Gradio Chatbot and Hugging Face Spaces App Husting

In the Chatbot interface design, after the research, there are several different approaches or solutions for online hosting option. Gradio chatbot interface and Hugging Face Spaces Demo have been selected for the project. Gradio is a Python library that allows users to easily create user interfaces (UI) for machine learning models. Its primary goal is to provide a simple way to showcase and interact with models, often for the purposes of debugging, demonstration, or fine-tuning. Gradio can be particularly helpful when User wants to share their model with non-technical users or quickly prototype an interface without delving deep into front-end development (PyPI., n.d.).

Some key features and uses of Gradio include:

1. Easy Integration: Users can wrap almost any machine learning model with a Gradio interface in just a few lines of code.
2. Support for Various Inputs and Outputs: Gradio offers a wide range of built-in input and output interface components, such as text boxes, images, audio clips, and more.
3. Interpretability: Gradio has features that allow for model interpretability, letting users see how different inputs affect model outputs.
4. Web Interface: Once the user has set up a Gradio interface, it launches a web app that the user can share with others. Users can interact with the model directly through this web app.
5. Sharing and Deployment: Users can easily share their Gradio app with others using a public link, and Gradio can host this link.
6. Integration with Popular Frameworks: Gradio can be used with popular machine learning and deep learning frameworks like TensorFlow, PyTorch, and scikit-learn (PyPI., n.d.).

With a study on Ograbek, K.’s YouTube video and Colab notebook (2023), he has pointed out how to handle the history message and system prompt that shows the full potential of the chatbot. This project has adopted his code for formatting the message with history and system prompts. Also, this project has learned pipeline options from his Colab notebook code.

The Chatbot interface can be run with T4 GPU, where it can use Hugging Face Spaces hosting, which we set up for the project demo. Hugging Face Spaces, GPU hosting is $0.60 per hour. The APP build time is around 20-25 minutes. To ensure the user experience, the suggestion is to keep GPU hosting running continuously for the user. There is not much difference between the Colab notebook and a running APP. Colab notebook is split into the requirement file for all the dependent libraries and the APP file for the function. As long as the Hugging Face Spaces is created under the Gradio environment, then all the setup is completed.

Reference:

Ograbek, K. (2023, September 7). *How to Create Llama 2 Chatbot with Gradio and Hugging Face in Free Colab* [Video]. YouTube. <https://www.youtube.com/watch?v=lSBX-nMQ8cE> Colab notebook: <https://colab.research.google.com/drive/1SSv6lzX3Byu50PooYogmiwHqf5PQN68E>

PyPI. (n.d.). Gradio. from <https://pypi.org/project/gradio/>