Diffusion Model API: A Full-Stack ML Deployment Project

Abrar Zahin

This project demonstrates the end-to-end deployment of an image generation service using a diffusion model. It uses the pre-trained Stable Diffusion model `CompVis/stable-diffusion-v1-4` from Hugging Face to convert text prompts into realistic images.

The application is served via FastAPI, a modern and fast web framework for building APIs with Python. A single POST endpoint `/generate` accepts a JSON payload containing a text prompt, which is passed to the model. The output is a PNG image returned as a streaming HTTP response.

To ensure environment consistency and portability, the application is packaged using Docker. The Dockerfile sets up the environment, installs required packages like `diffusers`, `torch`, and `Pillow`, and starts the FastAPI server with Uvicorn.

For real-world deployment, Kubernetes manifests are included to create:

- A `Deployment` with two replicas of the containerized API
- A `Service` of type `LoadBalancer` for external access
- A `Horizontal Pod Autoscaler (HPA)` to automatically scale pods based on CPU usage

The system is designed to simulate production environments where large-scale, on-demand generation of images is required. While this version runs on limited hardware (e.g., Google Colab), the architecture is robust and scalable to high-performance cloud infrastructure with GPUs, supporting much larger models like 7B or 70B parameter LLMs.

Overall, the project demonstrates practical experience in machine learning model serving, API development, containerization with Docker, and orchestration with Kubernetes. It highlights the ability to move beyond model training to full production-grade deployment pipelines.