

LLaVA

LLaVA

- 1. Model scale
 - 2. Pull LLaVA
 - 3. Use LLaVA
 - 3.1. Run LLaVA
 - 3.2. Have a conversation
 - 3.3. End the conversation
 - 4. 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	Run directly with Ollama	Run with 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 that aims to achieve general vision and language understanding by combining visual encoders and large-scale language models.

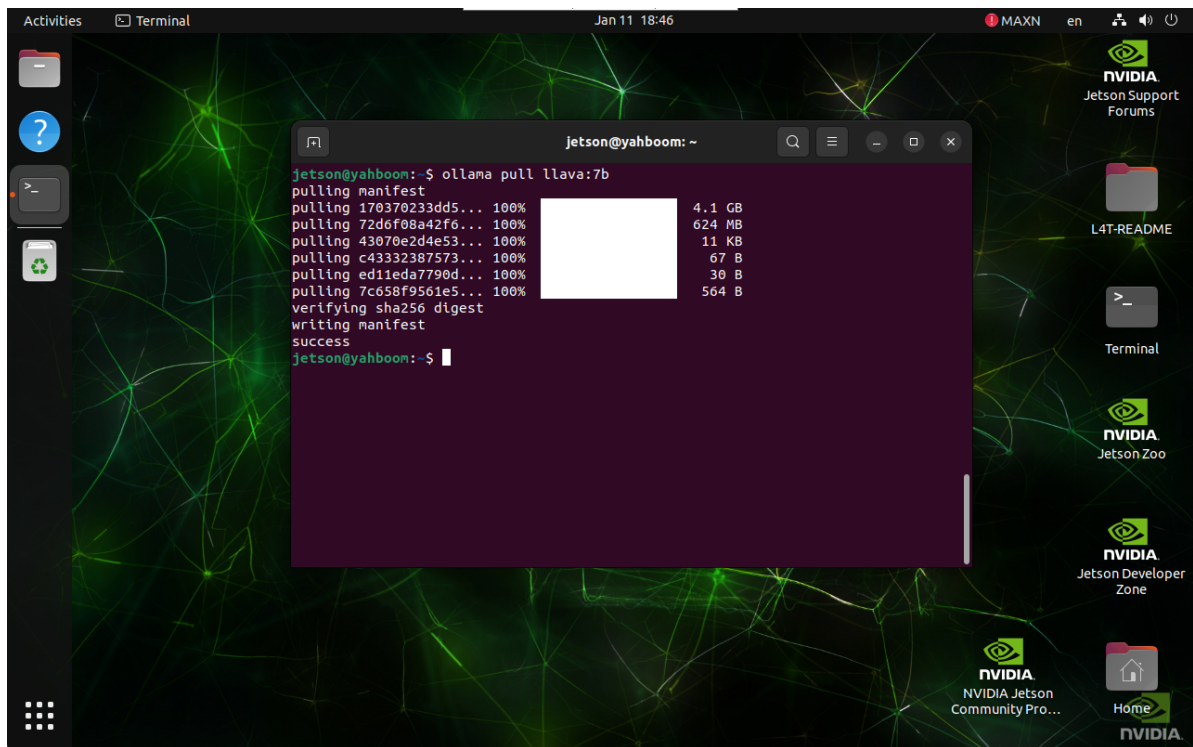
1. Model scale

Model	Parameters
LLaVA	7B
LLaVA	13B
LLaVA	34B

2. Pull LLaVA

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

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



3. Use LLaVA

Use LLaVA to identify local image content.

3.1. Run LLaVA

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

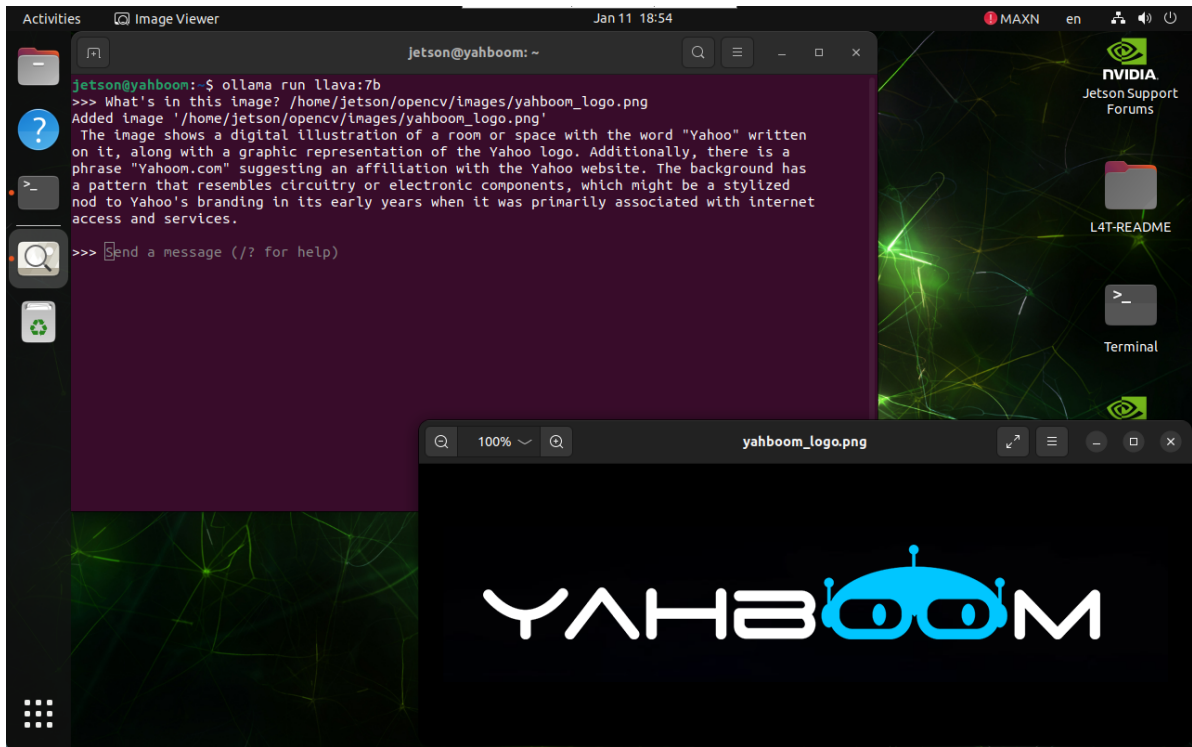
```
ollama run llava:7b
```

3.2. Have a conversation

```
what's in this image? /home/jetson/opencv/images/yahboom_logo.png
```

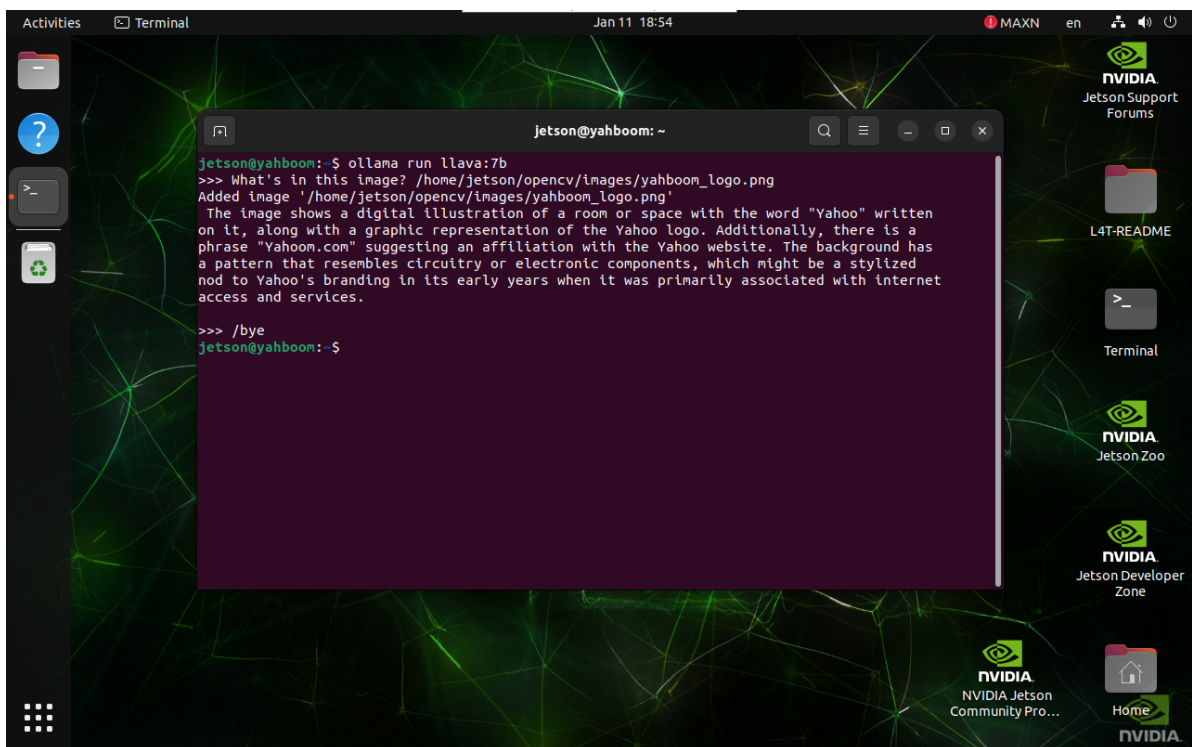
The time to reply to the question depends on the hardware configuration, so be patient!

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



3.3. End the conversation

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



4. Memory optimization

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

For users without display screen, the command line mode requires knowing the IP address of the motherboard 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 (graphical interface)

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

After running, restart the system to take effect and restore the 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>