컴퓨터구조 과제 #3
ICT융합공학부 202204010 공성택

1. NPU의 기본 개념 및 구조

NPU란 Neural Processing Unit으로 우리의 뇌처럼 정보를 학습하고 처리한다. 즉, 인공지능과 그에 필요한 딥러닝 작업을 위한 가속기 하드웨어이다. NPU는 인공지능과 딥러닝 모델 개발에 특화된 장치로써 해당 작업을 실행할 때 뛰어난 모습을 보인다. 우선 고속 연산 처리가 가능하기 때문에 딥러닝의 학습과 추론 속도를 더 빠르게 증가시킬 수 있다. 또한 이런 작업을 할 때, CPU나 GPU에 비해 전력 소모가 더 적어 효율적이다. 이런 전력 소모는 모바일 기기에서 특히 중요하게 작용하며, 실제로 '삼성전자'는 모바일 AP '엑시노스9'에 NPU를 탑재해 모바일 기기의 인공지능 연산 능력을 향상 시켰다. 이처럼 NPU가 CPU에 비해 인공지능 연산에서 더 뛰어난 퍼포먼스를 보이는 이유는 병렬 처리를 하는 데에 더 특화되어 있기 때문이다. NPU는 수많은 연산을 병렬 처리를 통해 동시에 수행하기 때문에, 연산 속도도 더 빠르고, 효율적이다.

NPU는 기본적으로 GPU와 같이 병렬처리를 하기 때문에 많은 연산 유닛으로 이루어진 구조를 보이지만 딥러닝 계산에 최적화된 MAC 유닛 등으로 구성되어 있다. 또한 SoC에서 더 나아가 NoC 네트워크 온 칩이라는 구조를 사용하여 더 빠른 데이터 전송을 할 수 있다. 구조와 관련해서는 뒤에서 더 자세히 다룰 것이다.

2. NPU 개발 배경 및 필요성

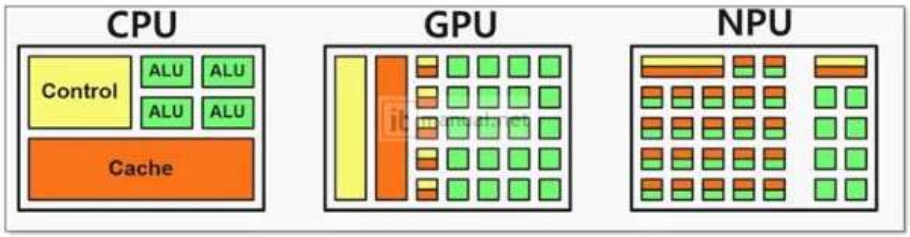
앞서 말했듯 NPU는 기존 CPU와 GPU보다 인공지능, 딥러닝 모델 개발에 특화된 장치이다. 4차 산업혁명으로 인공지능이 빠르게 발전하고 있으며, 발전하는 신경망에 따라서 구조를 구현할 수 있는 하드웨어의 발전도 필수적이다. 과거에는 인공지능이 이미지 기반 객체 검출 등 단순한 작업에 사용됐지만, 인공지능 시대로 접어들면서 카메라 화질, 음성 서비스 등 더 많은 연산을 요구하기 때문에 더 높은 성능을 요구하고 있다. 기존에 사용하던 GPU 또한 범용적으로 병렬 처리를 사용하기 때문에, CPU에 비해서는 효율적이었지만, NPU는 그보다 더 나아가 인공지능 연산이라는 목적에 최적화하여 특화된 하드웨어가 필요했다. NPU는 AI 모델의 연산 패턴을 최적화하기 때문에 GPU에 비해 훨씬 전력 소비와 처리 속도 부분에서 뛰어난 성능을 보여준다.

인공지능이 큰 컴퓨터나 대규모 장치에만 탑재되는 것이 아니라, 요즘 사회적으로 VR, AR 기기, 스마트폰 등 휴대용 기기에도 인공지능 기능을 탑재하는 경우가 많기 때문에 성능을 최적화하고, 전력 소비를 줄이기 위해서는 NPU의 추가적인 개발이 필수적이라고 할 수 있다.

3. NPU 구조 분석 – 주요 구성 요소, 작동 원리

NPU에서 가장 중요한 구조는 연산 유닛이라고 할 수 있다. NPU는 GPU와 같은 병렬 처리를 하지만, NPU는 연산 유닛을 딥러닝 연산에 최적화된 MAC 유닛과 텐서 코어를 사용한다. MAC 유닛은 곱셉과 덧셈 연산을 효과적으로 처리하기 때문에 CNN의 합성곱 연산 등의 신경망의 기본 연산을 병렬로 효율적으로 처리할 수 있다. 텐서 코어는 딥러닝에서 중요한 매트릭스 곱셈과 같은 텐서 연산을 효과적으로 처리한다. 이외에도 특화된 연산을 하는 신경망 가속기, 비트 연산 유닛 등도 존재한다. 연산 유닛 외에도 GPU와 비슷하게 캐시 메

모리, 메모리 컨트롤러도 존재한다. 연산 중간중간 계산 결과와 가중치, 입력 데이터 등을 저장하여 데이터 전송 시간을 줄이기 위해서 캐시 메모리를 사용하는 온 칩 메모리를 사용하였다. 또한 이런 온 칩 메모리와 외부 메모리 간의 데이터 전송을 관리하기 위한 메모리 컨트롤러도 존재한다. 데이터 경로를 관리하기 위해서 버퍼를 통해 입출력 데이터의 전달을 관리하고 데이터를 연산 유닛에 분배하고, 연산 유닛에서 나온 결과를 집계하는 데이터 유닛도 존재한다. 또한 디코더를 통해 명령어를 해석하여 각 연산 유닛에 작업을 시킬 수 있다. 마지막으로 CPU, GPU에 비해 인공지능 연산을 할 때 전력 소비를 적게 하도록 기능하는 전력 관련 유닛도 존재한다.
입출력 버퍼를 통해 데이터를 입력 받아, 디코더와 같은 제어 유닛을 통해 명령을 해석하고, 병렬 처리된 수많은 연산 유닛을 통해 연산을 하고, 데이터를 집계해서 출력하는 구조라고 할 수 있다.



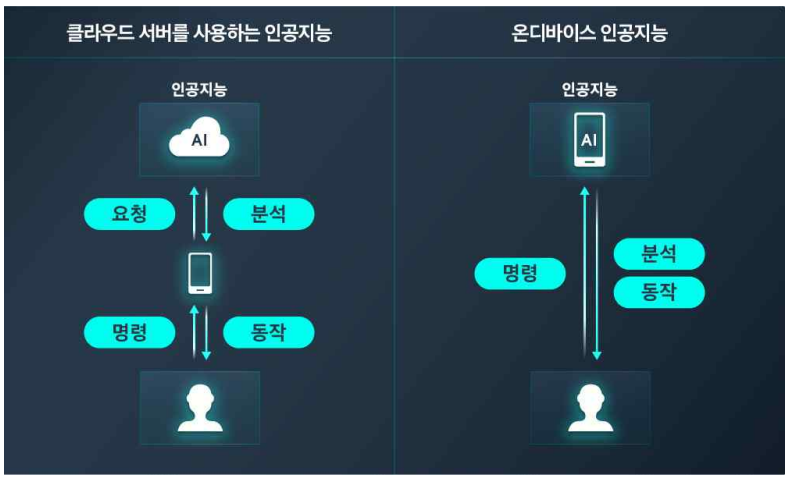
▲CPU, GPU, NPU 구조 비교

NPU는 SoC 구조에서 더 나아가 네트워크 온 칩이라는 NoC 구조를 사용한다. 네트워크 온 칩 구조는 SoC 구조처럼 한 칩에 여러 유닛들이 들어가지만, 그 다양한 유닛들이 네트워크를 통해 데이터를 통신하면서 더 빠른 데이터 전송을 할 수 있다.
이런 구조들을 통해서 NPU는 빠른 연산과 전력소비감소로 고성능, 고효율의 인공지능 연산 작업을 수행 할 수 있어서, 다양한 분야에서 사용될 수 있다.

4. NPU 응용 분야

앞서 말했듯이, NPU는 인공지능 연산 작업에서 낮은 전력 소비를 통한 고성능, 고효율의 퍼포먼스를 보여주기 때문에 다양한 분야에서 사용된다.
그 중 가장 대표적인 예시가 스마트폰이다. 맨 처음에도 언급했던 것처럼 현재 개발되고 있는 스마트폰들은 인공지능 기능들이 탑재되어 나오는 추세이다. 그 예시로는 사진 이미지에서 특정 부분만 감쪽같이 제거하는 기능도 있고, AI 비서 등도 있을 수 있다. 이처럼 스마트폰 같이 작은 모바일 기기에서 인공지능 작업 성능을 효율적으로 높이기 위해서는 전력 소비가 낮아야하고, 크기도 작아야한다. 따라서 '삼성전자'의 AP, '엑시노스9'처럼 요즘 스마트폰 AP에는 NPU가 탑재되어 개발되고 있는 추세이다.
스마트폰에서 인공지능 작업이 되어야 하는 이유 중 중요한 한가지 예시는 바로 추가적인 작업을 하지 않아도 된다는 것이다. 아래 사진은 클라우드 서버를 사용하는 인공지능과 온디바이스 인공지능의 작업 처리 과정이다. 클라우드 서버를 사용하는 인공지능은 사용자가 스마트폰을 통해 어떤 작업을 하면 스마트폰에서 클라우드 서버를 통해 인공지능 작업을 한 후 다시 사용자에게 제공하는 과정을 거쳤다. 하지만 온디바이스 인공지능처럼 스마트폰에 NPU를 통한 인공지능 작업을 할 수 있다면 사용자가 스마트폰에 작업을 요청하면 스마트폰에서 바로 서비스를 제공할 수 있다는 것이다. 이처럼 NPU의 개발은 연산 작업의 효율

뿐만 아닌 사용자에게 편의 또한 제공한다.



▲'삼성전자' 클라우드 서버를 사용하는 인공지능과 온디바이스 인공지능의 비교

이외에도 자율 주행 자동차의 실시간 객체 인식, 경로 계획 등에도 NPU가 사용되면 더 좋은 효율을 낼 수 있고, 스마트 시티의 실시간 교통 흐름 관리, 신호 제어 등 현재 인공지능이 사용되고 있는 다양한 분야에서 NPU가 사용된다면 효율과 성능이 무조건적으로 좋아질 것이다. 따라서 NPU 개발은 인공지능이 전세계적으로 많이 사용되고 있는 현대 사회에서 필수적으로 개발해야 할 과제인 것이다.

5. NPU와 GPU 장단점

위에서 계속 언급했듯이 NPU는 GPU에 비해 인공지능 연산 작업에 특화되어 있기 때문에 딥러닝 연산을 빠르게 할 수 있게 전력 소비도 더 효율적으로 할 수 있어서 모바일 기기에서 효과적이다. 하지만 모든 방면에서 NPU가 GPU보다 나은 것은 아니다. NPU는 딥러닝 연산에 특화되어 있기 때문에 범용적인 연산 작업에서는 GPU나 CPU만큼 다양한 작업을 할 수 없다. 반면 GPU는 딥러닝에서도 특히 학습 작업에 강점을 갖고 있고, 이외에도 그래픽 처리, 계산, 데이터 분석 등 다양한 병렬 처리를 할 수 있는 범용 프로세서로 사용할 수 있다. 또한 NPU는 아직 개발된 지 얼마되지 않았기 때문에 GPU에 비해 소프트웨어 생태계가 제한적이다. GPU는 TensorFlow, PyTouch 등의 딥러닝 프레임워크를 사용할 수 있기 때문에 소프트웨어 생태계가 잘 발달되어 있다고 할 수 있다. 하지만 GPU는 성능이 높을 수록 가격이 비싸고, 전력 소비가 크기 때문에 가격과 전력 소비를 줄여야하는 모바일 기기 시장에서는 NPU에 비해 불리하다.

요약하자면, NPU와 GPU 모두 병렬 처리를 통한 빠른 연산을 할 수 있다. GPU는 범용성이 높아서 다른 작업들도 수행할 수 있고, 딥러닝 학습 작업에서 강점을 갖는다. 하지만 가격이 비싸고, 전력 소비가 크다는 단점이 있다. 반면에 NPU는 전력 소비가 효율적이고, 딥러닝 연산에 최적화되어 실시간 AI 연산에 강점을 갖는다. 하지만 범용성이 떨어져 AI 작업에만 사용할 수 있고 아직 관련 소프트웨어 지원이 부족할 수 있다. 따라서 어느 한 쪽이 무조건 압도적으로 좋다고 할 수 없지만 NPU의 개발이 앞으로의 인공지능 성능에 큰 영향을 끼칠 것이기 때문에 많은 관심을 갖고 지켜봐야 한다고 생각한다.

참고 자료)

삼성 뉴스룸 - ['엑시노스' 개발 리더들이 SoC를 말하다] ② CPU · NPU 알아보기
https://news.samsung.com/kr/%EC%97%91%EC%8B%9C%EB%85%B8%EC%8A%A4-%EA%B0%9C%EB%B0%9C-%EB%A6%AC%EB%8D%94%EB%93%A4%EC%9D%B4-soc%EB%A5%BC-%EB%A7%90%ED%95%98%EB%8B%A4-%E2%91%A1-cpu-%C2%B7-npu-%EC%95%8C%EC%95%84

삼성전자 - NPU (Neural Processing Units)
https://semiconductor.samsung.com/kr/support/tools-resources/dictionary/the-neural-processing-unit-npu-a-brainy-next-generation-semiconductor/