Google Developer Student Clubs
Dongguk University

모델 경량화란?

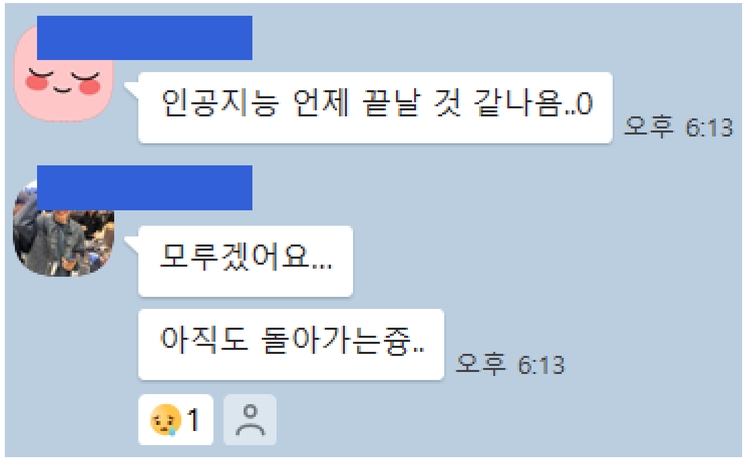
AI/ML
산업시스템공학과
김현진

Google Developer Student Clubs
Dongguk University

Contents

경량화의 필요성

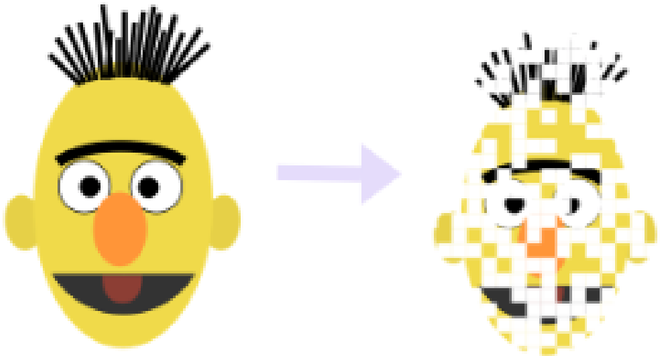
Google Developer Student Clubs
Dongguk University



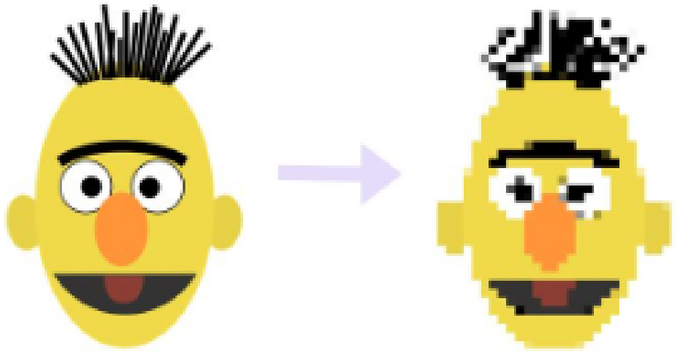
- **높은 비용**: 거대한 모델을 학습시키는 데는 엄청난 계산 자원이 필요하므로, 높은 경제적 비용이 발생함
- **추론 비용**: 추론 과정에서도 많은 계산 자원이 필요하므로, 실시간으로 빠른 응답이 요구되는 서비스에 모델을 적용하는 것을 어렵게 함
- **접근성 문제**: 거대한 모델은 모바일폰이나 임베디드 기기에서 활용불가

모델 경량화란?

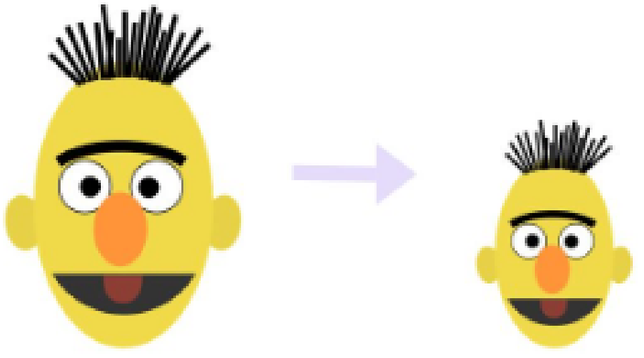
Google Developer Student Clubs
Dongguk University



▲ Pruning



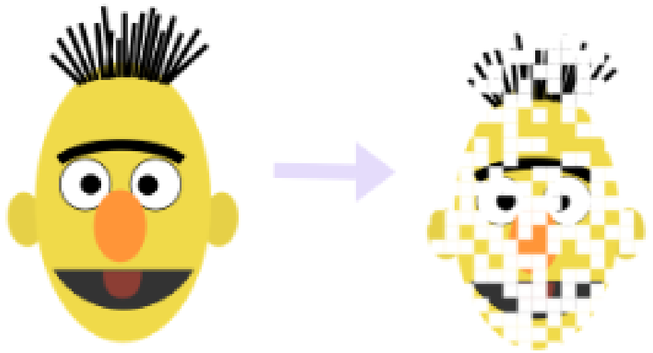
▲ Quantization



▲ Distillation

- **Pruning:** 중요하지 않은 부분을 적절히 가지치기를 해서 줄일 것이냐
- **Quantization:** 해상도를 낮춰서 작게 만들 것이냐
- **Distillation:** 사이즈 자체를 작게 만들 것이냐

Pruning



▲ Pruning

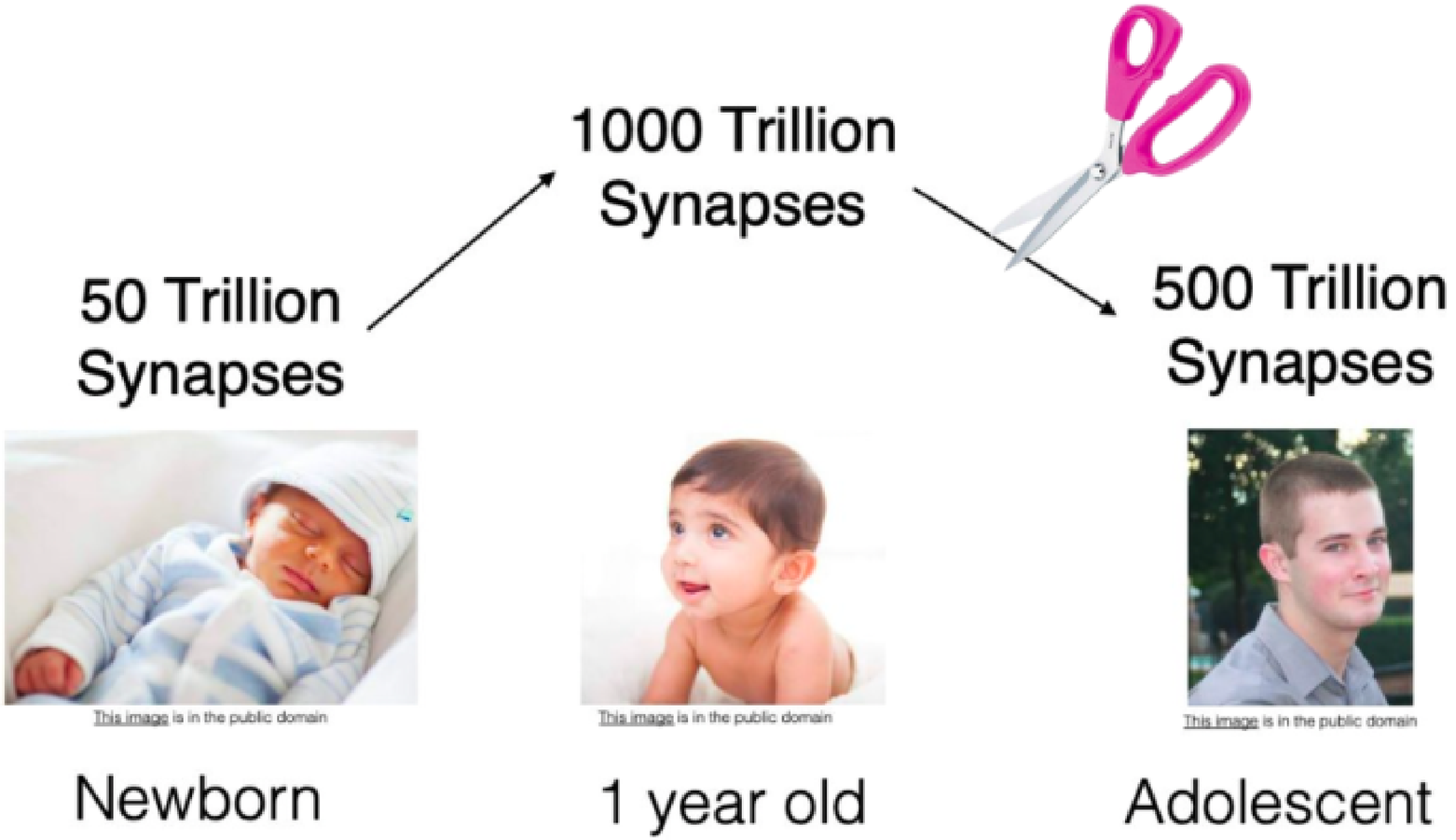
개요

딥러닝 모델의 많은 레이어 중 중요하지 않은 파라미터를 지워 모델을 경량화하는 방법

Weighted Sum for Pruning

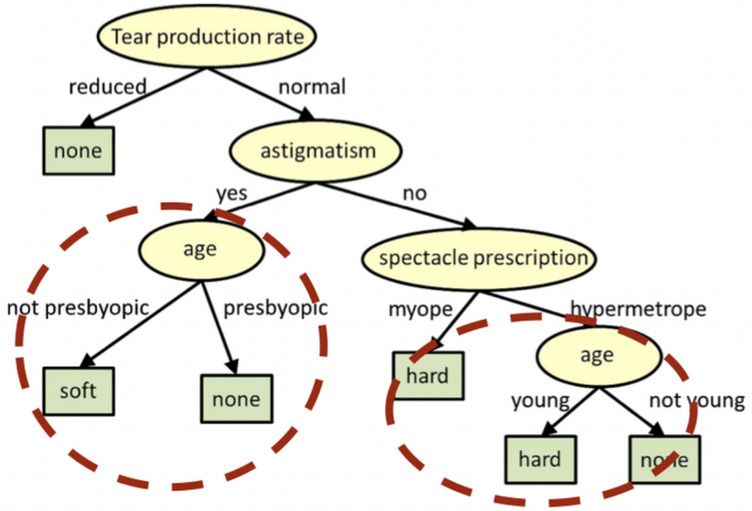
한 레이어에 존재하는 여러 파라미터의 중요도에 따라 다른 가중치를 적용하여 중요한 부분과 중요하지 않은 부분의 크기를 다르게 만듦

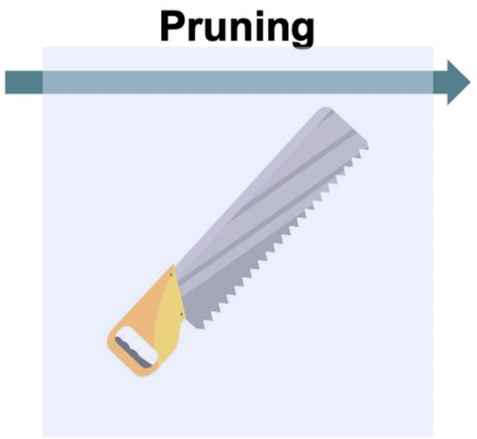
Pruning

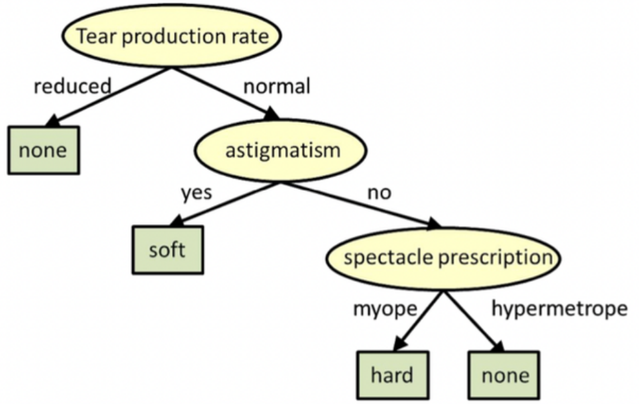


사람은 성장하면서 뉴런의 수 감소

Pruning







얻는 것

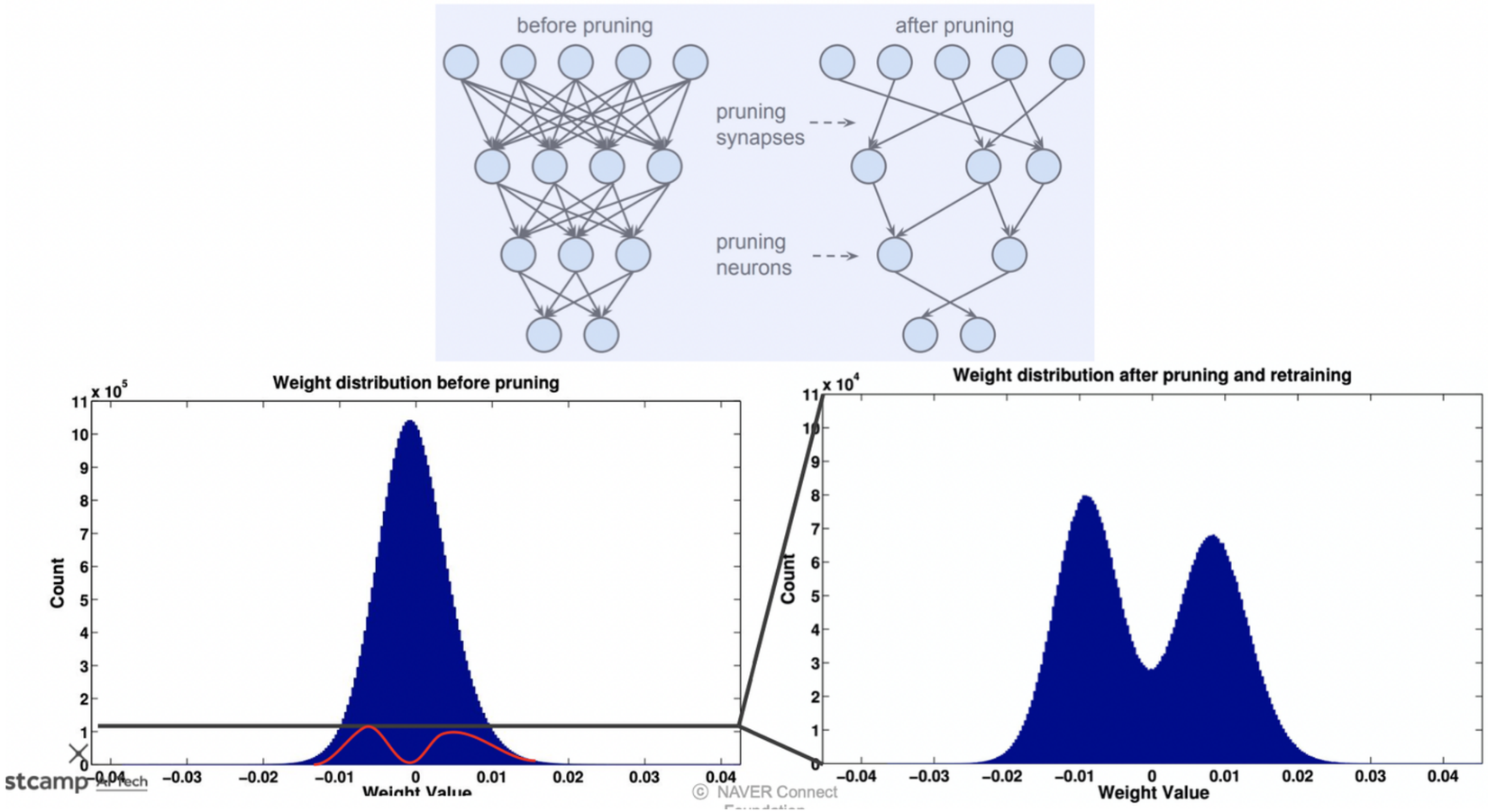
- Inference speed
- Regularization (lessen model complexity) brings generalization

잃는 것

- Information loss
- The granularity affects the efficiency of hardware accelerator design

사람과 마찬가지로 많은 뉴런 중 중요한 부분만 선별하여 크기를 줄임

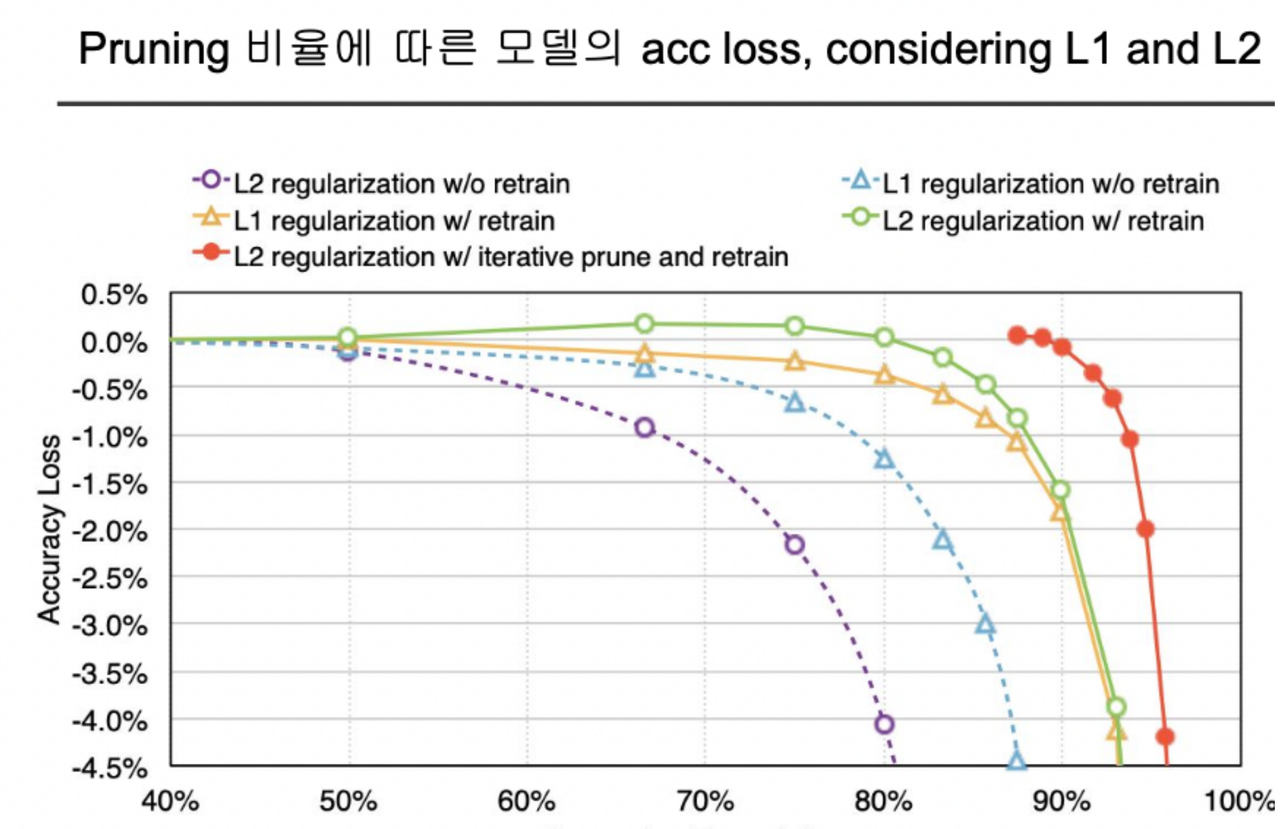
Pruning



파라미터의 값에 따라 가중치가 적용되기 때문에

아래 pruning 결과 분포와 같이 0 주변의 웨이트들이 많이 사라진 것을 볼 수 있음

Pruning



Pruning 결과

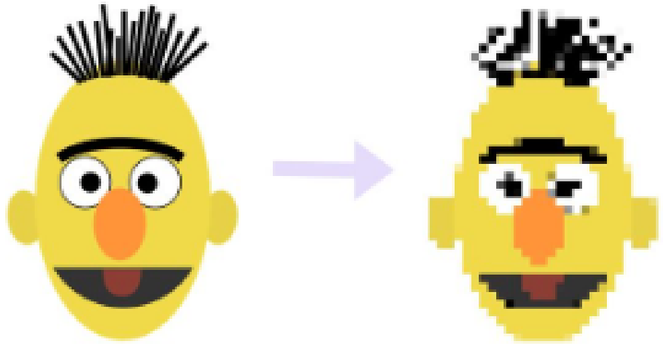
- iterative pruning을 적용한 모델은 약 90%의 웨이트를 날렸음에도 성능차이가 거의 없음

- pruning이 적용된 적은 수의 파라미터로 L1이나 L2 norm을 이용한 regularization(파라미터값 낮추기)을 수행하므로 좋은 성능을 유지할 수 있음

NIPS 2015, Learning both Weights and Connections for Efficient Neural Networks
https://arxiv.org/abs/1506.02626

Quantization(양자화)

Google Developer Student Clubs
Dongguk University



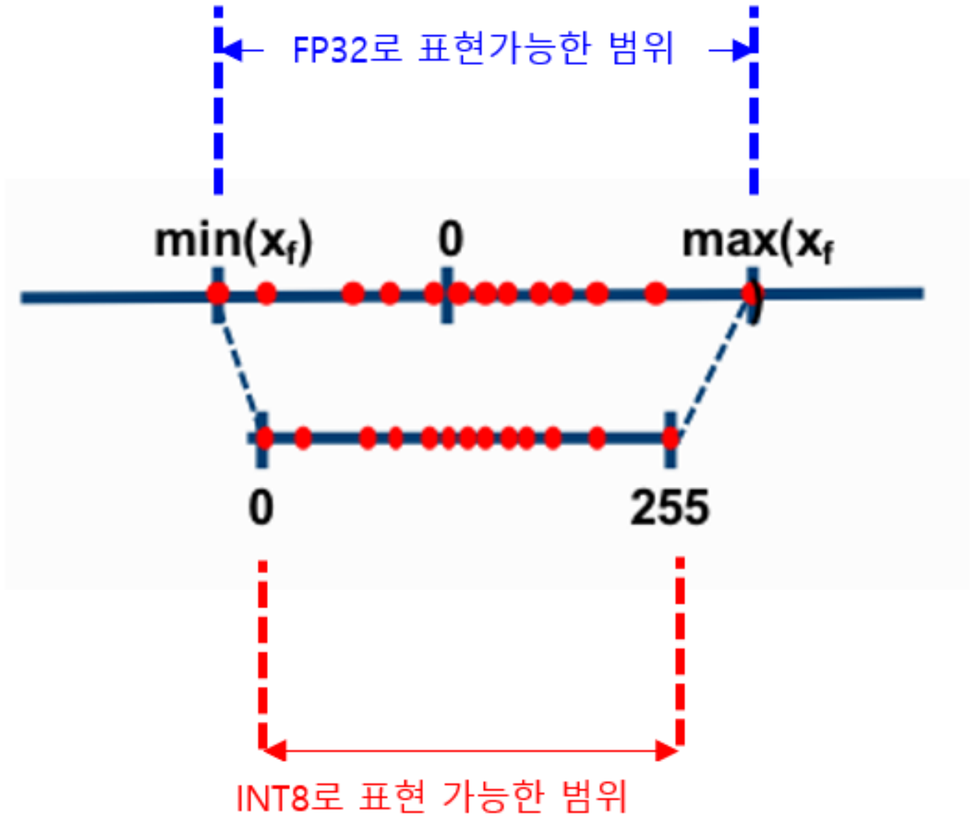
▲ Quantization

개요

학습된 딥러닝 모델이 weight값을 저장할 때 사용하는 비트의 수를 줄여서 모델 크기를 줄이는 방법

ICLR 2015, Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding
https://arxiv.org/abs/1510.00149

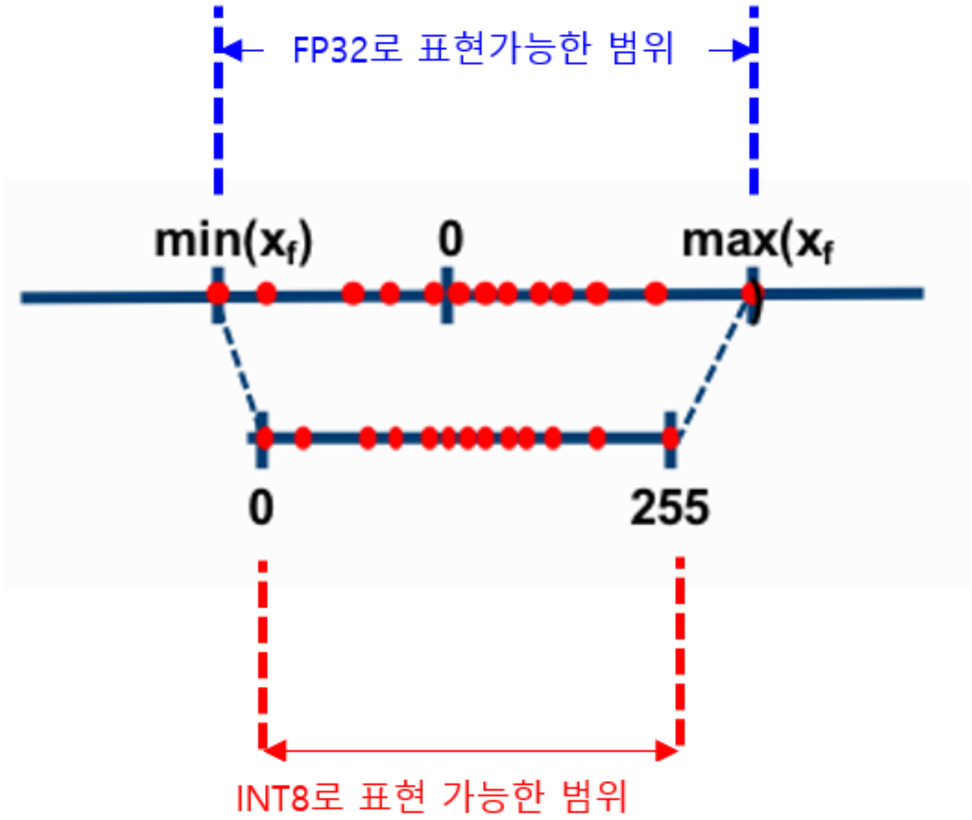
Quantization



- 딥러닝에서는 숫자를 저장하고 연산할 때 주로 32개의 비트를 사용하는 32-bit floting point(or FP32)를 사용

- 만약 weight의 값이 더 적은 수의 비트로 표현해도 값의 차이가 작다면 더 적은 비트를 사용하는 편이 적은 메모리를 사용할 것임

Quantization



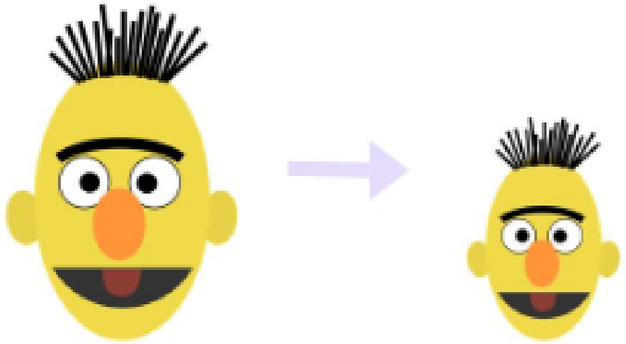
- weight값을 저장할 때 FP16 또는 INT8로 표현 가능한 범위의 숫자로 변환한 뒤 해당 비트 수만큼의 메모리에 저장하는 방법

- 더 적은 비트를 사용하기 때문에 모델의 메모리 사용량이 줄어드며, 모델을 사용해 추론할 때 동작시간을 단축할 수 있음

NVIDA DEVLEOPER, Technical Blog,
https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/

Distillation(Knowledge Distillation)

Google Developer Student Clubs
Dongguk University



▲ Distillation

개요

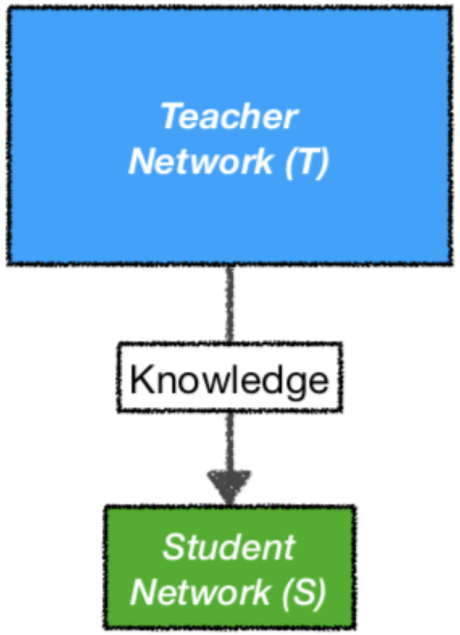
거대한 모델로부터 얻은 지식을 더 작은 모델에 전달하는 과정으로, 작은 모델이 큰 모델과 유사한 성능을 내면서도 훨씬 적은 자원으로 동작할 수 있게 됨

어떤 모델을 사용하는 것이 적합할까?

- 복잡한 모델 T : 예측 정확도 99% + 예측 소요 시간 3시간
- 단순한 모델 S : 예측 정확도 90% + 예측 소요 시간 3분

Introduction

Knowledge Distillation



1. **Teacher Network (T)**
   - **cumbersome model**
     ex) ensemble / a large generalized model
   - (pros) excellent performance
   - (cons) computationally expansive
   - can not be deployed when limited environments

2. **Student Network (S)**
   - **small model**
   - suitable for deployment
   - (pros) fast inference
   - (cons) lower performance than T

NIPS 2015, Distilling the Knowledge in a Neural Network
https://arxiv.org/abs/1503.02531

Distillation

Google Developer Student Clubs
Dongguk University

1) Soft Label

cow	dog	cat	car	
0	1	0	0	original hard targets
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble
.05	.3	.2	.005	softened output of ensemble

Softened outputs reveal the dark knowledge in the ensemble.

- 신경망에서 이미지 분류 작업을 할 때, softmax 레이어는 각 클래스에 대한 확률을 출력
- 예측 클래스 외 다른 클래스의 확률도 중요한 정보를 제공함. 하지만, 이러한 값들은 softmax에 의해 너무 작아 모델에 반영하기 쉽지 않을 수 있음
- 이를 위해 출력값의 분포를 좀 더 soft하게 만들어 이 값들을 이용

Distillation

Google Developer Student Clubs
Dongguk University

2) distillation loss

- 앞에서 정의한 soft target은 결국 큰 모델(T)의 지식을 의미
- 큰 모델(T)을 학습을 시킨 후 작은 모델(S)을 손실함수를 통해 학습시킴

Benefits

- 전력 소모 및 비용 절감
- 직관적이며 안정적인 Output

Risks

- 일반화의 어려움
- 아직 성숙하지 않은 분야

참고문헌

Google Developer Student Clubs
Dongguk University

https://rasa.com/blog/compressing-bert-for-faster-prediction-2/#motivation

https://velog.io/@qtly_u/%EB%AA%A8%EB%8D%B8-%EA%B2%BD%EB%9F%89%ED%99%94-%EA%B8%B0%EB%B2%95-Knowledge-Distillation

https://baeseongsu.github.io/posts/knowledge-distillation/

https://tilnote.io/pages/6480a73ee92fe5ef635f4d77

https://tech.scatterlab.co.kr/ml-model-optimize/

https://www.youtube.com/watch?v=NVNCPGWe5Ss

https://velog.io/@woojinn8/LightWeight-Deep-Learning-0.-%EB%94%A5%EB%9F%AC%EB%8B%9D-%EB%AA%A8%EB%8D%B8-%EA%B2%BD%EB%9F%89%ED%99%94

감사합니다