-NLP Study-

경량화

발표자: 박무재

AI명예학회

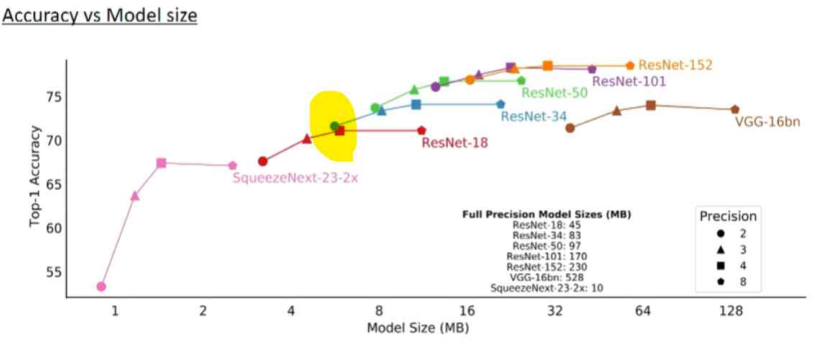
SKHU

목차

- 경량화
- 모델 구조 변경 경량화(CV)
- 양자화
- 가지치기
- 지식 증류
- 코드(QLoRA + 4bit quantization + LDCC-SOLAR-10.7B)

경량화

모델 구조를 바꾸거나 기 학습된 모델을 압축 또는 증류하는 기법



[Learned Step Size quantization, ICLR 2020]

경량화

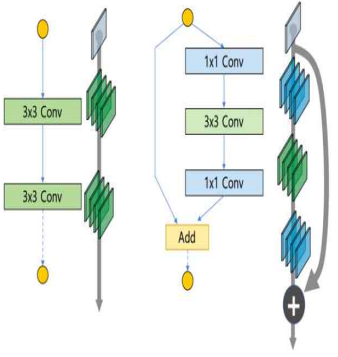
알고리즘을 경량화 VS 기 학습된 모델을 경량화

〈표 1〉 경량 딥러닝(Lightweight Deep Learning) 연구 동향

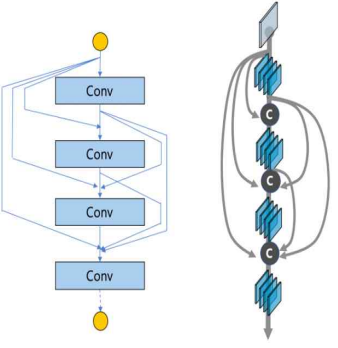
	접근방법	연구 방향
경량 알고리즘 연구	모델 구조 변경	잔여 블록, 병목 구조, 밀집 블록 등 다양한 신규 계층 구조를 이용하여 파라미터 축소 및 모델 성능을 개선하는 연구(ResNet, DenseNet, SqueezeNet)
	합성곱 필터 변경	합성곱 신경망의 가장 큰 계산량을 요구하는 합성곱 필터의 연산을 효율적으로 줄이는 연구(MobileNet, ShuffleNet)
	자동 모델 탐색	특정 요소(지연시간, 에너지 소모 등)가 주어진 경우, 강화 학습을 통해 최적 모델을 자동 탐색하는 연구(NetAdapt, MNasNet)
알고리즘 경량화 연구	모델 압축	가중치 가지치기, 양자화/이진화, 가중치 공유 기법을 통해 파라미터의 불필요한 표현력을 줄이는 연구(Deep Compression, XNOR-Net)
	지식 증류	학습된 기본 모델을 통해 새로운 모델의 생성 시 파라미터값을 활용하여 학습시간을 줄이는 연구(Knowledge Distillation, Transfer Learning)
	하드웨어 가속화	모바일 기기를 중심으로 뉴럴 프로세싱 유닛(NPU)을 통해 추론 속도를 향상시키는 연구
	모델 압축 자동 탐색	알고리즘 경량화 연구 중 일반적인 모델 압축 기법을 적용한 강화 학습 기반의 최적 모델 자동 탐색 연구(PocketFlow, AMC)

모델 구조 변경 경량화

Resnet

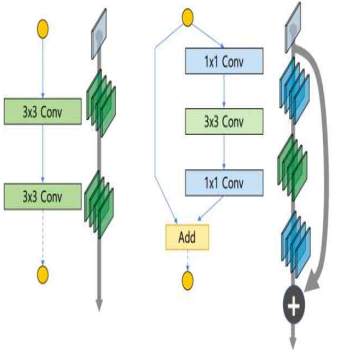


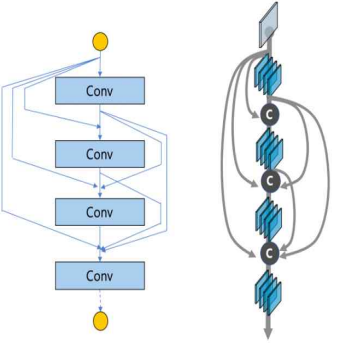
Densenet



모델 구조 변경 경량화

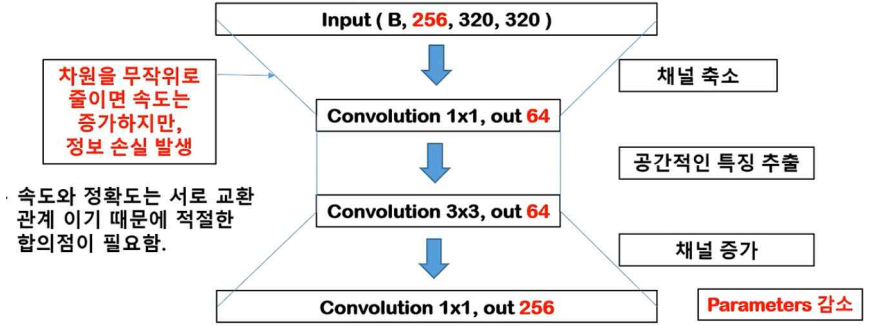
1x1 Conv(병목레이어)가 뭔 소용?





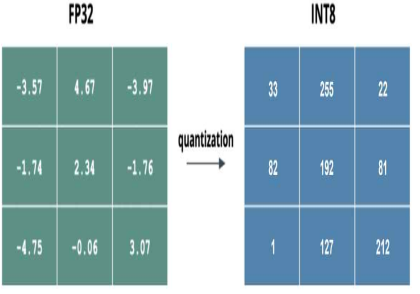
모델 구조 변경 경량화

그냥 3x3conv하는 것보다 사이즈를 줄여서 3x3 conv 후 다시 채널 사이즈를 키우는 방식이 연산량이 적음



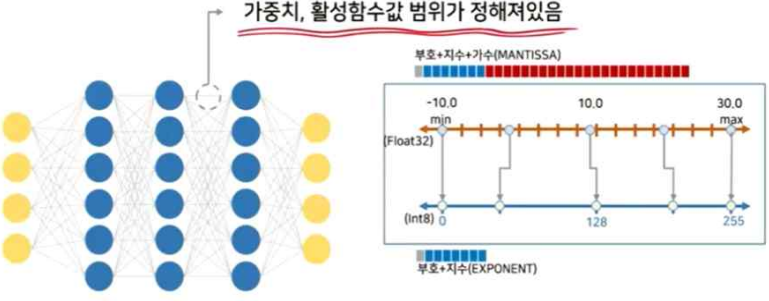
양자화(Quantization)

양자화: 대부분 딥러닝 모델의 가중치들은 float32인데 이것을 int8로 변환



양자화(Quantization)

값이 어느 정도 범위 안에 있다는 것을 가정



Disk사용
75%절감

메모리 사용량
25%만 필요

처리속도
2.3배 향상

양자화(Quantization)

값이 어느 정도 범위 안에 있다는 것을 가정

$$f_q(x, s, z) = \text{Clip}(\text{round}(\frac{x}{s}) + z)$$

$f_q(x, s, z)$: quantized value

clip() : clip the values in a range (ex. 0 ~ 255)

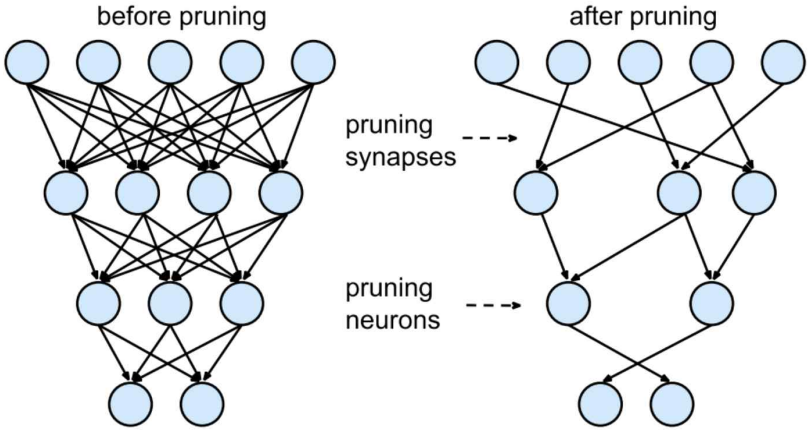
x : real value (float32)

s : scale

z : zero-point integer

가지치기

dropout x



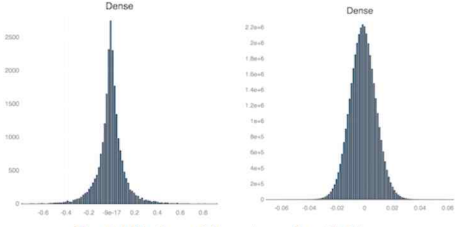
가지치기

Magnitude Pruner

$$thresh(w_i) = \begin{Bmatrix} w_i : if|w_i| > \lambda \\ 0 : if|w_i| \leq \lambda \end{Bmatrix}$$

weight값이 기준값 이하 라면 0으로 만들고, 기준값 보다 크다면 그대로 두는 것

Sensitivity Pruner

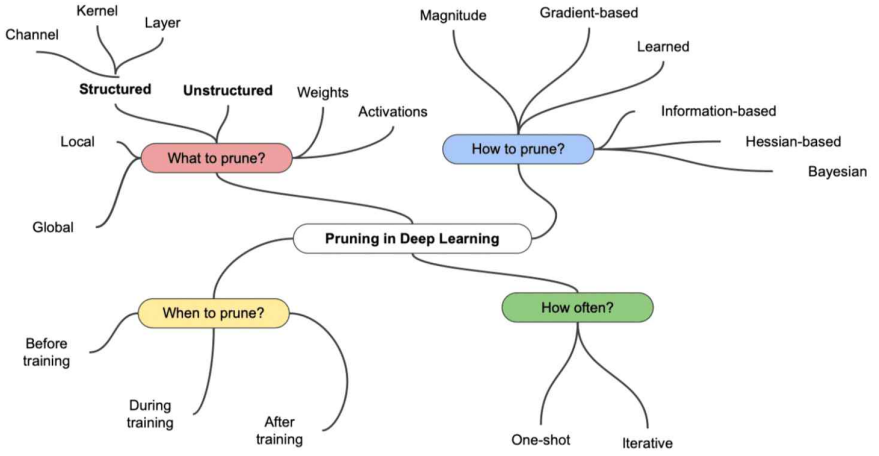


The distributions of Alexnet conv1 and fc1 layers

Sensitivity Pruning 방법은 Convolution layer 그리고 Fully connected layer가 가우시안 분포를 갖고있다는 것을 활용

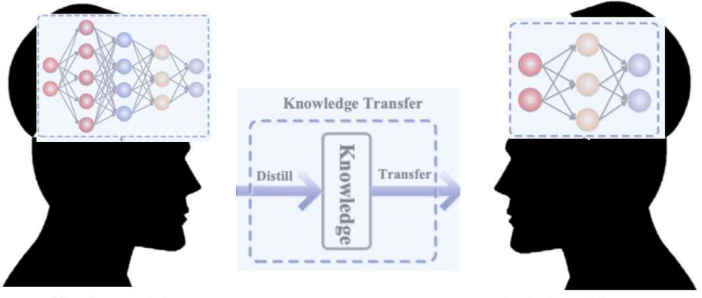
$\lambda = s * \sigma_l$ where σ_l is std of layer l as measured on the dense model

가지치기


Kernel
Channel
Layer
Structured
Unstructured
Weights
Activations
Local
What to prune?
Global
Magnitude
Gradient-based
Learned
Information-based
How to prune?
Hessian-based
Bayesian
Pruning in Deep Learning
When to prune?
Before training
During training
After training
How often?
One-shot
Iterative

지식증류

지식 증류기술은 (앙상블 기법을 통해 학습된) 큰 네트워크(들)로부터 작은 하나의 네트워크에 지식을 전달하는 방법론



지식증류

cow	dog	cat	car	
0	1	0	0	original hard targets

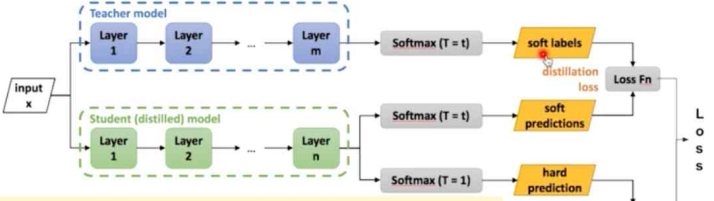
cow	dog	cat	car	
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble

cow	dog	cat	car	
.05	.3	.2	.005	softened output of ensemble

Softened outputs reveal the dark knowledge in the ensemble.

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

지식증류



$$\mathcal{L}(x;W) = \alpha * \mathcal{H}(y, \sigma(z_s; T=1)) + \beta * \mathcal{H}(\sigma(z_t; T=\tau), \sigma(z_s, T=\tau))$$

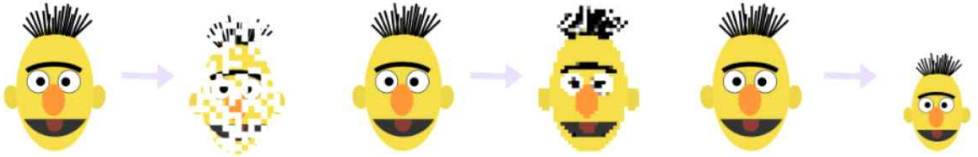
$$p_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})}$$

x	input	α	coefficient
W	student model parameters	β	coefficient
y	ground truth label	z_s	logit of the student
$\mathcal{H}$	cross-entropy loss function	z_t	logit of the teacher
σ	softmax function		
T	temperature		

Let α be much smaller than β (Hinton et al., 2015)

경량화

Pruning vs Quantization vs Distillation



▲ Pruning ▲ Quantization ▲ Distillation

QRoLA

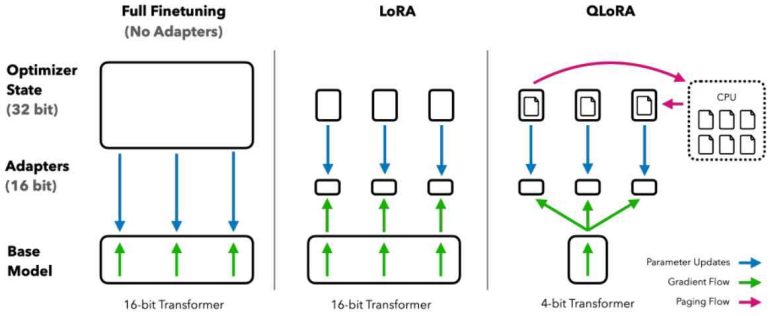


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

4-bit로 quantize하는 high-precision technique
-> GPU 하나로 fine-tuning 가능

reference

- https://blogik.netlify.app/BoostCamp/U_stage/45_pruning/
- https://gaussian37.github.io/dl-concept-quantization/
- https://ettrends.etri.re.kr/ettrends/176/0905176005/34-2_40-50.pdf
- https://intellabs.github.io/distiller/pruning.html#han-et-al-2015
- https://baeseongsu.github.io/posts/knowledge-distillation/
- https://sofar-sogood.tistory.com/entry/QLoRA-%EB%A6%AC%EB%B7%B0-Qlora-Efficient-finetuning-of-quantized-llms