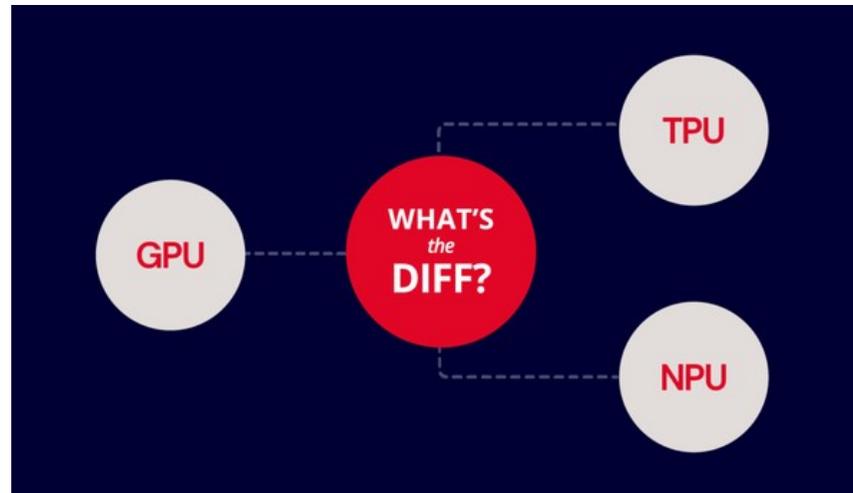


[techtube.co.kr](https://www.techtube.co.kr)

CPU•GPU•NPU•TPU의 차이

~2분

인공지능(AI) 개발에 필수적인 머신러닝에는 GPU, NPU, TPU 등의 프로세싱 칩이 사용되고 있는데, 각각의 차이점을 알기 쉽지 않다. 구글과 클라우드 스토리지 서비스를 제공하는 'Backblaze'가 CPU, GPU, NPU, TPU의 차이점을 정리했다.



출처 : Backblaze

CPU란?

CPU는 'Central Processing Unit'의 약자로, PC에서 문서 작성, 로켓

의 경로 계산, 은행의 거래 처리 등 다양한 용도로 사용되고 있다. CPU에서도 기계학습을 실시하는 게 가능하지만 계산할 때마다 메모리에 접속하는 특징이 있어, 기계학습에 필요한 대량의 계산을 실행할 때에는 메모리 통신 속도가 병목현상이 돼 처리속도가 느려진다.

GPU란?

GPU는 'Graphics Processing Unit'의 약자로, GPU와 메모리, 입출력 장치 등을 세트로 구성한 그래픽 보드가 시중에 널리 유통되고 있으며, GPU에는 수천 개의 산술논리연산유닛(ALU)이 탑재돼 '량의 연산을 동시에 수행하는 처리가 CPU에 비해 압도적으로 빠른 기계학습이 가능하다. 하지만 GPU는 게임이나 CG 처리 등에도 활용할 수 있는 범용 칩이기 때문에 머신러닝 전용으로 설계된 칩에 비해 효율이 떨어진다.

NPU란?

NPU는 'Neural Processing Unit' 또는 'Neural network Processing Unit'의 약자로, GPU와 마찬가지로 대량의 연산을 동시에 수행하는 처리에 특화되어 있다. 또한, NPU는 머신러닝 전용으로 설계된 칩이기 때문에 GPU보다 더 효율적으로 연산을 수행할 수 있다. 반면 NPU는 머신러닝 전용 설계 칩이기 때문에 CPU나 GPU처럼 다양한 용도로 사용할 수 없다는 단점이 있다.

최근 출시되는 스마트폰에는 NPU를 탑재한 제품들이 많이 등장하고 있으며, 아이폰 시리즈에는 'Neural Engine'이라는 이름으로 탑재되어 있고, 스마트폰용 SoC인 '스냅드래곤(Snapdragon)'에도 NPU를 탑재한 모델이 존재한다.

TPU란?

TPU는 'Tensor Processing Unit'의 약자로 구글이 개발한 NPU의 일종이다. 구글은 클라우드 컴퓨팅 서비스 'Google Cloud'를 통해 사용자에게 TPU의 처리 능력을 제공하고 있다. 사용자는 스스로 하드웨어를 준비하지 않아도 머신러닝 관련 처리를 고효율로 수행할 수 있다.

구글은 자사가 제공하는 'TPU v4'가 엔비디아의 머신러닝 관련 처리 특화 GPU 'A100'보다 1.2~1.7배 빠르고, 전력 효율도 1.3~1.9배 우수하다고 주장하고 있다.

(참고 : backblaze.com/blog)