Suggestions for vLLM on EC2

DA 510

Richard Kelley

Things to try

- Remember that this is a CPU-only demonstration
  - You'd do things differently if you had GPUs, but this assignment will get you surprisingly far.
- Use an m7i-flex.large instance
- Use tmux
- Use docker – don't try to build vLLM yourself for the CPU-only case.
- Use the recommended model: `HuggingFaceTB/SmolLM2-135M-Instruct`

Docker install reminder

```
$ sudo apt update
$ sudo apt install -y docker.io
$ sudo systemctl enable --now docker
$ sudo usermod -aG docker $USER
$ newgrp docker
```

vLLM

```
docker run \
  -v ~/.cache/huggingface:/root/.cache/huggingface \
  -p 8000:8000 \
  vllm/vllm-openai-cpu:latest-x86_64 \
  --model HuggingFaceTB/SmolLM2-135M-Instruct \
  --dtype float32
```

Testing vLLM locally on your instance

```
$ curl http://localhost:8000/v1/models
$ curl http://localhost:8000/v1/chat/completions \
  -H "Content-Type: application/json" \
  -d '{
    "model": "HuggingFaceTB/SmolLM2-135M-Instruct",
    "messages": [
      {"role": "user", "content": "Explain cloud computing in
one paragraph."}
    ],
    "max_tokens": 100,
    "temperature": 0.7
  }'
```

Requiring an API key

```
$ docker run \
  -v ~/.cache/huggingface:/root/.cache/huggingface \
  -p 8000:8000 \
  -e VLLM_API_KEY=secret123 \
  vllm/vllm-openai-cpu:latest-x86_64 \
  --model HuggingFaceTB/SmolLM2-135M-Instruct \
  --dtype float32 \
  --api-key $VLLM_API_KEY
```

Requiring an API Key

```
curl http://localhost:8000/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer secret123" \
  -d '{
    "model": "HuggingFaceTB/SmolLM2-135M-Instruct",
    "messages": [
      {"role": "user", "content": "Hello"}
    ]
  }'
```