voidint.com

GPU vs NPU 구조 및 NPU 목적 - voidint.com

Jay

4~5분

GPU NPU 구조 및 이러한 하드웨어를 사용할 때의 딥러닝 방식은 어떤 차이를 유발할까요.

대략 2016년 정도부터 본격적으로 딥러닝 기반 AI가 학계, 산업계에서 주목받기 시작했던 것 같은데 벌써 여러 반도체 기업들이 NPU 또는 주변장치들을 묶어 NPU 기반 SoM 이나 개발보드 등을 출시하고 있습니다.

대체 NPU 가 무엇이고 왜 AI 에 필요한 것인지 좀 더 자세히 살펴보도록 하겠습니다.

아래 앞전에 작성한 글에서 CPU, GPU, NPU, TPU 가 무엇인지 설명했었는데, NPU 소개가 너무 짧아서 약간 더 자세하게 살펴보려고 합니다. 혹시 딥러닝을 통한 AI 구현에서 왜 GPU 를 사용하는지 아직 모르겠다면 아래 글도 한번 일독해 보시기를 권합니다.

NPU 는 Neural Processing Unit 의 약어입니다. Neural Network 즉 인공신경망을 통한 인공지능 연산을 좀 더 효율적으로 해보자 라는 목적으로 설계된 프로세서입니다. 개념적으로는 이렇지만 실제 유닛을 구성하는 요소들이나 인터페이스는 설계 회사 또는 수요 기업군에 따라 매우 다양합니다.

GPU vs NPU 구조

그렇다면 GPU 에 비해서 NPU 가 가지고 있는 장점은 무엇일까요?

GPU 도 이미 병렬 대량 연산을 위한 구조로 설계된 프로세서인데 말이지요.

GPU 는 부동소수점 곱셈을 동시에 엄청나게 많이 처리할 수 있는 유닛입니다. 게임렌더링도 마찬가지이겠지만, 딥러닝 학습이나 추론 과정에서 곱셈이 많이 발생한다 하더라도 곱셈기만 있다고 딥러닝 학습이 되는 것은 아닐 겁니다.

곱셈을 어떻게 처리하고 결과는 어떻게 정리할 것인지 GPU 에게 지시를 하는 방법은 필요합니다.

CUDA & cuDNN

GPU 를 이용해 일련의 과정을 수행하기 위해서 NVIDIA 에서는 GPU 를 사용하기 위한 CUDA 와 cuDNN (CUDA Deep Neural Network library) 을 제공합니다. CUDA 는 C 언어를 이용하여 GPU 의 계산을 스케줄링할 수 있도록 제공하는 GPU 제조사 NVIDIA 의 일종의 SDK 입니다.

CUDA 를 이용하여 병렬 계산 알고리즘을 구현할 수 있고, cuDNN 을 이용하면 기본적인 딥러닝 primitive 들을 편리하게 구현할 수 있습니다. 그리고 바로 여기에 NPU 의 개념에 대한 힌트가 있습니다.

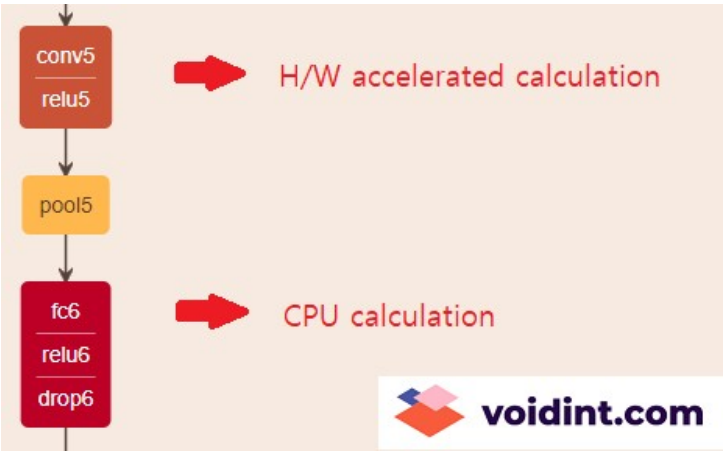
NPU 구현 방식

NPU 는 딥러닝 네트워크를 구성하는 각 layer 를 실리콘으로 구현한 칩셋입니다. 예컨대 Convolution, Fully Connected 와 같은 인공 신경망을 구성하는 뉴런을 하드웨어적으로 구현했다는 의미입니다.

NPU 를 이용하는 딥러닝 모델은 GPU 와 cuDNN 을 사용할 때와 같은 100% 자유도를 가지고 네트워크를 구성할 수는 없습니다. 실리콘으로 구현한 뉴런은 칩 설계시 정해져 있고 NPU 에서 구현하지 않은 뉴런이 딥러닝 네트워크에 포함되어 있다면, 그 부분은 하드웨어 가속을 타지 못하고 CPU 로 연산을 수행해야 하기 때문에 NPU 를 사용하는 이점이 약화됩니다.

만약에 어떤 가상의 NPU 가 convolution 은 구현하였지만 fully connected 계층은 구현하지 않았다고 한다면, 다음 그림과 같은 네트워크에서는 CPU 성능과 경우에 따라 메모리 대역폭도 소모해야

하기 때문에 성능이 저하될 것입니다.



NPU 에서 FC 를 지원하지 않는다면 이런 네트워크는 효율이 떨어짐

NPU 를 왜 사용하는가

그렇다면 왜 NPU 를 사용하는 것이고, 수많은 글로벌 실리콘 회사들이 왜 앞다투어 NPU 를 제작하고 있는 것일까요?

기존의 딥러닝이 너무 비싸기 때문입니다

기존의 PC 또는 워크스테이션과 고가의 GPU 를 이용한 딥러닝은 비싸다는 점을 제외하면 단점이 없습니다.

특히 AI 분야는 계산 능력이 곧 품질로 직결되기 때문에 비용을 투입하는 것에 비례해서 빠르게 좋은 모델이 학습되는 것이 일반적입니다. 그리고 새로 설계되는 뉴런/텐서/레이어가 발생하면 cuDNN 과 같은 소프트웨어 라이브러리를 구현하면 GPU 에서 가속을 태울 수 있어 유연성도 갖추고 있습니다.

그런데 아래의 애플 아이폰 12 프로에 대한 글에 잠시 언급한 것과 같은 작업을 하기 위해, 핸드셋에서 딥러닝 추론을 수행하고자 한다면 이 때 GPU 를 사용할 수 있을까요?

딥러닝 추론을 edge 로 가져 오기 위해

기존 GPU 시스템은 너무 비싸고 휴대할 수 없으며, 전력 소모도 매

우 큽니다. 예컨대 휴대폰에 GPU 를 넣을 수는 없습니다. 그렇다고 휴대폰 AP에 포함된 CPU 만 가지고 추론을 하기에는 연산 능력이 아쉽습니다. 물론 배터리도 문제일 것이고요.

애플 아이폰처럼 저조도 사진의 품질을 극대화하거나, 테슬라처럼 전기차로 자율주행을 하고자 할 때에도, 길거리 CCTV 에서 테러범을 찾아내거나 상점의 방문자 추이를 파악하는 데에도 딥러닝을 통한 추론은 그 동안 사용해 왔던 모든 방법을 압도하는 성능과 비용절감 효과를 가져다 줄 수 있습니다.

그리고 이 모든 일들은 말단 장비 자체에서 이루어져야만 합니다. 각 말단 장비마다 초당 수 메가바이트의 영상을 생산하고 있을텐데, 이를 클라우드나 워크스테이션 서버로 업로드 하는 것 자체가 불가능하기 때문입니다.

GPU + NPU 구조

다시 처음으로 돌아와서 GPU vs NPU 가 아닌 GPU + NPU 로 제목을 써 봤습니다. 딥러닝 추론은 NPU 를 탑재한 말단 장비에서 이루어져야 합니다. 하지만 NPU 를 이용해 구동하고자 하는 딥러닝 모델은 GPU 를 탑재한 대형 서버에서 학습하는 것이 좋을겁니다. 즉, 모델 학습을 빠르게 하려면 GPU 도 필요하고, 말단 장비에서 저전력으로 추론하려면 NPU 도 필요합니다.

그런데 구글 사례를 보면 자체 클라우드 시스템에서 TPU 병렬 구성을 통해 학습도 TPU 를 이용해 하고 있는 것으로 보입니다. 이런 추세를 볼 때 점차 학습도 병렬 구성으로 클라우드에 올라간 NPU 가 담당하게 되지 않을까 예상해 봅니다.

물론 그렇게 되기 위해서는 일반적으로 사용되는 뉴럴 네트워크 및 네트워크를 구성하는 개별 노드 형태가 일정 수준 이상 정형화 되어야 할 것입니다. NPU 는 실리콘으로 구현하기 때문에 소프트웨어의 유연함을 기대하는 데이 한계가 있기 때문입니다.

NPU 시장 상황

GPU 시장은 NVIDIA 독주 체제가 오랜 기간 지속되어 왔기 때문에 말할 필요가 없을 것입니다. 그런데 NPU 시장은 약간 상황이 재미있게 돌아가고 있는 것 같습니다.

NVIDIA

역시 NVIDIA 가 가장 발빠르게 Autonomous 타깃으로 Jetson 시리즈를 발표한 바 있으며, 올해 여름에 대략 60만원대의 NX 를 발표하였습니다. 기존 100만원을 넘어가던 Xavier 보다 가격은 크게 떨어뜨리고 성능은 가격만큼 큰 차이가 나지는 않게 출시했습니다. 현존하는 NPU SoM 중에서 가장 밸런스가 좋은 제품이라고 생각합니다. 하지만 비록 기존 Xavier 보다 가격을 낮추기는 했지만 60만원대의 가격도 받아들이기에 부담스러운 수준입니다.

Google

Google 도 자체 클라우드 시스템에도 탑재하는 TPU 를 만들어서 상용화에 성공했습니다. Coral 이라는 브랜드로 TPU 와 SoM 을 비롯해 각종 인터페이스 제품을 판매하고 있습니다. 가격은 SoM 기준 99.99$ 정도로 NX보다 가격은 매력이 있습니다. 그런데 성능 벤치마크가 4TOPS 정도로 NVIDIA NX 기준 대략 20% 수준입니다. TPU 2세대 모델이 출시 대기중이며, 구글에서 인공지능 대회에 출전할 때 DDR이 들어있는 2세대 모델을 가지고 나왔는데 평가 방식에 논란은 좀 있었지만 그래도 성능이 상당히 좋은 것으로 평가되었습니다.

Qualcomm

퀄컴도 스냅드래곤 845 와 855, 865 칩셋에 NPU 를 탑재한 모델을 출시했습니다. 아무래도 휴대폰에 많이 사용되는 칩셋이기 때문에 전력 사용에 민감하게 최적화 했을 것으로 예상됩니다. 865 칩셋에 장착한 NPU가 제조사 벤치마크 15TOPS 정도 성능입니다. 자세한 스펙은 제조사 데이터시트 참고하시기 바랍니다.

Huawei (Hisilicon)

Huawei 자회사 Hisilicon 에서도 2018년에 NPU 를 출시했었지만, 현재는 미국의 제재로 제품 로드맵 발표를 중단하였습니다. 2018년에 발표한 칩셋중 하이엔드급이 대략 4~6TOPS 정도 성능으로 알려졌습니다.

SKT

국내 기업으로 오늘자 뉴스에 발표된 내용이 있어 링크합니다.