theory of Stochastic Processes
- Provided technical support for students learning Neural Networks and Machine Learning

Engage-AI.org
Remote

Data Engineer - Contractor — *May 2023 - Jan 2024*

- Spearheaded the initial development and deployment of the Engage AI Data Platform, leveraging Cloudflare R2 for storage, DuckDB for query processing, and Apache Superset for data visualization
- Achieved a zero-cost initial deployment and laid the groundwork for a cost-efficient, cloud-based data management solution, setting the foundation for Engage AI's data platform
- Collaborated with data analysts to understand their data requirements, refining and optimizing the platform based on feedback

Truveta
Seattle, WA

Research Intern — *Jan 2022 - May 2022*

- Developed and deployed scalable Named Entity Recognition (NER) pipelines for clinical notes information extraction and de-identification using John Snow Labs' SparkNLP free model and a custom PyTorch model
- Analyzed cloud infrastructure costs and optimized potential savings by calculating cluster size, resulting in $2M annual savings and a 75% reduction in operating costs
- Conducted threat modeling using OWASP Threat Dragon and recommended mitigation strategies for pipeline deployment
- Created a clinical notes annotation tool prototype based on Label Studio and INCEpTION for internal use
- Navigated complex organizational processes and collaborated with multiple stakeholders to ensure timely access to resources and maintain project momentum

Graduate Intern — *May 2021 - Aug 2021*

- Developed a data quality measurement toolkit using Spark for the Truveta Health Data Model (THDM), which assessed data completeness, conformance to THDM, and plausibility based on OHDSI criteria
- The toolkit garnered trust from participating health providers, leading to multiple partnerships with new healthcare providers, a collaboration with Microsoft, and integration into Truveta Studio, Truveta's first product
- Established the foundational approach and model for Synthetic Patient Health Data, leveraging probabilistic theory of document retrieval, representation learning, and deep learning, with applications in generating synthetic health data, filling missing patient data, and predicting doctor-patient encounters and diagnoses/medications from health history
- Designed and implemented a synthetic patient data model using Monte Carlo sampling for stress-testing and bottleneck identification in the ETL process, capable of generating millions of records in 1 hour, ensuring robust data ingestion and querying systems
- Developed an annotation recommender system for medical concept normalization, reducing annotators' workload by 80%
- Deployed and integrated the data quality measurement toolkit into the data ingestion pipeline using Azure DevOps Pipelines, leveraging Azure resources such as Databricks, Azure VM, Storage Blob, and Data Factory (ADF)

Education

Florida Tech *Melbourne, FL*

Ph.D. Operations Research *Aug 2020 - May 2024*

- Determine bacteria mutation rate with double stochastic branching process with random offspring
- Research privacy-focused, longitudinal (temporal) generation of synthetic Electronic Health Records (EHR) with Differential Privacy
- Dissertation: *"Representation Learning for Generative Models with Applications to Healthcare, Astronautics, and Aviation"*. Advisor: Dr. Ryan T. White.

M.S. Operations Research, GPA: 4.0 *Aug 2018 - May 2020*

B.S. Biochemistry (Biology Emphasis), GPA: 3.6 *Aug 2014 - May 2018*

Skills

AI/ML Development	Python, PyTorch, TensorFlow, R, C#, ONNX, NVIDIA AI (TensorRT, TRT-LLM, PhysicsNeMo/Modulus, NeMo, Numba)
Data Science & Big Data	NVIDIA RAPIDS, Spark/PySpark, Dask, Data Mining, Statistical & Stochastic Modeling
MLOps & Deployment	Docker, Kubernetes (on-prem), Triton Inference Server, NVIDIA NIM, MLFlow, Kubeflow, Slurm, Azure Pipelines
Cloud & Databases	Microsoft Azure, Databricks, SQL (PostgreSQL, ClickHouse, MariaDB), Vector DB (Milvus), DuckDB

Research Experience

Florida Tech *Melbourne, FL*

Satellite 3D Reconstruction Research - Department of Mathematics and Systems Engineering *Aug 2023 - Present*

- Adapted Instant NeRF and D-NeRF algorithms for high-definition 3D modeling of resident space objects (RSOs) to assist with functionality identification and on-orbit servicing (OOS)
- Evaluated the algorithms for 3D reconstruction quality and hardware requirements using datasets of spacecraft mock-up images captured under various lighting and motion conditions at the Orbital Robotic Interaction, On-Orbit Servicing and Navigation (ORION) Laboratory
- Demonstrated the feasibility of training Instant NeRF on on-board computers to learn high-fidelity 3D models with manageable computational costs
- Developed an approach for mapping geometries of satellites on orbit based on 3D Gaussian splatting, capable of running on current spaceflight hardware
- Achieved nearly 2 orders of magnitude faster rendering of higher quality novel views of unknown satellites compared to previous NeRF-based algorithms, enabling on-board training and downstream machine intelligence tasks for autonomous guidance, navigation, and control
- Presented a novel approach for mapping geometries and high-confidence detection of components of unknown, non-cooperative satellites on orbit by combining accelerated 3D Gaussian splatting, virtual view rendering, and ensemble YOLOv5 object detection (submitted to CVPR 2024)

Suicide Prevention Research - Department of Computer Engineering and Sciences *Jan 2023 - Jan 2024*

- Enhanced data scraping pipelines for Twitter and Reddit, reducing ingestion time by 60 times using pure CLI tools (ripgrep, jq, awk, etc.) instead of databases like MySQL or MongoDB
- Employed a model explanation method, Layer Integrated Gradients, on top of a fine-tuned state-of-the-art encoder language model to assign attribution scores to tokens from Reddit users' posts for predicting suicidal ideation
- Proposed a methodology for preliminary screening of social media posts for suicidal ideation using the extracted token attributions, without relying on large language models during inference
- Developed a novel approach using off-the-shelf generative large language models (LLaMA2) to generate natural language explanations for suicide risk from users' Reddit posts
- Benchmarked various language models utilizing annotations and explanations by psychology experts, demonstrating the effectiveness of LLMs in classifying and responding with helpful reasoning for suicidal risk diagnosis

Deep Reinforcement Learning for Robotics - Department Of Mechanical And Civil Engineering *Jan 2022 - Sep 2022*

- Investigated the feasibility of using transformers in off-policy Deep Reinforcement Learning problems, starting with Q-learning
- Explored the potential of transformers in solving Partially Observable Markov Decision Process (POMDP) problems
- Implemented a spatio-temporal transformer module and action-memory within Soft Actor-Critic (SAC) and Twin Delayed Deep Deterministic (TD3) policy gradient architectures
- Enabled the agent to "remember" its previous actions and effectively predict the next ones, enhancing decision-making capabilities
- Benchmarked the transformer-enhanced SAC and TD3 models against TD3-LSTM on *highway-env*, an OpenAI Gym environment for autonomous driving decision-making tasks

Publications

Journal Articles

Characterizing Satellite Geometry via Accelerated 3D Gaussian Splatting
Van Minh Nguyen, Emma Sandidge, Trupti Mahendrakar, Ryan T. White
Aerospace 11.3 (2024). 2024

promSEMBLE: Hard Pattern Mining and Ensemble Learning for Detecting DNA Promoter Sequences
Bindi M. Nagda, Van Minh Nguyen, Ryan T. White
IEEE/ACM Transactions on Computational Biology and Bioinformatics 21.1 (2023) pp. 208–214. 2023

Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS)

Archana Tikayat Ray, Anirudh Prabhakara Bhat, Ryan T. White, Van Minh Nguyen, Olivia J. Pinon Fischer, Dimitri N. Mavris
Aerospace 10.9 (2023). 2023

Determination of Mutation Rates with Two Symmetric and Asymmetric Mutation Types
Jewgeni H. Dshalalow, Van Minh Nguyen, Richard R. Sinden, Ryan T. White
Symmetry 14.8 (2022). 2022

Conference Proceedings

Natural Language Explanations for Suicide Risk Classification Using Large Language Models
William Stern, Seng Jhing Goh, Nasheen Nur, Patrick J Aragon, Thomas Mercer, Siddhartha Bhattacharyya, Chiradeep Sen, Van Minh Nguyen
Machine Learning for Cognitive and Mental Health Workshop (ML4CMH 2024) at AAAI 2024, 2024

3D Reconstruction of Non-Cooperative Resident Space Object using Instant NeRF and D-NeRF
Basilio Caruso, Trupti Mahendrakar, Van Minh Nguyen, Ryan T. White, Todd Steffen
33rd AAS/AIAA Space Flight Mechanics Meeting 33.417 (2023). 2023

Conceptualizing Suicidal Behavior: Utilizing Explanations of Predicted Outcomes to Analyze Longitudinal Social Media Data
Van Minh Nguyen, Nasheen Nur, William Stern, Thomas Mercer, Chiradeep Sen, Siddhartha Bhattacharyya, Victor Tumbiolo, Seng Jhing Goh
2023 22nd IEEE International Conference on Machine Learning and Applications (ICMLA), 2023