Google Summer of Code 2025 Proposal

Project Title:

Optimizing AI Model Deployment for Open-Source Applications

Personal Information:

Name: Ikram Elahi Hashmi

GitHub: github.com/ikramhashmi

LinkedIn: linkedin.com/in/ikram-hashmi

Email: ikramhashmi1111@gmail.com

Project Summary:

Your AI models in open-source applications is often challenging due to high computational costs, inefficiencies in model inference, and difficulties in scaling deployments. Many developers face obstacles when transitioning trained models from research environments to real-world production systems. This project aims to optimize AI model deployment by developing a lightweight, scalable, and resource-efficient framework. Leveraging my experience in machine learning, deep learning, and model deployment, as seen in my work on brain tumor detection, spam detection, and ML pipelines, I will implement techniques such as model quantization, pruning, and TensorFlow Lite optimization to enhance performance. Additionally, I will integrate Flask/FastAPI for API-based deployment, Docker for containerization, and CI/CD pipelines for automated updates. The final deliverables will include a fully optimized deployment framework, APIs for seamless integration, Dockerized setups, performance benchmarks, and comprehensive documentation. This project will significantly benefit the open-source community by making AI deployment more efficient, scalable, and accessible.

Benefits to the Open-Source Community:

This project will help open-source developers by providing an optimized, scalable, and easy-to-integrate AI deployment framework. Many AI models fail to transition from research to production due to inefficiencies in inference and high resource consumption. By implementing optimization techniques such as quantization and model pruning, this framework will enable developers to deploy AI models more efficiently. The use of Flask/FastAPI and Docker will ensure seamless API-based deployment, making AI accessible to a wider range of applications and platforms. Furthermore, integrating CI/CD pipelines will automate deployment, reducing manual effort and ensuring up-to-date models.

Deliverables & Timeline (Weekly Plan)

- **Week 1-2:** Understanding organization requirements, finalizing project scope, setting up development environment.
- **Week 3-4:** Research and implement model quantization and pruning techniques.
- **Week 5-6:** Develop API-based deployment using Flask/FastAPI.
- **Week 7-8:** Implement Docker-based containerization and optimize performance benchmarks.
- **Week 9-10:** Integrate CI/CD pipelines for automated model updates and deployment.
- **Week 11-12:** Conduct extensive testing and validation of deployment framework.
- **Week 13-14:** Write detailed documentation, tutorials, and final project submission.

Technical Details:

- **Programming Languages:** Python
- **Machine Learning Frameworks:** TensorFlow, Keras, Scikit-Learn
- **Model Optimization Techniques:** Quantization, Pruning, TensorFlow Lite
- **Deployment & APIs:** Flask, FastAPI
- **Containerization & Scaling:** Docker
- **Model Evaluation & Performance Tuning:** NumPy, Pandas, Matplotlib
- **Version Control & Collaboration:** Git, GitHub
- **Automation & CI/CD:** GitHub Actions, Docker Compose

Why Me?

I have a strong background in AI, machine learning, and deployment frameworks, with hands-on experience in various ML projects, including brain tumor detection, spam detection, and ML pipeline creation. My expertise in model optimization, API development, and cloud deployment makes me an excellent candidate for this project. Additionally, my experience with Flask, Docker, and CI/CD automation ensures that I can deliver a scalable and efficient AI deployment solution for open-source applications. I am passionate about contributing to open-source communities and am eager to leverage my skills to enhance AI model deployment efficiency.