Deep Learning Implementation - Lab 8 : Running LLMs on Raspberry PI

20230683 박한비, 20230308 최승희

본 과제에서는, Raspberry Pi 5에서 LLM 모델을 구동하여, 클라우드 지원이 없는 모바일 프레임워크 상 엣지 디바이스에서의 LLM을 구현한다. Raspberry Pi 5 환경에 Ollama를 설치하고 GGUF 포맷의 모델을 설치 및 실행한다. 그 다음으로, Project 1에서 Raspberry Pi 5에 구현된 LLM을 다른 환경에서도 이용할 수 있도록 REST API call을 사용해 Raspberry Pi 5 로컬 LLM 챗봇과 리모트 환경의 클라이언트 간 양방향 인터랙션이 가능한 가벼운 웹 서버를 구현하는 것을 목적으로 한다.

1. Run your custom model on Ollama

huggingface에서 google의 gemma-2-2b 모델을 내 repo로 가져온 후 다운로드하여 Raspberry Pi 5에 보냈다. 지침에 따라 명령어를 입력하여 gemma2_2b Raspberry Pi 5 로컬 모델을 생성하였다.

2. Project 1 : lightweight web server

1) Implementations

`server.py` 파일로 웹서버를 구현하였다. `server.py` 내부 코드 설명은 아래와 같다. 먼저 알아둬야 할 것은 `server.py` 에서는 각각 비동기 HTTP 클라이언트 / 웹프레임워크 / 입력 검증을 위해 외부 라이브러리 httpx / FastAPI / pydantic을 import하여 사용한다.

아래는 server.py에서 정의한 구조체 및 변수들이다.

- OLLAMA_HOST : Ollama의 로컬 주소를 저장한다. 환경변수 OLLAMA_HOST 값을 읽어오거나 환경변수가 없다면 "http://localhost:11434" 값으로 한다.

- class ChatMessage : /chat으로 받는 메세지를 저장 및 관리하는 pydantic 구조체이다.

- class ChatRequest : /chat으로 받는 전체 요청을 저장 및 관리하는 pydantic 구조체이다.

- class ChatResponse : /chat으로 받은 요청에 대해 보낼 답변을 저장 및 관리하는 pydantic 구조체이다.

- app : FastAPI 객체로, 웹 어플리케이션의 본체를 담당한다.

아래는 server.py에서 정의한 함수들이다.

- health
: `/health`, 헬스 체크에 대해 게이트웨이 서버가 살아있고 **Ollama(11434)** 에도 연결되는지 **상태를 체크해 반환한다**. 내부에서 `GET http://OLLAMA_HOST/api/tags`를 찍어보고 성공 시 `{"status":"ok","ollama":"connected"}` 를 리턴하고 unreachable 이유, 혹은 error 메세지를 넣어 리턴한다.

- proxy_all
: 랩 지침 파일의 API call과 같은 형태로 명령어를 입력해도 되도록 처리해주는 함수이다. `/api/{path:path}` 형태로 들어온 모든 메세지에 대해 들어온 요청을 그대로 Ollama의 `http://OLLAMA_HOST/api/{path}`로 프록시 및 전달하고 전달에 대한 Ollama의 반환값을 그대로 반환한다.

- chat
: Ollama의 문답에서 OIllama가 보낸 메세지 중 답변에 해당하는 부분만 추출해주는 함수이다. 클라이언트에서 /chat request를 보내면 내부에서 `POST {OLLAMA_HOST}/api/chat` 을 `stream:false`로 호출해 JSON 응답에서 `message.content` 만을 빼내어 response로 보내준다.

- load_model
: 해당 요청 시 `{OLLAMA_HOST}/api/generate` 를 빈 메세지와 함께 Ollama에 전달해 Ollama 모델을 미리 로컬 메모리에 올려둔다. 이를 통해 다음에 올 첫 response의 응답 시간을 줄일 수 있다. 성공 및 실패 결과는 Ollama의 답변을 그대로 반환하여 반환된 메세지에서 확인 가능하도록 하였다.

- unload_model
: 해당 요청 시 `{OLLAMA_HOST}/api/generate` 를 `"keep_alive": 0`과 함께 Ollama에 전달해 Ollama 모델을 로컬 메모리에서 해제한다. 결과는 Ollama의 답변을 그대로 반환하여 반환된 메세지에서 확인 가능하도록 하였다.

2) Result

pip install로 uvicorn, httpx, FastAPI, pydantic를 설치한 후 `server.py` 를 실행하였다. 이후 uvicorn 명령로 FastAPI 어플리케이션 객체 app을 실행하며 포트 8080으로 외부에서 접속 가능하도록 하였다. 그 결과는 아래와 같고 메세지 출력으로 보아 성공적으로 기동되었다는 것을 알 수 있다.

```
(exp) dl08@raspberrypi:~ $ uvicorn server:app --host 0.0.0.0 --port 8080
INFO:     Started server process [6061]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:8080 (Press CTRL+C to quit)
INFO:     172.30.1.86:52976 - "GET /health HTTP/1.1" 200 OK
INFO:     172.30.1.86:52980 - "POST /load HTTP/1.1" 200 OK
INFO:     172.30.1.86:52981 - "POST /api/generate HTTP/1.1" 200 OK
INFO:     172.30.1.86:52982 - "POST /api/chat HTTP/1.1" 200 OK
INFO:     172.30.1.86:53008 - "POST /api/chat HTTP/1.1" 200 OK
INFO:     172.30.1.86:53028 - "POST /unload HTTP/1.1" 200 OK
```

Raspberry Pi 5와 연결된 ssh 서버가 아닌 노트북 로컬 환경에서 REST API call로 Raspberry Pi 5에 메세지를 보내보았다. 그 결과는 아래와 같다.

```
(base) ✘ 🐍 base xxhee 🐶 ~
❯ curl http://172.30.1.66:8080/health
{"status":"ok","ollama":"connected"}%
(base) 🐍 base xxhee 🐶 ~
❯ curl -X POST 'http://172.30.1.66:8080/load' \
  -H 'Content-Type: application/json' \
  -d '{"model":"llama3.2","messages":[{"role":"user","content":""}]}'

{"model":"llama3.2","created_at":"2025-11-10T14:18:39.551768517Z","response":"","done":true,"done_reason":"load"}%
(base) 🐍 base xxhee 🐶 ~
❯ curl -X POST 'http://172.30.1.66:8080/api/generate' \
  -H 'Content-Type: application/json' \
  -d '{"model":"llama3.2","prompt":""}'
{"model":"llama3.2","created_at":"2025-11-10T14:18:46.40117184Z","response":"","done":true,"done_reason":"load"}%
```

챗봇 모델을 로드 요청하는 메세지를 보냈고 그에 답하는 메세지가 도착해 출력된 것을 보아 성공적으로 로드가 이루어진 것을 알 수 있다.

```
(base) 🐍 base xxhee 🐶 ~
❯ curl -X POST 'http://172.30.1.66:8080/api/chat' \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "llama3.2",
    "stream": false,
    "messages": [
      {"role": "user", "content": "Why is the sky blue?"}
    ]
  }'
{"model":"llama3.2","created_at":"2025-11-10T14:22:51.079188407Z","message":{"role":"assistant","content":"The sky appears blue because of a phenomeno
n called Rayleigh scattering, named after the British physicist Lord Rayleigh. He discovered that shorter (blue) wavelengths of light are scattered mo
re than longer (red) wavelengths when they pass through tiny molecules in the atmosphere.\n\nHere's what happens:\n\n1. Sunlight enters Earth's atmosp
here and encounters tiny molecules like nitrogen (N2) and oxygen (O2).\n2. The sunlight is made up of a spectrum of colors, with shorter wavelengths (
like blue and violet) being more energetic than longer wavelengths (like red and orange).\n3. When the sunlight hits these tiny molecules, it scatters
 in all directions.\n4. Since blue light has a shorter wavelength, it is scattered more than other colors by the smaller molecules in the atmosphere.\
n5. As a result, the blue light is dispersed throughout the sky, giving it that characteristic blue hue.\n\nOther colors, like red and orange, have lo
nger wavelengths and are not scattered as much by the tiny molecules. This is why we don't see these colors dominating the sky.\n\nIt's also worth not
ing that the sky can appear different shades of blue at various times of day and in different atmospheric conditions (like pollution or dust particles
). However, Rayleigh scattering remains the primary reason for the sky's blue coloration."},"done":true,"done_reason":"stop","total_duration":51730903
965,"load_duration":240714927,"prompt_eval_count":31,"prompt_eval_duration":175552487,"eval_count":262,"eval_duration":51013789487}%
```

챗봇 모델과 채팅을 요청하는 메세지를 보냈고 그에 답하는 메세지가 도착해 출력된 것을 보아 챗봇이 성공적으로 답변을 도출한 것을 알 수 있다.

```
(base) ✘ 🐍 base xxhee 🐶 ~
❯ curl -X POST 'http://172.30.1.66:8080/unload' \
  -H 'Content-Type: application/json' \
  -d '{"model":"llama3.2","messages":[{"role":"user","content":""}]}'
{"model":"llama3.2","created_at":"2025-11-10T14:25:08.360416617Z","response":"","done":true,"done_reason":"unload"}%
(base) 🐍 base xxhee 🐶 ~
❯
```

챗봇 모델을 언로드 요청하는 메세지를 보냈고 그에 답하는 메세지가 도착해 출력된 것

을 보아 성공적으로 로드가 끝난 것을 알 수 있다.