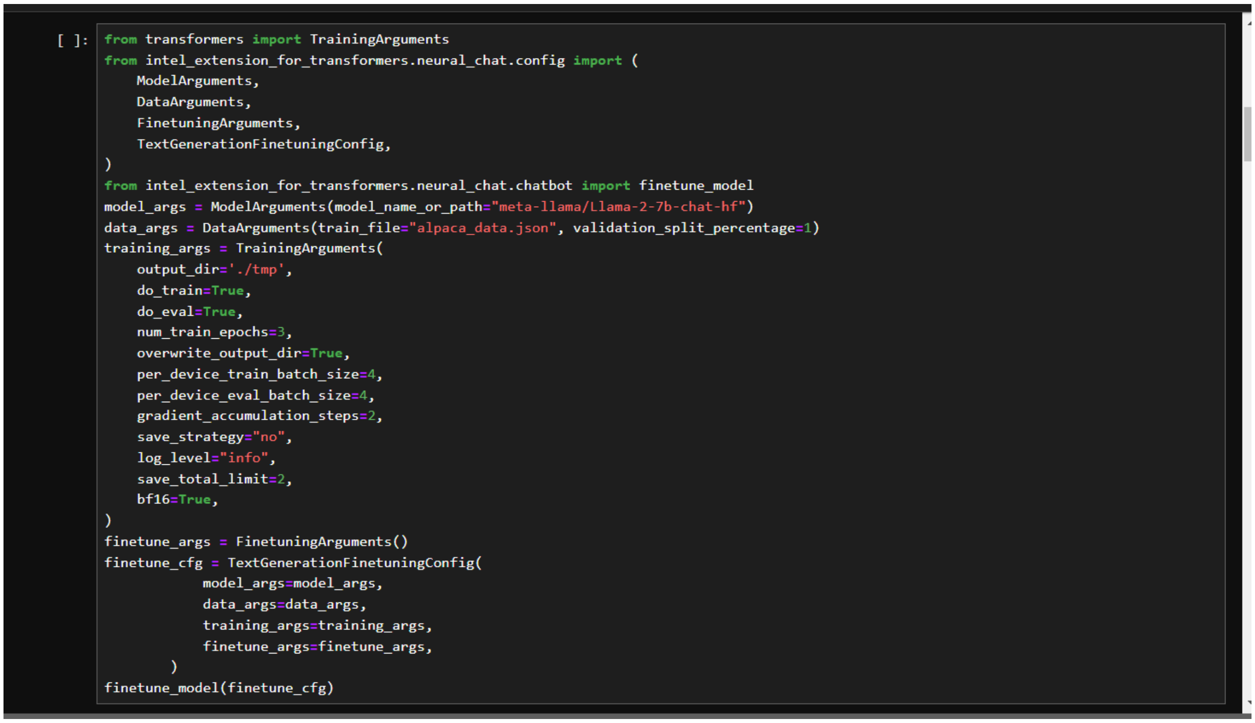
Single_node_finetuning_on_spr

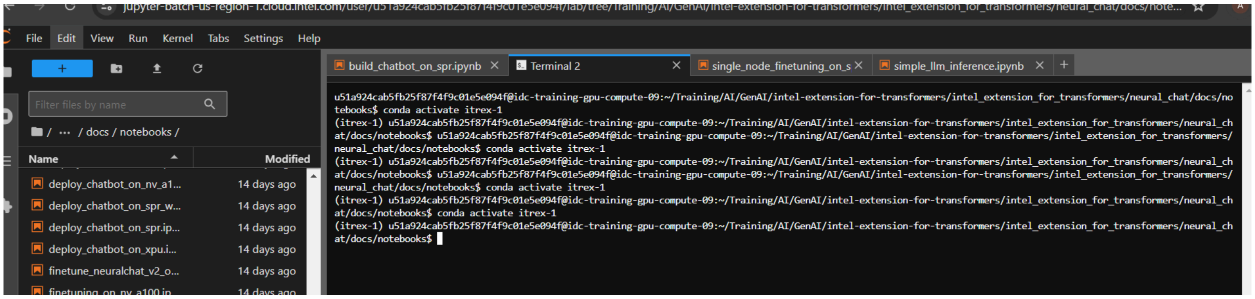
What is Finetuning?

Fine-tuning is a process in which a data is fed to the model and tells the models internal weights to get closer to responding how we would like it. For example we can fine tune a model for code generation , text generation and summarisation.In simple terms finetuning is a process in which we can train a model to do specific task by providing the right dataset



```python
from transformers import TrainingArguments
from intel_extension_for_transformers.neural_chat.config import (
    ModelArguments,
    DataArguments,
    FinetuningArguments,
    TextGenerationFinetuningConfig,
)
from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
data_args = DataArguments(train_file="alpaca_data.json", validation_split_percentage=1)
training_args = TrainingArguments(
    output_dir='./tmp',
    do_train=True,
    do_eval=True,
    num_train_epochs=3,
    overwrite_output_dir=True,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    gradient_accumulation_steps=2,
    save_strategy="no",
    log_level="info",
    save_total_limit=2,
    bf16=True,
)
finetune_args = FinetuningArguments()
finetune_cfg = TextGenerationFinetuningConfig(
            model_args=model_args,
            data_args=data_args,
            training_args=training_args,
            finetune_args=finetune_args,
        )
finetune_model(finetune_cfg)
```

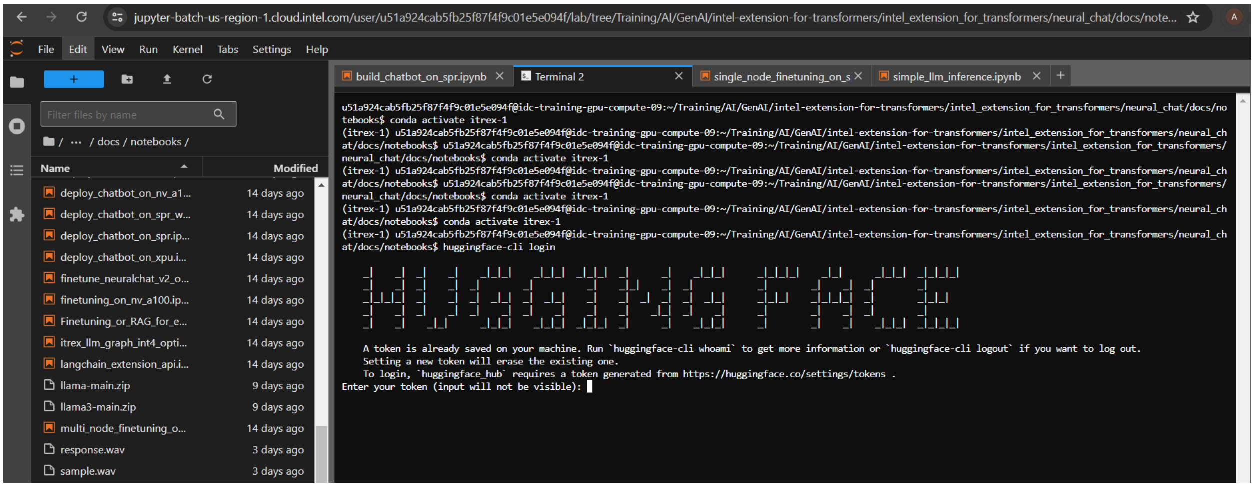
Step 1:
Activate the conda environment


jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/note...
File
Edit
View
Run
Kernel
Tabs
Settings
Help
+
Filter files by name
/ … / docs / notebooks /
Name
Modified
deploy_chatbot_on_nv_a1... 14 days ago
deploy_chatbot_on_spr_w... 14 days ago
deploy_chatbot_on_spr.ip... 14 days ago
deploy_chatbot_on_xpu.i... 14 days ago
finetune_neuralchat_v2_o... 14 days ago
finetuning_on_nv_a100.ip... 14 days ago
build_chatbot_on_spr.ipynb
Terminal 2
single_node_finetuning_on_s
simple_llm_inference.ipynb

```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/no
tebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ 
```

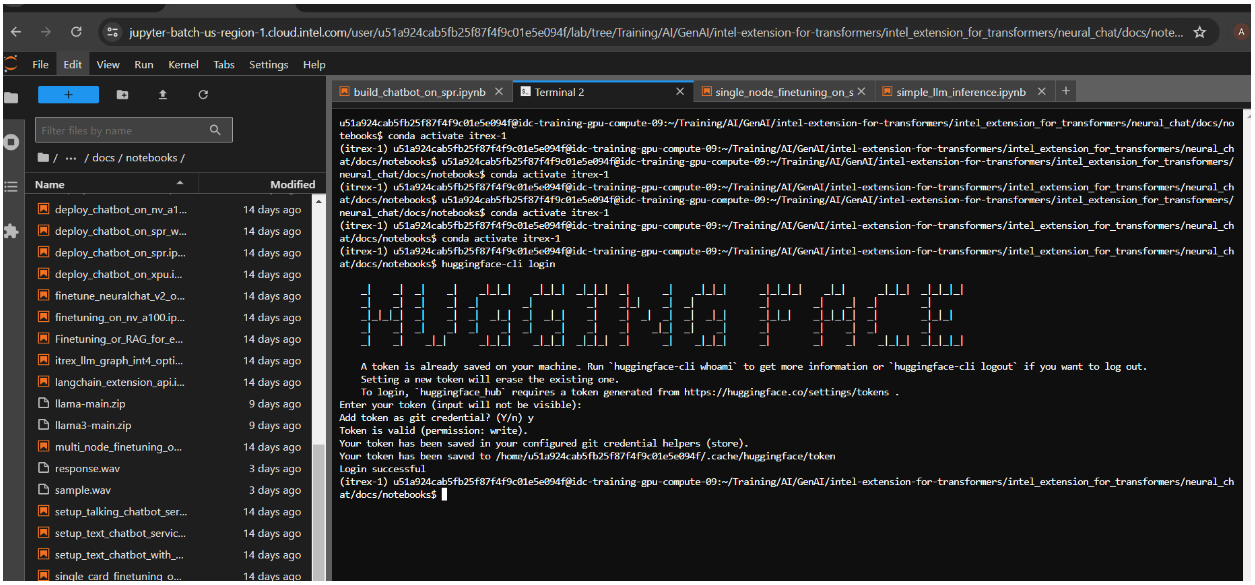
Step2:

Activate the hugging face hub using huggingface-cli login


jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/note...
File
Edit
View
Run
Kernel
Tabs
Settings
Help
+
Filter files by name
/ … / docs / notebooks /
Name
Modified
deploy_chatbot_on_nv_a1... 14 days ago
deploy_chatbot_on_spr_w... 14 days ago
deploy_chatbot_on_spr.ip... 14 days ago
deploy_chatbot_on_xpu.i... 14 days ago
finetune_neuralchat_v2_o... 14 days ago
finetuning_on_nv_a100.ip... 14 days ago
Finetuning_or_RAG_for_e... 14 days ago
itrex_llm_graph_int4_opti... 14 days ago
langchain_extension_api.i... 14 days ago
llama-main.zip 9 days ago
llama3-main.zip 9 days ago
multi_node_finetuning_o... 14 days ago
response.wav 3 days ago
sample.wav 3 days ago
build_chatbot_on_spr.ipynb
Terminal 2
single_node_finetuning_on_s
simple_llm_inference.ipynb

```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/no
tebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ huggingface-cli login

    _|    _|  _|    _|    _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|_|_|_|    _|_|      _|_|_|  _|_|_|_|
    _|    _|  _|    _|  _|        _|          _|    _|_|    _|  _|            _|        _|    _|  _|        _|
    _|_|_|_|  _|    _|  _|  _|_|  _|  _|_|    _|    _|  _|  _|  _|  _|_|      _|_|_|    _|_|_|_|  _|        _|_|_|
    _|    _|  _|    _|  _|    _|  _|    _|    _|    _|    _|_|  _|    _|      _|        _|    _|  _|        _|
    _|    _|    _|_|      _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|        _|    _|    _|_|_|  _|_|_|_|

    A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
    Setting a new token will erase the existing one.
    To login, `huggingface_hub` requires a token generated from https://huggingface.co/settings/tokens .
Enter your token (input will not be visible): 
```

Step 3:
Input the user token

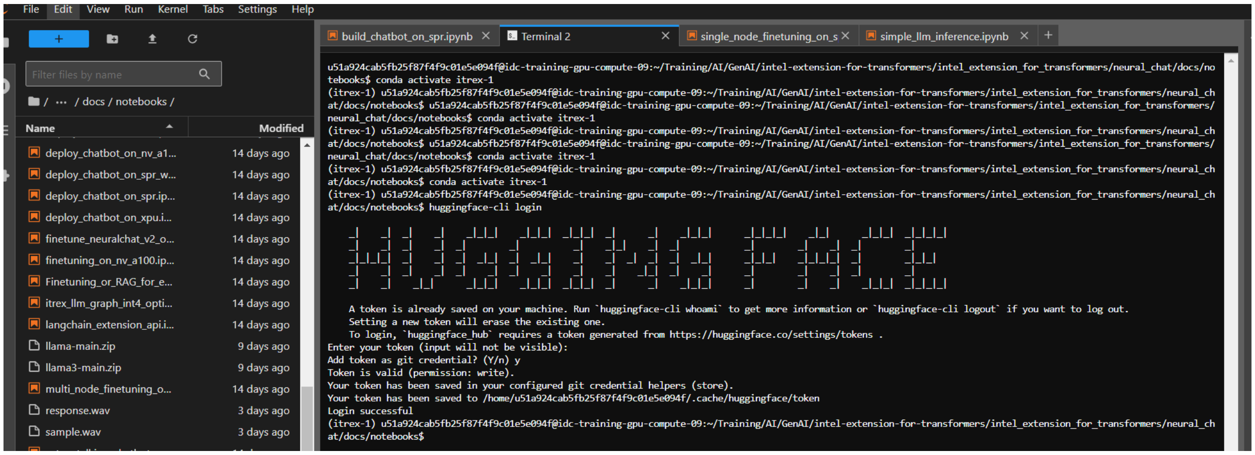

jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/note...
File
Edit
View
Run
Kernel
Tabs
Settings
Help
+
Filter files by name
/ ··· / docs / notebooks /
build_chatbot_on_spr.ipynb
Terminal 2
single_node_finetuning_on_s
simple_llm_inference.ipynb

Name	Modified
deploy_chatbot_on_nv_a1...	14 days ago
deploy_chatbot_on_spr_w...	14 days ago
deploy_chatbot_on_spr.ip...	14 days ago
deploy_chatbot_on_xpu.i...	14 days ago
finetune_neuralchat_v2_o...	14 days ago
finetuning_on_nv_a100.ip...	14 days ago
Finetuning_or_RAG_for_e...	14 days ago
itrex_llm_graph_int4_opti...	14 days ago
langchain_extension_api.i...	14 days ago
llama-main.zip	9 days ago
llama3-main.zip	9 days ago
multi_node_finetuning_o...	14 days ago
response.wav	3 days ago
sample.wav	3 days ago
setup_talking_chatbot_ser...	14 days ago
setup_text_chatbot_servic...	14 days ago
setup_text_chatbot_with_...	14 days ago
single_card_finetuning_o...	14 days ago

```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/no
tebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ huggingface-cli login

    _|    _|  _|    _|    _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|_|_|_|    _|_|      _|_|_|  _|_|_|_|
    _|    _|  _|    _|  _|        _|          _|    _|_|    _|  _|            _|        _|    _|  _|        _|
    _|_|_|_|  _|    _|  _|  _|_|  _|  _|_|    _|    _|  _|  _|  _|  _|_|      _|_|_|    _|_|_|_|  _|        _|_|_|
    _|    _|  _|    _|  _|    _|  _|    _|    _|    _|    _|_|  _|    _|      _|        _|    _|  _|        _|
    _|    _|    _|_|      _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|        _|    _|    _|_|_|  _|_|_|_|

    A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
    Setting a new token will erase the existing one.
    To login, `huggingface_hub` requires a token generated from https://huggingface.co/settings/tokens .
Enter your token (input will not be visible):
Add token as git credential? (Y/n) y
Token is valid (permission: write).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/token
Login successful
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ 
```

Step 4:

Once the token is accepted the user has the id to use hugging face datasets

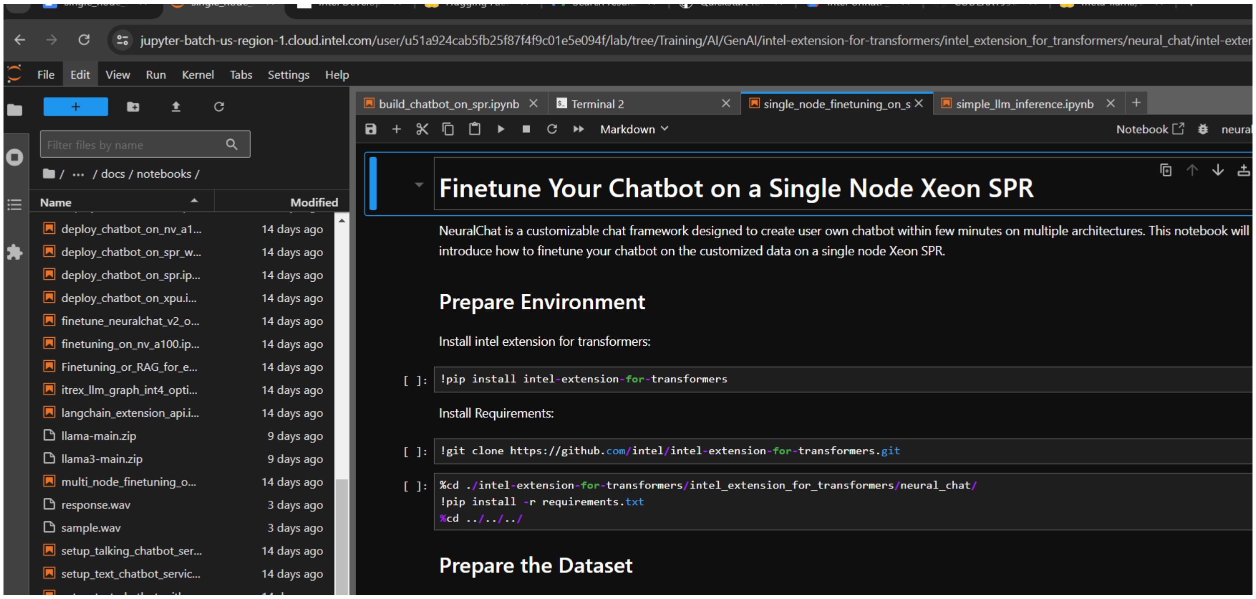

File
Edit
View
Run
Kernel
Tabs
Settings
Help
+
Filter files by name
/ ··· / docs / notebooks /
build_chatbot_on_spr.ipynb
Terminal 2
single_node_finetuning_on_s
simple_llm_inference.ipynb

Name	Modified
deploy_chatbot_on_nv_a1...	14 days ago
deploy_chatbot_on_spr_w...	14 days ago
deploy_chatbot_on_spr.ip...	14 days ago
deploy_chatbot_on_xpu.i...	14 days ago
finetune_neuralchat_v2_o...	14 days ago
finetuning_on_nv_a100.ip...	14 days ago
Finetuning_or_RAG_for_e...	14 days ago
itrex_llm_graph_int4_opti...	14 days ago
langchain_extension_api.i...	14 days ago
llama-main.zip	9 days ago
llama3-main.zip	9 days ago
multi_node_finetuning_o...	14 days ago
response.wav	3 days ago
sample.wav	3 days ago

```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/no
tebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/
neural_chat/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$ huggingface-cli login

    _|    _|  _|    _|    _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|_|_|_|    _|_|      _|_|_|  _|_|_|_|
    _|    _|  _|    _|  _|        _|          _|    _|_|    _|  _|            _|        _|    _|  _|        _|
    _|_|_|_|  _|    _|  _|  _|_|  _|  _|_|    _|    _|  _|  _|  _|  _|_|      _|_|_|    _|_|_|_|  _|        _|_|_|
    _|    _|  _|    _|  _|    _|  _|    _|    _|    _|    _|_|  _|    _|      _|        _|    _|  _|        _|
    _|    _|    _|_|      _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|        _|    _|    _|_|_|  _|_|_|_|

    A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
    Setting a new token will erase the existing one.
    To login, `huggingface_hub` requires a token generated from https://huggingface.co/settings/tokens .
Enter your token (input will not be visible):
Add token as git credential? (Y/n) y
Token is valid (permission: write).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/token
Login successful
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-09:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$
```

Step 5:

Open the single_node_finetuning_on_spr on idc


jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/intel-exter
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Filter files by name
/ ··· / docs / notebooks /
build_chatbot_on_spr.ipynb
Terminal 2
single_node_finetuning_on_s
simple_llm_inference.ipynb
Markdown
Notebook
neural

Name	Modified
deploy_chatbot_on_nv_a1...	14 days ago
deploy_chatbot_on_spr_w...	14 days ago
deploy_chatbot_on_spr.ip...	14 days ago
deploy_chatbot_on_xpu.i...	14 days ago
finetune_neuralchat_v2_o...	14 days ago
finetuning_on_nv_a100.ip...	14 days ago
Finetuning_or_RAG_for_e...	14 days ago
itrex_llm_graph_int4_opti...	14 days ago
langchain_extension_api.i...	14 days ago
llama-main.zip	9 days ago
llama3-main.zip	9 days ago
multi_node_finetuning_o...	14 days ago
response.wav	3 days ago
sample.wav	3 days ago
setup_talking_chatbot_ser...	14 days ago
setup_text_chatbot_servic...	14 days ago

Finetune Your Chatbot on a Single Node Xeon SPR

NeuralChat is a customizable chat framework designed to create user own chatbot within few minutes on multiple architectures. This notebook will introduce how to finetune your chatbot on the customized data on a single node Xeon SPR.

Prepare Environment

Install intel extension for transformers:

```
[ ]: !pip install intel-extension-for-transformers
```

Install Requirements:

```
[ ]: !git clone https://github.com/intel/intel-extension-for-transformers.git
```

```
[ ]: %cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/
     !pip install -r requirements.txt
     %cd ../../../
```

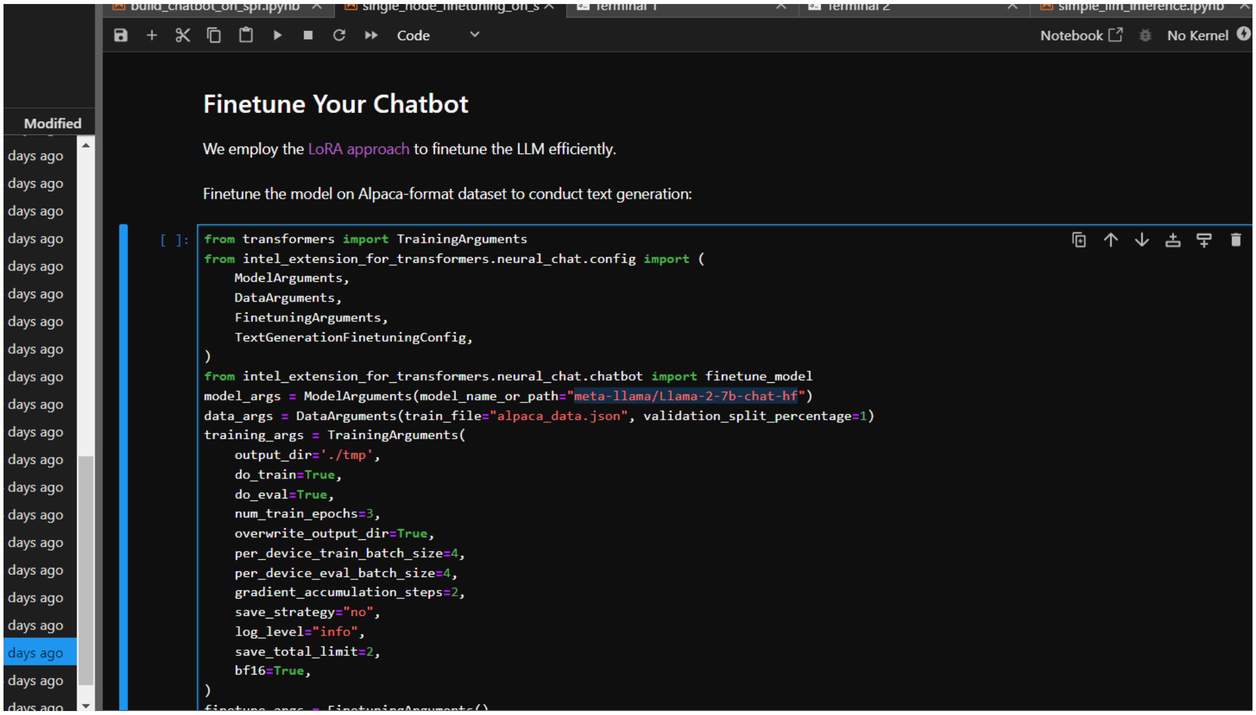
Prepare the Dataset

Step 6:
Inorder to finetune the 3 notebooks we need 3 different datasets required in the notebook . In the given notebook 3 different datasets are provided.   The model mention here is the meta-llama/Llama-2-7b-chat-hf

The model required to run the given notebook is

1. Alcapa dataset
2. Cnn_dailymail
3. theblackcat102/evol-codealpaca-v1

Step 7:


build_chatbot_on_spr.ipynb
single_node_finetuning_on_s
Terminal 1
Terminal 2
simple_llm_inference.ipynb
Code
Notebook
No Kernel
Modified
days ago

Finetune Your Chatbot

We employ the LoRA approach to finetune the LLM efficiently.

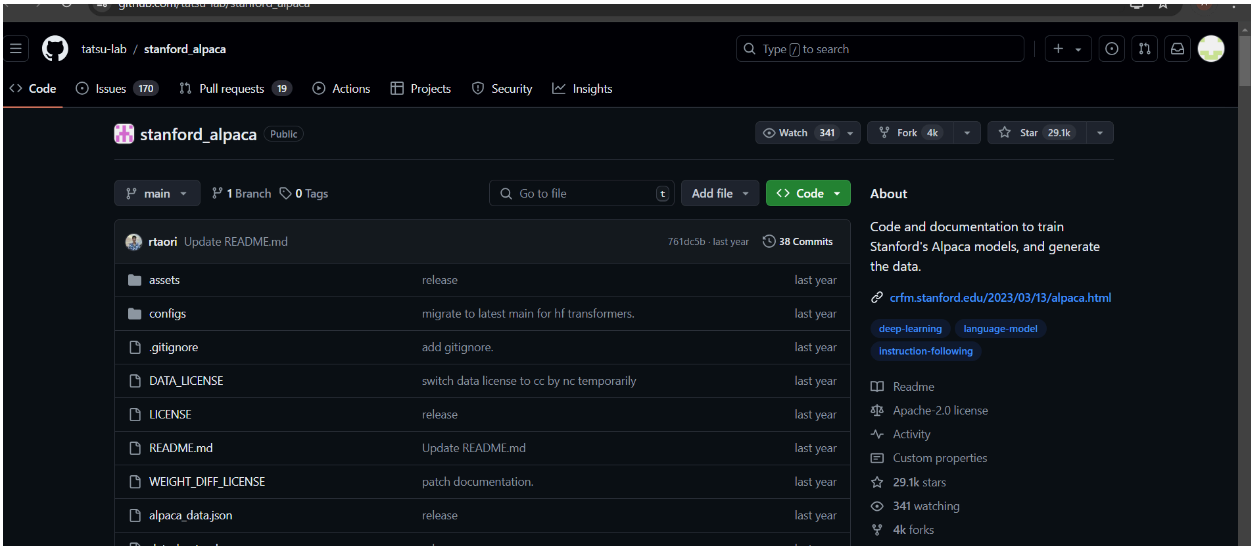
Finetune the model on Alpaca-format dataset to conduct text generation:

```
[ ]: from transformers import TrainingArguments
     from intel_extension_for_transformers.neural_chat.config import (
         ModelArguments,
         DataArguments,
         FinetuningArguments,
         TextGenerationFinetuningConfig,
     )
     from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
     model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
     data_args = DataArguments(train_file="alpaca_data.json", validation_split_percentage=1)
     training_args = TrainingArguments(
         output_dir='./tmp',
         do_train=True,
         do_eval=True,
         num_train_epochs=3,
         overwrite_output_dir=True,
         per_device_train_batch_size=4,
         per_device_eval_batch_size=4,
         gradient_accumulation_steps=2,
         save_strategy="no",
         log_level="info",
         save_total_limit=2,
         bf16=True,
     )
     finetune_args = FinetuningArguments()
```

In the given notebook we have specified the model name and path as well as the dataset required for the finetuning The dataset mentioned in the notebook is the alpaca dataset.

Step 8:

Click the alpaca dataset in the notebook . Once its clicked it will redirect us to a github rep having the alpaca dataset



rtaori Update README.md 761dc5b · last year 38 Commits

assets	release	last year
configs	migrate to latest main for hf transformers.	last year
.gitignore	add gitignore.	last year
DATA_LICENSE	switch data license to cc by nc temporarily	last year
LICENSE	release	last year
README.md	Update README.md	last year
WEIGHT_DIFF_LICENSE	patch documentation.	last year
alpaca_data.json	release	last year

About

Code and documentation to train Stanford's Alpaca models, and generate the data.

crfm.stanford.edu/2023/03/13/alpaca.html

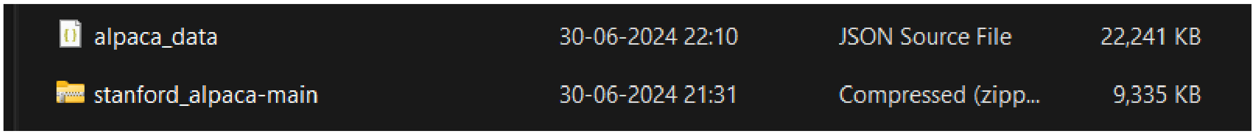
deep-learning language-model instruction-following

Readme
Apache-2.0 license
Activity
Custom properties
29.1k stars
341 watching
4k forks

Step 9:

Inorder to use the dataset we need to download it as a zip file

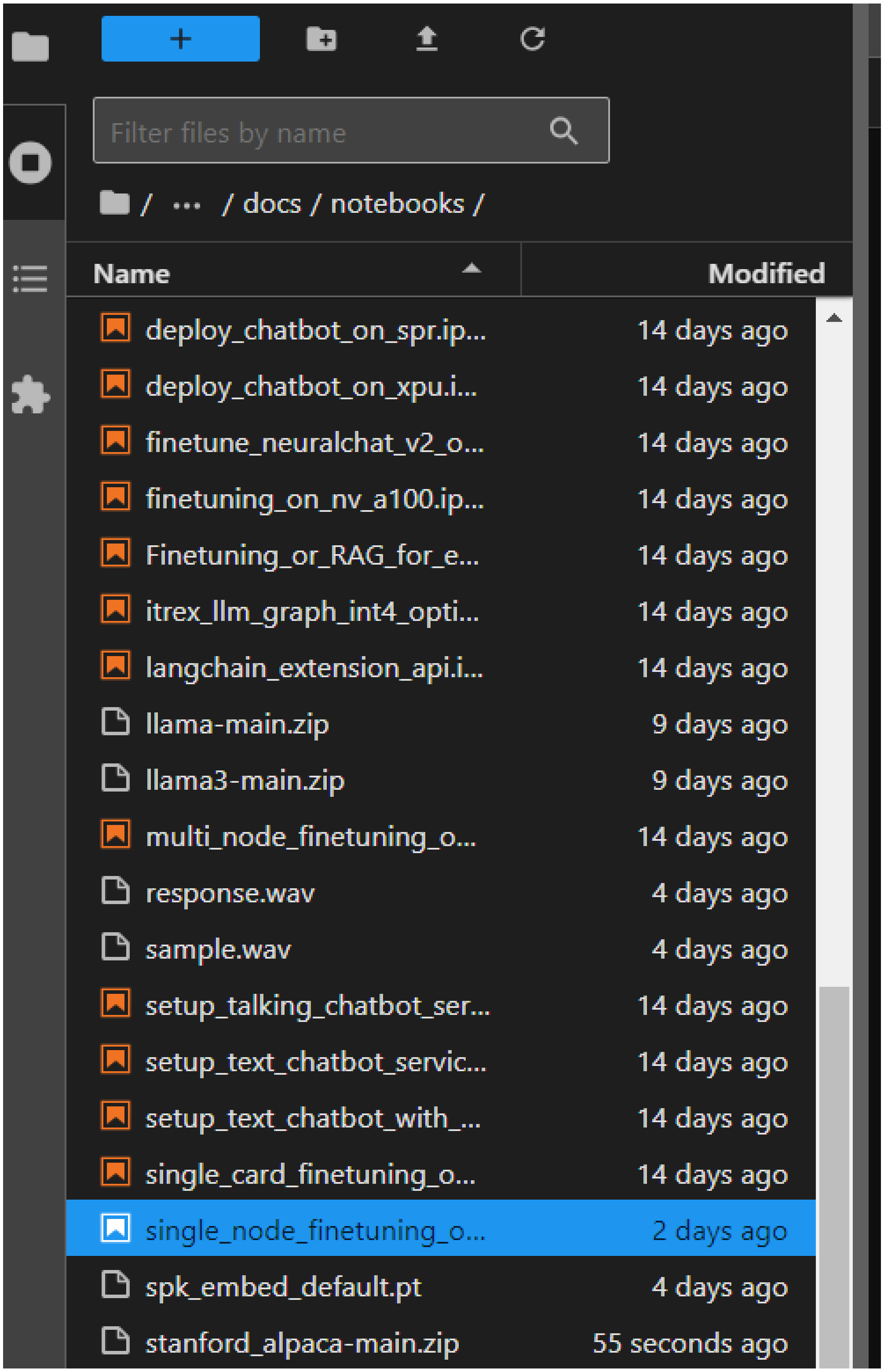
Step 10:
 Once the file is downloaded it needs to be uploaded in idc or the local terminal for unzipping the zip file .



alpaca_data	30-06-2024 22:10	JSON Source File	22,241 KB
stanford_alpaca-main	30-06-2024 21:31	Compressed (zipp...	9,335 KB

Step 11:

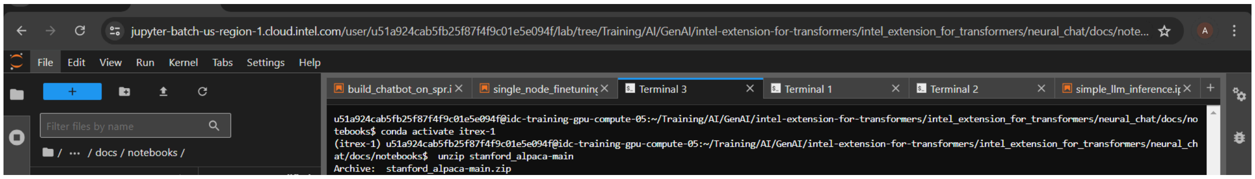
Once the file is uploaded we can unzip it in the terminal in idc server



Name	Modified
deploy_chatbot_on_spr.ip...	14 days ago
deploy_chatbot_on_xpu.i...	14 days ago
finetune_neuralchat_v2_o...	14 days ago
finetuning_on_nv_a100.ip...	14 days ago
Finetuning_or_RAG_for_e...	14 days ago
itrex_llm_graph_int4_opti...	14 days ago
langchain_extension_api.i...	14 days ago
llama-main.zip	9 days ago
llama3-main.zip	9 days ago
multi_node_finetuning_o...	14 days ago
response.wav	4 days ago
sample.wav	4 days ago
setup_talking_chatbot_ser...	14 days ago
setup_text_chatbot_servic...	14 days ago
setup_text_chatbot_with_...	14 days ago
single_card_finetuning_o...	14 days ago
single_node_finetuning_o...	2 days ago
spk_embed_default.pt	4 days ago
stanford_alpaca-main.zip	55 seconds ago

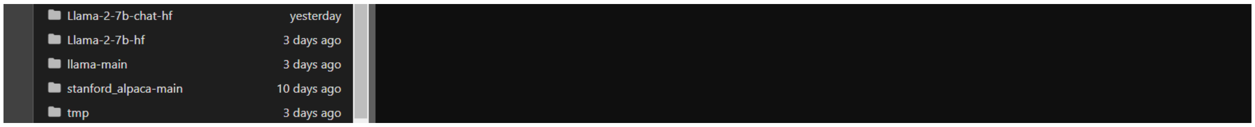
Step 12:

Unzip it in the idc terminal



jupyter-batch-us-region-1.cloud.intel.com/user/u51a924cab5fb25f87f4f9c01e5e094f/lab/tree/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/note...
File
Edit
View
Run
Kernel
Tabs
Settings
Help
build_chatbot_on_spr.i
single_node_finetuninc
Terminal 3
Terminal 1
Terminal 2
simple_llm_inference.ip
Filter files by name
/ … / docs / notebooks /

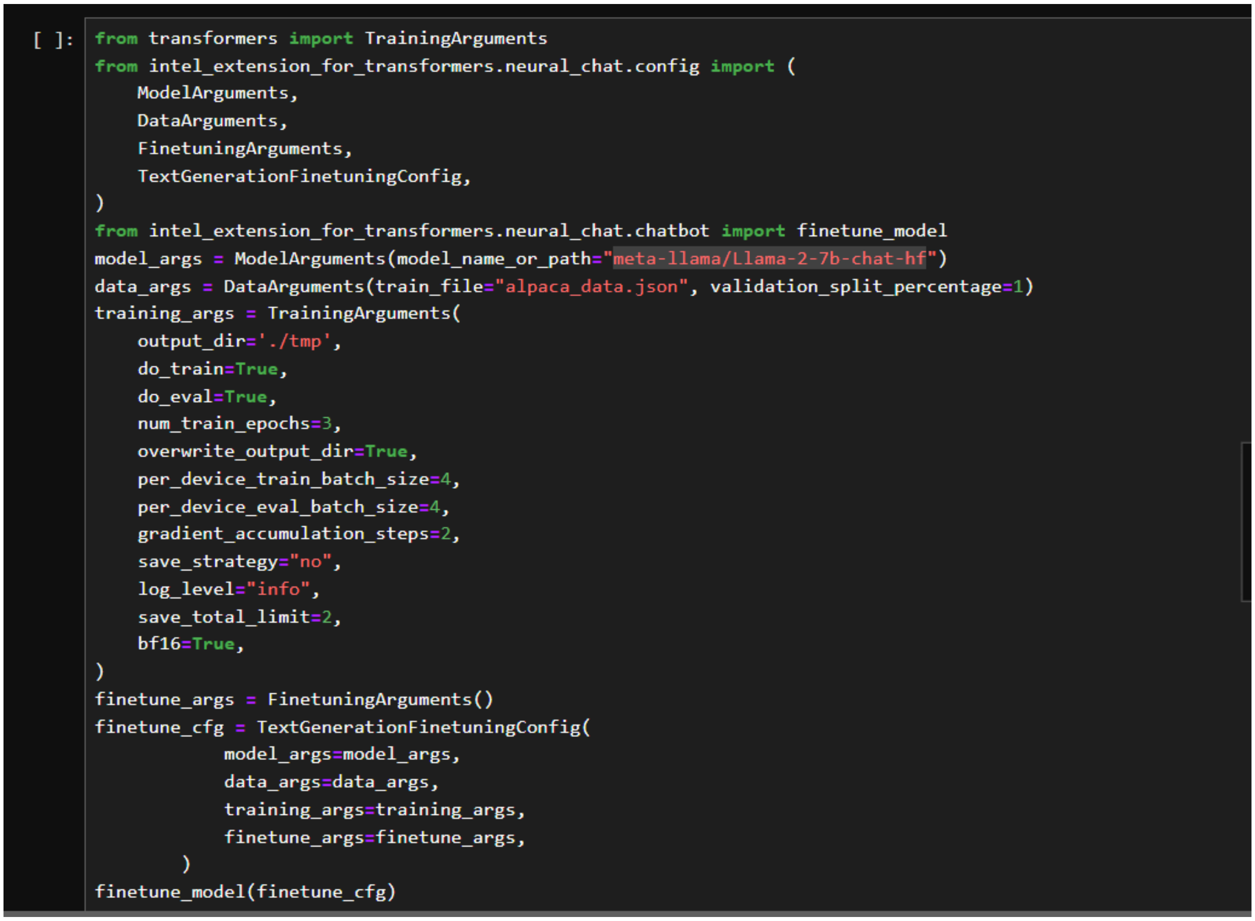
```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-05:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/no
tebooks$ conda activate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-05:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_ch
at/docs/notebooks$  unzip stanford_alpaca-main
Archive:  stanford_alpaca-main.zip
```



Llama-2-7b-chat-hf	yesterday
Llama-2-7b-hf	3 days ago
llama-main	3 days ago
stanford_alpaca-main	10 days ago
tmp	3 days ago

Step 13:

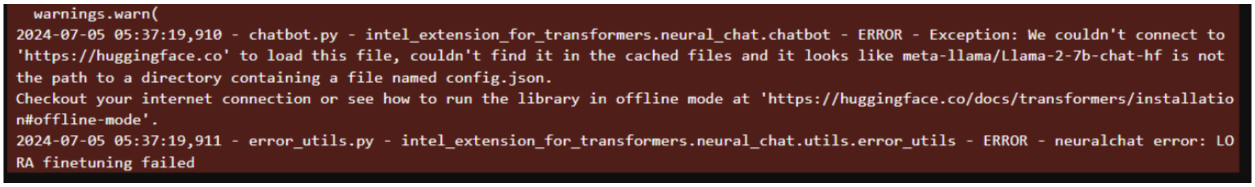
Run the notebook



```
[ ]: from transformers import TrainingArguments
     from intel_extension_for_transformers.neural_chat.config import (
         ModelArguments,
         DataArguments,
         FinetuningArguments,
         TextGenerationFinetuningConfig,
     )
     from intel_extension_for_transformers.neural_chat.chatbot import finetune_model
     model_args = ModelArguments(model_name_or_path="meta-llama/Llama-2-7b-chat-hf")
     data_args = DataArguments(train_file="alpaca_data.json", validation_split_percentage=1)
     training_args = TrainingArguments(
         output_dir='./tmp',
         do_train=True,
         do_eval=True,
         num_train_epochs=3,
         overwrite_output_dir=True,
         per_device_train_batch_size=4,
         per_device_eval_batch_size=4,
         gradient_accumulation_steps=2,
         save_strategy="no",
         log_level="info",
         save_total_limit=2,
         bf16=True,
     )
     finetune_args = FinetuningArguments()
     finetune_cfg = TextGenerationFinetuningConfig(
                 model_args=model_args,
                 data_args=data_args,
                 training_args=training_args,
                 finetune_args=finetune_args,
             )
     finetune_model(finetune_cfg)
```

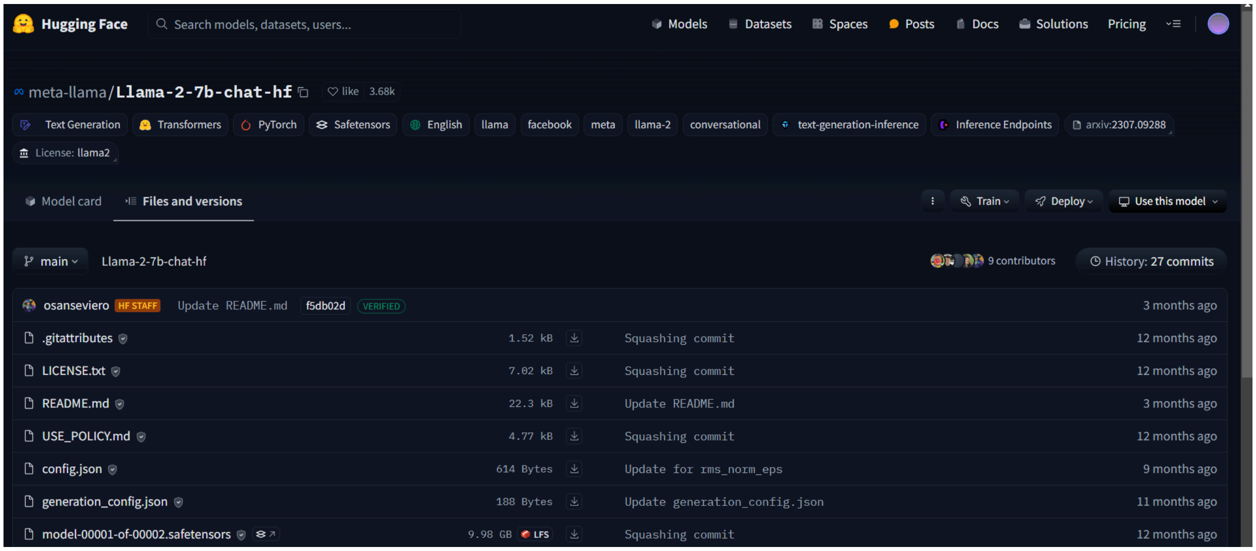
Step 14:

In certain cases the notebook may show an error saying the huggingface token is not able to connect to the meta-llama/Llama-2-7b-chat-hf model
The error msg indicates



```
  warnings.warn(
2024-07-05 05:37:19,910 - chatbot.py - intel_extension_for_transformers.neural_chat.chatbot - ERROR - Exception: We couldn't connect to
'https://huggingface.co' to load this file, couldn't find it in the cached files and it looks like meta-llama/Llama-2-7b-chat-hf is not
the path to a directory containing a file named config.json.
Checkout your internet connection or see how to run the library in offline mode at 'https://huggingface.co/docs/transformers/installatio
n#offline-mode'.
2024-07-05 05:37:19,911 - error_utils.py - intel_extension_for_transformers.neural_chat.utils.error_utils - ERROR - neuralchat error: LO
RA finetuning failed
```

The only way to run the notebook is by loading the meta lama mode in offline mode ; inorder to do so we need to download the entire files of lama model



meta-llama/Llama-2-7b-chat-hf

Text Generation · Transformers · PyTorch · Safetensors · English · llama · facebook · meta · llama-2 · conversational · text-generation-inference · Inference Endpoints · arxiv:2307.09288

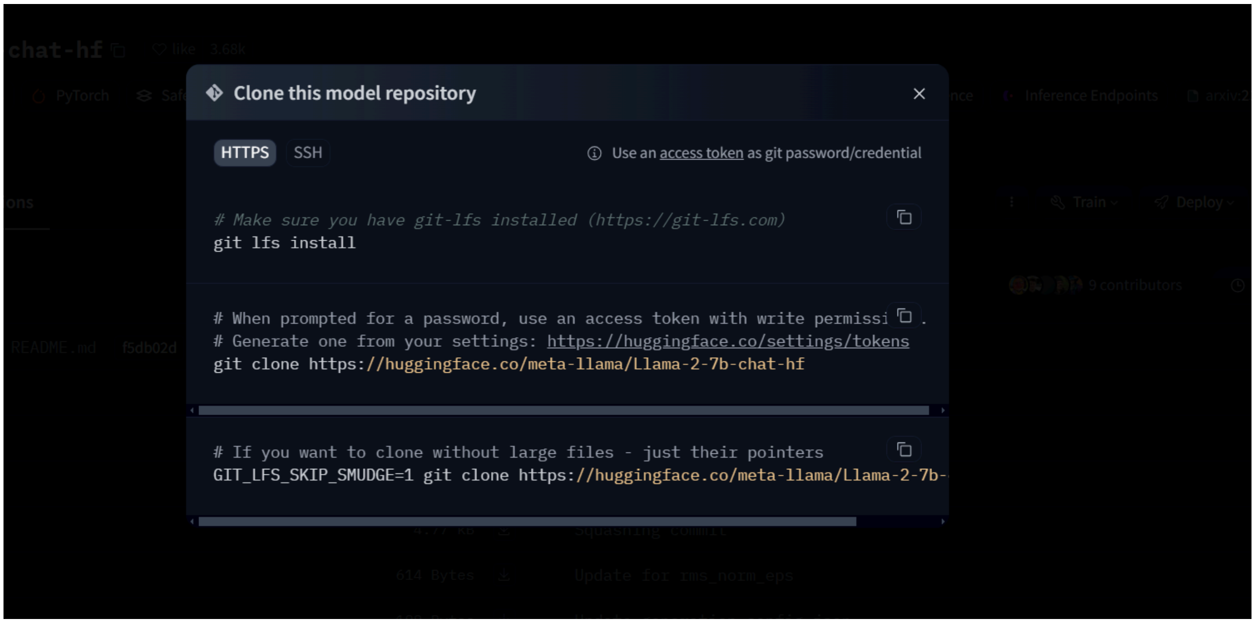
License: llama2

Llama-2-7b-chat-hf · 9 contributors · History: 27 commits

osanseviero HF STAFF Update README.md f5db02d VERIFIED 3 months ago

.gitattributes	1.52 kB	Squashing commit	12 months ago
LICENSE.txt	7.02 kB	Squashing commit	12 months ago
README.md	22.3 kB	Update README.md	3 months ago
USE_POLICY.md	4.77 kB	Squashing commit	12 months ago
config.json	614 Bytes	Update for rms_norm_eps	9 months ago
generation_config.json	188 Bytes	Update generation_config.json	11 months ago
model-00001-of-00002.safetensors	9.98 GB LFS	Squashing commit	12 months ago

Or we can clone the entire meta lama repo

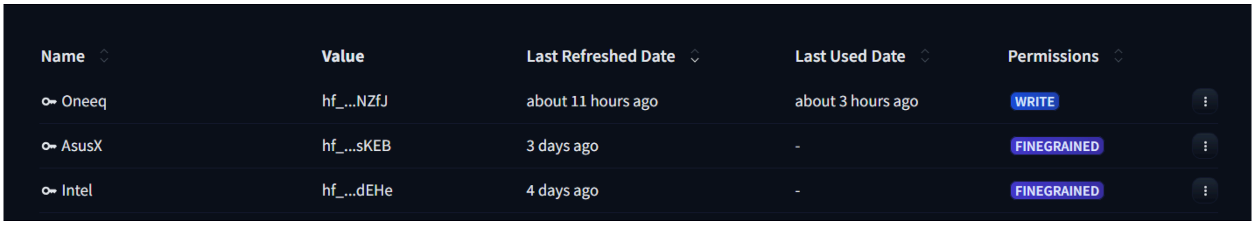

Clone this model repository
HTTPS
SSH
Use an access token as git password/credential

```
# Make sure you have git-lfs installed (https://git-lfs.com)
git lfs install
```

```
# When prompted for a password, use an access token with write permissi .
# Generate one from your settings: https://huggingface.co/settings/tokens
git clone https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
```

```
# If you want to clone without large files - just their pointers
GIT_LFS_SKIP_SMUDGE=1 git clone https://huggingface.co/meta-llama/Llama-2-7b-
```

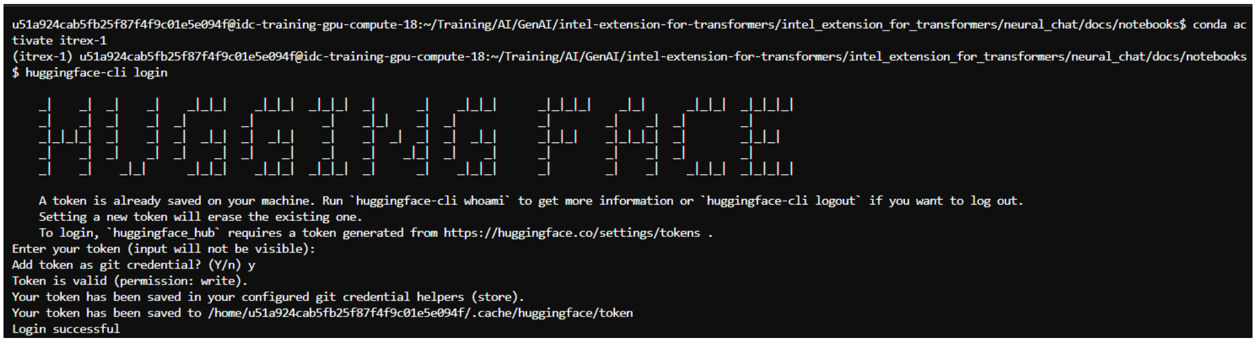
Inorder to clone this repo we need a Hugging face token with write permission



Name	Value	Last Refreshed Date	Last Used Date	Permissions
Oneeq	hf_...NZfJ	about 11 hours ago	about 3 hours ago	WRITE
AsusX	hf_...sKEB	3 days ago	-	FINEGRAINED
Intel	hf_...dEHe	4 days ago	-	FINEGRAINED

Step 16:

Now we need to login to the hugging face idd using the write permitted token

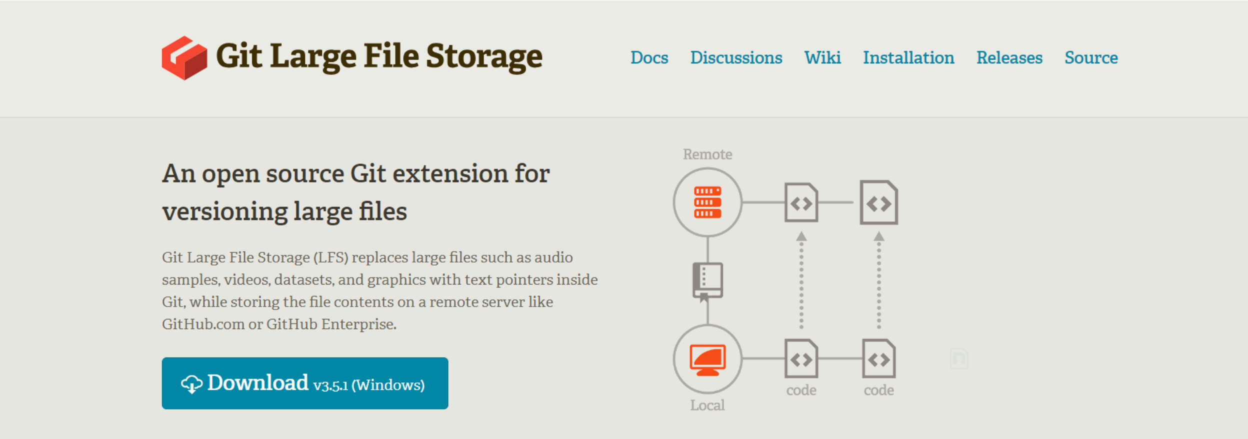


```
u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks$ conda ac
tivate itrex-1
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ huggingface-cli login

    _|    _|  _|    _|    _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|_|_|_|    _|_|      _|_|_|  _|_|_|_|
    _|    _|  _|    _|  _|        _|          _|    _|_|    _|  _|            _|        _|    _|  _|        _|
    _|_|_|_|  _|    _|  _|  _|_|  _|  _|_|    _|    _|  _|  _|  _|  _|_|      _|_|_|    _|_|_|_|  _|        _|_|_|
    _|    _|  _|    _|  _|    _|  _|    _|    _|    _|    _|_|  _|    _|      _|        _|    _|  _|        _|
    _|    _|    _|_|      _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|        _|    _|    _|_|_|  _|_|_|_|

    A token is already saved on your machine. Run `huggingface-cli whoami` to get more information or `huggingface-cli logout` if you want to log out.
    Setting a new token will erase the existing one.
    To login, `huggingface_hub` requires a token generated from https://huggingface.co/settings/tokens .
Enter your token (input will not be visible):
Add token as git credential? (Y/n) y
Token is valid (permission: write).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/token
Login successful
```

Step 17:

Since meta lama model repo huge we cannot clone it directly ; in order to clone it in windows we need to install git lfs :

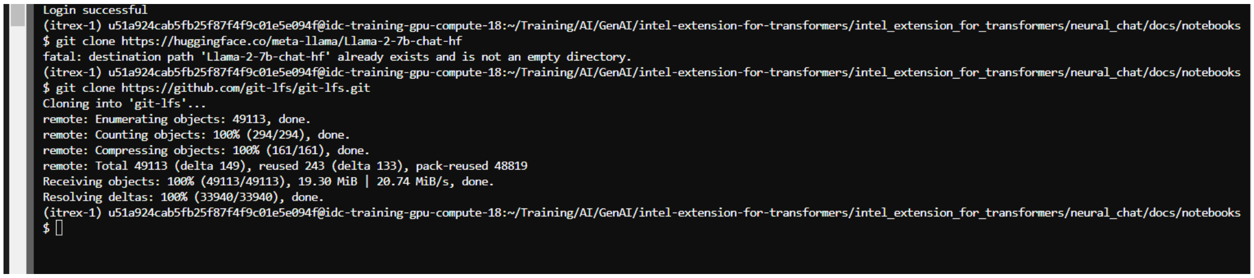
(Git large file storage) which can be done by manually installing it from the git website



An open source Git extension for versioning large files

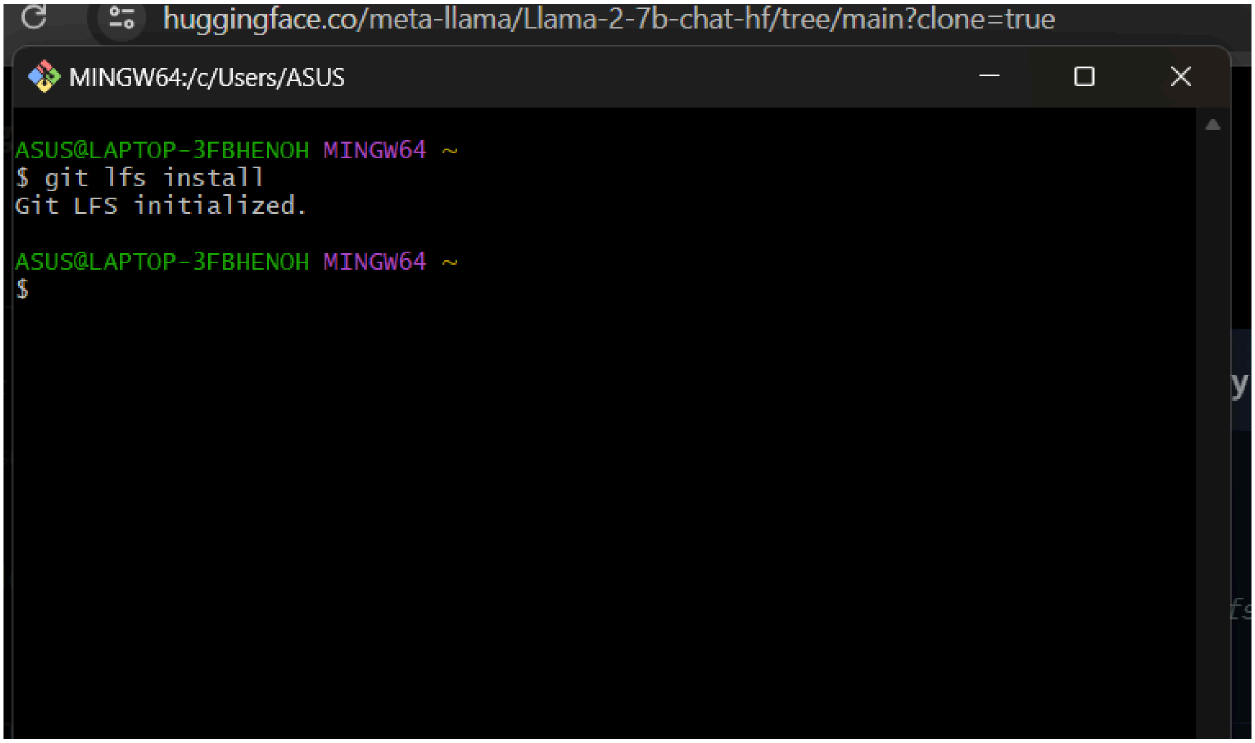
Git Large File Storage (LFS) replaces large files such as audio samples, videos, datasets, and graphics with text pointers inside Git, while storing the file contents on a remote server like GitHub.com or GitHub Enterprise.

Or by clone git lfs repo into the terminal



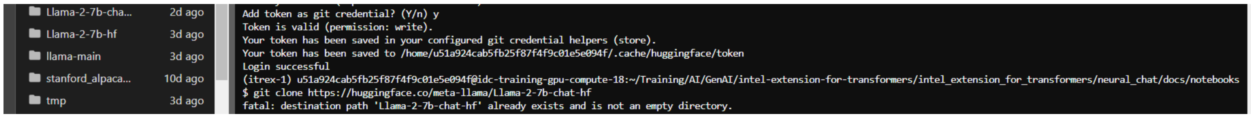
```
Login successful
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ git clone https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
fatal: destination path 'Llama-2-7b-chat-hf' already exists and is not an empty directory.
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ git clone https://github.com/git-lfs/git-lfs.git
Cloning into 'git-lfs'...
remote: Enumerating objects: 49113, done.
remote: Counting objects: 100% (294/294), done.
remote: Compressing objects: 100% (161/161), done.
remote: Total 49113 (delta 149), reused 243 (delta 133), pack-reused 48819
Receiving objects: 100% (49113/49113), 19.30 MiB | 20.74 MiB/s, done.
Resolving deltas: 100% (33940/33940), done.
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ 
```

Step 18:
Once the installation or cloning is completely done
In case of manual installation we may need to activate lfs in git bash


huggingface.co/meta-llama/Llama-2-7b-chat-hf/tree/main?clone=true
MINGW64:/c/Users/ASUS

```
ASUS@LAPTOP-3FBHENOH MINGW64 ~
$ git lfs install
Git LFS initialized.

ASUS@LAPTOP-3FBHENOH MINGW64 ~
$
```

In case of cloning lfs ; we can directly clone the meta lama model



```
Add token as git credential? (Y/n) y
Token is valid (permission: write).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /home/u51a924cab5fb25f87f4f9c01e5e094f/.cache/huggingface/token
Login successful
(itrex-1) u51a924cab5fb25f87f4f9c01e5e094f@idc-training-gpu-compute-18:~/Training/AI/GenAI/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/docs/notebooks
$ git clone https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
fatal: destination path 'Llama-2-7b-chat-hf' already exists and is not an empty directory.
```

Step 19:

Once the model is completely cloned or installed we can run the notebook with the alpaca dataset in the correct directory . Similarly we can finetune 3 notebooks by accessing the hugging face tokens or by manually cloning or installing the required datasets in the correct directories