Lightweight Deep Learning

박태우

경량 딥러닝이란?

인공 신경망이 상당히 deep한 모델들을 연산량을 줄이고 효율적으로 만든 딥러닝 모델.

딥러닝의 성능을 향상시키기 위해 최근 모델들의 연산량은 기하급수적으로 커지고 있음.
하지만 이와 반대로 하드웨어의 성능은 상당히 느리게 증가하는 중.

다양한 분야에서 딥러닝을 사용하면서 임베디드 관련 분야에서도 딥러닝을 사용하게 됨.
이에 따른 하드웨어의 한계를 극복하기 위해 딥러닝 경량화에 많은 연구를 진행중임.
(핸드폰, 차동차 임베디드 등)

딥러닝 경량화 방법

1) Pruning

학습 후에 가중치가 상당히 낮은 부분을 제거함으로써 연산량을 줄임.

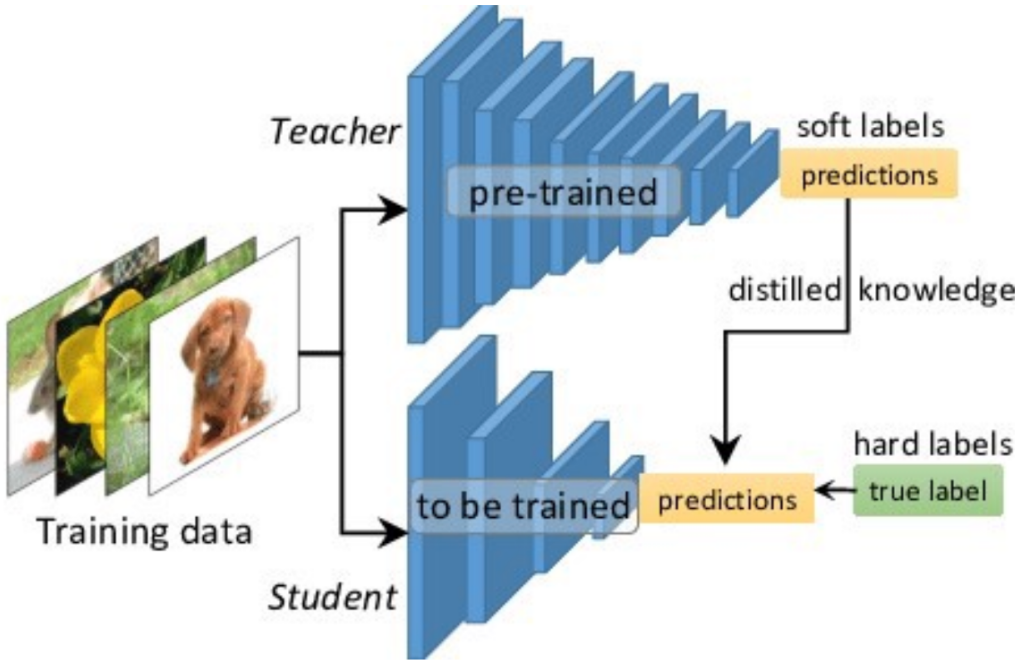
2) Knowledge Distillation

미리 잘 학습된 네트워크(Teacher)로부터 사용하고자 하는 네트워크(Student)를 학습시키는 방법.

1) Soft Label로 Teacher의 결과와 Student 결과 비교.

2) 실제 Label과 Student 결과 비교.

1), 2) 모두 cross entropy 사용.



딥러닝 경량화 방법

3) Weight Factorization

가중치 행렬을 분해하여 두 개의 작은 행렬의 곱으로 근사하는 방법.

4) Weight Sharing

모델의 일부 가중치들을 다른 파라미터들과 공유하는 방식.

5) Quantization

부동 소수점 값을 잘라내서 32bit, 64bit가 아닌 16bit, 8bit 수준의 precision을 사용해 적은 정확도 손실로 고속 연산을 가능하게 함.

References

https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

https://light-tree.tistory.com/196

https://blog.est.ai/2020/03/%EB%94%A5%EB%9F%AC%EB%8B%9D-%EB%AA%A8%EB%8D%B8-%EC%95%95%EC%B6%95-%EB%B0%A9%EB%B2%95%EB%A1%A0%EA%B3%BC-bert-%EC%95%95%EC%B6%95/