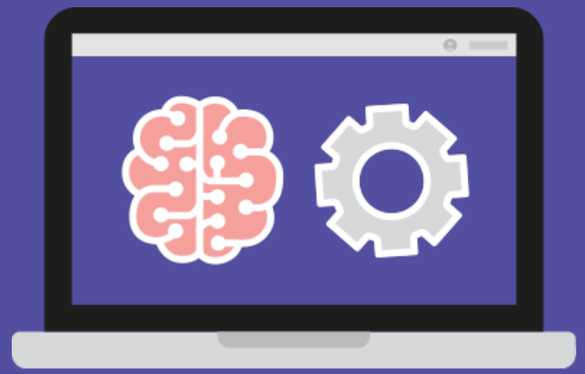
/* elice */

양재 AI School 인공지능 캠프

Lecture 18

Batch Normalization and Transfer Learning



박상수 선생님

커리큘럼

1

배치 정규화 (Batch Normalization)

학습의 성능을 개선하는 배치 정규화에 대해 파악하고
기본적인 이론과 방법에 대해 학습합니다.

2

전이 학습 (Transfer Learning)

전이학습의 개념을 파악하고
전이학습의 구현 방법에 대해 학습합니다.

목차

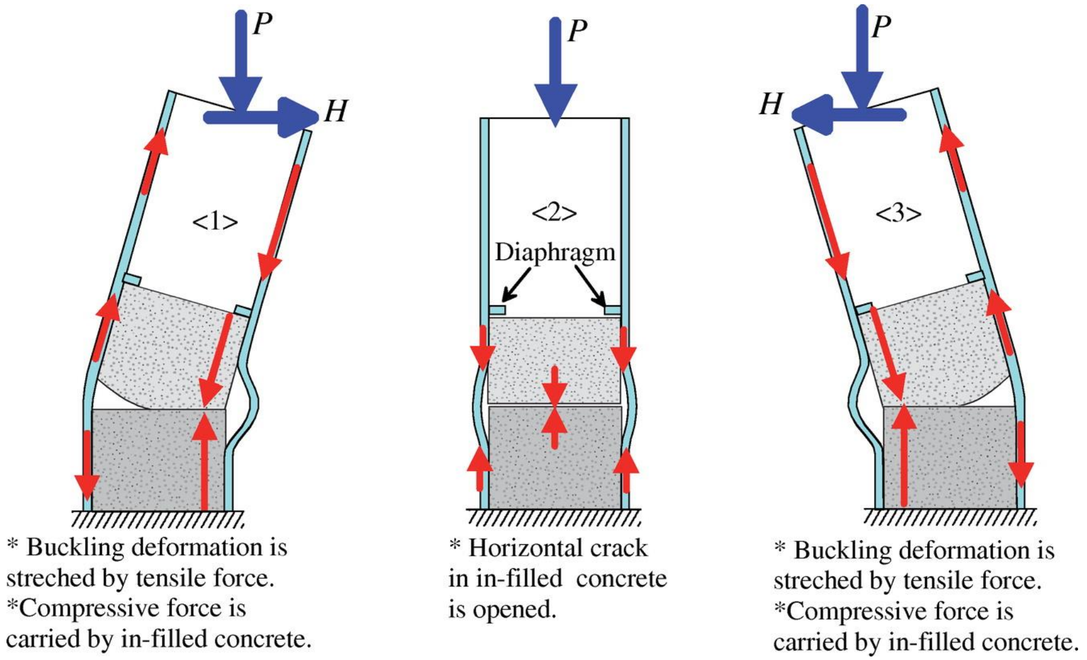
배치 정규화

딥러닝 모델에 필수적인 존재가 되는 것

Covariate Shift

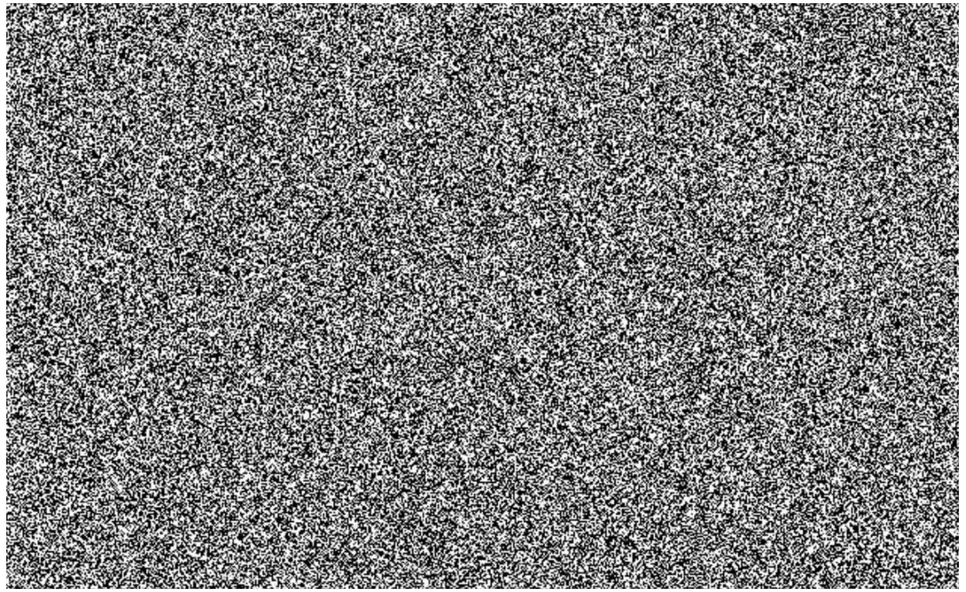
학습하는 과정에서 이전 Layer의 파라미터 변화로 인해
현재 Layer의 분포가 바뀌는 현상
학습과정에서는 분포가 비슷할수록 좋음

건축현상에서의 휘어짐 (Buckling)



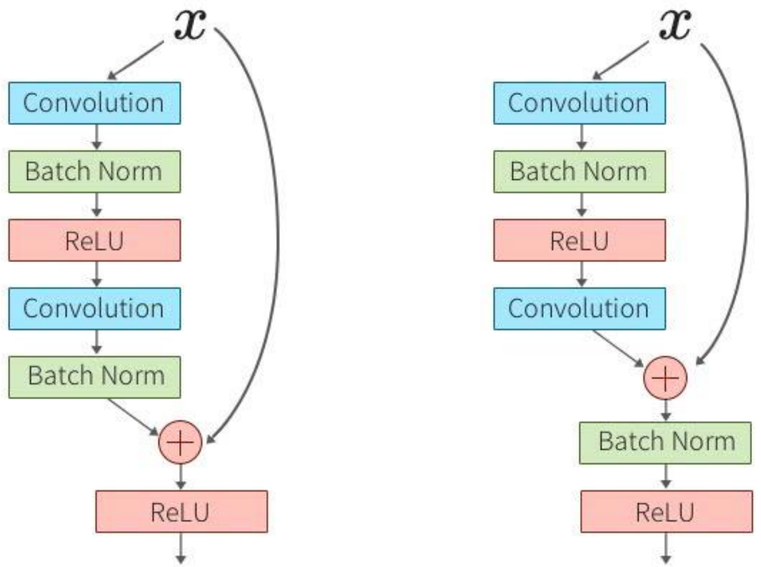
기둥이 휘어지는 것을 막으려면 휘어짐을 방지하는 수단이 필요
Batch Normalization을 통해 휘어짐 방지

Covariate Shift를 줄이기 위해서는



각 Layer로 들어가는 입력을 Whitening하는 방법
Whitening: 입력을 평균 0, 분산 1로 바꾸는 것
(일정한 스펙트럼, 패턴을 갖도록 하는 과정)
Backpropagation과 무관하게 Whitening을 한다면
특정 파라미터가 계속 커지는 형태로 진행 될 수 있음

Whitening과 Batch Norm의 차이점



평균과 분산을 조절하는 과정이 별도로 있는 것이 아니라
신경망 안에 포함되어 학습과정에서 평균과 분산을 조절

Batch Norm 방법

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

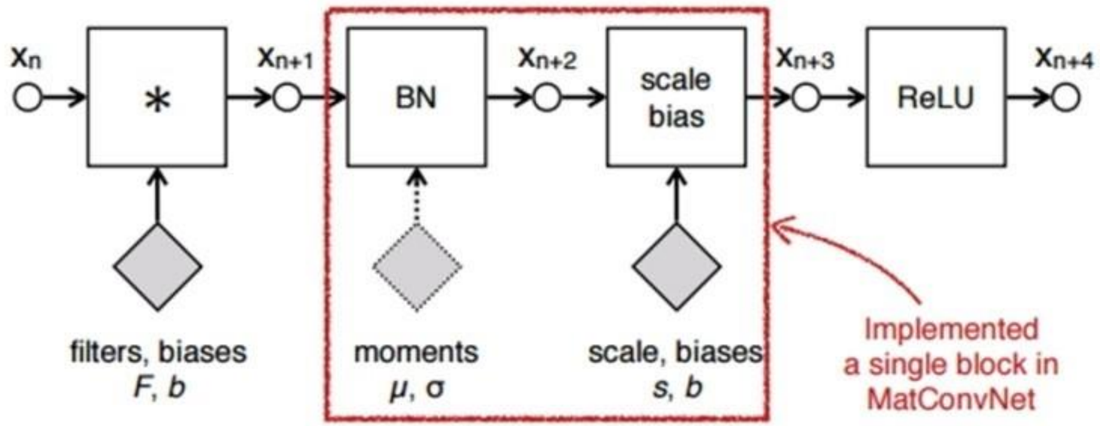
$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

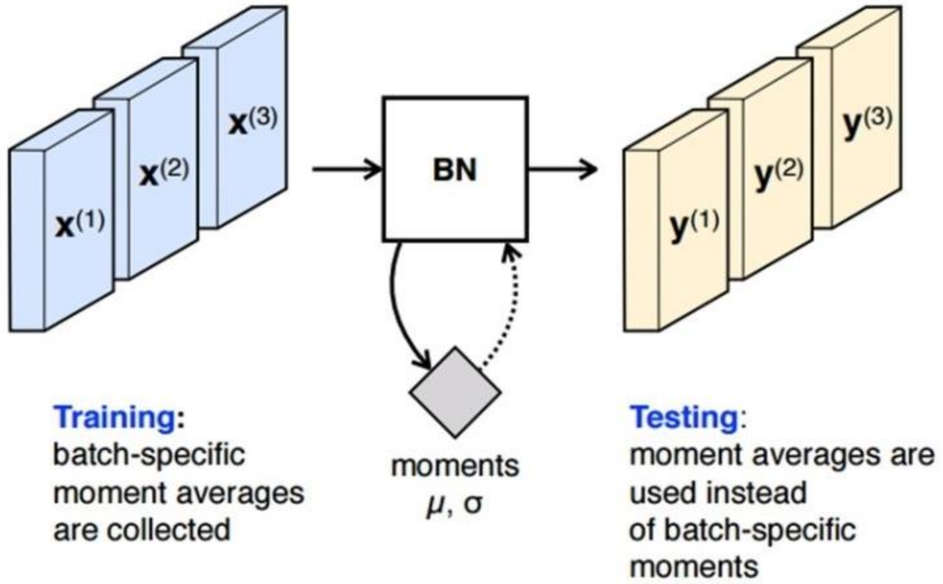
정규화는 학습 전체 과정에 대해서 적용하는 것이 최고이지만
SGB를 사용한다면, 파라미터의 업데이트가 Mini Batch 단위로 진행
따라서, Mini Batch 단위로 Batch Norm

Batch Norm 과정



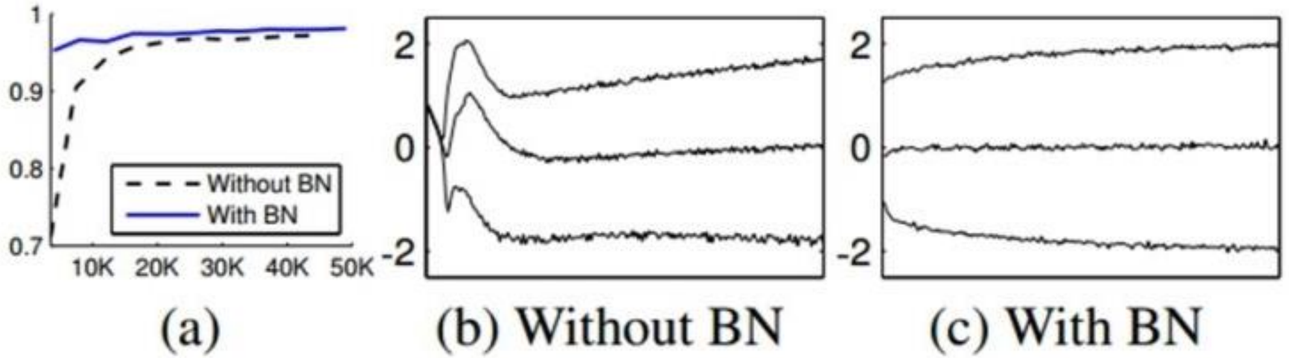
Batch Norm은 평균과 분산을 구한 후 정규화
정규화 이후에 Scale (α)과 Shift (β)연산을 진행

Batch Norm의 학습과 테스트 차이



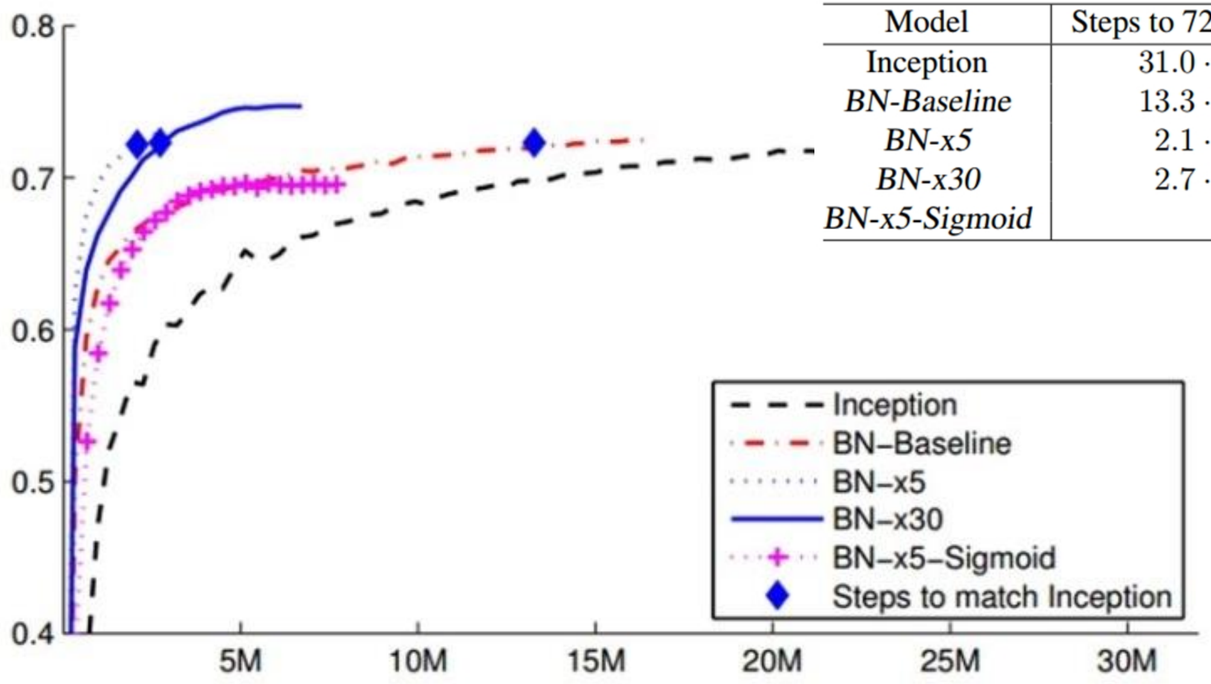
학습 과정에서는 Mini Batch 마다 Scale (α)과 Shift (β)를
구하고 그 값을 저장, 테스트 과정에서는 구한
Scale (α)과 Shift (β)의 평균을 사용

Batch Norm의 FC에 적용한다면 ?



Fully-connected로만 구성된 신경망에 적용하면
(MNIST, 28*28-100*10)
Batch Norm을 적용한 경우가 학습 속도, 결과에서 더 좋음

Batch Norm의 CNN에 적용한다면 ?



Model	Steps to 72.2%	Max accuracy
Inception	$31.0 \cdot 10^6$	72.2%
BN-Baseline	$13.3 \cdot 10^6$	72.7%
BN-x5	$2.1 \cdot 10^6$	73.0%
BN-x30	$2.7 \cdot 10^6$	74.8%
BN-x5-Sigmoid		69.8%

BN-Baseline (Nonlinearity 앞에 Batch Norm), BN-x*N* (학습 진도율 N)
BN-x5-Sigmoid (BN-x5와 동일하지만, ReLU 대신 Sigmoid 사용)

전이학습

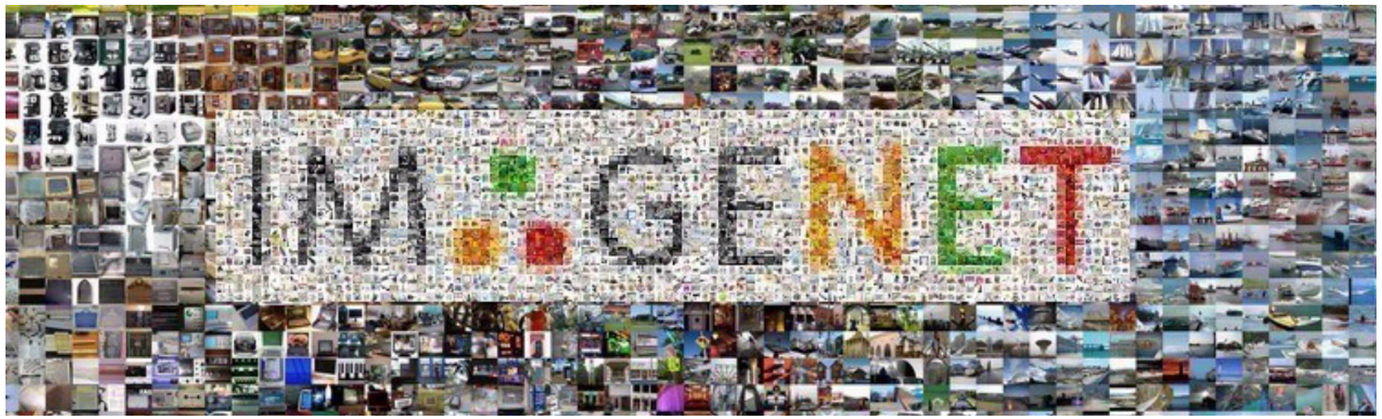
지식이 전이되는 새로운 학습 방법

전이학습 이란 ?

<사과> <인공지능> <배>

잘 훈련된 모델을 사용하여 유사한 문제를 해결하는 방법
(예: 사과를 깎는 방법을 배운 AI에게 배를 깎는 방법을 학습)
데이터가 부족한 분야에도 적용 가능

전이학습이 필요한 이유 #1



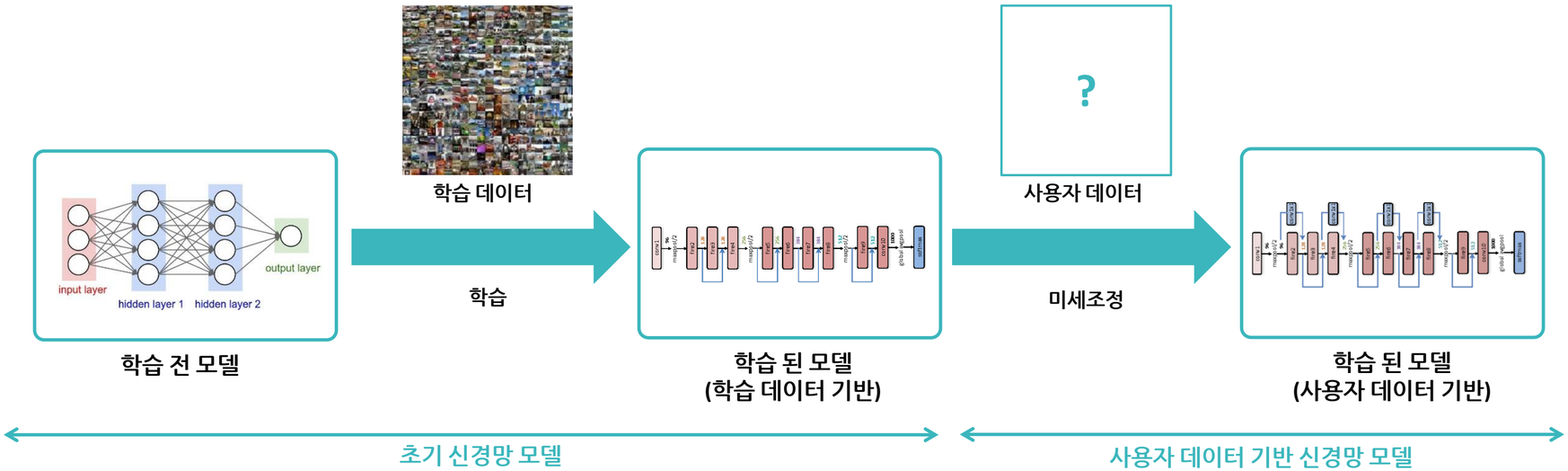
학습을 위해서는 많은 데이터가 필요
(예: 이미지넷 벤치마크는 150GB 필요)
학습데이터가 많아도 학습이 잘 안되는 경우가 많음

전이학습이 필요한 이유 #2

GPU 종류	GPU 개수	이미지 개수 (Batch)	처리량 (Images/sec)	소요시간 (240 Epochs 기준)
GTX 1080	1	16	48.8	24h
Titan X	1	32	40.2	34h
Titan X	2	32	70.8	20h
Titan X	4	32	110	12h

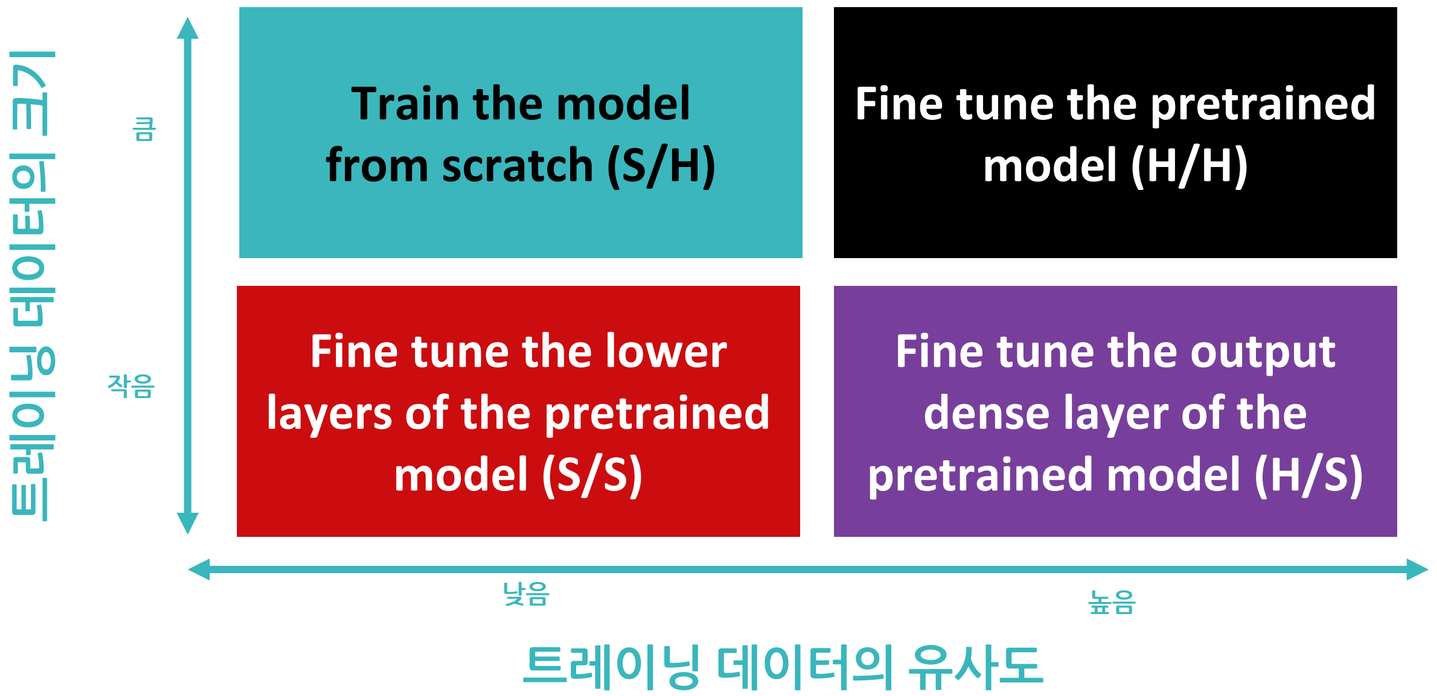
모델의 일부분을 변경했을 때, 다시 전체를 학습하는 과정 필요
좋은 컴퓨터 (워크스테이션)을 사용해도 많은 시간이 소요

전이학습의 동작 과정



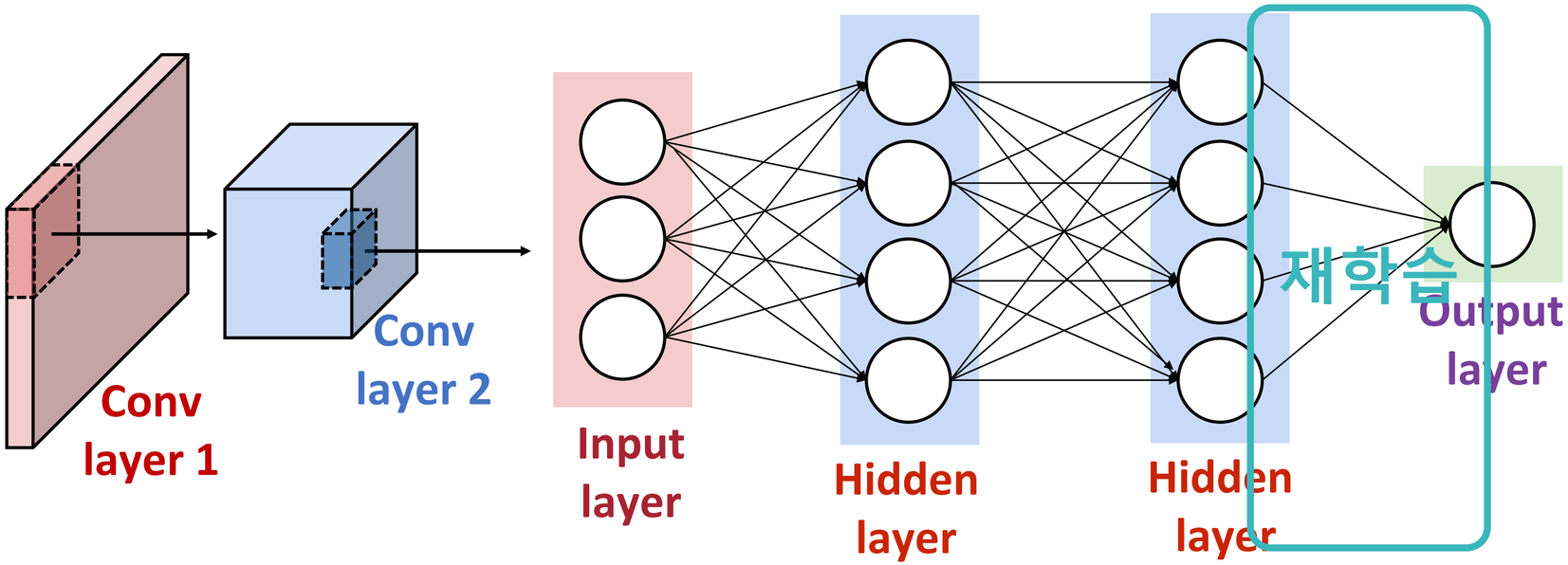
빅 데이터 (학습 데이터 셋)을 사용하여 초기 신경망 모델 학습
초기 신경망 모델에 사용자 데이터를 사용하여 학습

전이학습 학습 방법



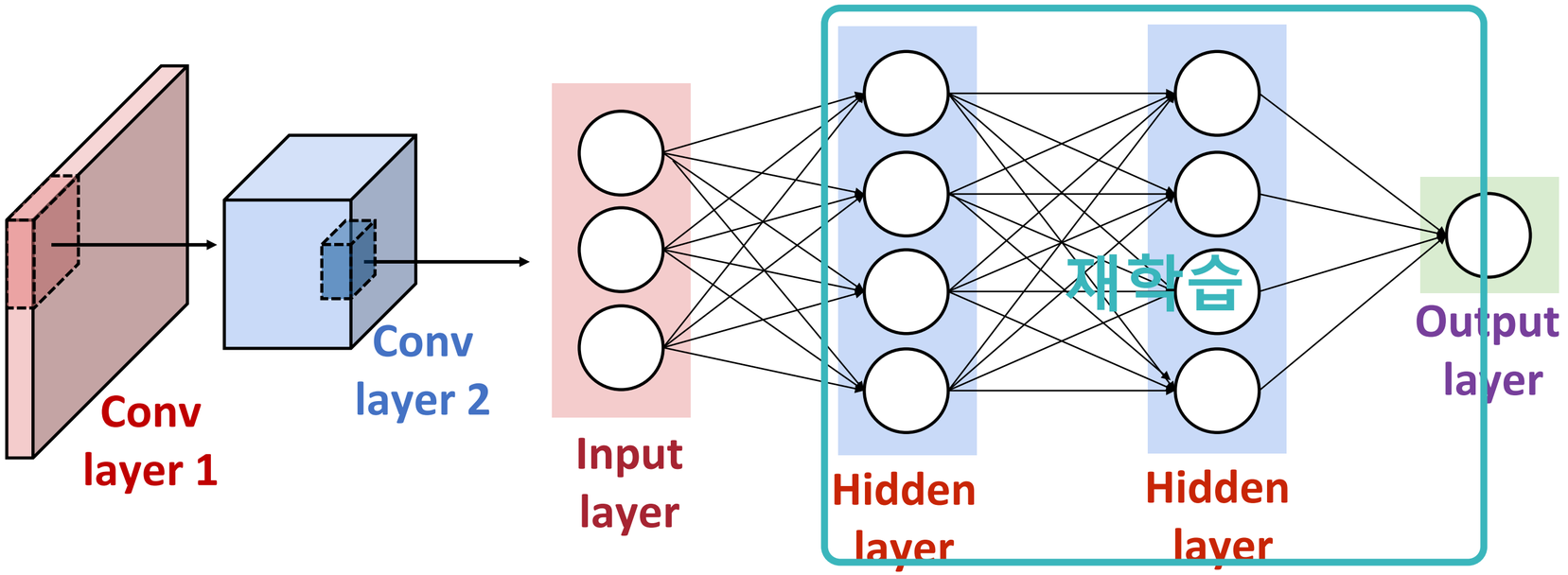
새로 학습에 사용할 데이터 크기, 유사성에 의해 구분

Fine tune the output dense layer (H/S)



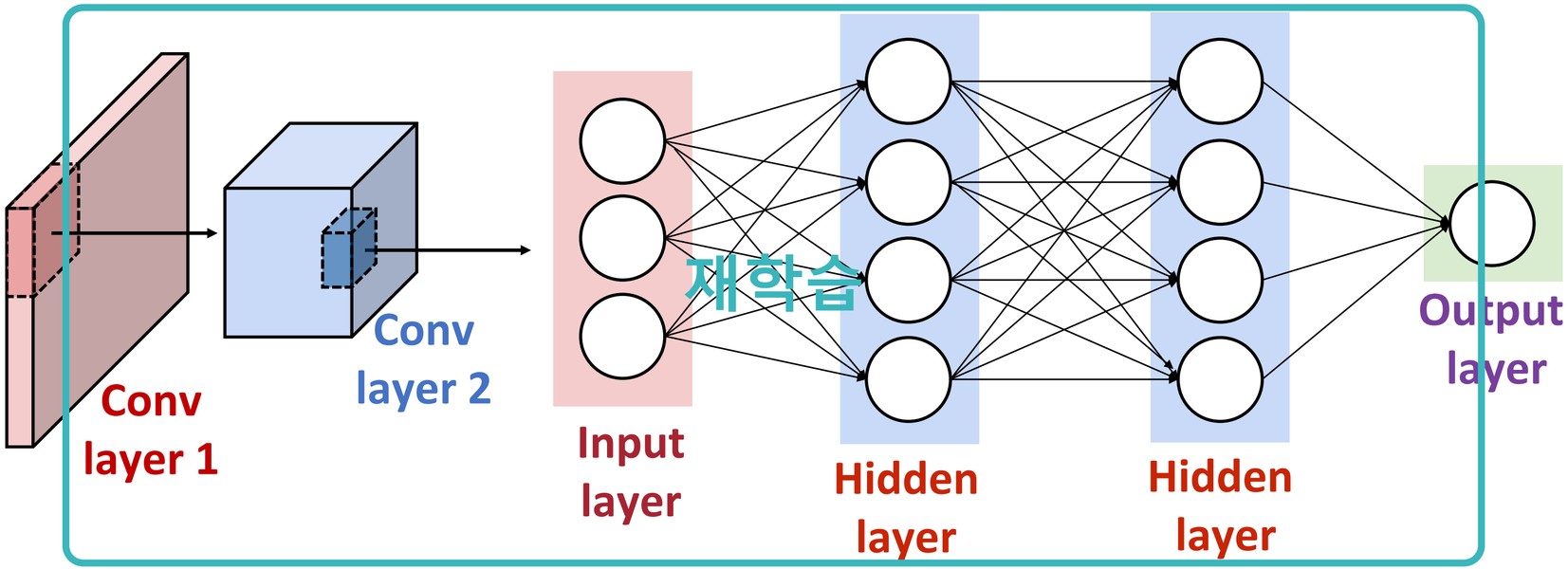
유사성이 높고, 데이터 셋의 개수가 작은 경우
유사성을 활용하여 문제 해결 가능 !
분류기의 마지막 레이어를 재학습

Fine tune the lower layers (S/S)



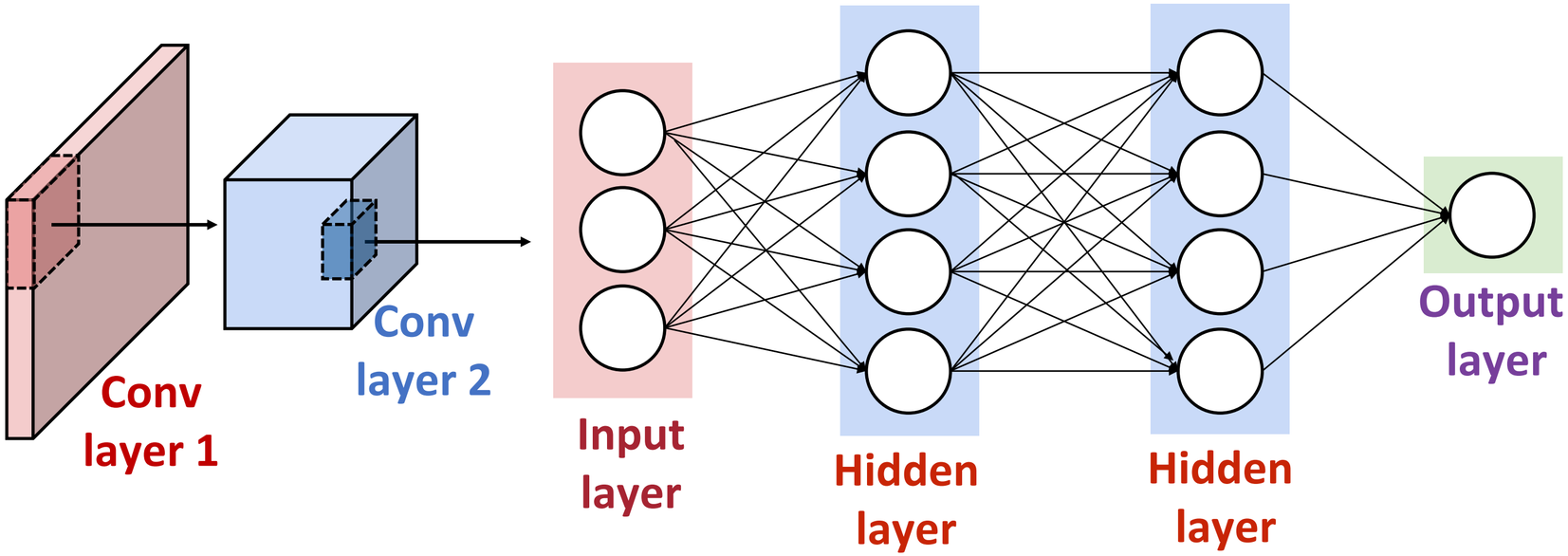
유사성이 낮고 데이터 셋의 개수가 적은 경우
학습된 모델의 특징 추출기를 그대로 사용
분류기 전체를 재학습

Fine tune the output dense layer (S/H)



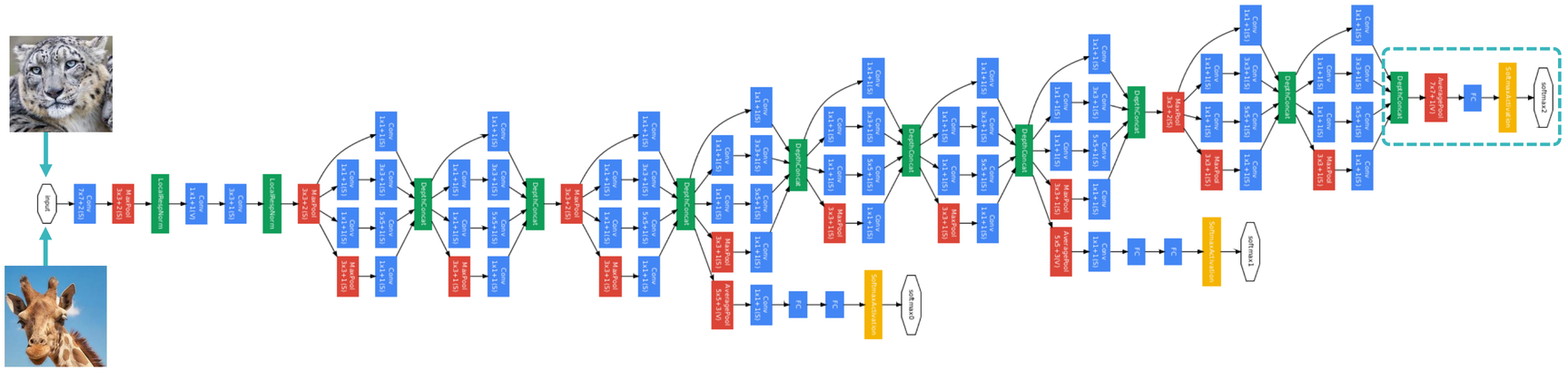
유사성은 낮지만, 데이터의 개수가 많은 경우
전이학습을 하기 보다는 전체를 다시 학습하는 것이 유리함

Fine tune the pretrained model (H/H)



유사성이 높고 데이터의 개수가 많은 경우
전이학습을 통해 얻는 성능이 가장 극대화 되는 경우
사전에 학습된 모델을 사용하여 전이학습을 진행

GoogLeNet을 활용한 전이학습



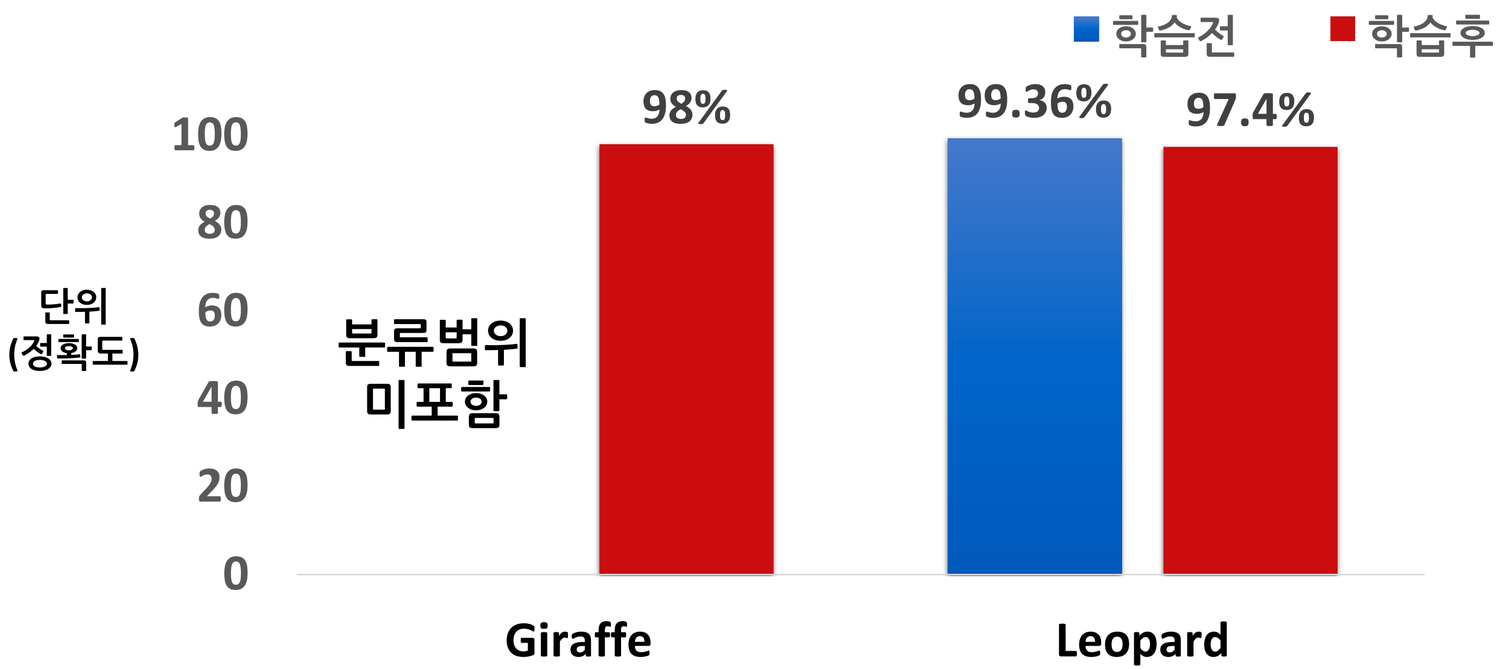
CNN (ImageNet 학습)에 새로운 데이터 (Leopard, Giraffe) 학습

Leopard (유사성 높음), Giraffe (유사성 없음)

Leopard (415개), Giraffe (394개)를 학습과 검증에 사용

80% 학습, 20% 검증

GoogLeNet을 활용한 전이학습 결과

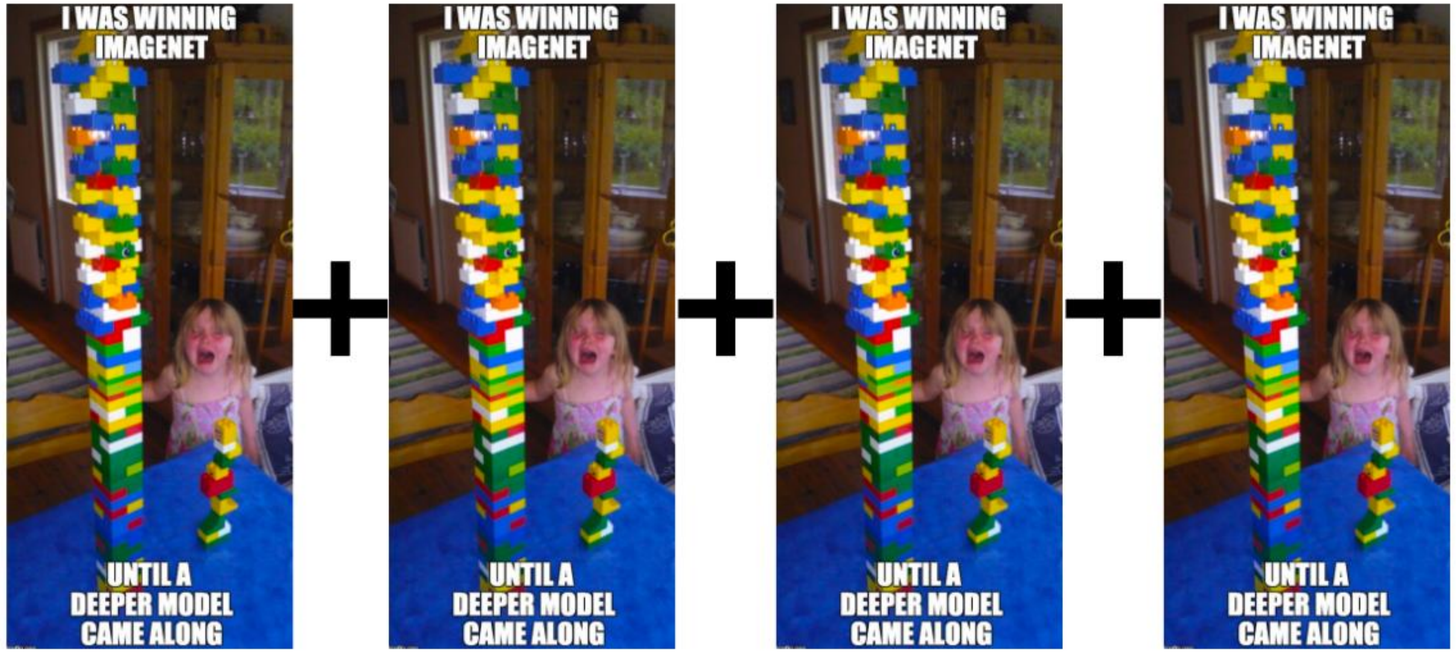


Batch (64)에서 학습에 500번의 Iteration이 소요됨
약 30초 (Ryzen7, 8C), 5분 (라즈베리파이 3)

전이학습 방법

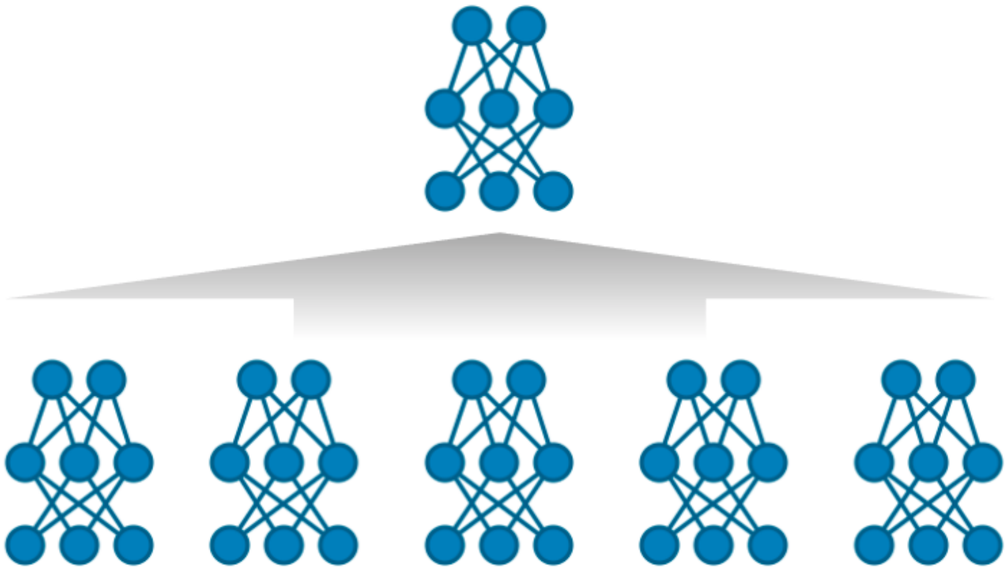
Knowledge Distillation (지식 증류)

Ensemble



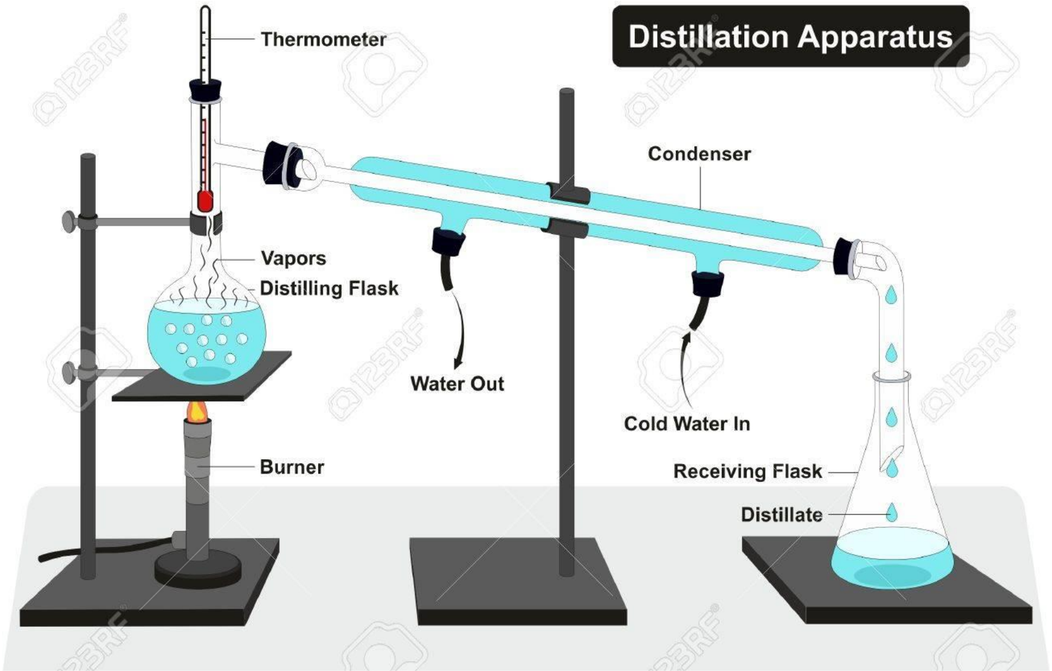
동일한 구조에서 Initialization을 다르게 여러 번 학습하거나
다른 구조의 NN을 학습하고 나오는 결과물을 합치는 과정
여러 모델을 돌리기 위한 컴퓨팅 자원이 필요

Ensemble



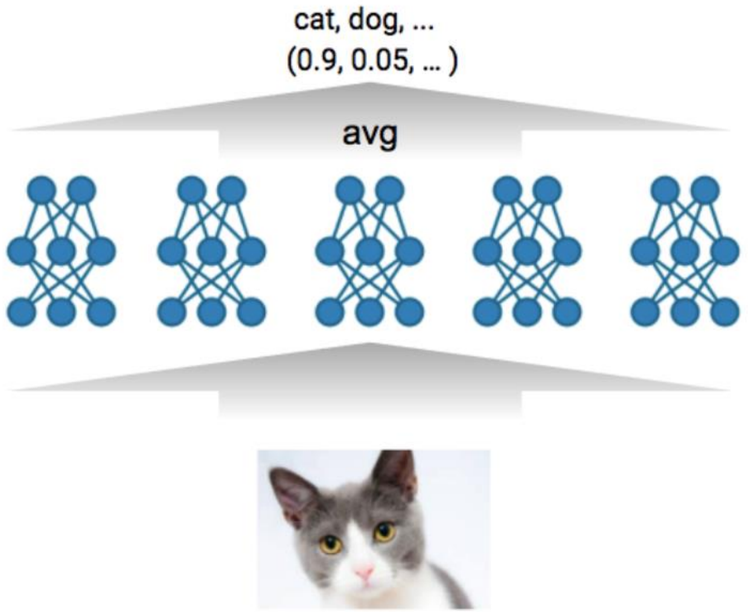
Generalization 능력을 가진 신경망이 Single Neural Network에게 학습한 지식을 전달 (Transfer) 할 수 있다면 ?

Distilling the Ensemble Knowledge



증류 (혼합물에서 특정 성분을 분리시키는 방법)
신경망에의 지식 증류는 불필요한 파라미터가 사용되는 Ensemble 모델에서 Generalization 성능을 향상 할수 있는 지식들을 분리하는 것

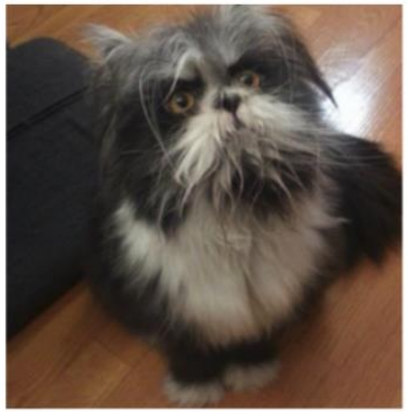
Distilling the Ensemble Knowledge



Softmax output from the model ensembles

이미지 분류에서 최종 출력은 Softmax 함수를 사용
Softmax 함수는 여러 카테고리에 대한 합이 1이 되는 확률 사용
Ensenble 모델의 Softmax 결과를 새로운 NN이 잘 전달 받는다면
새로운 NN가 학습에 활용하여 기존과 비슷한 성능 가능

Softmax 결과 = 지식 = Soft Label



dog

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

Comparison with the 'hard label' and the 'soft label'

학습 과정은 Softmax 출력이 정답 (Label)과 최대한 비슷해 지도록 Softmax Cross-entropy Loss를 최소화 하는 방식으로 학습
학습된 신경망에는 많은 정보가 담겨져 있음

Softer Softmax

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

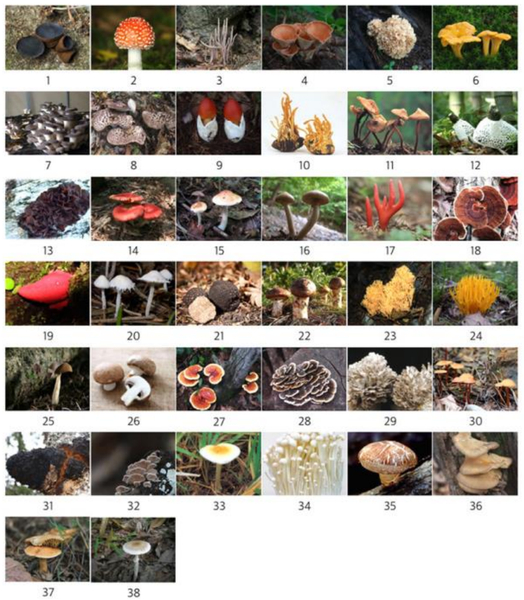
Softmax with
temperature T

Temperature라는 파라미터 T를 추가하여, T가 높을수록
기존보다 더 Soft한 확률 분포를 얻을 수 있도록 함
불순물을 가열하여 물질을 추출하는 과정에서 온도를 조절

전이학습 적용 사례

비전과 자연어 처리에 적용된 케이스

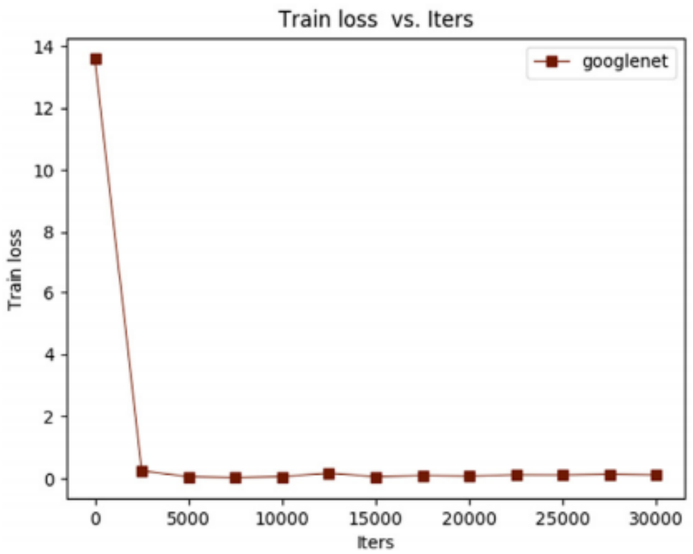
전이학습: Vision

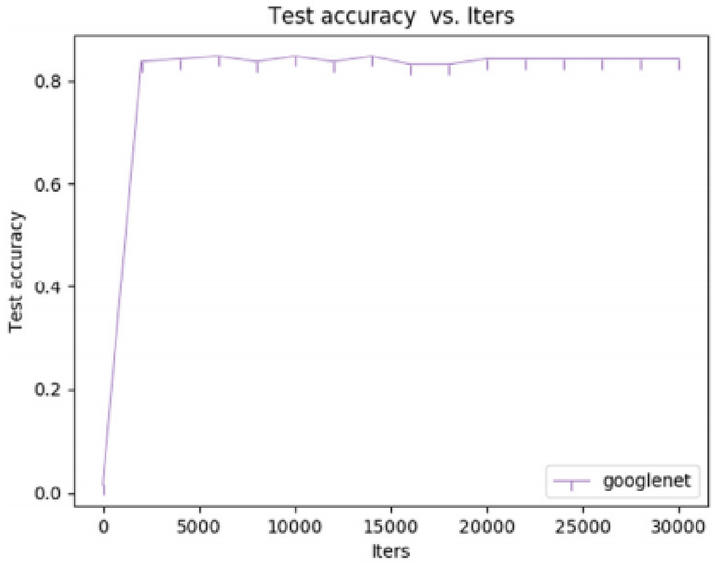


버섯의 종류를 인식하는 신경망[1]
학습된 신경망 (VGG-16, AlexNet)에 38종의 버섯 데이터 학습
1,478장의 데이터에서 1,288장 (학습), 190 (테스트)

[1] 컨볼루션 신경망과 전이 학습을 이용한 버섯 영상 인식, 정보과학회 컴퓨팅의 실제 논문지, 2018

전이학습: Vision





분류 범위에 없기 때문에 처음에는 정확도가 낮지만
2epochs 정도로 높은 정확도 가능

전이학습: NLP

인공지능 AI 언어모델 BERT

(Bidirectional Encoder Representations from Transformers)

전이학습을 사용하여 자연어 처리 모델을 학습

전이학습: NLP

위키피디아

책

학습된 모델
(대용량 Unlabeled data)

스탠포드 Q&A

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

전이학습
(특정 Task를 가지고 있는
Labeled data)

인공지능 AI 언어모델 BERT

(Bidirectional Encoder Representations from Transformers)

전이학습을 사용하여 자연어 처리 모델을 학습

EfficientNet

Table 5. **EfficientNet Performance Results on Transfer Learning Datasets**. Our scaled EfficientNet models achieve new state-of-the-art accuracy for 5 out of 8 datasets, with 9.6x fewer parameters on average.

	Comparison to best public-available results						Comparison to best reported results					
	Model	Acc.	#Param	Our Model	Acc.	#Param(ratio)	Model	Acc.	#Param	Our Model	Acc.	#Param(ratio)
CIFAR-10	NASNet-A	98.0%	85M	EfficientNet-B0	98.1%	4M (21x)	†Gpipe	**99.0%**	556M	EfficientNet-B7	98.9%	64M (8.7x)
CIFAR-100	NASNet-A	87.5%	85M	EfficientNet-B0	88.1%	4M (21x)	Gpipe	91.3%	556M	EfficientNet-B7	**91.7%**	64M (8.7x)
Birdsnap	Inception-v4	81.8%	41M	EfficientNet-B5	82.0%	28M (1.5x)	GPipe	83.6%	556M	EfficientNet-B7	**84.3%**	64M (8.7x)
Stanford Cars	Inception-v4	93.4%	41M	EfficientNet-B3	93.6%	10M (4.1x)	‡DAT	**94.8%**	-	EfficientNet-B7	94.7%	-
Flowers	Inception-v4	98.5%	41M	EfficientNet-B5	98.5%	28M (1.5x)	DAT	97.7%	-	EfficientNet-B7	**98.8%**	-
FGVC Aircraft	Inception-v4	90.9%	41M	EfficientNet-B3	90.7%	10M (4.1x)	DAT	92.9%	-	EfficientNet-B7	**92.9%**	-
Oxford-IIIT Pets	ResNet-152	94.5%	58M	EfficientNet-B4	94.8%	17M (5.6x)	GPipe	**95.9%**	556M	EfficientNet-B6	95.4%	41M (14x)
Food-101	Inception-v4	90.8%	41M	EfficientNet-B4	91.5%	17M (2.4x)	GPipe	93.0%	556M	EfficientNet-B7	**93.0%**	64M (8.7x)
Geo-Mean						**(4.7x)**						**(9.6x)**

†GPipe (Huang et al., 2018) trains giant models with specialized pipeline parallelism library.

‡DAT denotes domain adaptive transfer learning (Ngiam et al., 2018). Here we only compare ImageNet-based transfer learning results.

Transfer accuracy and #params for NASNet (Zoph et al., 2018), Inception-v4 (Szegedy et al., 2017), ResNet-152 (He et al., 2016) are from (Kornblith et al., 2019).

논문에서는 정말 학습이 잘 되는데...

직접 실험을 했을 때는 ?

모델	신경망 종류	학습 종류	정확도
Tiny- darknet	경량 신경망	사전학습	51.19%
		전이학습	65.02%
ResNet-50	비경량 신경망	사전학습	41.78%
		전이학습	71.58%

신경망의 종류, 신경망의 특성에 따라 전이학습의 효과가 상이함
그래도 전이학습을 사용하면 유사한 문제 해결에 대처 가능

/* elice */

문의 및 연락처

academy.elice.io

contact@elice.io

facebook.com/elice.io

medium.com/elice