이기종 컴퓨팅과 복수 신경망 추론 환경에서 높은 처리량을 위한 스케줄러 관한 연구

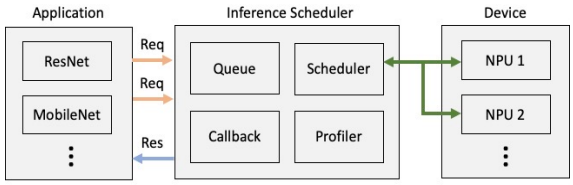
차주형, 박제만, 권용인
aoikazto@naver.com, (jeman, yongin.kwon)@etri.re.kr

연구 동기

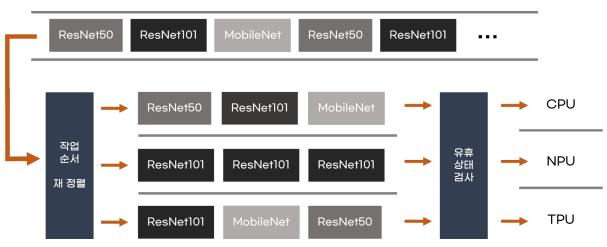
1. 높은 정확도와 다양한 정보를 추출하기 위해 레이어가 깊고 복수 개의 신경망 모델을 사용하고 있음.
2. 딥 러닝 연산을 가속하기 위해 NPU나 GPU를 사용하게 되는데 단일 연산 장치에서 모두 수행하게 된다면 높은 부하가 발생하며 처리량이 고정적.
3. 이를 위해 이종 시스템을 통한 추론 성능 향상하는 방향으로 연구가 진행.
4. 그러나 이종 시스템을 관리하고 연산 장치에 특화된 연산을 수행이 가능하도록 작업 분배하는 스케줄러를 비교 분석한 선행 연구가 부족.
5. 본 논문에서는 실시간으로 신경망 추론이 필요한 상황에서 최적화된 연산 수행하도록 작업을 재배열 했을 때 변화되는 성능 비교.

스케줄러의 정의 및 구현 방안

1. 연산 장치는 특화되어 있는 연산과 컴퓨팅 성능마다 상이하므로 신경망 모델마다 처리량이 상이함
2. 따라서 M 개의 신경망 모델과 N 번 추론을 수행해야할 때 K 개의 연산 장치가 최단 시간으로 작업 분배하는 스케줄러가 필요함.
3. 딥 러닝 연산 작업을 스케줄러를 통해 분배하기 앞서 응용 계층과 스케줄링 계층, 하드웨어 계층으로 3개의 모듈로 추상화가 가능함.
   - 응용 계층은 신경망 모델을 스케줄러에게 연산 요청하는 계층.
   - 스케줄링 계층은 아래의 4개의 요소로 구현함.
     - 응용 계층으로 부터 입력된 요청을 저장하는 대기열(Queue).
     - 딥 러닝 추론 결과를 반환하는 콜백(Callback).
     - 연산 장치의 추론 성능을 분석하는 프로파일러.(Profiler)
     - 연산 장치 마다 작업을 할당하는 스케줄러(Scheduler).
   - 하드웨어 계층은 작업 할당된 연산을 수행.



실시간으로 최적의 추론 처리량을 탐색하는 스케줄러 구조



신경망 모델 작업과 연산 장치 간 작업 분배하는 스케줄링 알고리즘

실험 환경

1. 스케줄러 구성

1. FCFS(First Come First Served)
   - 대기열에 입력되는 작업을 순차적으로 연산 장치에 할당.
2. Affinity
   - 신경망 모델과 연산 장치 1 대 1 사상
3. SJF(Short Job First)
   - 대기열에서 연산 비용이 낮은 모델 부터 순차 처리
4. Priority
   - 임의로 지정되거나 프로파일러에서 생성된 우선순위 기준으로 처리

2. 하드웨어 구성

1. 실험 보드 : Asus Tinker Edge R (SoC : RK3399Pro)
2. Intel NCS2 는 2 개 이상의 신경망 모델 적재가 불가능함.

표. 연산 장치별 지원하는 API

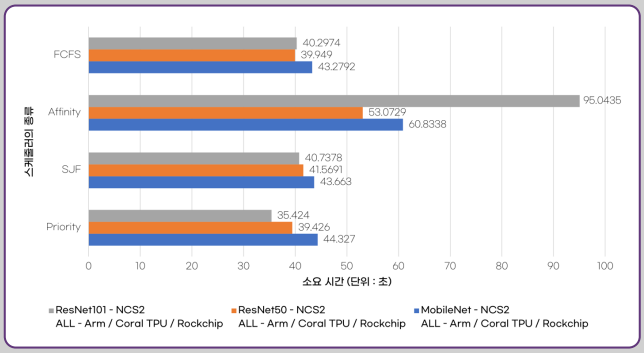
프로세싱 유닛 명	백엔드	데이터 타입	복수 추론
CPU (Cortex A72)	ArmCL	FP 32	O
Intel NCS 2	ncAPI	FP 16	X
Google Coral TPU	Edgetpu	INT 8	O
Rockchip NPU	rknn	INT 8	O

3. 신경망 모델 구성

표. 신경망 모델과 연산 장치 간 평균 추론 시간 및 우선순위 정보

	MobileNetV2	ResNet50	ResNet101
신경망 파라미터의 수	3,538,984	25,636,712	44,707,176
CPU (A72)	86.355 ms (1)	582.072 ms (3)	1201.923 ms (2)
Intel NCS 2	27.459 ms	50.743 ms	94.509 ms
Google Coral USB	3.393 ms (1)	54.086 ms (3)	108.260 ms (2)
Rockchip NPU	10.499 ms (2)	21.345 ms (1)	31.770 ms (3)

스케줄러 성능 비교



ETRI
Electronics and Telecommunications Research Institute