인공지능플랫폼최적화 연구논문요약

Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge

2021234640 이종현

1. Intro

최근 모바일 디바이스가 보편화되고 각종 모바일 디바이스에서 구동되는 AI 어플리케이션이 급증하고 있다. 그러나 모바일 디바이스가 갖는 가장 큰 문제 중 하나는 전원을 배터리를 통해 공급받는다는 점과 크기가 제한되기 때문에 컴퓨팅 파워가 낮다는 점이다. 따라서 대부분의 모바일 디바이스에서의 AI 어플리케이션은 인풋으로 받은 데이터를 네트워크를 통해 데이터 센터로 전달하고 이를 데이터 센터에서 처리한 후, 결과를 다시 edge 디바이스로 전달하는 status quo 방식이 일반적이다.

그러나 여기에서 다음과 같은 질문을 던져볼 수 있다. 과연 모든 데이터를 cloud 환경에 의존해서 처리하는 것이 가장 좋은 방법인가. 본 연구는 이 의문에 대해 cloud 환경과 edge 디바이스 환경에서의 최적의 협력 관계의 가능성을 제시하고 있다.

1. Experiments & Solutions

대부분의 모바일 환경에서 cloud only 인 환경, 즉 status quo를 선택할 수밖에 없는 것은 모바일 디바이스의 배터리 문제와 부족한 컴퓨팅 파워에서 기인한다. 배터리는 상시 전원을 공급받는 방식이 아니기에 연산량이 많아져 배터리 소모가 많아지면 기기의 사용 시간이 줄어들게 되어 사용자에게 큰 불편을 야기한다. 또한 모바일 디바이스의 특성상 컴퓨팅 파워가 제한된 환경이기에 무거운 AI 모델을 구동하는 것도 computing latency를 발생시킨다. 반면에 데이터를 datacenter로 전송하여 처리하는 cloud 방식은 앞서 언급한 모바일 디바이스의 한계를 모두 상쇄한다. 그러나 데이터를 전송하는 과정에서 발생하는 communication latency, 즉 network I/O가 존재한다. 동시에 여러 디바이스에서 cloud로 데이터를 전달하게 되면 물리적 한계가 존재하는 datacenter에서도 무리가 발생할 수 있다.

Neurosurgeon은 2개의 regression모델을 도입한다. 각각 computing efficiency에 대한 모델과 power efficiency에 대한 모델이다. 이는 사용자의 목적에 따라 최적화 방향이 다를 수 있기 때문이다. 각 regression 모델은 딥러닝 모델이 edge 디바이스에서 구동되었을 때, cloud 환경에서 구동되었을 때의 소요되는 시간과 에너지를 계산한다. 이 결과를 바탕으로 Neurosurgeon은 연산량 또는 에너지 소모량이 최소화되는 layer를 결정하게 된다.

Convolution 연산이 반복적으로 실시되는 VGG 16 모델에서 computing latency를 최적화하는 지점은 #5 pooling layer이다. 이는 fully connected layer의 연산이 상대적으로 무겁고 pooling layer 를 연속으로 거치며 data의 dimension이 줄어 communication latency가 줄기 때문이다. 반면 energy efficiency에 대한 의사결정은 input layer 였다. 시작하자마자 데이터를 datacenter에서 처리하는 것이 edge device에서 CNN 연산을 수행하는 것보다 energy efficient하다고 판단했기 때문이다.

A picture containing timeline

Description automatically generated

그림 01. VGG 16 모델에서 (상) layer별 연산량, input-output dimension, (중) computing efficiency, (하) energy efficiency.

연구자들은 Neurosurgeon 방식을 computer vision 방식 뿐만 아니라 다양한 과제들에 대해서 적용하였다. Face recognition, NER, POS 등의 과제를 수행하는 모델들 역시 각각의 목표에 따른 최적 partitioning포인트를 찾아낼 수 있었다. Neurosurgeon 방식은 기존 cloud-only 방식에 비해서 최대 40.7배와 평균3.1 배의 속도 개선, 최대 94.7%, 평균 59.5%의 배터리 절감 성과를 보였다.

1. Discussion

본 연구는 Neurosurgeon이라는 단어에 드러나듯, 뉴런의 구조를 흉내낸 DNN 모델을 잘 ‘수술’하여 성능 개선을 이뤄냈다. 개인적으로 가장 흥미로웠던 점은regression 모델을 이용하여 최적점을 찾아내고 이를 통해 기존 status quo 방식보다 큰 개선 효과를 보였다는 점이다. 가장 큰 장점은 Neurosurgeon 방식이 특정 모바일 AP, cloud 환경, 혹은 네트워크 조건에서만 작동하는 것이 아니라 범용적으로 모든 상황에서 적용이 가능하다는 점이다. 향후 모바일 AP의 발전은 지속되겠지만 이에 맞추어 DNN 역시 복잡하고 더 많은 연산을 요구할 것이다. 데이터 센터는 물리적 제약이 있기 때문에 Neurosurgeon은 edge 디바이스와 cloud 환경을 가장 효율적으로 쓸 수 있는 방법으로 기대된다. 또한 가장 큰 장점은 현재 널리 쓰이는 모델을 별도의 변형 과정 없이 사용이 가능하다는 점이다. 추가적으로 다른 최적화 방법과 병행 실시하게 되면 그 성능은 기하급수적으로 향상될 것이다.

본 연구를 보고 발생한 의문은 모델을 partitioning한 본 연구와 다르게 데이터를 partitioning하는 방법에 대한 가능성이다. 즉 데이터를 parallel로 처리할 수 있지 않을까에 대한 의문이다. 상대적으로 연산 능력이 떨어지는 edge 디바이스는 전체 데이터의 40% 정도만 처리하고 네트워크 병목을 줄이기 위해 cloud에는 전체 데이터의 60%만 전달하고 해당 결과는 다시 edge 디바이스에서 합치는 방식도 가능할 것 같다는 생각을 해보았다.

향후 무선 네트워크의 성능이 개선될 것으로 기대되는 상황이다. (예를 들어 5G) 네트워크 상황이 개선되면 cloud-only 접근이 오히려 더 좋아질 수 있지만, 3G, LTE 등이 혼재되어 있는 최근의 상황에서 가장 적절하다고 생각된다.