PAVAN KANDAPAGARI

Tech Lead | Foundation Models & Embodied AI Munich, Germany | +49 1573 9467478 | 785pavan@gmail.com linkedin.com/in/kandapagari | kandapagari.vercel.app

PROFESSIONAL SUMMARY

R&D Tech Lead specializing in **Vision-Language-Action (VLA)** models and **Imitation Learning** for autonomous agents. Expert in architecting distributed training pipelines for large foundation models on **AWS HyperPod**, utilizing deep knowledge of Transformers and multi-modal learning. Proven track record of bridging the gap between research innovation and production, scaling infrastructure to handle **75TB+** datasets, and deploying high-performance models for real-world robotic systems.

TECHNICAL SKILLS

- **Core AI Domains:** Vision-Language-Action (VLA) Models, Foundation Models, Imitation Learning, Generative AI (LLMs, Diffusion, GANs), Behavior Cloning, Robot Control.
- **Deep Learning Architectures:** Transformers (BERT, GPT, ViT), Attention Mechanisms, CNNs, RNNs/LSTMs, Multi-modal Fusion.
- **Frameworks & Libraries:** PyTorch (Expert), TensorFlow, JAX, Hugging Face, LangChain, OpenCV, NumPy, Pandas, Gym.
- **Infrastructure & MLOps:** AWS (SageMaker, HyperPod, EC2, S3), Docker, Kubernetes, GitLab CI/CD, MLflow, Ray Serve.
- **Performance Engineering:** Distributed Training (Data/Model Parallelism), CUDA, TensorRT, ONNX, Quantization, Mixed-Precision.
- **Languages:** Python (Expert), C++, Java, Rust, SQL, Bash.

PROFESSIONAL EXPERIENCE

Agile Robots SE | Munich, Germany **Tech Lead – Foundation Models for Intelligent Agents** | *Aug 2025 – Present Leading technical strategy for VLA models, pre-training infrastructure, and data strategy for next-gen robotic agents.*

- **Foundation Model Architecture:** Pioneered a novel VLA architecture by attaching action-decoding heads to pre-trained Vision-Language Models (VLMs), enabling zero-shot generalization across complex manipulation tasks.
- **Distributed Systems at Scale:** Orchestrated end-to-end pre-training on **AWS SageMaker HyperPod** across distributed GPU clusters. Implemented advanced data and model parallelism to maximize throughput, achieving **>85% GPU utilization**.
- **Big Data Strategy:** Managed a **75TB** multimodal dataset (video, sensor streams, language). Engineered data loading pipelines with intelligent caching and S3 optimization, eliminating I/O bottlenecks and reducing data loading latency by **60%**.
- **Training Stability:** Resolved convergence issues in massive-scale cluster training (100+ GPUs) using gradient accumulation, mixed-precision training, and custom learning rate scheduling.
- **Production Deployment:** Led the transition of research models to production, optimizing

via ONNX and quantization for real-time inference and implementing drift detection monitoring.

Senior Deep Learning Engineer | *Dec 2023 – Aug 2025 Spearheaded initiatives in Imitation Learning and Generative AI while building robust data infrastructure.*

- **Robot Transformer (Imitation Learning):** Engineered a production-grade Transformer policy with causal masking, enabling agents to generalize tasks from limited demonstrations (5-20 shots). Achieved **85%+ success rates** on unseen tasks, outperforming behavior cloning baselines.
- **RAG Knowledge System:** Designed an internal knowledge management system using **ChatGPT API**, **LangChain**, and **RAG**. Reduced research information retrieval time by **65%** via semantic search and prompt engineering.
- **Data Infrastructure:** Built a production-ready validation framework using GYM and ZeroMQ. Integrated ReRun API for interactive visual validation, ensuring data quality prior to training to prevent "garbage-in-garbage-out" failures.
- **Inference Optimization:** Applied knowledge distillation to reduce model size by **40%** while retaining **98%** accuracy, enabling feasible deployment on robot hardware.

Deep Learning Engineer | *Aug 2021 – Dec 2023 Full-stack ML development spanning research, production systems, and MLOps infrastructure.*

- **Semi-Supervised Learning:** Implemented a framework using pseudo-labeling and consistency regularization, reducing data labeling overhead by **60%** while maintaining accuracy within **2%** of fully supervised baselines.
- **Edge Computer Vision:** Developed a modular PyTorch object detection library (YOLO/Faster R-CNN). Achieved **45 FPS** on edge hardware through pruning and quantization; this library is now the backbone of the robot's perception system.
- **MLOps Transformation:** Overhauled GitLab CI/CD pipelines for HPC systems. Reduced iteration time from hours to minutes via Docker containerization, parallel scheduling, and smart caching.
- **Model Compression:** Deployed models to TensorRT and ONNX Runtime, achieving **3-5x inference speedups** on resource-constrained hardware.

Robert Bosch GmbH | Hildesheim, Germany **Master Thesis: Deep Learning for Multiple Object Tracking (MOT)** | *Sept 2020 – March 2021*

- **Novel Architecture:** Proposed extensions to the SHAMANN (Shared Memory Augmented Neural Networks) paradigm to improve tracking stability in crowded scenes.
- **Performance:** Achieved state-of-the-art stability metrics on **MOT17/MOT20** benchmarks while reducing computational complexity by **30%** through custom attention mechanisms.

Auvisus GmbH | Karlsruhe, Germany **Deep Learning Intern** | *Mar 2020 – Aug 2020*

- **Edge Optimization:** Optimized PyTorch classification models for mobile deployment, achieving **4-7x speedups** using ONNX Runtime and TensorRT while maintaining **>92% accuracy**.
- **Transfer Learning:** Utilized MobileNetV2 transfer learning to build high-accuracy classifiers with sparse data (<500 samples), overcoming significant labeling bottlenecks.

PATENTS & PUBLICATIONS

- **Patent:** P. Kandapagari, et al. "Animal Physical Parameter Estimation by Image Processing." *European Patent Application No. 23207432.8*, May 2024.
- **Publication:** A. Niemann, P. Kandapagari, et al. "Tissue Segmentation in Histologic Images of Intracranial Aneurysm Wall." *Interdisciplinary Neurosurgery*, Vol. 26, April 2021.

EDUCATION

M.Sc. Computer Science (Digital Engineering) | Otto-von-Guericke University Magdeburg
Focus: Deep Learning, Computer Vision, Neural Networks | *GPA: 1.8 (German Scale)*
B.Tech. Mechanical Engineering | JNTUA College of Engineering, Ananthapuram | 75%