Installation and Use of Mistral-7B-Instruct-v0.3

Katherine Simon

December 1, 2024

**Abstract**

The Mistral-7B-Instruct-v0.3 LLM is a model that enables an interactive chat-bot functionality within a development project. In this project, I attempted to install and use the Mistral-7B-Instruct-v0.3 LLM. The goals were to successfully install and demonstrate use of various functions within the LLM.

**Methods**

For the installation of the Mistral 7B Instruct LLM, I followed the installation instructions listed on the HuggingFace website (Mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face). To perform the download, I first needed to create a free HuggingFace account. Once created, I was able to generate a personal Access Token that would later be used to download and install the LLM. Once the Access Token was available, I was able to run both the pip install commands as well as the download statements for the LLM. I first tested a simple chat feature using the InferenceClient library. To use some of the basic functionality like the chat-bot feature, Cuda needs to be installed. Without Cuda, the LLM will produce errors related to the use of the python packages without Cuda enabled. I followed the Cuda Installation Guide for Windows (NVIDIA) to verify device compatibility and install Cuda correctly. Once Cuda was installed, the test using this singular library with a simple pre-defined prompt was successful.

Next, I attempted use of the instruct library within the LLM. Before running any tests, I first referenced the LLMs documentation for Problems were encountered using this part of the LLM due to lack of computing resources. Upon executing a test command of "Explain Machine Learning to me in a nutshell.", using a PC equipped with 2 GPUs having dedicated memory and 16 GB of RAM, I executed a test prompt of "Explain Machine Learning to me in a nutshell." After several minutes the PC became unresponsive and the display showed a blank screen which required a hard reset to fix.

After encountering issues with the instruct library of the LLM, I decided to execute a basic command line command using the mistral-chat function from the LLM. This produced additional errors about missing dependencies. Installing those missing dependencies caused other libraries to be automatically updated with a newer version. The newer version of some of the existing libraries caused errors for some of the LLMs functions which required the older version of said library. The only way to resolve the cascading error of library incompatibility would be to uninstall python and re-install, unfortunately, the documentation is insufficient to know which version of each dependency is needed or if it’s even feasible to complete a fresh install of all versions necessary.

**Conclusions**

The Mistral-7B-Instruct-v0.3 LLM is fairly straight forward to implement, involving minimal coding to get started. The documentation is OK, but could be improved. Some of the features are more difficult to implement strictly due to compute hardware limitations. For example, a simple chat bot function could be implemented and run without issue. However, the instruct library uses significant compute resources that require GPU capabilities and a good amount of memory. The real issue I see with this LLM seems to be the lack of code hardening. The open-source nature of the LLM has made it susceptible to too many incompatible changes. The maintainers of this software might need to consider something synonymous to an “renv lock” file in which they can save a runnable and distributable environment for this software.

**Statement of Business Value**

Use of the Mistral-7B-Instruct-v0.3 LLM for chat functions has broad application. The ease of implementation reduces the cost of maintenance for a company while also allowing developers to add some AI features to a project with little experience. Most businesses use a chat-bot feature on their website to answer easy questions, so the addition of even the simplest features from the LLM lowers the threshold for this feature for smaller businesses. Unfortunately, more advanced features are still somewhat more restricted due to the compute requirements and fragility of the software.

# References

“Mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face.” *Mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face*, huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed 30 Nov. 2024.

NVIDIA. "Cuda Installation Guide for Microsoft Windows." Vers. 12.6. n.d. *NVIDIA Web Site.* Web Browser. 30 November 2024. <https://docs.nvidia.com/cuda/cuda-installation-guide-microsoft-windows/index.html#>.