LLaMA Locally on CPU

1. Prerequisites

Before proceeding with the setup, ensure that the following software and tools are installed on your system:

1. **Miniconda or Anaconda**: Used for managing virtual environments.
2. **Python 3.7+**: Required for running Python code.
3. **Visual Studio Code (VS Code)**: Code editor to write and execute Python code.

2. Setting Up the Environment

Follow these steps to set up a new Conda virtual environment:

1. Open the terminal (or Anaconda Prompt).
2. Create a new Conda environment with the following command:

   `conda create --name llms2`

3. Activate the environment:

   `conda activate llms2`

3. Installing Dependencies

After activating the environment, install the required dependencies by running the following commands:

1. Install `pip` in the Conda environment:

   `conda install pip`

2. Install the necessary packages:

   `pip install --upgrade llama-cpp-python==0.1.78`

   `pip install langchain`

   `pip install langchain-community`

4. Running the Code

Once the environment is set up and dependencies are installed, follow these steps to run the code:

1. Open VS Code and set the Python interpreter to the Conda environment (`llms2`).
2. Create a new Python file (e.g., `llama_model.py`) and copy the provided code into the file.
3. Run the file by opening the terminal and using the command:

   `python llama_model.py`

5. Understanding the Code

The code uses the Langchain library to interact with the LLaMA model. Here's a breakdown of the code:

1. **Callbacks**: `CallbackManager` and `StreamingStdOutCallbackHandler` are used to handle token-wise streaming and display the output.

2. **LlamaCpp**: This is the interface to the LLaMA model, initialized with the model path and other parameters like `temperature`, `max_tokens`, and `top_p`.

3. **Question**: The model generates a response to the question: 'What is the largest country on Earth?'

4. **Response**: The response is printed to the terminal.

5. **Model Path**: Ensure that the model path points to the correct file location on your system.

6. Expected Output

When you run the code, the model will generate a response based on the input question. An example output may look like:

```

Response: The largest country by land area is Russia.

```

Note: The model may return different answers based on the training and context provided.

Step-by-Step Guide to Set Up the Environment

Step 1: Install Miniconda (if not already installed)

1. **Download Miniconda**:
   - Go to the Miniconda download page and download the installer for your operating system.
2. **Install Miniconda**: Follow the installation instructions for your system.

Step 2: Install VS Code

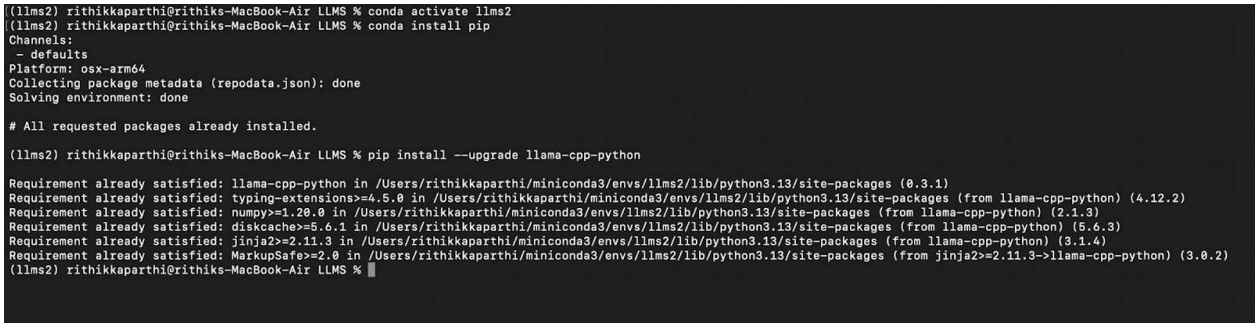
1. **Download VS Code**:
   - Visit the VS Code download page and install it for your operating system.
2. **Install Python Extension in VS Code**:
   - Open VS Code, go to the Extensions view (`Cmd+Shift+X` on Mac, `Ctrl+Shift+X` on Windows), and search for **"Python"**.
   - Install the official Python extension by Microsoft.

Step 3: Create Conda Environment

Open your terminal and **create a new Conda environment**:

```
conda create --name llms
```

```
(base) rithikkaparthi@rithiks-MacBook-Air ~ % cd LLMS
(base) rithikkaparthi@rithiks-MacBook-Air LLMS % code .
(base) rithikkaparthi@rithiks-MacBook-Air LLMS % conda create --name llms2
Channels:
 - defaults
Platform: osx-arm64
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /Users/rithikkaparthi/miniconda3/envs/llms2



Proceed ([y]/n)? y

Preparing transaction: done
Verifying transaction: done
Executing transaction: done
#
# To activate this environment, use
#
#     $ conda activate llms2
#
# To deactivate an active environment, use
#
#     $ conda deactivate
```



This creates a new environment named `llms`

Activate the Conda environment:
`conda activate llms`

Step 4: Install Required Packages

Now that your Conda environment is active, you can install the necessary packages for your project.

Install `pip` (if not already installed):
`conda install pip`

Install the required Python libraries:
`pip install llama-cpp-python==0.1.78`
`pip install langchain`
`pip install langchain-community`

Step 5: Download the Llama Model

Download the model file `llama-2-7b-chat.ggmlv3.q8_0.bin` from the source. Make sure the model is placed in a directory that is accessible, such as `/Users/rithikkaparthi/Desktop/`.

Step 6: Configure VS Code to Use the Conda Environment

1. **Open VS Code.**
2. **Open Command Palette** (`Cmd+Shift+P` on Mac, `Ctrl+Shift+P` on Windows) and type **"Python: Select Interpreter".**
3. **Select the `llms` environment** from the list (it should look like `llms: Python 3.x.x (conda)`).

Step 7: Verify the Installation

In your terminal, inside the `llms` environment, check if the installed packages are correctly installed:

```
pip show llama-cpp-python
pip show langchain
pip show langchain-community
```

These commands will show the details of the installed packages. If any package is missing, you can install it again using `pip install <package-name>`.

Step 8: Run the Code

Create a Python file (e.g., `llama_model.py`) and paste the following code into it:
Python

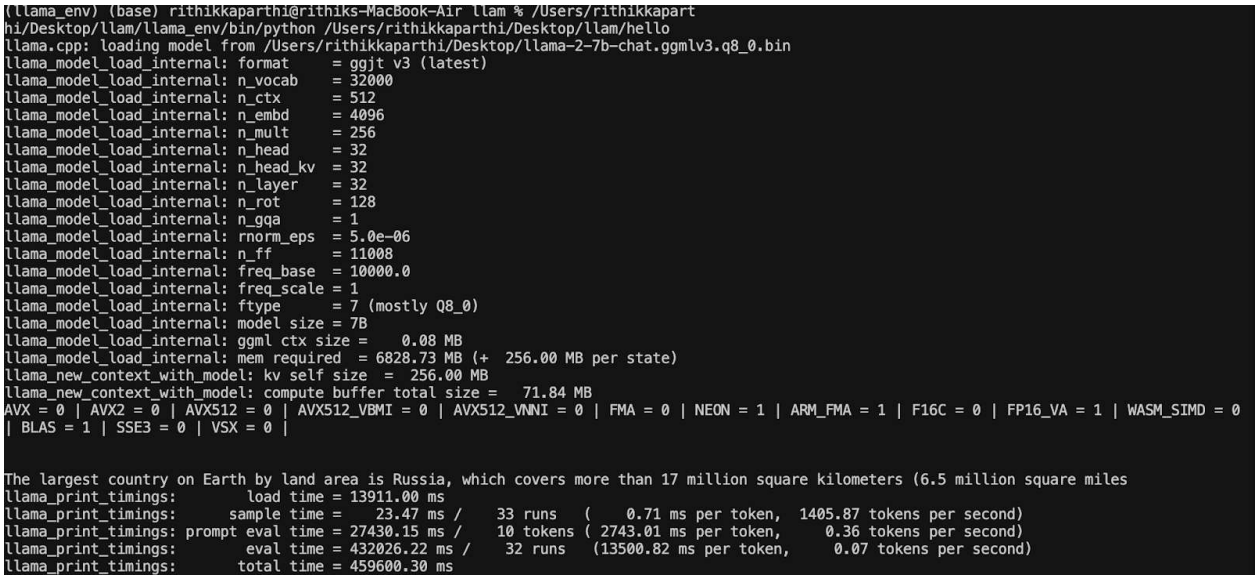
```
from langchain_community.llms import LlamaCpp
from langchain_core.callbacks import CallbackManager,
StreamingStdOutCallbackHandler
from langchain_core.prompts import PromptTemplate

# Callbacks support token-wise streaming
callback_manager = CallbackManager([StreamingStdOutCallbackHandler()])

# Initialize the model with proper configuration
llm = LlamaCpp(
    model_path="/Users/rithikkaparthi/Desktop/llama-2-7b-chat.ggmlv3.q8_0.bin",  #
Ensure the model path is correct
    temperature=0.75,
    max_tokens=2000,
    top_p=1,
    callback_manager=callback_manager,
    verbose=True,  # Verbose is required to pass to the callback manager
)

# Define the question
question = "What is the largest country on Earth?"

# Use the generate method instead of invoke()
response = llm.invoke(question)

# Print the response
```

print(response['text'])

Run the Python file: In the terminal, navigate to the folder where the Python file is located and run it:

```
python llama_model.py
```

This should print the model's response to the question.



Step 9: Troubleshooting (If Needed)

If you encounter any issues:

- **Check Model Path**: Ensure that the model file path (`model_path`) is correct and the file is located at `/Users/rithikkaparthi/Desktop/llama-2-7b-chat.ggmlv3.q8_0.bin` or the path you have specified.
- **Check Package Installations**: Make sure all required packages are installed inside the Conda environment. If any package is missing, reinstall using `pip`.
- **Memory/Performance**: If you face memory issues, make sure your system has enough resources to handle the model. The model can be large, and running it on machines with limited RAM may cause issues.