Rohit Gupta

Delhi, India | +91 886-011-7496 | rohitgr1998@gmail.com | linkedin/rohitgr7 | github/rohitgr7 | twitter/imgrohit | kaggle/rohitgr

Professional Experience

Shopadvisor AI (Full-Time) Remote
Founding ML Engineer *Aug 2023 - Current*

- Shopadvisor AI: Building an AI shopping advisor for e-commerce companies using LLMs.
- Leading everything in product and engineering.

Mazaal AI (Part-Time) Remote
ML Lead *Dec 2022 - July 2023*

- Mazaal AI: Building a no-code ML platform for users to build, train and deploy their ML models.
- Created all kinds of pipelines in the image and text domain and connected these services with AWS and Runpod for deployment.
- Deployed ML models on serverless Runpod platform as inference APIs.
- Deployed zero-shot models for labeling tasks such as image object detection and segmentation.
- Added few-shot pipelines for text-related tasks to enable training models with a small amount of data.

Freelance Remote
ML Engineer *March 2023 - July 2023*

- A platform for users to create chatbots using documents and website content using Rafiq AI.
- A chatbot for US citizens to get updated information of their respective townships. Implemented the whole backend and LLM integration: Munichat.

Lightning AI (Formerly PyTorch Lightning) Remote
Research Engineer (L3) *Aug 2022 - Dec 2022*

- Muse App: Deployed a text-to-image generation model (stable-diffusion) in production, which handles thousands of concurrent users. Implemented a load balancer with dynamic batching to improve the model serving performance from 10 to 500 concurrent users using micro-services. On the launch, the App hit around 8000 requests in 2 days without any failures.
- Led the stable diffusion research with 2 team members, where we explored the limitation of the Lightning framework when a foundational model is deployed in production.
- Collaborated with the Colossal AI team to integrate the Colossal AI engine that implements different parallelism algorithms that are especially interesting for developing SOTA transformer models.
- Additional notable contributions were Tuner callbacks, FSDP auto-wrappers, and DeepSpeed integration improvements.
- All of my contributions to the framework are available here.

Research Engineer (L2) *Oct 2021 – Aug 2022*

- The project I worked on and maintained is lightning which is open-source and currently has 20k stars and 7510+ contributors. By the time I joined, I was already in the top 10 as a core contributor, and now I am in the top-5. It is built on top of PyTorch from Meta.
- Advise on, assist, and support the implementation, optimization, and development of those improvements and new technologies in the framework, along with timely releases.
- Integration and maintenance of SOTA deep-learning optimization techniques developed by research labs into the framework.
- Ensured proper maintenance of documentation that is completely user-facing by ensuring everything is updated and explained with real-world examples.
- Ensured continuous community engagement and discussions that helped increase the reach of the framework and helped in its development.
- Also did a livestream on how to make their first contributions to OSS :)

Episource LLC *Chennai, India*
Associate Data Scientist *July 2020 – Sep 2021*

- Slashed the processing time from days to an hour by building a complete automated pipeline to generate patient profiles for HRA along with model deployment on AWS and GitHub CI/CD pipeline.

- Collaborated in the in-house development of the Datalake warehouse using PySpark. This helped to ensure data integrity and correctness in various verticals of business where it is being used.
- Analyzed the HRA and Telehealth services thus reducing the cost incurred, operation optimization, and managed the planning & deployment of design and procedures for weekly metric reports. The analysis helped in taking major business decisions to improve the processes. Lead the team with 2 interns.

Data Science Intern *Jan. 2020 - June 2020*

- Improved the contact rate for OCs by 30% to non-approachable customers by developing an automated pipeline to mine historical data and find relevant contact information for upcoming/ongoing projects.
- Built a machine learning model that can predict the outreach consultant headcount for an incoming project.

Projects

Lightning AI (Open Source) | *Python, PyTorch, Git* | *Link* *June 2020 – Sep 2021*

- A lightweight PyTorch wrapper for high-performance AI research. Working as a core maintainer, fixing bugs and adding new features.

Earwise | *Python, Pytorch, Whisper, Git* | *Link*

- An application that lets you search within Audio files and YT videos. The backend is deployed on GCP and frontend is based on streamlit.

GitMate | *Python, OpenAI, Git* | *Link*

- GitMate is your companion to generate commit messages, PR titles and descriptions using ChatGPT.

PotterHead | *Python, Git, ChatGPT, Streamlit* | *Link*

- An application that lets you ask questions to Harry Potter, about his life. This is built using ChatGPT and frontend is based on streamlit.

Triton cc | *Python, Pytorch, Triton, Docker, Git* | *Link*

- It's a collection of notebooks for beginners to understand the Triton Inference server and how to deploy ML models efficiently.

TVmodels | *Python, Pytorch, Git* | *Link*

- Python implementation of various vision models. They are readily available for use as a PyPI package.

Education

Maharaja Agrasen Institute of Technology *Delhi, India*
Bachelor of Technology (CSE) | 7.92 CGPA *Aug. 2016 – May 2020*

Technical Skills

Languages: Python, SQL, Bash, C/C++
Frameworks/Libraries: PyTorch, Pandas, NumPy, Matplotlib, Seaborn, PySpark, Scikit-learn, LightGBM
Developer Tools: Jupyter, Git, Docker, AWS, VS Code

Achievements

- Top 3% out of 3,115 teams in Kaggle's Intruder Detection through Webpage Session Tracking.
- Ranked 40 out of 1,600 participants in mlcourse.ai 2019 session.
- Top 10% out of 4,127 teams in Kaggle's Elo Merchant Recommendation Challenge.
- Top 12% out of 2,038 teams in Kaggle's NFL Big Data Bowl Challenge.
- Kaggle notebooks expert (top 1% out of 166,000 members).

Publications

- Gupta, R. and Sharma, A. Super-Resolution using GANs for Medical Imaging Procedia Computer Science 173 (2020): 28-35. | *Link*