Pranjal Datta

Machine Learning Engineer Background in ML Inference systems and Tooling, DAG Orchestration workloads and Model Optimisation

Core Skills

- *Programming Languages*: Golang, Python (Proficient), C/C++ (Intermediate)
- *Primary Skills*: ML workload orchestration, MLOps Tooling, System and API Design
- *Databases*: SQL and SQL-like Databases
- *Software/Tools*: Argo Workflows, Linux, Docker, Kubernetes, Pytorch, PostgreSQL, GRPC

Professional Experience

Pixxel Space, Bangalore, India

Software Engineer, Machine Learning; *June, 2022 - Present.*

- **ML Inference Orchestration Engine** - Lead technical efforts to build the system from scratch.
  - Supports execution of *independently developed internal and external models* as *an user-defined dynamic* DAG workflow.
  - Introduced *Type System*, *Async Processing*, *Artifact Management*, *model version management* and *Realtime Monitoring*.
  - System manages *20+ unique models*, **200+ unique workflows and 90+ hours of execution time per week**, and *30+ GB of model data produced per week*.
- **Model Development Python SDK**: Built from scratch and currently, the *de-facto* tool used by internal and external customers.
  - Supports multiple execution modes, *local*, *async remote* and *sync remote*.
  - SDK handles *type inference and validation*, *artifact management* for different execution modes at runtime.
  - SDK allows users to one-click deploy models across multiple environments and manage multiple versions.
  - As a result, client model code is fundamentally simpler. Reduced **lines of code by 80%**.
- Deployed in-house *monitoring and analytics* tooling with Grafana, Prometheus and OpenTelemetry tools.
- Optimised both *Deep Learning* and *Physics* based models to extract maximum performance and reliability with minimal resources.
- **High Level Software stack**: Golang, Python, Argo Workflows, Kubernetes, GRPC, Pytorch, NVIDIA SDKs, Terraform.

Mad Street Den, Chennai, India

Machine Learning and Platforms Intern; *Jan, 2022 - June, 2022.*

- Working to optimize a distributed platform that enables custom orchestration of ML/Data workloads and make it more reliable, scalable and performant. Increased request-resolution time by 27%
- Increased turnaround time of the platform scheduler by 20%
- Optimized platform costs by introducing new scaling features for deployed applications.

Synopsys Inc, Mumbai, India

Technical Engineering Intern; *May, 2021 - Jan, 2022*.

- Designed novel ML-powered cloud-native service to identify bottlenecks during chip manufacturing, processing over 100+ GB of data in realtime.
- Rolling out the complete feature into production as an elegant, distributed software utilizing Apache Spark with request-resolution time ~11% faster than legacy systems for large datasets.

Open Source

- Pytorch-Lightning [25k+ stars]: Contributed image processing metrics (code+docs+tests) here and here.

Education

SRM Institute of Science and Technology, Chennai, India

Bachelor of Technology in Computer Science and Engineering, **GPA: 9.69 / 10.0**.

Projects

- PyVision: Built a toolbox of the latest computer vision algorithms, ready to use in 3 lines of code to reduce development time and allow for quick iteration.
- Image2Sketch: Utilized Generative Adversarial Networks (GANs) to convert Selfies into artist sketches by training them on a custom dataset built specifically for the project.
- DenseDepth-PyTorch: Implemented the research paper to reproduce and verify the results of depth-estimation in 2D while modularizing the code following OOP principles for simpler usage.

Competitive Achievements

- **HackerEarth Hackathon, 2020** - Placed 3rd among 75+ teams. Designed an almost no-code solution to quickly deploy ML Models onto the cloud for rapid sharing and collaboration.