finetune-gemma2b

August 18, 2024

[1]:
```
!pip install -q peft==0.4.0 bitsandbytes==0.40.2  trl==0.4.7 datasets==2.17.0
```

[2]:
```
!pip install accelerate==0.27.2
```

```
Collecting accelerate==0.27.2
  Downloading accelerate-0.27.2-py3-none-any.whl.metadata (18 kB)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from accelerate==0.27.2) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from accelerate==0.27.2) (24.1)
Requirement already satisfied: psutil in /usr/local/lib/python3.10/dist-packages
(from accelerate==0.27.2) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.10/dist-packages
(from accelerate==0.27.2) (6.0.2)
Requirement already satisfied: torch>=1.10.0 in /usr/local/lib/python3.10/dist-
packages (from accelerate==0.27.2) (2.3.1+cu121)
Requirement already satisfied: huggingface-hub in
/usr/local/lib/python3.10/dist-packages (from accelerate==0.27.2) (0.23.5)
Requirement already satisfied: safetensors>=0.3.1 in
/usr/local/lib/python3.10/dist-packages (from accelerate==0.27.2) (0.4.4)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from torch>=1.10.0->accelerate==0.27.2) (3.15.4)
Requirement already satisfied: typing-extensions>=4.8.0 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(4.12.2)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages
(from torch>=1.10.0->accelerate==0.27.2) (1.13.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-
packages (from torch>=1.10.0->accelerate==0.27.2) (3.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from torch>=1.10.0->accelerate==0.27.2) (3.1.4)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages
(from torch>=1.10.0->accelerate==0.27.2) (2023.10.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(12.1.105)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
```

```
(12.1.105)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(12.1.105)
Requirement already satisfied: nvidia-cudnn-cu12==8.9.2.26 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(8.9.2.26)
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(12.1.3.1)
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(11.0.2.54)
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(10.3.2.106)
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(11.4.5.107)
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(12.1.0.106)
Requirement already satisfied: nvidia-nccl-cu12==2.20.5 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(2.20.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.27.2)
(12.1.105)
Requirement already satisfied: triton==2.3.1 in /usr/local/lib/python3.10/dist-
packages (from torch>=1.10.0->accelerate==0.27.2) (2.3.1)
Requirement already satisfied: nvidia-nvjitlink-cu12 in
/usr/local/lib/python3.10/dist-packages (from nvidia-cusolver-
cu12==11.4.5.107->torch>=1.10.0->accelerate==0.27.2) (12.6.20)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from huggingface-hub->accelerate==0.27.2) (2.32.3)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.10/dist-
packages (from huggingface-hub->accelerate==0.27.2) (4.66.5)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from
jinja2->torch>=1.10.0->accelerate==0.27.2) (2.1.5)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub->accelerate==0.27.2) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->huggingface-hub->accelerate==0.27.2) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub->accelerate==0.27.2) (2.0.7)
```

```
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub->accelerate==0.27.2) (2024.7.4)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from
sympy->torch>=1.10.0->accelerate==0.27.2) (1.3.0)
Downloading accelerate-0.27.2-py3-none-any.whl (279 kB)
                         280.0/280.0 kB
7.7 MB/s eta 0:00:00
Installing collected packages: accelerate
  Attempting uninstall: accelerate
    Found existing installation: accelerate 0.21.0
    Uninstalling accelerate-0.21.0:
      Successfully uninstalled accelerate-0.21.0
Successfully installed accelerate-0.27.2
```

[3]:

```
!pip install transformers==4.38.2
```

```
Requirement already satisfied: transformers==4.38.2 in
/usr/local/lib/python3.10/dist-packages (4.38.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers==4.38.2) (3.15.4)
Requirement already satisfied: huggingface-hub<1.0,>=0.19.3 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.38.2) (0.23.5)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from transformers==4.38.2) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.38.2) (24.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from transformers==4.38.2) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.38.2) (2024.5.15)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers==4.38.2) (2.32.3)
Requirement already satisfied: tokenizers<0.19,>=0.14 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.38.2) (0.15.2)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.38.2) (0.4.4)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-
packages (from transformers==4.38.2) (4.66.5)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.19.3->transformers==4.38.2) (2023.10.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.19.3->transformers==4.38.2) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers==4.38.2)
```

```
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers==4.38.2) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers==4.38.2)
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers==4.38.2)
(2024.7.4)
```

[4]:

```python
import os
import transformers
import torch
from google.colab import userdata
from datasets import load_dataset
from trl import SFTTrainer
from peft import LoraConfig
from transformers import AutoTokenizer, AutoModelForCausalLM
from transformers import BitsAndBytesConfig
```

[5]:

```python
os.environ["HF_TOKEN"]= userdata.get("HF_TOKEN")
```

[6]:

```python
model_id = "google/gemma-2b"
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
```

[7]:

```python
tokenizer = AutoTokenizer.from_pretrained(model_id, token=os.
  ↪environ["HF_TOKEN"])
model = AutoModelForCausalLM.from_pretrained(
    model_id, quantization_config=bnb_config,
    device_map={"":0},
    token=os.environ["HF_TOKEN"]
)
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132:
FutureWarning: `resume_download` is deprecated and will be removed in version
1.0.0. Downloads always resume when possible. If you want to force a new
download, use `force_download=True`.
  warnings.warn(
```

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

```
You are calling `save_pretrained` to a 4-bit converted model, but your
`bitsandbytes` version doesn't support it. If you want to save 4-bit models,
make sure to have `bitsandbytes>=0.41.3` installed.
```

[8]:
```
text = "Quote: Imagination is more,"
device= "cuda:0"
inputs = tokenizer(text,return_tensors="pt").to(device)

outputs = model.generate(**inputs, max_new_tokens=50)
print(tokenizer.decode(outputs[0], skip_special_tokens=True))
```

```
Quote: Imagination is more, than knowledge.

I am a self-taught artist, born in 1985 in the beautiful city of Porto Alegre,
Brazil.

I have always been interested in art, but I never thought I would be able to
make a living
```

[9]:
```
lora_config = LoraConfig(
    r=8,
    target_modules = [
        "q_proj","o_proj","k_proj","v_proj","gate_proj","up_proj","down_proj"
    ],
    task_type="CAUSAL_LM",
)
```

[10]:
```
from datasets import load_dataset

data = load_dataset("Abirate/english_quotes")
data = data.map(lambda samples: tokenizer(samples["quote"]), batched=True)
```

[11]:
```
data['train']
```

[11]:
```
Dataset({
    features: ['quote', 'author', 'tags', 'input_ids', 'attention_mask'],
    num_rows: 2508
})
```

[12]:
```
def formating_func(example):
  text = f"Quote: {example['quote'][0]}\nAuthor: {example['author'][0]}"
  return [text]
```

[13]:
```
training_arguments = transformers.TrainingArguments(

    per_device_train_batch_size= 1,
    gradient_accumulation_steps= 4,
    warmup_steps=2,
    max_steps=100,
    learning_rate=2e-4,
    fp16= True,
    logging_steps=1,
    output_dir="outputs",
    optim="paged_adamw_8bit"
)
# Setting sft parameters
trainer = SFTTrainer(
    model=model,
    train_dataset=data['train'],
    peft_config=lora_config,
    max_seq_length= None,
    tokenizer=tokenizer,
    args=training_arguments,
    packing= False,
    formatting_func=formating_func
)
```

```
/usr/local/lib/python3.10/dist-packages/peft/utils/other.py:102: FutureWarning:
prepare_model_for_int8_training is deprecated and will be removed in a future
version. Use prepare_model_for_kbit_training instead.
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/trl/trainer/sft_trainer.py:159:
UserWarning: You didn't pass a `max_seq_length` argument to the SFTTrainer, this
will default to 1024
  warnings.warn(
```

[14]:

```
trainer.train()
```

```
You are using 8-bit optimizers with a version of `bitsandbytes` < 0.41.1. It is
recommended to update your version as a major bug has been fixed in 8-bit
optimizers.
`use_cache=True` is incompatible with gradient checkpointing. Setting
`use_cache=False`.
```

```
<IPython.core.display.HTML object>
```

[14]:

```
TrainOutput(global_step=100, training_loss=0.13250487179029732,
metrics={'train_runtime': 99.4017, 'train_samples_per_second': 4.024,
'train_steps_per_second': 1.006, 'total_flos': 54994550906880.0, 'train_loss':
0.13250487179029732, 'epoch': 66.67})
```

[16]:

```
text = "Quote: The opposite of love is not hate"
device = "cuda:0"
inputs = tokenizer(text, return_tensors="pt").to(device)

outputs= model.generate(**inputs, max_new_tokens=20)
print(tokenizer.decode(outputs[0], skip_special_tokens=True))
```

```
Quote: The opposite of love is not hate
```

```
Author: Aung San Suu Kyi
Source: 19Quote: Be yourself;
```

[ ]:

```
```