phi2_merged_gguf

March 1, 2024

[2]:
```
!pip install datasets
!pip install -q -U torch=='2.0.0'
!pip install -q -U accelerate=='0.25.0' peft=='0.7.1' bitsandbytes=='0.41.3.
 ↪post2' trl=='0.7.4'
!pip install -q -U transformers einops
```

```
Collecting datasets
  Downloading datasets-2.17.1-py3-none-any.whl (536 kB)
                             536.7/536.7

kB 11.4 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from datasets) (3.13.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from datasets) (1.25.2)
Requirement already satisfied: pyarrow>=12.0.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (14.0.2)
Requirement already satisfied: pyarrow-hotfix in /usr/local/lib/python3.10/dist-
packages (from datasets) (0.6)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                             116.3/116.3

kB 10.5 MB/s eta 0:00:00
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-
packages (from datasets) (1.5.3)
Requirement already satisfied: requests>=2.19.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (4.66.2)
Requirement already satisfied: xxhash in /usr/local/lib/python3.10/dist-packages
(from datasets) (3.4.1)
Collecting multiprocess (from datasets)
  Downloading multiprocess-0.70.16-py310-none-any.whl (134 kB)
                             134.8/134.8

kB 13.0 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]<=2023.10.0,>=2023.1.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (2023.6.0)
```

```
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-
packages (from datasets) (3.9.3)
Requirement already satisfied: huggingface-hub>=0.19.4 in
/usr/local/lib/python3.10/dist-packages (from datasets) (0.20.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from datasets) (23.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (6.0.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.19.4->datasets)
(4.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests>=2.19.0->datasets) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(2024.2.2)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->datasets) (2023.4)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.1->pandas->datasets) (1.16.0)
Installing collected packages: dill, multiprocess, datasets
Successfully installed datasets-2.17.1 dill-0.3.8 multiprocess-0.70.16
                           619.9/619.9
MB 2.3 MB/s eta 0:00:00
                           21.0/21.0 MB
65.0 MB/s eta 0:00:00
```

```
                            849.3/849.3
kB 55.2 MB/s eta 0:00:00
                            11.8/11.8 MB
74.8 MB/s eta 0:00:00
                            557.1/557.1
MB 2.9 MB/s eta 0:00:00
                            317.1/317.1
MB 3.1 MB/s eta 0:00:00
                            168.4/168.4
MB 9.8 MB/s eta 0:00:00
                            54.6/54.6 MB
30.0 MB/s eta 0:00:00
                            102.6/102.6
MB 10.9 MB/s eta 0:00:00
                            173.2/173.2
MB 5.8 MB/s eta 0:00:00
                            177.1/177.1
MB 5.5 MB/s eta 0:00:00
                            98.6/98.6 kB
14.4 MB/s eta 0:00:00
                            63.3/63.3 MB
25.5 MB/s eta 0:00:00
                            153.0/153.0
kB 20.2 MB/s eta 0:00:00
  Installing build dependencies … done
  Getting requirements to build wheel … done
  Installing backend dependencies … done
  Preparing metadata (pyproject.toml) … done
  Building wheel for lit (pyproject.toml) … done
```

```
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
torchaudio 2.1.0+cu121 requires torch==2.1.0, but you have torch 2.0.0 which is
incompatible.
torchdata 0.7.0 requires torch==2.1.0, but you have torch 2.0.0 which is
incompatible.
torchtext 0.16.0 requires torch==2.1.0, but you have torch 2.0.0 which is
incompatible.
torchvision 0.16.0+cu121 requires torch==2.1.0, but you have torch 2.0.0 which
is incompatible.
                              265.7/265.7
kB 6.6 MB/s eta 0:00:00
                              168.3/168.3
kB 22.0 MB/s eta 0:00:00
                              92.6/92.6 MB
17.4 MB/s eta 0:00:00
                              133.9/133.9
kB 15.5 MB/s eta 0:00:00
                              79.8/79.8 kB
10.1 MB/s eta 0:00:00
                              44.6/44.6 kB
1.6 MB/s eta 0:00:00
```

[3]:
```python
# Data Exploration Imports
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split

import os
os.environ["CUDA_VISIBLE_DEVICES"] = "0"
os.environ["TOKENIZERS_PARALLELISM"] = "false"

import warnings
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd
from tqdm import tqdm
```

```
import bitsandbytes as bnb
import torch
import torch.nn as nn
import transformers
from datasets import Dataset
from peft import LoraConfig, PeftConfig
from trl import SFTTrainer
from transformers import (AutoModelForCausalLM,
                          AutoTokenizer,
                          BitsAndBytesConfig,
                          TrainingArguments,
                          pipeline,
                          logging)
from sklearn.metrics import (accuracy_score,
                             classification_report,
                             confusion_matrix)
```

[4]:
```
from google.colab import drive

drive.mount('drive')
```

```
Mounted at drive
```

[5]:
```
from huggingface_hub import notebook_login

notebook_login()
```

```
VBox(children=(HTML(value='<center> <img\nsrc=https://huggingface.co/front/
↪assets/huggingface_logo-noborder.sv…
```

[6]:
```
from peft import PeftModel

model_name = "microsoft/phi-2"

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    trust_remote_code=True,
    device_map="auto",
    torch_dtype=torch.float16
)

peftmodel = PeftModel.from_pretrained(model, "pavankumarbalijepalli/
↪phi2-nl2sql-lora")
peftmodel = peftmodel.merge_and_unload()

tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
```

```
config.json:   0%|          | 0.00/863 [00:00<?, ?B/s]
```

```
configuration_phi.py:   0%|          | 0.00/9.26k [00:00<?, ?B/s]

A new version of the following files was downloaded from
https://huggingface.co/microsoft/phi-2:
- configuration_phi.py
. Make sure to double-check they do not contain any added malicious code. To
avoid downloading new versions of the code file, you can pin a revision.

modeling_phi.py:   0%|          | 0.00/62.7k [00:00<?, ?B/s]

A new version of the following files was downloaded from
https://huggingface.co/microsoft/phi-2:
- modeling_phi.py
. Make sure to double-check they do not contain any added malicious code. To
avoid downloading new versions of the code file, you can pin a revision.

model.safetensors.index.json:   0%|          | 0.00/35.7k [00:00<?, ?B/s]

Downloading shards:   0%|          | 0/2 [00:00<?, ?it/s]

model-00001-of-00002.safetensors:   0%|          | 0.00/5.00G [00:00<?, ?B/s]

model-00002-of-00002.safetensors:   0%|          | 0.00/564M [00:00<?, ?B/s]

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]

generation_config.json:   0%|          | 0.00/124 [00:00<?, ?B/s]

adapter_config.json:   0%|          | 0.00/617 [00:00<?, ?B/s]

adapter_model.safetensors:   0%|          | 0.00/94.4M [00:00<?, ?B/s]

PeftModelForCausalLM(
  (base_model): LoraModel(
    (model): PhiForCausalLM(
      (model): PhiModel(
        (embed_tokens): Embedding(51200, 2560)
        (embed_dropout): Dropout(p=0.0, inplace=False)
        (layers): ModuleList(
          (0-31): 32 x PhiDecoderLayer(
            (self_attn): PhiAttention(
              (q_proj): lora.Linear(
                (base_layer): Linear(in_features=2560, out_features=2560,
bias=True)
                (lora_dropout): ModuleDict(
                  (default): Dropout(p=0.05, inplace=False)
                )
                (lora_A): ModuleDict(
                  (default): Linear(in_features=2560, out_features=16,
bias=False)
                )
                (lora_B): ModuleDict(
                  (default): Linear(in_features=16, out_features=2560,
bias=False)
```

```
          )
          (lora_embedding_A): ParameterDict()
          (lora_embedding_B): ParameterDict()
        )
        (k_proj): lora.Linear(
          (base_layer): Linear(in_features=2560, out_features=2560,
bias=True)
          (lora_dropout): ModuleDict(
            (default): Dropout(p=0.05, inplace=False)
          )
          (lora_A): ModuleDict(
            (default): Linear(in_features=2560, out_features=16,
bias=False)
          )
          (lora_B): ModuleDict(
            (default): Linear(in_features=16, out_features=2560,
bias=False)
          )
          (lora_embedding_A): ParameterDict()
          (lora_embedding_B): ParameterDict()
        )
        (v_proj): lora.Linear(
          (base_layer): Linear(in_features=2560, out_features=2560,
bias=True)
          (lora_dropout): ModuleDict(
            (default): Dropout(p=0.05, inplace=False)
          )
          (lora_A): ModuleDict(
            (default): Linear(in_features=2560, out_features=16,
bias=False)
          )
          (lora_B): ModuleDict(
            (default): Linear(in_features=16, out_features=2560,
bias=False)
          )
          (lora_embedding_A): ParameterDict()
          (lora_embedding_B): ParameterDict()
        )
        (dense): lora.Linear(
          (base_layer): Linear(in_features=2560, out_features=2560,
bias=True)
          (lora_dropout): ModuleDict(
            (default): Dropout(p=0.05, inplace=False)
          )
          (lora_A): ModuleDict(
            (default): Linear(in_features=2560, out_features=16,
bias=False)
          )
```

```
                (lora_B): ModuleDict(
                  (default): Linear(in_features=16, out_features=2560,
bias=False)
                )
                (lora_embedding_A): ParameterDict()
                (lora_embedding_B): ParameterDict()
              )
              (rotary_emb): PhiRotaryEmbedding()
            )
            (mlp): PhiMLP(
              (activation_fn): NewGELUActivation()
              (fc1): lora.Linear(
                (base_layer): Linear(in_features=2560, out_features=10240,
bias=True)
                (lora_dropout): ModuleDict(
                  (default): Dropout(p=0.05, inplace=False)
                )
                (lora_A): ModuleDict(
                  (default): Linear(in_features=2560, out_features=16,
bias=False)
                )
                (lora_B): ModuleDict(
                  (default): Linear(in_features=16, out_features=10240,
bias=False)
                )
                (lora_embedding_A): ParameterDict()
                (lora_embedding_B): ParameterDict()
              )
              (fc2): lora.Linear(
                (base_layer): Linear(in_features=10240, out_features=2560,
bias=True)
                (lora_dropout): ModuleDict(
                  (default): Dropout(p=0.05, inplace=False)
                )
                (lora_A): ModuleDict(
                  (default): Linear(in_features=10240, out_features=16,
bias=False)
                )
                (lora_B): ModuleDict(
                  (default): Linear(in_features=16, out_features=2560,
bias=False)
                )
                (lora_embedding_A): ParameterDict()
                (lora_embedding_B): ParameterDict()
              )
            )
            (input_layernorm): LayerNorm((2560,), eps=1e-05,
elementwise_affine=True)
```

```
            (resid_dropout): Dropout(p=0.1, inplace=False)
          )
        )
        (final_layernorm): LayerNorm((2560,), eps=1e-05,
elementwise_affine=True)
      )
      (lm_head): Linear(in_features=2560, out_features=51200, bias=True)
    )
  )
)
```

```
tokenizer_config.json:   0%|          | 0.00/7.34k [00:00<?, ?B/s]
```

```
vocab.json:   0%|          | 0.00/798k [00:00<?, ?B/s]
```

```
merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
```

```
tokenizer.json:   0%|          | 0.00/2.11M [00:00<?, ?B/s]
```

```
added_tokens.json:   0%|          | 0.00/1.08k [00:00<?, ?B/s]
```

```
special_tokens_map.json:   0%|          | 0.00/99.0 [00:00<?, ?B/s]
```

```
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
```

[7]:
```
peftmodel
```

[7]:
```
PhiForCausalLM(
  (model): PhiModel(
    (embed_tokens): Embedding(51200, 2560)
    (embed_dropout): Dropout(p=0.0, inplace=False)
    (layers): ModuleList(
      (0-31): 32 x PhiDecoderLayer(
        (self_attn): PhiAttention(
          (q_proj): Linear(in_features=2560, out_features=2560, bias=True)
          (k_proj): Linear(in_features=2560, out_features=2560, bias=True)
          (v_proj): Linear(in_features=2560, out_features=2560, bias=True)
          (dense): Linear(in_features=2560, out_features=2560, bias=True)
          (rotary_emb): PhiRotaryEmbedding()
        )
        (mlp): PhiMLP(
          (activation_fn): NewGELUActivation()
          (fc1): Linear(in_features=2560, out_features=10240, bias=True)
          (fc2): Linear(in_features=10240, out_features=2560, bias=True)
        )
        (input_layernorm): LayerNorm((2560,), eps=1e-05,
elementwise_affine=True)
        (resid_dropout): Dropout(p=0.1, inplace=False)
      )
    )
```

```
    (final_layernorm): LayerNorm((2560,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=2560, out_features=51200, bias=True)
)
```

[8]:
```
peftmodel.save_pretrained("phi2-nl2sql-lora-merged")
tokenizer.save_pretrained("phi2-nl2sql-lora-merged")
```

[8]:
```
('phi2-nl2sql-lora-merged/tokenizer_config.json',
 'phi2-nl2sql-lora-merged/special_tokens_map.json',
 'phi2-nl2sql-lora-merged/vocab.json',
 'phi2-nl2sql-lora-merged/merges.txt',
 'phi2-nl2sql-lora-merged/added_tokens.json',
 'phi2-nl2sql-lora-merged/tokenizer.json')
```

[ ]:
```
import os

if not os.path.exists('out'):
  os.makedirs('out')
```

[ ]:
```
!git clone https://github.com/ggerganov/llama.cpp
!cd llama.cpp && make
!python3 -m pip install -r requirements.txt

!python llama.cpp/convert-hf-to-gguf.py phi2-nl2sql-lora-merged --outfile out/
 ↪phi2-nl2sql-lora-merged-f16.gguf --outtype f16
```

[9]:
```
from zipfile import ZipFile
import os

def get_all_file_paths(directory):
    file_paths = []
    for root, directories, files in os.walk(directory):
        for filename in files:
            filepath = os.path.join(root, filename)
            file_paths.append(filepath)
    return file_paths

def zip_it(directory: str, file_name: str):
    file_paths = get_all_file_paths(directory)
    print('Following files will be zipped:')
    for file_path in file_paths:
        print(file_path)
    with ZipFile(file_name,'w') as zip:
        for file in file_paths:
            zip.write(file)
    print('All files zipped successfully!')
```

[17]:
```
zip_it("phi2-nl2sql-lora-merged", "merged.zip")

from datetime import datetime
import shutil

name = 'merged_' + datetime.now().strftime("%Y_%m_%d_%H_%M_%S") + '.zip'
shutil.move("/content/merged.zip", "/content/drive/MyDrive/phi2_finetune/" +␣
 ↪name)
```

```
Following files will be zipped:
pavankumarbalijepalli/phi2-nl2sql-lora-merged/tokenizer_config.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/tokenizer.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/model-00001-of-00002.safetensors
pavankumarbalijepalli/phi2-nl2sql-lora-merged/added_tokens.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/model-00002-of-00002.safetensors
pavankumarbalijepalli/phi2-nl2sql-lora-merged/special_tokens_map.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/generation_config.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/model.safetensors.index.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/config.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/vocab.json
pavankumarbalijepalli/phi2-nl2sql-lora-merged/merges.txt
All files zipped successfully!
```

[17]:
```
'/content/drive/MyDrive/phi2_finetune/merged_2024_02_21_09_19_47.zip'
```

[13]:
```
zip_it("out", "phi2-nl2sql-lora-merged-f16.zip")

from datetime import datetime
import shutil

name = 'merged_gguf_' + datetime.now().strftime("%Y_%m_%d_%H_%M_%S") + '.zip'
shutil.move("/content/phi2-nl2sql-lora-merged-f16.zip", "/content/drive/MyDrive/
 ↪phi2_finetune/" + name)
```

[13]:
```
'/content/drive/MyDrive/phi2_finetune/merged_gguf_2024_02_21_10_35_00.zip'
```

[ ]: