기말고사 대체 보고서

인공지능과 미래산업 특강

고려대학교
기계공학과
2022011038
장대윤

반도체 업계 리딩기업들이 집중하는 AI 모델 최적화 및 HW-aware MLOps 기술과 그 적용 사례 - 12th lecture <Nota AI 채명수 대표님>

1. 서론

카메라와 같은 이미지 추출 장치들의 크기는 줄어들고 그 성능은 향상되고 있다. 특히 NVIDIA와 같이 그래픽카드에 사용되는 메인 칩셋, 즉 MCU의 최적화에도 그 영향을 주기 시작했다. 각종 모듈에 사용되는 메인 MCU의 크기는 작으면서 그 성능을 최대로 끌어올리려 한다는 것이다. 이런 상황에 따라 개발과 운영을 따로 나누지 않고 의 생산성과 운영의 안정성을 최적화하기 위한 방법론에 머신러닝 기술을 적용한 MLOps가 프로세스 발전에 박차를 가하고 있다. 사실 기존의 AI 연구에서는 실제 모델들의 학습구조를 어떻게 하면 자신들의 데이터에 최적화될지를 꽤 많이 신경 썼다. 물론 이 부분은 궁극적으로 중요한 부분이다. 하지만 앞서 말한 MLOps 방법론에 따라 반도체 업체들이 특별하게 최적화 부분에서 포커스하는 새로운 부분이 생겼다. 바로 '인공지능 압축 기술'이다. 이는 연산량과 알고리즘 모델 등 관련된 환경 모두를 최적화하여 같은 성능을 낼 수 있도록 하는 기술을 말한다.

위의 내용에 따라 우리는 '인공지능 압축 기술'에는 어떠한 것들이 있는지, 그리고 이 기술을 사용한 실제 국내·국외 사용 사례에 확인하고 결과를 살펴보며, 긍·부정적 영향, 전망에 대해서 분석해본다.

2. 본론

2.1. 인공지능 압축 기술

- 가중치 가지치기

기존 신경망이 가지고 있는 가중치 중 실제 추론을 위해 필요한 값은 비교적 작은 값들에 대한 내성을 가지므로, 작은 가중치를 모두 0으로 하여 네트워크의 모델 크기를 줄이는 기술이다.

- 양자화 및 이진화

양자화와 이진화는 기존의 신경망의 부동 소수점 수를 줄이는 데 그 목적이 있으며, 양자화의 경우 특정 비트 수만큼으로 줄여서 계산하는 방식이다.

- 가중치 공유

낮은 정밀도에 대한 높은 내성을 가진 신경망의 특징을 활용해 가중치를 근사하는 방법이다.

2.2. 사례

– 국내

NetsPresso는 딥러닝 모델 자동 경량화 플랫폼이다. 사용자가 이미 딥러닝 모델을 가지고 있다는 전제 하에, 해당 모델을 입력으로 넣으면 NetsPresso는 더 작은 모델로 경량화하여 출력해준다. 가지치기, 양자화 등의 경량화 작업을 진행하는데, 여기서 특별한 것은 사용자가 입력값으로 넣어주는 HW의 특성이다. NetsPresso는 모델과 HW의 조합성능치까지 측정해주는데, 이를 통해 연산량에서 오는 문제를 해결해주고, 비용과 온디바이스 AI에서도 사용 가능하도록 안내해 줄 수 있다.

평택시의 교통실험도 진행하였다. 실제 교통신호의 제어에 압축 인공지능 모델을 얹은 것인데, 기존의 클라우드 기반의 데이터 수집이 아닌 객체 탐지 기술을 활용해 실시간으로 혼잡도를 분석하고 교통신호를 생성시켜 준다. 이 결과로 통행의 속도가 약 3배 증가하였다.

대한영상의학회에서는 의료 영상에 최적화된 모델 최적화 기법을 적용하였다. 실제 임상에 적용하기 위한 모델을 만들어야하는데, 방대한 양의 데이터를 정확하지만 빠르게 처리시켜야 한다. 특히 의료계에서는 환자에게 정확한 진단과 치료를 제공해야 하기 때문에 AI의 기술을 적용하기 위해서는 그 신뢰도를 다른 어떠한 분야보다 높게 가져가야 한다. 허나 모델 경량화를 통해 영상의 처리 정확성은 동일하지만 실시간성과 다양한 임상 결과를 향상시켰다.

– 국외

퀄컴은 차세대 비디오 압축표준 VVC(Versatile Video Coding) 압축으로 비디오 파일 크기가 약 1000배 감소하는 것과 같은 기술을 개발하였다. 또한 음성 압축도 같이 진행하여, 기존의 ‘피드백 순환 변형 자동 인코더’에 비해 전송률이 2.6배 향상되었다. 이는 압축 과정에서 인공지능 모델을 압축하는 기술을 적용하였다. 레이블이 지정되지 않은 훈련 데이터를 가져와 동일한 분포에서 새 샘플을 생성하는 비지도 학습을 통해 심층 모델을 생성하고 적용하였다.

3. 결론

현재는 개인 데스크탑 수준에서도 인공지능 학습이 가능한 환경이 되었다. 이것도 다양한 AI 이론이 활성화되면서 생긴 현상이다. 허나 위처럼 높은 수준으로 인공지능 모델의 압축이 가능해지면서, 스마트폰과 같은 소형 모듈에서도 객체

인식 등의 프로세스를 진행시킬 수 있게 되었다.

긍정적인 부분을 살펴보자면, 당연히 Big Data 시대에 더욱 효과적인 프로세스를 진행할 수 있다는 것이다. 인공지능 학습모델은 넘쳐흐를 정도로 새롭게 나오고 있는데, 이 모델이 얼마나 하드웨어 리소스를 잡아먹을지는 모른다. 허나 최적화 및 압축 기술이 적용될 경우 적은 에너지로도 학습이 가능할 것이다. 물론 줄어든 모델의 용량만큼 데이터의 품질을 높일 수도 있을 것이다.

반대로 부정적인 부분을 살펴보자. 최적화가 기술적으로 보편화되면 다양한 플랫폼과 모듈에서 인공지능을 쉽게 사용할 수 있을 것이다. 하지만 이는 모든 상황에서 인공지능만 활용하려 하는 경향성을 보일 수 있으며, 이에 따라 사용자에게 제공되기 위한 기준정책이 변화할 수 있다. 즉, 안정성이 저하될 수 있는 상태가 된다는 의미이다. 문제가 발생한 상태에서, 해당 환경에 애매함이 존재할 경우 사용자와 기업 중에서 누가 책임이 있을지에 대한 공방 문제가 빈번하게 발생할 수 있다.

부정적인 영향이 발생할 수 있음에도 불구하고 최적화 및 경량화에 대한 긍정적 부분이 더욱 앞으로의 사회와 기술 발전에 더 유효한 의마와 작용이 있을 것이라 생각된다. 최적화뿐만 아니라 데이터를 수집하는 장치도 더욱 정교해질 것이고, 기술들의 단점은 계속 보완될 것이다. 물론 모든 연구와 기술은 단점이 보완되어야 한다. 본인이 말하는 것은, 인공지능이 현재 생각 이상으로 가속화된 상태이기 때문에 발전이 그만큼 더 빠르게 진행될 것이라는 것이다. 앞으로는 인공지능 전문가뿐만 아니라 AI의 발전에 맞춰 데이터를 처리할 기본적인 장치의 향상을 이루어줄 인력들도 늘어나야 하지 않을까라는 생각도 든다. 다양한 업종에 인력을 투자할 수 있도록 교육과 가치 방향성을 설정하고 대비하는 것이 중요한 관점일 것이다.

References

1. M.Y.LEE, J.H. Chung, J.H. Lee, J.H, Han, Y.S. Kwon, Trends in AI Processor Technology, 2020, Electronics and telecommunications trends v.35 no.3 , pp. 66-75
2. Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High Fidelity Neural Audio Compression. arXiv preprint arXiv:2210.13438.