Deepa K

Pleasanton, CA · deepausc6@gmail.com · 925-236-0197 ·
https://www.linkedin.com/in/deepa-kiran-a4862322b/details/experience/

Education

University of Southern California — Los Angeles, CA
MS Computer Science

Work Experience

Lucid Motors — Newark, CA
Senior Machine Learning Engineer — Oct 2022 - Present

- Developed chatbots which use Generative AI (GenAI) and Large Language Models (LLMs) to deliver personalized and context-aware responses, enhancing user engagement and satisfaction.
- Fine-tuned LLMs using domain-specific datasets to optimize the chatbot's performance, ensuring high accuracy and relevance in conversations.
- Implemented a Retrieval-Augmented Generation (RAG) framework to improve the chatbot's response generation by integrating external knowledge sources, thereby enhancing the richness of the interaction.
- Utilized Vector Databases to store and retrieve embeddings efficiently, facilitating quick and accurate similarity searches for improved user query handling.
- Designed and executed robust model evaluation protocols, including A/B testing and user feedback integration, to continuously refine and improve the chatbot's performance.
- Deployed the chatbot on AWS using SageMaker for scalable model training and inference, ensuring reliable and efficient performance under varying load conditions.
- Used Databricks for managing the entire ML lifecycle, from data preprocessing and model training to deployment, monitoring, and maintenance, ensuring a seamless end-to-end workflow.
- Integrated Ray for distributed model training and hyperparameter tuning, significantly reducing training time and improving model performance through parallel processing.
- Utilized LangChain for advanced conversational capabilities and workflow automation, enhancing the chatbot's ability to manage complex dialogues and tasks.
- Collaborated closely with data scientists and software engineers to align on model requirements, data pipelines, and deployment strategies, fostering a cohesive and productive development environment.

CIGNA Health Insurance — Phoenix, AZ
Machine Learning Engineer — Jan 2021 - Sept 2022

- Developed, deployed, and maintained machine learning models using PyTorch for real-time applications, leveraging advanced deep learning techniques to drive business outcomes.
- Conducted extensive data collection using web scraping tools like BeautifulSoup and Scrapy, and integrated APIs via Requests and RESTful services, storing the data in MySQL and MongoDB databases, as well as data lakes like Amazon S3.
- Performed comprehensive data preparation and processing using pandas and NumPy for data manipulation, OpenRefine for data cleaning, and Scikit-learn for feature engineering, managing ETL pipelines with Apache Airflow and Apache NiFi.
- Executed thorough exploratory data analysis (EDA) utilizing Matplotlib, Seaborn, and Plotly for data visualization, along with statistical analysis using SciPy and Statsmodels to uncover insights and inform model development.
- Developed and fine-tuned deep learning models with PyTorch, employing experiment tracking tools like MLflow and TensorBoard, and optimized model performance using hyperparameter tuning frameworks such as Optuna and Ray Tune.
- Trained models on distributed systems using PyTorch Distributed and Horovod, leveraging GPU acceleration technologies like NVIDIA CUDA, cuDNN, and TensorRT to enhance training efficiency and model performance.
- Evaluated model performance using Scikit-learn and PyTorch-Ignite for metrics calculation, applied cross-validation techniques, and conducted rigorous hyperparameter tuning with Optuna and Ray Tune to ensure model robustness and reliability.

- Deployed machine learning models on cloud platforms including AWS SageMaker, Google AI Platform, and Azure ML, utilizing Docker for containerization, Kubernetes for orchestration, and TorchServe for scalable model serving.
- Implemented comprehensive monitoring and logging solutions using Prometheus and Grafana for performance monitoring, and the ELK Stack (Elasticsearch, Logstash, Kibana) for logging and troubleshooting in production environments.
- Collaborated with cross-functional teams, including data scientists, software engineers, and product managers, utilizing tools like Jira and Trello for project management, and ensuring seamless integration and API development with Flask and FastAPI.

E-Trade San Fransisco, CA
Data Engineer Sept 2019 - Dec 2020

- Designed and implemented data ingestion pipelines using Apache NiFi and Talend, ensuring efficient extraction, transformation, and loading (ETL) of data from diverse sources.
- Built and managed real-time data streaming applications with Apache Kafka and Apache Flink, enabling real-time data processing and analysis.
- Developed custom Python scripts and used Scrapy for web scraping and API integration, automating the collection of data from various web services and websites.
- Optimized data storage solutions using PostgreSQL, MySQL, and MongoDB, balancing structured and unstructured data storage requirements for high performance and scalability.
- Used and maintained data warehouses with Amazon Redshift and Google BigQuery, providing scalable storage and fast query performance for large datasets.
- Used Apache Spark and Hadoop to process and analyze large datasets, achieving significant improvements in data processing efficiency and scalability.
- Implemented batch processing workflows using Apache Beam and Apache Flink, facilitating efficient and reliable data processing and transformation.
- Managed complex data pipelines with Apache Airflow, automating workflows and ensuring timely data availability for downstream processes.
- Deployed containerized applications using Docker and Kubernetes, enhancing the scalability and reliability of data processing environments.
- Utilized AWS Glue and AWS EMR to build and orchestrate data pipelines on AWS, integrating machine learning models and advanced analytics into data workflows for comprehensive data engineering solutions.

Frontier Airlines Denver, CO
Senior Data Analyst Jan 2016 - Aug 2019

- Led end-to-end development of NLP projects using TensorFlow, encompassing data collection, preprocessing, EDA, text representation, model building, training, evaluation, deployment, and monitoring, resulting in high-performing and scalable NLP solutions.
- Implemented web scraping pipelines using BeautifulSoup and Scrapy to collect large-scale text data from various online sources and stored the data in a PostgreSQL database.
- Developed text preprocessing scripts with TensorFlow Text to handle tokenization, normalization, and vectorization of raw text data, ensuring consistent input format for model training.
- Conducted exploratory data analysis (EDA) using pandas and Seaborn to visualize text data distributions, identify patterns, and detect outliers, leading to data-driven decisions for model improvements.
- Utilized pre-trained word embeddings like GloVe and Word2Vec, and integrated transformer-based embeddings from TensorFlow Hub to create dense vector representations of textual data for enhanced model performance.
- Designed and built various NLP models using TensorFlow and Keras, including LSTM, GRU, and transformer-based architectures, tailored for tasks such as text classification, named entity recognition, and sentiment analysis.
- Executed extensive hyperparameter tuning using Keras Tuner and TensorFlow Model Optimization Toolkit to enhance model accuracy and reduce overfitting in NLP tasks.

- Evaluated model performance with TensorFlow Metrics, employing metrics like accuracy, precision, recall, and F1-score, and visualized results using confusion matrices and ROC curves.
- Deployed trained NLP models using TensorFlow Serving, creating RESTful APIs with Flask to enable real-time inference and scalable deployment in production environments.
- Implemented monitoring and logging systems with Prometheus, Grafana, and TensorBoard to track model performance, detect anomalies, and facilitate model maintenance and updates.
- Developed intuitive web applications using Flask and React to provide user-friendly interfaces for interacting with deployed NLP models.

PeerVU, Inc San Francisco, CA
Data Analyst Feb 2013 - Dec 2015

- Developed automated data pipelines using SQL, Apache Airflow, and APIs to collect and integrate data from multiple sources, ensuring timely and accurate data availability for analysis.
- Managed and optimized relational and NoSQL databases, including PostgreSQL, MySQL, and MongoDB, to support data storage and retrieval operations, improving query performance by 30%.
- Designed and implemented ETL workflows using Talend and Python scripts, facilitating seamless data transformation and loading into Snowflake and Google BigQuery data warehouses.
- Optimized data storage solutions using PostgreSQL, MySQL, and MongoDB, balancing structured and unstructured data storage requirements for high performance and scalability.
- Used and maintained data warehouses with Amazon Redshift and Google BigQuery, providing scalable storage and fast query performance for large datasets.
- Used Apache Spark and Hadoop to process and analyze large datasets, achieving significant improvements in data processing efficiency and scalability.
- Implemented batch processing workflows using Apache Beam and Apache Flink, facilitating efficient and reliable data processing and transformation.
- Managed complex data pipelines with Apache Airflow, automating workflows and ensuring timely data availability for downstream processes.
- Deployed containerized applications using Docker and Kubernetes, enhancing the scalability and reliability of data processing environments.
- Utilized AWS Glue and AWS EMR to build and orchestrate data pipelines on AWS, integrating machine learning models and advanced analytics into data workflows for comprehensive data engineering solutions.