Anup Raj Niroula

+1(646)463-9965 a.r.niroula@gmail.com Anup Raj Niroula ARNiroula

Education

New York University, New York, USA **August 2024 – Ongoing**
Master in Computer Engineering (M.S.) *Current CPGA: 4.0/4.0*

Kathmandu University, Dhulikhel, Nepal **August 2017 – March 2022**
Bachelor in Computer Engineering (B.E.) *CGPA: 3.83/4.0*

Relevant Skills

Programming Languages

- Python, Golang, Javascript, Typescript, C, C++, R, SAS 9.4

Frameworks

- Web Framework — FastAPI, Django, Flask, Express.js, React, Next JS
- Data Science — PyTorch, Tensorflow, Pandas, Polars, Scikit-learn, Tidyverse, data.tables
- Database — PostgreSQL, MySQL, Redis, MongoDB, Cassandra

Tools

- Git, Github, Bitbucket, AWS, GCP, Docker, Kubernetes, Kafka, RabbitMQ, Jenkins

Experience

MedLaunch Concepts **May 2025– Present**
AI Engineer Intern *Florida, United States*

- Integrated an AI-powered chatbot using Retrieval-Augmented Generation via the LangChain framework, which reduced manual query response time by approximately 80%
- Formulating a scalable document parsing mechanism using AWS services, currently managing 1,000+ concurrent parses on average

New York University **January 2025– May 2025**
Graduate Teaching Assistant *New York, United States*

- Facilitated office hours to answer questions, explain concepts, and provide additional support to students who needed help understanding course material
- Revised machine learning assignments by integrating real-world datasets and clearer instructions, leading to a 20% improvement in student scores and positive feedback

Kathmandu University **December 2023– April 2024**
Teaching Assistant *Kavre, Nepal*

- Conducted lab sessions for Computer Programming in C for over 120 students
- Demonstrated flexibility in adapting curriculum based on real-time performance data, and enabling the instructor to deliver 96% of planned lessons

Nimble Clinical Research **October 2021– May 2023**
Software Engineer *New Jersey, United States*

- Developed and maintained communication between Django and R service
- Led the R Programming Team and managed a team of 6 people
- Engineered Report Automation System, which could output reports of more than 100000 rows within a minute. Reduced report generation time by approx. 9x on average
- Trained a Supervised Learning Model for automated clinical data mapping, achieving 70%+ accuracy, improving efficiency in data standardization for clinical trials

Achievements

- 1st Place at Open Software Competition IT MEET 2018
- 1st Place at Fourth SNB Memorial Inter-School Science Quiz Contest

Projects

Image Classification using Custom ResNet Model | Python, Pandas, PyTorch, Tensorflow
Github

- Created a Custom ResNet Model for image classification with around 5M parameters using the CIFAR-10 dataset
- Achieved test accuracy of more than 96%
- Secured top 20% position in CIFAR 10 Classification Kaggle competition using the Custom ResNet model

Spotify Buddies: an organic playlist recommender | MLflow, FastAPI, Prometheus, Grafana, Apache Airflow, Docker
Github

- Trained and deployed a collaborative filtering recommendation model on Chameleon Cloud, generating up to 200,000 predictions per minute while maintaining high accuracy
- Redesigned the deployment pipeline, reducing model server launch time from 300s to 30s, and keeping the server image size under 2 GB
- Integrated real-time monitoring using Grafana and Prometheus to visualize model performance metrics and detect cold-start scenarios
- Automated retraining and model management using MLflow and Apache Airflow, improving system scalability and maintenance

Parameter Efficient Finetuning of LLMs with Low-Rank Adaption for Custom News Data | Pytorch, Transformers, Tens
Github

- Fine-tuned LLMs using LoRA on news datasets with HuggingFace Transformers
- Applied domain adaptation techniques to enhance performance on target-specific news data
- Achieved test accuracy of more than 93%
- Secured 1st place in an internal Kaggle-style competition on LoRA-based fine-tuning organized by New York University