Global Weight:
심층 신경망의 압축을 위한
네트워크 수준의 가중치 공유

2020. 07. 13

신은섭 배성호

Machine Learning and Visual Computing Lab.

Kyung Hee University

INDEX

- 배경 및 동기
- 방법
- 실험 결과
- 결론 및 향후 연구

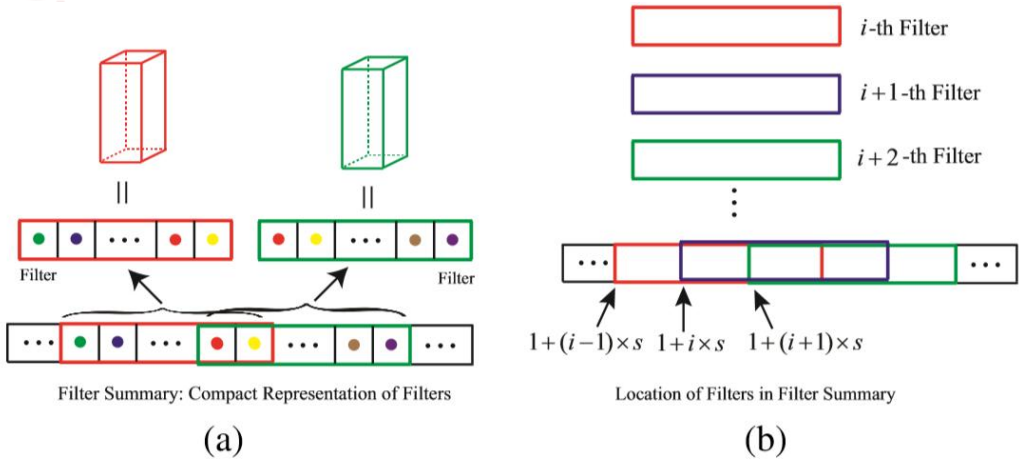
배경 및 동기

배경

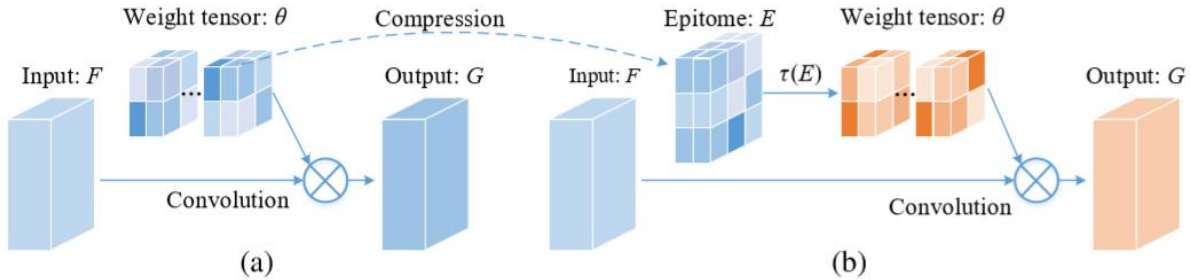
모델	Top 5 on Image Net	매개변수	FLOPs
AlexNet	84.70%	62M	1.5B
VGGNet	92.30%	138M	19.6B
Inception	93.30%	6.4M	2B
ResNet-152	95.51%	60.3M	11B

- 심층신경망은 Computer Vision분야에서 매우 좋은 성능을 보이고 있음
- 그러나 사용하는 리소스가 매우 많기 때문에 하드웨어 제한이 있는 장치에서는 실행하기 어려움
- 이를 해결하기 위해 모델 압축이라는 분야가 활발하게 연구 중

배경



FSNet[2]



NES[3]

- 그 중 가중치 공유기법은 학습된 가중치를 여러 곳에서 공유하는 기법
- 그러나 기존의 선행연구들[1, 2, 3]은 **레어어 단에서만 가중치를 공유**

[1] WSNet: Compact and Efficient Networks with Weight Sampling. ICLR Workshop, 2018.
[2] FSNet: Compression of Deep Convolutional Neural Networks by Filter Summary, ICLR, 2
[3] Neural Epitome Search for Architecture-Agnostic Network Compression, ICLR, 2020.

동기

- 그러나 **Weight 중복**은 레이어 내부에만 있는 것이 아니라 **네트워크 전체**에 있음
- 본 논문에서는 레이어 내부에서만 가중치를 공유하는 것이 아닌 **전체 네트워크에서 가중치를 공유**하는 **Global Weight**방법을 제안

Global Weight

- Global Weight
- Global Weight Convolution
- Global Weight Networks

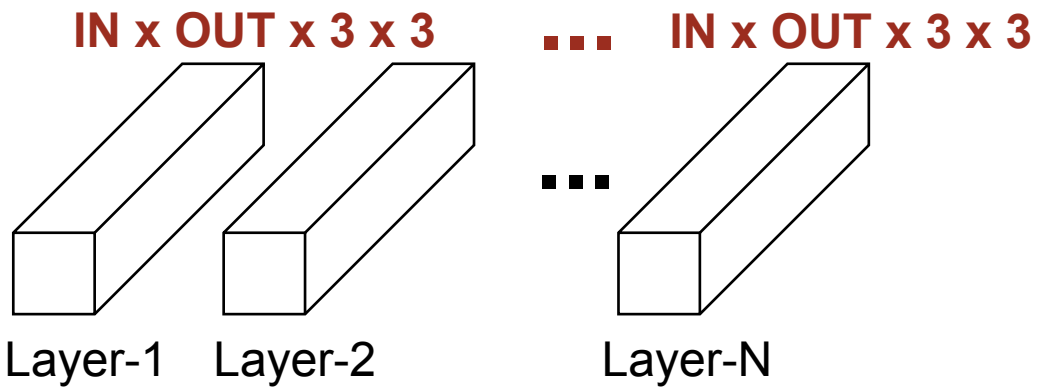
Global Weight

- **Global Weight**는 기존의 Layer-wise 가중치 공유 방법에서 벗어나, **전체 네트워크**에서 **하나의 가중치 셋을 공유**하는 패러다임.

- 장점
  - 가중치의 **중복성**을 획기적으로 줄일 수 있음
  - 매개변수도 효율적으로 **압축**할 수 있음

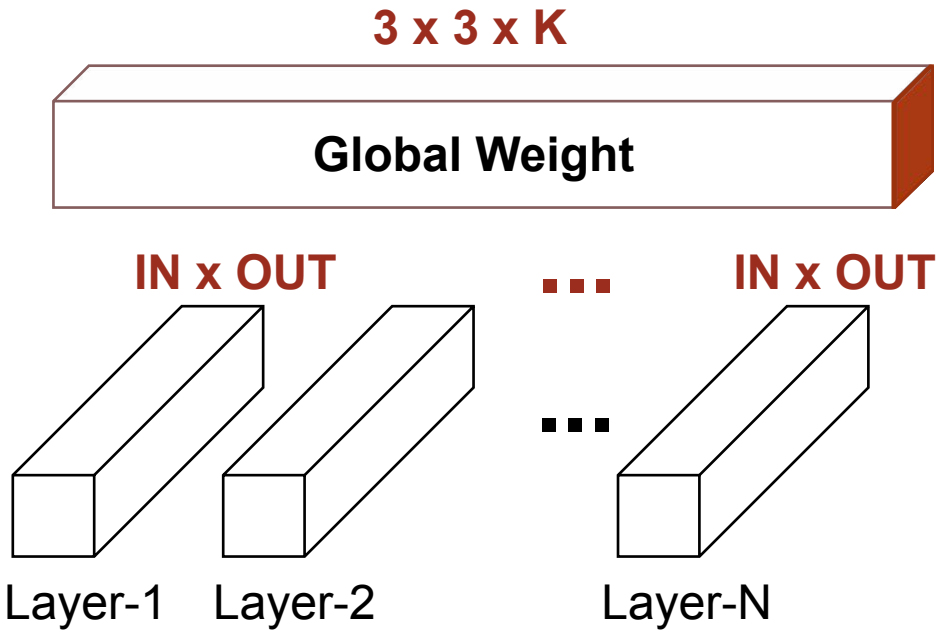
Global Weight Convolution

- Global weight 패러다임을 적용한 convolution연산



가중치 공유를 하지 않는 방식

각 레이어마다 필터가 독립적으로 존재

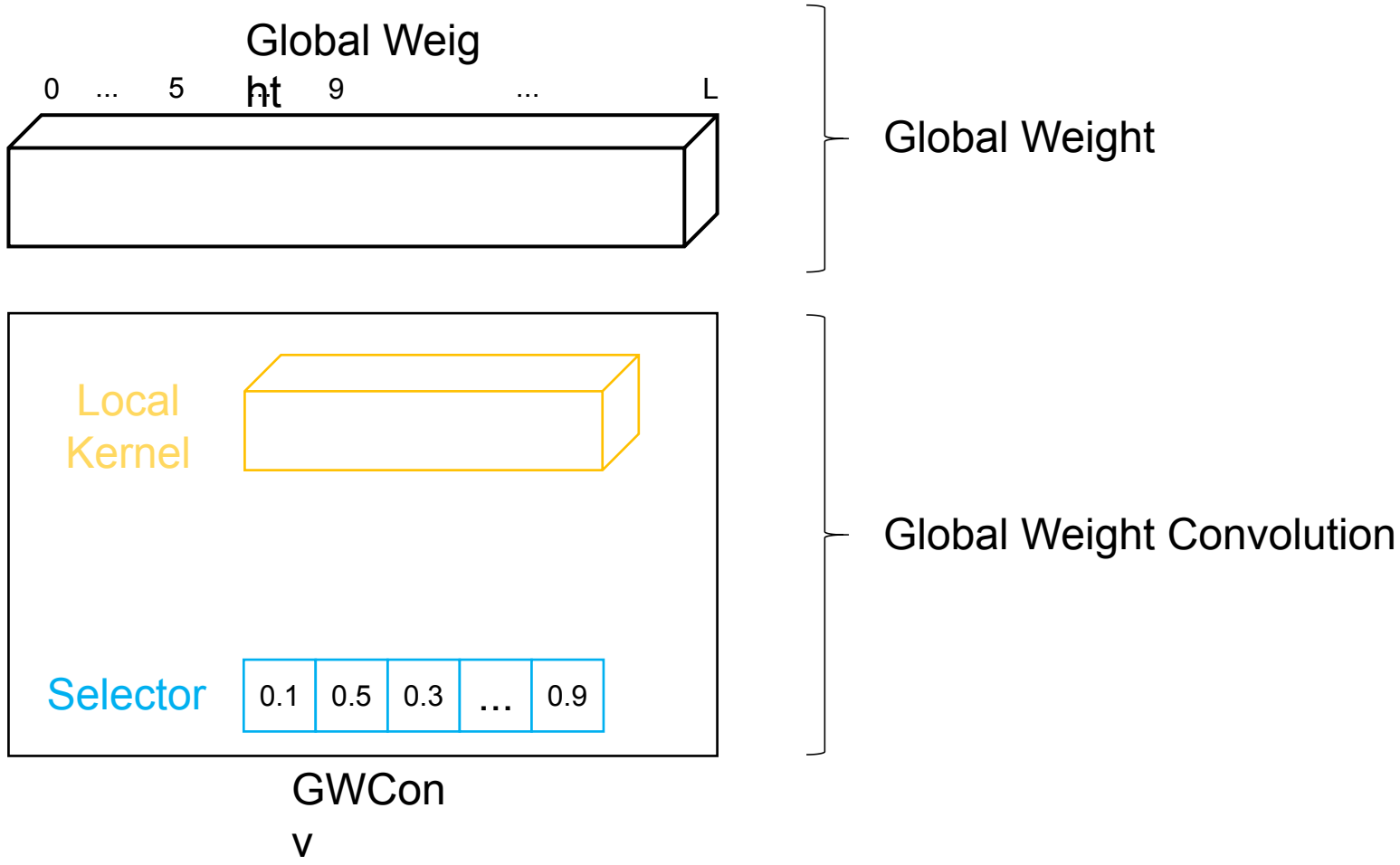


Global Weight를 적용한 방식

각 레이어는 필터의 인덱스만 저장하고
Global Weight에 저장된 실제 필터를 사용

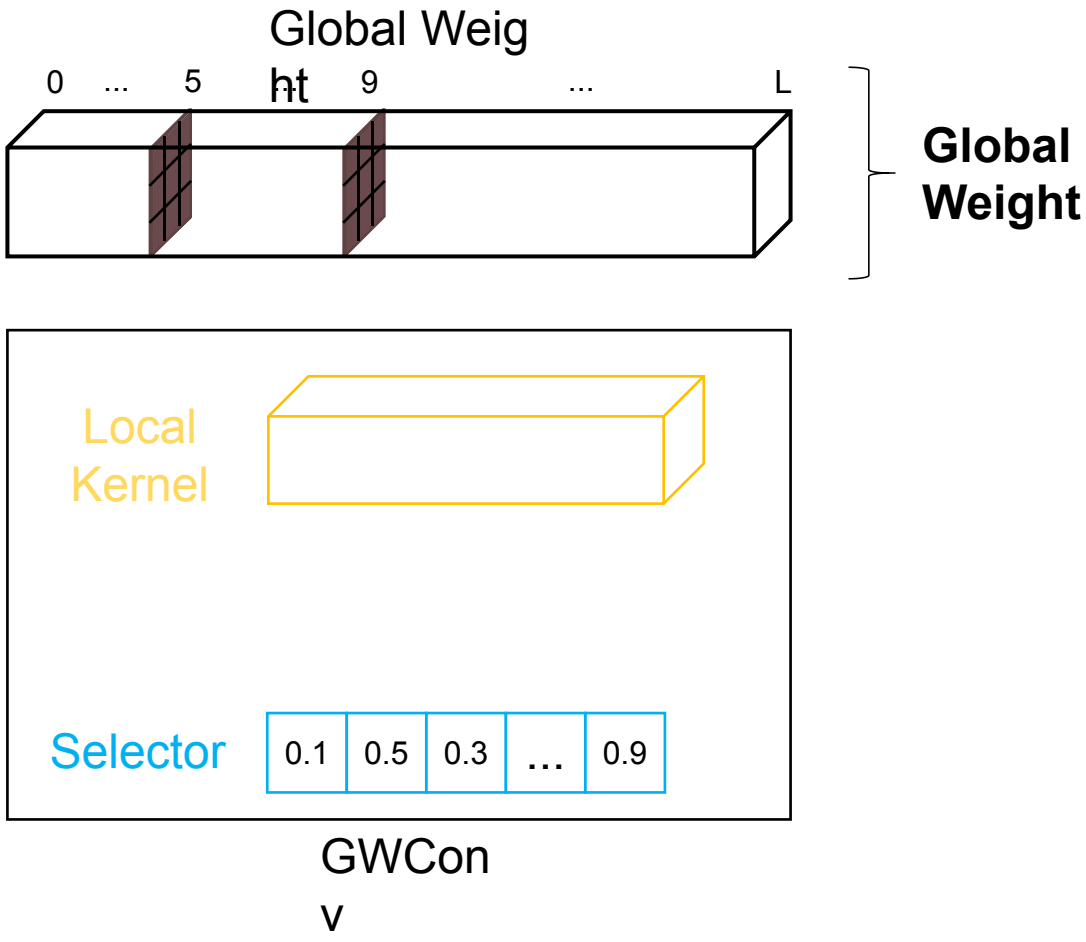
Global Weight Convolution

- GWConv의 작동 원리



Global Weight Convolution

- GWConv의 작동 원리



Global Weight

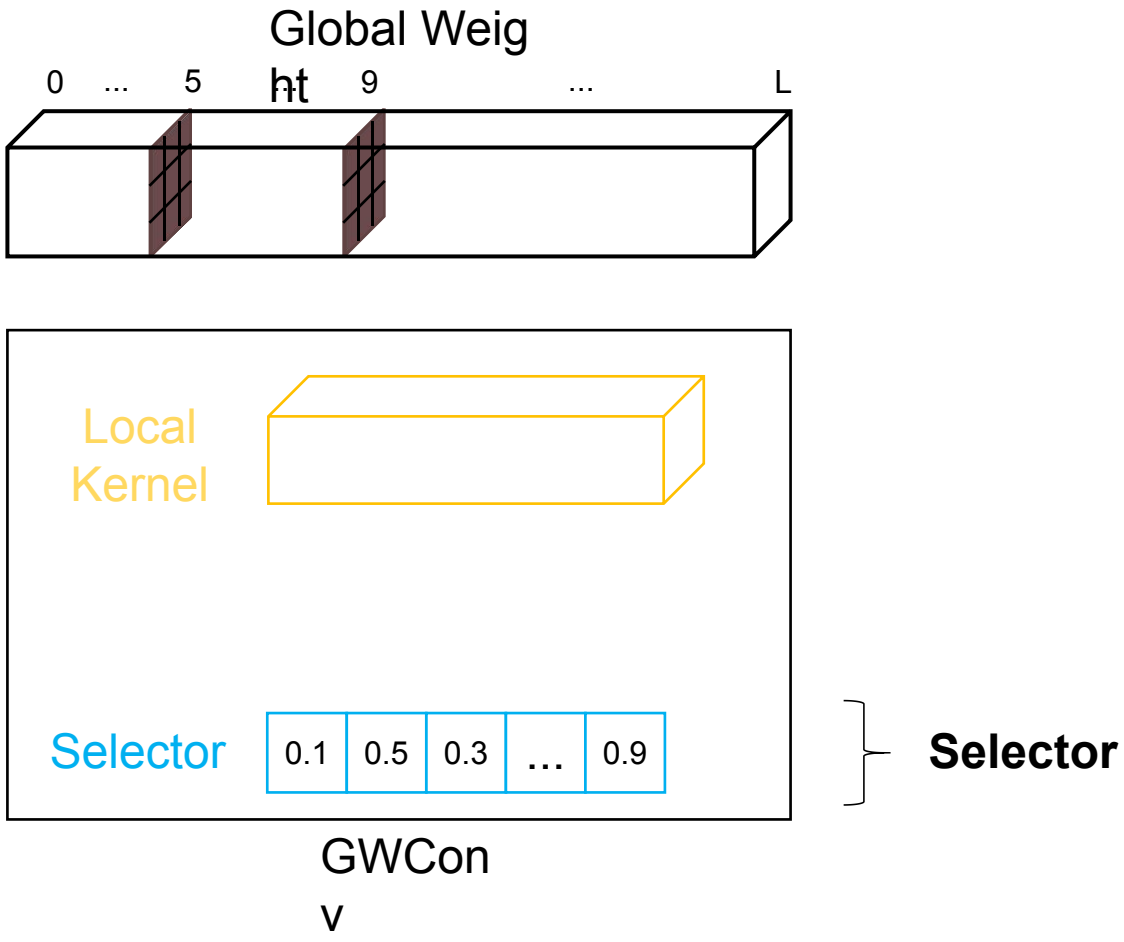
실제 Convolution의 **Filter가 학습** 되는 부분
전체 네트워크에서 **공유**됨

크기: K x K x L

(K: 커널 크기, L: GW 크기)

Global Weight Convolution

- GWConv의 작동 원리



Selector

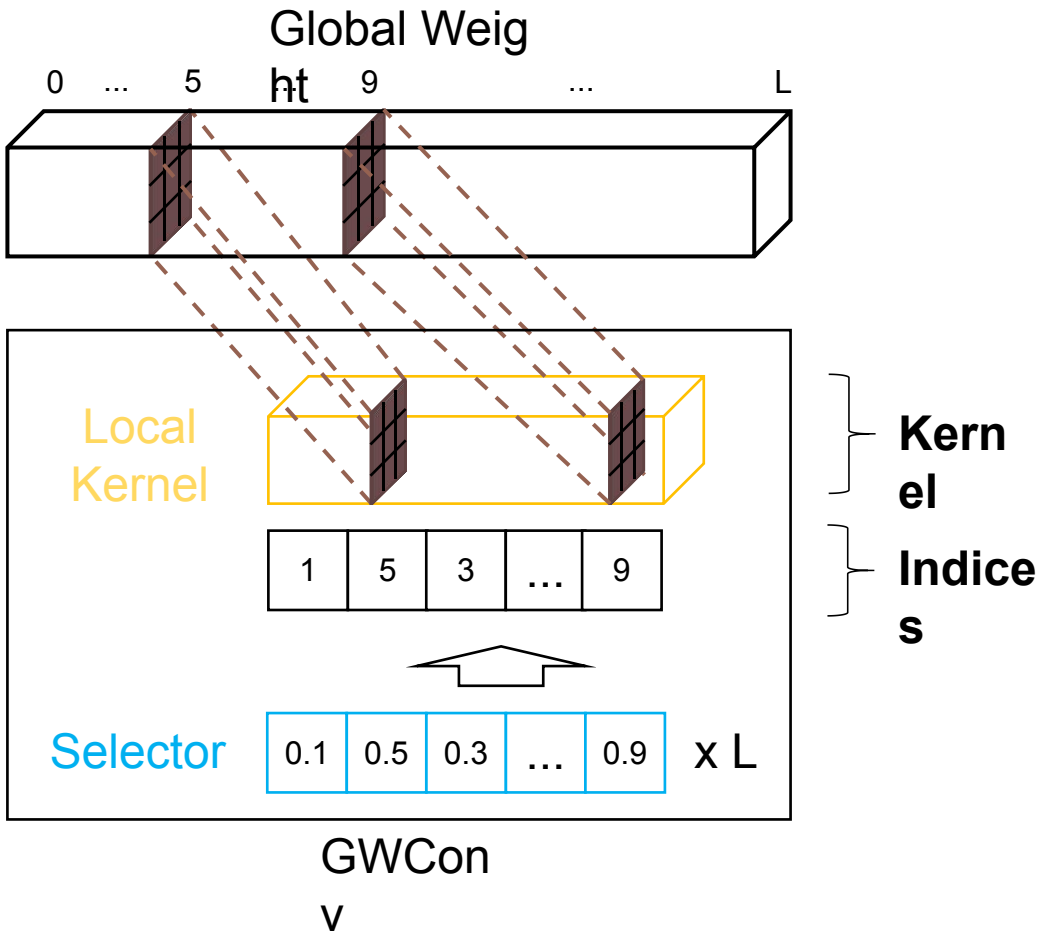
Global Weight의 **몇 번 째 Weight를 사용할지**를 학습
GWConv에서 **학습되는 유일한 부분**

Selector의 크기: in_channel x out_channel

Selector의 값: 0 ~ 1

Global Weight Convolution

- GWConv의 작동 원리



Local Kernel

Forward시 생성되는 **중간물**
Selector와 Global Kernel의 조합으로 구성
LK를 이용하여 실제 Convolution 수행

구성 방법:

1. Selector에 GW 크기 L을 곱하고 반올림하여 Indices 계산
2. 해당 Indices위치에 있는 GW실제 Weight를 이용하여 Local Kernel을 구성

Global Weight Convolution

- 매개변수 크기

Baseline	GWConv
$\mathrm{K}\sum_{\mathrm{n}=0}^{\mathrm{N}} I_n O_n$	$KL + \sum_{\mathrm{n}=0}^{\mathrm{N}} I_n O_n$

Notation

N: Number of Layer
K: kernel size
L: Number of Global Kernel
$\mathrm{I_n}$: number of input channel
$\mathrm{O_n}$: number of input channel

Global Weight Convolution

- 압축비

$$Let) \quad \sum_{n=0}^{N} I_n O_n = H$$

$$CR = \frac{KH}{KL + H}$$

$$= \frac{1}{\frac{L}{H} + \frac{1}{K}}$$

$$\approx \frac{H}{L}$$

압축비 CR은 왼쪽과 같이 구해 짐

K는 일반적으로 1, 9, 25와 같이 L과 H에 비해 매우 작은 값임으로 무시가 가능

즉, GWConv는 L을 조절 함으로써 압축비를 조절

Notation

N: Number of Layer
K: kernel size
L: Number of Global Kernel
I_n: number of input channel
O_n: number of input channel

Global Weight Networks

- GWConv는 추가적인 연결이나, 학습에서만 사용되는 매개변수가 없기 때문에 **기존의 Convolution연산을 바로 대체하여 사용이 가능**

- 기존의 Convolution연산을 GWConv로 대체한 네트워크를 **GWNet** 이라 함

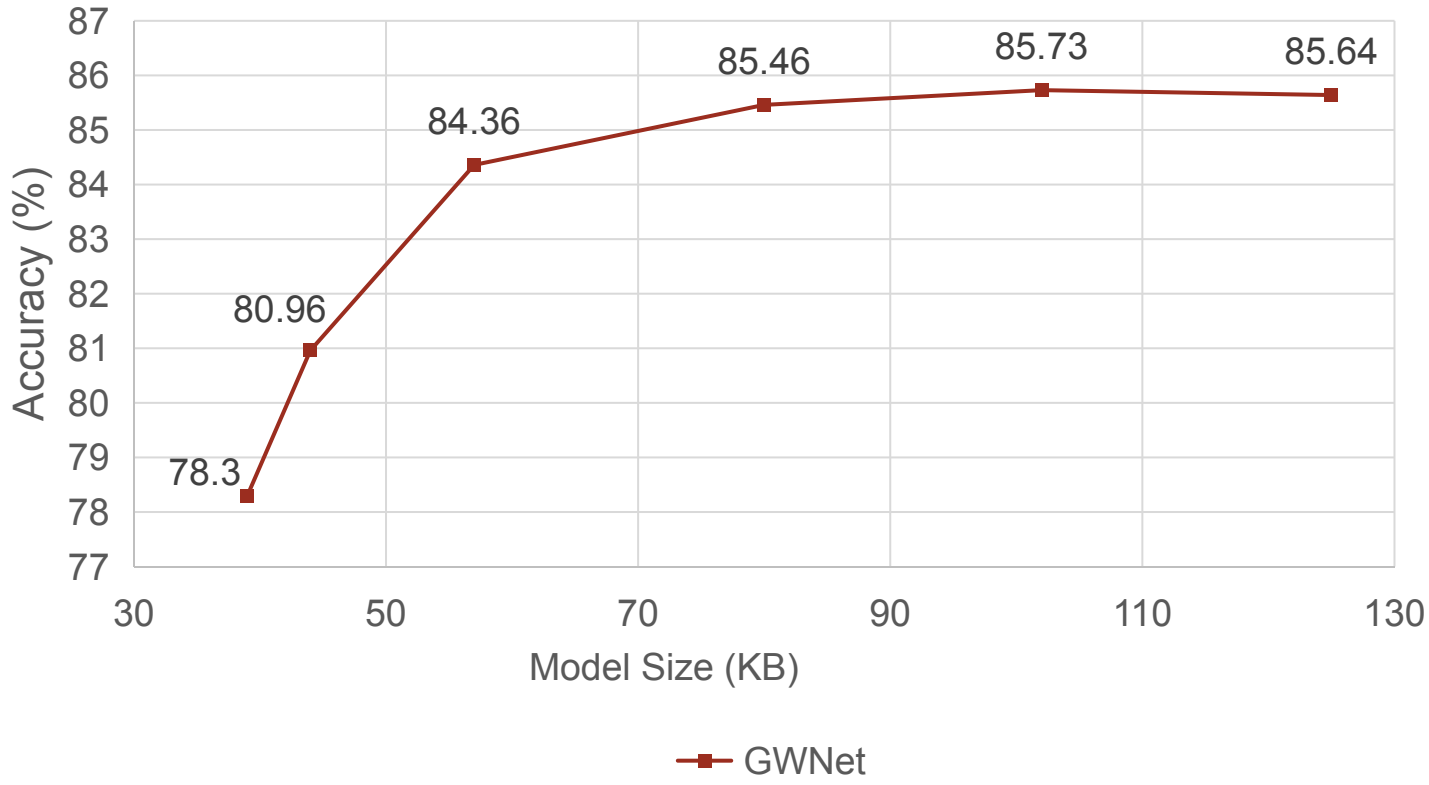
실험 결과

실험 결과

Model	# Params	Top-1	CR
ResNet-20	272,474	92.02%	1.00
GWNet-10000	124,810	85.64%	2.18
GWNet-7500	102,310	85.73%	2.66
GWNet-5000	79,810	85.46%	3.41
GWNet-2500	57,310	84.36%	4.75
GWNet-1000	43,810	80.96%	6.22
GWNet-500	39,310	78.30%	6.93

CIFAR10에서 GW의 크기를 변화해가며 실험

실험 결과



CIFAR10에서 GW의 크기를 변화해가며 실험

결론 및 향후 연구

결론 및 향후 연구

- 심층신경망의 가중치를 네트워크 전체에서 공유하는 방법 **Global Weight** 제시
- Global Weight로 **가중치를 공유하며 압축이 가능**함을 보임

- 압축이 됨에 따라 정확도가 약 2배 압축에서 약 **6% 하락**함
- 겹침 패턴 가중치 공유방식, 가중치 증강 방식을 추가로 도입하여 정확도는 올리고 압축율은 높이는 방식으로 연구 진행 예정

Thank you

Any Question ?

Machine Learning and Visual Computing Lab.

Kyung Hee University