Intel 18T I5 BMv1 인공지능 테스트 및 활용 정보

NPU(Neural Processing Unit)는 인공지능(AI) 및 머신러닝(ML) 작업에 특화된 전용 프로세서입니다. 주로 딥러닝 모델의 추론(inference) 작업을 가속화하기 위해 설계되었으며, CPU나 GPU보다 효율적으로 계산을 수행할 수 있습니다.

Green 18T I5 BMv1 상품에는 NPU 가 탑재되어 있습니다. 일반적으로 AI 에 사용되는 고성능 장비의 GPU 서버가 아니기 때문에 서비스를 위한 개발 용도로 사용은 어렵지만 몇가지 관련 정보를 제공해 드리오니 관심 있으신 분은 내용을 확인해 보시기 바랍니다. 추가적인 모듈 설치 및 테스트 요청, 컨설팅 등은 어려우므로 참고용으로만 봐주시면 감사하겠습니다.

1. 드라이버 및 라이브러리 설치

1) Intel NPU 드라이버 및 라이브러리 설치

- 시스템 버전과 커널은 공식 요구 사항에 따라 설치해야 하며, 해당 버전은 아래와 같습니다.
- 운영 체제 및 커널 요구사항: Ubuntu 22.04, 6.8.0-47-generic
- 드라이버 버전: 1.10.0설치 가이드: Intel NPU 드라이버 공식 페이지

2) Intel GPU 드라이버 설치

- 설치 가이드: Intel GPU 드라이버 공식 문서

3) Intel NPU 가속 라이브러리

- 설명: Intel NPU를 사용하여 고속 계산을 수행할 수 있는 Python 라이브러리
- 설치: pip install intel-npu-acceleration-library
- GitHub: NPU 가속 라이브러리

4) OpenVINO GenAI 가속 라이브러리

- 설명: 생성형 AI 모델 실행을 위한 OpenVINO 기반의 C++/Python API 라이브러리
- 설치: pip install openvino-genai
- GitHub: OpenVINO GenAI

2. 성능 테스트 결과 (NPU 와 GPU 비교 테스트)

1) NPU 성능

- 하드웨어: Green 18T I5 BMv1
- 모델: Qwen2-Math-7B
- NPU 테스트: 공식 스크립트
  - ㄴ 총 실행 시간: 175.32초
  - ㄴ 초당 처리 토큰: 약 1.11 tokens/sec
- NPU와 CPU 비교 테스트: 공식 스크립트
- 모델: TinyLlama-1.1B-Chat-v1.0
  - ㄴ NPU: 초당 3 tokens
  - ㄴ CPU: 초당 14.27 tokens
- 메모리 사용량
  - ㄴ NPU: 약 7GB
  - ㄴ CPU: 약 1GB

2) GPU 성능

- 하드웨어: Nvidia T4
- 모델: Qwen2-Math-7B
  - ㄴ총 실행 시간: 15,128.13ms
  - ㄴ초당 처리 토큰: 37.02 tokens/sec

3) 결론

- NPU 가속 라이브러리를 사용할 경우, CPU 의 사용률은 80%. CPU 단독 실행 시 사용률은 30% 정도 입니다.
- Qwen 모델 테스트에서 NPU는 초당 약 1 token, TinyLlama 모델에서는 초당 3 tokens, CPU 는 TinyLlama 모델에서 초당 14.27 tokens 를 기록됩니다.
  - ㄴ TinyLlama: 메모리 사용량: NPU 로 모델 변환 후 메모리 사용률 약 7GB, CPU 로 변환 후 메모리 사용률은 약 1GB 입니다.
  - ㄴ OpenVINO GenAI 라이브러리를 사용하여 변환된 TinyLlama 모델을 처리할 경우, CPU 와 GPU 에서 매우 빠르게 실행됩니다.
- 다중 모달 테스트
  - ㄴ NPU: 초당 1.86 tokens
  - ㄴ CPU: 초당 0.85 tokens
- 동일한 모델을 NPU 용 Intel NPU Acceleration Library 와 CPU 용 OpenVINO 라이브러리에서 실행하기 위해서는 모델 변환이 필요합니다. 변환 과정에서 발생하는 차이로인해, 아래 테스트 데이터는 참고용으로만 사용해 주세요.

3. 모델 테스트

1) Intel NPU 가속 라이브러리

- 모델: Qwen2-Math-7B (스크립트)
  ㄴ 결과: NPU: 초당 1.11 tokens / CPU: 초당 14 tokens
- 모델: Intel/Llava-Gemma-2B (스크립트)
  ㄴ 테스트: 이미지 요약 작업 (이미지)
  ㄴ 결과: NPU: 초당 1.86 tokens
- 모델: TinyLlama-1.1B-Chat-v1.0 (스크립트)
  ㄴ 결과: NPU: 초당 3.74 tokens
- 모델: Microsoft/Phi-2 (스크립트)
  ㄴ 결과: NPU: 초당 0.97 tokens
- 모델: microsoft/Phi-3-mini-4k-instruct (스크립트)
  ㄴ 결과: NPU: 초당 1.35 tokens
- 모델: google/flan-t5-small (스크립트)
  ㄴ 결과: 실행 실패

2) OpenVINO GenAI 라이브러리

- 이 라이브러리는 PC 와 노트북에서 실행 가능하며, 자원 소비를 최적화했습니다.
  생성 모델을 실행하기 위해 외부 종속성이 필요하지 않으며, 핵심 기능 (예: openvino-tokenizers 를 통한 토크나이징)을 이미 포함하고 있습니다.
- 모델: NPU 에서 OpenVINO GenAI Flavor 가 적용된 LLM 실행 (스크립트)
  ㄴ 결과: 실행 실패. 관련 문제는 Github 에 보고 됨 (내용)
- 모델: TinyLlama-1.1B-Chat-v1.0 (스크립트)
  ㄴ 결과: NPU 에서 실행되지 않았으나, CPU 와 GPU 에서는 매우 빠르게 실행 되어짐
- 모델: Dreamlike-art/Dreamlike-anime-1.0 (스크립트)
  ㄴ 결과: CPU: real 6m13.358s / user 44m54.039s / sys 0m28.786s
  GPU: real 0m36.230s / user 0m14.074s / sys 0m10.571s
  NPU: 실행 실패
- 모델: OpenAI/Whisper-base (스크립트)
  ㄴ 결과: CPU: real 0m10.612s / user 0m28.134s / sys 0m3.303s
  GPU: real 0m9.046s / user 0m11.476s / sys 0m3.248s
  NPU: 실행 실패

3) OpenVINO 라이브러리

- 모델: llava-hf/Llava-1.5-7B-hf (스크립트)
  ㄴ 테스트: 이미지를 텍스트로 요약하는 작업 (이미지)
  ㄴ 결과: CPU: 초당 0.85 tokens
- 모델: TinyLlama/TinyLlama-1.1B-Chat-v1.0 (스크립트)
  ㄴ 결과: CPU: 초당 14.27 tokens

4) 참고 자료

- Intel NPU Acceleration Library GitHub
- OpenVINO GenAI GitHub
- OpenVINO Interactive Tutorials