Hemanth Kumar Gonela

📍 Hyderabad 500040, Telangana, India ☎ +91-7674904052 ✉ ghemanthkumar0001@gmail.com

EDUCATION

Vellore Institute of Technology, Vellore
Bachelor of Engineering, Computer Science (GPA: 8.28) *Vellore, Tamil nadu*

PROJECTS

Education Platform

- Led a team of two during a 36-hour hackathon to develop create an educational platform. Leveraged React, Tailwind CSS, JavaScript, and Firebase. The platform includes user authentication, chat rooms, file sharing, whiteboard snapshots, study materials, course overviews, profiles, and messaging. Our solution secured 1st prize in the competition.

Image dehazing using Encoder-Decoder architecture

- Developed an image dehazing model using an autoencoder architecture with 6 encoder and 6 decoder layers, effectively restoring clear images from hazy inputs.
- Trained the model on the Dense-Haze CVPR 2019 dataset, leveraging its high-quality image pairs to enhance the network's haze removal capability.
- Optimized the model for end-to-end image restoration, enabling efficient feature extraction and high-clarity image reconstruction with minimal distortions.

Text-to-Image Generation using LoRA-based Generative AI

- Developed and trained a text-to-image generative model using Low-Rank Adaptation (LoRA) to fine-tune a pre-trained diffusion model, enabling high-fidelity image synthesis from textual descriptions.
- Utilized large-scale datasets to train the model, optimizing latent space representations for improved text-image alignment, detail preservation, and realism.
- Implemented advanced conditioning techniques, including classifier-free guidance and cross-attention mechanisms, to enhance generation quality and coherence with textual prompts.
- Fine-tuned hyperparameters and incorporated adversarial training strategies to mitigate artifacts, ensuring high-resolution, photorealistic outputs with efficient inference.

Discord Server Moderation Bot

- Programmed a Discord bot to detect/filter inappropriate messages, manage users, monitor chat activity, send welcome messages, track server activity, and provide reports. It includes anti-spam measures, custom commands, poll creation, and server status updates.
- Improved server management efficiency by automating moderation tasks, reducing manual workload, enhancing user experience, and ensuring a safer and more organized community environment.

TECHNICAL SKILLS

- **Programming Languages**: C, C++, Python, JavaScript, HTML, Tailwind CSS, Next.js, FastAPI framework
- **Operating Systems**: Windows, Linux
- **Machine Learning**: Neural networks, Natural Language Processing, Computer Vision, Model Optimization, Reinforcement Learning
- **Certification:** Aws Cloud Practitioner, Eccouncil - Certificated Ethical Hacker (CEH)

PROFESSIONAL EXPERIENCE

Jio Platforms Limited | *Intern - Machine Learning Engineer* **Nov 2023 - May 2024**

- Developed real-time object detection models for a CCTV-based AI surveillance system, integrating ONNX Runtime and PaddleOCR to enhance vehicle and number plate recognition, later deployed commercially.
- Built a high-speed Python image scraper, extracting 500 images per minute, and developed an auto-annotator using YOLOv7, creating a 50,000-image dataset to improve the surveillance model.
- Worked on a real-world AI-powered baby monitor, leveraging stable diffusion models to generate human face datasets, aiding in its commercial deployment as an advanced vision-based monitoring tool.
- Optimized deep learning models, reducing inference latency by 35%, using quantization, pruning, and model distillation for seamless edge deployment.
- Enhanced object detection accuracy from 82% to 91%, fine-tuning YOLOv8 with self-supervised learning for improved real-time surveillance.
- Streamlined model deployment, cutting serving time from 20 min to under 5 min by integrating ONNX Runtime and TensorRT for faster rollouts.

RESEARCH EXPERIENCE

LLM model Compression

- Conducting research on a bipartite graph-based approach to optimize Large Language Models (LLMs) by reducing model size and computational time while maintaining performance, leveraging vocabulary pruning and transformer pruning to eliminate redundant parameters and improve efficiency.
- Developing an innovative compression model that integrates pipelined pruning, combining vocabulary and transformer pruning, along with structured pruning and sparsity-aware training, to enhance inference speed, lower energy consumption, and enable scalable real-time AI applications; research paper currently in the process of publication.