Krishna Teja Reddy Chinnakotla

+1 5516897755 chinnakotlaktr@gmail.com

Education

Stony Brook University **Aug 2021 – Dec 2022**
Master's in Computer Science *NY, USA*

Indian Institute of Technology, Bhubaneswar (IIT) **2015 – 2019**
Bachelor's in Computer Science *Odisha, India*

Technical Skills

Languages and Frameworks: Python, C++, Rust, SQL, Java, Django, Shell Scripting
Libraries: Pytorch, Pyspark, Tensorflow, Scikit-Learn, Langchain, Huggingface, OpenCV, spaCy, Pandas, NumPy, CUDA
Data Tools and Others: Hadoop, MapReduce, Grafana, Prometheus, NoSQL, Apache Spark, Airflow, Kafka, Redis, Cassandra, MongoDB, Hive, BigQuery, Elastic Search, Jenkins, Splunk, Git, Swagger, Docker, Kubernetes
Cloud: Google Cloud (GCP), Amazon Web Services (AWS)

Patents and Publications

- "Systems and methods for predicting and preventing patient collisions" - **Patent**.
- "Systems and Methods for AI based unconscious patient fall prediction." - **Patent**.
- "Selective Federated Transfer Learning using Representation Similarity" - **NeurIPS-SpicyFL 2020**.
- "Multi-modal estimation of the properties of containers and their content" - **IEEE Journal 2022**.

Experience

Petco **Jan 2024 -**
Senior Data Scientist (NLP, LLMs, Pytorch, AWS, GCP, MLOps) *Milpitas, CA*

- Lead cross-functional AI team driving enterprise-wide LLM adoption, architecting production AI systems with RAG infrastructure, AI agents and store employee assistants across retail locations, web and mobile.
- Built scalable AI infrastructure on Petco's GPU clusters using VLLM, optimizing inference pipelines for high throughput and cost efficiency while maintaining production stability.
- Developed and fine-tuned proprietary domain-specific Large Language Models (LLMs) using Petco's database, improving semantic search quality by 13% and personalized product discovery by 8%.
- Designed and implemented a production recommendation system using a microservice architecture with two-tower neural networks, deploying knowledge graph services with Redis caching and GraphQL APIs, improving inventory discovery by 14% while maintaining sub-100ms latency at scale.
- Built resilient ML pipelines for continuous model deployment with automated A/B testing framework, integrating model monitoring, feature stores, and CI/CD automation that scaled to 40K+ RPM with 99.9% availability while enabling rapid experimentation cycles for recommendation strategies.

Flexera Global **Nov 2023 – Jan 2024**
Software Developer (Python, Pyspark, Pytorch, GCP) *Remote, CA*

- Engineered robust, scalable data pipelines for ML systems handling terabytes of image and video data on GCP, using Python and PySpark for ETL processes, and orchestrated data certification to enhance model training efficiency.

Bonsai Robotics **Jul 2023 – Aug 2023**
Senior Computer Vision Engineer (Python, C++, Pytorch, MLFlow, AWS) *San Jose, CA*

- Designed and developed an automatic labeling pipeline by integrating Grounded DINO and SAM, allowing for automatic annotations using label word and refining annotations using CLIP scores for high quality labelling.
- Working towards building collision avoidance safety features for an autonomous vehicle. Implemented multiple deep learning architectures and classical image processing techniques to tailor the system specifically for dusty agricultural field scenarios, ensuring safety and efficiency.
- Enabled the continuous end-to-end training workflow of Machine Learning models by automating large data collection, data transformation and data feeding to train the models using Python, AWS.

Oishii **Jan 2023 – Jul 2023**
Senior Robotics Software Engineer (LLMs, Pytorch, AWS, Kubernetes, Kubeflow, MLOps) *Jersey City, NJ*

- Created a Chatbot leveraging Langchain for preprocessing, LLM (GPT) for context-aware response generation, and FAISS for textual emphasis; seamlessly integrated with backend, enabling dynamic context-based interactions.
- Developed traditional deep learning models and advanced vision transformer models for classification, 3D segmentation, object detection, tracking and key point detection. Used data augmentation methods to handle data imbalance.

- Leveraged Multi-GPU and Distributed deep learning techniques to train all the above models. Skillfully deployed scalable machine learning models onto the AWS to build an efficient ag-tech AI system.
- Automated ML workflows to accelerate data labelling, data preparation, model building, training, and experiments. Primarily responsible for maintenance of critical pipelines that support analytic systems and A/B experiments.
- Built APIs and continuous integration and delivery (CI/CD) pipelines to reduce model management overhead. Monitored quality of ML models by automatically detecting bias, model drift, and concept drift.

General Electric (GE) Healthcare **July 2019 – July 2021**

Software Engineer - AI *Bangalore, India*

- Architected and implemented a Natural Language Procesing (NLP) based medical semantic search engine to extract relevant information of the patient records. Developed deep language models using Transformers (BERT) for information extraction and semantic search in clinical notes.
- Built Machine Learning systems for patient safety in ICUs such as predicting patient fall from hospital beds and patient collision with medical apparatus. Designed deep learning architectures to tackle multimodal data and detect, monitor and analyse patient behaviours, patient poses and their restlessness in hospitals. This work is part of **two US patents**.
- Performed post training model optimization such as model pruning, model quantization(half precision) and deployment on target hardware using C++, TensorRT and ONNX.
- Created training jobs on AWS Sagemaker using Scikit-learn, Pytorch and deployed them using Python, Flask on AWS Elastic Beanstalk. Simplified the deployment process using Kubeflow and KFServing. This helped the team to fasten the training, deployment process and perform hyper-parameter tuning very efficiently.
- Devised predictive analytics systems for medical imaging products (MRI, X-ray) to predict failure and reduce unplanned equipment downtime. Experimented with models such as Random Forest, XGBoost and Stacking classifer. Used Regression Modeling and LSTM techniques to predict typical repair times and distress thresholds.
- Compiled and cleaned multimodal health data from disparate sources (large databases, APIs, flat files) to aid Health Sciences Strategy team with their projects and engagements. Optimized and automated data pipelines and created new pipelines using Hadoop, Spark and AWS that performed the necessary data quality checks and anomaly detection steps.

Orchard Robotics **June 2022 – Aug 2022**

Machine Learning Engineer Intern *Ithaca, NY*

- Built an end-to-end Computer Vision pipeline from data acquisition, data ingestion to inference and model monitoring for apple detection in apple orchards.
- Experimented with multiple architectures like YoloV5, Faster RCNN and Mask RCNN. Implemented the Ellipse RCNN model from scratch for accurate detection of occluded apples as well as calculation of apple volume/size.
- Developed an efficient 3D clustering method for fruit localization and counting. Built dashboards for visualization of statistics such as fruits count, fruit sizes and weights distribution and fruit yield prediction.
- Created training jobs on AWS leveraging Pytorch, OpenCV, OpenVINO and ONNX. Programmed web services and monitored them using Flask, MongoDB, Docker, Kubernetes.

Projects

OpenMined **Mar 2020 – Dec 2020**

Open Source Research Engineer (Privacy ML) *Remote*

- Proposed a framework, Selective Federated Transfer Learning(SFTL) to address the problem of source model selection during transfer learning in Federated scenarios. **Published this work at NeurIPS-SpicyFL 2020.**
- Experimented on CNN based architectures using Pytorch and PySyft by leveraging the concepts of representation learning to develop SFTL, which provides accurate selection and transfer of model parameters on the edge devices.

Multi-modal AI approach to Analyze Unseen Containers (Vision + Audio)

- Secured 4th position in Multi-modal fusion Corsmal challenge, 2020 which involved estimation of capacity, filling level and filling type of unseen containers. Published this work as part of a **Journal in IEEE Transactions on Multimedia**.
- Leveraged MFCC based audio features and CNN model for sound based filling type classification. Developed container capacity estimation method with RGB-D data's 3D point cloud and Mask R-CNN.

Achievements

- Secured All India Level Rank 12 in SIMO, South Indian Mathematics Olympiad.
- Got awarded prestigious Young Scientist Promotion Fellowship (KVPY) by IISc, Bangalore.
- Recipient of National Talent Search (NTSE) Scholarship by The National Council of Educational Research and Training (NCERT),Government of India.