Sai Manohar Vemuri

(312) 687-5356 | vemurimanohar6642@gmail.com | linkedin.com/in/vemuri02 | github.com/vemuri02 | vemuri02.github.io

EDUCATION

PhD in Computer Science & Engineering, Illinois Institute of Technology — Aug 2022 – Present

B.Tech in Computer Science, Jawaharlal Nehru Technological University — Jun 2018 – May 2022

SKILLS

Data Science & AI: Image Classification, Object Detection, Semantic Segmentation, Feature Extraction, Augmentation
Programming Languages: Python, R, C++, Java, SQL, NoSQL, CUDA
Frameworks & Tools: Numpy, matplotlib, scikit-image, PyTorch, Keras, TensorFlow, OpenCV, NLTK, LangChain, spaCy
Big Data & Cloud: AWS (S3, EC2, Glue, SageMaker, Lambda), PySpark, Databricks, Hadoop, Apache Airflow, Azure
Model Optimization: Neural Architecture Search (NAS), Pruning, Quantization, Knowledge Distillation
Math & Optimization: Linear Algebra, Multivariate Calculus, Optimization Algorithms, PCA
Generative AI Expertise: Transformer Models, BERT, LLM, Reinforcement Learning, GANs, VAEs, Diffusion Models

WORK EXPERIENCE

Research and Teaching Assistant, Illinois Institute of Technology — Oct 2024 – Present

Researching HW/SW co-design for ML/DL on edge devices, focusing on object detection, semantic segmentation, and LiDAR-camera fusion for ADAS and autonomous driving. Working on model optimization using NAS, quantization, knowledge distillation, and pruning, implementing low-power RTL design for efficient FPGA deployment. TA for Hardware Acceleration for ML course covering GPUs, NPUs, and FPGAs for efficient ML model training and inference optimization.

AI Research Intern, SoundSafe.ai — Dec 2024 – Aug 2025

Leading development of AI-driven audio security solutions, focusing on speech authentication and audio forensics using self-supervised learning, while managing Azure cloud environments, model training, and scalability.

AI Engineer Intern, Aster Ramesh Hospitals — Dec 2020 – Aug 2022

Developed deep learning models for segmentation of CT/MRI scans targeting gliomas, meningiomas, and cardiac conditions, using YOLOv4, Faster R-CNN, and Mask R-CNN for ischemic stroke and myocardial infarction detection. Improved diagnostic accuracy with GAN-based augmentation, attention mechanisms, and texture analysis (Gabor filters, Haralick features), achieving 95% F1 score, 93% accuracy, and 90% precision in medical condition detection.

Network Engineer Industrial Trainee, Sir C R Reddy Polytechnic — Nov 2018 – Apr 2019

Gained practical experience in TCP/IP, OSI model, subnetting, and network configuration (routers, switches) for LANs and WANs, troubleshooting with Wireshark. Assisted in implementing SMTP, POP, and FTP protocols, while supporting network security, firewall configurations, and VPN setups.

PUBLICATIONS

"CB-DistillGrad: A Class-Balanced Knowledge Distillation Loss for Long-Tailed Visual Recognition," *Under Review.* Vemuri, S.M.; Gundrapally, A.; Kim, T.; Kim, J. (Senior Member, IEEE); Choi, K. (Senior Member, IEEE).

"Hardware Accelerator Design by Using RT-Level Power Optimization Techniques on FPGA for Future AI Mobile Applications," *Electronics*, vol. 14, no. 16, p. 3317, 2025. Gundrapally, A.; Shah, Y.A.; Vemuri, S.M.; Choi, K. DOI: 10.3390/electronics14163317.

PROJECTS

EfficientNet Distillation for Real-Time Road Surface Classification on Edge Devices: Compressed EfficientNet-B0 by 31× with knowledge distillation and Langevin dynamics regularization, achieving 88% Top-1 accuracy with 95% retention, optimized for edge deployment using PEDE, Haar approximations for Swish activation, mixed precision and QAT.

Swin Transformer: Fine tuned SOTA Swin Transformer with hierarchical features and shifted window-based multi-head ViT architecture for high-resolution images, achieving competitive efficiency with 30% fewer parameters and linear complexity.

Customer Churn Risk Prediction: Deployed Unsupervised Learning model on SageMaker for real time ecommerce streaming data through Cloud 9 to estimate churn risk, processed with Apache Flink and Kinesis with focus on end-to-end ML pipeline.

Seismic Image Segmentation using U-Net: Developed a custom U-Net with optimized loss functions (BCE, focal, dice) for seismic segmentation, achieving an 84.16 IOU score and outperforming ResNet50, ResNet101, and VGG16 with fewer parameters.

Object Detection using Transformers: Engineered ViT model with multi-head self-attention layers for object detection, achieving an 85% classification accuracy and mean IOU of 0.82. Optimized architecture for efficiency on Edge Devices and accommodated different object sizes.

Image Translation using Self Attention-GAN: Designed state-of-the-art techniques, including GANs, Spectral Normalization, and self-attention mechanisms within the SAGAN architecture, enhancing cross-domain image translation with a 9% improvement in FID score, indicating highly realistic and coherent results while preserving semantic content.

CERTIFICATIONS

AWS Certified Machine Learning Specialty (Udemy) • Machine Learning from Stanford University (Coursera)