

深度學習

期末作業程式執行過程截圖

四資四甲

吳宥錚

B10090012

目錄:

Bert_embedding.ipynb3

Finetune_Chinese_Weibo.ipynb.....4

PEFT_SFTTrainer.ipynb7

Pipeline_with_Deepseek_R1.ipynb.....13

Bert_embedding.ipynb

```
11 秒
!pip install transformers

Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.50.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.26.0->transformers) (2025.3.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.26.0->transformers) (4.13.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.31)
```

transformer為Huggingface的套件，需要Huggingface的token授權

- 請至huggingface網站註冊並申請token，並把token加入colab中，名為**HF_TOKEN**的環境變數
- 可參考筆記[Huggingface與Transformer套件](#)
- 從列印出來的模型參數中，可以看到Bert的embedding token數量有30522個(類似字典裏面的單字數量)，每個token有768維的隱向量空間(Latent vector space)

```
50 秒
[2] from transformers import AutoTokenizer, AutoModel
model=AutoModel.from_pretrained("bert-base-uncased"):
print(model)

config.json: 100% ██████████ 570/570 [00:00<00:00, 43.7kB/s]
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download. For better performance, instal
WARNING:huggingface_hub.file_download:Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP do
model.safetensors: 100% ██████████ 440M/440M [00:06<00:00, 54.8MB/s]
BertModel(
  (embeddings): BertEmbeddings(
    (word_embeddings): Embedding(30522, 768, padding_idx=0)
    (position_embeddings): Embedding(512, 768)
    (token_type_embeddings): Embedding(2, 768)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (encoder): BertEncoder(
    (layer): ModuleList(
      (0-11): 12 x BertLayer(
        (attention): BertAttention(
          (self): BertSdpaSelfAttention(
            (query): Linear(in_features=768, out_features=768, bias=True)
            (key): Linear(in_features=768, out_features=768, bias=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (output): BertSelfOutput(
            (dense): Linear(in_features=768, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
        (intermediate): BertIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation()
        )
        (output): BertOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
  )
  (pooler): BertPooler(
    (dense): Linear(in_features=768, out_features=768, bias=True)
    (activation): Tanh()
  )
)
```

Bert的參數總量為109,482,240，約一億個參數

```
0 秒
[3] print(model.num_parameters())

109482240

0 秒
[4] from transformers import AutoTokenizer
tokenizer=AutoTokenizer.from_pretrained("bert-base-uncased")
raw_inputs = [
    "I've been waiting for a HuggingFace course my whole life.",
    "I hate this so much!",
]
inputs = tokenizer(raw_inputs, padding=True, truncation=True, return_tensors="pt")
print(inputs)

tokenizer_config.json: 100% ██████████ 48.0/48.0 [00:00<00:00, 2.79kB/s]
vocab.txt: 100% ██████████ 232k/232k [00:00<00:00, 3.87MB/s]
tokenizer.json: 100% ██████████ 466k/466k [00:00<00:00, 9.59MB/s]
{'input_ids': tensor([[ 101, 1045, 1005, 2310, 2042, 3403, 2005, 1037, 17662, 12172,
 2607, 2026, 2878, 2166, 1012, 102],
 [ 101, 1045, 5223, 2023, 2061, 2172, 999, 102, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]]), 'token_type_ids': tensor([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])}
```

Finetune_Chinese_Weibo.ipynb

利用中文微博評價資料進行Bert微調

```
[ ] | pip install transformers datasets
| pip install evaluate

Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.50.3)
Collecting datasets
  Downloading datasets-3.5.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Collecting xxhash (from datasets)
  Downloading xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2024.12.0,>=2023.1.0 (from fsspec[http]<=2024.12.0,>=2023.1.0->datasets)
  Downloading fsspec-2024.12.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from datasets) (3.11.15)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.2.0)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.18.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.26.0->transformers) (4.13.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Downloading datasets-3.5.0-py3-none-any.whl (491 kB)
----- 491.2/491.2 kB 12.8 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
----- 116.3/116.3 kB 9.7 MB/s eta 0:00:00
Downloading fsspec-2024.12.0-py3-none-any.whl (183 kB)
----- 183.9/183.9 kB 8.0 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
----- 143.5/143.5 kB 9.6 MB/s eta 0:00:00
Downloading xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
----- 194.8/194.8 kB 7.3 MB/s eta 0:00:00
Installing collected packages: xxhash, fsspec, dill, multiprocess, datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2
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
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cublas-cu12 12.5.3.2 which
```

下載微博評價資料

```
[ ] | wget https://github.com/shhuangmust/AI/raw/refs/heads/113-1/weibo_senti_100k.csv

—2025-04-09 09:07:51— https://github.com/shhuangmust/AI/raw/refs/heads/113-1/weibo_senti_100k.csv
Resolving github.com (github.com)... 140.82.116.3
Connecting to github.com (github.com)|140.82.116.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/shhuangmust/AI/refs/heads/113-1/weibo_senti_100k.csv [following]
—2025-04-09 09:07:52— https://raw.githubusercontent.com/shhuangmust/AI/refs/heads/113-1/weibo_senti_100k.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.111.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 19699818 (19M) [application/octet-stream]
Saving to: 'weibo_senti_100k.csv'

weibo_senti_100k.cs 100%[=====>] 18.79M 93.7MB/s in 0.2s

2025-04-09 09:07:53 (93.7 MB/s) - 'weibo_senti_100k.csv' saved [19699818/19699818]
```

讀取Weibo資料集

- 共有119988筆資料

```
from datasets import load_dataset, DatasetDict

ds = load_dataset("csv", data_files="weibo_senti_100k.csv")
print(ds)

Generating train split: 119988/0 [00:01<00:00, 132979.15 examples/s]
DatasetDict({
  train: Dataset({
    features: ['label', 'review'],
    num_rows: 119988
  })
})
```

分割資料集

- 80%訓練(train)資料
- 10%測試(test)資料
- 10%驗證(valid)資料

```
[ ] train_testvalid = ds['train'].train_test_split(test_size=0.2)
test_valid = train_testvalid['test'].train_test_split(test_size=0.5)
dataset = DatasetDict({
  'train': train_testvalid['train'],
  'test': test_valid['test'],
  'valid': test_valid['train']})
```


利用TrainingArguments設定微調參數

```
[ ] from transformers import TrainingArguments
import numpy as np
import evaluate

metric = evaluate.load("accuracy")
def compute_metrics(eval_pred):
    logits, labels = eval_pred
    predictions = np.argmax(logits, axis=-1)
    return metric.compute(predictions=predictions, references=labels)

training_args = TrainingArguments(output_dir="test_trainer_chinese", evaluation_strategy="epoch")
```

Downloading builder script: 100% 4.20k/4.20k [00:00<00:00, 443kB/s]
/usr/local/lib/python3.11/dist-packages/transformers/training_args.py:1611: FutureWarning: `evaluation_strategy` is deprecated and will be removed in
warnings.warn()

利用Trainer進行訓練

- 此處須輸入 wandb key

```
[ ] from transformers import Trainer

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=small_train_dataset,
    eval_dataset=small_eval_dataset,
    compute_metrics=compute_metrics,
)
```

wandb: WARNING The `run_name` is currently set to the same value as `TrainingArguments.output_dir`. If this was not intended, please specify a different
wandb: Using wandb-core as the SDK backend. Please refer to <https://wandb.me/wandb-core> for more information.
wandb: Logging into wandb.ai. (Learn how to deploy a W&B server locally: <https://wandb.me/wandb-server>)
wandb: You can find your API key in your browser here: <https://wandb.ai/authorize>
wandb: Paste an API key from your profile and hit enter:
wandb: WARNING If you're specifying your api key in code, ensure this code is not shared publicly.
wandb: WARNING Consider setting the WANDB_API_KEY environment variable, or running `wandb login` from the command line.
wandb: No netrc file found, creating one.
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
wandb: Currently logged in as: **candy533520** (candy533520-must) to <https://api.wandb.ai>. Use `wandb login --relogin` to force relogin
Tracking run with wandb version 0.19.9
Run data is saved locally in /content/wandb/run-20250409_091019-tntzbamg
Syncing run **test_trainer_chinese** to [Weights & Biases](#) (docs)
View project at <https://wandb.ai/candy533520-must/huggingface>
View run at <https://wandb.ai/candy533520-must/huggingface/runs/tntzbamg>
[3750/3750 1:02:58, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.122000	0.120404	0.979000
2	0.097800	0.078305	0.981400
3	0.074900	0.066147	0.982000

TrainOutput(global_step=3750, training_loss=0.10864674352010091, metrics={'train_runtime': 3825.795, 'train_samples_per_second': 7.842, 'train_steps_per_second': 0.98, 'total_flos': 789333166080000.0, 'train_loss': 0.10864674352010091, 'epoch': 3.0})

利用pipeline進行測試

- LABEL_0: 負面
- LABEL_1: 正面

```
[ ] from transformers import pipeline
pipe = pipeline("sentiment-analysis", model="test_trainer_chinese/checkpoint-1500", tokenizer=tokenizer)
```

Device set to use cuda:0

```
[ ] pipe("我喜歡這個產品")
```

```
[ ] [{"label": "LABEL_1", "score": 0.9995720982551575}]
```

PEFT_SFTTrainer.ipynb

✧ 使用QLORA訓練白話文和文言文互轉的模型(台大資工所2024年應用深度學習作業3)

1. 作業目標

- 本次作業目標是使用QLORA訓練一個白話文和文言文互轉的模型。
- 使用的基礎模型是 `zake7749/gemma-2-2b-it-chinese-kyara-dpo`，這是一個Google Gemma2經過Instruction Tuned及DPO之後的模型。
- 本次作業的資料集是台大資工作業提供的，並沒有經過資料清理，是用簡體硬轉成繁體的，資料集問題很多，這是一個包含白話文和文言文的資料集。

2. 作業步驟

- 本次作業的步驟如下：
 1. 資料前處理
 2. 使用QLORA訓練模型
 3. 模型測試
- 訓練監控
- 本次作業的訓練監控如下：
 1. 訓練過程中的loss
 2. 使用wandb記錄訓練過程
 3. 調整各種參數觀察訓練結果

```
[ ] !pip install transformers datasets torch bitsandbytes peft wandb trl flash-attn nvidia-ml-py3
```

Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.50.3)
Collecting datasets
 Downloading datasets-3.5.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (2.6.0+cu124)
Collecting bitsandbytes
 Downloading bitsandbytes-0.45.5-py3-none-manylinux_2_24_x86_64.whl.metadata (5.0 kB)
Requirement already satisfied: peft in /usr/local/lib/python3.11/dist-packages (0.14.0)
Requirement already satisfied: wandb in /usr/local/lib/python3.11/dist-packages (0.19.9)
Collecting trl
 Downloading trl-0.16.1-py3-none-any.whl.metadata (12 kB)
Collecting flash-attn
 Downloading flash_attn-2.7.4.post1.tar.gz (6.0 MB)
----- 6.0/6.0 MB 31.3 MB/s eta 0:00:00
 Preparing metadata (setup.py) ... done
Collecting nvidia-ml-py3
 Downloading nvidia-ml-py3-7.352.0.tar.gz (19 kB)
 Preparing metadata (setup.py) ... done
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.1)
Requirement already satisfied: numpy=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: tokenizers<0.22.0,>=0.21.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: pyarrow=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
 Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Collecting xxhash (from datasets)
 Downloading xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocessing<0.70.17 (from datasets)
 Downloading multiprocessing-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2024.12.0,>=2023.1.0 (from fsspec[http]<=2024.12.0,>=2023.1.0->datasets)
 Downloading fsspec-2024.12.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from datasets) (3.11.15)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.13.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch) (3.1.6)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch)
 Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch)
 Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch)
 Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch)
 Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch)
 Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch)
 Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch)
 Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch)
 Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)

- ✓ 抓取訓練資料，裡面包含五萬筆白話文->文言文資料

```
[ ] !wget https://github.com/shhuangmust/AI/raw/refs/heads/master/train.json
```

```

--2025-04-10 11:21:10--  https://github.com/shuangmust/Al/raw/refs/heads/master/train.json
Resolving github.com (github.com)... 140.82.112.3
Connecting to github.com (github.com)|140.82.112.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/shuangmust/Al/refs/heads/master/train.json [following]
--2025-04-10 11:21:10--  https://raw.githubusercontent.com/shuangmust/Al/refs/heads/master/train.json
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2940614 (2.8M) [application/octet-stream]
Saving to: 'train.json'

train.json          100%[=====] 2.80M  --.-KB/s  in 0.05s

2025-04-10 11:21:11 (59.6 MB/s) - 'train.json' saved [2940614/2940614]

```

- 採用gemma-2-2b-it-chinese-kyara-dpo基礎模型
- 4bit的量化模型來節省訓練記憶體
- 要產生文言文，因此採用AutoModelForCausalLM

```
[ ] import torch
from transformers import AutoTokenizer, AutoModelForCausalLM, BitsAndBytesConfig

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)

model_id = "zake7749/gemma-2-2b-it-chinese-kyara-dpo"

model = AutoModelForCausalLM.from_pretrained(model_id, quantization_config=bnb_config,
                                              attn_implementation='eager',
                                              cache_implementation=None,
                                              use_cache=False,)

tokenizer = AutoTokenizer.from_pretrained(model_id, add_eos_token=True)
model.to('cuda')
```

```
config.json: 100% ██████████ 818/818 [00:00<00:00, 39.4kB/s]
`low_cpu_mem_usage` was None, now default to True since model is quantized.
model.safetensors.index.json: 100% ██████████ 24.2k/24.2k [00:00<00:00, 2.18MB/s]
Fetching 3 files: 100% ██████████ 3/3 [02:38<00:00, 158.86s/it]
model-00002-of-00003.safetensors: 100% ██████████ 4.98G/4.98G [02:37<00:00, 130MB/s]
model-00001-of-00003.safetensors: 100% ██████████ 4.99G/4.99G [02:38<00:00, 196MB/s]
model-00003-of-00003.safetensors: 100% ██████████ 481M/481M [00:41<00:00, 13.2MB/s]
Loading checkpoint shards: 100% ██████████ 3/3 [00:49<00:00, 13.77s/it]
generation_config.json: 100% ██████████ 208/208 [00:00<00:00, 10.7kB/s]
tokenizer_config.json: 100% ██████████ 47.1k/47.1k [00:00<00:00, 2.59MB/s]
tokenizer.json: 100% ██████████ 34.4M/34.4M [00:00<00:00, 93.3MB/s]
special_tokens_map.json: 100% ██████████ 636/636 [00:00<00:00, 54.6kB/s]
Gemma2ForCausalLM(
  (model): Gemma2Model(
    (embed_tokens): Embedding(256000, 2304, padding_idx=0)
    (layers): ModuleList(
      (0-25): 26 x Gemma2DecoderLayer(
        (self_attn): Gemma2Attention(
          (q_proj): Linear4bit(in_features=2304, out_features=2048, bias=False)
          (k_proj): Linear4bit(in_features=2304, out_features=1024, bias=False)
          (v_proj): Linear4bit(in_features=2304, out_features=1024, bias=False)
          (o_proj): Linear4bit(in_features=2048, out_features=2304, bias=False)
        )
        (mlp): Gemma2MLP(
          (gate_proj): Linear4bit(in_features=2304, out_features=9216, bias=False)
          (up_proj): Linear4bit(in_features=2304, out_features=9216, bias=False)
          (down_proj): Linear4bit(in_features=9216, out_features=2304, bias=False)
          (act_fn): PytorchGELUTanh()
        )
        (input_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
        (post_attention_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
        (pre_feedforward_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
        (post_feedforward_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
      )
    )
    (norm): Gemma2RMSNorm((2304,), eps=1e-06)
    (rotary_emb): Gemma2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=2304, out_features=256000, bias=False)
```



```
[ ] print(model)
```

```
Gemma2ForCausalLM(
  (model): Gemma2Model(
    (embed_tokens): Embedding(256000, 2304, padding_idx=0)
    (layers): ModuleList(
      (0-25): 26 x Gemma2DecoderLayer(
        (self_attn): Gemma2Attention(
          (q_proj): Linear4bit(in_features=2304, out_features=2048, bias=False)
          (k_proj): Linear4bit(in_features=2304, out_features=1024, bias=False)
          (v_proj): Linear4bit(in_features=2304, out_features=1024, bias=False)
          (o_proj): Linear4bit(in_features=2048, out_features=2304, bias=False)
        )
        (mlp): Gemma2MLP(
          (gate_proj): Linear4bit(in_features=2304, out_features=9216, bias=False)
          (up_proj): Linear4bit(in_features=2304, out_features=9216, bias=False)
          (down_proj): Linear4bit(in_features=9216, out_features=2304, bias=False)
          (act_fn): PytorchGELUTanh()
        )
        (input_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
        (post_attention_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
        (pre_feedforward_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
        (post_feedforward_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
      )
    )
    (norm): Gemma2RMSNorm((2304,), eps=1e-06)
  )
  (lm_head): Linear(in_features=2304, out_features=256000, bias=False)
)
```

讀取資料

- 將資料轉換成gemma2的讀取格式

```
[ ] from datasets import load_dataset
dataset = load_dataset('json', data_files="train.json", split="train").shuffle(seed=42)

def generate_prompt(data_point):
    prefix_text = '你是一個使用繁體中文的人工智慧助理，下面是問題的描述，以及對應的答案，請照著問題並且回答答案。'
    text = f'<start_of_turn>user {prefix_text} {data_point["instruction"]} <end_of_turn>\n<start_of_turn>model {data_point["output"]} <end_of_turn>'
    return text
# Add the 'prompt' column to the dataset
text_column = [generate_prompt(data_point) for data_point in dataset]
dataset = dataset.add_column("prompt", text_column)
# Tokenize the dataset
dataset = dataset.shuffle(seed=1234)
dataset = dataset.map(lambda samples: tokenizer(samples["prompt"]), batched=True)
# Split the dataset into training and testing
dataset = dataset.train_test_split(test_size=0.2)
train_data = dataset["train"]
test_data = dataset["test"]
```

```
Generating train split: 10000/0 [00:00<00:00, 71395.69 examples/s]
Flattening the indices: 100% 10000/10000 [00:00<00:00, 41595.10 examples/s]
Map: 100% 10000/10000 [00:02<00:00, 5324.75 examples/s]
```

```
[ ] print(train_data[200])
```

```
{'id': '80cb6256-c2a3-4dc9-b381-89e78e067014', 'instruction': '萬方來賀，華夷充庭。翻譯成現代文：', 'output': '萬方諸侯前來慶賀，華夏蠻夷擠滿庭院。', 'prompt': '<sta
```

找出所有可以進行QLora訓練的層

```
[ ] def find_all_linear_names(peft_model, int4=False, int8=False):
    """Find all linear layer names in the model. reference from qlora paper."""
    cls = torch.nn.Linear
    if int4 or int8:
        import bitsandbytes as bnb
        if int4:
            cls = bnb.nn.Linear4bit
        elif int8:
            cls = bnb.nn.Linear8bitLt
    lora_module_names = set()
    for name, module in peft_model.named_modules():
        if isinstance(module, cls):
            # last layer is not add to lora_module_names
            if 'lm_head' in name:
                continue
            if 'output_layer' in name:
                continue
            names = name.split('.')
            lora_module_names.add(names[0] if len(names) == 1 else names[-1])
    return sorted(lora_module_names)
```

```
[ ] from peft import LoraConfig, PeftModel, prepare_model_for_kbit_training, get_peft_model

model.enable_input_require_grads()
model.gradient_checkpointing_enable()

model = prepare_model_for_kbit_training(model)
modules = find_all_linear_names(model) # Get modules to apply LoRA to
lora_config = LoraConfig(
    r=64,
    lora_alpha=32,
    target_modules=modules,
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM"
)
peft_model = get_peft_model(model, lora_config)
```

```
[ ] print(peft_model)

(lora_dropout): ModuleDict(
  (default): Dropout(p=0.05, inplace=False)
)
(lora_A): ModuleDict(
  (default): Linear(in_features=2304, out_features=64, bias=False)
)
(lora_B): ModuleDict(
  (default): Linear(in_features=64, out_features=9216, bias=False)
)
(lora_embedding_A): ParameterDict()
(lora_embedding_B): ParameterDict()
(lora_magnitude_vector): ModuleDict()
)
(up_proj): lora.Linear4bit(
  (base_layer): Linear4bit(in_features=2304, out_features=9216, bias=False)
  (lora_dropout): ModuleDict(
    (default): Dropout(p=0.05, inplace=False)
  )
  (lora_A): ModuleDict(
    (default): Linear(in_features=2304, out_features=64, bias=False)
  )
  (lora_B): ModuleDict(
    (default): Linear(in_features=64, out_features=9216, bias=False)
  )
  (lora_embedding_A): ParameterDict()
  (lora_embedding_B): ParameterDict()
  (lora_magnitude_vector): ModuleDict()
)
(down_proj): lora.Linear4bit(
  (base_layer): Linear4bit(in_features=9216, out_features=2304, bias=False)
  (lora_dropout): ModuleDict(
    (default): Dropout(p=0.05, inplace=False)
  )
  (lora_A): ModuleDict(
    (default): Linear(in_features=9216, out_features=64, bias=False)
  )
  (lora_B): ModuleDict(
    (default): Linear(in_features=64, out_features=2304, bias=False)
  )
  (lora_embedding_A): ParameterDict()
  (lora_embedding_B): ParameterDict()
  (lora_magnitude_vector): ModuleDict()
)
(act_fn): PytorchGELUTanh()
)
(input_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
(post_attention_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
(pre_feedforward_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
(post_feedforward_layernorm): Gemma2RMSNorm((2304,), eps=1e-06)
)
(norm): Gemma2RMSNorm((2304,), eps=1e-06)
(rotary_emb): Gemma2RotaryEmbedding()
)
(lm_head): Linear(in_features=2304, out_features=256000, bias=False)
)
)
```

▼ 列出可訓練的參數數目跟比例

```
[10] peft_model.print_trainable_parameters()
trainable params: 83,066,880 || all params: 2,697,408,768 || trainable%: 3.0795
```

```
[11] print(modules)
['down_proj', 'gate_proj', 'k_proj', 'o_proj', 'q_proj', 'up_proj', 'v_proj']
```

▼ 進行PEFT訓練

- 採用trl套件
- 可修改max_steps以調整訓練次數

```
[12] from trl import SFTConfig
import transformers
from trl import SFTTrainer

trainer = SFTTrainer(
    model=peft_model,
    train_dataset=train_data,
    eval_dataset=test_data,
    peft_config=lora_config,
    args=SFTConfig(
        per_device_train_batch_size=2,
        gradient_accumulation_steps=4,
        max_steps=100,
        learning_rate=2e-4,
        output_dir="output1",
        # dataset_text_field="prompt",
        optim="paged_adamw_32bit",
        save_strategy="steps",
        report_to=None,
        #report_to="wandb",
        logging_steps=1,
        packing=False,
        gradient_checkpointing=True,
    ),
    data_collator=transformers.DataCollatorForLanguageModeling(tokenizer, mlm=False),
)
trainer.train()
```

```
72 1.705100
73 1.766100
74 1.991800
75 1.586400
76 1.912700
77 1.867400
78 1.633500
79 1.675500
80 1.814900
81 2.127300
82 2.034600
83 1.552700
84 1.736600
85 1.919000
86 1.887400
87 1.982000
88 2.509200
89 1.755600
90 1.832700
91 2.085100
92 1.350800
93 1.860800
94 1.395900
95 1.889100
96 1.842000
97 1.758000
98 1.887300
99 1.859900
100 1.532100
```

```
TrainOutput(global_step=100, training_loss=2.077938541173935, metrics={'train_runtime': 749.8471, 'train_samples_per_second': 1.067, 'train_steps_per_second': 0.133, 'total_flos': 1192471461734400.0, 'train_loss': 2.077938541173935})
```

儲存Adapter

- Huggingface的Token, 一定要包含【write】權限
- ????????/peft-model-repo, ??????請輸入自己在huggingface的帳號

```
[13] peft_model.save_pretrained("peft_model")
repo_name = "Candy04/peft-model-repo" # 替換為你的 repository 名稱
save_directory = "./peft_model" # 模型儲存的本地路徑

# 上傳模型到 Hugging Face
peft_model.push_to_hub(repo_name)
```

README.md: 100% 5.17k/5.17k [00:00<00:00, 551kB/s]

adapter_model.safetensors: 100% 332M/332M [00:21<00:00, 27.8MB/s]

CommitInfo(commit_url='https://huggingface.co/Candy04/peft-model-repo/commit/f1d9c9c29947fab741dd5c88015cc269b1627962', commit_message='Upload model', commit_description='', oid='f1d9c9c29947fab741dd5c88015cc269b1627962', pr_url=None, repo_url=RepoUrl('https://huggingface.co/Candy04/peft-model-repo', endpoint='https://huggingface.co', repo_type='model', repo_id='Candy04/peft-model-repo'), pr_revision=None, pr_num=None)

```
[14] base_model = AutoModelForCausalLM.from_pretrained(
    model_id,
    low_cpu_mem_usage=True,
    return_dict=True,
    torch_dtype=torch.float16,
    device_map="auto",
)

merged_model = PeftModel.from_pretrained(base_model, "output1/checkpoint-100")
merged_model = merged_model.merge_and_unload()
```

Loading checkpoint shards: 100% 3/3 [00:45<00:00, 12.66s/it]

一般來說只需要上傳Adapter即可, 要使用LLM時, Adapter+基礎模型即可。這邊直接把兩者合併為(merged_model)

```
[15] merged_model.save_pretrained("merged_model", safe_serialization=True, push_to_hub=True)
tokenizer.save_pretrained("merged_model", push_to_hub=True)
```

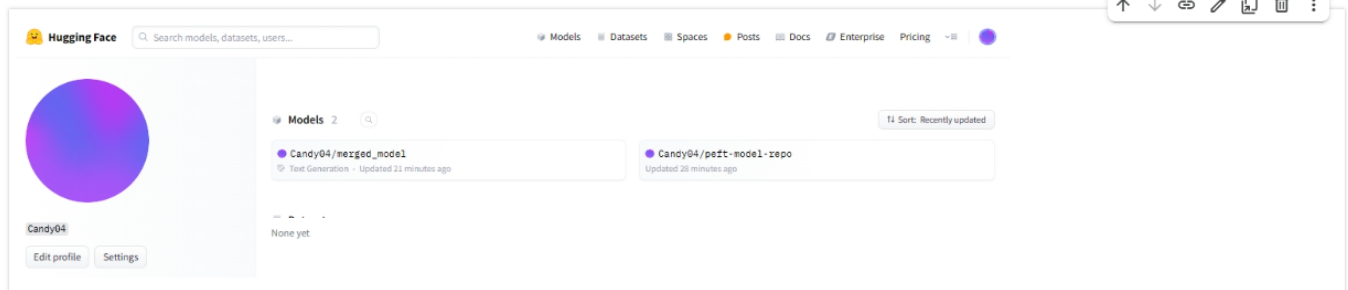
Upload 2 LFS files: 100% 2/2 [03:32<00:00, 121.38s/it]

model-00002-of-00002.safetensors: 100% 241M/241M [00:19<00:00, 12.8MB/s]

model-00001-of-00002.safetensors: 100% 4.99G/4.99G [03:31<00:00, 20.0MB/s]

tokenizer.json: 100% 34.4M/34.4M [00:03<00:00, 18.2MB/s]

('merged_model/tokenizer_config.json',
'merged_model/special_tokens_map.json',
'merged_model/tokenizer.json')



Pipeline_with_Deepseek_R1.ipynb

```

!pip install transformers torch

Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.50.3)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (2.6.0+cu124)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex<=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.13.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch) (3.4.2)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch) (2025.3.62)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cuspars-cu12==12.3.1.170 (from torch)
  Downloading nvidia_cuspars-cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch) (3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.31)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl (363.4 MB)
-----
363.4/363.4 MB 2.8 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (13.8 MB)
-----
13.8/13.8 MB 24.8 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (24.6 MB)
-----
24.6/24.6 MB 17.7 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
-----
883.7/883.7 kB 19.0 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl (664.8 MB)
-----
82.7/664.8 MB 62.7 MB/s eta 0:00:10

```

```
from transformers import pipeline

generator = pipeline("text-generation", model="deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B", device_map="auto")

result=generator(
    "一個農場有雞跟兔，共有5個頭16隻腳，請問雞跟兔各有幾隻?請用繁體中文回答",
    truncation=True,
    max_length=30000,
    num_return_sequences=1,
)

print(result)
```

```
config.json: 100% ██████████ 679/679 [00:00<00:00, 55.4kB/s]
model.safetensors: 100% ██████████ 3.55G/3.55G [00:30<00:00, 184MB/s]
Sliding Window Attention is enabled but not implemented for `sdpa`; unexpected results may be encountered.
generation_config.json: 100% ██████████ 181/181 [00:00<00:00, 9.98kB/s]
tokenizer_config.json: 100% ██████████ 3.07k/3.07k [00:00<00:00, 224kB/s]
tokenizer.json: 100% ██████████ 7.03M/7.03M [00:00<00:00, 34.9MB/s]
Device set to use cpu
[{'generated_text': '一個農場有雞跟兔，共有5個頭16隻腳，請問雞跟兔各有幾隻?請用繁體中文回答。\\n\\n要解這道題，首先我需要了解雞和兔的頭和腳的數量 relation。已知每個雞有1個頭，2只腳；每
```