後端模型和前端網頁串聯測試

Jim

6 min read · 4 days ago

Share ••• More

使用fastapi + unvicorn啟動。從前端傳入圖片，先由florence給出圖片內容，phi(LLM)轉化為中文，再以josn回傳至前端。

```
from fastapi import FastAPI, File, UploadFile
from im2text import getimg2text
from PIL import Image
import torch
from fastapi.middleware.cors import CORSMiddleware  # 導入 CORS 中間件
from ollamaConect import getResponse
import io
from fastapi.responses import JSONResponse
app = FastAPI()# FastAPI 物件
# 添加 CORS 中間件
app.add_middleware(
    CORSMiddleware,
    allow_origins=["*"],  # 允許所有來源，或者指定具體的來源（例如 ["http://localhost
    allow_credentials=True,
    allow_methods=["*"],  # 允許所有 HTTP 方法
    allow_headers=["*"],  # 允許所有 HTTP 頭
)
@app.post("/upload/")# 裝飾器
async def upload_image(file: UploadFile = File(...)):  # 接收前端上傳的圖片
    try:
        image_data = await file.read()
        image = Image.open(io.BytesIO(image_data)).convert("RGB")
        text = getimg2text(image)
        response = getResponse(text)
        return JSONResponse(content=response)

    except Exception as e:
        return JSONResponse(content={"error": str(e)}, status_code=400)
```

```
from ollama import chat

# 指定模型名稱
model = "phi4"

# 生成回應
def getResponse(input):
    response = chat(model=model, messages=[{"role": "user", "content": "請將下文
    return response["message"]["content"]


# 輸出結果
if __name__ == "__main__":
    print(getResponse("hello how are you"))
```

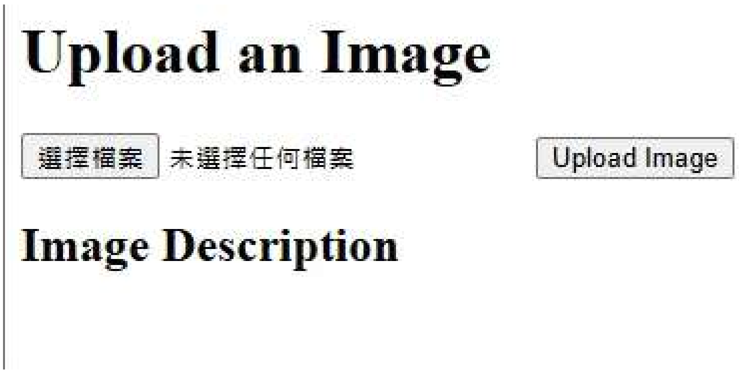
```
import requests

import torch
```

Open in app ↗

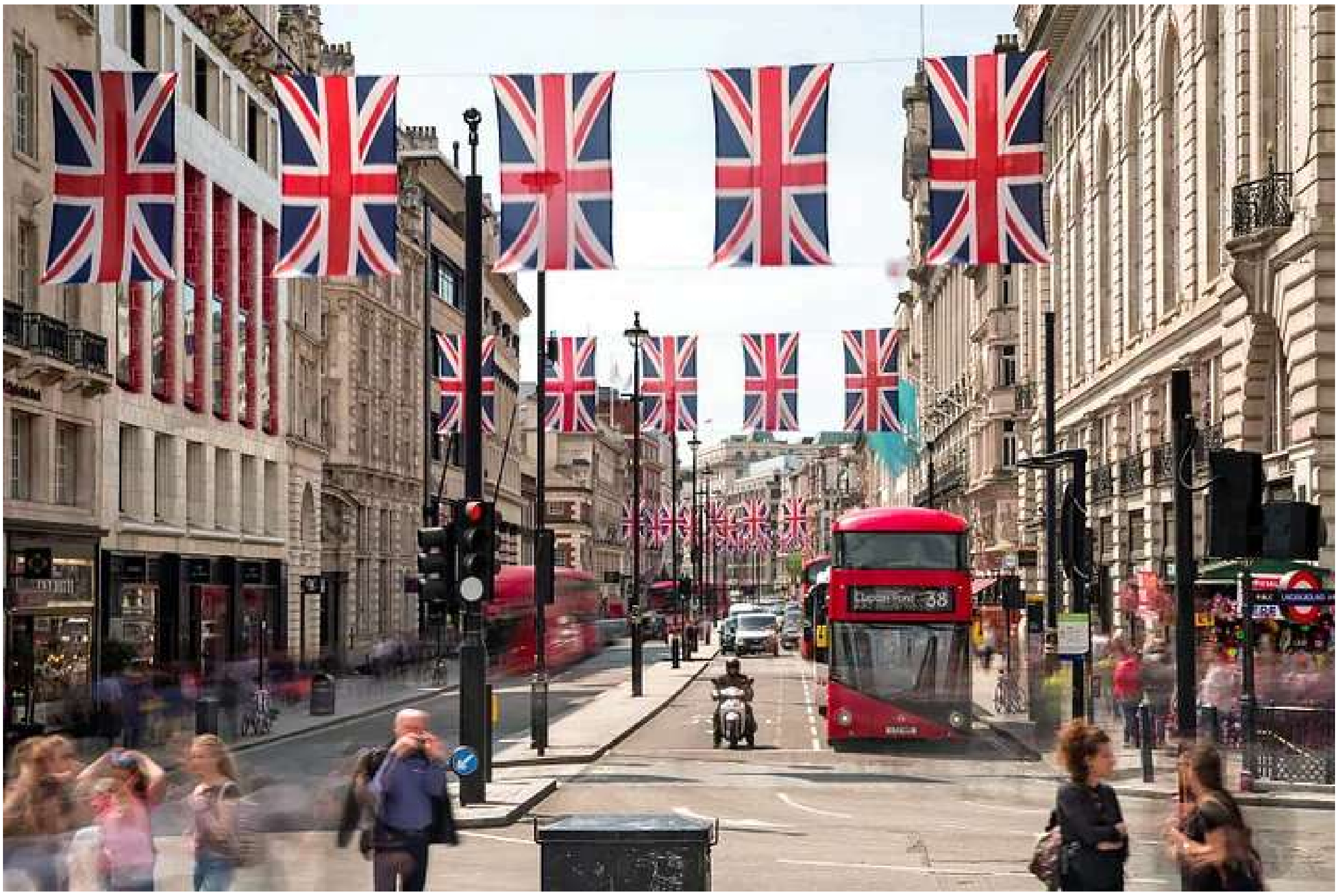
Medium Search

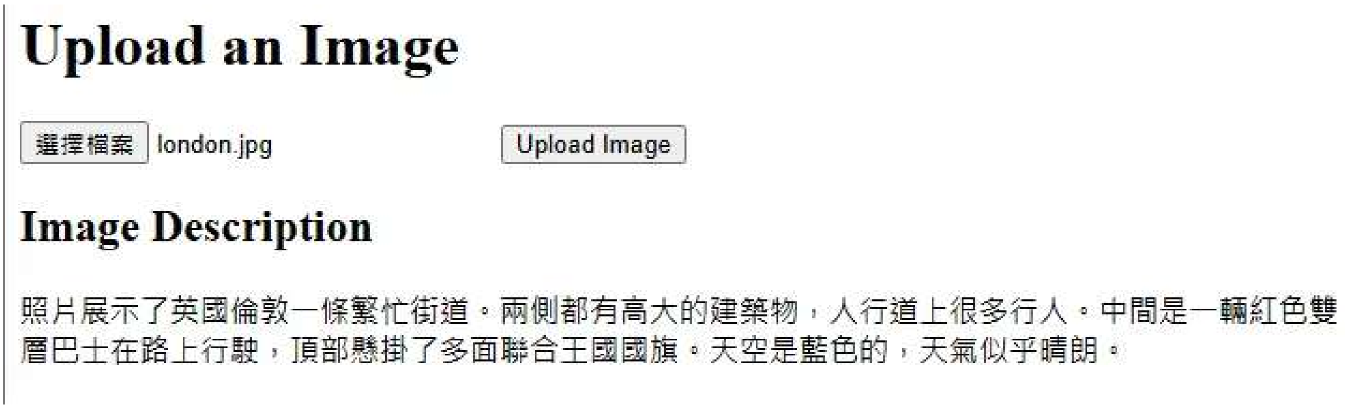
```
    device = "cuda:0" if torch.cuda.is_available() else "cpu"
    torch_dtype = torch.float16 if torch.cuda.is_available() else torch.float32
    model = AutoModelForCausalLM.from_pretrained("microsoft/Florence-2-large",
    processor = AutoProcessor.from_pretrained("microsoft/Florence-2-large", tru
    prompt = "<MORE_DETAILED_CAPTION>"
    inputs = processor(text=prompt, images=image, return_tensors="pt").to(devic
    generated_ids = model.generate(
    input_ids=inputs["input_ids"],
    pixel_values=inputs["pixel_values"],
    max_new_tokens=4096,
    num_beams=3,
    do_sample=False
)
    generated_text = processor.batch_decode(generated_ids, skip_special_tokens=
    output = processor.post_process_generation(generated_text, task="<MORE_DETA
    return output['<MORE_DETAILED_CAPTION>']
if __name__ == "__main__":
    print(getimg2text(image))
```



Upload an Image

Image Description



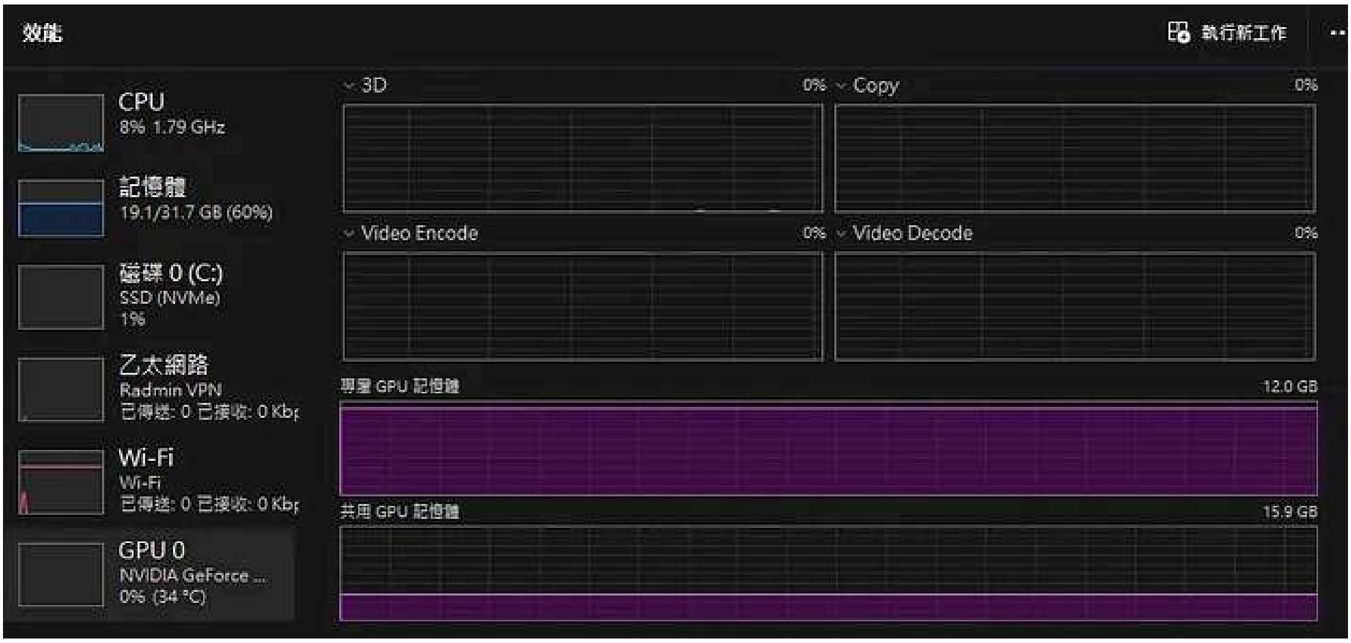


Upload an Image

Image Description

照片展示了英國倫敦一條繁忙街道。兩側都有高大的建築物，人行道上很多行人。中間是一輛紅色雙層巴士在路上行駛，頂部懸掛了多面聯合王國國旗。天空是藍色的，天氣似乎晴朗。

未來改進方向



模型結束後記憶體釋放問題。 掛上gork連接線上

Fastapi

Backend

Python

Edit profile

Written by Jim

1 Follower · 1 Following

No responses yet

Jim

What are your thoughts?