

The slide features decorative geometric shapes in the corners. The top-left corner has several overlapping triangles in shades of blue, green, and red. The bottom-right corner has several overlapping triangles in shades of light gray.

Big Transfer(BiT)


General Visual Representation Learning

(<https://arxiv.org/abs/1912.11370>)

한성대학교 1971336 김태민

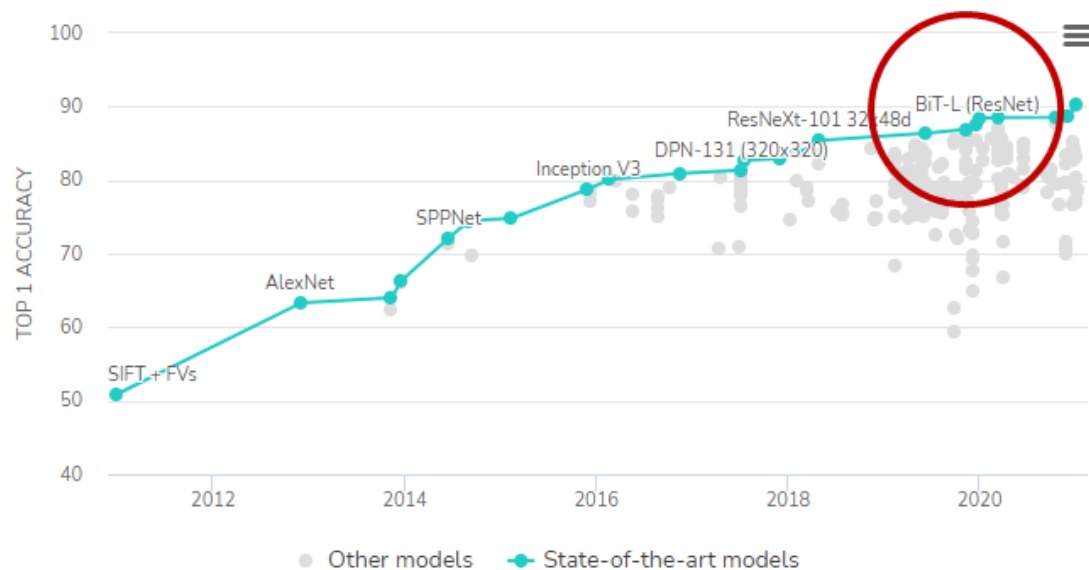


목차

- BiT
 - **Group Normalization + Weight Standardization**
 - HyperRule
 - 결과 분석
- 

BiT

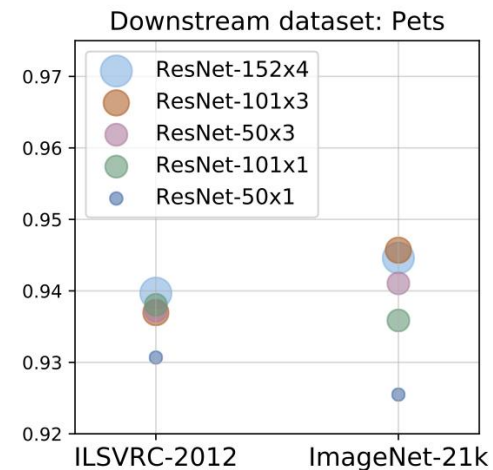
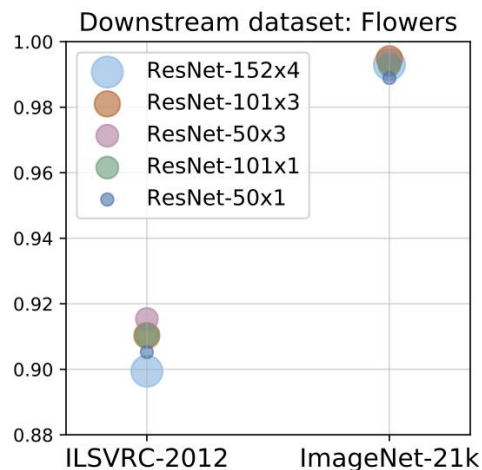
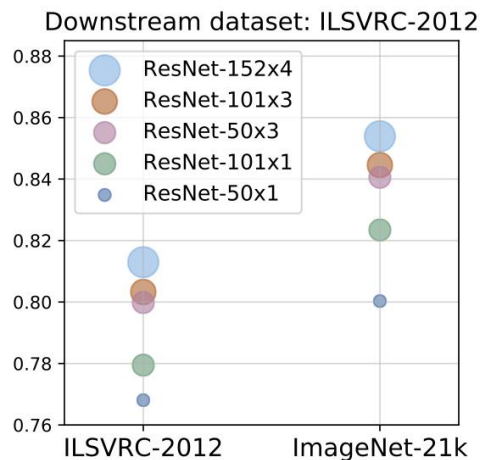
What is BiT



- 2019년도 imagenet에서 1등을 기록한 모델
- 학습으로 JFT dataset을 pre-training
- 20개의 데이터 셋에 대해 fine-tuning
- 모델로 ResNet-152x4를 사용

BiT

What is BiT



- 각종 데이터 셋에 대하여 높은 성능을 달성
- X축은 업스트림 데이터 세트이고 Y축은 다운스트림 작업에 미치는 성능이다.
- 두 번째 그림에서 ILSVRC-2012의 데이터를 학습한 ResNet-152x4보다 ResNet-50x3이 더 높은 성능을 달성
- 결론 : 데이터세트 또는 모델의 크기만 늘렸을 시 결과는 예상과 다를 수가 있다.

BiT

What is BiT

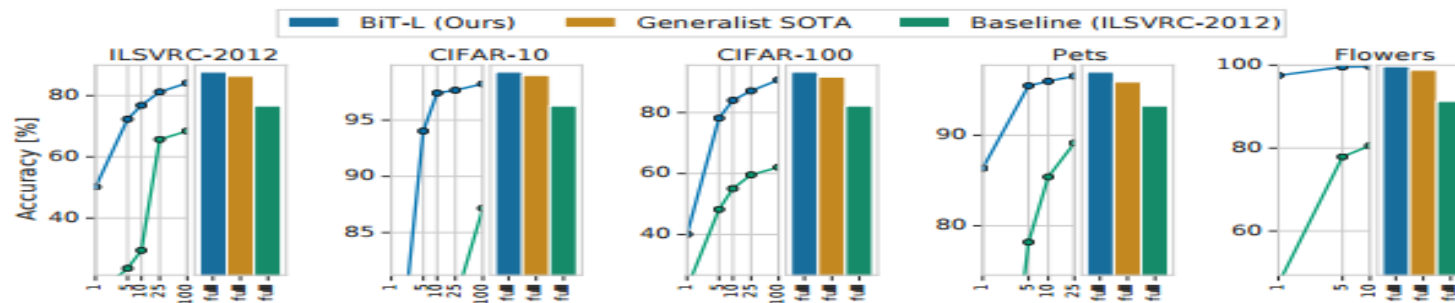
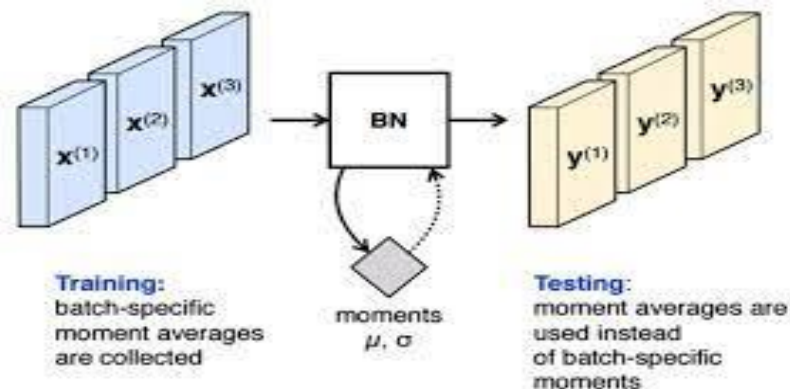


Fig. 1: Transfer performance of our pre-trained model, BiT-L, the previous state-of-the-art (SOTA), and a ResNet-50 baseline pre-trained on ILSVRC-2012 to downstream tasks. Here we consider only methods that are pre-trained independently of the final task (generalist representations), like BiT. The bars show the accuracy when fine-tuning on the full downstream dataset. The curve on the left-hand side of each plot shows that BiT-L performs well even when transferred using only few images (1 to 100) per class.

- 기존의 SOTA와 비교하여도 각종 데이터 셋에 대하여 높은 성능을 달성
- 결과적으로 어떻게 Pre-Training와 Fine-Tuning에 대한 고찰이다.
- 전이 학습에 대해 좋은 참고 자료이다.

Group Normalization + Weight Standardization

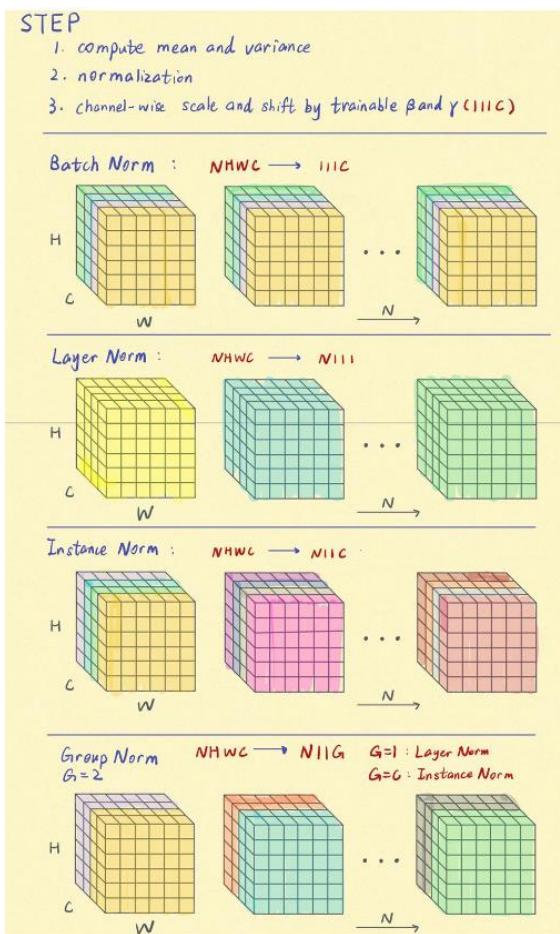
The problem with batch normalization



- BN 사용 시 모델의 크기에 따라 GPU에 들어갈 이미지 수가 적기 때문에 적은 양의 배치 사이즈를 사용
- 적은 양의 배치사이즈를 사용할 경우 배치사이즈의 영향을 받는BN의 성능 악화
- Transfer에는 통계량을 업데이트하기 때문에 안 좋은 영향

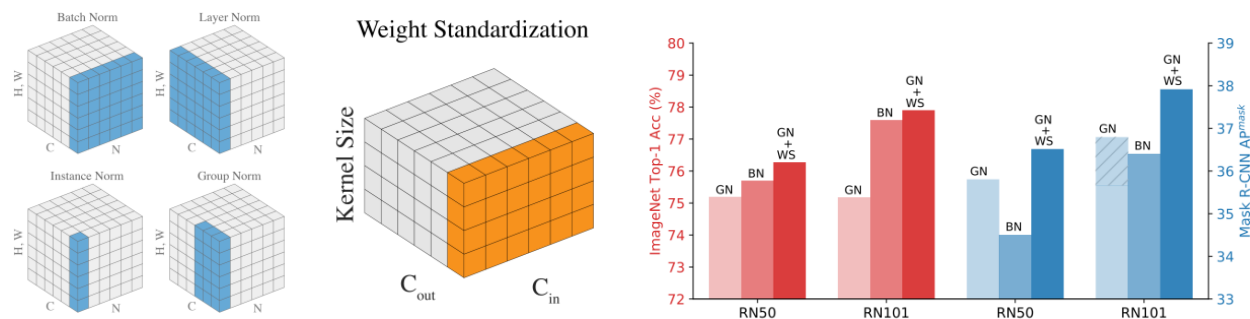
Group Normalization + Weight Standardization

Replace BN with GN+WS



N = batch size H,W,C = image


- Batch Norm
- 미니 배치 전체에 걸쳐 feature의 특정 채널을 모두 합쳐 normalize함
- Group Norm
- 개별 데이터에서 나온 feature의 채널들을 N개의 그룹으로 묶어 normalize함
[Group Normalization\(Yuxin Wu, Kaiming He\)](#)
- Weight Standardization
- 단순 하게 weight(convolution filter)을 대상을 normalization을 수행
- (Filter의 mean을, 0 variance를 1로 조정)
- [Micro-Batch Training with Batch-Channel Normalization and Weight Standardization](#)





Hyper Rule

Upstream Pre-Training

- 모델 : ResNet-v2
 - 모델 크기 : ResNet 152x4d
 - Optimizer : SGD with momenuum ($Lr = 0.03, m=9$)
 - 학습 데이터 :
 - BiT-S – 130만장 (ILSVRC-2012)
 - BiT-M – 1400만장 (ImageNet-21k)
 - BiT-L - 3억 장 (JFT)
 - 이미지 크기 : 224x224
 - Warm-up : 5000 steps(매 스텝마다 lr 을 batch size / 256 씩 곱함)
 - Batch Size : 4090 (사용 규모 TPU 512장 (chip 당 8 이미지 학습))
- 



Hyper Rule

Downstream Fine-Training

*Small task = 2만 장 / Medium task = 50만장 미만 / Large task = 50만장 이상

- Optimizer : SGD with momentum (Lr = 0.003,m=0.9)
- Batch Size = 512
- 이미지 크기 :
 - 96x96 이하 = 160x160변환 후 128 crop
 - 448x448 이상 = 384 crop
- Fine-Tuning Schedule : 공통 - 30, 60, 90%에서 Lr decay (10 factor)
 - S task - **500 steps** (최소 약 12.8 epoch)
 - M task - **1만 steps** (최소 약 10.2 epoch)
 - L task - **2만 steps** (최대 20.5 epoch)
- Mixup 유무 : M과 L task에서만 Mixup 사용(alpha = 0.1)

결과 분석

BiT 결과 분석

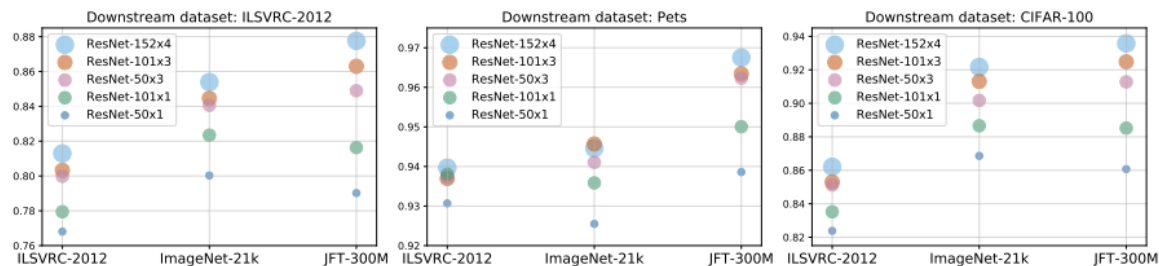


Fig. 5: Effect of upstream data (shown on the x-axis) and model size on downstream performance. Note that exclusively using more data or larger models may hurt performance; instead, both need to be increased in tandem.

- 확실히 성능이 입증되었으며 결론으로 큰 모델에 작은 데이터를 적용시 역효과가 발생 가능
- 반대로 작은 모델에 큰 데이터도 좋지 않다.
- 큰 데이터 + 큰 모델 사용 시 확실히 성능을 보장

결과 분석

BiT 결과 분석

Table 1: Top-1 accuracy for BiT-L on many datasets using a single model and single hyperparameter setting per task (BiT-HyperRule). The entries show median \pm standard deviation across 3 fine-tuning runs. Specialist models are those that condition pre-training on each task, while generalist models, including BiT, perform task-independent pre-training. (*Concurrent work.)

	BiT-L	Generalist SOTA	Specialist SOTA
ILSVRC-2012	87.54 \pm 0.02	86.4 [57]	88.4 [61]*
CIFAR-10	99.37 \pm 0.06	99.0 [19]	-
CIFAR-100	93.51 \pm 0.08	91.7 [55]	-
Pets	96.62 \pm 0.23	95.9 [19]	97.1 [38]
Flowers	99.63 \pm 0.03	98.8 [55]	97.7 [38]
VTAB (19 tasks)	76.29 \pm 1.70	70.5 [58]	-

BiT-L의 성능

Table 4: Top-1 accuracy of ResNet-50 trained from scratch on ILSVRC-2012 with a batch-size of 4096.

	Plain Conv	Weight Std.
Batch Norm.	75.6	75.8
Group Norm.	70.2	76.0

Table 2: Improvement in accuracy when pre-training on the public ImageNet-21k dataset over the “standard” ILSVRC-2012. Both models are ResNet152x4.

	ILSVRC-2012	CIFAR-10	CIFAR-100	Pets	Flowers	VTAB-1k (19 tasks)
BiT-S (ILSVRC-2012)	81.30	97.51	86.21	93.97	89.89	66.87
BiT-M (ImageNet-21k)	85.39	98.91	92.17	94.46	99.30	70.64
Improvement	+4.09	+1.40	+5.96	+0.49	+9.41	+3.77

BiT 종류에 따른 성능

Table 5: Transfer performance of the corresponding models from Table 4 fine-tuned to the 19 VTAB-1k tasks.

	Plain Conv	Weight Std.
Batch Norm.	67.72	66.78
Group Norm.	68.77	70.39

BN,GN,WS에 대한 실험

