How to run the Flux-1-dev model on server

1. **python3 -m venv venv**
   source venv/bin/activate
2. From pytorch website:
   pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118
3. **pip install –upgrade diffusers accelerate transformers**
4. From Huggingface website flux model python file:-

```
import torch
from diffusers import FluxPipeline

pipe = FluxPipeline.from_pretrained("black-forest-labs/FLUX.1-dev", torch_dtype=torch.bfloat16)
pipe.enable_model_cpu_offload() #save some VRAM by offloading the model to CPU. Remove this if you have enough GPU power

prompt = "A cat holding a sign that says hello world"
image = pipe(
        prompt,
        height=1024,
        width=1024,
        guidance_scale=3.5,
        num_inference_steps=50,
        max_sequence_length=512,
        generator=torch.Generator("cpu").manual_seed(0)
).images[0]
image.save("flux-dev.png")
```

5. Changes to make:-

```
import torch
from diffusers import FluxPipeline

pipe = FluxPipeline.from_pretrained("black-forest-labs/FLUX.1-dev", token=" ",torch_dtype=torch.bfloat16)
pipe.enable_model_cpu_offload() #save some VRAM by offloading the model to CPU. Remove this if you have enough GPU power

prompt = "A cat holding a sign that says hello world"
image = pipe(
        prompt,
        height=1024,
        width=1024,
        guidance_scale=3.5,
        num_inference_steps=50,
```

max_sequence_length=512,
generator=torch.Generator("cpu").manual_seed(0)
).images[0]
image.save("flux-dev.png")

Note:- In token's double codes write the huggingface token

6. Run:- python main.py
7. On system folder :- huggingface>hub>locks
   In **locks** you can see that **flux-1-dev** folder is there and there are two cache folders so delete them and then run the main.py
8. Inside the environment only you can write the code in main.ipynb file and then try to run the file there.
9. In main.ipynb:- pip install sentencepiece
   Add the line **import sentencepiece** in the starting and then try to run the cell.
10. Import Error:- protobuf not found.
11. Try the python file in jupyter notebook and try uninstalling and installing sentencepiece.
12. pip install protobuf
13. In terminal outside VS CODE:
    watch -n1 nvidia-smi
14. To check the memory usage:-
    In the code change
    pipe = FluxPipeline.from_pretrained("black-forest-labs/FLUX.1-dev", torch_dtype=torch.bfloat16)
    this line to
    pipe = FluxPipeline.from_pretrained("black-forest-labs/FLUX.1-dev", torch_dtype=torch.bfloat16).to("cuda")

15. So, it is currently using:-
    GPU Memory :- 33754 MB / 49140 MB
    Consumption :- 298 W