

LLaVA

LLaVA

- Model size
- Pull LLaVA
- Use LLaVA
 - Run LLaVA
 - Have a conversation
 - End the conversation
- Memory optimization
- References

Demo Environment

Development board: Jetson Orin series motherboard

SSD: 128G

Tutorial application scope: Whether the motherboard can run is related to the available memory of the system. The user's own environment and the programs running in the background may cause the model to fail to run

| Motherboard model | Ollama | Open WebUI |
|----------------------|--------|------------|
| Jetson Orin NX 16GB | √ | √ |
| Jetson Orin NX 8GB | √ | √ |
| Jetson Orin Nano 8GB | √ | × |
| Jetson Orin Nano 4GB | × | × |

LLaVA (Large-scale Language and Vision Assistant) is a multimodal model designed to achieve general vision and language understanding by combining visual encoders and large-scale language models.

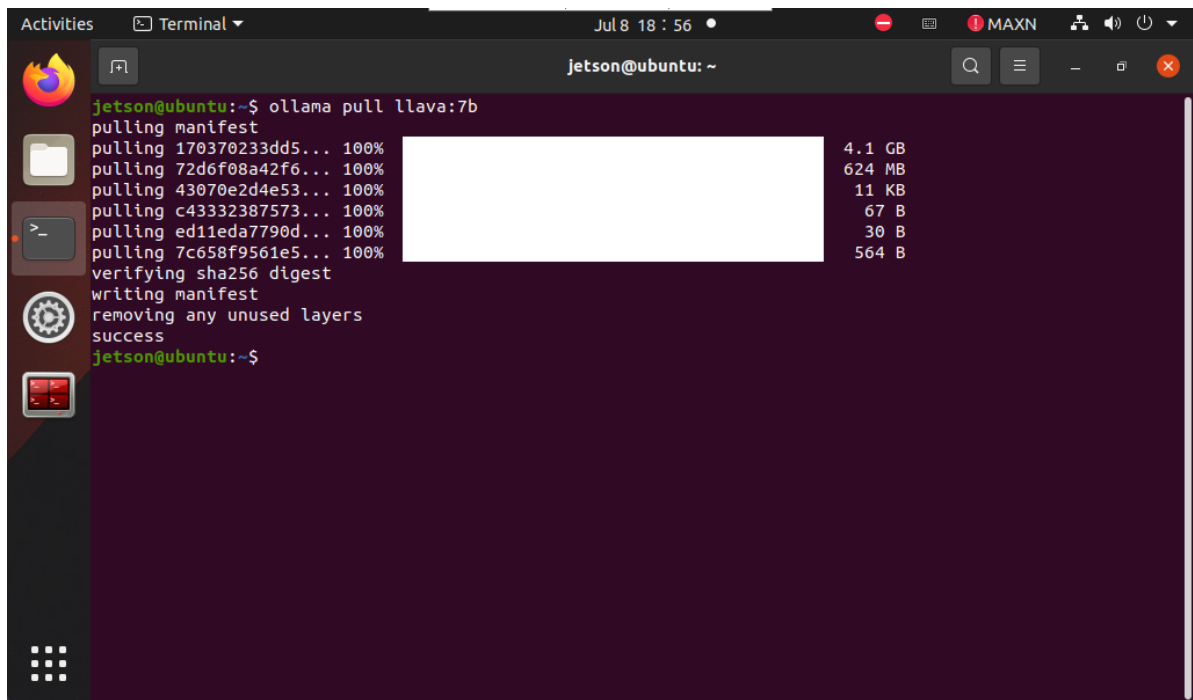
Model size

| Model | Parameters |
|-------|------------|
| LLaVA | 7B |
| LLaVA | 13B |
| LLaVA | 34B |

Pull LLaVA

Using the pull command will automatically pull the model of the Ollama model library:

```
ollama pull llava:7b
```

A terminal window on a Jetson Ubuntu system showing the command 'ollama pull llava:7b' being executed. The output shows the progress of pulling the model layers, with a progress bar and a table of layer sizes. The layers are: 170370233dd5... (4.1 GB), 72d6f08a42f6... (624 MB), 43070e2d4e53... (11 KB), c43332387573... (67 B), ed11eda7790d... (30 B), and 7c658f9561e5... (564 B). The process completes with 'success' and the prompt returns to 'jetson@ubuntu:~\$'.

```
jetson@ubuntu:~$ ollama pull llava:7b
pulling manifest
pulling 170370233dd5... 100%
pulling 72d6f08a42f6... 100%
pulling 43070e2d4e53... 100%
pulling c43332387573... 100%
pulling ed11eda7790d... 100%
pulling 7c658f9561e5... 100%
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$
```

Use LLaVA

Use LLaVA to recognize local image content.

Run LLaVA

If the system does not have a running model, the system will automatically pull the LLaVA 7B model and run it:

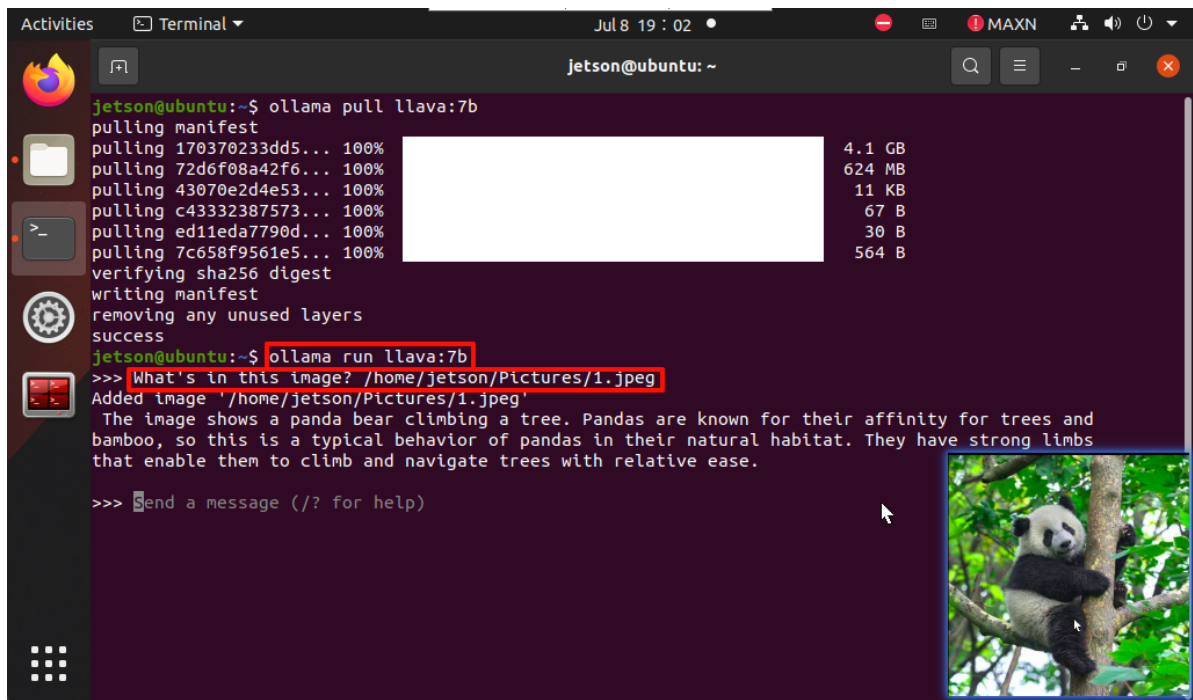
```
ollama run llava:7b
```

Have a conversation


```
what's in this image? /home/jetson/Pictures/1.jpeg
```

The time to reply to the question depends on the hardware configuration, so be patient!

If the image does not have a corresponding image, you can download the image yourself (the resolution should not be too large) and put the image path after the question!

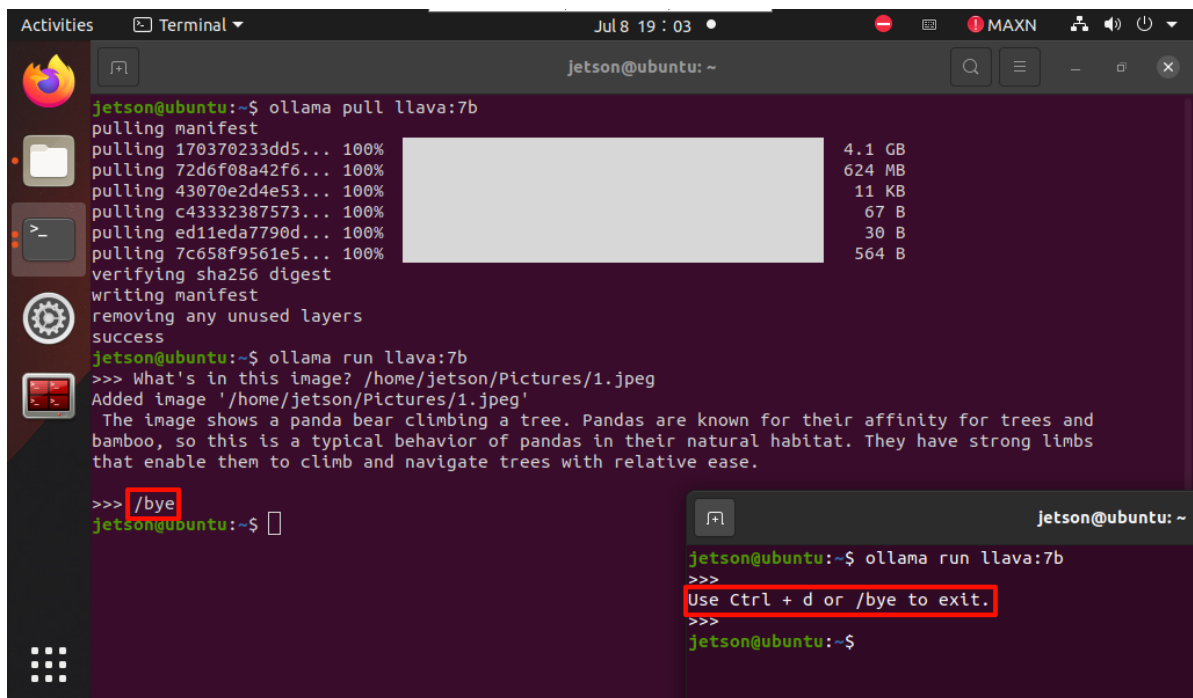


```
jetson@ubuntu:~$ ollama pull llava:7b
pulling manifest
pulling 170370233dd5... 100%
pulling 72d6f08a42f6... 100%
pulling 43070e2d4e53... 100%
pulling c43332387573... 100%
pulling ed11eda7790d... 100%
pulling 7c658f9561e5... 100%
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$ ollama run llava:7b
>>> What's in this image? /home/jetson/Pictures/1.jpeg
Added image '/home/jetson/Pictures/1.jpeg'
The image shows a panda bear climbing a tree. Pandas are known for their affinity for trees and bamboo, so this is a typical behavior of pandas in their natural habitat. They have strong limbs that enable them to climb and navigate trees with relative ease.
```



End the conversation

Use the `Ctrl+d` shortcut key or `/bye` to end the conversation!



```
jetson@ubuntu:~$ ollama pull llava:7b
pulling manifest
pulling 170370233dd5... 100%
pulling 72d6f08a42f6... 100%
pulling 43070e2d4e53... 100%
pulling c43332387573... 100%
pulling ed11eda7790d... 100%
pulling 7c658f9561e5... 100%
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$ ollama run llava:7b
>>> What's in this image? /home/jetson/Pictures/1.jpeg
Added image '/home/jetson/Pictures/1.jpeg'
The image shows a panda bear climbing a tree. Pandas are known for their affinity for trees and bamboo, so this is a typical behavior of pandas in their natural habitat. They have strong limbs that enable them to climb and navigate trees with relative ease.
>>> /bye
jetson@ubuntu:~$
```

```
jetson@ubuntu:~$ ollama run llava:7b
>>> Use Ctrl + d or /bye to exit.
>>>
jetson@ubuntu:~$
```

Memory optimization

Since the local model has very high memory requirements, those who cannot run the model can follow the tutorial below to close the graphical interface and run the model in command line mode.

For users without a display screen, the command line mode requires knowing your motherboard IP in advance

- Command line mode

```
sudo systemctl set-default multi-user.target
```

After running, restart the system to take effect, and then use SSH remote system to run the model.

- Desktop mode (graphic interface)

```
sudo systemctl set-default graphical.target
```

After running, restart the system to take effect and restore desktop mode.

References

Ollama

Official website: <https://ollama.com/>

GitHub: <https://github.com/ollama/ollama>

LLaVA

GitHub: <https://github.com/haotian-liu/LLaVA>

Ollama corresponding model: <https://ollama.com/library/llava>